DiversityNet: a collaborative benchmark for generative AI models in chemistry

Mostapha Benhenda[1], Esben Jannik Bjerrum[2], hsiao yi[3], and chintan zaveri[4]

[1]StartCrowd, online AI lab
[2]Wildcard Pharmaceutical Consulting
[3]National Taiwan University
[4]The Maharaja Sayajirao University of Baroda

March 5, 2018

Commenting on the document is possible without registration, but for editing, you need to:

- *Register on Authorea: https://www.authorea.com/*
- *Join the DiversityNet group: https://www.authorea.com/inst/18886*
- *Come back here*

- Code: https://github.com/startcrowd/DiversityNet
- Blog post: https://medium.com/the-ai-lab/diversitynet-a-collaborative-benchmark-for-generative-ai-models-in-chemistry-f1b9cc669cba
- Telegram chat: https://t.me/joinchat/Go4mTw0drJBrCdal0JWu1g

Generative AI models in chemistry are increasingly popular in the research community. They have applications in drug discovery and organic materials (solar cells, semi-conductors). Their goal is to generate virtual molecules with desired chemical properties (more details in this blog post).

However, this flourishing literature still lacks a unified benchmark. Such benchmark would provide a common framework to evaluate and **compare** different generative models. Moreover, this would allow to formulate **best practices** for this emerging industry of 'AI molecule generators': how much training data is needed, for how long the model should be trained, and so on.

That's what the **DiversityNet benchmark** is about. DiversityNet continues the tradition of data science benchmarks, after the MoleculeNet benchmark (Stanford) for predictive models in chemistry, and the ImageNet challenge (Stanford) in computer vision.

Diversity metrics

Designing evaluation metrics is an important part of the challenge. These metrics assess the quality and diversity of generated samples. Here, contributions from medicinal chemists and statisticians are especially welcome.

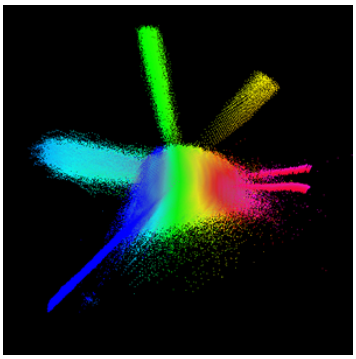

Figure 1: Chemical space

Measures of diversity are based on distance metrics in the chemical space. This distance tells when two molecules are chemically close to each other. The most popular distance is the Tanimoto distance on Morgan fingerprints. It's not necessary to get into details of the definition, the point is that those fingerprints are hand-crafted features, and it's probably better to replace them with deep learning features, as suggested in the MoleculeNet benchmark.

Let's denote:

- T_d the distance in the chemical space.
- A the set of generated molecules with desired properties. Its size is noted $|A|$.
- B the training set.

Let's define:

- **Nearest neighbor diversity**: it's the average distance between a generated molecule in A and its nearest neighbor in the training set B. The formula is:

$$NN(A, B) = \frac{1}{|A|} \sum_{x \in A} \min_{y \in B} T_d(x, y)$$

- **Internal diversity**: it's the average distance of desired generated molecules with each other. The formula is:

$$I(A) = \frac{1}{|A|^2} \sum_{(x,y) \in A \times A} T_d(x, y)$$

For more discussion about those two metrics, see my previous paper.

Variance vs. entropy

This internal diversity formula is essentially a **variance** (without the square). However, variance can be a poor measure of diversity, when data is clustered in few distant regions.

In this case, another measure of diversity is better: **entropy**.

To see why, it helps looking at the simplified case when data labels are discrete (like in classification situations). In this case, entropy is higher when data is spread in a lot of categories: for N equi-distributed categories, the entropy is equal to $\log(N)$, which is an increasing function of N.

This reasoning can be generalized to our setting, where data lives in a continuous and high-dimensional space, using differential entropy.

Let the **kernel density estimator** be:

$$\hat{f}_h(x) = \frac{1}{|A|} \sum_{y \in A} K_h(T_d(x, y))$$

where $K_h(x) \simeq \exp(-x^2/h^2)$ is the Gaussian Kernel of bandwidth $h > 0$ (the value of h remains to be determined).

Then consider an **embedding** of the chemical space into the **unit sphere of dimension n**, (typically, $n \tilde{} 300$). For example, Mol2Vec or Smiles2Vec (PNNL). Such embedding is analogous to Word2Vec in Natural Language Processing. Then we can identify a molecule in the chemical space with its image in the sphere by the embedding, and we can compute the entropy of A with the formula:

$$H(A) = -\int_{\mathbb{S}^\kappa} \hat{f}_h(x) \log \hat{f}_h(x)\, dx$$

where dx is the spherical measure on $\mathbb{S}^\kappa$. We need an embedding into the unit sphere because there is no probability measure on the chemical space.

Note that this entropy depends on the choices of the bandwidth h and of the embedding.

Earth Mover Distance with a reference dataset

Another measure of internal diversity is to compare the set of generated samples with a reference set, which is known to be diverse *a priori*. For example, the ZINC dataset seems suitable. Chemists can propose alternative reference datasets.

The idea is to take a random subset of the reference set with the same size as the generated set. Then to consider those two sets as two piles of sand in the chemical space, and measure the energy necessary to move the first pile into the second pile (this measure is known as Earth Mover Distance in statistics, and Wasserstein metric in mathematics).

Get inspiration from computer vision

To find out better metrics in chemistry, it helps having a look at related metrics in computer vision (GAN and stuff). However, it is important to keep in mind that goals are different: in chemistry, the goal is to generate molecules with new and nice properties, while in computer vision, the goal is to reconstruct training data.

- **Inception Score** (OpenAI): This metric uses the Inception predictive model, which is a standard image classifier (a winner of the ImageNet challenge). A generative model has a high Inception score when the Inception model is very confident that generated images belong to a particular ImageNet category, and when all categories are equally represented. This suggests that the generative model has both high quality and diversity.

- **Fréchet Inception Distance** (Linz University): it computes a distance between distributions of the training data and of the generated data. See their Fréchet ChEMBLNet distance.

There are many other evaluation metrics, and even evaluations of evaluations metrics (Cornell).

Recreation of fully enumerated chemical libraries

The Reymond research group has created a number of fully enumerated chemical libraries with different limits on number of compound atoms and atom types. `https://pubs.acs.org/doi/abs/10.1021/ci300415d`

The compounds are not "drug-like", but the systematic coverage allows for systematic experimentation and simple metrics to explore the properties of the generative methods. They are available for download `http://gdb.unibe.ch/downloads/`. The smaller ones could support fast experimentation and the larger ones for more extended exploration. Using a training set, it can be measured how much of the remaining space is recreated as a simple percentage (probably as a function of number of samples molecules). It can also be measured if the networks create molecules outside the training space (e.g. creating molecules with 7 or 9 atoms when they are trained on GDB-8.) The recreation of the remaining space can be tuned using hyper parameter search, and it is possible to make experiments of how much of the chemical space needs to be covered to get a decent recreation of the remainder.

Tasks

To perform the benchmark, it's good to start with tasks already done in the literature. Also, it is interesting to evaluate the same model across a large variety of tasks (to avoid overfitting a particular task).

Multi-objective tasks are more realistic, but more difficult than single-objective tasks (for example, getting molecules which are active, non-toxic, and synthetizable). It has been tried recently (Peking University).

Here's a list of tasks (tell me if I omitted your paper):

Drug discovery tasks

- Cancer (In SilicoMedicine 1 and InsilicoMedicine 2 (use Sci-Hub to bypass the paywall))
- Targeting the 5-HT2A Receptor (antidepressants), Malaria, staph aureus (AstraZeneca 1)
- Activity on the Dopamine receptor D2: antipsychotics (AstraZeneca 2, AstraZeneca 3)
- Activity on PPAR and RXR: lowers triglycerides and blood sugar (ETH Zurich 1, ETH Zurich 2).

- Inhibition of JAK2: cancer, inflammatory diseases, various skin conditions, and autoimmune diseases (University of North Carolina)
- Joint inhibition of JNK3 and ΓΣK3β: Alzheimer disease (Peking University)

Organic materials tasks

- Organic solar cells: Power Conversion Efficiency (Harvard 3)
- Organic semi-conductors: HOMO–LUMO gap (University of Tokyo, Denmark Tech (use Sci-Hub for the paywall))

Participants can also propose their own favorite objectives. In any case, I think it is better to consider at least one specific real-world application, and not just generate 'drug-like' molecules, as in those preliminary papers by Harvard 1, Google/Cambridge, Paris-Saclay, Wildcard, Harvard 2, Novartis, Georgia Tech.

Data

This DiversityNet benchmark is based on publicly available data, like all the papers cited above. In most papers, data is taken from:

- PubChem
- ChEMBL
- ExCAPE-DB, which aggregates PubChem and ChEMBL.
- ZINC

Many papers only use small datasets (including mine), and in some way, that's bad. Model pre-training should be made on a large dataset.

Even better, different pre-training set sizes could be tested (5K, 10K, 15K, 30K, 50K, 100K, 250K, 1M) to understand how performance of the generative model changes (that's a suggestion from an anonymous referee of my paper).

Besides small molecules chemistry, the same generative models can be used for other tasks related to drug discovery: for RNA sequences (University of Tokyo), for DNA sequences (University of Toronto) and for proteins (Harvard 4, ETH Zurich 3). However, I think it is better to keep those non-chemistry tasks for separate benchmarks: DiversityNet-genetics and DiversityNet-proteins.

Generative models

It's good to start with models already tried in the literature:

- Variational auto-encoder: Harvard 1, Alan Turing Institute, AstraZeneca 3, Georgia Tech, Denmark Tech (use Sci-Hub for the paywall)
- Adversarial auto-encoder: In SilicoMedicine 1, InsilicoMedicine 2(DruGAN) (use Sci-Hub to bypass the paywall), AstraZeneca 3
- Recurrent Neural Networks(RNN): Paris-Saclay, Wildcard
- Reinforcement Learning (RL)+ RNN: Google, AstraZeneca 1, AstraZeneca 2, University of Tokyo, ETH Zurich 1, University of North Carolina, Novartis, ETH Zurich 2
- RL+ RNN+ Generative Adversarial Networks (GAN): Harvard 2 (ORGAN), Harvard 3 (ORGANIC)
- Conditional Graphs: Peking University

For GAN, there are different flavors: Wasserstein-GAN (Facebook), Cramer-GAN (DeepMind), Optimal Transport-GAN (OpenAI), Coulomb-GAN (Linz University), although at the end, maybe they are all equal (Google).

You can also find more in the **Natural Language Processing** literature (and apply them to SMILES):

- Texygen benchmark (Shanghai University)
- MaskGAN (Google)
- ACtuAL (University of Montreal)
- ARAE (New York University)
- Adversarial Generation of Natural Language (University of Montreal) (and don't miss the adversarial review)
- MaliGAN (University of Montreal)
- RankGAN (University of Washington)
- GSGAN (Alan Turing Institute)
- TextGAN (Duke University)
- LeakGAN (Shanghai University)

Finally, it will be interesting to design a systematic procedure for testing hyperparameter values. These methods are often very sensitive to hyperparameter choice (another suggestion from an anonymous referee of my previous paper).