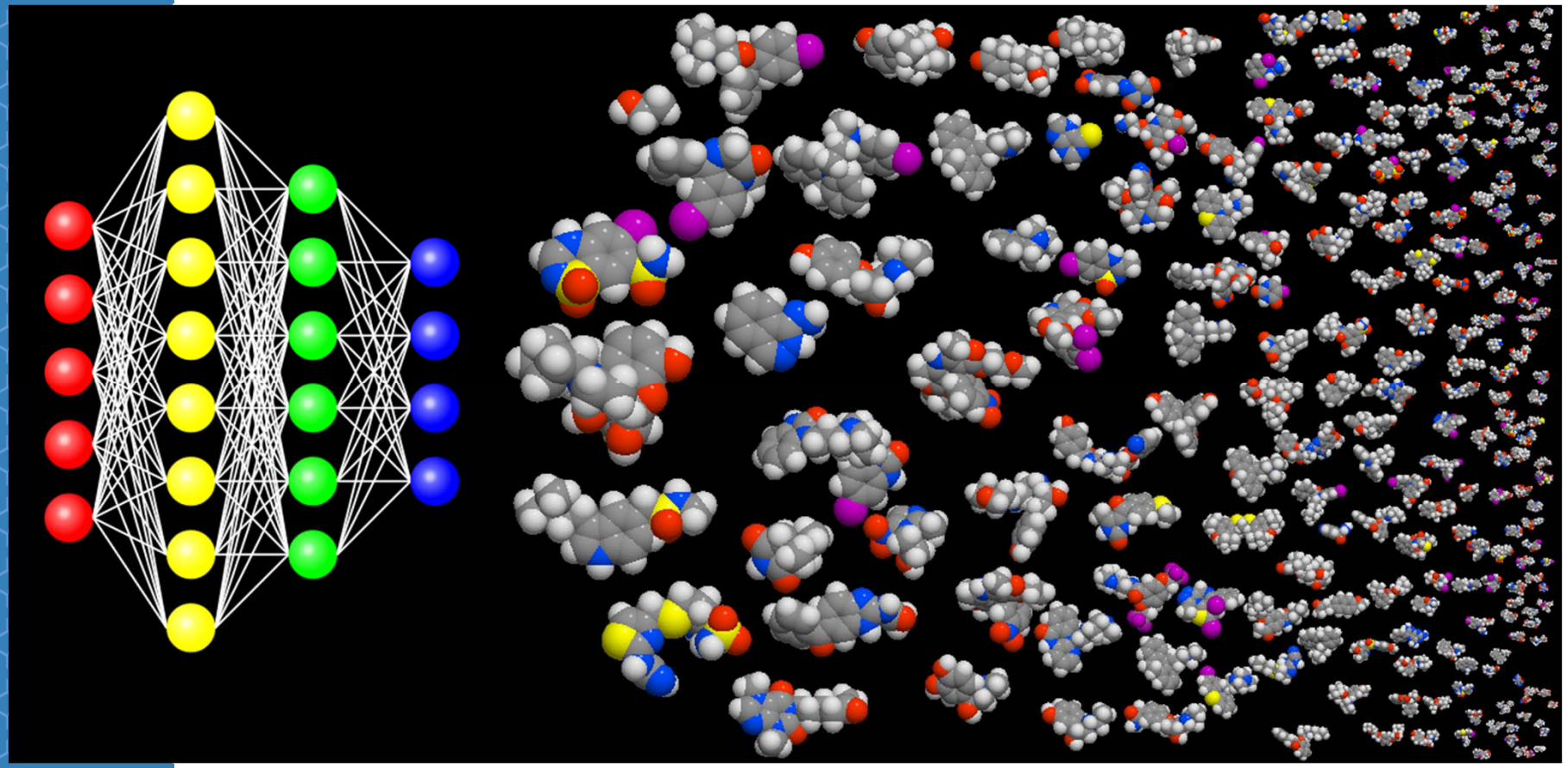
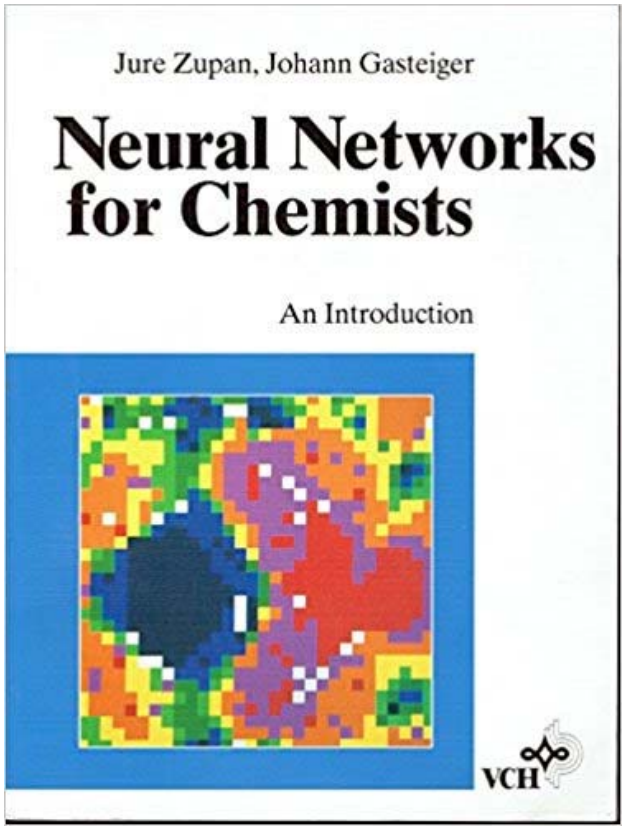


In silico generation of novel, drug-like chemical matter using the LSTM deep neural network

Peter Ertl
Novartis Institutes for BioMedical Research, Basel, CH
September 2018

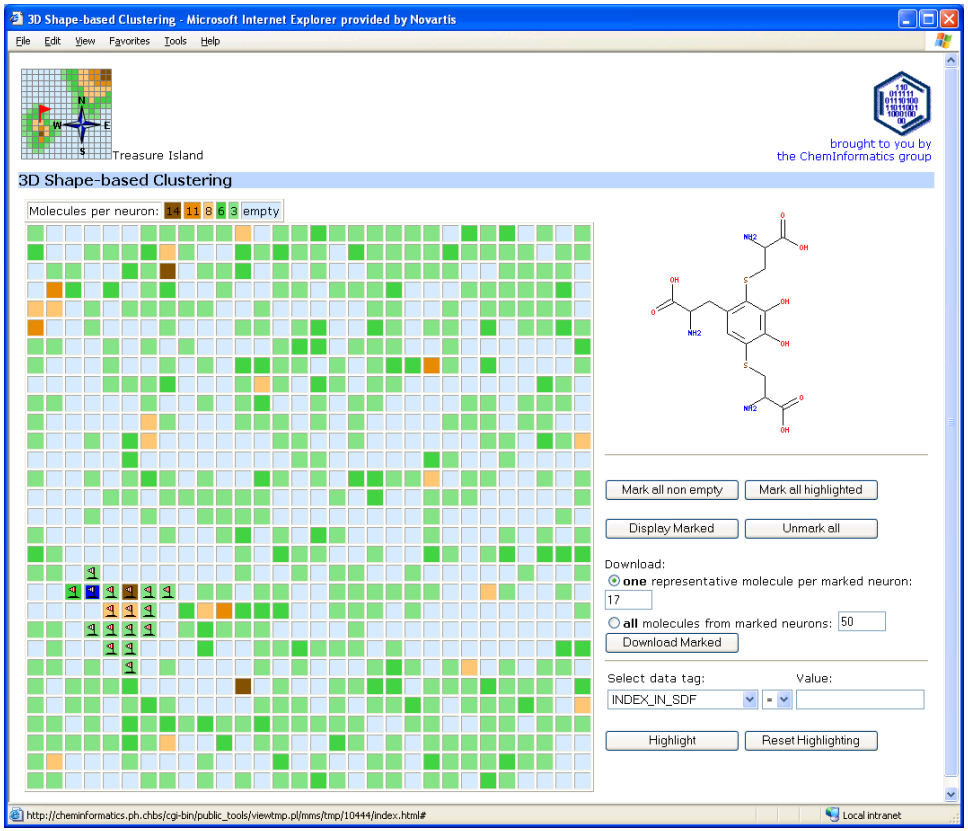
NOVARTIS

Neural networks in cheminformatics



Published in 1993

"Treasure Island" by P. Selzer and P. Ertl, 2001
analysis of the Novartis archive by Kohonen NNs



J. Chem. Inf. Model. 46, 2319 (2006)
Drug Disc.Today 2. 2001 (2005)

NOVARTIS

Neural networks 2.0

Dramatic increase in computation power, much large datasets and particularly novel network architectures available as open source caused a “quantum leap” in the NN applications.

LSTM (long short-term memory) recurrent neural network is one of these novel network types – it powers Google’s speech and image recognition, Apples’s Siri and Amazon’s Alexa.

Very simplistically expressed; the LSTM can learn to understand the “inner structure” or grammar of properly encoded objects and then answer questions about these object or generate objects similar to those in the training set.

Disadvantage of the LSTM networks is that they require very large training sets and also long learning time.

Major challenge
the proper network architecture

To design a network with the correct architecture is not easy; there are no general rules, one has to experiment and try different architectures and parameters.

Numerous network parameters need to be to set-up:

- number, types and size of network layers
- learning rate
- drop-out rate
- loss and activation functions

combination of these parameters makes the number of possible network architectures practically unlimited.

It took >3 weeks of heavy computational experiments to find the properly working network architecture to design new molecules.

The LSTM architecture used

```
Using TensorFlow backend.
corpus length: 23664668
total chars: 23
nb sequences: 7888210
```

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 40, 128)	77824
lstm_2 (LSTM)	(None, 56)	41440
dropout_1 (Dropout)	(None, 56)	0
dense_1 (Dense)	(None, 23)	1311
activation_1 (Activation)	(None, 23)	0

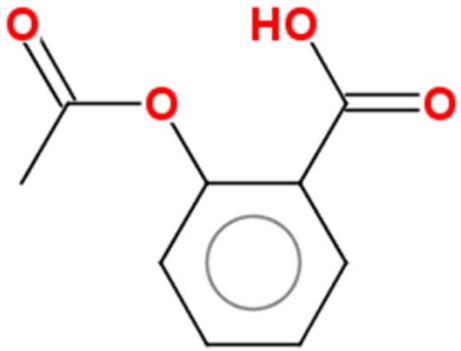
```
Total params: 120,575.0
Trainable params: 120,575.0
```

After several trials, the architecture shown above has been chosen. The following parameters have been used: dropout rate 0.2, RMSprop optimizer with learning rate 0.01 and categorical crossentropy as a loss criteria during the optimization.

The implementation was done using standard software: Python3 + Keras + TensorFlow

SMILES

Simplified Molecular-Input Line-Entry System



uppercase – aliphatic atom

= - double bond

CC(=O)Oc1ccccc1C(O)=O

parenthesis - branching

lowercase – aromatic atom

pair of numbers - ring closure

CC(CCC(=O)NCCS(O)(=O)=O)C1CCC2C4C(CC(O)C12C)C3(C)CCC(O)CC3CC4O

trained on 40 previous characters the output should be O

NOVARTIS

Training

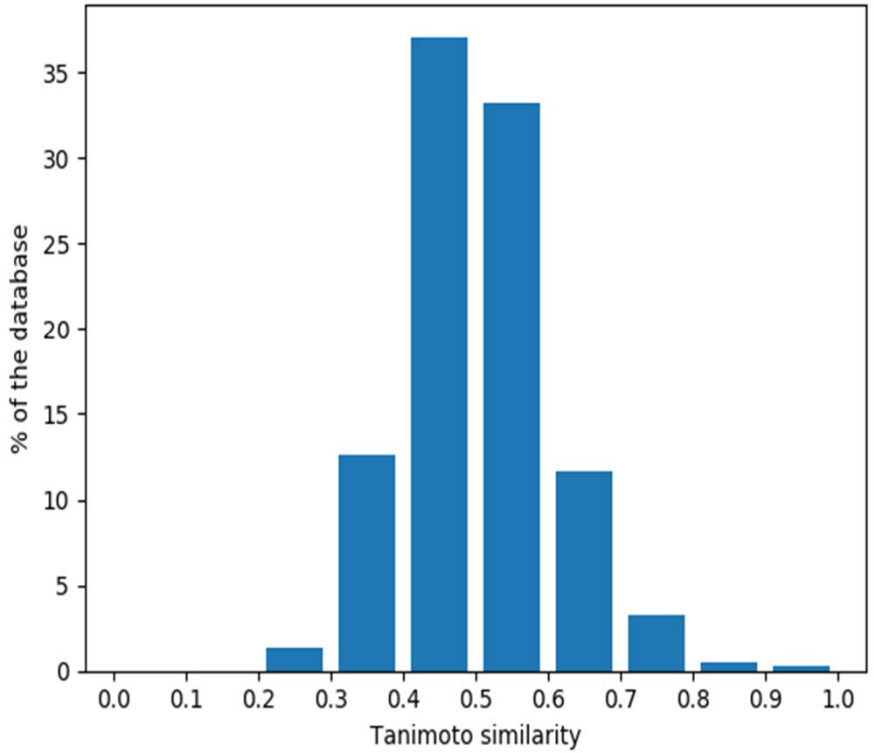
- the network was trained on ~550,000 bioactive SMILES's from ChEMBL
- the goal was to learn the grammar of SMILES's encoding bioactive molecules and then to use the trained network to generate the SMILES's for novel molecules
- some level of "randomness" had to be also added (we do not want to exactly reproduced the ChEMBL structures, but generate novel, ChEMBL-like molecules)
- training took ~1 week on a single CPUs - parallelization is not yet supported in Keras (according to my 1st experiments, using the GPUs would speed-up training about 5-times).

Generation of novel molecules

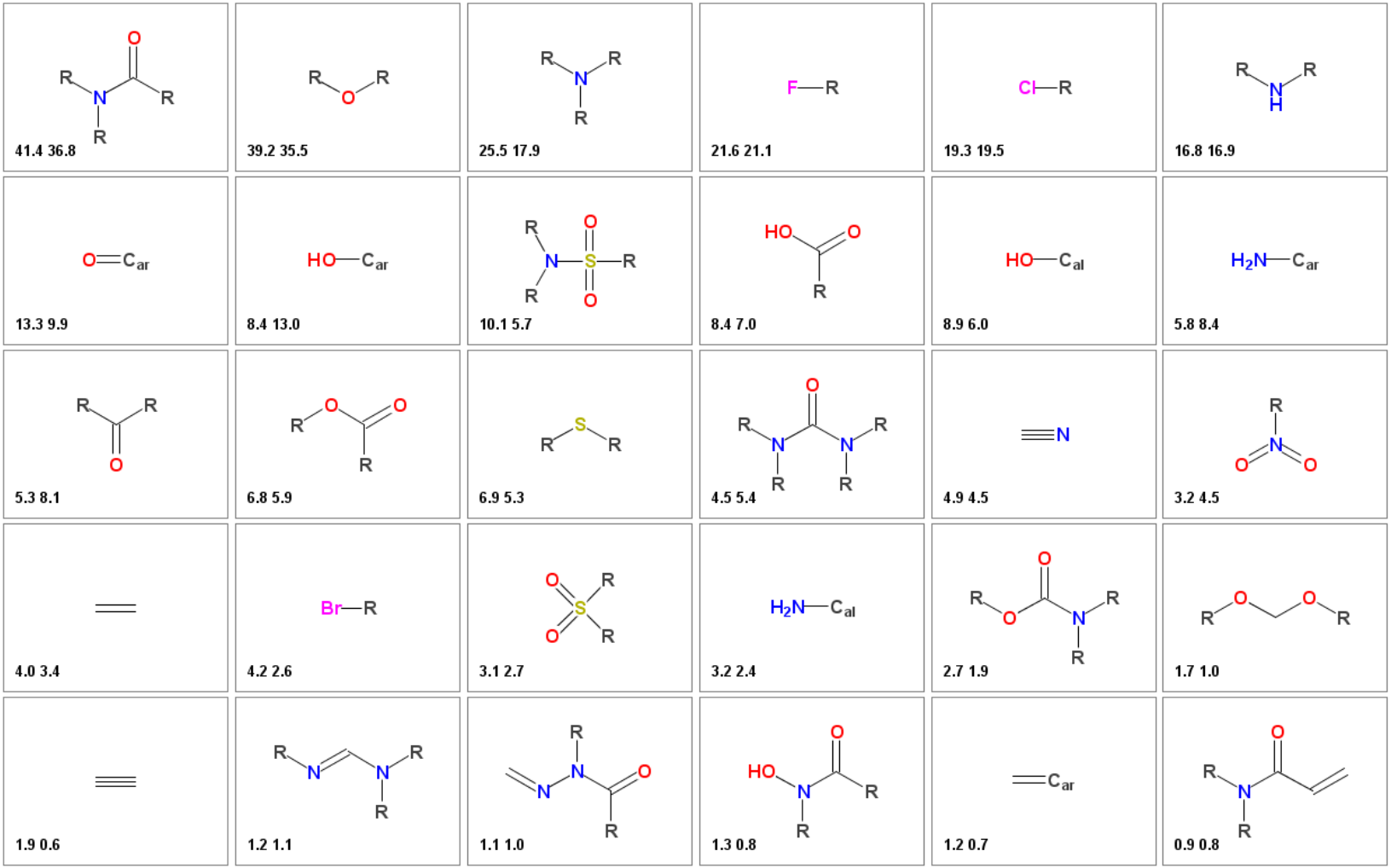
- the new SMILES's were generated character-by-character based on learned structure of ~550k ChEMBL SMILES's
- also incorrect SMILES's were generated (already one false character makes SMILES non-parsable); 54% of generated SMILES's could be discarded just by text check (not all brackets or rings paired), additional 14% were discarded after parsing (problems with aromaticity or incorrect valences); at the end 32% of SMILES's generated led to correct molecules (this ratio could be increased by longer training, but we do not want to reproduce exactly ChEMBL)
- generation of 1 million correct SMILES's required less than 2 hours on 300 CPUs using rather crude code; optimization, switch to GPUs and more processors would allow to generate 100s of millions, even billions of novel molecules

Novelty of generated structures

- out of 1 million generated molecules only 2774 were contained in the ChEMBL training set – the generated molecules are novel
- similarity to ChEMBL structures (standard RDKit similarity; distance to the closest neighbor) is medium
- 1 million generated structures contain 627k unique scaffolds, (550k ChEMBL structure contain 172k scaffolds), overlap between 2 sets is only 18k scaffolds = the generated structures are diverse and contain many new chemotypes



Functionalities of novel structures



Distribution of functional groups (% in the new set, % in ChEMBL) is very similar in both sets. The generated structures are novel, but of the same type as bioactive ChEMBL molecules.

P. Ertl, An algorithm to identify functional groups in organic molecules, *J. Cheminformatics* 9:36 2017

Substructure analysis

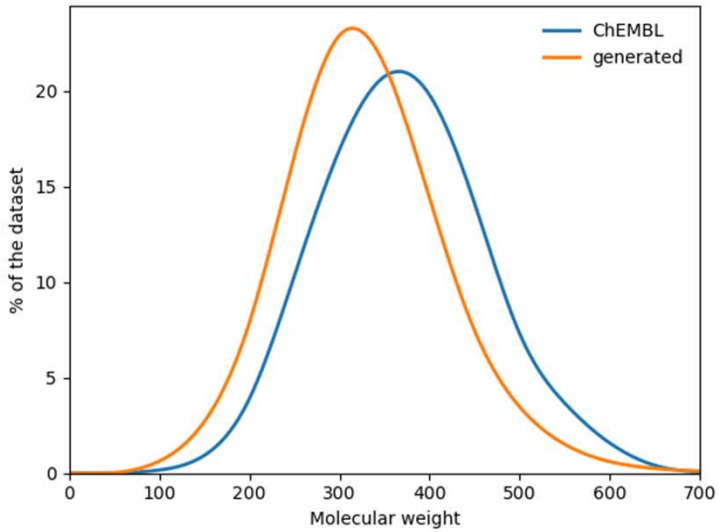
Feature	ChEMBL	generated	baseline	Feature	ChEMBL	generated	baseline
no rings	0.4	0.4	0.1	Large rings (>8)	0.4	1.8	75.9
1 ring	2.8	4.3	13.2	Spiro ring	1.9	0.6	0.6
2 rings	14.8	23.1	17.7	without N,O,S	0	0.2	2.6
3 rings	32.2	43.5	27.3	contains N	96.5	96.1	92.3
4 rings	32.7	23.9	25.2	contains O	93	92	85.5
>4 rings	17.2	4.8	16.5	contains S	35.6	27.9	39.6
fused ar. rings	38.8	30.9	0.2	contains halogen	40.7	38.8	49.4

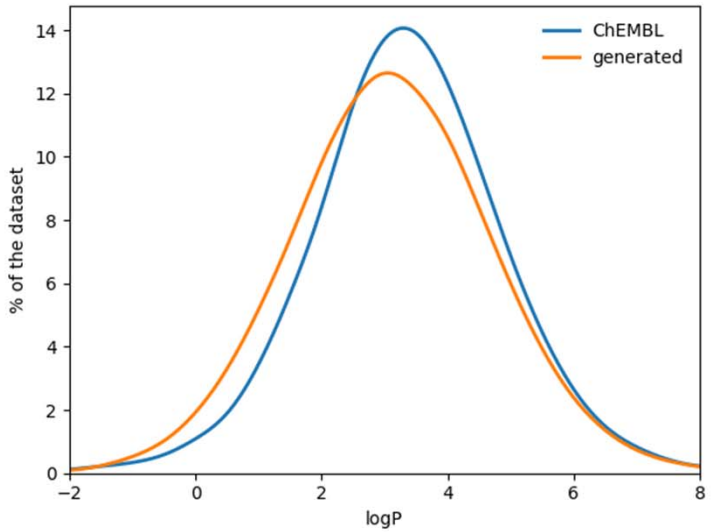
Average molecule formula:

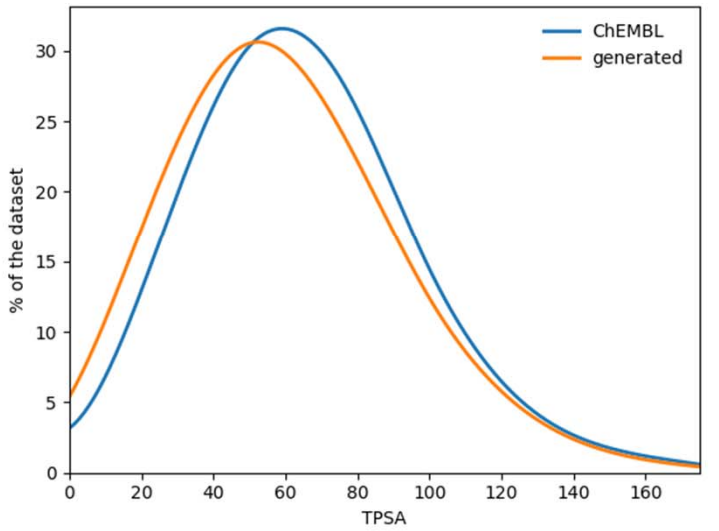
ChEMBL	C20.6 H22.1 N3.3 O2.8 S0.4 X0.8	
generated	C18.7 H19.7 N3.0 O2.5 S0.3 X0.6	(X = halogen)
baseline	C25.1 H35.0 N3.8 O2.3 S0.5 X0.7	

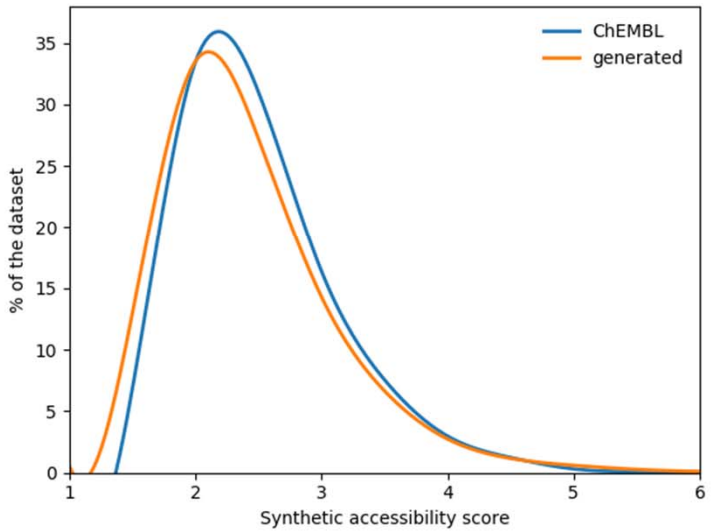
The substructure features of the generated molecules are practically identical with those of ChEMBL structures, the “baseline” structures are very different, contain many macrocycles.

Molecular properties









P Ertl | LSTM SMILES generator

Summary

- properly configured LSTM deep neural networks are able to learn the general structural features of bioactive molecules and then generate novel molecules of this type
- novelty, molecular properties, distribution of functional groups and synthetic accessibility of generated structures is very good
- the network is able to generated practically unlimited stream (100's of millions, even billions) of novel molecules
- obvious applications of the generated drug-like molecules are:
  - virtual screening
  - "identifying holes" in the corporate chemical space
  - generation of specialized molecule sets (natural product-like, target X-like, ...)
  - additional applications (possibly using other deep learning techniques) can be discussed

Acknowledgements

Niko Fechner
Brian Kelley
Richard Lewis
Eric Martin
Valery Polyakov
Stephan Reiling
Bernd Rohde
Gianluca Santarossa
Nadine Schneider
Ansgar Schuffenhauer
Lingling Shen
Finton Sirockin
Clayton Springer
Nik Stiefl
Wolfgang Zipfel

In silico generation of novel, drug-like chemical matter using the LSTM neural network

Peter Ertl, Richard Lewis, Eric Martin, Valery Polyakov

arXiv:1712:07499 (2017)

The Python code (BSD license) is available on request