Electronic Supplementary Material (ESI) for RSC Advances.
This journal is © The Royal Society of Chemistry 2020

Supplementary Information

cheML.io: An Online Database of ML-generated Molecules

R. Zhumagambetov, D. Kazbek, M. Shakipov, D. Maksut, V. A. Peshkov and S. Fazli

Moses benchmark comparison

To juxtapose the methods that have been employed, the MOSES benchmark framework was used[1]. It is a benchmark that encompasses several metrics that assess the generated molecules in order to characterize a given method. One of the important metrics is *novelty*. It is a proportion of generated molecules that are not seen in the initial database, i.e. the training set. *Filters* is a metric that measures the proportion of molecules that are passed through the custom medicinal chemistry filters (MCFs) and PAINS filters[2]. These filters were hand-picked to exclude molecules with specific undesired properties, such as reactivity and chelation. $IntDiv_1$ and $IntDiv_2$ assess the internal diversity of the generated molecules and their values range between 0 and 1. A low score indicates that the generated molecules are limited in the variety of scaffolds and a high score corresponds to a higher inner diversity of the molecules.

$IntDiv_1$ and $IntDiv_2$ can be calulated as follows

$$\mathrm{IntDiv}_p(G) = 1 - \sqrt[p]{\frac{1}{|G|^2} \sum_{m_1, m_2 \in G} T(m_1, m_2)^p},$$

where G stands for the set of molecules, T stands for Tanimoto distance, m_1 and m_2 stand for any pair of molecules in the set G and $p \in 1, 2$.

As can be seen from Table S1, aside from GrammarVae and MolCycleGAN the algorithms have shown they are able to produce a large portion of novel, never-seen-before molecules. The internal diversity score also shows high values across all 10 algorithms.

Model	# of molecules	IntDiv$_1$	IntDiv$_2$	Filters	Novelty
JT-VAE	1399265	0.861	0.856	0.733	1
RNN	962247	0.847	0.837	0.813	0.953
GrammarVAE	239262	0.871	0.865	0.598	0.196
ChemVAE	99344	0.879	0.874	0.589	1
MolCycleGan	60856	0.869	0.862	0.607	0.42
ORGAN	50268	0.86	0.852	0.71	0.902
ORGANIC	42610	0.855	0.842	0.621	0.999
SSVAE	42606	0.84	0.832	0.89	0.971
CDN	3639	0.886	0.878	0.620	0.997
CVAE	539	0.786	0.767	0.538	1

Table S1: Comparison of methods by means of the MOSES benchmark

Comparison of ML model outputs

To further analyze the properties of the molecules within the CheML database, molecules generated by each ML model were compared side-by-side. The main tools are again Levene test for equal variances and Two-Sample KS-test for equal distributions.

Levene test was conducted 270 times in total as per Table S2, and it showcases that for the most part, the differences in variances of each subset along each property is statistically significant. The only exception being again CVAE-generated molecules, which suggest statistically insignificant difference in variance along the number of bridgehead atoms in comparison to 7 of the 9 other subsets.

Two-Sample KS-tests were conducted 270 times as well as per Table S3. These tests reveal that when comparing distributions across methods and molecular properties, all distributions are statistically different

Model	JT-VAE	RNN	GrammarVAE	ChemVAE	MolCycleGAN	ORGAN	ORGANIC	SSVAE	CDN
JT-VAE									
RNN	RRRRRR								
GrammarVAE	RRRRNR	RRRRRR							
ChemVAE	RRRRRR	RRRRRR	RRRRRR						
MolCycleGAN	RRRRRR	RRRRRR	RRRNRR	RRRRRN					
ORGAN	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR				
ORGANIC	RRRRRR	RRRRRR	RRRRRR	RRRRRR	NNRRRR	RRRRNR			
SSVAE	RRNRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR		
CDN	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR	
CVAE	RRRRNR	RRRRNR	RRRRNR	RRRRRR	RRRRNR	RRRRNN	RRRRNR	RRRRNR	RRRRRR

Table S2: Levene test for variances between CheML model outputs across each property. The six letters inside each cell indicate whether the null hypothesis of equal variances was rejected or not (R – for being rejected, N – for not being rejected). Properties were ordered as such: number of molecules, exact molecular mass, number of atoms, number of chiral centers, number of rings, number of bridgehead atoms, number of heterocycles.

Model	JT-VAE	RNN	GrammarVAE	ChemVAE	MolCycleGAN	ORGAN	ORGANIC	SSVAE	CDN
JT-VAE									
RNN	RRRRRR								
GrammarVAE	RRRRNR	RRRRNR							
ChemVAE	RRRRRR	RRRRRR	RRRRRR						
MolCycleGAN	RRRRNR	RRRRRR	RRRRNR	RRRRNR					
ORGAN	RRRRNR	RRRRNR	RRRRNR	RRRRRR	RRRRNR				
ORGANIC	RRRRRR	RRRRNR	RRRRRR	RRRRRR	RRRRRR	RRRRRR			
SSVAE	RRNRNR	RRRRRR	RRRRNR	RRRRNR	RRRRNR	RRRRNR	RRRRRR		
CDN	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR	RRRRRR	
CVAE	RRRRNR	RRRRNR	RRRRNR	RRRRNR	RRRRNR	RRRRNN	RRRRNR	RRRRNR	RRRRNR

Table S3: Two-sample Kolmogorov-Smirnov test for equal distributions between CheML model outputs across each property. The six letters inside each cell indicate whether the null hypothesis of equal variances was rejected or not (R – for being rejected, N – for not being rejected). Properties were ordered as such: number of molecules, exact molecular mass, number of atoms, number of chiral centers, number of rings, number of bridgehead atoms, number of heterocycles.

with the only exception being the number of bridgehead atoms between each pair of the following sets: GrammarVAE, CVAE, ORGAN, MolCycleGAN and SSVAE.

References

[1] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik and A. Zhavoronkov, *arXiv*, 2020, preprint, arXiv:1811.12823v5, `https://arxiv.org/abs/1811.12823v5`.

[2] J. B. Baell and G. A. Holloway, *J. Med. Chem.*, 2010, **53**, 2719.