Vanilla-Flavored Vanilla RNN: Learning to Generate Food Molecules with Loop-Regularized Character RNN

Alex Shypula, Mert İnan, Shubhakar R. Tipireddy, Yi-Yuan Lee
Carnegie Mellon University
{ashypula, merti, stipired, yiyuanl}@andrew.cmu.edu

11-785 INTRODUCTION TO DEEP LEARNING

Introduction & Motivation

Historically, there have been few applications of deep learning in the space of food, odor, and smell. One significant challenge has been the bottleneck of methods and quality data.

One area within computational chemistry that has benefitted from deep learning recently has been de-novo drug discovery: utilizing generative models to create candidates for novel drugs.

However, one of the major challenges faced in real-world deep-learning applications of molecule generation; be it drug molecules, or scent/odor/food compounds is data shortages: there are only few compounds known to have a scent like vanilla; there are only a few compounds that can cure a rare form of cancer.

In our project we explored how existing methods to generate drugs would transfer over to food molecule generation. In doing so, we also hoped to explore the broader issue of how we can address the "small data" issue that all methods in drug discovery and AI-based molecule generation face.

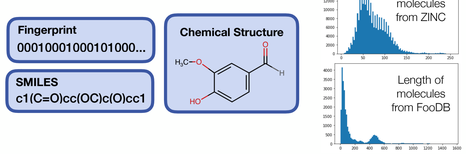
In doing so, we were able to successfully generate thousands of potential food-like compounds and created a novel data augmentation trick.

Related work

In our literature review we saw most papers focused on applying generative models for novel drug discovery. Within the scent space, Sanchez-Lengling et al. explored how chemical properties could be used to predict scent compounds; however did not explore scent generation models. Within the space of drug molecule generation, we reviewed RNN-based methods, adversarial models, and RL-based adversarial methods.

Dataset

The dataset of food molecules used came from FooDB, containing 24,171 molecules in SMILES format. We divided it into a 70/10/20 train/validation/test split .
Because the "small data" problem, we sampled 500,000 drug-like chemical compounds from the ZINC database for our pre-training experiment.



Methods

Models

We have applied three models as our baseline methods, and we improve on them using two strategies.

Character-level Recurrent Neural Network (CharRNN): SMILES data is considered as a regular string, and given a seed atom, it predicts the next atom until a stop character using a vanilla RNN architecture.

Adversarial Autoencoder (AAE): SMILES data is inputted to an autoencoder that generates a latent space and a discriminator is used to decide between the prior distribution and the generated molecule.

Objective-Reinforced Generative Adversarial Network (ORGAN): First synthetic data is generated and then using reinforcement learning, objectives are enforced on the latent space.

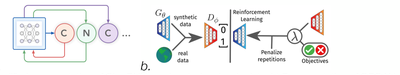


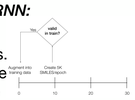
Figure 1: a) Simplified representation of CharRNN on chemical molecule data. b) Representation of ORGAN architecture [2,4].

Improvement with Pre-training Transfer Learning on CharRNN:

Due to the small size of the FooDB dataset, we experimented with a pre-training phrase to learn the "syntax" of molecules, and "transferred" it to learning the food chemicals. We first trained the CharRNN for 40 epochs, and selected a promising model, and retrained on the food dataset for another 40 epochs.

Improvement with Regularization Loop on CharRNN:

After pre-training, the CharRNN still suffered from instability in generating valid chemical compounds. We tried a novel regularization trick during the "fine tuning" phase of training, where for 10 epochs we generated 5,000 compounds, and fed only valid and novel compounds back into the training set.



Metrics

We use the following metrics to evaluate the similarity between generated compound and compounds in the database.

Fragment similarity calculates the similarity of fragment frequency between two compounds.

$$Frag(G,R) = 1 - cos(f_G, f_R)$$

Scaffold similarity calculates the similarity of scaffolds between two compounds.

$$Scaff(G,R) = 1 - cos(s_G, s_R)$$

Nearest neighbor similarity calculates the similarity of the average Tanimoto similarity between compounds.

$$SNN(G,R) = \frac{1}{G}\sum_{m_G \in G} \max_{m_R \in R} T(m_G, m_R)$$

Internal diversity evaluates the diversity of generated molecules. This metric detects mode collapse, which makes the model produce a limited variety of samples, ignoring some areas of chemical space.

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

Synthetic Accessibility Score estimates how easy (1) or how hard (10) a compound can be synthesized based on fragment and structural complexity.

$$SAscore = fragmentScore - complexityPenalty$$

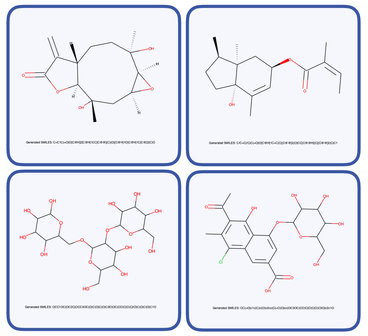
Results

	Baseline Models			Improvement Schemes	
	CharRNN	AAE	ORGAN	CharRNN & Pre-training	CharRNN & Regularization
Unique	78.1%	99.5%	99.5%	83.0%	96.9%
Novelty	26.5%	-	-	54.1%	87.2%
Valid Compounds	87.6%	99.7%	98%	87.9%	93.7%
Fragment Similarity	98.6%	99.5%	18%	98.7%	98.6%
Scaffold Similarity	70.3%	86.6%	NaN	70.7%	66.6%
NN Similarity	65.0%	57.5%	13.3%	63.1%	54.0%
Internal Diversity	88.1%	85.6%	55.5%	88.1%	89.1%
Synthetic Accessibility Score	0.072	0.0005	12.03	0.063	0.114

Table 1: This table shows the results and the evaluation metrics of the results for the three different baseline models and improvements on the CharRNN architecture.

The results show that the highest novel compounds are generated by the improved CharRNN model with regularization. It is also observable that ORGAN is the worst performing model among all. In terms of synthetic accessibility score, AAE performs the best. These all show that among the baseline models AAE has the highest scores in most categories and regularization of CharRNN mainly improves the novelty.

In addition to the evaluation metric results, here we present a sample of novel molecules generated by the CharRNN model,



It is observable in figure 3 that the regularized CharRNN is not running on its desired chemical space, but it is scalable nonetheless, as the novel molecules generated are slightly more after several epochs.

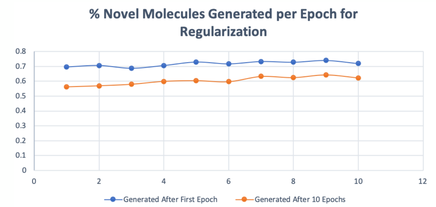


Figure 3: The plot of percentage of novel molecules generated after the first epoch of loop-regularized CharRNN and after 10 epochs.

Discussion

In the CharRNN experiments, we generally saw improvements in the quality of generated molecules through pre-training on drug-like compounds and additionally through our regularization trick.

It seems that the regularization trick enabled the CharRNN model to generate more novel compounds and more chemically valid compounds, which likely came from the increased number of samples in our training sample. However, we also witnessed degraded nearest neighbors and scaffold similarity, possibly because the model may have "remembered" drug like compounds from pre-training during the generation phase.

ORGAN generated very unique compounds, but they are neither similar to ZINC and FooDB chemical space nor easy to be synthesized. We hypothesize that this may be due to mode-collapse.

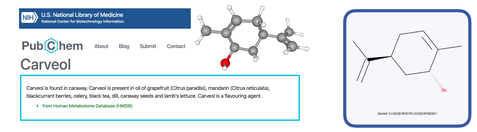
The adversarial autoencoder model displayed strong statistics; however, we hypothesize that it may have "remembered" the training set, and may potentially have degraded novelty potential.

Conclusion

Our experiments were able to show we could adapt state-of-the-art techniques for de-novo drug discovery to the domain of food compounds; below we show a compound we generated that our model was never trained on.

With our CharRNN experiments, we saw that despite the limited amount of food-compounds, through transfer learning and through our regularization trick we were able further stabilize our generative models without sacrificing much novelty or quality in generated output.

We believe that our regularization trick could be useful in other generative molecular tasks with "small data" issues



Reference

[1] Daniil Polykovskiy et al. "Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models", arXiv:1811.12823 [cs.LG]
[2] M. H. S. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks," ACS Central Science, vol. 4, no. 1, pp. 120–131, 2017.
[3] A. Zhavoronkov et al., "Deep learning enables rapid identification of potent DDR1 kinase inhibitors," Nature Biotechnology, vol. 37, no. 9, pp. 1038–1040, 2019.
[4] Gabriel Lima Guimaraes et al. "Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models"