Uncertainty-aware Labelled Augmentations for High Dimensional Latent Space Bayesian Optimization

Ekansh Verma *
Indian Institute of Technology Madras, India
vermaekansh55@gmail.com

Souradip Chakraborty*
University of Maryland, College Park
schakra3@umd.edu

Abstract

Black-box optimization problems are ubiquitous and of importance in many critical areas of science and engineering. Bayesian optimisation (BO) over the past years has emerged as one of the most successful techniques for optimising expensive black-box objectives. However, efficient scaling of BO to high-dimensional settings has proven to be extremely challenging. Traditional strategies based on projecting high-dimensional input data to a lower-dimensional manifold, such as Variational autoencoders (VAE) and Generative adversarial networks (GAN) have improved BO performance in high-dimensional regimes, but their dependence on excessive labeled input data has been widely reported. In this work, we target the data-greedy nature of deep generative models by constructing uncertainty-aware task-specific labeled data augmentations using Gaussian processes (GPs). Our approach outperforms existing state-of-the-art methods on machine learning tasks and demonstrates more informative data representation with limited supervision.

1 Introduction

Black-box optimization problems are extremely important in many critical areas of science and engineering including the bio-medical domain, healthcare, molecule generation etc. Black-box problems imply the absence of the explicit problem formulation, without any access to the model parameters or gradients and can only be accessed through the black-box evaluations. Bayesian optimisation (BO) over the past years has emerged as one of the most successful algorithms in this field of expensive black-box optimization. However, efficient scaling of the Bayesian optimization approach to high-dimensional settings has proven to be extremely challenging especially when the black-box function is expensive to evaluate.

The latest work in this field uses the latent space optimization framework in a model-based setting to address the high dimensionality issue in Bayesian optimization leveraging the power of deep generative models [1, 2, 3, 4, 5, 6, 7]. Deep Generative models like Variational autoencoders (VAE) [8] and Generative adversarial networks (GANs) [9] provide an immensely powerful approach to project the high dimensional input data to a lower dimensional manifold. Hence, in the first stage of the latent space optimization method, the high dimensional input is mapped to a low dimensional latent space using Variational autoencoders and subsequently BO is performed in this lower dimensional continuous latent space of the generative models. However, learning the unsupervised generative model without the possibility to update the learnt feature space during optimization results in non-discriminative features and subsequently slower convergence. Recent approach [10] tackles the above problem by implementing a weighted retraining mechanism, assigning more influence to regions of the latent space with favourable black-box function values in subsequent retraining of the deep generative models. However, the success of this approach is dependent on the availability of large amount of initial labelled data and also does not guarantee to produce an optimally discriminating

*equal contribution

NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications

latent space because latent points are not grouped according to their function value. [11] proposed a new method aimed at enhancing the discriminating power of the latent space leveraging the concepts of deep metric learning [12]. Deep metric learning based VAE BO [11] method outperformed the previous state of the art, emphasizing that metric learning based representations can enhance the surrogate model's ability to generalise to unseen regions of the latent space. However, the performance of metric learning based approaches are heavily dependent on efficient incorporation of invariances through prior knowledge and high quality data augmentations without which, their performances might lack the desired robustness as shown in [13, 14]. The hypothesis highlights the necessity and importance to learn meaningful data augmentations in order to enhance the sample efficiency of the VAE BO methods, as also pointed out in [11].

Hence, we try to solve the above mentioned problem by learning black-box task specific data augmentations with minimal supervision and enhancing the sample efficiency of VAE BO method. Our problem is especially hard since the explicit formulation for the black-box function is not present. Also, the task invariant augmentations are unknown which is in contrast with the popular self-supervised learning methods where the task invariant augmentations are easily identifiable. Our research presents a surprisingly simple yet effective methodology to extract uncertainty aware task-specific labelled augmentations leveraging the Gaussian process [15] in an iterative fashion.

We design the distribution of structured augmentations and systematically study the sensitivity of those augmentations to the task using the response predicted by our surrogate GP model which also predicts the uncertainty of the response/labels. We iteratively update the augmentation labels and enhance our dataset for training the generative model with certain and meaningful augmentations. Our uncertainty aware augmentation approach achieves State of the art results on the MNIST-Norm and Classification task with significantly low supervision compared to several other state of the art baselines. The proposed approach highlights the importance of learning task specific data augmentations adding meaningful data augmentations thereby learning much more informative representations with limited supervision.

2 Background

We consider the problem of global maximization defined as:

$$\mathbf{x}^\star = \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \tag{1}$$

with x as the global optimizer of the black box function $f(\cdot) : \mathcal{X} \to \mathbb{R}$ over a high-dimensional and structured input domain $\mathcal{X}$. Notably, we do not have the analytic form or derivative information available for $f(\cdot)$ and further it is expensive to evaluate $f(\cdot)$ at chosen input $\mathbf{x}$.

Bayesian Optimization : Bayesian optimization is a powerful tool for globally optimizing black-box functions [16, 17, 18] that are expensive to evaluate. There are two standard steps for BO : (1) Training a probabilistic surrogate model to estimate the black-box function and get the corresponding calibrated uncertainties, usually a Gaussian process (GP); (2) maximising the acquisition function that trades off exploration and exploitation according to uncertainty estimates derived from the surrogate model. Once surrogate model is trained, acquisition function recommends new input points $\mathbf{x}$ to evaluate black box function $f(\mathbf{x})$, the existing dataset is augmented with the new input points, function evaluations and surrogate model is retrained.

Latent Space Bayesian Optimization: In LSO, a low-dimensional data representation is learned with deep generative model (DGM), such as a variational autoencoder (VAE) or a generative adversarial network (GAN). VAE model is trained to map low dimensional high dimensional input $\mathcal{X}$ to latent space $\mathcal{Z} \subseteq \mathbb{R}^d$. Sample efficient BO is performed in the latent space of VAE which outputs,$\mathbf{z}^*$ in$\mathcal{Z}$. To evaluate the black box function $f(\cdot)$, $\mathbf{z}^*$ is projected back to original input space $\mathbf{x}$ using VAE decoder.

Latent Space Optimization with Weighted Retraining: LSO involves learning the VAE model offline without the possibility to update the latent space during the optimization process. This implies that the new points acquired during LSO can not be used to rectify the learnt latent space and expand the coverage for feasible regions generated using the VAE decoder. To address this challenge, [10] proposed a method for efficient black-box optimization over high-dimensional, structured input spaces, combining latent space optimization with weighted retraining of the VAE model. Weighted retraining modifies the latent space and uncovers new promising regions in $\mathcal{X}$ which the downstream

optimization algorithm can exploit. However, this method requires a large dataset of labelled data to train the VAE model and thus, weighted retraining is underwhelming for optimization problems in limited supervision scenarios.

Deep Metric Learning and High Dimensional Bayesian Optimization :

[11] incorporate the methods from deep metric learning to construct discriminative latent spaces for VAE Bayesian optimisation. They combined the metric learning loss functions : contrastive and triplet with VAE training to improve surrogate model's ability to generalise to unseen regions of the latent space and consequently improved generalisation performance in downstream Bayesian optimisation. Further, a semi-supervised scheme is adopted to achieve comparable or better performance compared to previous methods while using a small fraction of the data. However, even the performance of metric learning based approaches are dependent on the availability of good amount of data to learn the necessary invariances without which their performance might lack the desired robustness as pointed out in [13, 14].

3 Methodology : Learning Uncertainty-aware Black Box Augmentations

Our work aims at enhancing the sample-efficiency of the LSO with weighted retraining method by learning uncertainty-aware task specific data augmentations in this limited supervision setting. Incorporating invariances or prior knowledge in the machine learning models via data augmentation is extremely common and has gained a lot of popularity in the current years especially in limited supervision and unsupervised settings.

In standard settings, data augmentation refers to creating additional training examples by exploring the transformations for input data such that the predictions remain unchanged. However, in black box scenarios, we do not have prior knowledge about the various invariances in the task. Hence, generating relevant data augmentations becomes non-trivial, which leads us to learn invariances from the limited available data. Although there are works that aim to learn invariances from the data using generative models and inverse learning methods like normalizing flows, the majority of such methods need a large amount of data and knowledge about downstream tasks to learn the invariances. However, this is in direct contrast with the expensive black box optimization setting as the amount of labelled data is limited and hence these methods cannot be directly adopted in our setting.

In our work referred to as GP_LCB, we design the distribution of structured augmentations and systematically study the sensitivity of those augmentations to our black-box task. Since, our inputs are high-dimensional images, our augmentation set contains augmentation distributions, A, characterized by translation $T(x)$, rotation $R(x)$ and smoothing $S(x)$, where $x \in R^d$ is the input high-dimensional image. We understand that assuming invariance for the augmentations might be erroneous as we are unaware of the difference $||f(x) - f(a(x))||$ where $a \in A$. We also validate our understanding through a detailed empirical study as shown in the experimental section. Hence, we leverage the posterior predictive distribution of the iteratively trained Gaussian process to understand the effect of the augmentations on the task.

$$Z = q(X), Z_{aug} = q(a(X)) \tag{2}$$

$$P(f^*|Z, Y, Z_{aug}) = N(E[f^*|Z, Y, Z_{aug}], V[f^*|Z, Y, Z_{aug}) \tag{3}$$

$$E[f^*|Z, Y, Z_{aug}] = m(Z_{aug}) + k(Z_{aug}, Z)(K + \sigma_n^2 I)^{-1}(Y - m(Z)) \tag{4}$$

$$V[f^*|Z, Y, Z_{aug}] = k(Z_{aug}, Z_{aug}) - k(Z_{aug}, Z)(K + \sigma_n^2 I)^{-1}k(Z, Z_{aug}) \tag{5}$$

$E[f^*|Z, Y, Z_{aug}]$ represents the evaluation on the latent representation of the augmented images Z_{aug} and $V[f^*|Z, Y, Z_{aug}]$ represents the uncertainty in the prediction. We evaluate our entire augmentation distribution A with the GP posterior and implement systematic selective augmentation technique to enhance the dataset for weighted re-training of the generative model (VAE). [10] demonstrates that we can learn meaningful latent representation using VAE if the generative model is trained on a distribution that places more probability mass on high-scoring points. Hence, we sample the augmentations with higher value of the posterior expectation as shown in Eq (2) for our maximization problem. Our selection strategy also penalizes higher GP posterior variance, as we want to include the augmentation and the corresponding labels which exhibit higher confidence

Algorithm 1 Latent Space Weighted Retraining Optimization with Uncertainty-aware Labelled Augmentations

Input: Data $\mathcal{D} = (x_i, f(x_i))_{i=1}^{N}$, query budget M, objective function $f(x)$, latent objective model $h(z)$, generative/inverse model $g(z)/q(x)$, retrain frequency r, weighting function $w(x)$
Candidate Augmentations: $\mathcal{D}_{aug} = T(x) \cup R(x) \cup S(x); \mathcal{D}_{vae} = \mathcal{D}_{gp} = \mathcal{D}$
for $1, \ldots, M/r$ **do**
  Train VAE model with encoder/decoder : g/q on data $\mathcal{D}_{vae}$ weighted by $\mathcal{W} = \{w(x)\}_{x \in D_{vae}}$
  for $1, \ldots, r$ **do**
    Compute latent variables for data $\mathcal{D}_{gp}$; $\mathcal{Z} = \{z = q(x)\}_{x \in D}$
    Compute latent variables for augmentations $\mathcal{D}_{aug}$; $\mathcal{Z}_{aug} = \{z = q(x)\}_{x \in D_{aug}}$
    Fit surrogate GP h to $\mathcal{Z}$ and $\mathcal{D}$, and optimize h to obtain new latent query point $\mathbf{z}^*$
    Predictive mean and variance from GP for augmentations, based on uncertainty select $\mathbf{z}^*_{aug}$
    Obtain corresponding input $\mathbf{x}^*$ = g($\mathbf{z}^*$), $\mathbf{x}^*_{aug}$ = g($\mathbf{z}^*_{aug}$)
    Evaluate $f(\mathbf{x}^*)$ and augment data $\mathcal{D}_{gp}$
    Evaluate $f(\mathbf{x}^*_{aug})$ and augment data $\mathcal{D}_{vae}$
  end for
end for
Output: Augmented dataset $\mathcal{D}_{gp}$

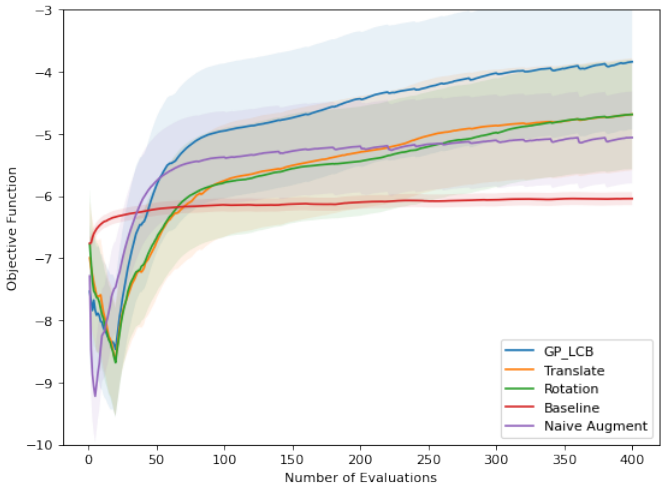


Figure 1: Objective function score on MNIST-Norm task

amongst the complete augmentation set, A along with high objective values. To achieve this, we incorporate a simple strategy similar to the Lower Confidence Bound (LCB) by computing the value of $E[f^*|Z, Y, Z_{aug}] - V[f^*|Z, Y, Z_{aug}]$ for $Z_{aug} = q(a(X))$ which ensures to select the augmentations $a(X), a \in A$ that maximize the objective but with minimal variance. This turns out to be a significant component in enhancing the performance of our data augmented weighted retraining method.

4 Experiments and Results

MNIST-Norm || Generate Digits with Minimum Norm Objective function : In this setting, we consider global minimization of the black box objective function defined as :

$$\mathbf{x}^\star = \arg\min_{\mathbf{x} \in \mathcal{X}} |\mathbf{x} - \mathbf{x}_{ref}| \tag{6}$$

where x_{ref} is the chosen reference image. Data used for this task are the normalized images from MNIST dataset. A low value of the objective function will lead to generating images that are closer to the reference image.

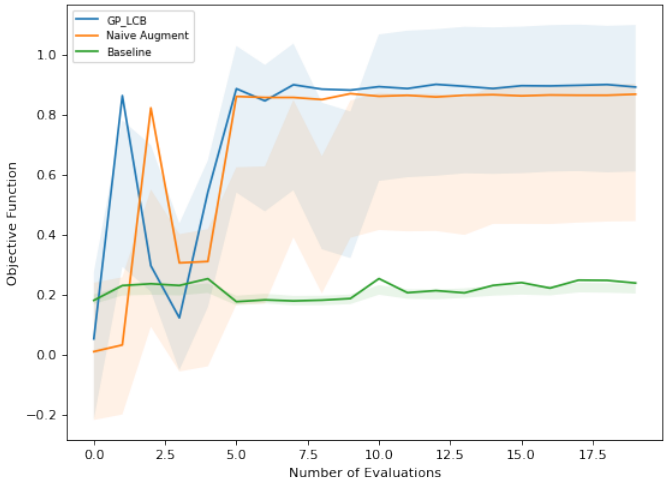


Figure 2: Objective function score on MNIST-Classifier task

MNIST-Classifier || Generate Digits with Expected Score objective function: Input to the black box are the images. We use an image classifier to get the probability distribution over the labels. A scoring function, modeled as a squared exponential centered at the digit '3', returns a score for each digit. The objective function is the expected score based on the image classifier and scoring function.

We compare our proposed approach with the methodologies mentioned below.

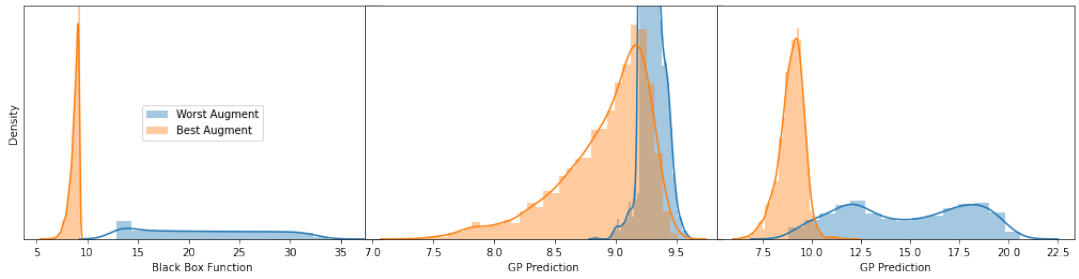
Naive Augment: Naive Augment method is the direct extension of self-supervised augmentation methods for our black box optimization setting. Here, objective function value for the augmented input A is same as the original input X. **Translate** : Consider only translation augmentation $T(x)$ and corresponding objective function values.

Rotate : Consider only rotation augmentation $R(x)$ and corresponding objective function values.

Baseline: LSO with weighted retraining as described in [10]

To understand the efficacy of our proposed method, we randomly chose to vary the distribution of initial data for 4 separate optimization runs and further, run all our experiments for 3 different seeds.

For MNIST-Norm task, GP_LCB method clearly outperforms other approaches considered for high dimensional BO. Notably, in MNIST-Classifier setting, Naive Augment method results in the correct objective function value for transformed inputs as image classification is invariant to the considered transformations A . However, GP_LCB method still remains competitive in the limited data setting even without any explicit information about the black box objective function.



(a) True Objective Function (b) GP Augmentation : Start (c) GP Augmentation : End

Figure 3: Density plot for Labelled augmentations from GP predictions across the high value and the low value image augmentations

(a) Reference Image

(b) GP Augmentation : Start

(c) GP Augmentation : Mid

(d) GP Augmentation : End

Figure 4: Top scoring image augmentations from GP predictions

5 Effect of Learning Labelled Augmentations on Optimization

For MNIST-Norm task, we perform the ablation studies to understand the effectiveness of learning labelled augmentations for high dimensional black box optimization. **We are interested in minimizing equation 6. Thus, high-scoring points will result in low objective function output.**

1. Methodology learns to identify high-scoring augmentations : We plot the learnt labelled augmentation values for the high-scoring and the low-scoring image augmentations. As the BO iteration progresses, our method shows success in discriminating between high-scoring and the low-scoring image augmentations.
2. Inspecting high-scoring augmentations : We plot the top-scoring augmentation from our method for three iterations: starting, mid and the end of the optimization. We note, our method augments the VAE training with high-scoring augmentations as the optimization progresses. This causes the feasible region to extend to high-scoring points, enabling the optimization to find better points as the optimization continues.

6 Conclusion and Discussion

The high dimensionality of the input feature space has been one of the most critical problems in Bayesian optimization while solving black-box optimization problems. The criticality further increases when the black box function is extremely expensive to evaluate. We propose a simplistic methodology to tackle the above problem and generate task-specific labelled augmentations using Gaussian processes (GP) and enhance the sample-efficiency of the VAE-BO LSO methods. Our uncertainty aware label-augmented weighted re-training method demonstrates superior performance in extremely limited supervision scenarios as shown in the experiments even starting with 20 datapoints in standard classification settings. Our approach directs towards a very important direction of learning task-specific data augmentations in limited supervision scenarios.

The experimental results indicate that incorporating invariances without understanding the effect of such invariances can hinder the performance on the downstream task and delay the convergence significantly. Also, our research highlights the importance of uncertainty estimation and calibration while performing data augmentation in limited supervision scenarios. As a part of our future work, we plan to incorporate the aspects of the VAE uncertainty as shown in [19] and metric VAE BO [11] along with the GP posterior variance to improve the quality of augmentations. We would also like to

extend our method for real world challenging tasks with non-image inputs, namely chemical design [20, 21] and expression reconstruction [22].

Acknowledgments and Disclosure of Funding

We would like to express our gratitude to Haitham Bou Ammar and Rasul Tutunov for mentorship and helpful feedback throughout the work.

References

[1] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science.*, 11:577–586, 2020.

[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] Xiaoyu Lu, Javier Gonzalez, Zhenwen Dai, and Neil Lawrence. Structured variationally auto-encoded optimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3267–3275. PMLR, 10–15 Jul 2018.

[4] Stephan Eissman, Daniel Levy, Rui Shu, Stefan Bartzsch, and Stefano Ermon. Bayesian optimization and attribute adjustment. In *Proc. 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

[5] Riccardo Moriconi, Marc P. Deisenroth, and K. S. Sesh Kumar. High-dimensional bayesian optimization using low-dimensional feature spaces, 2020.

[6] Eero Siivola, Javier Gonzalez, Andrei Paleyes, and Aki Vehtari. Good practices for bayesian optimization of high dimensional structured spaces, 12 2020.

[7] Rika Antonova, Akshara Rai, Tianyu Li, and Danica Kragic. Bayesian optimization in variational latent spaces with dynamic compression. In *Conference on Robot Learning*, pages 456–465, 2020.

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

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] Austin Tripp, Erik Daxberger, and José Miguel Hernández-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. *Advances in Neural Information Processing Systems*, 33, 2020.

[11] Antoine Grosnit, Rasul Tutunov, Alexandre Max Maraval, Ryan-Rhys Griffiths, Alexander Imani Cowen-Rivers, Lin Yang, Lin Zhu, Wenlong Lyu, Zhitang Chen, Jun Wang, Jan Peters, and Haitham Bou-Ammar. High-dimensional bayesian optimisation with variational autoencoders and deep metric learning. *CoRR*, abs/2106.03609, 2021.

[12] Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pages 521–528, 2002.

[13] Byungsoo Ko and Geonmo Gu. Embedding expansion: Augmentation in embedding space for deep metric learning, 2020.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[15] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA, 2006.

[16] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 1(104):148–175, 2016.

[17] Peter I Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[18] Ryan Turner, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. *CoRR*, abs/2104.10201, 2021.

[19] Pascal Notin, José Miguel Hernández-Lobato, and Yarin Gal. Improving black-box optimization in VAE latent space using decoder uncertainty. *CoRR*, abs/2107.00096, 2021.

[20] Jianhui Chen, Zheng Zhao, Jieping Ye, and Huan Liu. Nonlinear adaptive distance metric learning for clustering. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 123–132, 2007.

[21] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2323–2332. PMLR, 10–15 Jul 2018.

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