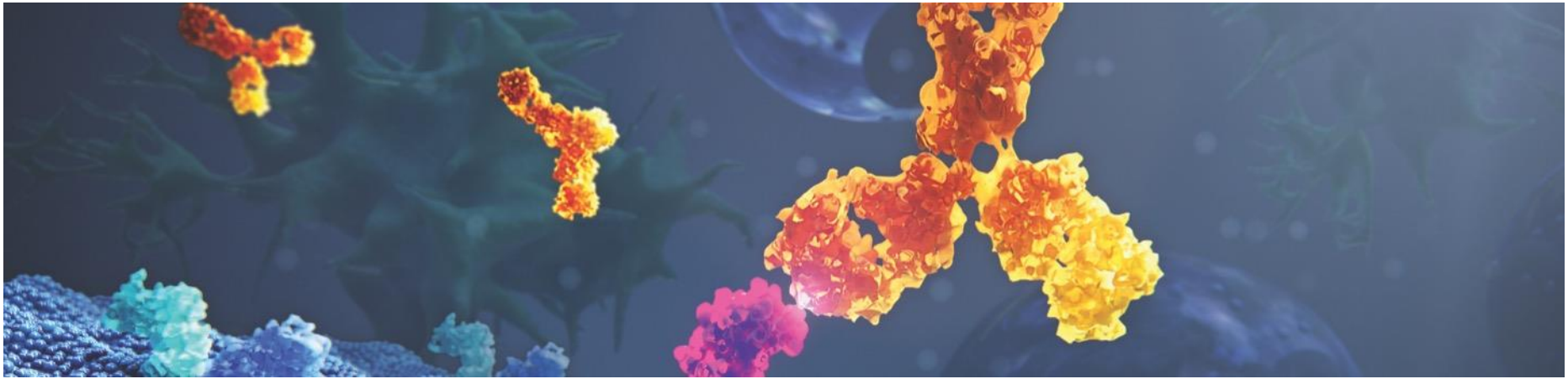
SMILES, RNNs and RDKit, - To the molecular universe and beyond

Esben Jannik Bjerrum, Molecular AI Group

RDKit UGM 2019

Company Restricted

25 – September – 2019

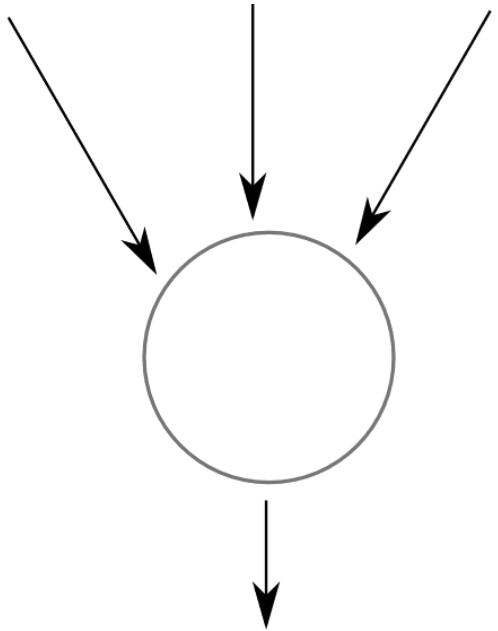


Outline

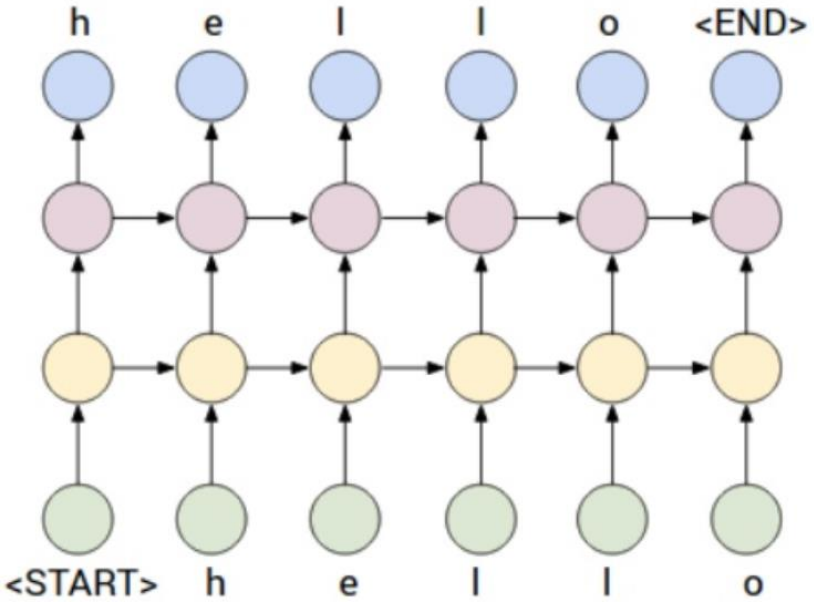
- Artificial Neural Networks 0=>120mph
  - RNNs, SMILES, LSTM
- Combining elements for molecular tasks
  - Encoders (QSAR task)
  - Generators (de novo design task)
  - Autoencoders and Heteroencoders (Both)
- Conditional Recurrent Neural Networks (cRNNs)
  - Generated Molecules
  - Control of Properties
  - Limitiations

Artificial Neurons

Input	0.1	1	-0.9
weights	20	-1	-2
Weighted sum	20*0.1 + -1*1 + -2*-0.9 = 2.8		
Activation Function	Tanh(2.8)=0.992		
Output	0.992		

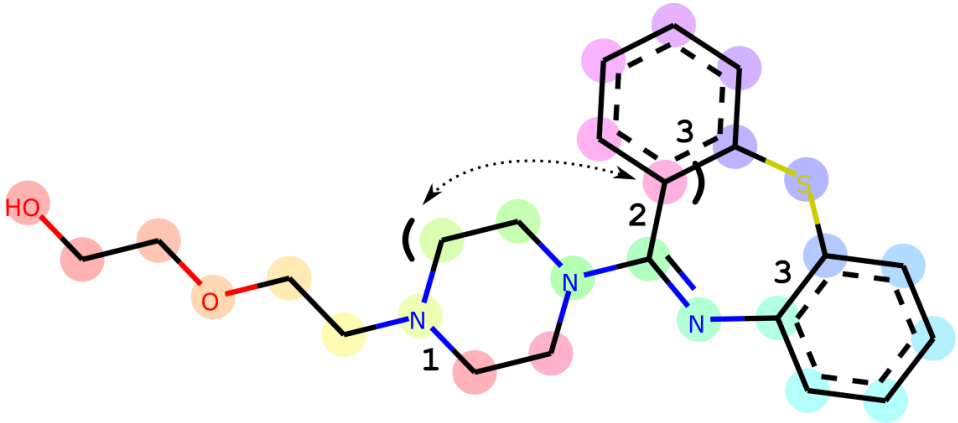


Recurrent Neural Networks (RNN)



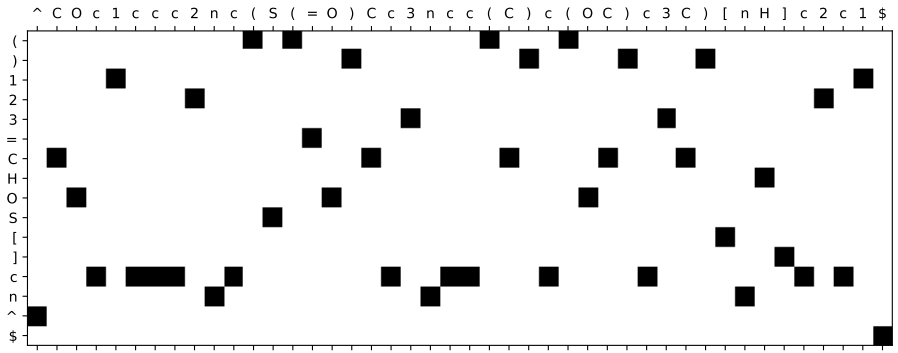
- Sequences of features as inputs
- The same task for every element of a sequence, with the output being affected by the previous computations
- Modeling of sequences such as text, tweets, time series etc.

SMILES, a Chemical Language and Information System

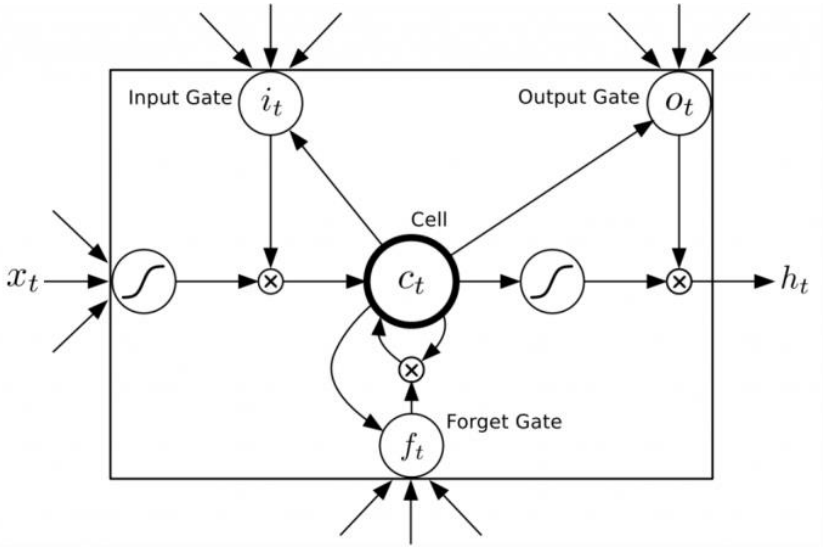


One hot encoding

- Neural Networks learns on vectors, matrices or tensors.
- One-hot encoding with a defined vocabulary converts SMILES strings into 2D matrices



Long Short-Term Memory cells (LSTM)

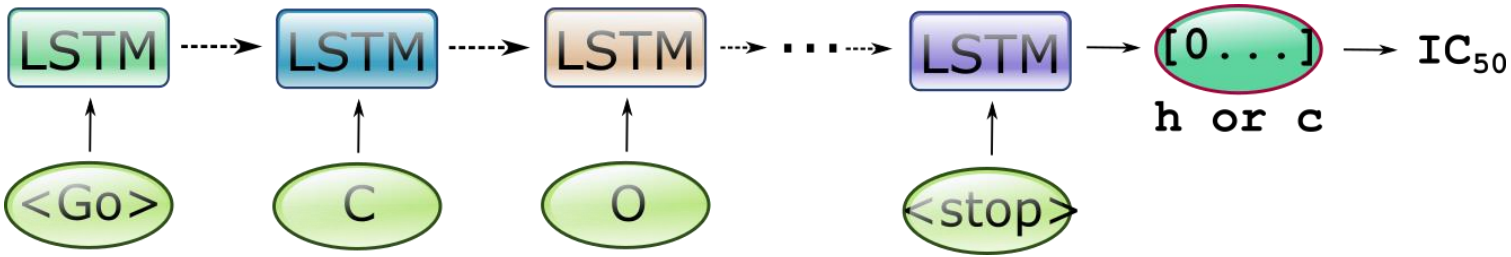


=

LSTM

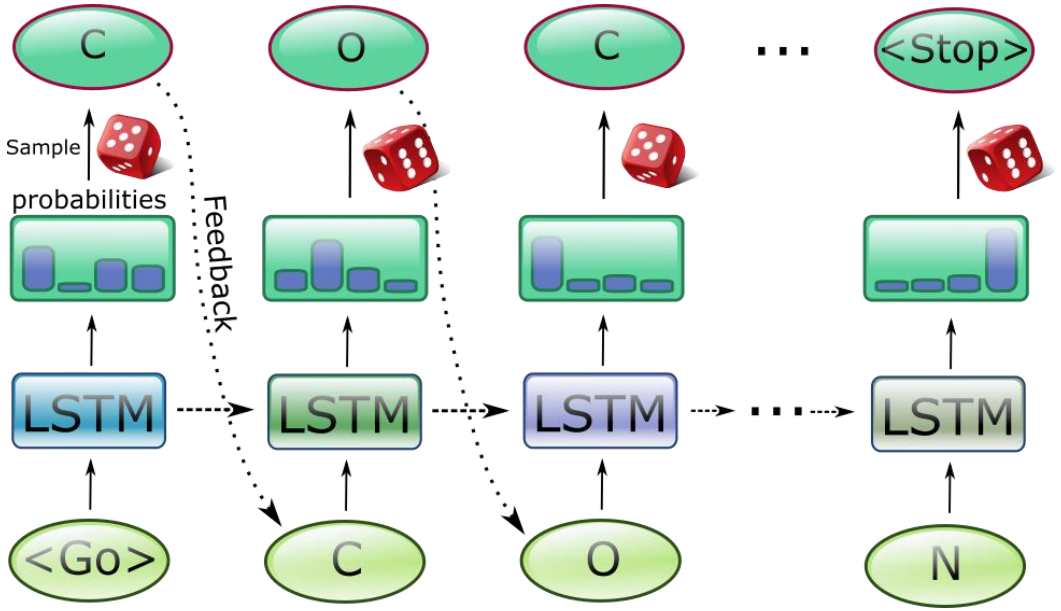
Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Computation*.

RNNs as an encoder

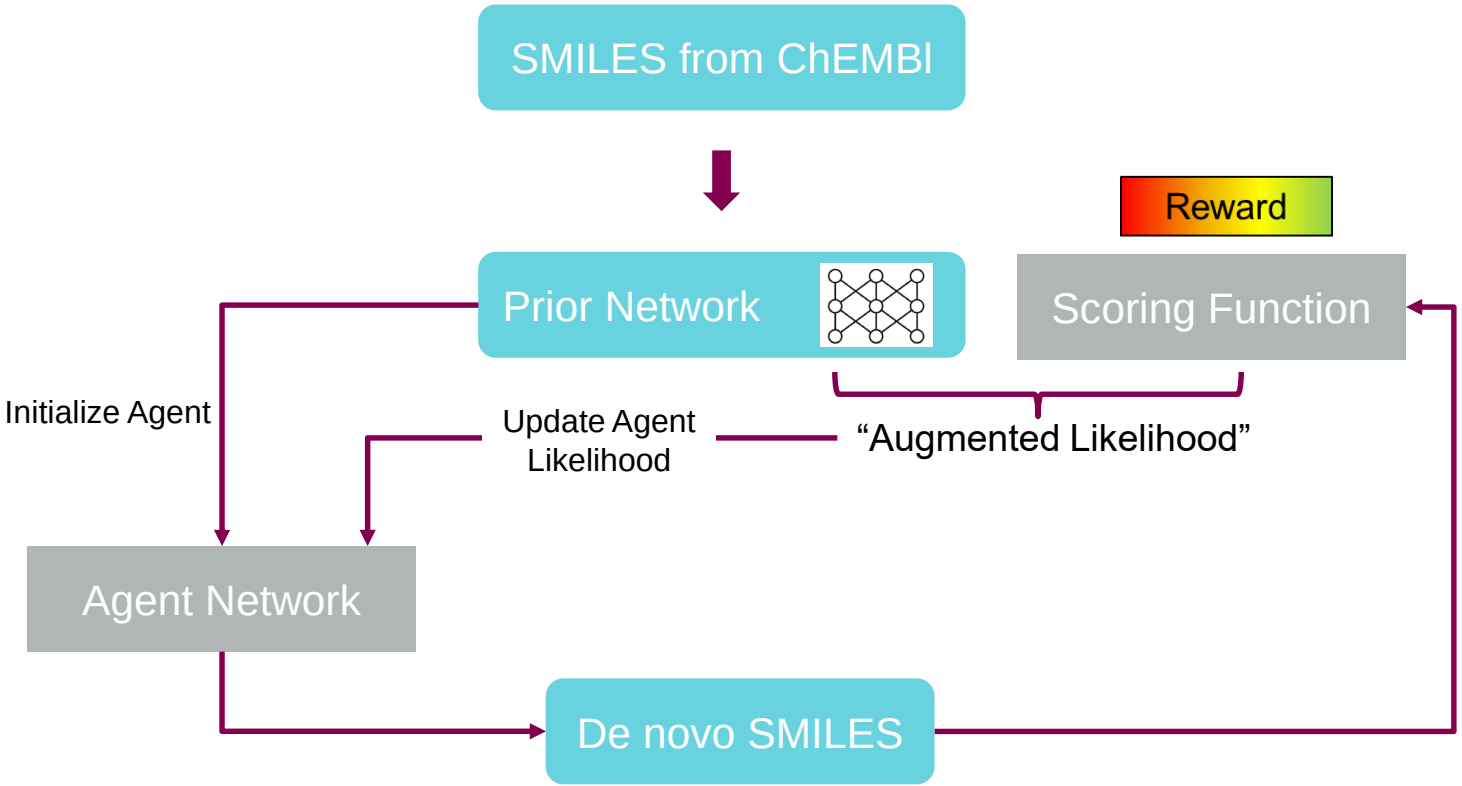


Internal LSTM states gets changed from step to step. The full sequence influences the final vector used for prediction task. QSAR from raw SMILES possible

RNNs as generators

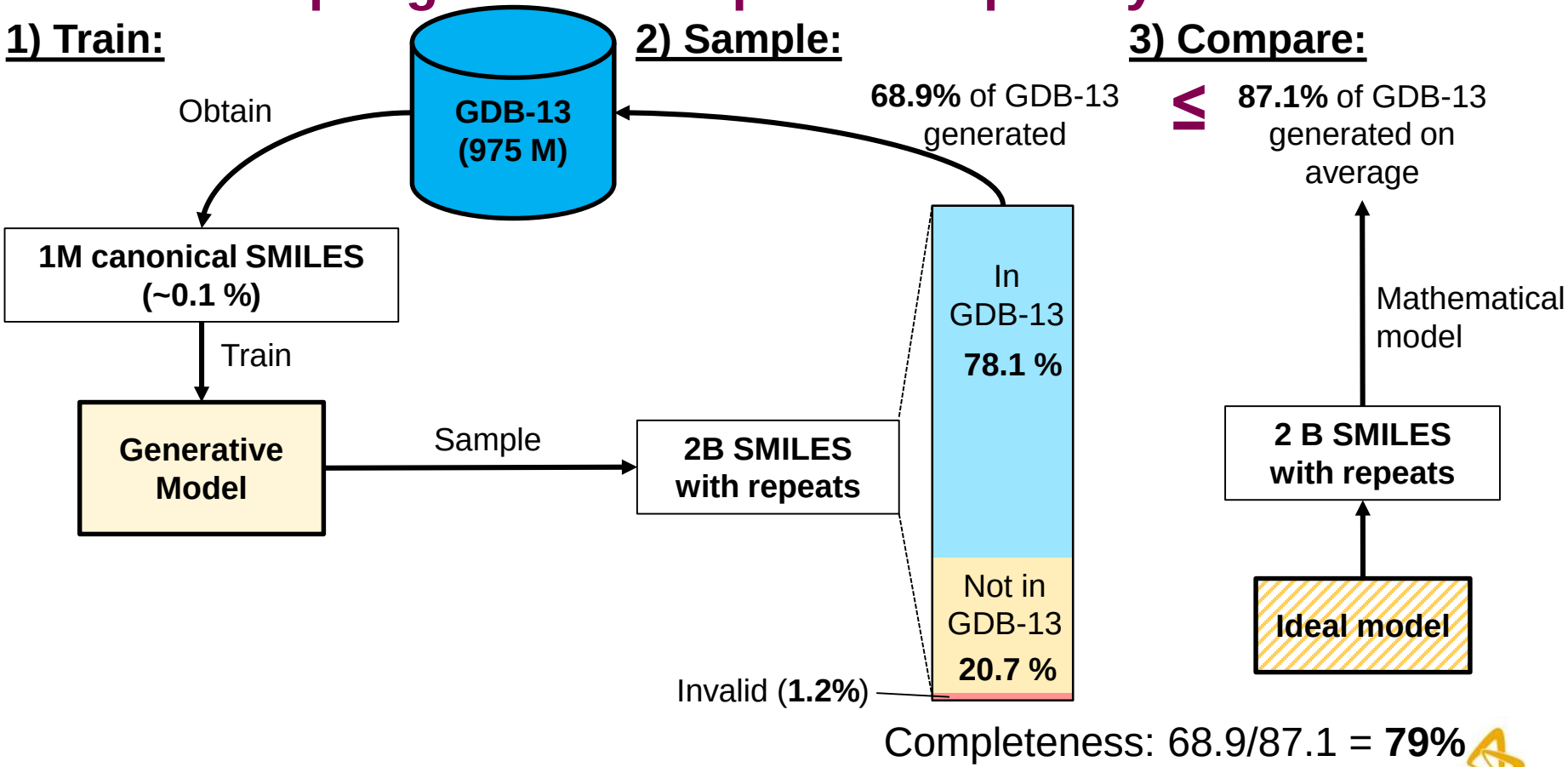


Optimizing molecular generation via reinforcement learning



REINVENT: Olivecrona et. al. Molecular de-novo design through deep reinforcement learning, J. Cheminf. 2017

Are we sampling chemical space completely?

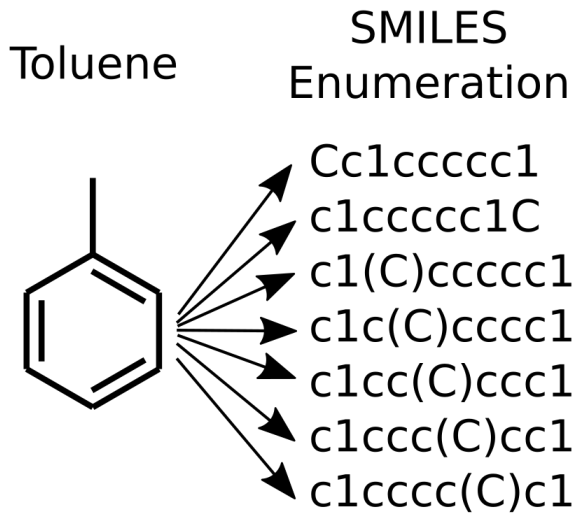


Completeness: 68.9/87.1 = **79%**

11 (1) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J.-L. L.; Chen, H.; Engkvist, O. Exploring the GDB-13 Chemical Space Using Deep Generative Models. *J. Cheminform.* **2019**, *11* (1), 20. https://doi.org/10.1186/s13321-019-0341-z.

Enumeration of non-canonical SMILES

- Canonical SMILES ensures a 1:1 relationship between molecule and SMILES
- I go the other way and generate multiple SMILES for the same molecule
- Works as data augmentation => Improves Deep Learning models



Bjerrum, Esben Jannik. 2017. “SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules.” http://arxiv.org/abs/1703.07076

Random SMILES in practice

```
[1]: from rdkit import Chem
     from rdkit.Chem.Draw import IPythonConsole

[2]: drugname = "Omeprazol"
     mol = Chem.MolFromSmiles("CC1=CN=C(C(=C1OC)C)CS(=O)C2=NC3=C(N2)C=C(C=C3)OC")
     mol
```

[2]:

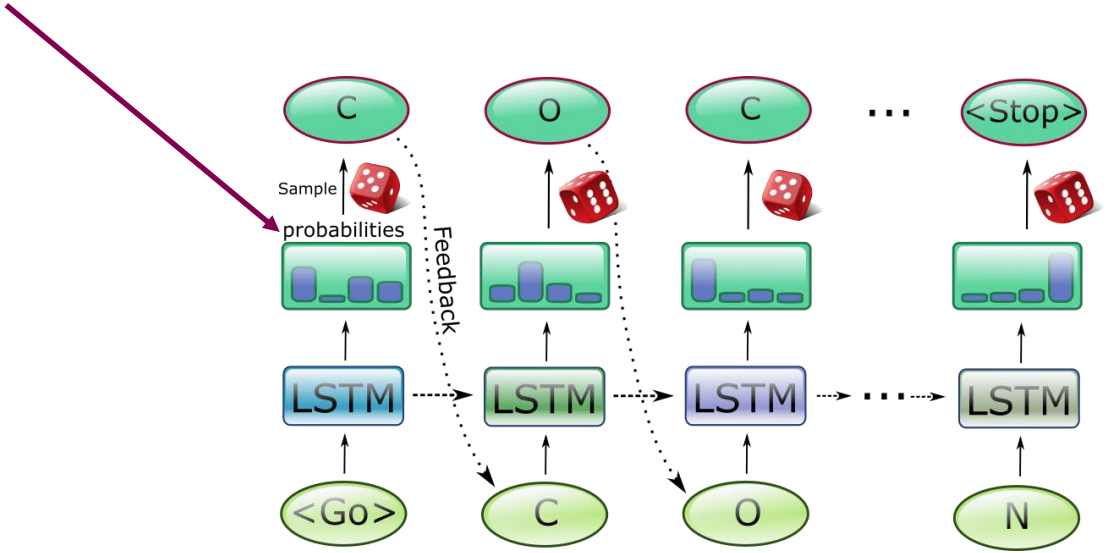

```
[3]: print("%i Atoms, %i Rings"%(mol.GetNumAtoms(), Chem.GetSSSR(mol)))

     24 Atoms, 3 Rings

[ ]: s = set()
     i = 0

     while True:
         l = len(s)

         smiles = Chem.MolToSmiles(mol, doRandom = True)

         s.add(smiles)
         if len(s) > l:
             print("\r %i \t %s"% ( len(s), smiles), end = '')
             i = 0
         else:
             i = i + 1
         if i > 1000:
             break
     print()
     print("Done")

[ ]:
```

- doRandom flag in Chem.MolToSmiles randomizes all decisions during path traversal
- ”Old-style” atom order shuffling gives less SMILES forms

```
ans = list(range(mol.GetNumAtoms()))
np.random.shuffle(ans)
rmol = Chem.RenumberAtoms(mol,ans)
Chem.MolToSmiles(rmol, canonical=False)
```

Finding the sampling probability of a single SMILES

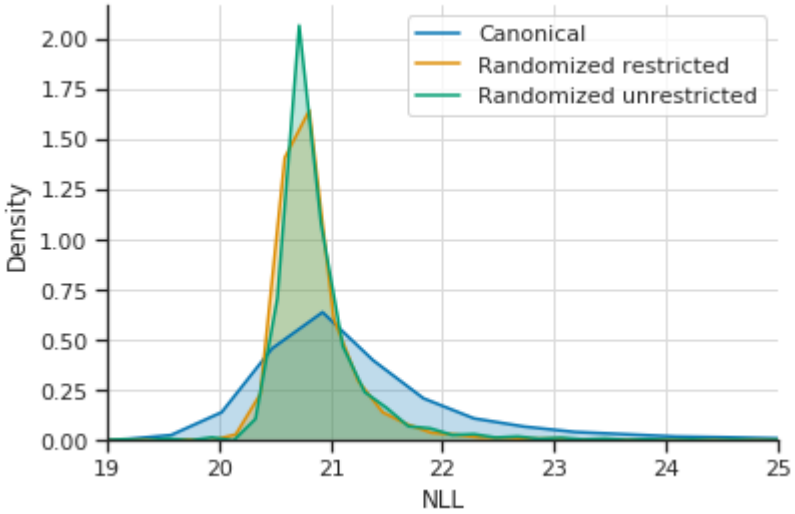
Multiplying the probability for given characters will yield the probability of sampling that exact sequence.



Sum of negative log likelihoods used for numerical stability
Low probability => High NLL

SMILES enumeration increases Chemical Space Coverage

More uniform



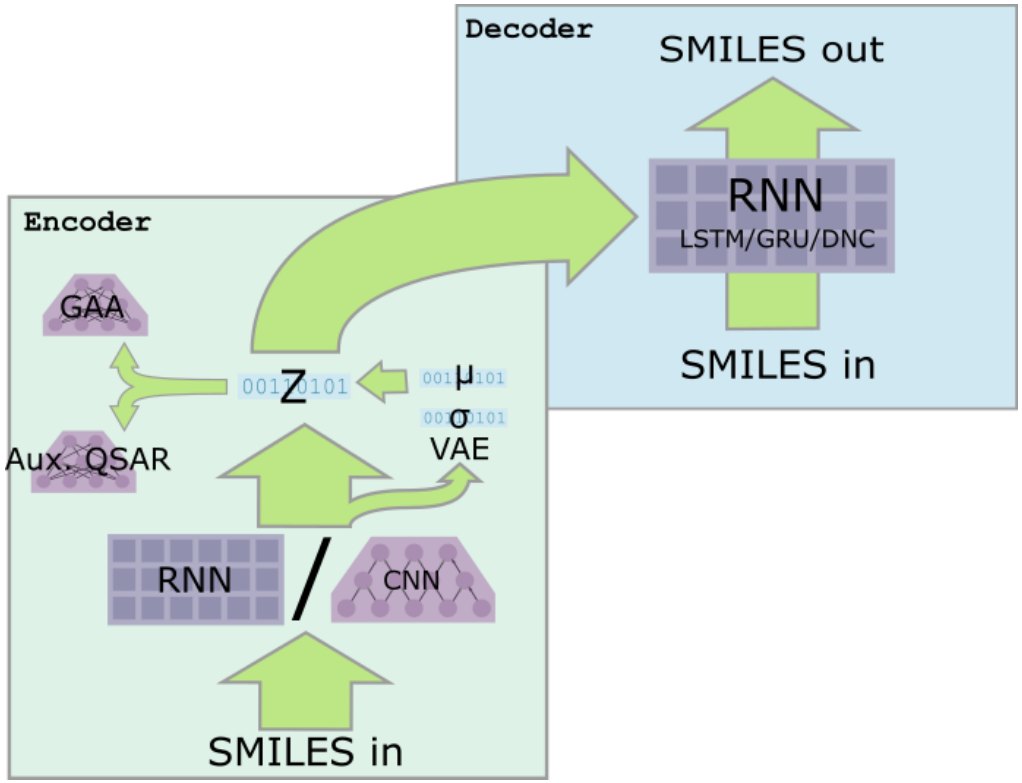
More Complete

Set	SMILES	Validity	Completeness
1M	Canonical	0.994	0.836
	Randomized	**0.999**	**0.953**
10K	Canonical	0.905	0.445
	Randomized	0.974	0.715
1K	Canonical	0.504	0.167
	Randomized	0.812	0.392

GDB-13 is 975 million molecules

Arús-Pous, Josep et al. 2019. https://chemrxiv.org/articles/Randomized_SMILES_Strings_Improve_the_Quality_of_Molecular_Generative_Models/8639942.

SMILES Based Autoencoders



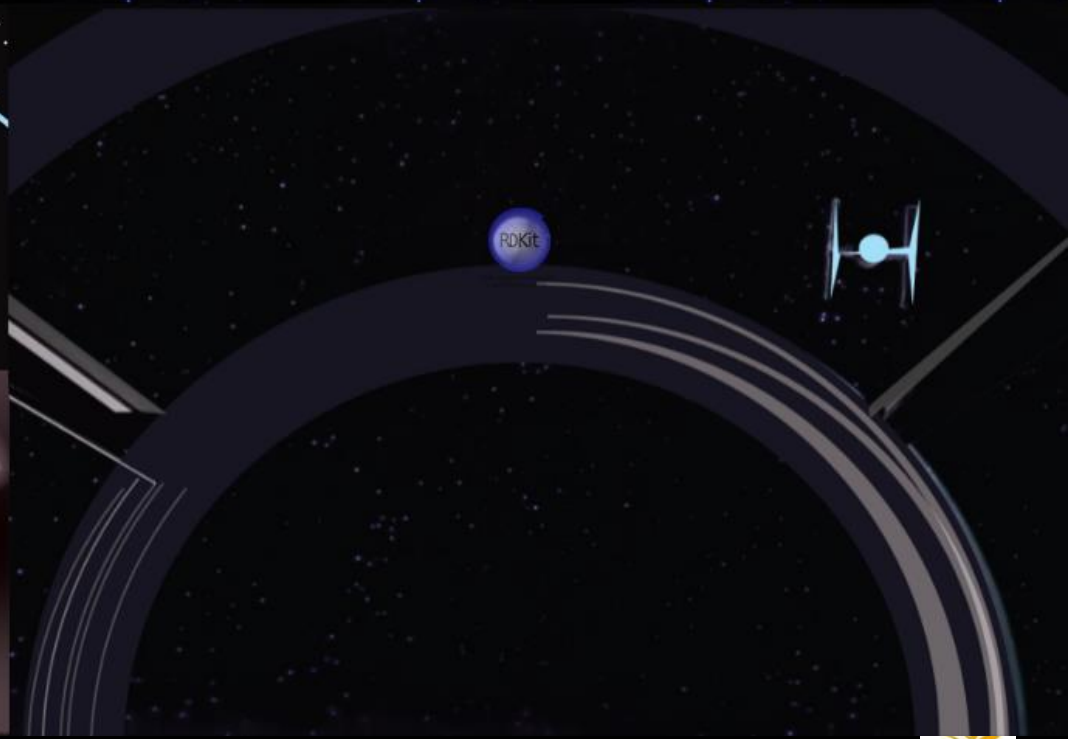
Gómez-Bombarelli, Rafael et al. 2018. "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules." ACS Central Science 4(2): 268–76.

Winter et al. 2018. "Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations." *Chemical Science* 10(6): 1692–1701.

Bjerrum, Esben Jannik, and Boris Sattarov. 2018. "Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders." *Biomolecules*.

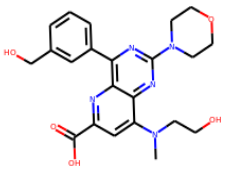
MEANWHILE IN DEEP MOLECULAR LATENT SPACE

DEEP LEARNING REBELS ARE PURSUING A MOLECULE...

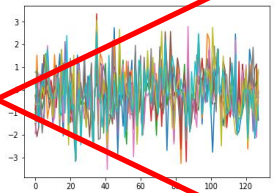


Conditional RNN's

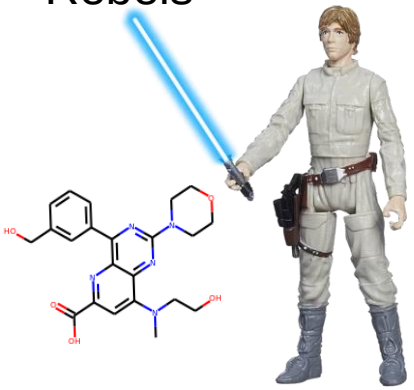
Deep Learning Rebels

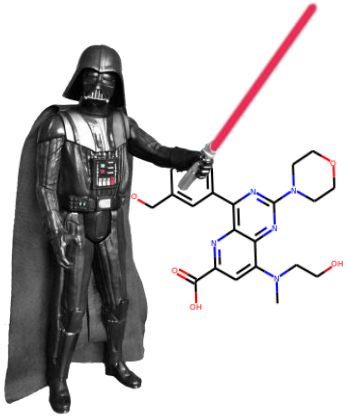


Encoder



Decoder



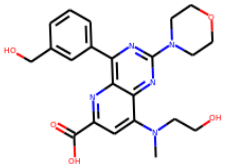


RDKit

Open-Source Cheminformatics and Machine Learning

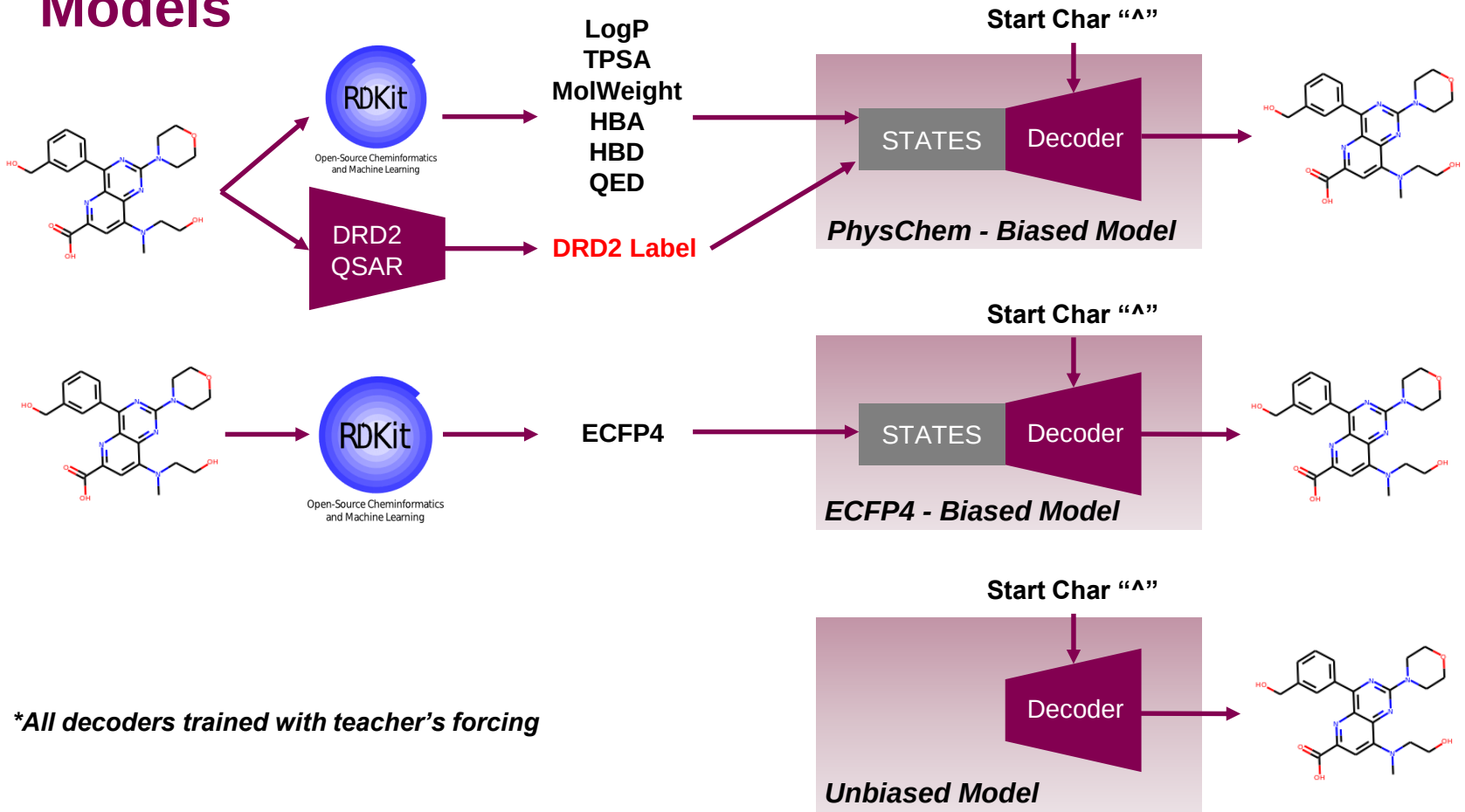
LogP
TPSA
MolWeight
HBA
HBD

Decoder



Darth Cheminformaniac

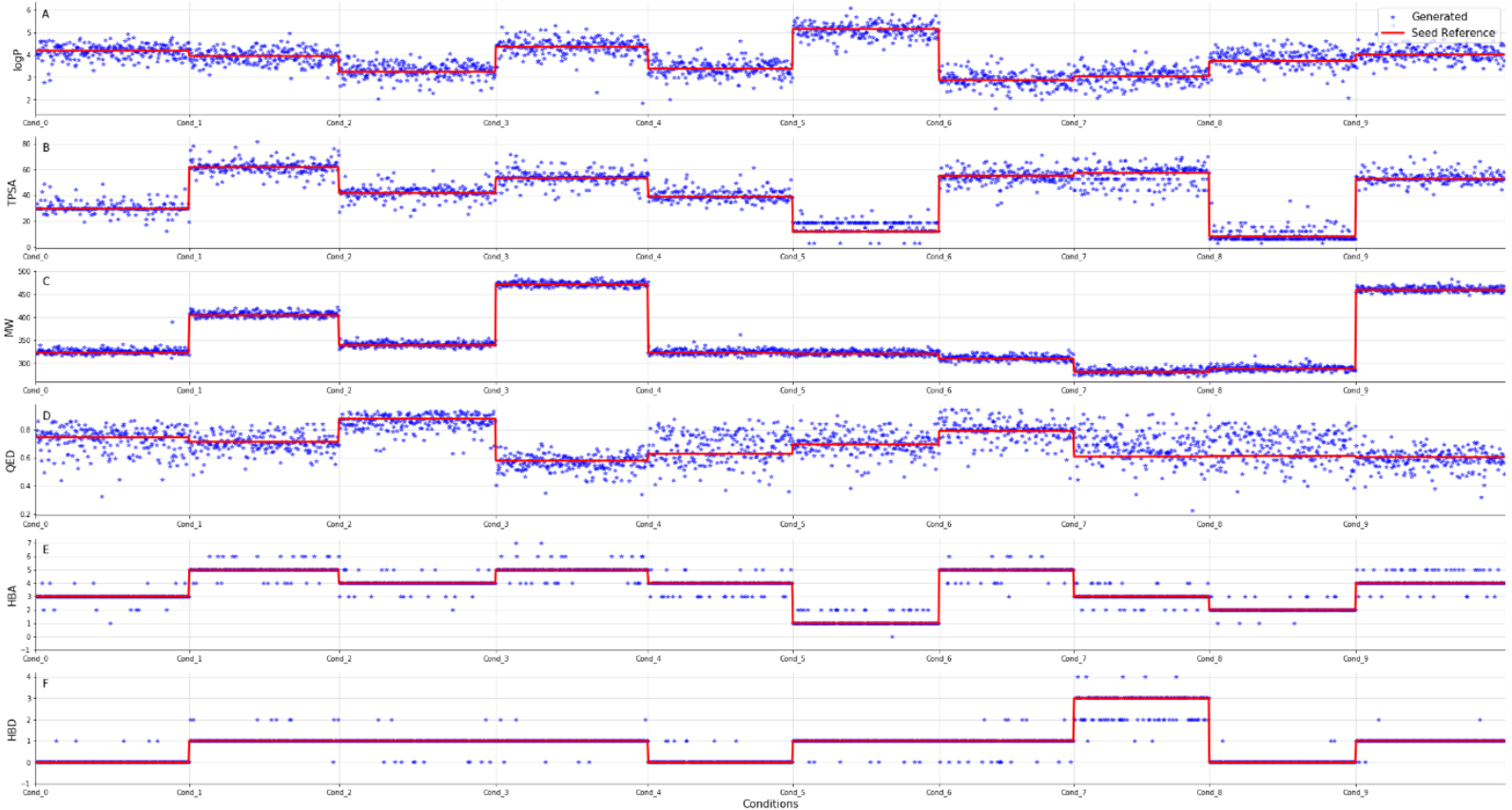
Models



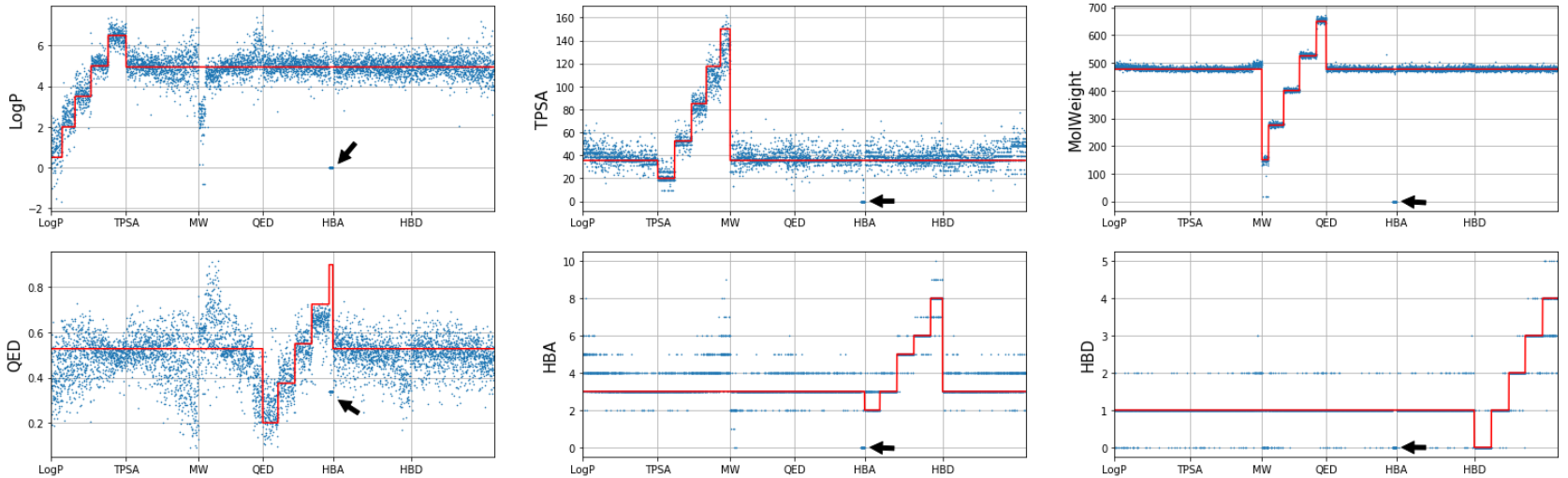
**All decoders trained with teacher's forcing*

https://chemrxiv.org/articles/Direct_Steering_of_de_novo_Molecular_Generation_using_Descriptor_Conditional_Recurrent_Neural_Networks_cRNNs_/9860906

Control of Properties

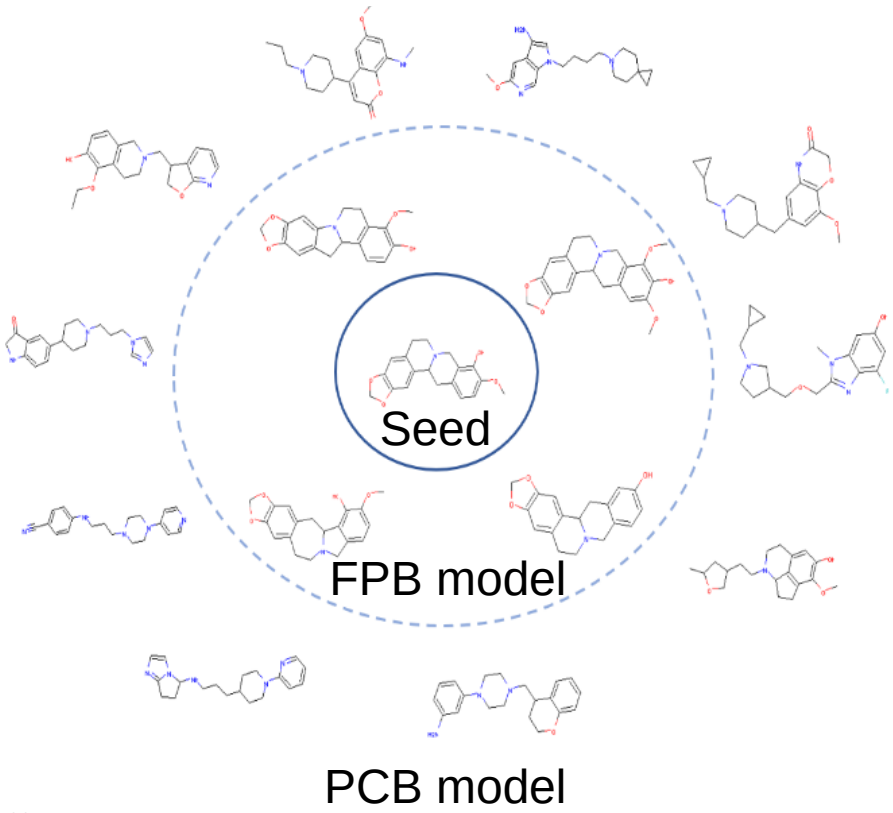


Independent control of properties

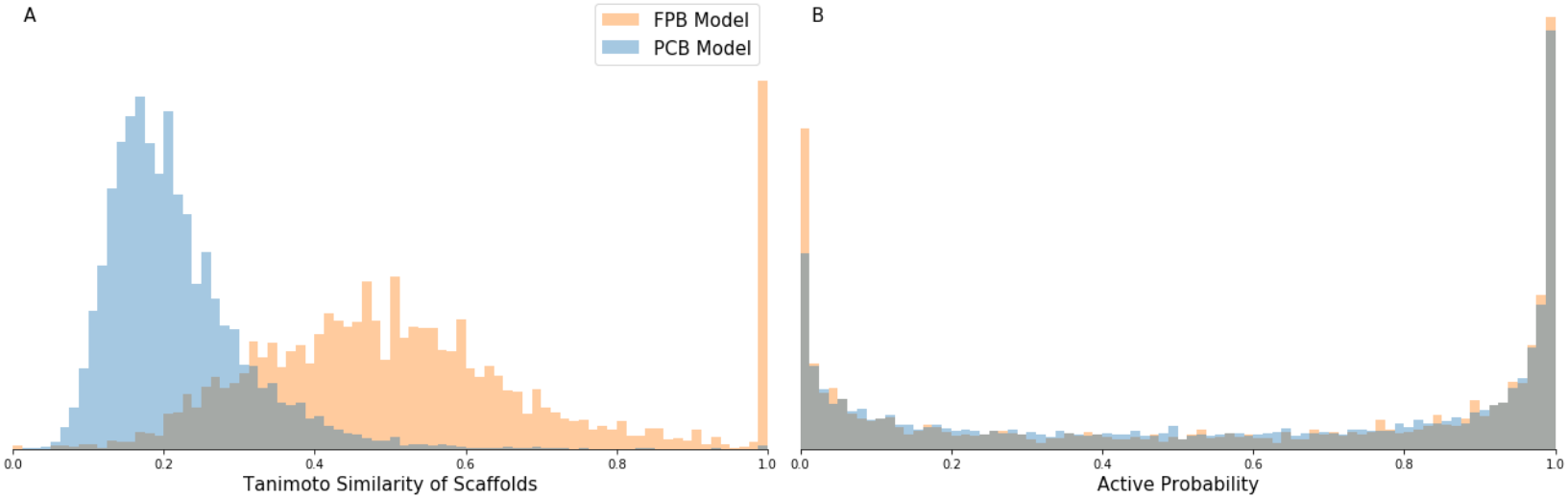


Molecule generation breaks down outside of applicability domain (shown with arrow)

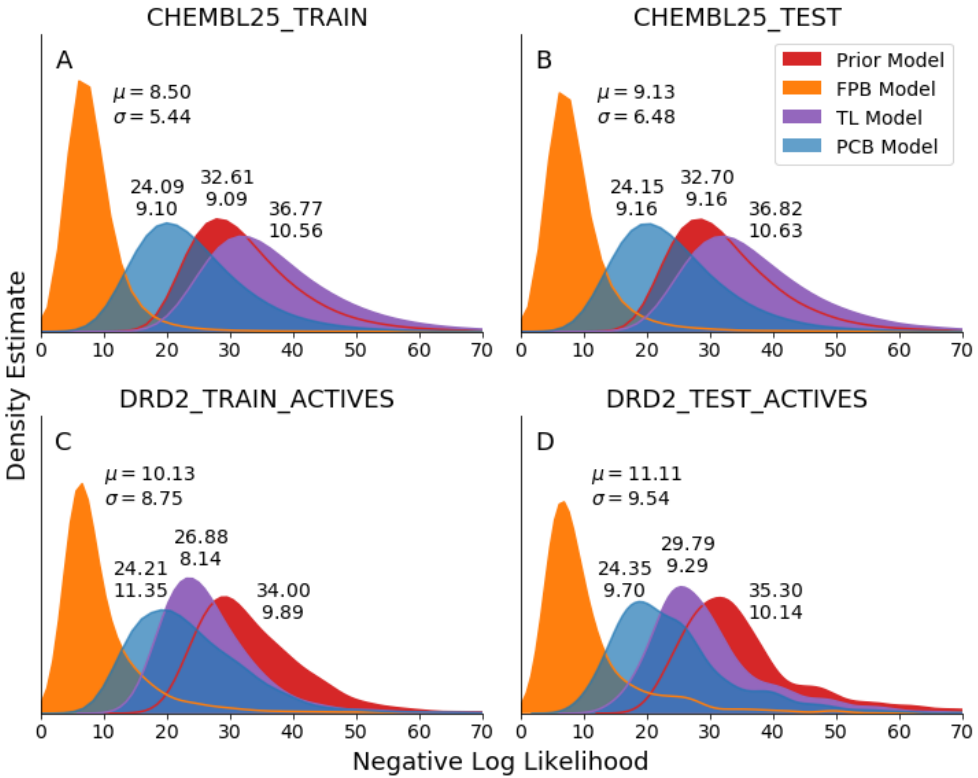
Some molecules generated



cRNN as Scaffold jumper



Generative probability distributions (neg.log.likelihood)



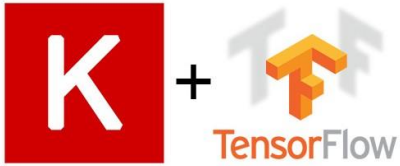
If completely uniform (Dirac distribution)
NLL 0, One SMILES possible
NLL 11, tens of thousands
NLL 24, Billions
NLL 35, Quadrillions

Toolkits – Source code - Links

Blogposts: www.wildcardconsulting.com

github.com/MarcusOlivecrona/REINVENT

github.com/Ebjerrum/molvegen



Deep Drug Coder github.com/pcko1/Deep-Drug-Coder

Conclusions

- Using SMILES and RNNs in different architectures can solve many different tasks
- Its fast to develop! RDKit is central for molecular/SMILES handling and LSTM cells are standard in most NN frameworks
- SMILES enumeration tricks gives better performance in many applications, but sampling becomes a bit more ”fuzzy”
- Conditioinal Recurrent Neural Networks are a direct inverse QSAR model which have intermediate properties between ideal autoencoders and unbiased RNNs

Acknowledgements

Josep Arús-Pous, Ph.D student, BIGCHEM

Panagiotis-Christos Kotsias, Graduate Scientist, IMED Graduate Programme

De Novo Design group

Hongming Chen, Principal Scientist, Molecular AI
Christian Tyrchan, Team Leader - Computational Chemistry
Ola Engkvist, Associate Director, Molecular AI
Atanas Patronov, Associate Principal Scientist, Molecular AI
Michael Withnall, Ph.D student, Molecular AI
Rocio Mercado, Post.doc. Molecular AI
Jiazhen He, post.doc. Molecular AI
Josep Arús-Pous, Ph.D student, BIGCHEM
Dhanushka Weerakoon, Graduate Scientist, IMED Graduate Programme
Simon Johansson, Master student
Oleksii Prydkhodko, Master student

Thank you for listening, Feedback, Questions

Confidentiality Notice