Deep Learning for Bayesian Optimization of Scientific Problems with High-Dimensional Structure

Samuel Kim SAMKIM@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Peter Y. Lu
Department of Physics
Massachusetts Institute of Technology

Charlotte Loh
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Jamie Smith
Google Research

Jasper Snoek
Google Research

Marin Soljačić SOLJACIC@MIT.EDU
Department of Physics
Massachusetts Institute of Technology

Abstract

Bayesian optimization (BO) is a popular paradigm for global optimization of expensive black-box functions, but there are many domains where the function is not completely a black-box. The data may have some known structure and/or the function can yield useful intermediate information. We propose performing BO on complex, structured problems by using multi-output Bayesian neural networks. We demonstrate BO on a number of realistic problems in physics and chemistry, including topology optimization of photonic crystal materials using convolutional neural networks, and chemical property optimization of molecules using graph neural networks. On these complex tasks, we show that neural networks often outperform GPs as surrogate models for BO in terms of both sampling efficiency and computational cost.

1. Introduction

Bayesian optimization (BO) is a methodology well-suited for global optimization of expensive, black-box functions. However, in many domains, the system is not a complete black box. For example, complex, high-dimensional input spaces such as images or molecules have some known structure, symmetries and invariances. In addition, the function may be a composite function in which it may provide intermediate or auxiliary information from which the objective function can be cheaply computed. For example, a scientific experiment or simulation may produce a high-dimensional observation or multiple measurements

simultaneously, such as the optical scattering spectrum of a nanoparticle over a range of wavelengths, or multiple quantum chemistry properties of a molecule from a single density functional theory (DFT) calculation. All of these physically-informed insights into the system are potentially useful and important factors for designing surrogate models through inductive biases, but they are often not fully exploited in existing methods and applications.

BO relies on specifying a surrogate model which are typically Gaussian Processes (GPs), as the posterior distribution of GPs can be expressed analytically. Multi-output GPs have been used to apply BO to synthetic problems that can be decomposed into composite functions (Astudillo and Frazier, 2019). GP kernels have also been formulated for complex input spaces including convolutional kernels (Van der Wilk et al., 2017; Novak et al., 2020; Wilson et al., 2016) and graph kernels (Shervashidze et al., 2011; Walker and Glocker, 2019). However, (1) time complexity of GPs scales with the number of observations and output dimensionality, limiting their use to smaller problems, and (2) GPs operate most naturally over continuous input spaces, so kernels for high-dimensional, structured data must be carefully formulated and tuned by hand for each new domain.

In contrast, neural networks and Bayesian neural networks have been proposed as an alternative to GPs in BO due to their scalability and flexibility (Snoek et al., 2015; Springenberg et al., 2016). This approach also enables BO in more complex settings including transfer learning across multiple tasks and modeling of auxiliary signals to improve performance (Perrone et al., 2018).

This work demonstrates the use of deep learning to enable BO for complex, real-world scientific datasets. In particular, (1) we take advantage of auxiliary or intermediate information from composite functions, (2) we demonstrate BO on complex input spaces including images and molecules using convolutional and graph neural networks, respectively, and (3) we apply BO to several realistic scientific datasets, including topology optimization of a photonic crystal material, and chemical property optimization of molecules from the QM9 dataset. We show that neural networks are often able to significantly outperform GPs as surrogate models on these problems, and we believe that these strong results will also generalize to other contexts and enable BO to be applied to a wider range of problems.

We present limited results here for brevity, and more complete results can be found in Kim et al. (2021).

2. Bayesian Optimization

We formulate our optimization task as a maximization problem in which we wish to find the input $\mathbf{x}^* = \arg\max_{\mathbf{x}} f(\mathbf{x})$. In iteration N, a Bayesian surrogate model $\mathcal{M}$ is trained on a labeled dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$. An acquisition function α then uses $\mathcal{M}$ to suggest the next data point to label, $\mathbf{x}_{N+1} = \arg\max_{\mathbf{x}\in\mathcal{X}} \alpha(\mathbf{x}; \mathcal{M}, \mathcal{D}_{\text{train}})$, which is then evaluated and added to $\mathcal{D}_{\text{train}}$.

We use the expected improvement (EI) acquisition function α_{EI} (Jones et al., 1998). When the posterior predictive distribution of the surrogate model $\mathcal{M}$ is a normal distribution, EI can be expressed analytically. For surrogate models that do not give an analytical form for the posterior predictive distribution, we sample from the posterior N_{MC} times and use a Monte Carlo approximation of EI (Wilson et al., 2018).

2.1 Continued Training with Learning Rate Annealing

To minimize the training time of BNNs in each optimization loop, we use the model that has been trained in the Nth optimization loop iteration as the initialization (also known as a "warm start") for the $(N+1)$th iteration, rather than training from a random initialization. In particular, we use the cosine annealing learning rate proposed in Loshchilov and Hutter (2016).

2.2 Auxiliary Information

Here we consider the case where f is a composite function and can be decomposed as $f(\mathbf{x}) = h(g(\mathbf{x}))$ where $g\colon \mathcal{X} \to \mathcal{Z}$ is the expensive labeling process, and $h\colon \mathcal{Z} \to \mathcal{Y}$ is a known objective function that can be cheaply computed (Astudillo and Frazier, 2019; Balandat et al., 2020). The evaluation function thus provides some intermediate or auxiliar information $\mathbf{z} \in \mathcal{Z}$. In this case, we train $\mathcal{M}\colon \mathcal{X} \to \mathcal{Z}$ to model g, and the approximate EI acquisition function becomes

$$\alpha_{\text{EI-MC-aux}}(\mathbf{x}) = \frac{1}{N_{\text{MC}}} \sum_{i=1}^{N_{\text{MC}}} \max\left(h\left(\mu^{(i)}(\mathbf{x})\right) - y_{\text{best}}, 0\right). \tag{1}$$

which can be seen as a Monte Carlo version of the acquisition function presented in Astudillo and Frazier (2019). We denote models trained using auxiliary information with the suffix "-aux."

3. Results

For the BNN, we use an **ensemble** of $N_{\text{MC}} = 10$ neural networks with identical architectures, and use Eq. 1 for acquisition. We have experimented with multiple BNN approximations including variational inference (VI) approaches (such as Bayes by Backprop (Blundell et al., 2015) and Multiplicative Normalizing Flows (MNF) (Louizos and Welling, 2017)), SGHMC as implemented by BOHAMIANN (Springenberg et al., 2016), Neural Linear (Snoek et al., 2015), and infinite-width/ensemble approximations (Novak et al., 2020), and found that ensembles perform better and more consistently over other types of BNNs.

For our baselines, we use **GP** to refer to a specific, standard specification that uses a Matérn 5/2 kernel. To operate on images, we use a convolutional kernel, labeled as **ConvGP**, where the implementation is the infinite-ensemble limit of a convolutional neural network (Novak et al., 2020). Finally, to operate directly on graphs, we use the Weisfeiler-Lehman (WL) kernel as implemented by (Ru et al., 2021). Additionally, we compare against **GP-aux** which use multi-output GPs for composite functions (Astudillo and Frazier, 2019).

All BO results are averaged over multiple trials, and the shaded area in the plots represents ± one standard error over the trials.

3.1 Photonic Crystal Topology

Next we look at a more complex, high-dimensional domain that contains symmetries not easily exploitable by GPs. Photonic crystals (PCs) are nanostructured materials that are engineered to exhibit exotic optical properties not found in bulk materials, including photonic

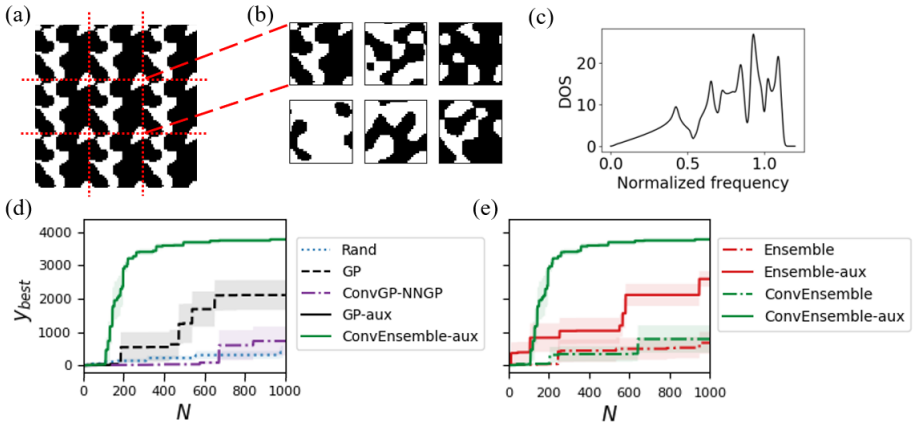

Figure 1: (a) A 2D photonic crystal (PC). The black and white regions represent different materials, and the periodic unit cells are outlined in red. Examples of PC unit cells. (c) Example of a PC density of states (DOS). (d, e) BO results.

band gaps and negative index of refraction (John, 1987; Yablonovitch, 1987; Joannopoulos et al., 2008). As advanced fabrication techniques are enabling smaller and smaller feature sizes, there has been growing interest in inverse design and topology optimization to design even more sophisticated PCs (Jensen and Sigmund, 2011; Men et al., 2014; Piggott et al., 2015; Lin et al., 2019).

Here we consider 2D PCs consisting of periodic unit cells represented by a 32×32 pixel image, as shown in Figure 1(a), with white and black regions representing vacuum (or air) and silicon, respectively. Because optimizing over raw pixel values may lead to pixel-sized features or intermediate pixel values that cannot be fabricated, we have parameterized the PCs with a level-set function that converts $\mathbf{x} \in \mathbb{R}^{51}$ into an image $\mathbf{v} \in \mathbb{R}^{32\times32}$ that represents the PC.

The optical properties of PCs can be characterized by their photonic density of states (DOS), e.g. see Figure 1(c). We choose an objective function h that aims to minimize the DOS in a certain frequency range while maximizing it everywhere else, which corresponds to opening up a photonic band gap in said frequency range. While we train GPs directly on the level-set parameters $\mathcal{X}$, we can train the Bayesian convolutional NNs (BCNNs) on the more natural unit cell image space $\mathcal{V}$. BCNNs can also be trained to predict the full DOS as auxiliary information $\mathbf{z} \in \mathbb{R}^{500}$, in which we use Equation 1 for acquisition.

The BO results in Figure 1(d), show that BCNNs outperform GPs by a significant margin on both datasets, which is due to both the auxiliary information and the inductive bias of the convolutional layers, as shown in Figure 1(e). Because the behavior of PCs is determined by their topology rather than individual pixel values or level-set parameters, BCNNs are much better suited to analyze this dataset compared to GPs. Additionally, BCNNs can be made much more data-efficient since they directly encode translation invariance and thus learn the behavior of a whole class of translated images from a single image. Because GP-aux is extremely expensive compared to GP (500× longer on this dataset), we

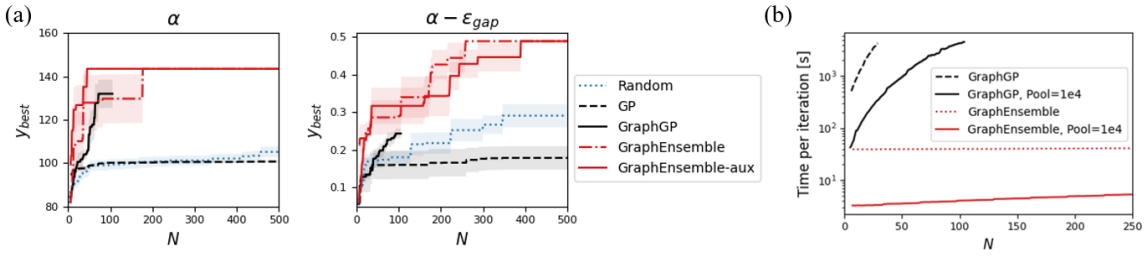

Figure 2: (a) Quantum chemistry task BO resultsfor various properties. (b) Time per BO iteration. (Note the logarithmic scale on the y-axis.) GraphGP takes orders of magnitudes longer than BGNNs for moderate N.

are only able to run GP-aux for a small number of iterations, where it performs comparably to random sampling. ConvGP only performs slightly better than random sampling, which is likely due to a lack of auxiliary information and inflexibility to learn the most suitable representation for this dataset.

3.2 Organic Molecule Quantum Chemistry

Finally, we optimize the chemical properties of molecules. Chemical optimization is of huge interest with applications in drug design and materials optimization (Hughes et al., 2011; Gómez-Bombarelli et al., 2018; Korovina et al., 2020). This is a difficult problem where computational approaches such as density functional theory (DFT) can take days for simple molecules and are intractable for larger molecules; synthesis is expensive and time-consuming, and the space of synthesizable molecules is large and complex.

Here we focus on the QM9 dataset (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014), which consists of 133,885 small organic molecules along with their geometric, electronic, and thermodynamics quantities that have been calculated with DFT. Instead of optimizing over a continuous space, we draw from the fixed pool of available molecules and iteratively select the next molecule to add to $\mathcal{D}_{\text{train}}$. We use a Bayesian graph neural network (BGNN) for our surrogate model. For the GP baseline, we encode the molecules as a continuous vector using the Smooth Overlap of Atomic Positions (SOAP) descriptor (De et al., 2016; Himanen et al., 2020).

We compare two different optimization objectives derived from the QM9 dataset: the isotropic polarizability α and $(\alpha - \epsilon_{\text{gap}})$ where ϵ_{gap} is the HOMO-LUMO energy gap. Because many of the chemical properties in the QM9 dataset can be collectively computed by a single DFT or molecular dynamics calculation, we can treat a group of labels from QM9 as auxiliary information $\mathbf{z}$ and train our BGNN to predict this entire group simultaneously. The objective function h then simply picks out the property of interest.

As shown in Figure 2(c), BGNNs and GraphGPs significantly outperform GPs, showing that the inductive bias in the graph structure leads to a much more natural representation of the molecule and its properties. In the case of maximizing the polarizability α, including the auxiliary information improves BO performance, showing signs of positive transfer. As

seen in Figure 2(b), we also note that the GraphGP is relatively computationally expensive ($15\times$ longer than GPs for small N and $800\times$ longer than BGNNs for $N = 100$) and so we are only able to run it for a limited N in a reasonable time frame. BGNNs perform comparably or better than GraphGPs despite incurring a fraction of the computational cost.

4. Discussion

We have demonstrated global optimization on multiple tasks using a combination of deep learning and BO. In particular, we have shown how BNNs can enables the scaling of BO to large datasets and provides the flexibility to incorporate a wide variety of inductive biases and auxiliary information. We note that our method is not necessarily tied to any particular application domain, and can lower the barrier of entry for design and optimization.

We conjecture that the additional information forces the BNN to learn a more consistent physical model of the system since it must learn features that are shared across the multi-dimensional auxiliary information, thus enabling the BNN to generalize better. It is also possible that the loss landscape for the auxiliary information is smoother than that of the objective function and that the auxiliary information acts as an implicit regularization that improves generalization performance.

There is an interesting connection between how well BNNs are able to capture and explore a multi-modal posterior distribution and their performance in BO. For example, we have noticed that larger batch sizes tend to significantly hurt BO performance. On the one hand, larger batch sizes may be resulting in poorer generalization as the model finds sharper local minima in the loss landscape. Another explanation is that the stochasticity inherent in smaller batch sizes allows the BNN to more easily explore the posterior distribution, which is known to be highly multi-modal (Fort et al., 2019). Indeed, BO often underperforms for very small dataset sizes N but quickly catches up as N increases, indicating that batch size is an important hyperparameter which must be balanced with computational cost.

When comparing BNN architectures, we find that ensembles tend to consistently perform among the best, which is supported by previous literature showing that ensembles capture uncertainty much better than variational methods (Ovadia et al., 2019; Gustafsson et al., 2020) especially in multi-modal loss landscapes (Fort et al., 2019). Ensembles are also attractive because they require no additional hyperparameters and they are simple to implement.

Future work will consider using stochastic training approaches such as SG-MCMC methods for exploring posterior distributions (Welling and Teh, 2011; Zhang et al., 2019) as well as other continual learning techniques to further minimize training costs, especially for larger datasets (Parisi et al., 2019). Future work will also investigate more complex BNN architectures with stronger inductive biases. For example, output constraints can be placed through unsupervised learning (Karpatne et al., 2017) or by variationally fitting a BNN prior (Yang et al., 2020). Custom architectures have also been proposed for partial differential equations (Raissi et al., 2017; Lu et al., 2020), many-body systems (Cranmer et al., 2020), and generalized symmetries (Hutchinson et al., 2020), which will enable effective BO on a wider range of tasks. The methods and experiments presented here enable BO to be effectively applied in a wider variety of settings.

Acknowledgments

The authors would like to acknowledge Rodolphe Jenatton, Thomas Christensen, Andrew Ma, Rumen Dangovski, Joy Zeng, Charles Roques-Carmes, and Mohammed Benzaouia for fruitful conversations. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper. This work is supported in part by the the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, `http://iaifi.org/`). This research was also sponsored in part by the Department of Defense through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. This material is based upon work partly supported by the Air Force Office of Scientific Research under the award number FA9550-21-1-0317, as well partly supported by the US Office of Naval Research (ONR) Multidisciplinary University Research Initiative (MURI) grant N0 0014-20-1-2325 on Robust Photonic Materials with High-Order Topological Protection It is also based upon work supported in part by the U.S. Army Research Office through the Institute for Soldier Nanotechnologies at MIT, under Collaborative Agreement Number W911NF-18-2-0048. Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Raul Astudillo and Peter Frazier. Bayesian optimization of composite functions. In *International Conference on Machine Learning*, pages 354–363. PMLR, 2019.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. Botorch: a framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538, 2020.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. *arXiv preprint arXiv:2006.11287*, 2020.

Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

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.

Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020.

Lauri Himanen, Marc O. J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020. ISSN 0010-4655. doi: 10.1016/j.cpc.2019.106949. URL `https://doi.org/10.1016/j.cpc.2019.106949`.

James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

Michael Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. *arXiv preprint arXiv:2012.10885*, 2020.

Jakob Søndergaard Jensen and Ole Sigmund. Topology optimization for nano-photonics. *Laser & Photonics Reviews*, 5(2):308–321, 2011.

John D. Joannopoulos, Steven G. Johnson, Joshua N. Winn, and Robert D. Meade. *Photonic Crystals: Molding the Flow of Light (Second Edition)*. Princeton University Press, 2 edition, 2008. ISBN 0691124566.

Sajeev John. Strong localization of photons in certain disordered dielectric superlattices. *Physical review letters*, 58(23):2486, 1987.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, Dec 1998. ISSN 1573-2916. doi: 10.1023/A:1008306431147. URL `https://doi.org/10.1023/A:1008306431147`.

Anuj Karpatne, William Watkins, Jordan Read, and Vipin Kumar. Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv preprint arXiv:1710.11431*, 2017.

Samuel Kim, Peter Y Lu, Charlotte Loh, Jamie Smith, Jasper Snoek, and Marin Soljačić. Deep learning for bayesian optimization of scientific problems with high-dimensional structure. *arXiv preprint arXiv:2104.11667*, 2021.

Ksenia Korovina, Sailun Xu, Kirthevasan Kandasamy, Willie Neiswanger, Barnabas Poczos, Jeff Schneider, and Eric Xing. Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations. In *International Conference on Artificial Intelligence and Statistics*, pages 3393–3403. PMLR, 2020.

Zin Lin, Victor Liu, Raphaël Pestourie, and Steven G Johnson. Topology optimization of freeform large-area metasurfaces. *Optics express*, 27(11):15765–15775, 2019.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. *arXiv preprint arXiv:1703.01961*, 2017.

Peter Y Lu, Samuel Kim, and Marin Soljačić. Extracting interpretable physical parameters from spatiotemporal systems using unsupervised learning. *Physical Review X*, 10(3): 031056, 2020.

Han Men, Karen YK Lee, Robert M Freund, Jaime Peraire, and Steven G Johnson. Robust topology optimization of three-dimensional photonic-crystal band-gap structures. *Optics express*, 22(19):22632–22648, 2014.

Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A. Alemi, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*, 2020. URL `https://github.com/google/neural-tangents`.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's

uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13991–14002, 2019.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Valerio Perrone, Rodolphe Jenatton, Matthias Seeger, and Cédric Archambeau. Scalable hyperparameter transfer learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6846–6856, 2018.

Alexander Y Piggott, Jesse Lu, Konstantinos G Lagoudakis, Jan Petykiewicz, Thomas M Babinec, and Jelena Vučković. Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer. *Nature Photonics*, 9(6):374–377, 2015.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1): 1–7, 2014.

Binxin Ru, Xingchen Wan, Xiaowen Dong, and Michael Osborne. Interpretable neural architecture search via bayesian optimisation with weisfeiler-lehman kernels. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=j9Rv7qdXjd`.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*, pages 2171–2180, 2015.

Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization with robust bayesian neural networks. *Advances in neural information processing systems*, 29:4134–4142, 2016.

Mark Van der Wilk, Carl Edward Rasmussen, and James Hensman. Convolutional gaussian processes. *arXiv preprint arXiv:1709.01894*, 2017.

Ian Walker and Ben Glocker. Graph convolutional gaussian processes. In *International Conference on Machine Learning*, pages 6495–6504. PMLR, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR, 2016.

James Wilson, Frank Hutter, and Marc Deisenroth. Maximizing acquisition functions for bayesian optimization. *Advances in Neural Information Processing Systems*, 31:9884–9895, 2018.

Eli Yablonovitch. Inhibited spontaneous emission in solid-state physics and electronics. *Physical review letters*, 58(20):2059, 1987.

Wanqian Yang, Lars Lorch, Moritz A Graule, Himabindu Lakkaraju, and Finale Doshi-Velez. Incorporating interpretable output constraints in bayesian neural networks. *arXiv preprint arXiv:2010.10969*, 2020.

Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902.03932*, 2019.