Automatic Chemical Design using Variational Autoencoders

Rafael Gómez-Bombarelli* Harvard University
David Duvenaud* University of Toronto
José Miguel Hernández-Lobato* University of Cambridge

Timothy D. Hirzel Harvard University
Jorge Aguilera-Iparraguirre Harvard University

Ryan P. Adams Harvard University, Twitter
Alán Aspuru-Guzik Harvard University

Abstract

We train a variational autoencoder to convert discrete representations of molecules to and from a multidimensional continuous representation. This continuous representation allow us to automatically generate novel chemical structures by performing simple operations in the latent space, such as decoding random vectors, perturbing known chemical structures, or interpolating between molecules. Continuous representations also allow the use of powerful gradient-based optimization to efficiently guide the search for optimized functional compounds. We demonstrate our method in the design of drug-like molecules as well as organic light-emitting diodes.

1 Introduction

The goal of drug and material design is to propose novel molecules that optimally achieve various measurable desiderata. However, optimization in molecular space is extremely challenging, because the search space is large, discrete, and unstructured. Making and testing new compounds is costly and time consuming, and the number of potential candidates is overwhelming. Only about 10^8 substances have ever been synthesized,[1] whereas the commonly reported range of potential drug-like molecules is 10^{23}- 10^{60}.[2]

Computation offers a way to speed up this search.[3, 4, 5, 6] Virtual libraries containing thousands to hundreds of millions of candidates can be assayed with computational methods, and the most promising leads are selected and tested experimentally.

However, even with accurate simulations,[7] computational molecular design is limited by the search strategies available to explore chemical space. Current methods are either an exhaustive search through a fixed library,[8, 9] or the use of a discrete local search method such as a genetic algorithm[10, 11, 12, 13, 14, 15] or a similar discrete interpolation technique.[16]

A differentiable, reversible, and data-driven representation has several advantages over existing systems. First, hand-specified mutation rules are unnecessary, and new compounds can be generated automatically by modifying the vector representation and decoding. Large chemical databases typically contain millions of molecules, but most properties are nevertheless unknown for most molecules. A data-driven representation can leverage a large set of unlabeled chemical compounds to automatically build an even larger implicit library, and then use the smaller set of labeled examples to build a regression model from the continuous representation to the desired properties. Having a differentiable representation allows the use of gradient-based optimization to leverage geometric information and make larger jumps in chemical space. We can also use Bayesian optimization

methods to select compounds that are likely to be informative about the global optimum. These methods can be combined into a closed loop that proposes new compounds, tests their properties, and uses this new information to suggest even better compounds.

2 Methods

To leverage the power of recent advances in sequence-to-sequence autoencoders for modeling text[17], we use the SMILES[18] representation, a commonly-used text encoding for organic molecules. We also tried InChI[19] as an alternative string representation, found it to perform significantly worse than SMILES, presumably due to a more complex syntax that includes counting and arithmetic.

To enable molecular design, the chemical structures encoded in the continuous representation of the autoencoder need to be correlated to the target properties that need to be optimized. Therefore, based on the autoencoder results, we train a third model to predict molecular properties based on the latent representation of a molecule. To propose promising new candidate molecules, latent vectors of encoded molecules are moved in the direction most likely to improve the desired attribute and these new candidate vectors are decoded.

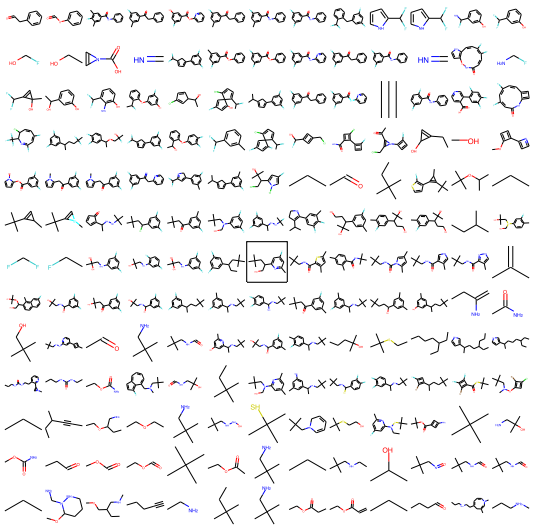

Figure 1: Starting from the molecule in the center, two random, unit-length vectors were followed in latent space for increasingly large displacements. This defines a random two-dimensional plane in the 56-dimensional latent space. A each location in in this two-dimensional subspace, we show the molecule most likely to be decoded at that point in the latent space. Nearby points decode to similar molecules, and distant points decode to a wide variety of compounds.

Bayesian optimization of molecules We trained a sparse Gaussian process (GP) model [20] with 500 inducing points to predict the cost of each molecule from the molecule's feature vector. After this, we perform 10 iterations of Bayesian optimization using the expected improvement (EI) heuristic [21]. On each iteration, we select a batch of 50 latent feature vectors by sequentially maximizing the EI acquisition function. To account for pending evaluations in the batch selection process we use the Kriging Believer Algorithm [22]. That is, after selecting each new data point in the batch, we add that data point as a new inducing point in the sparse GP model with associated target variable equal to the mean of the GP predictive distribution. Once a new batch of 50 latent feature vectors has been selected, each point in the batch is transformed into its corresponding SMILES string using the decoder network. From the SMILES string, we then obtain the corresponding score value using (1).

3 Results

Using variational autoencoders to produce a compact representation. To ensure that every point in the latent space corresponds to a valid molecule, we modified our autoencoder and its

objective into a *variational* autoencoder (VAE) [23]. Using variational autoencoders with RNN encoder and decoder networks was first tried by Bowman *et al.* and we follow their approach closely.[17],

The autoencoder was trained on a dataset with approximately 250,000 drug-like commercially available molecules extracted from the ZINC database.[24] We also tested this approach on approximately 100,000 OLED molecules that have been generated only computationally.[9]

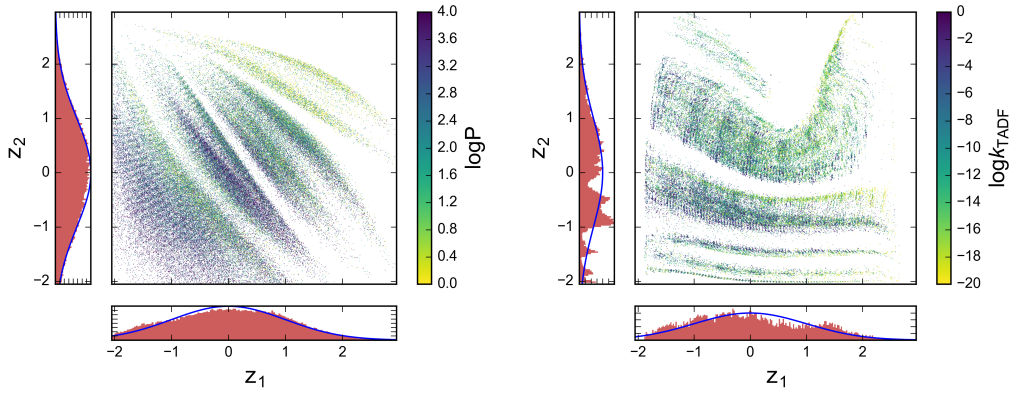

Figure 2: Projection of the molecular training sets onto learned two-dimensional latent spaces. The one-dimensional histograms show the distribution of the training data along each dimension, overlaid with the Gaussian prior imposed in the variational autoencoder. The points are colored along a chemical property that is relevant to their function, and will be the target of optimization experiments. *Left*: A natural library of drug-like molecules, colored by their predicted water-octanol partition coefficient. *Right*: A combinatorially-generated library of organic LED molecules, colored by their predicted delayed fluorescent emission rate (k_{TADF} in μs^{-1}).

Bayesian optimization of drug-like molecules The proposed molecule autoencoder can be used to discover new molecules with desired properties.

As a simple example, we first attempt to maximize the water-octanol partition coefficient (logP), as estimated by RDkit.[25] logP is an important element in characterizing the drug-likeness of a molecule, and is of interest in drug design. To ensure that the resulting molecules to be easy to synthesize in practice, we also incorporate the synthetic accessibility[26] (SA) score into our objective.

Our initial experiments, optimizing only the logP and SA scores, produced novel molecules, but ones having unrealistically large rings of carbon atoms. To avoid this problem, we added a penalty for having carbon rings of size larger than 6 to our objective.

Thus our preferred objective is, for a given molecule m, given by:

$$J(m) = \mathrm{logP}(m) - \mathrm{SA}(m) - \text{ring-penalty}(m)\,, \tag{1}$$

where the scores $\mathrm{logP}(m)$, $\mathrm{SA}(m)$, and $\text{ring-penalty}(m)$ are normalized to have zero mean and unit standard deviation across the training data.

More than half of the 500 latent feature vectors selected by the above process produced a valid SMILES string. Among the resulting molecules, the two best had objective values of 5.02 and 4.68, higher than the best objective value in the training data, 4.52. The right part of Figure 4 shows the empirical distribution of objective values for the molecules in the training data. The two new molecules are shown in the left part of Figure 4.

A high value of the objective (1) does not necessarily translate into a high logP score. However, the logP scores for the molecules from Figure 4 are 8.07 and 8.51, while the highest logP score in the training data is 8.25. Therefore, the second molecule has higher logP score than any other molecule in the training set. This shows that the molecule autoencoder can be combined with Bayesian optimization to discover new molecules with better properties than those found in the training set.

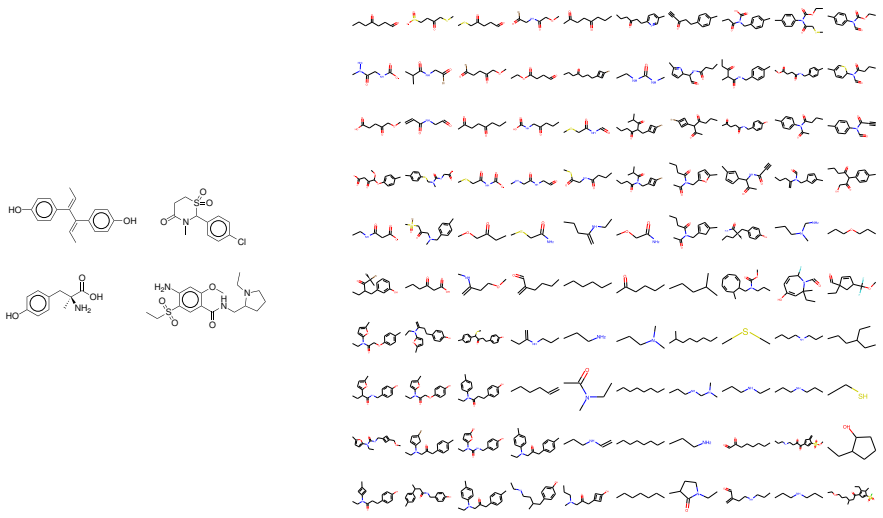

Figure 3: Interpolation. Two-dimensional interpolation between four random drugs. *Left* Starting molecules encoded, whose decodings correspond to the respective four corners of the figure to the right. *Right* Decodings of interpolating linearly between the latent representations of the four molecules to the right.

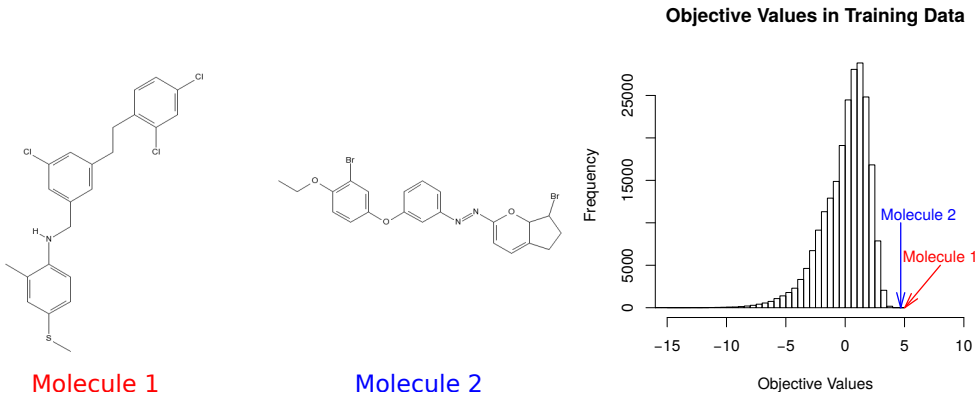

Figure 4: *Left*: Molecules generated by the optimization process with better score values than any other molecule in the training data. *Right*: Histogram of objective values in the training data.

4 Limitations

One problem with the current two-stage learning approach is that the latent representation from unsupervised training might not smoothly map to the property being optimized. A straightforward way to address this problem would be to jointly train on both objectives. Jointly training would encourage the model to find a latent representation which is both easily decoded, and easy to predict with.[27] In addition, we also expect to obtain better generalization by training a larger deep autoencoder with more data. The chemical structures of close to one hundred million chemical compounds are known, and could be used to train a single unified embedding of known chemistry. Software packages that use multiple graphical processing units are being applied to this task.

In this work, we used a text-based encoding of molecules, but using a graph-based autoencoder would have several advantages. However, building a neural network which can output arbitrary graphs is an open problem.

References

[1] Sunghwan Kim et al. “PubChem Substance and Compound databases”. In: *Nucleic Acids Res.* 44.D1 (2016), D1202–D1213. ISSN: 0305-1048.

[2] P. G. Polishchuk, T. I. Madzhidov, and A. Varnek. “Estimation of the size of drug-like chemical space based on GDB-17 data”. In: *J. Comput.-Aided Mol. Des.* 27.8 (2013), 675–679. ISSN: 0920-654X.

[3] Brian K. Shoichet. “Virtual screening of chemical libraries”. eng. In: *Nature* 432 (7019 2004), pp. 862–5.

[4] Thomas Scior, Andreas Bender, Gary Tresadern, Jose L. Medina-Franco, Karina Martinez-Mayorga, Thierry Langer, Karina Cuanalo-Contreras, and Dimitris K. Agrafiotis. “Recognizing Pitfalls in Virtual Screening: A Critical Review”. In: *J. Chem. Inf. Model.* 52.4 (2012), 867–881. ISSN: 1549-9596.

[5] Tiejun Cheng, Qingliang Li, Zhigang Zhou, Yanli Wang, and Stephen H. Bryant. “Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review”. In: *AAPS J.* 14.1 (2012), 133–141. ISSN: 1550-7416.

[6] Edward O. Pyzer-Knapp, Changwon Suh, Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, and Alán Aspuru-Guzik. “What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery”. In: *Annu. Rev. Mater. Res.* 45.1 (2015), pp. 195–216.

[7] Gisbert Schneider. “Virtual screening: an endless staircase?” In: *Nat. Rev. Drug Discov.* 9.4 (2010), pp. 273–276. ISSN: 1474-1776.

[8] Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M Brockway, and Alán Aspuru-Guzik. “The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid”. In: *J. Phys. Chem. Lett.* 2.17 (2011), pp. 2241–2251.

[9] Rafael Gómez-Bombarelli et al. “Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach”. In: *Nat. Mater.* 15 (Aug. 8, 2016), pp. 1120–1127. ISSN: 1476-4660.

[10] Aaron M. Virshup, Julia Contreras-García, Peter Wipf, Weitao Yang, and David N. Beratan. “Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds”. In: *J. Am. Chem. Soc.* 135.19 (2013), pp. 7296–7303.

[11] Chetan Rupakheti, Aaron Virshup, Weitao Yang, and David N. Beratan. “Strategy To Discover Diverse Optimal Molecules in the Small Molecule Universe”. In: *J. Chem. Inf. Model.* 55.3 (2015), pp. 529–537.

[12] Jean-Louis Reymond. “The Chemical Space Project”. In: *Acc. Chem. Res.* 48.3 (2015), pp. 722–730.

[13] Jean-Louis Reymond, Ruud van Deursen, Lorenz C. Blum, and Lars Ruddigkeit. “Chemical space as a source for new drugs”. In: *Med. Chem. Commun.* 1.1 (2010), p. 30.

[14] Ilana Y. Kanal, Steven G. Owens, Jonathon S. Bechtel, and Geoffrey R. Hutchison. “Efficient Computational Screening of Organic Polymer Photovoltaics”. In: *J. Phys. Chem. Lett.* 4.10 (2013). PMID: 26282968, pp. 1613–1623.

[15] Noel M. O’Boyle, Casey M. Campbell, and Geoffrey R. Hutchison. “Computational Design and Selection of Optimal Organic Photovoltaic Materials”. In: *J. Phys. Chem. C* 115.32 (2011), pp. 16200–16210.

[16] Ruud van Deursen and Jean-Louis Reymond. “Chemical Space Travel”. In: *ChemMedChem* 2.5 (2007), pp. 636–640.

[17] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. “Generating Sentences from a Continuous Space”. In: *arXiv preprint arXiv:1511.06349* (2015).

[18] David Weininger. “SMILES a chemical language and information system. 1. Introduction to methodology and encoding rules”. In: *J. Chem. Inf. Model.* 28.1 (1988), pp. 31–36.

[19] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. “InChI - the worldwide chemical structure identifier standard”. eng. In: *J. Cheminf.* 5 (2013), p. 7.

[20] Edward Snelson and Zoubin Ghahramani. “Sparse Gaussian processes using pseudo-inputs”. In: *Advances in neural information processing systems*. 2005, pp. 1257–1264.

[21] Donald R Jones, Matthias Schonlau, and William J Welch. “Efficient global optimization of expensive black-box functions”. In: *J. Global Optim.* 13.4 (1998), pp. 455–492.

[22] Noel Cressie. “The origins of kriging”. In: *Math. Geol.* 22.3 (1990), pp. 239–252.

[23] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[24] John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. “ZINC: A Free Tool to Discover Chemistry for Biology”. In: *J. Chem. Inf. Model.* 52.7 (2012). PMID: 22587354, pp. 1757–1768.

[25] Scott A. Wildman and Gordon M. Crippen. “Prediction of Physicochemical Parameters by Atomic Contributions”. In: *J. Chem. Inf. Comput. Sci.* 39.5 (1999), pp. 868–873.

[26] Peter Ertl and Ansgar Schuffenhauer. “Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions”. In: *J. Cheminf.* 1.1 (2009), pp. 1–11. ISSN: 1758-2946.

[27] Jasper Snoek, Ryan P Adams, and Hugo Larochelle. “Nonparametric guidance of autoencoder representations using label information”. In: *Journal of Machine Learning Research* 13.Sep (2012), pp. 2567–2588.