Genenerative AI Models In Small Molecule Drug Discovery: The Open Challenge To Create A Unified Benchmark

Feb. 11, 2018 by Mostapha Benhenda

Generative AI models in chemistry are increasingly popular in the research community, mainly, due to their interest for drug discovery applications. They generate virtual molecules with desired chemical and biological 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, it would help 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.

In this post, I will provide a brief review of the current example of using generative models for drug discovery applications and introduce **DiversityNet benchmark,** which might be the answer to some of the current research needs in this area. 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.

However, there's a novelty in this challenge: in traditional data science competitions (ImageNet, Kaggle…), participants are divided into small teams, who compete against each other to submit the best model. This 'best' model is defined relative to a well-defined metric, and with well-defined tasks and datasets.

On the other hand, DiversityNet is a **challenge** '**in the wild'**, which means that part of the challenge is to properly define those tasks, metrics, and datasets. As a result, instead of competing against each other, participants are invited to **collaborate like one big team**, for both the design and execution of the benchmark. Their joint output is a research paper, together with open-source code and data. The challenge is open to all, and potential participants include data scientists and medicinal chemists, whose skills are complementary.

Competitive challenges are a winner-take-all game, while in a **collaborative challenge**, sponsors distribute prize shares proportionally to individual contributions. For more details, see the **call for**

sponsors below.

Writing a paper collaboratively has been done recently for a review in deep learning for medicine, using GitHub. But for writing the DiversityNet paper, GitHub is an inappropriate tool, because GitHub does not natively support math formulas.

There are many alternatives (GitLab, Overleaf…), and I suggest trying Authorea, which allows real-time collaboration between LaTeX users (data scientists, for example) and non-LaTeX users (medicinal chemists, for example). Here is the draft, editable and citable:

DiversityNet: a collaborative benchmark for generative AI models in chemistry

Diversity metrics

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

Measures of diversity are based on measures of distance in the chemical space. These distances tell when two molecules are 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:

- *Td* the distance in the chemical space.
- *A* the set of generated molecules with desired properties.
- *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:

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

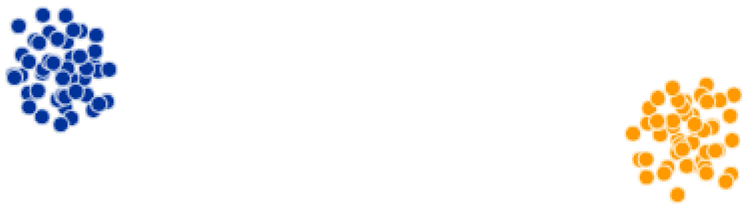
- **Internal diversity**: it’s the average distance of generated molecules in *A* with each other. The formula is:

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

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.



High variance, but data is clustered

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

To see why it helps let's look at the simplified case when data labels are discrete (like in classification tasks). 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. To learn more, take a look at the **DiversityNet draft paper**.

Earth Mover Distance with a reference dataset

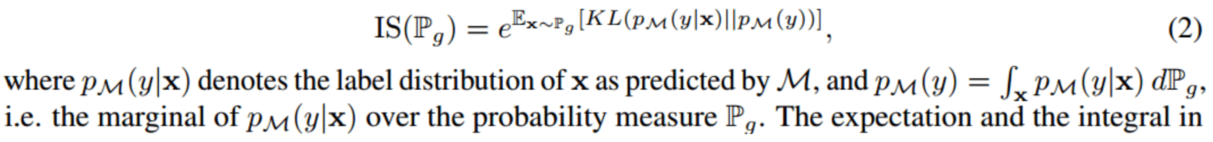
Another way to measure internal diversity is to compare the set of generated samples with a reference set, which is known to be diverse *a priori*. For example, the ZINC dataset seems suitable. Chemists can propose alternative reference datasets.

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 to 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.

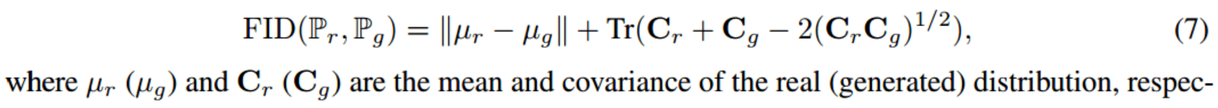
- **Inception Score** (OpenAI): This metric uses the Inception predictive model, which is a standard image classifier (winner of the 2014 ImageNet challenge). A generative model has a high Inception Score when the Inception model is very confident that generated images belong to a particular ImageNet category, and when all categories are equally represented. This suggests that the generative model has both high quality and diversity.

- 

$$\mathrm{IS}(\mathbb{P}_g) = e^{\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_g}[KL(p_{\mathcal{M}}(y|\mathbf{x})||p_{\mathcal{M}}(y))]}, \qquad (2)$$

where $p_{\mathcal{M}}(y|\mathbf{x})$ denotes the label distribution of $\mathbf{x}$ as predicted by $\mathcal{M}$, and $p_{\mathcal{M}}(y) = \int_{\mathbf{x}} p_{\mathcal{M}}(y|\mathbf{x})\, d\mathbb{P}_g$, i.e. the marginal of $p_{\mathcal{M}}(y|\mathbf{x})$ over the probability measure $\mathbb{P}_g$. The expectation and the integral in

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



$$\mathrm{FID}(\mathbb{P}_r, \mathbb{P}_g) = \|\mu_r - \mu_g\| + \mathrm{Tr}(\mathbf{C}_r + \mathbf{C}_g - 2(\mathbf{C}_r\mathbf{C}_g)^{1/2}), \qquad (7)$$

- where μ_r (μ_g) and $\mathbf{C}_r$ ($\mathbf{C}_g$) are the mean and covariance of the real (generated) distribution, respec-

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

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 synthesizable). It has been tried recently (Peking University). For a general introduction to Multi-objective deep reinforcement learning, see (Oxford).

Here’s a list of tasks (write in the comments below if I omitted your paper):

Drug discovery tasks

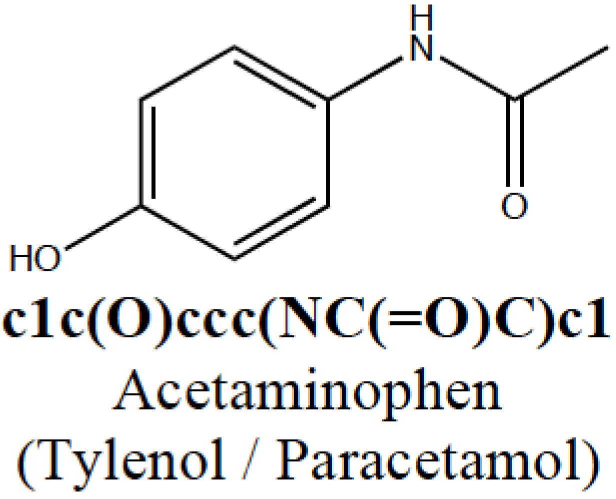
- Cancer (In SilicoMedicine 1 and InsilicoMedicine 2 (use Sci-Hub to bypass the paywall))
- Targeting the 5-HT2A Receptor (antidepressants), Malaria, staph aureus (AstraZeneca 1)
- Activity on the Dopamine receptor D2: antipsychotics (AstraZeneca 2, AstraZeneca 3)
- Activity on PPAR and RXR: lowers triglycerides 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 GSK3β: Alzheimer disease (Peking University)

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 1, 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



Typical small molecule

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. This will require intensive computations, and here, the generosity of cloud sponsors is important.

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 two separate benchmarks (that can be run in parallel, if participants and sponsors ask): **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 1, 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 (University of Montreal), Cramer-GAN (DeepMind), Optimal Transport-GAN (OpenAI), Coulomb-GAN (Linz University), although at the end, maybe they are all equal (Google 2).

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

- Texygen benchmark (Shanghai University)
- MaskGAN (Google)
- VGAN (Beihang University)
- 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)

Due to this flood of publications from non-chemistry fields, I think it would be inefficient to build a dedicated library for DiversityNet. That's another difference with MoleculeNet, who is building the DeepChem library. The problem with it is that there are long delays between model publication and integration into the chemistry-specific library. Instead, I suggest that practitioners should learn to use a general library like TensorFlow, PyTorch or Keras. There are many beginner-friendly courses and tutorials online.

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).

Computational resources

For computations, you can use GPU from cloud sponsors, when these resources will be available. In the meantime, you can start small experiments using up to 2 GPU for free, with Microsoft Azure $200/30 days trial, if you have a Visa/Mastercard payment card. (notably, Google Cloud excludes GPU of their free trial [edit: but apparently, you can use the Google ML engine]). I used Microsoft Azure in my paper (and their Data science Linux VM), and it was fine.

Academic prize: publish in a high-impact journal

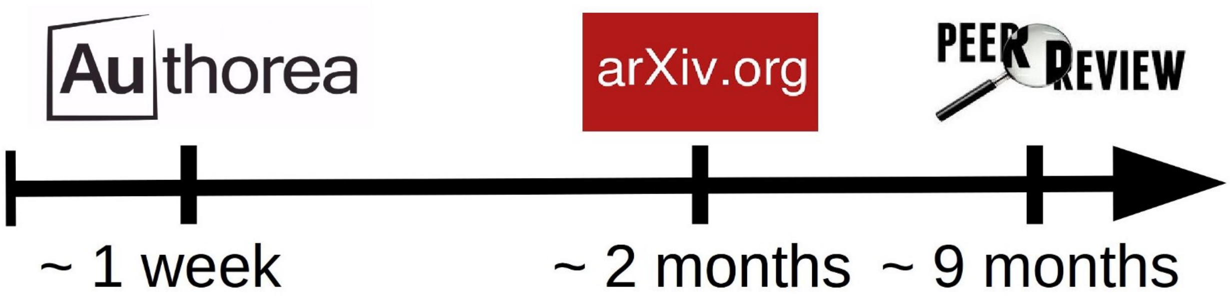
If goals are met, the DiversityNet paper has reasonable chances to get published in a high-impact journal (for example, the Journal of Cheminformatics). Being a co-author of a good paper can be useful at any stage of the academic career, from undergraduate students applying for graduate school, up to tenured professors applying for a research grant. Scientific publications are the lifeblood of academic life.

However, since this challenge is 'in the wild', this academic prize is not 100% guaranteed: it's still possible to get scooped by a traditional lab, posting a preprint earlier on ArXiv (for example, Harvard is working on this topic for months).

However, this risk remains pretty low. At the flag-planting game, Authorea (or any 'GitHub for papers') is a better weapon than ArXiv: it allows to release micro-contributions quickly and often (like on GitHub), whereas ArXiv requires to write a whole PDF paper, which is a hassle.

Moreover, papers can be 'forked', which makes easier to build upon them. With ArXiv, a new paper needs to be written from scratch.

As a result, iteration cycle is shorter, and idea dissemination is accelerated.



Iteration cycle is shorter with open collaborative writing

To avoid "idea stealing" within the (potentially large) community of co-authors, it is strongly recommended to use public and timestamped communication channels (GitHub, Telegram…), so that **collaboration is possible without the need of mutual trust**. These communication records can be used to attribute credit individually (i.e. who planted his micro-flag first).

Financial prizes: call for sponsors

Financial incentives are helpful, especially for non-academics, who don't have the pressure to publish research articles. **Money matters**: there are huge crowds of data scientists competing for prizes on platforms like Kaggle. On the other hand, the MoleculeNet benchmark does not attract participants beyond its core contributors at Stanford (who are paid to do this job), due to the absence of financial prize.

Non-financial contributions are also welcome, especially GPU resources on the cloud. It would also be nice to have a cloud infrastructure for multiple GPUs, like big labs have (see OpenAI).

Sponsors can **fill the information form.**

Non-profits & philanthropists

Generative AI models in chemistry have the potential to **benefit humanity** in many different ways. As a result, they can attract funding from various non-profit foundations and philanthropic persons.

At some point, AI might provide new treatments for various incurable diseases. Moreover, by keeping research free to the maximum, this might hopefully reduce the price of those AI-generated drugs, and avoid a situation where Pharma and biotech companies are too dependent on costly proprietary platforms.

For-profits

Like non-profit organizations and philanthropists, private companies can also want to benefit humanity. At the same time, sponsorship can help their business agenda for:

- **Precompetitive collaboration**: open innovation is a way to pool resources. For Pharma and biotech companies, it can mean better virtual screening and better leads.
- **Recruitment** and **brand exposure**: like an online hackathon, it helps to identify, attract and recruit talent. Sponsorship can also increase engagement with a developer ecosystem (API, cloud…).

How money will be spent

It’s a collaboration, so there’s no sense to elect a ‘winner’ like in competitions. Here, money should be spent to encourage fast sharing of information and accumulation of knowledge.

For example, sponsors can nominate a jury of experts, who **split the money among participants**, proportionally to contributions. Experts make evaluations based on public and timestamped communication records.

The jury needs to dig into the project, and **such review is subjective**. For evaluating complex tasks, human judgment is probably unavoidable, and can't be replaced by an automated metric yet.

On the other hand, this human evaluation is not arbitrary either: to maximize impact, sponsors should **treat participants fairly**, in order to maximize their motivation and **productivity** (see Equity theory). That's the way to get big bangs for bucks.

Successful Organizations
Have an ethos of 'Fair Treatment' in how they deal with people
This enabies the organization to thrive as employees are more content

Conclusion

To contribute to the DiversityNet collaborative benchmark, you can edit the draft on **Authorea**, edit the code on **Github**, or talk on **Telegram**.