Automatic chemical design using a data-driven continuous representation of molecules

Rafael Gómez-Bombarelli*[1], David Duvenaud*[2,3], José Miguel Hernández-Lobato*[3,4], Jorge Aguilera-Iparraguirre[1], Timothy D. Hirzel[1], Ryan P. Adams[3,5], and Alán Aspuru-Guzik[1,†]

[1]Department of Chemistry and Chemical Biology, Harvard University
[2]Department of Computer Science, University of Toronto
[3]John A. Paulson School of Engineering and Applied Sciences, Harvard University
[4]Department of Engineering, University of Cambridge
[5]Twitter Inc.
*Equal contributions
†Corresponding author, alan@aspuru.com

October 11, 2016

Abstract

We report a method to convert discrete representations of molecules to and from a multidimensional continuous representation. This generative model allows efficient search and optimization through open-ended spaces of chemical compounds.

We train deep neural networks on hundreds of thousands of existing chemical structures to construct two coupled functions: an encoder and a decoder. The encoder converts the discrete representation of a molecule into a real-valued continuous vector, and the decoder converts these continuous vectors back to the discrete representation from this latent space.

Continuous representations 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]

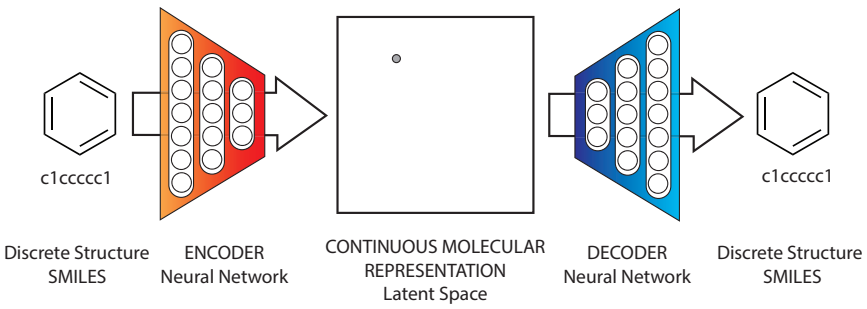

Figure 1: A diagram of the proposed autoencoder for molecular design. Starting from a discrete molecular representation, such as a SMILES string, the encoder network converts each molecule into a vector in the latent space, which is effectively a continuous molecular representation. Given a point in the latent space, the decoder network produces a corresponding SMILES string.

Computation offers a way to speed up this search. [3–6] 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 [10–15] or a similar discrete interpolation technique. [16]

Although these techniques have led to useful new molecules, both approaches still face large challenges. Fixed libraries are monolithic, costly to explore fully and require controlled hand-crafted assembly to avoid impractical chemistries. The genetic generation of compounds requires the empirical, manual specification of heuristics for mutation and crossover rules. Discrete optimization methods have difficulty effectively searching large areas of chemical space, because geometric cues such as gradients are not available to guide the search.

The representation of molecular graphs is at the heart of this challenge. Molecular representations are discrete: they can be considered as either undirected graphs with atoms as nodes and bonds as edges or as three-dimensional arrangement of atoms. For cheminformatic purposes, these representations are usually converted to numerical representations that, ideally, should be rotationally- and translationally-invariant. Representations derived from the graph include chemical fingerprints, [17] convolutional networks on graphs [18] and similar graph-convolutions, [19] or the bag-of-bonds [20] approach. The Coulomb matrix approach, [21], scattering transforms [22], atomic distances [23], ... are based on 3D geometries, usually obtained at an affordable level of theory. Even recently-developed representations based on neural networks cannot be driven in reverse to decode molecules from optimal vectors.

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.

Recent advances in machine learning have resulted in powerful probabilistic generative models that, after being trained on real examples, are able to produce realistic synthetic samples. Such models usually also produce low-dimensional continuous representations of the data being modeled, allowing interpolation or analogical reasoning for natural images [24], text [25], speech, and music [26].

In this work, we propose the use of continuous optimization to produce novel compounds, by building a data-driven, vector-valued representation of molecules. We transform between discrete and continuous representations using a pair of neural networks trained together as an autoencoder. We apply this technique to the design of drug-like molecules and organic light-emitting diodes (OLED).

2 Methods

Initial representation of molecules Before building an encoder to a continuous representation, we must choose which discrete representation to use to represent molecules both before and after encoding. To leverage the power of recent advances in sequence-to-sequence autoencoders for modeling text [25], we use the SMILES [27] representation, a commonly-used text encoding for organic molecules. We also tried InChI [28] 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.

Training an autoencoder Starting from a large library of string-based representation of molecules, we train a pair of complementary recurrent neural networks: an encoder network to convert each string into a fixed-dimensional vector, and a decoder network to convert vectors back into strings. Such encoder-decoder pairs are known as *autoencoders*. The autoencoder is trained to minimize error in reproducing the original string, i.e., it attempts to learn the identity function. The key insight of the autoencoder is to subject this identity map to an *information bottleneck*. This bottleneck — here the fixed-length continuous vector — induces the network to learn a compressed representation that captures the most statistically salient information in the data. We call the vector-encoded molecule the *latent representation* of the molecule.

The character-by-character nature of the SMILES representation and the relative fragility of its internal syntax (opening and closing cycles and branches, allowed valences, etc.) can sometimes result in the output of invalid molecules from the decoder. We employ the open

source cheminformatics suite RDKit [29] to validate the chemical output molecules and discard invalid ones.

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.

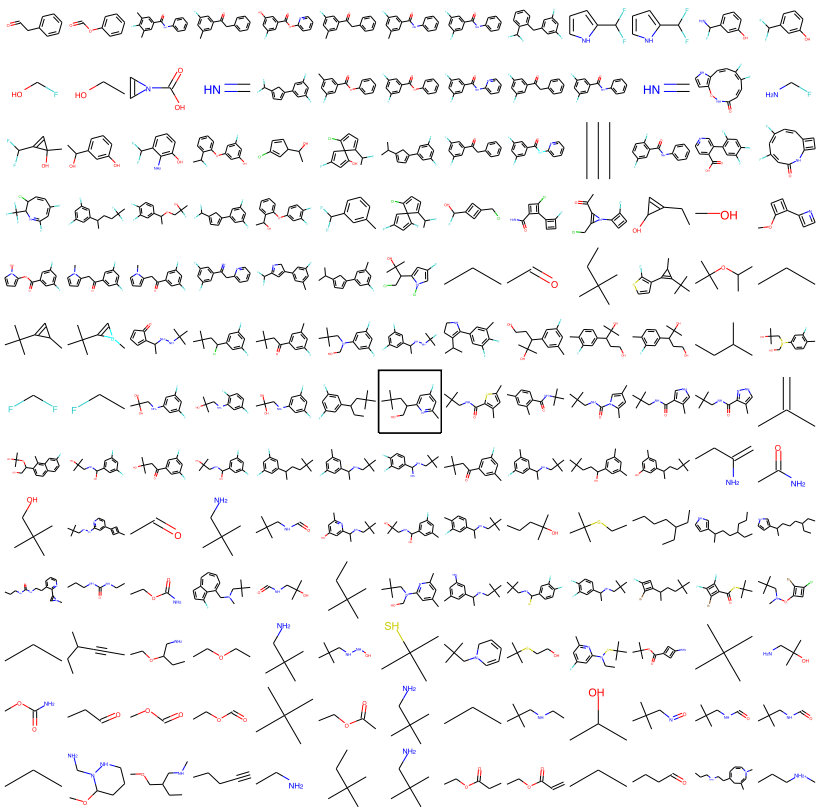

Figure 2: Starting from the molecule in the center, two random, unit-length vectors were followed in latent space for increasingly large displacements. This defines a random two-dimensional plane in the 56-dimensional latent space. A each location in in this two-dimensional subspace, we show the molecule 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.

Autoencoder architecture Strings of characters can be encoded into vectors using recurrent neural networks (RNNs). An encoder RNN can be paired with a decoder RNN to perform sequence-to-sequence learning. [30] We also experimented with using convolutional networks for string encoding [31] and observed improved performance. This is explained by the presence of repetitive, translationally-invariant substrings that correspond to chemical

substructures, e.g., cycles and functional groups.

The SMILES text encoding uses 35 different characters, including blank spaces for padding short strings. We encoded strings up to a maximum length of 120 characters. The structure of the VAE deep network was as follows: The encoder used three 1D convolutional layers of filter sizes 9, 9, 11 and 9, 9, 10 convolution kernels, respectively. Followed by two fully connected layers of dimensions 435 and 292. The decoder started with a fully-connected layer of width 292, fed into three layers of gated recurrent unit networks [32] with hidden dimension of 501. We used the Keras [33] and Theano [34, 35] packages to build and train this model.

The last layer of the RNN decoder defines a probability distribution over all possible characters at each position in the SMILES string. This means that the writeout operation is stochastic, and the same latent point may map to different SMILES strings, depending on the random seed used to sample characters.

Bayesian optimization of molecules We trained a sparse Gaussian process (GP) model [36] 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 [37]. 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 [38]. 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 perform unconstrained optimization in the latent space, most points in the latent space should decode into valid SMILES strings. However, the training objective of the autoencoder does not enforce this constraint, potentially leading to large "dead areas" in the latent space, which decode to invalid or nonsensical SMILES strings.

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) [39]. VAEs were developed as a principled approximate-inference method for latent-variable models, in which each datapoint has a corresponding, but unknown, latent representation. VAEs generalize autoencoders, adding stochasticity to the encoder, and adding a penalty term encouraging all areas of the latent space to correspond to a valid decoding. The intuition is that adding noise to the encoded molecules forces the decoder to learn how to decode a wider variety of latent points. In addition, since two different molecules can have their encodings stochastically brought close in the latent space, but still need to decode to different molecular graphs, this constraint also encourages the encodings to spread out over the entire latent

space to avoid overlap. Using variational autoencoders with RNN encoder and decoder networks was first tried by Bowman *et al.* and we follow their approach closely. [25],

The autoencoder was trained on a dataset with approximately 250,000 drug-like commercially available molecules extracted from the ZINC database. [40] We also tested this approach on approximately 100,000 OLED molecules that have been generated only computationally. [9]

We performed Bayesian optimization over multiple hyperparameters (such as the choice between RNN or CNN encoder, number of hidden layers, layer size, regularization and learning rate) to search the optimal deep autoencoder configuration. [41] We also ran an outer loop of optimization to determine the how small the latent dimension could be while still producing reasonable reconstruction error. After training naïve and variational autoencoders on drug-like molecules, the same structure was also use with organic-light emitting (OLED) molecules.

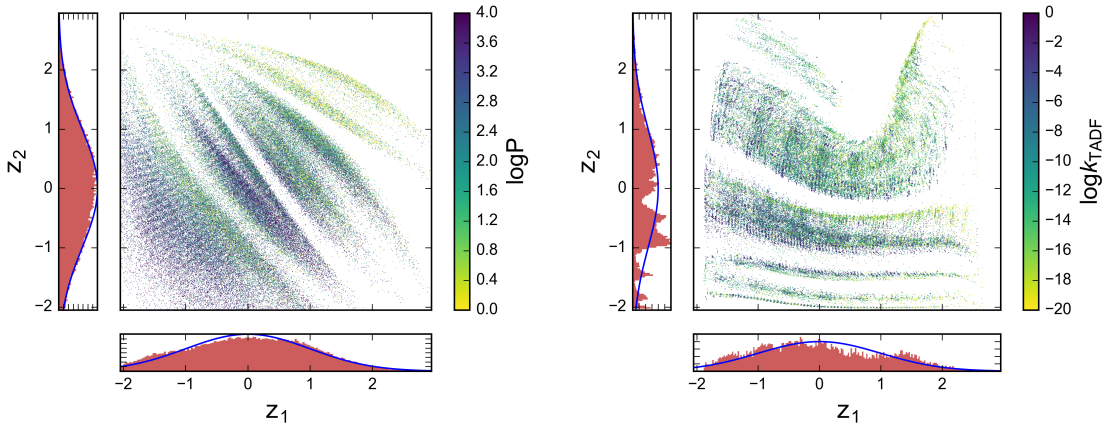

Figure 3: Projection of the molecular training sets onto learned two-dimensional latent spaces. The one-dimensional histograms show the distribution of the training data along each dimension, overlaid with the Gaussian prior imposed in the variational autoencoder. The points are colored along a chemical 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}).

Figure 3 shows the results of training an autoencoder with a latent space of dimension 2. In the case of the drug-like molecules, since they are a diverse, natural training set, we observed that the molecules are spread quite evenly across the space and they follow the Gaussian prior very closely along both dimensions. On the other hand, the OLED molecules, even in only two dimensions, are tightly clustered and leave larger gaps in the latent representation. The distribution of the encoded molecules could not be effectively regularized into a Gaussian shape and are highly structured along each dimension. This is a consequence of the combinatorial donor-bridge-acceptor way in which they were generated

and explains the difficulty of the autoencoder to efficiently learn a latent representation.

In both projections, even in the well-regularized drug-like set, we observe banding and structure in the 2D space at multiple scales. This showcases the ability of the deep autoencoder to address molecular similarity while mapping a discrete to a continuous representation. Interestingly, the plotted molecular properties show a marked dependency on the latent coordinates; molecules clustered closely tend to show similar properties, in the case of the emitter library, both at a local and at more general scale. That similar molecules have similar properties is the foundation of structure-property relationships in chemistry, and a desired feature for any molecular representation. This result is particularly encouraging since the autoencoder was trained in an unsupervised way, independently of target properties.

Table 1 compares the ability of our best drug and OLED autoencoder to reconstruct the train and test sets both for naive and varïational configurations, confirming the greater challenge of learning a latent representation with the machine-generated OLED library.

Molecular family	Autoencoder training loss	Latent dimension	Training set reconstruction %	Test set reconstruction %
drug-like	naïve	56	99.1	98.3
drug-like	variational	292	96.4	95.3
OLED	naïve	56	96.7	91.2
OLED	variational	292	91.4	79.4

Table 1: Reconstruction accuracy for the deep autoencoders used in this work. Accuracy is defined as the percentage of correct characters in decoded SMILES strings. An autoencoder with a large enough latent dimension could achieve perfect reconstruction, but exploration of the latent space tends to become more difficult as the latent dimension increases.

Perturbation and interpolation in molecular space We analyzed the performance of the VAE in a number of tasks involving encoding and decoding molecular structures. Figures 10 and 11 show multiple decodings of eight randomly-sampled points in the latent space. The random points decode into mid-sized realistic molecules, suggesting that the VAE is effective at ensuring that points in the latent space map to valid molecules. The slight variation within the molecules decoded for each point in latent space is due to the stochastic nature of the decoder.

Random drugs from a list of 1386 FDA approved drugs were selected, [42] encoded by sampling the VAE and their latent representations were decoded. Figures 4, 8 and 9 show the result of such process. Multiple chemical variations of the original compounds were obtained as a consequence of both the probabilistic sampling of the VAE and stochastic nature of the SMILES string decoding.

In addition to sampling points in latent space, we analyzed the meaning of random directions in the latent space. Starting from a random drug molecule, two random unitary vectors in latent space were followed and decoded into molecules. Figures 2-13 show the starting molecule in the center, and the most probable decodings of the extrapolated molecules on the horizontal and vertical direction. Most points in latent space correspond

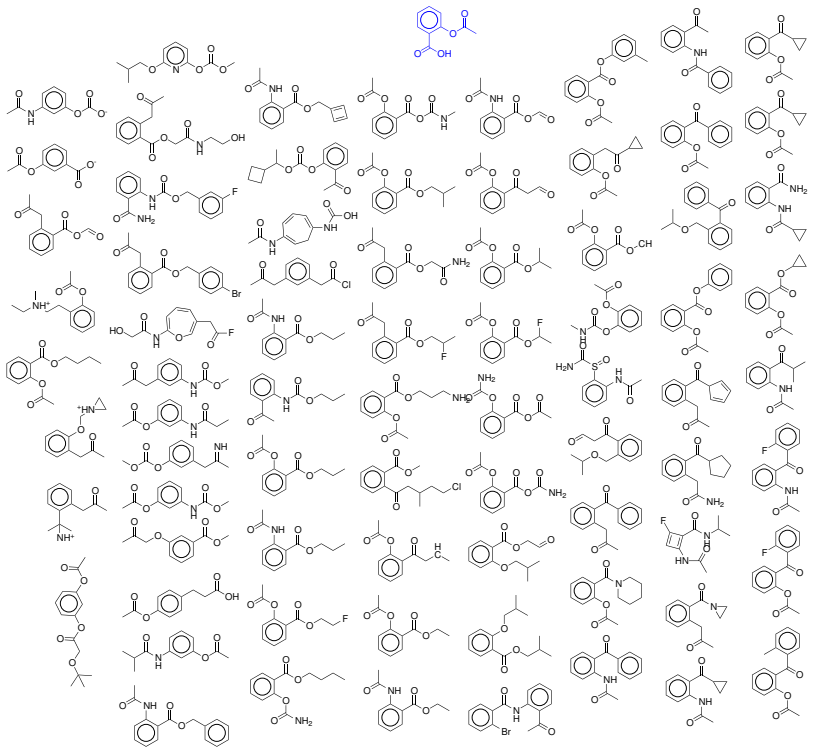

Figure 4: Molecules decoded from randomly-sampled points in the latent space of a variational autoencoder, near to a given molecule (aspirin [2-(acetyloxy)benzoic acid], highlighted in blue).

to realistic drug-like molecules. In a related experiment, and following the success of other generative models of images, we performed interpolations in chemical space. Random drugs from the list of FDA approved molecules were selected and encoded by sampling the mean of the VAE. We then performed a linear grid interpolation over two dimensions. We decoded each point in latent space multiple times and report the one whose latent representation, once re-encoded, is the closest to the sampled point (Figures 14-5)

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. [43] 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 [44] (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.

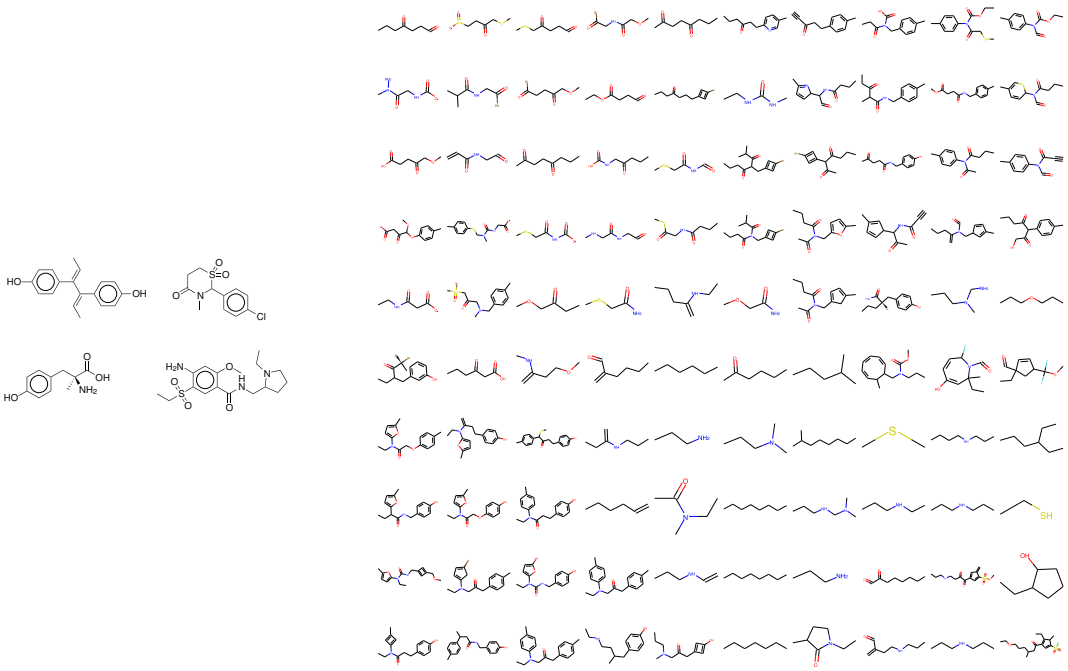

Figure 5: Interpolation. Two-dimensional interpolation between four random drugs. *Left* Starting molecules encoded, whose decodings correspond to the respective four corners of the figure to the right. *Right* Decodings of interpolating linearly between the latent representations of the four molecules to the right.

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

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

where the scores $\text{logP}(m)$, $\text{SA}(m)$, 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 7 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 7.

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 7 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.

Experiments with OLEDs Because of the large design space and promising properties of organic molecules as a replacement for solid-state inorganic materials, the field of organic electronics is very active in simulation-based molecular design, particularly in organic light

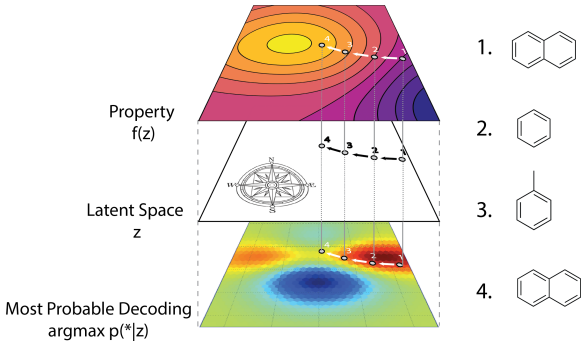

Figure 6: Gradient-based optimization in continuous latent space. After training a surrogate model $f(z)$ to predict the properties of molecules based on their latent representation z, we can optimize $f(z)$ with respect to z to find new latent representation expected to have high values of desired properties. These new latent representations can then be decoded into SMILES strings, at which point their properties can be tested emprically.

emitting diodes. OLEDs are organic molecules that can emit light in response to an electric current. They are currently applied in small displays, mostly smartphones, and have the potential to become the new paradigm of display, replacing liquid-crystal displays with LED backpanels in larger area displays. The newest OLED technology, thermally-assisted delayed fluorescence (TADF), [45] relies on combining fluorescent character of the emitter with a low energy difference between its lowest triplet and singlet excited state. Design proxies for these properties, together with with characteristics such as emission color, are amenable to simple quantum calculations and a variety of recent works have addressed the computer-driven design of new OLEDs both in TADF [9, 11, 46] and phosphorescence. [47] For this reason, we tested the performance of an autoencoder system on this class of molecules.

We used a training set with about 150,000 initial molecules generated by fragment combination. [9] Different properties of these molecules (emission color, delayed fluorescence decay rate (k_{TADF}), etc.) were estimated using time-dependent density functional theory. After that, we produced new molecules by optimizing these properties in the latent space.

Unfortunately, in this case, we found that the latent vectors selected by the Bayesian optimization procedure either did not produce valid SMILE strings, or the resulting SMILES were already found in the training data. We believe that the molecule autoencoder failed in this case to learn a generalizable latent representation of chemical space. This is probably due to the fact that the molecules used to train the autoencoder were very similar to each other. This resulted in the lowered accuracy that reported in Table 1. These problems could potentially be addressed by training the molecule autoencoder using even larger amounts of diverse, unlabeled molecules.

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.

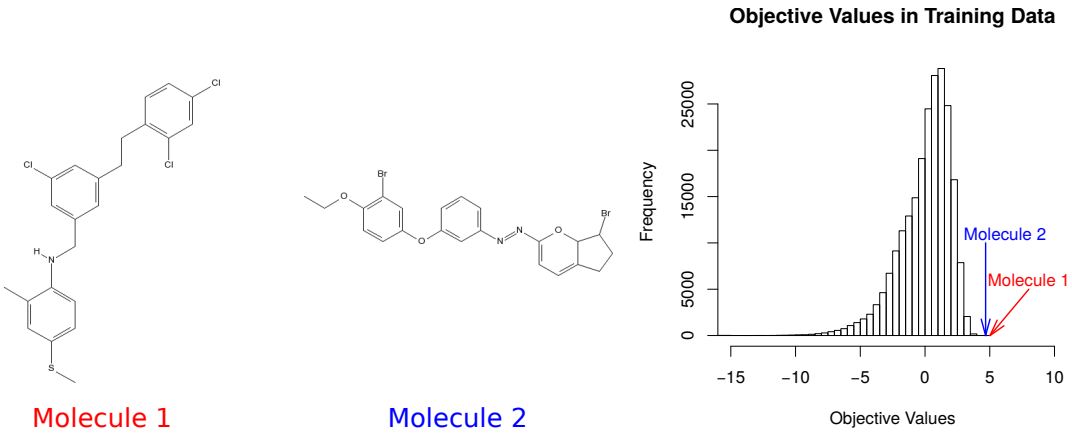

Figure 7: *Left*: Molecules generated by the optimization process with better score values than any other molecule in the training data. *Right*: Histogram of objective values in the training data.

Jointly training would encourage the model to find a latent representation which is both easily decoded, and easy to predict with. [48] 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. Forcing the decoder to produce valid SMILES strings makes the learning problem unnecessarily hard, since the decoder must also implicitly learn which strings are valid SMILES. An autoencoder that directly outputs molecular graphs is appealing, since it would directly address issues of graph isomorphism and the problem of strings that do not correspond to valid molecular graphs. An encoder which takes in molecular graphs is straightforward, with off-the-shelf molecular fingerprinting method, such as ECFP [17] or a continuously-parameterized variant of ECFP such as neural molecular fingerprints. [18] However, building a neural network which can output arbitrary graphs is an open problem.

Another issue with the string-encoding approach is that the decoder sometimes produces invalid SMILES strings. Since the autoencoder is only presented with valid molecules as inputs, the decoder lacks precise knowledge of chemical rules about which molecules are invalid. A similar issue in machine-learning the results of chemical experiments has been addressed by using both valid and invalid data (failed experiments) to train a discriminator [49].

5 Discussion

We have proposed a new family of methods for exploring chemical space based on continuous encodings of molecules. These methods can eliminate the need to hand-build libraries of compounds, and allow a new type of directed, gradient-based search through chemical

space. We have observed good fidelity and optimization ability when training with diverse representative data, but less so with machine-generated combinatorial libraries.

6 Acknowledgements

This work was supported financially by the Samsung Advanced Institute of Technology. The authors acknowledge the use of the Harvard FAS Odyssey Cluster and support from FAS Research Computing. JMHL acknowledges support from the Rafael del Pino Foundation.

References

[1] Kim, S. *et al.* Pubchem substance and compound databases. *Nucleic Acids Res.* **44**, 1202–1213 (2016).

[2] Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like chemical space based on gdb-17 data. *J. Comput.-Aided Mol. Des.* **27**, 675–679 (2013).

[3] Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862–5 (2004).

[4] Scior, T. *et al.* Recognizing pitfalls in virtual screening: A critical review. *J. Chem. Inf. Model.* **52**, 867–881 (2012).

[5] Cheng, T., Li, Q., Zhou, Z., Wang, Y. & Bryant, S. H. Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J.* **14**, 133–141 (2012).

[6] Pyzer-Knapp, E. O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J. & Aspuru-Guzik, A. What is high-throughput virtual screening? a perspective from organic materials discovery. *Annu. Rev. Mater. Res.* **45**, 195–216 (2015).

[7] Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* **9**, 273–276 (2010).

[8] Hachmann, J. *et al.* The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2**, 2241–2251 (2011).

[9] Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).

[10] Virshup, A. M., Contreras-García, J., Wipf, P., Yang, W. & Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **135**, 7296–7303 (2013).

[11] Rupakheti, C., Virshup, A., Yang, W. & Beratan, D. N. Strategy to discover diverse optimal molecules in the small molecule universe. *J. Chem. Inf. Model.* **55**, 529–537 (2015).

[12] Reymond, J.-L. The chemical space project. *Acc. Chem. Res.* **48**, 722–730 (2015).

[13] Reymond, J.-L., van Deursen, R., Blum, L. C. & Ruddigkeit, L. Chemical space as a source for new drugs. *Med. Chem. Commun.* **1**, 30 (2010).

[14] Kanal, I. Y., Owens, S. G., Bechtel, J. S. & Hutchison, G. R. Efficient computational screening of organic polymer photovoltaics. *J. Phys. Chem. Lett.* **4**, 1613–1623 (2013).

[15] O'Boyle, N. M., Campbell, C. M. & Hutchison, G. R. Computational design and selection of optimal organic photovoltaic materials. *J. Phys. Chem. C* **115**, 16200–16210 (2011).

[16] van Deursen, R. & Reymond, J.-L. Chemical space travel. *ChemMedChem* **2**, 636–640 (2007).

[17] Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

[18] Duvenaud, D. K. *et al.* Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, 2215–2223 (2015).

[19] Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: Moving beyond fingerprints. *arXiv preprint arXiv:1603.00856* (2016). 1603.00856.

[20] Hansen, K. *et al.* Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).

[21] Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108** (2012).

[22] Hirn, M., Poilvert, N. & Mallat, S. Quantum energy regression using scattering transforms. *arXiv preprint arXiv:1502.02077* (2015). 1502.02077.

[23] Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *arXiv preprint arXiv:1609.08259* (2016). 1609.08259.

[24] Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[25] Bowman, S. R. *et al.* Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349* (2015).

[26] van den Oord, A. *et al.* Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016). 1609.03499.

[27] Weininger, D. *J. Chem. Inf. Model.* **28**, 31–36 (1988).

[28] Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. Inchi - the worldwide chemical structure identifier standard. *J. Cheminf.* **5**, 7 (2013).

[29] Rdkit: Open-source cheminformatics. `http://www.rdkit.org`.

[30] Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112 (2014).

[31] Kalchbrenner, N., Grefenstette, E. & Blunsom, P. A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (2014).

[32] Chung, J., Gülçehre, Ç., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[33] Chollet, F. keras. `https://github.com/fchollet/keras` (2015).

[34] Bastien, F. *et al.* Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop (2012).

[35] Team, T. D. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688* (2016).

[36] Snelson, E. & Ghahramani, Z. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, 1257–1264 (2005).

[37] Jones, D. R., Schonlau, M. & Welch, W. J. Efficient global optimization of expensive black-box functions. *J. Global Optim.* **13**, 455–492 (1998).

[38] Cressie, N. The origins of kriging. *Math. Geol.* **22**, 239–252 (1990).

[39] Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[40] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. Zinc: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **52**, 1757–1768 (2012).

[41] Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. In *Neural Information Processing Systems 25* (2012).

[42] Law, V. *et al.* Drugbank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, 1091–1097 (2014).

[43] Wildman, S. A. & Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).

[44] Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **1**, 1–11 (2009).

[45] Uoyama, H., Goushi, K., Shizu, K., Nomura, H. & Adachi, C. Highly efficient organic light-emitting diodes from delayed fluorescence. *Nature* **492**, 234–238 (2012).

[46] Shu, Y. & Levine, B. G. Simulated evolution of fluorophores for light emitting diodes. *J. Chem. Phys.* **142** (2015).

[47] Kwak, H. S. *et al.* Virtual screening and evaluation of highly efficient organometallic light-emitting materials. vol. 9941, 994119 (2016).

[48] Snoek, J., Adams, R. P. & Larochelle, H. Nonparametric guidance of autoencoder representations using label information. *J. Mach. Learn. Res.* **13**, 2567–2588 (2012).

[49] Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).

7 Supplementary Material

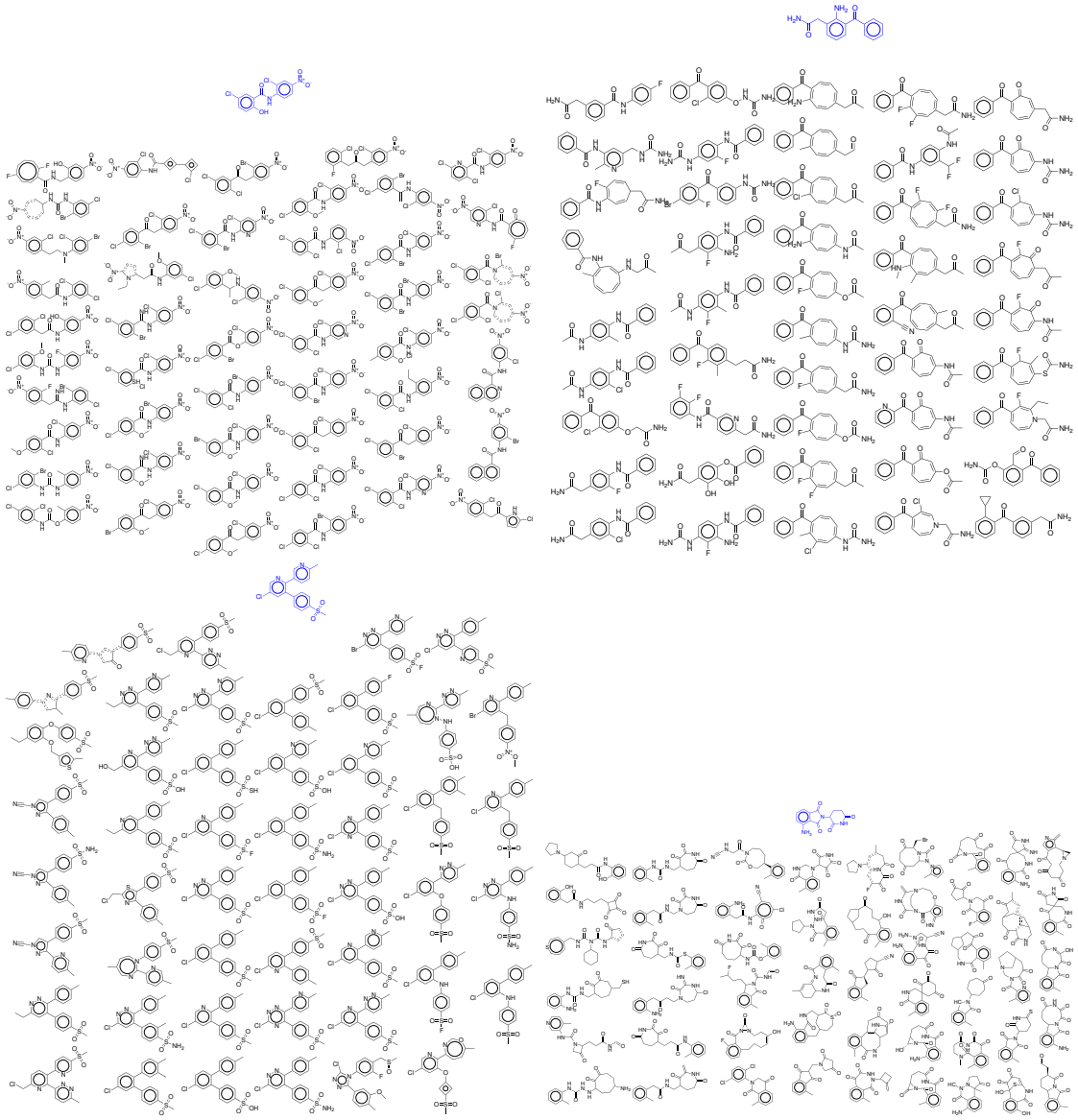

Figure 8: Molecules which were used as inputs to the variational autoencoder, presented with decodings of multiple samples from the encoding distribution

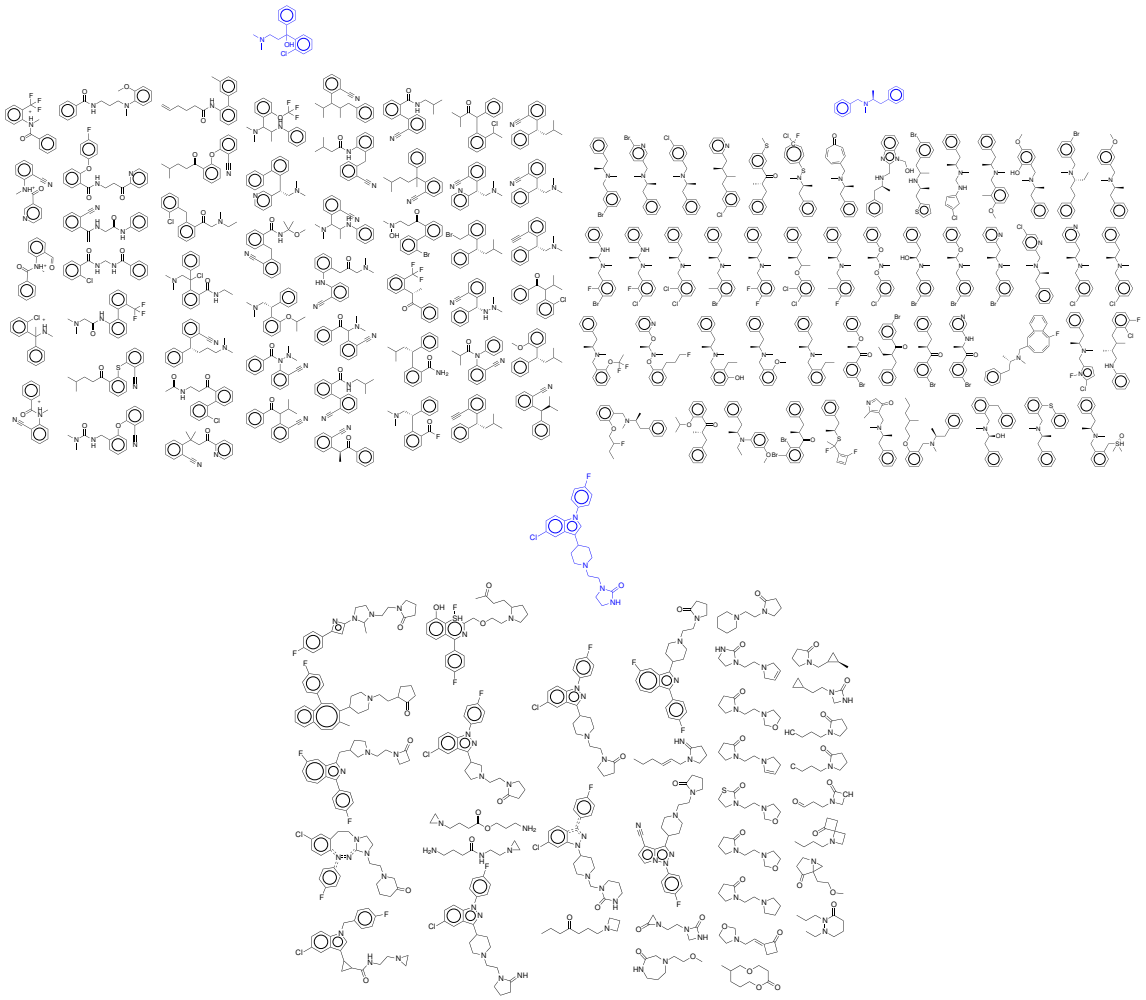

Figure 9: In blue, molecules which were used as inputs to the variational autoencoder, presented with decodings of multiple samples from the encoding distribution.

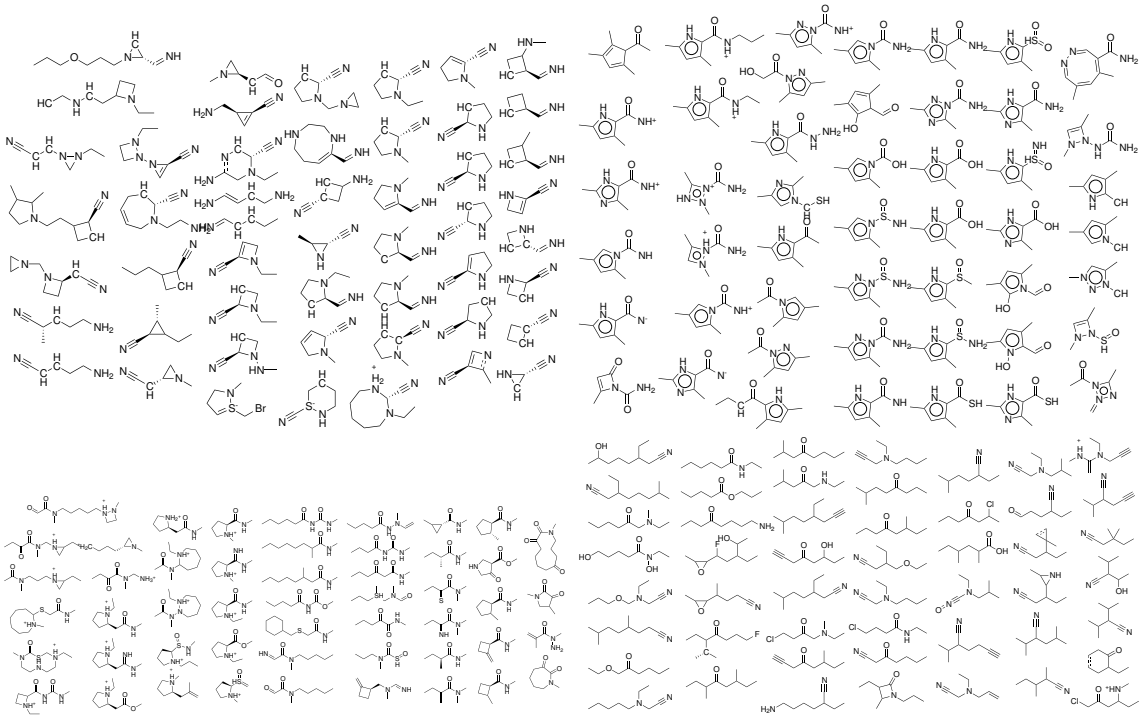

Figure 10: Molecules decoded from randomly-sampled points in the latent space of a variational autoencoder.

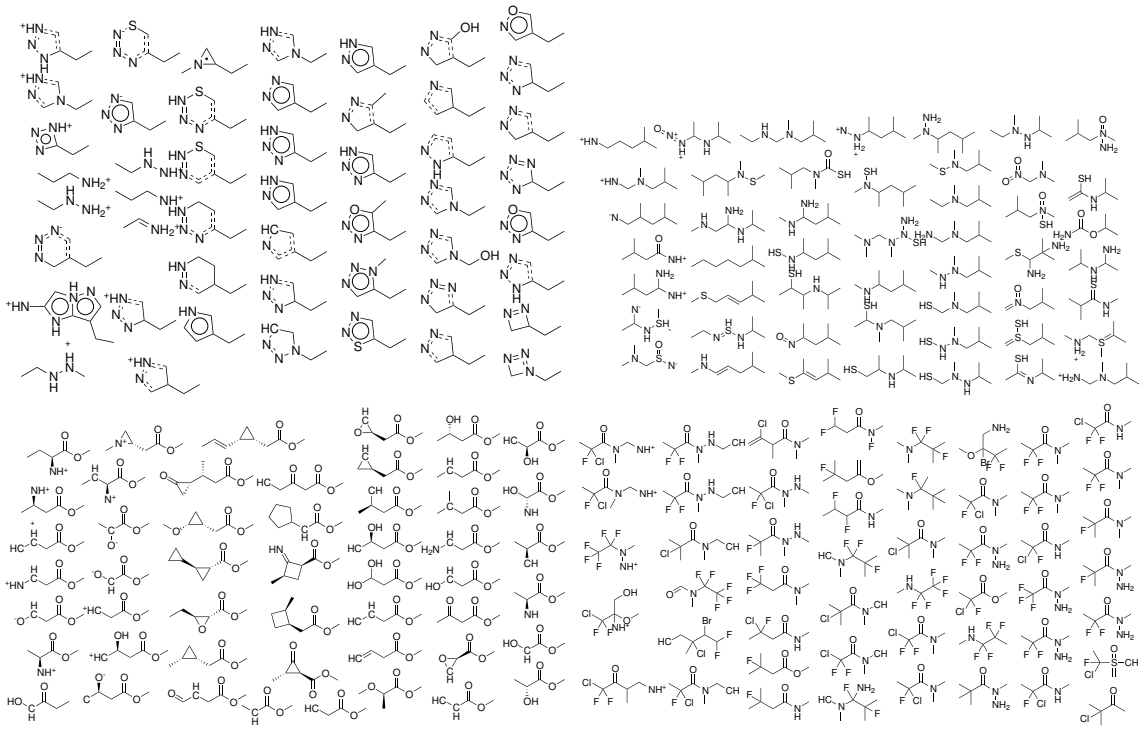

Figure 11: Molecules decoded from randomly-sampled points in the latent space of a variational autoencoder.

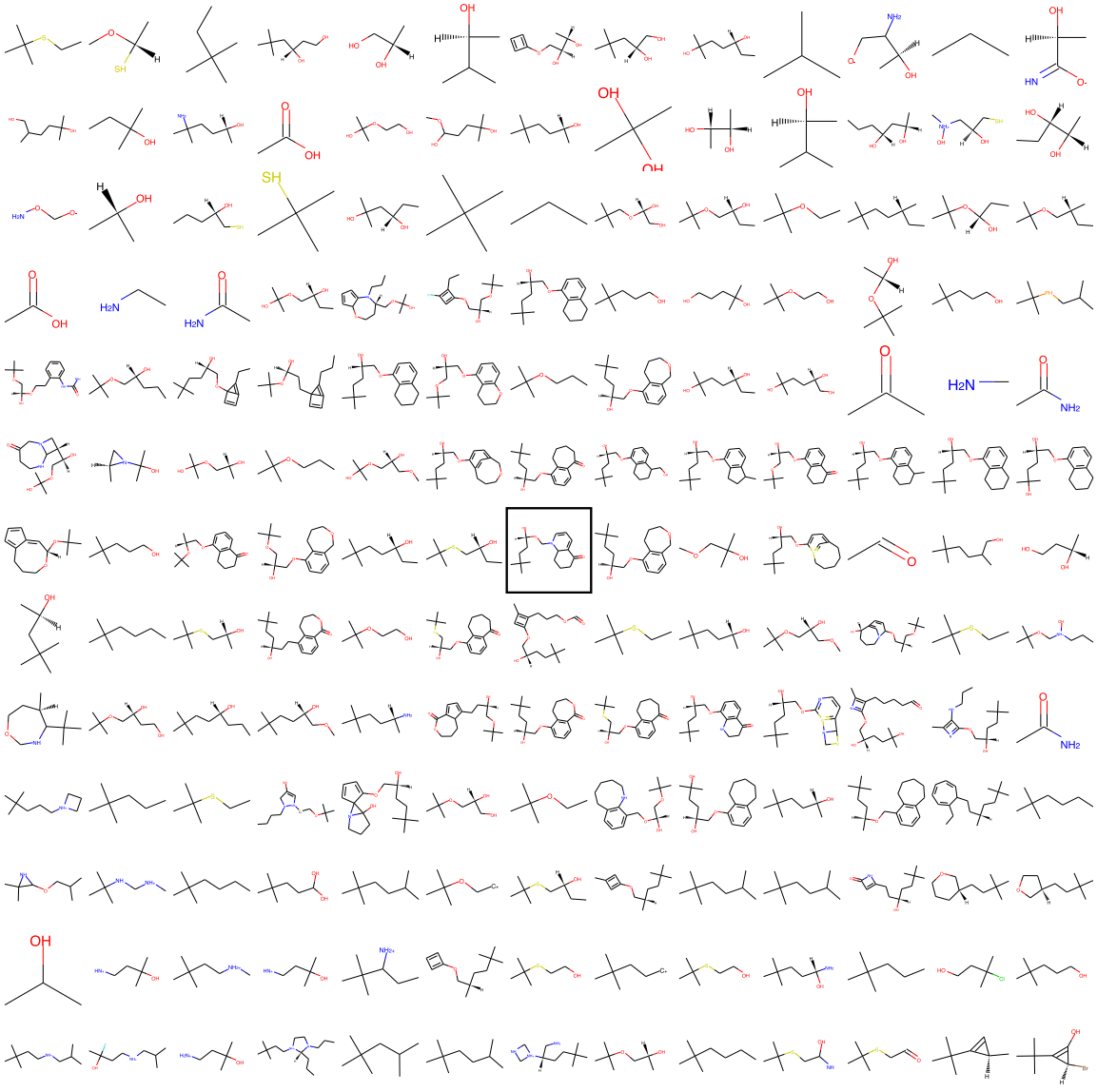

Figure 12: Starting from the molecule in the center, two random, unit-length vectors were followed in latent space for increasingly large displacements.

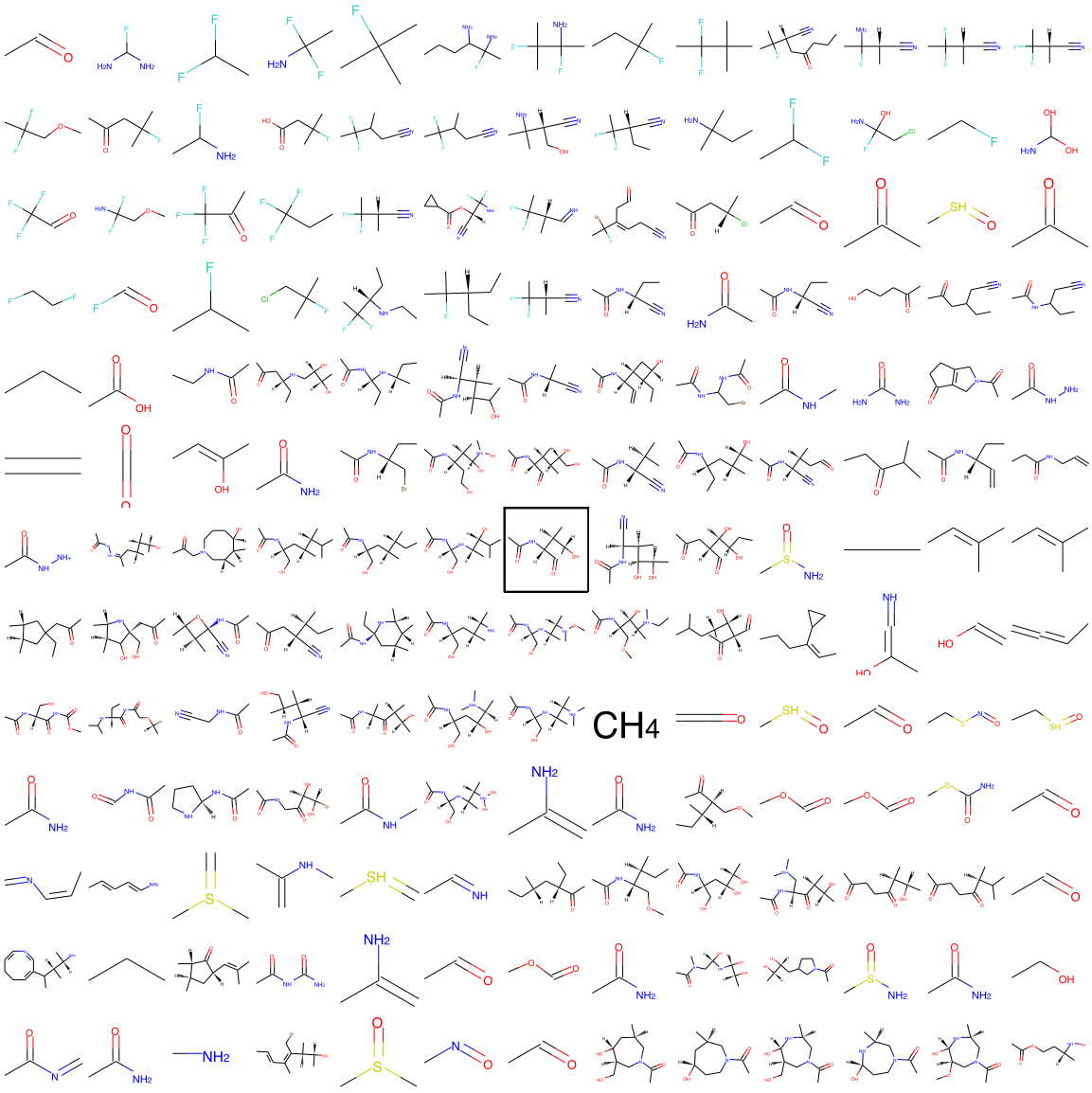

Figure 13: Starting from the molecule in the center, two random, unit-length vectors were followed in latent space for increasingly large displacements.

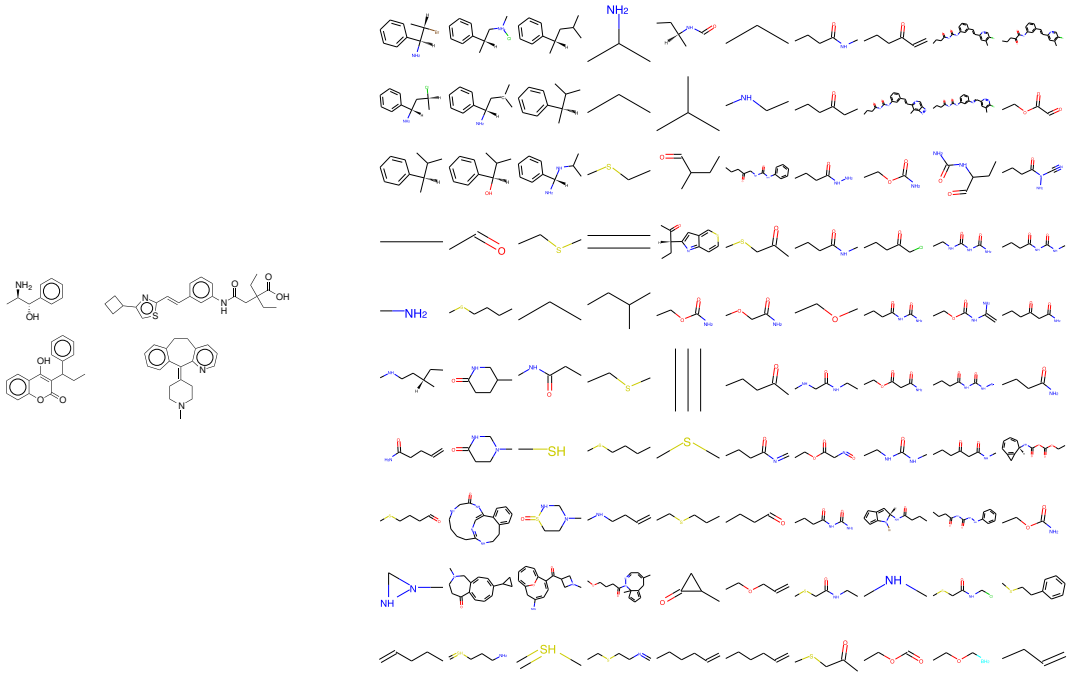

Figure 14: Interpolation. Two-dimensional interpolation between four random drugs. *Left:* Starting molecules, whose encodings defined the four corners of a place in the latent space. *Right:* Decodings of linearly-interpolated points between the latent representations of the four molecules to the right.

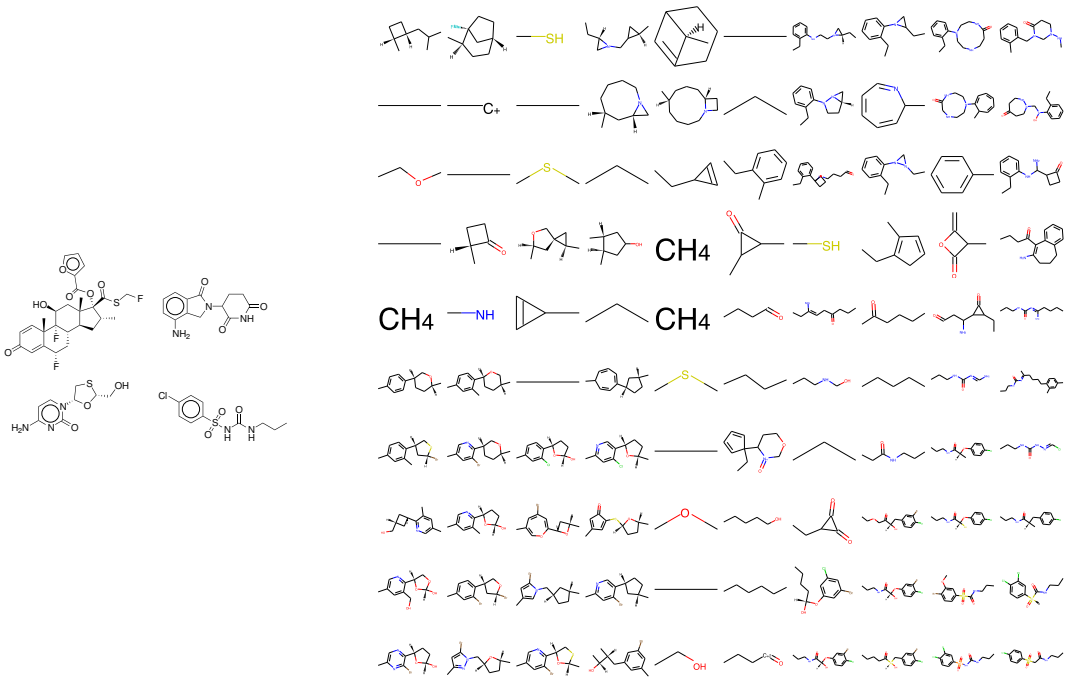

Figure 15: Interpolation. Two-dimensional interpolation between four random drugs. *Left:* Starting molecules, whose encodings defined the four corners of a place in the latent space. *Right:* Decodings of linearly-interpolated points between the latent representations of the four molecules to the right.