Deep Molecular Dreaming: Inverse machine learning for de-novo molecular design with surjective representations

Cynthia Shen [1,⋆]
cynt.shen@mail.utoronto.ca

Mario Krenn[1,2,3,⋆]
mario.krenn@utoronto.ca

Sagi Eppel[1,2,3]
sagieppel@gmail.com

Alán Aspuru-Guzik[1,2,3,4]
alan@aspuru.com

[1]Department of Computer Science, University of Toronto, Canada.
[2]Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Canada.
[3]Vector Institute for Artificial Intelligence, Toronto, Canada.
[4]Canadian Institute for Advanced Research (CIFAR) Senior Fellow, Toronto, Canada.
⋆These authors contributed equally

1 Introduction

The de-novo design of new functional chemical compounds can bring enormous scientific and technological advances. For this reason, researchers in cheminformatics have developed a plethora of A.I. methodologies for the challenging inverse molecular design task [1]. They include deep learning techniques such as variational autoencoders (VAE) [2, 3, 4], generative adversarial networks (GAN) [5, 6], reinforcement learning (RL) [7, 8], and evolutionary techniques such as genetic algorithms (GA) [9, 10, 11].

These methods belong to a category with one particular attribute: the model *indirectly* optimizes molecules for a target property. For example, VAEs and GANs learn to mimic a distribution of molecules from a training set, constructing a latent space that is then scanned to find molecules that optimize an objective function. In the case of RL, the agent learns from rewards in the environment in order to build a policy for generating molecules, which is subsequently used to maximize an objective function. Finally, in GAs, the population is optimized iteratively by applying mutations and selections.

Here, we present preliminary results for PASITHEA[1], a new generative model for molecules inspired by inceptionism techniques [12] in computer vision. PASITHEA is a gradient-based method that optimizes a discrete molecular structure for a target property. We train a neural network to predict chemical properties using a molecular string representation. We then invert the training of the network to generate new variants of molecules. This approach has two significant novelties:

- Molecules are *directly* optimized to a given objective function, sidestepping the learning of distributions and policies, or the application of mutations to a population.
- We can analyse what the regression network has learned about the chemical property by probing its inverse training with test molecules.

[1] PASITHEA is the goddess of relaxation, meditation, hallucinations, and wife of Hypnos, the god of sleep.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Figure 1: Deep dreaming is well-known for creating new dream-like images.

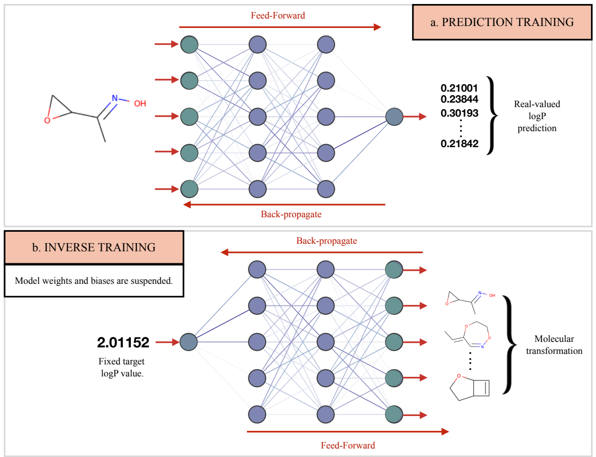


Figure 2: Two-step training for PASITHEA.

Furthermore, PASITHEA does not require expensive function evaluations for quantum chemistry calculations, provided that we use a pre-calculated dataset, since costly chemical properties can be directly optimized.

This method is made possible by the application of SELFIES, a 100% robust molecular string representation [13]. In contrast to SMILES, for which a large fraction of generable strings do not map to valid molecular graphs, SELFIES is a surjective map between molecular strings and molecular graphs.

We train PASITHEA on a subset of QM9 to predict logP values. We then initialize the inverse training with a molecule and optimize it for a logP value. We clearly confirm a shift in the logP distribution of generated molecules. We can observe how the model changes a molecule quasi-continuously over several steps to a final, optimized chemical structure. Finally, we indicate how this technique can be used to probe concepts learned by PASITHEA.

2 Methodology

Inceptionism [12] has drawn considerable attention as an artistic method for rendering images. By using a neural network trained to classify an image (i.e., dog, car or house), the network can perform deep 'dreaming' on an image in order to mutate it gradually to fit a different class while retaining features of the original image. For example, it may enhance animal features in the image of a chemistry lab while the general structure of a lab is still visible (Figure 1). The rendered images have dream-like properties that make them a popular artistic style in the media.

We generalize this methodology to the inverse-design task of functional molecules. PASITHEA uses a fully-connected neural network consisting of four layers, each with 500 nodes, and takes as input the one-hot encoding of the SELFIES representation of each molecular graph.

Prior to deep dreaming, the network learns to predict a specific real-valued property for each molecule in a given dataset (i.e., logarithm of partition coefficient, or logP) from the molecular graph. The training involves the standard feedforward and backpropagation process. For a set of fixed inputs and outputs, the network iteratively improves its predictions by updating the weights through gradient descent. (Figure 2a).

In deep dreaming, an input molecule with a property value predicted by the network is incrementally mutated to a similar molecule with the desired value. The weights and biases of each layer of the network are now fixed; the neural network is no longer adjusting its logP prediction for each molecule, but applying its prediction to the dreaming process. Through backpropagation, we minimize the error between the predicted properties of each input molecule and the desired target property (Figure 2b). The computed error is then used to compute the gradient with respect to the one-hot encoding of the input. This effectively transforms the input gradually to a molecule that matches the target property. Once the loss function has been minimized, the gradient evaluates approximately to zero, which terminates the training. In this process, the same standard feedforward and backpropagation algorithm is used, but the input molecule is adjusted while the weights and biases remain constant.

It is an ongoing study to find the best numerical conversion from the SELFIES string into an appropriate input for deep dreaming. We introduce noise in the one-hot encoding as input to deep dreaming as one effective approach. Every zero in the one-hot encoding is altered to a random number between zero and a specific upper-bound, which is typically set to a value between 0.5 and 0.9. Using this method, we observe an incremental optimization for each given molecular input, as required.

Another important contribution to the model is the application of SELFIES. This method requires a continuous space in which all points are valid, a criterion met by the recently developed SELFIES, which is proven to be 100 % valid [13]. Our findings here highlight only one of the many potential applications of SELFIES.

3 Results

Our experiments clearly indicate that deep dreaming achieves both a direct, gradient-based design of novel functional molecules and the explainability of neural networks for molecules. We did not require an exhaustive search for the ideal training hyperparameters. In this analysis, PASITHEA is trained to predict the logarithm of partition coefficient (logP), obtained from the RDKit library [14], on a set of 10,000 molecules randomly selected from the QM9 dataset. We demonstrate how PASITHEA transforms molecules in a stepwise, quasi-continuous fashion and shifts the distribution of logP in the molecular dataset toward set targets. We then analyse what PASITHEA has learned regarding the relationship between logP and molecular structure.

3.1 Evolution of individual molecules

Of particular interest is the gradual progression of each molecule through inverse training. Over hundreds of training epochs, the gradient with respect to input SELFIES produces minor adjustments in the molecule that increments to a pronounced transmutation (Figure 3). The behaviour of these adjustments are stepwise due to the discrete, textual nature of the molecules represented by strings, but continuous in terms of real-valued one-hot encodings.

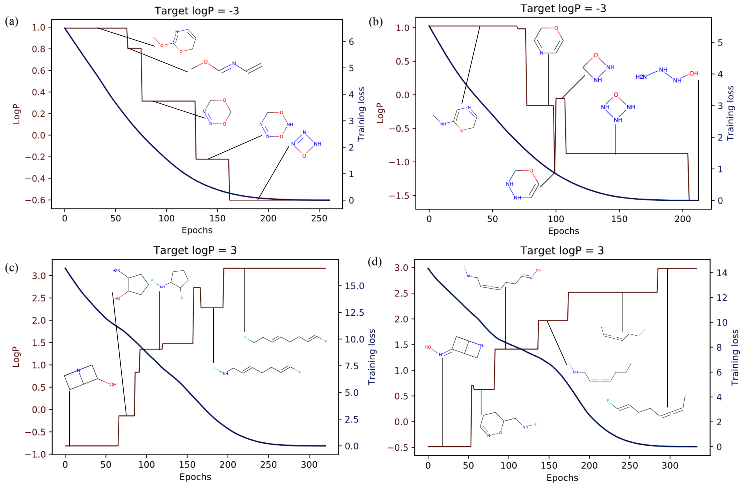
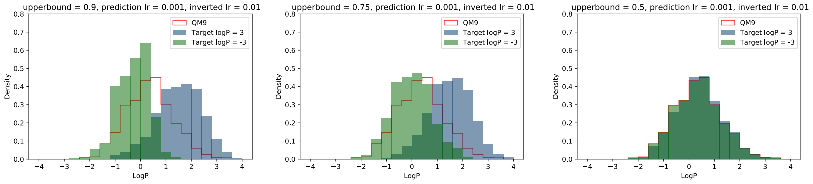


Figure 3: Stepwise molecule transformations, optimized for a higher and lower target logP.



(a) (b) (c)

Figure 4: Shifts in distribution.

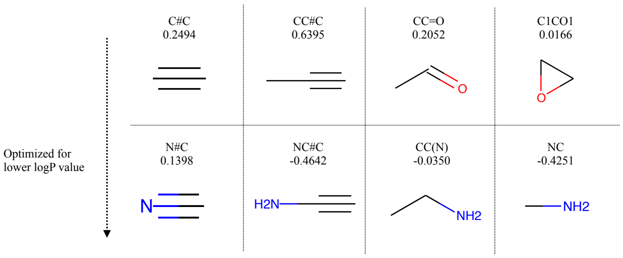


Figure 5: Graphical and SMILES representation of nitrogen and fluorine appendages, with corresponding logP.

3.2 Shift in distribution

In order to observe a large-scale pattern over the entire dataset, we disregard the intermediate molecules and restrict our analysis to the initial and fully optimized molecules. From these results, there is a clear shift in the distribution of logP values in the set of molecules as they transmute toward a given target value (Figure 4). Although the training learning rates have little effect on the quality of training, the addition of more noise to one-hot encoded inputs (higher upper-bound values in Figure 4) has a large influence on the shifts in distribution curves. We furthermore observe from these distribution shifts in Figure 4 that there are some molecules generated with logP values exceeding the lowest and highest values in the original dataset. For instance, notice that in Figure 4a, the left tail of the left-shifted (green) distribution extends beyond the left tail of the original (red) distribution and the right tail of the right-shifted (blue) distribution extends beyond the right tail of the original distribution. Demonstrably, Pasithea is generating novel molecules with properties outside the limits of the original training set of molecules, which attests to the large potential of this method.

3.3 Probing the neural network's intuitions

Inspired by explainable representations in image recognition [15] and rediscovery of concepts in physics [16], we can *understand the internal molecular representation by inverting it*. For that, we probe the neural network with specific test molecules and observe patterns in how it changes them. For example, the composition of atoms after inverse training follows a predictable pattern, such as the appendage of a few non-carbon atoms, fluorine and nitrogen. Take, for example, the transmutations of the simplest molecules in the QM9 dataset (Figure 5), which suggest that PASITHEA interprets these non-carbon atoms as correlated with lower logP values. A similar trend persists for more complex molecules, in which more than one atom may be replaced with nitrogen (Figure 3a), though this persists to a lower extent for fluorine.

The intermediate states during continuous transformation can be used as additional insights into the network's understanding of chemical property. In particular, by observing a single test molecule, there are instances where an additional iteration in inverse-training transforms the molecule with a repeated 'strategy' that has been used in previous iterations. The neural network appears to persist with a single strategy until the training terminates. We demonstrate this behaviour in Figure 3b, which shows a gradual process of reducing length, and in Figure 3a, which shows an initial molecule containing a single nitrogen atom, an intermediate molecule containing two nitrogen atoms, and a final molecule containing three nitrogen atoms. These cases validate that the network is charting deliberate, non-arbitrary paths toward the target logP; it has a non-trivial understanding of features corresponding to higher and lower logP values.

4 Outlook

In the immediate future, we will verify our results on larger datasets and more complex molecules, such as PubChem. Furthermore, we plan to test PASITHEA on molecular properties that require expensive quantum chemistry calculations. We see much potential in discovering other 'strategies' the network may use in order to optimize molecules with different properties.

References

[1] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[2] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[3] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. *arXiv preprint arXiv:1802.04364*, 2018.

[4] Tengfei Ma, Jie Chen, and Cao Xiao. Constrained generation of semantically valid graphs via regularizing variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 7113–7124, 2018.

[5] Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.

[6] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[7] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):1–10, 2019.

[8] Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in neural information processing systems*, pages 6410–6421, 2018.

[9] Jan H Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

[10] AkshatKumar Nigam, Pascal Friederich, Mario Krenn, and Alán Aspuru-Guzik. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. *arXiv preprint arXiv:1909.11655*, 2019.

[11] Emilie S Henault, Maria H Rasmussen, and Jan H Jensen. Chemical space exploration: how genetic algorithms find the needle in the haystack. *PeerJ Physical Chemistry*, 2:e11, 2020.

[12] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015.

[13] Mario Krenn, Florian Hase, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 2020.

[14] Greg Landrum et al. Rdkit: Open-source cheminformatics. 2006.

[15] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[16] Alireza Seif, Mohammad Hafezi, and Christopher Jarzynski. Machine learning the thermodynamic arrow of time. *Nature Physics*, pages 1–9, 2020.