De Novo Molecular and Crystal Design with Latent Space Bayesian Optimization

Onur Boyar

Department of Mechanical Systems Engineering
Nagoya University
boyaronur@gmail.com

Abstract

This thesis explores Latent Space Bayesian Optimization (LSBO) for the generation and optimization of de novo molecules and crystal materials. Our goal is to develop practical, sample-efficient de novo discovery algorithms with a focus on real-world applicability, and our results so far demonstrate significant progress toward practical implementation.

Introduction

The design of de novo molecules and materials is crucial for scientific and industrial applications. Creating de novo molecules and discovering novel drugs can result in groundbreaking medical treatments, while developing new materials with specific properties has the potential to revolutionize industries such as semiconductors. However, designing such high-dimensional objects presents inherent challenges due to the complexity and high dimensionality of molecular and material structures. The vast search space makes finding optimal candidates computationally expensive and time-consuming. Most optimization algorithms struggle with the vast search space. This makes conventional approaches impractical for discovering truly novel chemical entities. *Latent Space Bayesian Optimization* (LSBO) is a promising methodology frequently used in the AI for Science community that addresses these challenges by reducing search space complexity through generative models like Variational Autoencoders (VAEs). By mapping high-dimensional structures into lower-dimensional latent representations, LSBO enables Bayesian Optimization (BO) to operate more efficiently thanks to lower dimensional search space obtained via the VAE. This strategy has proven effective in applications such as molecular generation and material discovery.

Problem Definition

Our objective is to generate a molecule (or crystal) that optimizes a specific property $\mathcal{P}$, such as drug-likeness, determined by the Black-Box (BB) function f^{BB}. This function is typically costly to evaluate (e.g. wet-lab experiment), therefore we aim to optimize f^{BB} with as low evaluations as possible. BO, known for its sample-efficient optimization through a balance of exploration and exploitation, is an effective tool for optimizing such BB functions. The molecule exists in the input space $\mathcal{X}$, and our optimization challenge is to find $\boldsymbol{x}^* \in \mathcal{X}$ that maximizes $f^{\text{BB}}(\boldsymbol{x})$:

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}\in\mathcal{X}} f^{\text{BB}}(\boldsymbol{x}). \tag{1}$$

Due to the high dimensionality of $\mathcal{X}$ and the cost of evaluating f^{BB}, direct optimization in $\mathcal{X}$ is impractical, as BO does not work as expected in such high dimensional search spaces. Instead, we train a VAE with encoder f^{enc} and decoder f^{dec} on a dataset of related instances and perform optimization in the d-dimensional latent space of VAE, $\mathcal{Z} \in \mathbb{R}^d$, using BO. The latent space optimization problem is defined as:

$$\boldsymbol{z}^* = \arg\max_{\boldsymbol{z}\in\mathcal{Z}} g(\boldsymbol{z}), \tag{2}$$

where $g(\boldsymbol{z}) = f^{\text{BB}}(f^{\text{dec}}(\boldsymbol{z}))$ maps a latent point back to the input space to evaluate its property $\mathcal{P}$ via the BB function.

Background

LSBO has emerged as a popular framework for optimizing expensive-to-evaluate black-box functions in the context of generative modeling. Seminal work by Gómez-Bombarelli et al. (2018) first introduced latent space optimized based approach for de novo molecular design, where VAE is combined with a gradient-based search methodology to identify molecules with desired properties. Following this, several studies expanded on LSBO's capabilities. For instance, Tripp, Daxberger, and Hernández-Lobato (2020) proposed weighted retraining to ensure the model undergoes a retraining after each iteration of BO where weights of each training instance decided based on the target task, and Grosnit et al. (2021) extended their idea by incorporating metric learning loss into the retraining process to organize the latent space based on the property values of the molecules.

However, these methods encounter significant challenges. The direct combination of VAEs and BO often leads to suboptimal results due to the inherent mismatch between their objectives. VAEs typically generate instances similar to the training data, focusing on dense regions of the latent space, while BO seeks to explore new, often sparsely populated regions to identify *de novo* instances—ones that differ substantially from those seen during training. Existing studies address this by retraining the VAE with new instances

queried during LSBO to update the model, but this requires many costly BB function evaluations for meaningful updates, making these methods impractical in real settings.

Proposed Methodology

To address the challenges discussed, we introduce an LSBO approach enhanced by latent data augmentations (Boyar and Takeuchi 2023, 2024), referred to as *Latent Consistency-Aware LSBO* (LCA-LSBO). We use augmented latent points from a distribution $p(\hat{z}) \sim \mathcal{N}(\mu, \sigma)$, focusing on regions identified as promising by LSBO, such as those with high acquisition values or has favorable properties. The mean μ is set based on these regions, and augmented instances are generated by sampling with a variance $\sigma^2 < 1$ to focus the augmentations on the promising region. These synthetic instances are then re-encoded, and their deviations are penalized. LCA-LSBO uses a variant of a VAE model, LCA-VAE, that uses these augmentated instances in the retraining process. The objective function of LCA-VAE is:

$$J_{\text{LCA-VAE}}(\phi, \theta) = J_{\text{VAE}}(\phi, \theta) - \gamma \mathbb{E}_{\hat{z} \sim p(\hat{z})}[\text{LCL}(\hat{z})] \quad (3)$$

where J_{VAE} is the standard VAE loss, and γ controls the contribution of the latent consistency loss (LCL). The LCL is defined as $\text{LCL}(\hat{z}) = \|\hat{z} - f^{\text{enc}}(f^{\text{dec}}(\hat{z}))\|^2$. LCL ensures that latent points remain consistent after reconstruction, effectively reinforcing the model's learning by penalizing inconsistencies. This approach allows LCA-VAE to learn from diverse synthetic points that are from promising regions in the latent space, thereby reducing reliance on expensive real data queries and enhancing exploration in a sample-efficient manner, making the LCA-LSBO framework suitable for real-world applications.

Application to Molecular and Crystal Design: The proposed LCA-LSBO framework showed significant improvements in generating molecules with desired properties. Our experiments targeted molecules with high docking affinity to an ion-channel protein, which can aid drug discovery for conditions such as chronic pain and cardiovascular diseases. Our model successfully identified a molecule with a docking score better than those of 50,000 randomly sampled molecules, achieving this result in under 350 iterations. This highlights the efficiency of our approach in navigating the chemical space and discovering promising candidates with fewer evaluations compared to standard methods.

Building on this foundation, we extended the LCA-LSBO framework to de novo crystal material design in our work on *Crystal-LSBO* (Boyar et al. 2024). Applying LSBO to crystal design posed unique challenges due to the inherent complexity of crystal materials, prompting us to develop a novel VAE structure specifically tailored for this setting. In our de novo design setting, we focused on optimizing formation energy as the target property, which is a key property for material stability. Crystal-LSBO, which utilizes the idea of LCA-LSBO in its optimization process, discovered novel crystal structures that significantly outperformed existing methods. Our results demonstrated that Crystal-LSBO effectively explored a broader latent space, discovering crystal structures that conventional VAEs and diffusion model-based approaches could not, showcasing its performance. Overall, our results underline the potential of the LCA-LSBO framework in both molecular and material design.

Research Progress and Future Directions: Our research began by identifying the challenges in LSBO and developing a methodology suitable for real-world settings with a focus on sample efficiency. This led to the development of LCA-LSBO (Boyar and Takeuchi 2023, 2024), which was later applied to crystal design (Boyar et al. 2024). Despite this progress, assessing the *real-world applicability* of newly generated molecules or crystals remains a challenge, as domain experts struggle to evaluate the synthesizability of these novel structures, reducing their chances of being produced and used in real settings. To address this, our current research aims to develop an LSBO-based methodology that modifies existing molecules to optimize a target property, as modifying known molecules is more likely to yield synthesizable compounds. We have already achieved promising results, including the successful wet-lab synthesis of an optimized molecule. By the time of the doctoral consortium, I aim to submit a paper on our methodology to a machine learning journal and conduct research to extend our framework for Human-in-the-Loop use by domain experts through a web application. Additionally, I expect to receive reviews for our Crystal-LSBO paper and will refine it accordingly.

Contributions: Our work in (Boyar and Takeuchi 2023, 2024) was developed with my supervisor, while research in (Boyar et al. 2024) conducted in collaboration with material scientists, who contributed domain expertise, while I led the development and implementation of the methodology.

References

Boyar, O.; Gu, Y.; Tanaka, Y.; Tonogai, S.; Itakura, T.; and Takeuchi, I. 2024. Crystal-LSBO: Automated Design of De Novo Crystals with Latent Space Bayesian Optimization. arXiv:2405.17881.

Boyar, O.; and Takeuchi, I. 2023. Enhanced Exploration in Latent Space Bayesian Optimization. In *International Workshop on Pattern Recognition in Healthcare Analytics, Asian Conference on Machine Learning*.

Boyar, O.; and Takeuchi, I. 2024. Latent Space Bayesian Optimization With Latent Data Augmentation for Enhanced Exploration. *Neural Computation*, 1–33.

Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; and Aspuru-Guzik, A. 2018. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2): 268–276.

Grosnit, A.; Tutunov, R.; Maraval, A. M.; Griffiths, R.-R.; Cowen-Rivers, A. I.; Yang, L.; Zhu, L.; Lyu, W.; Chen, Z.; Wang, J.; Peters, J.; and Bou-Ammar, H. 2021. High-Dimensional Bayesian Optimisation with Variational Autoencoders and Deep Metric Learning. *ArXiv*, abs/2106.03609.

Tripp, A.; Daxberger, E.; and Hernández-Lobato, J. 2020. Sample-Efficient Optimization in the Latent Space of Deep Generative Models via Weighted Retraining. In *Advances in Neural Information Processing Systems*.