Constrained Bayesian Optimization for Automatic Chemical Design

Ryan-Rhys Griffiths  
Department of Engineering  
University of Cambridge  
rrg27@cam.ac.uk

José Miguel Hernández-Lobato  
Department of Engineering  
University of Cambridge  
jmh233@cam.ac.uk

Abstract

Automatic Chemical Design provides a framework for generating novel molecules with optimized molecular properties. The current model suffers from the pathology that it tends to produce invalid molecular structures. By reformulating the search procedure as a constrained Bayesian optimization problem, we showcase improvements in both the validity and quality of the generated molecules. We demonstrate that the model consistently produces novel molecules ranking above the 90th percentile of the distribution over training set scores across a range of objective functions. Importantly, our method suffers no degradation in the complexity or the diversity of the generated molecules.

1 Introduction

There are two fundamental ways in which Machine Learning can be leveraged in chemistry:

1. To evaluate a molecule for a given application.
2. To find a promising molecule for a given application.

There has been much progress in the first use-case through the development of quantitative structure activity relationship (QSAR) models using deep learning [Ma et al., 2015]. These models have achieved state-of-the-art results in predicting properties of known molecules.

The second use-case, finding new molecules that are useful for a given application, is arguably more important however. One existing approach for finding molecules that maximise an application-specific metric involves searching a large library of compounds, either physically or virtually Pyzer-Knapp et al. [2015]. This has the disadvantage that the search is not open-ended; if the molecule is not in the library you specify, the search won’t find it.

A second method involves the use of genetic algorithms. In this approach, a known molecule acts as a seed and a local search is performed over a discrete space of molecules. Although these methods have enjoyed success in producing biologically active compounds, an approach featuring a search over an open-ended, continuous space would be beneficial. The use of geometrical cues such as gradients to guide the search in continuous space could accelerate both drug [Pyzer-Knapp et al., 2015, Gómez-Bombarelli et al., 2016a] and material [Hachmann et al., 2011, 2014] discovery by functioning as a high-throughput virtual screen of unpromising candidates.

Recently, Gómez-Bombarelli et al. [Gómez-Bombarelli et al., 2016b] presented Automatic Chemical Design, a VAE architecture capable of encoding continuous representations of molecules. In continuous latent space, gradient-based optimization is leveraged to find molecules that maximize a design metric.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.
Although a strong proof of concept, Automatic Chemical Design possesses a deficiency in so far as it fails to generate a high proportion of valid molecular structures. Gómez-Bombarelli et al. [2016b] hypothesise that molecules selected by Bayesian Optimization lie in “dead regions” of the latent space far away from any data that the VAE has seen in training, yielding invalid structures when decoded.

Although there have been many attempts to address the issue of generating valid molecules, they all rely on model assumptions that can negatively impact either the complexity [Jaques et al., 2017, Janz et al., 2017, Kusner et al., 2017] or the diversity [Jin et al., 2018] of the generated molecules.

The principle contribution of this paper will be to present an approach based on constrained Bayesian optimization that generates a high proportion of valid sequences without relying on model assumptions that affect the complexity or the diversity of the generated molecules.

2 Methods

2.1 SMILES Representation

SMILES strings Weininger [1988] are a means of representing molecules as a character sequence. This text-based format facilitates the use of tools from natural language processing for applications such as organic chemistry reaction prediction Schwaller et al. [2017], Jin et al. [2017]. For use with the variational autoencoder, the SMILES strings in turn are converted into one-hot vectors indicating the presence or absence of a particular character within a sequence as illustrated in Figure 1.

2.2 Variational Autoencoder

Variational autoencoders Kingma and Welling [2013], Kingma et al. [2014] allow us to map molecules x to and from continuous values z in a latent space. The encoding z is interpreted as a latent variable in a probabilistic generative model over which there is a prior distribution p(z). The probabilistic decoder is defined by the likelihood function pθ(x|z). The posterior distribution pθ(x|z) is interpreted as the probabilistic encoder. The parameters of the likelihood pθ(x|z) as well as the parameters of the approximate posterior distribution qφ(z|x) are learned by maximizing the evidence lower bound (ELBO)

\[
\mathcal{L}(\phi, \theta; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x, z) - \log q_\phi(z|x)].
\]

Variational autoencoders have been coupled with recurrent neural networks by Bowman et al. [2015] to encode sentences into a continuous latent space. This approach is followed for the SMILES format both by Gómez-Bombarelli et al. [2016b] and here. The SMILES variational autoencoder is shown in Figure 2.
2.3 Bayesian Optimization of Molecules

Bayesian Optimization is performed in the latent space of the variational autoencoder in order to find molecules that score highly under a specified objective function. We assess molecular quality on the following objectives:

\[
J_{\text{comp}}^{\log P}(m) = \log P(m) - \text{SA}(m) - \text{ring-penalty}(m),
\]

\[
J_{\text{comp}}^{\text{QED}}(m) = \text{QED}(m) - \text{SA}(m) - \text{ring-penalty}(m),
\]

\[
J_{\text{QED}}^{\text{comp}}(m) = \text{QED}(m).
\]

\(m\) denotes a molecule, \(\log P(m)\) is the water-octanol partition coefficient, \(\text{QED}(m)\) is the quantitative estimate of drug-likeness [Bickerton et al., 2012] and \(\text{SA}(m)\) is the synthetic accessibility [Ertl and Schuffenhauer, 2009]. The ring penalty term is as featured in [Gómez-Bombarelli et al., 2016b]. The “comp” subscript is designed to indicate that the objective function is a composite of standalone metrics.

It is important to note, that the first objective, a common metric of comparison in this area Gómez-Bombarelli et al. [2016b], Kusner et al. [2017], Jin et al. [2018], is mis-specified. From a chemical standpoint it is undesirable to maximize the \(\log P\) score as is being done here. Rather it is preferable to optimize \(\log P\) to be in a range that is in accordance with the Lipinski Rule of Five Lipinski et al. [1997]. We use the penalized \(\log P\) objective here because regardless of its relevance for chemistry, it serves as a point of comparison against other methods.

2.4 Constrained Bayesian Optimization of Molecules

We now describe our extension to the Bayesian Optimization procedure followed by Gómez-Bombarelli et al. [2016b]. Expressed formally, the constrained Bayesian optimization problem is

\[
\max_x f(x) \text{ s.t. } \Pr(C(x)) \geq 1 - \delta
\]

where \(f(x)\) is a black-box objective function, \(\Pr(C(x))\) denotes the probability that a boolean constraint \(C(x)\) is satisfied and \(1 - \delta\) is some user-specified minimum confidence that the constraint is satisfied [Gelbart et al., 2014]. The constraint is that a latent point must decode successfully a large fraction of the times decoding is attempted. The black-box objective function is noisy because a single latent point may decode to multiple molecules when the model makes a mistake, obtaining different values under the objective. In practice, \(f(x)\) is one of the objectives described in section 2.3

2.5 Expected Improvement with Constraints (EIC)

EIC may be thought of as expected improvement (EI) that offers improvement only when a set of constraints are satisfied [Schonlau et al., 1998]:

\[
\text{EIC}(x) = \text{EI}(x) \Pr(C(x)).
\]
The implicit incumbent solution $\eta$ suppressed in EI($x$), may be set in an analogous way to vanilla expected improvement [Gelbart, 2015] as either:

1. The best observation in which all constraints are observed to be satisfied.
2. The minimum of the posterior mean such that all constraints are satisfied.

The latter approach is adopted for the experiments performed in this paper. If at the stage in the Bayesian optimization procedure where a feasible point has yet to be located, the form of acquisition function used is that defined by [Gelbart, 2015]

$$EIC(x) = \begin{cases} Pr(C(x))EI(x), & \text{if } \exists x, Pr(C(x)) \geq 1 - \delta \\ Pr(C(x)), & \text{otherwise} \end{cases}$$

with the intuition being that if the probabilistic constraint is violated everywhere, the acquisition function selects the point having the highest probability of lying within the feasible region. The algorithm ignores the objective until it has located the feasible region. It would also have been possible to adopt the methodology introduced by Rainforth et al. [2016] under the assumption that the constraint is cheap to evaluate.

3 Related Work

Gómez-Bombarelli et al. [2016b], Segler et al. [2017] constructed character-level generative models of SMILES strings using recurrent decoders. These models both suffered from the problem that they produced a high proportion of invalid SMILES strings. Blaschke et al. [2017] compared a range of autoencoder architectures in terms of their ability to produce valid molecules. Kusner et al. [2017], Dai et al. [2018] addressed the issue of generating valid molecules explicitly by imposing syntactic and semantic constraints using context free and attribute grammars while Janz et al. [2017], Guimaraes et al. [2017] leveraged additional training signal to enforce molecular validity.

A potential drawback of these methods is that the constraints may favour the generation of simple molecules. Simonovsky and Komodakis [2018], Li et al. [2018] build generative models of molecular graphs but do not impose any constraints for molecular validity. Jin et al. [2018] build a generative model of molecular graphs with constraints in order to achieve 100% validity. One possible disadvantage of this approach is that each molecule is assumed to be constructed from a fixed vocabulary of subgraphs and so it will be impossible to generate a molecule comprised of subgraphs outside this vocabulary. Direct comparison with the aforementioned methods is complicated due to the orthogonality of the assumptions made for each model. As such, we compare out method against Gómez-Bombarelli et al. [2016b] only.

Our principle contribution is in showing that valid molecules can be generated without encoding assumptions into the design model that limit the complexity and diversity of the generated molecules.

4 Experiment II: Drug Design

In this section we conduct an empirical test of the hypothesis from [Gómez-Bombarelli et al., 2016b] that the decoder’s lack of efficiency is due to data point collection in “dead regions” of the latent space far from the data on which the VAE was trained. We use this information to construct a binary classification BNN to serve as a constraint function that outputs the probability of a latent point being valid. Secondly, we compare the performance of our constrained Bayesian optimization implementation against the original model (baseline) in terms of the numbers of valid and drug-like molecules generated. Thirdly, we compare the quality of the molecules produced by constrained Bayesian optimization with those of the baseline model.

4.1 Implementation

The implementation details of the encoder-decoder network as well as the sparse GP for modelling the objective remain unchanged from [Gómez-Bombarelli et al., 2016b]. For the constrained Bayesian optimization algorithm, the BNN is constructed with 2 hidden layers each 100 units wide with ReLU
activation functions and a logistic output. Minibatch size is set to 1000 and the network is trained for 5 epochs with a learning rate of 0.0005. 20 iterations of parallel Bayesian optimization are performed using the Kriging-Believer algorithm in all cases. Data is collected in batch sizes of 50. The same training set as [Gómez-Bombarelli et al., 2016b] is used, namely 249,456 drug-like molecules drawn at random from the ZINC database [Irwin et al., 2012].

4.2 Diagnostic Experiments and Labelling Criteria

These experiments were designed to test the hypothesis that points collected by Bayesian optimization lie far away from the training data in latent space. In doing so, they also serve as labelling criteria for the data collected to train the BNN acting as the constraint function. The resulting observations are summarized in Figure 3.

There is a noticeable decrease in the percentage of valid molecules decoded as one moves further away from the training data in latent space. Points collected by Bayesian optimization do the worst in terms of the percentage of valid decodings. This would suggest that these points lie farthest from the training data. The decoder over-generates methane molecules when far away from the data. One hypothesis for why this is the case is that methane is represented as ‘C’ in the SMILES syntax and is by far the most common character. Hence far away from the training data, combinations such as ‘C’ followed by a stop character may have high probability under the distribution over sequences learned by the decoder.

Given that methane has far too low a molecular weight to be a suitable drug candidate, a third plot in Figure 3(c), shows the percentage of decoded molecules such that the molecules are both valid and have a tangible molecular weight. The definition of a tangible molecular weight was interpreted somewhat arbitrarily as a SMILES length of 5 or greater. Henceforth, molecules that are both valid and have a SMILES length greater than 5 will be loosely referred to as drug-like. This is not to imply that a molecule comprising five SMILES characters is likely to be drug-like, but rather this SMILES length serves the purpose of determining whether the decoder has been successful or not.

As a result of these diagnostic experiments, it was decided that the criteria for labelling latent points to initialize the binary classification neural network for the constraint would be the following: if the
latent point decodes into drug-like molecules in more than 20% of decode attempts, it should be classified as drug-like and non drug-like otherwise.

4.3 Molecular Validity

The BNN for the constraint was initialized with 117,440 positive class points and 117,440 negative class points. The positive points were obtained by running the training data through the decoder assigning them positive labels if they satisfied the criteria outlined in the previous section. The negative class points were collected by decoding points sampled uniformly at random across the 56 latent dimensions of the design space. Each latent point undergoes 100 decode attempts and the most probable SMILES string is retained. \( J_{\log P} (m) \) is the choice of objective function. The relative performance of constrained Bayesian optimization and unconstrained Bayesian optimization (baseline) [Gómez-Bombarelli et al., 2016b] is compared in 4(a).

The results show that greater than 80% of the latent points decoded by constrained Bayesian optimization produce drug-like molecules compared to less than 5% for unconstrained Bayesian optimization. One must account however, for the fact that the constrained approach may be decoding multiple instances of the same novel molecules. Constrained and unconstrained Bayesian optimization are compared on the metric of the percentage of unique novel molecules produced in 4(b).

One may observe that constrained Bayesian optimization outperforms unconstrained Bayesian optimization in terms of the generation of unique molecules, but not by a large margin. A manual inspection of the SMILES strings collected by the unconstrained optimization approach showed that there were many strings with lengths marginally larger than the cutoff point, which is suggestive of partially decoded molecules. As such, a fairer metric for comparison should be the quality of the new molecules produced as judged by the scores from the black-box objective function. This is examined next.

4.4 Molecular Quality

The results of Figure 5 indicate that constrained Bayesian optimization is able to generate higher quality molecules relative to unconstrained Bayesian optimization across the three drug-likeness metrics introduced in section 2.3. Over the 5 independent runs, the constrained optimization procedure in every run produced new drug-like molecules ranked in the 100th percentile of the distribution over training set scores for the \( J_{\log P} (m) \) objective and over the 90th percentile for the remaining objectives. Table 1 gives the percentile that the averaged score of the new molecules found by each process occupies in the distribution over training set scores.
Figure 5: The best scores for new molecules generated from the baseline model (blue) and the model with constrained Bayesian optimization (red). The vertical lines show the best scores averaged over 5 separate train/test splits of 90/10. For reference, the histograms are presented against the backdrop of the top 10% of the training data in the case of Composite LogP and QED, and the top 20% of the training data in the case of Composite QED.

Table 1: Percentile of the averaged new molecule score relative to the training data. The results of 5 separate train/test set splits of 90/10 are provided.

| Objective     | Baseline        | Constrained     |
|---------------|-----------------|-----------------|
| LOGP Composite| 36±14           | 92±4            |
| QED Composite | 14±3            | 72±10           |
| QED           | 11±2            | 79±4            |

5 Experiment 3: Material Design

In order to show that the constrained Bayesian optimization approach is extensible beyond the realm of drug design, we trained the model on data from the Harvard Clean Energy Project [Hachmann et al., 2011, 2014] to generate molecules optimized for power conversion efficiency (PCE).

5.1 Implementation

A neural network was trained on 200,000 molecules drawn at random from the Harvard Clean Energy Project dataset using 512-bit Morgan circular fingerprints Rogers and Hahn [2010] as input features with bond radius of 2 using the RDKit package RDKit.

5.2 Results

The results are given in Figure 6. The averaged score of the new molecules generated lies above the 90th percentile in the distribution over training set scores. Given that the objective function in this instance was learned using a neural network, advances in predicting chemical properties from data Duvenaud et al. [2015], Ramsundar et al. [2015] are liable to yield concomitant improvements in the optimized molecules generated through this approach.
Figure 6: The best scores for novel molecules generated by the constrained Bayesian optimization model optimizing for PCE. The results are averaged over 3 separate runs with train/test splits of 90/10.

6 Conclusion and Future Work

6.1 Contributions

The reformulation of the search procedure in the Automatic Chemical Design model as a constrained Bayesian optimization problem has led to concrete improvements on two fronts:

1. Validity - The number of valid molecules produced by the constrained optimization procedure offers a marked improvement over the original model. Notably, by applying constraints solely to the latent space of the variational autoencoder, the method does not require model assumptions that compromise either the complexity or the diversity of the generated molecules.

2. Quality - For five independent train/test splits, the scores of the best molecules generated by the constrained optimization procedure consistently ranked above the 90th percentile of the distribution over training set scores for all objectives considered. The ability to find new molecules that are competitive with the very best of a training set of already drug-like molecules is a powerful demonstration of the model’s capabilities. As a further point, the generality of the approach should be emphasised. This approach is liable to work for a large range of objectives encoding countless desirable molecular properties.

6.2 Future Work

Future work could investigate whether performance gains can be achieved through the implementation of a more accurate constraint model. Recent work by Blaschke et al. [2017] has featured a more targeted search for novel compounds with predicted activity against dopamine receptor type 2. This represents a move towards more industrially-relevant objective functions for Bayesian Optimization which should ultimately replace the chemically mis-specified objectives, such as the penalized logP score, identified here.
References

Junshui Ma, Robert P Sheridan, Andy Liaw, George E Dahl, and Vladimir Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 55(2): 263–274, 2015.

E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, and A. Aspuru-Guzik. What is high-throughput virtual screening? a perspective from organic materials discovery. *Annual Review of Materials Research*, 45:195–216, 2015.

R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D-G. Ha, T. Wu, and G. Markopoulos. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials*, 15(10):1120–1127, 2016a.

J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera., A. Gold-Parker, L. Vogt, A. M. Brockway, and A. Aspuru-Guzik. The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *The Journal of Physical Chemistry Letters*, 2(17):2241–2251, 2011. doi: 10.1021/jz200866s.

J. Hachmann, R. Olivares-Amaya, A. Jinich, A. L. Appleton, M. A. Blood-Forsythe, L. R. Seress, C. Roman-Salgado, K. Treppe, S. Atahan-Evrenk, S. Er, S. Shrestha, R. Mondal, A. Sokolov, Z. Bao, and A. Aspuru-Guzik. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry - the harvard clean energy project. *Energy Environ. Sci.*, 7:698–704, 2014.

R. Gómez-Bombarelli, D. Duvenaud, J. M. Hernández-Lobato, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *arXiv preprint arXiv:1610.02415*, 2016b.

N. Jaques, S. Gu, D. Bahdanau, J. M. Hernández-Lobato, R. E. Turner, and D. Eck. Sequence tutor: Conservative fine-tuning of sequence generation models with KL-control. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1645–1654, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

David Janz, Jos van der Westhuizen, Brooks Paige, Matt J Kusner, and Jose Miguel Hernandez-Labato. Learning a generative model for validity in complex discrete structures. *arXiv preprint arXiv:1712.01664*, 2017.

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International Conference on Machine Learning*, pages 1945–1954, 2017.

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

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. “found in translation”: Predicting outcome of complex organic chemistry reactions using neural sequence-to-sequence models. *arXiv preprint arXiv:1711.04810*, 2017.

Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. In *Advances in Neural Information Processing Systems*, pages 2604–2613, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

R. G. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

P. Ertl and A. Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1(1):8, 2009.
C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

M. A. Gelbart, J. Snoek, and R. P. Adams. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*, 2014.

M. Schonlau, W. J. Welch, and D. R. Jones. Global versus local search in constrained optimization of computer models. *Lecture Notes-Monograph Series*, pages 11–25, 1998.

M. A. Gelbart. *Constrained Bayesian Optimization and Applications*. PhD thesis, Harvard University, 2015.

Tom Rainforth, Tuan Anh Le, Jan-Willem van de Meent, Michael A Osborne, and Frank Wood. Bayesian optimization for probabilistic programs. In *Advances in Neural Information Processing Systems*, pages 280–288, 2016.

Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 2017.

Thomas Blaschke, Marcus Olivecrona, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Application of generative autoencoder in de novo molecular design. *Molecular informatics*, 2017.

Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.

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

Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. *arXiv preprint arXiv:1802.03480*, 2018.

Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03524*, 2018.

J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–1768, 2012.

D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

RDKit. Rdkit: Open-source cheminformatics. URL http://www.rdkit.org.

D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS’15*, pages 2224–2232, Cambridge, MA, USA, 2015. MIT Press.

B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande. Massively Multitask Networks for Drug Discovery. *ArXiv e-prints*, February 2015.