EVALUATION OF GENERATIVE MODELS FOR MOLECULES

Philipp Renz renz@ml.jku.at

October 22, 2021

ELLIS Unit Linz, LIT AI Lab, Institute for Machine Learning,
Johannes Kepler University, Linz, Austria

JKU

Lots of attention on automatic drug design

Cell
Volume 180, Issue 4, 20 February 2020, Pages 688-702.e13

Article

A Deep Learning Approach to Antibiotic Discover

Jonathan M. Stokes [1, 2, 3], Kevin Yang [3, 4, 10], Kyle Swanson [3, 4, 10], Wengong Jin [3, 4], Andres Cubillos-Ruiz [1, 2, 5], Nina M. Donghia [1, 5], Craig R. MacNair [6], Shawn French [6], Lindsey A. Carfrae [6], Zohar Bloom-Ackermann [2, 7], Victoria M. Tran [2], Anush Chiappino-Pepe [5, 7], Ahmed H. Badran [2], Ian W. Andrews [1, 2, 5], Emma J. Chory [1, 2], George M. Church [5, 7, 8], Eric D. Brown [6], Tommi S. Jaakkola [3, 4], Regina Barzilay [3, 4, 9], James J. Collins [1, 2, 5, 9, 11]

Learning to Navigate The Synthetically Accessible Chemical Space Using Reinforcement Learning

Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Shengchao Liu, Simon Blackburn, Karam Thomas, Connor Coley, Jian Tang, Sarath Chandar, Yoshua Bengio *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:3668-

Brief Communication | Published: 02 September 2019

Deep learning enables rapid identification of potent DDR1 kinase inhibitors

Alex Zhavoronkov, Yan A. Ivanenkov, Alex Aliper, Mark S. Veselov, Vladimir A. Aladinskiy, Anastasiya V. Aladinskaya, Victor A. Terentiev, Daniil A. Polykovskiy, Maksim D. Kuznetsov, Arip Asadulaev, Yury Volkov, Artem Zholus, Rim R. Shayakhmetov, Alexander Zhebrak, Lidiya I. Minaeva, Bogdan A. Zagribelnyy, Lennart H. Lee, Richard Soll, David Madge, Li Xing, Tao Guo & Alán Aspuru-Guzik

Nature Biotechnology **37**, 1038–1040 (2019) | Cite this article

Introduction

- Many generative models for de-novo drug design in recent years (2016-ongoing)
- Aim is to "invent" new molecules.
- No test set for testing the algorithm
- Evaluation strategy?

Evaluating the results

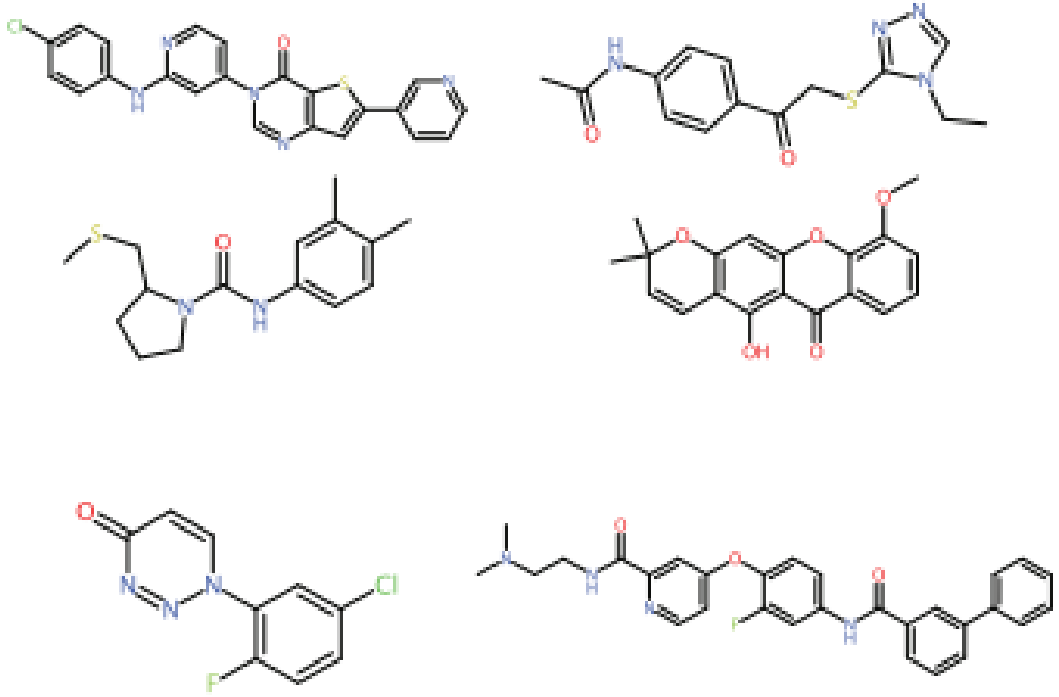
- Important to critically review methods
- Could the results be achieved by more simple means?
- Could an expert come up with similar ideas in less time?
- How useful are results in the first place?

Distribution-learning

The aim is to generate samples that resemble the training set in distribution.

- Novelty?
- Quality?
- Diversity?
- Distribution match?

Distribution-learning



- Novelty?
- Quality?
- Diversity?
- Distribution match?

Guacamol metrics

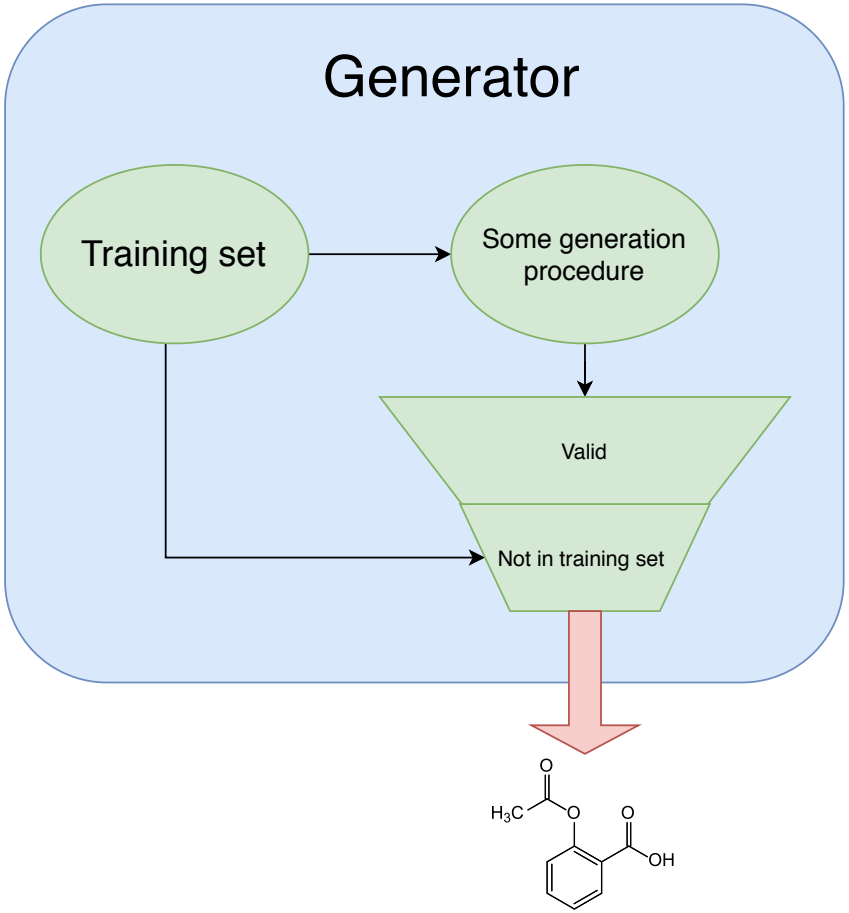
Brown et al. (2019) used the following metrics to evaluate a generated set:

- Validity: Percentage of viable molecules (correct valences,...)
- Uniqueness: Fraction of non-repeated molecules in set
- Novelty: Fraction of molecules not in training set.
- KL divergence: mean KL-divergence between property (MolLogP, MolWt, TPSA,..)
- FCD: Frechet distance between ChemNet representations of train set and generated set.

Problems

- Validity: Discard invalid molecules
- Uniqueness: Discard non-unique molecules
- Novelty: Discard non-novel molecules

Model sketch



Filtering will slow model down, but speed is usually not measured in the first place.

JKU

AddCarbon[2]

- Take random compound from training set
- Add a carbon at some point in the SMILES string, such that it minimally changes the canonical SMILES
- Only output it if it is novel and valid

Table 1: Comparing the AddCarbon model to the baselines in[1]. RS randomly samples from the training set.

Benchmark	RS	LSTM	GraphMCTS	AAE	ORGAN	VAE	AddCarbon
Validity	1.000	0.959	**1.000**	0.822	0.379	0.870	**1.000**
Uniqueness	0.997	**1.000**	**1.000**	**1.000**	0.841	0.999	0.999
Novelty	0.000	0.912	0.994	0.998	0.687	0.974	**1.000**
KL divergence	0.998	**0.991**	0.522	0.886	0.267	0.982	0.982
FCD	0.929	**0.913**	0.015	0.529	0.000	0.863	0.871

[1] Brown et al. 2019.

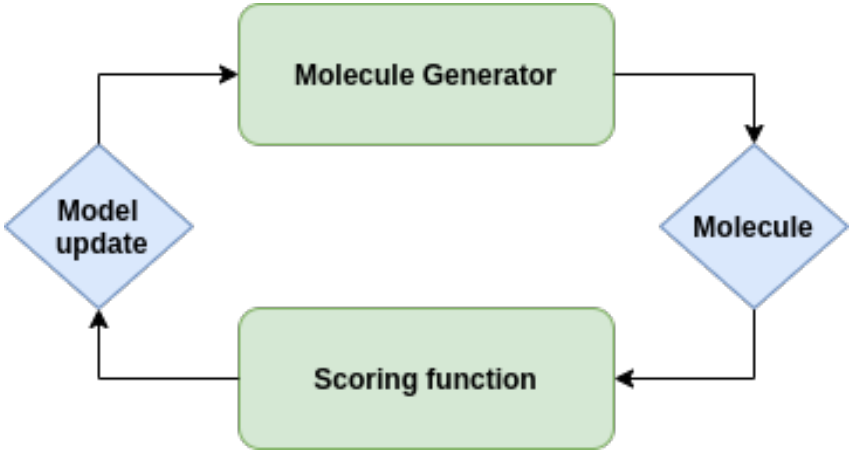
[2] Renz et al. 2019.

Takeaways

- Simple model performs relatively well.
- Casts doubt on how expressive the metrics are.
- Consider cross-entropy (Bits per character) on a test set if applicable

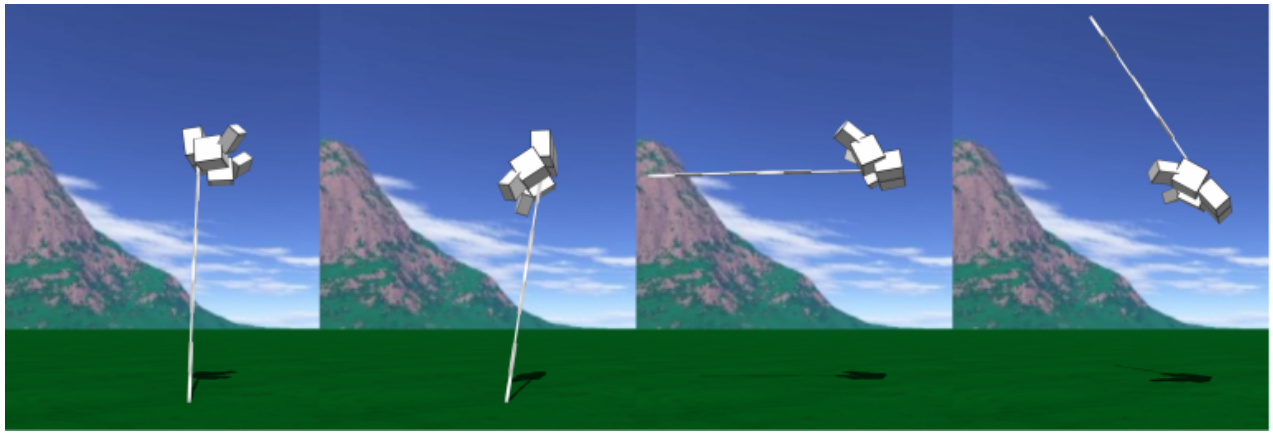
Goal-directed generation

- Aim is to generate molecules that satisfy some property profile (bioactivity, physchem, ADME)
- Aim encoded as a scoring function: Molecule in, score out
- Hard to encode complex properties as scoring functions



Imperfect scoring function

Evolving a body that can jump. Score is determined highest point reached by any part of the body[3].



[3] Lehman et al. 2019.

ML models as scoring functions

GSK-3β and JNK3 are potential targets in the treatment of Alzheimer's disease. Their corresponding property predictors are random forests trained on real-world experimental data using Morgan fingerprint features (Rogers & Hahn, 2010). In our experiment, we consider all properties by combining their scores into a unified scoring function[1]: 4

Biological objectives. Following Jin et al. (2020), we consider the following inhibition scores against two Alzheimer-related target proteins as the biological activity objectives. The score is given by a random forest model [2] that predicts based on Morgan fingerprint features of a molecule (Rogers & Hahn, 2010).

- GSK3β: Inhibition against glycogen synthase kinase-3β.
- JNK3: Inhibition against c-Jun N-terminal kinase-3. 5

4.1. Predictive Modeling

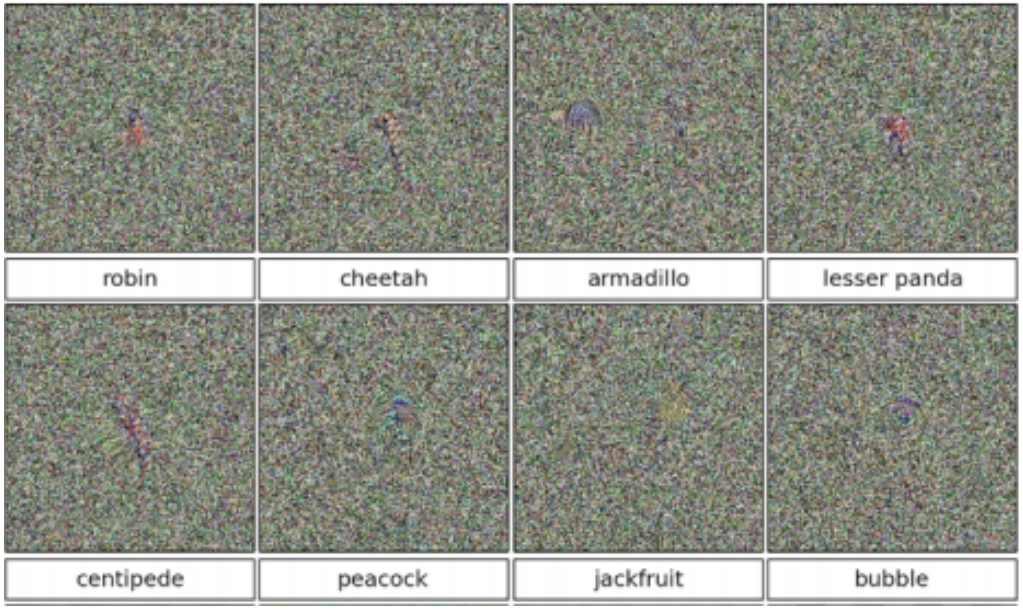
To test the applicability of PGFS in an in-silico proof-of-concept for de novo drug design, we develop predictive models against three biological targets related to the human immunodeficiency virus (HIV) - as scoring functions. The biological activity data available in the public domain allowed us to develop ligand-based machine learning models using the concept of quantitative structure-activity relationship modeling (QSAR). 6

A support vector machine (SVM) classifier with a Gaussian kernel was built in Scikit-learn [40] on the training set as a predictive model for DRD2 activity. The optimal C and Gamma values utilized in the final model were obtained from a grid search for the highest ROC-AUC performance on the validation set. 7

[4] Chen et al. 2021.
[5] Xie et al. 2021.
[6] Gottipati et al. 2020.
[7] Olivecrona et al. 2017.

JKU

An image analogy



Generating images by maximizing outputs of a classifier gives unpleasing results[8].

[8] Nguyen, Yosinski, and Clune 2015.

Problems

- Optimizing output of ML models can be problematic.
- How relevant are predictions outside of training domain?
- Do molecules "overfit" to scoring function?

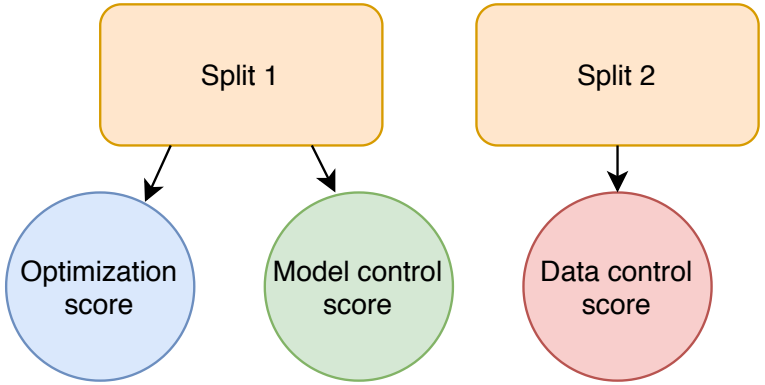
Data

Target	ChEMBL ID	Active	Inactive	AUC
JAK2	CHEMBL3888429	140	527	0.78±0.03
EGFR	CHEMBL1909203	40	802	0.76±0.05
DRD2	CHEMBL1909140	59	783	0.86±0.03

Table 2: Information on the data sets. AUC shows the performance of the trained classifiers

- The ratio of actives to inactives in both splits is kept equal.
- ECFP4 used as features.
- Random forest classifiers

Optimization / control scores[9]



We train three classifiers to obtain three different scoring functions:

- Optimization score (OS): Classifier trained on Split 1
- Model control score (MCS): Classifier trained on Split 1, but with different random seed
- Data control score (DCS): Classifier trained on Split 2

[9] Renz et al. 2019.

Data specific biases

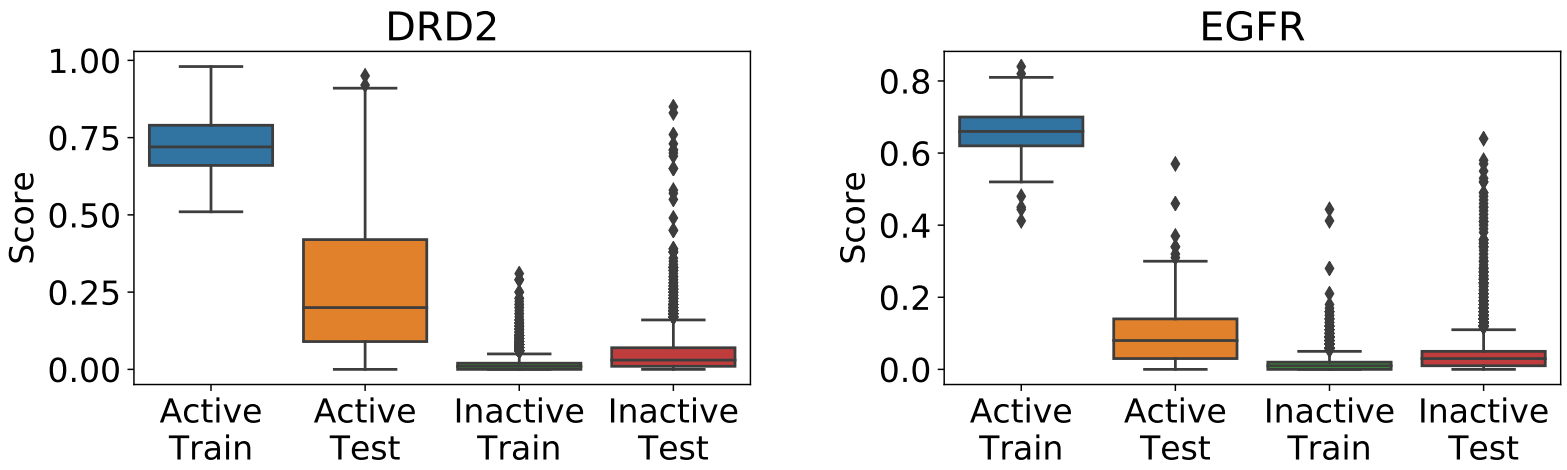


Figure 1: Random forest predictions on bioactivity datasets described below.

- Classifiers fit on data usually exhibit a bias to those exact data.
- During optimization we might prioritize compounds similar to train actives.

JKU

Optimization

We employ the two top-performing methods in[10]:

- GA: graph based genetic algorithm (GA)[11].
  1. Start with random molecules from ChEMBL.
  2. Make random changes to them.
  3. Keep the best ones
  4. Back to 2
- LSTM: Next character LSTM combined with hill-climbing[12]
  1. Pretrain SmilesLSTM on ChEMBL
  2. Sample molecules
  3. Add best to buffer
  4. Finetune on buffer
  5. Back to 2

[10] Brown et al. 2019.
[11] Jensen 2019.
[12] Segler et al. 2017.

Quantifying model/data specific exploits

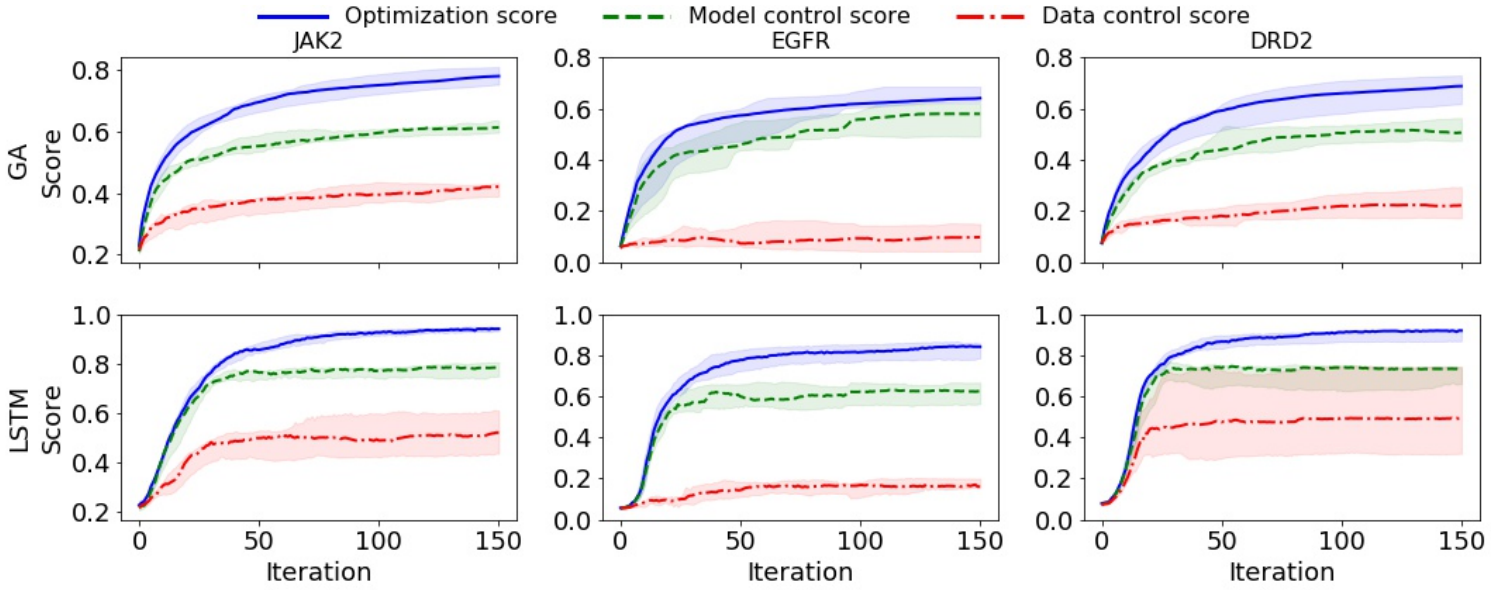


Figure 2: The bold line corresponds to the median while the shaded areas correspond to the interquartile range.

- OS and MCS grow in sync initially and later diverge.
- OS always grows, while control scores sometimes fall

JKU

Similarity embedding 1

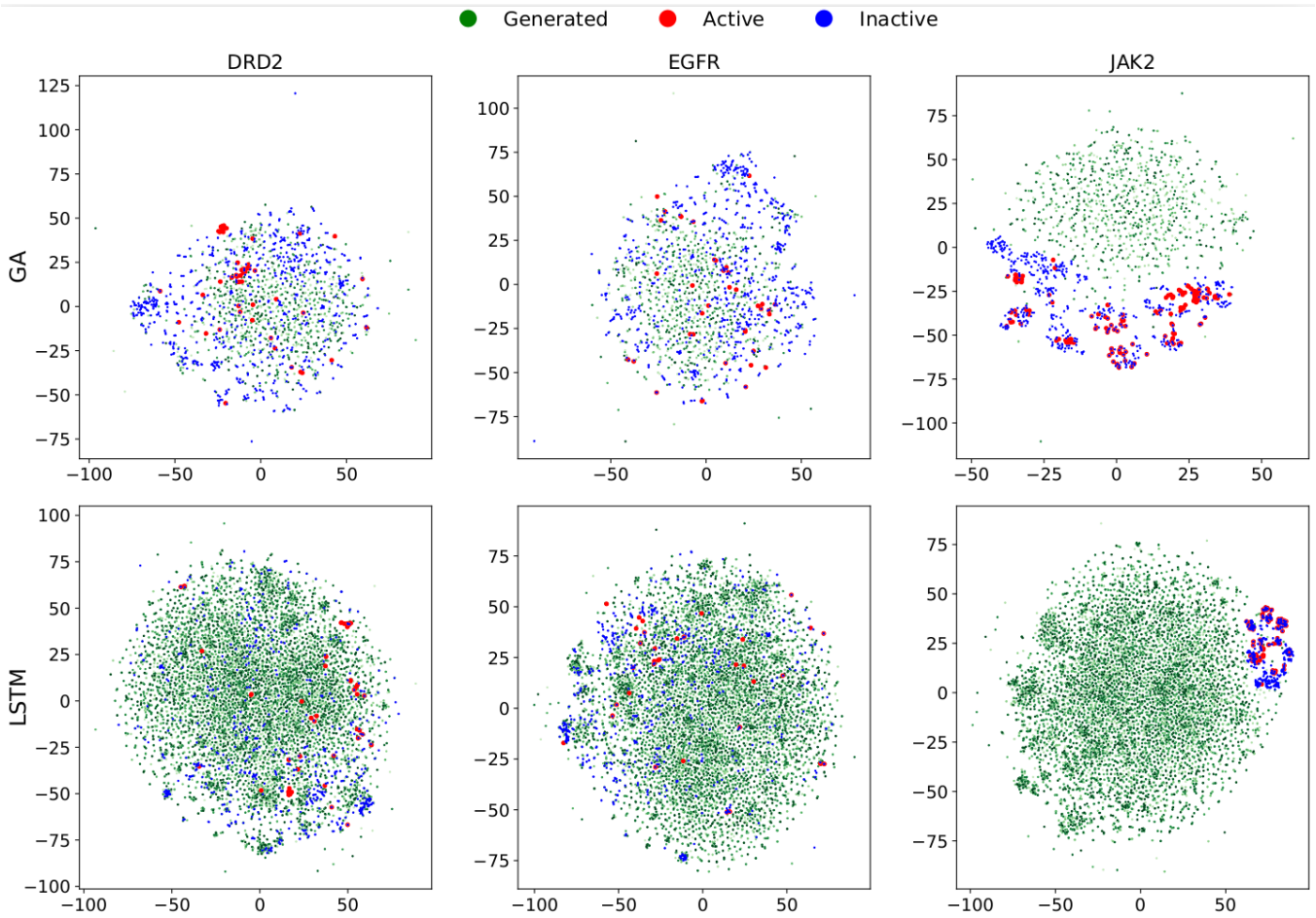


Figure 3: t-SNE embedding at the start of optimization

Similarity embedding 2

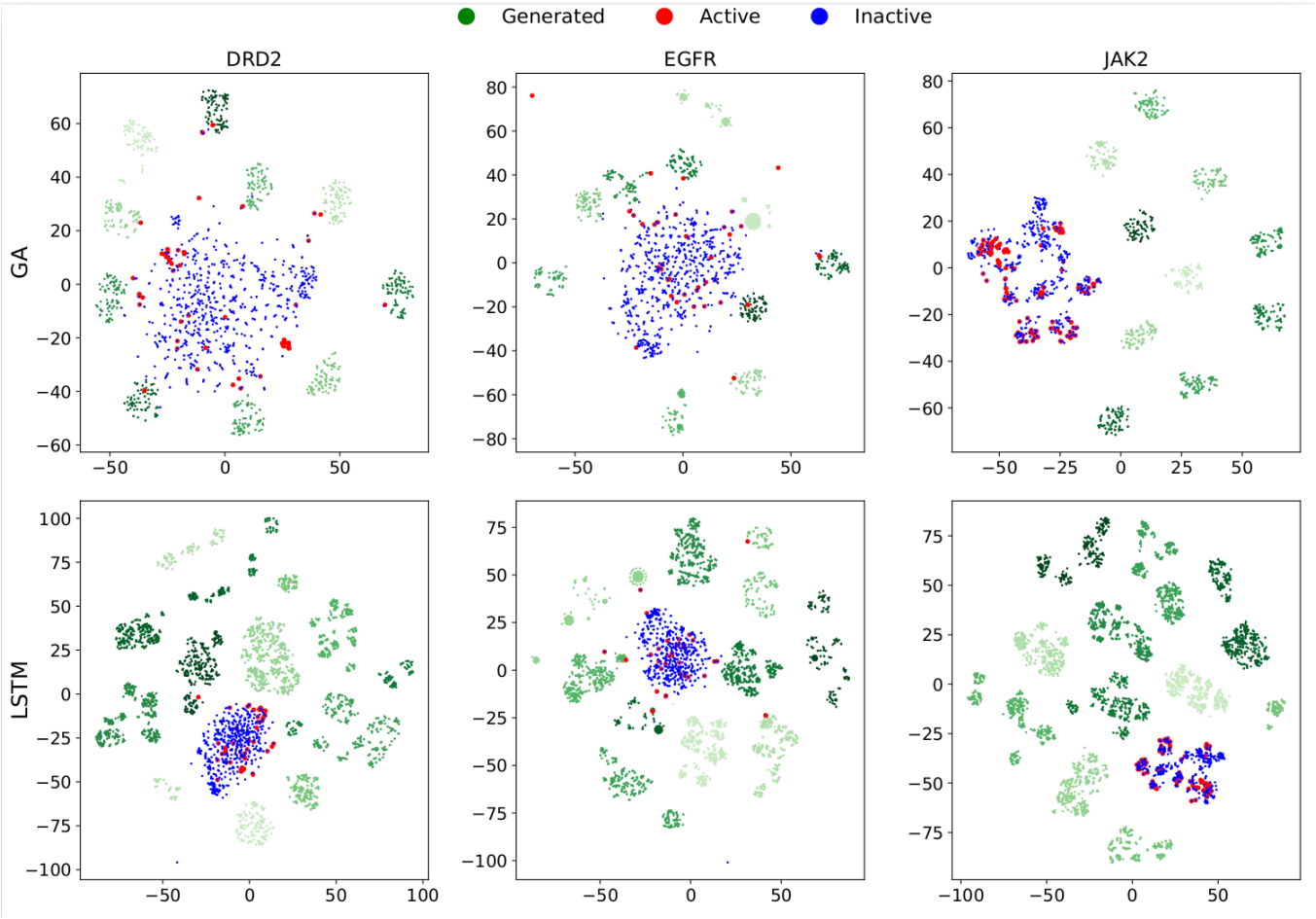


Figure 4: t-SNE embedding at the end of optimization

More questions

One should state the desiderata clearly:

- Do we want one top-scoring molecule or many?
- If we report absolute number of "good" molecules found are we satisfied with trivial variatons?
- Could we have achieved results with simpler methods (virtual screening)?
- Did we provide the same compute budget to the simpler methods?

Summary

- Distribution-learning
  - Current distribution-learning evaluation is not really sufficient.
  - Likelihood on test set would give better evaluation if possible.
  - More relevant measures of novelty important.
- Goal-directed learning
  - Optimization methods show both
    - Model specific biases
    - Data specific biases
  - Control scores might help to better evaluate generated compounds.
  - Predictive accuracy of scoring function on generated compounds unknown.

Brown, Nathan et al. (Mar. 2019). “GuacaMol: Benchmarking Models for de Novo Molecular Design”. In: Journal of Chemical Information and Modeling 59.3, pp. 1096–1108. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.8b00839.

Chen, Binghong et al. (2021). “MOLECULE OPTIMIZATION BY EXPLAINABLE EVOLUTION”. en. In: p. 15.

Gottipati, Sai Krishna et al. (Apr. 26, 2020). “Learning To Navigate The Synthetically Accessible Chemical Space Using Reinforcement Learning”. In: arXiv:2004.12485 [cs]. tex.ids= gottipati2020learninga, gottipatilearning. arXiv: 2004.12485. URL: http://arxiv.org/abs/2004.12485 (visited on 04/29/2020).

Jensen, Jan H. (Mar. 2019). “A Graph-Based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space”. en. In: Chemical Science 10.12, pp. 3567–3572. ISSN: 2041-6539. DOI: 10.1039/C8SC05372C.

Lehman, Joel et al. (Nov. 2019). “The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities”. In: arXiv:1803.03453 [cs]. arXiv: 1803.03453 [cs]. URL: http://arxiv.org/abs/1803.03453 (visited on 03/16/2020).

Nguyen, Anh, Jason Yosinski, and Jeff Clune (Apr. 2015). "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images". In: arXiv:1412.1897 [cs]. arXiv: 1412.1897 [cs]. URL: http://arxiv.org/abs/1412.1897 (visited on 03/10/2020).

Olivecrona, Marcus et al. (Sept. 2017). "Molecular De-Novo Design through Deep Reinforcement Learning". In: Journal of Cheminformatics 9.1, p. 48. ISSN: 1758-2946. DOI: 10.1186/s13321-017-0235-x.

Renz, Philipp et al. (Dec. 1, 2019). “On failure modes in molecule generation and optimization”. In: Drug Discovery Today: Technologies. Artificial Intelligence 32-33. tex.ids= renz2019failurea, renzFailureModesMolecule2019a, pp. 55–63. ISSN: 1740-6749. DOI: 10.1016/j.ddtec.2020.09.003. URL: https://www.sciencedirect.com/science/article/pii/S1740674920300159 (visited on 03/08/2021).

Segler, Marwin H. S. et al. (Jan. 2017). “Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks”. In: arXiv:1701.01329 [physics, stat]. arXiv: 1701.01329 [physics, stat]. URL: http://arxiv.org/abs/1701.01329 (visited on 09/17/2018).

Xie, Yutong et al. (Mar. 2021). “MARS: Markov Molecular Sampling for Multi-Objective Drug Discovery”. In: arXiv:2103.10432 [cs, q-bio]. arXiv: 2103.10432 [cs, q-bio].