# Dimension Deficit: Is 3D a Step Too Far for Optimizing Molecules?

**Andres Guzman-Cordero**[*,1,4] **Luca Thiede**[2,4] **Gary Tom**[2,4] **Alán Aspuru-Guzik**[2,4]
**Felix Strieth-Kalthoff**[3] **Agustinus Kristiadi**[*,4]

[1]University of Amsterdam
[2]University of Toronto
[3]University of Wuppertal
[4]Vector Institute

## Abstract

The discovery of new materials with desirable properties is essential for technological advancements, from pharmaceuticals to renewable energy. Traditional simulation methods like Density Functional Theory (DFT) provide ab initio quantum calculation estimates of common properties but are computationally expensive, prompting the need for carefully selecting candidates for the calculation. Bayesian optimization (BO) is commonly used to efficiently find and screen candidates. However, choosing the right vector representations for a Bayesian regressor is challenging: While molecules are 3-dimensional, obtaining 3D features is computationally intensive, so 1D and 2D features are typically used. In this work, we study this discrepancy. Are 3D features worth considering for BO over molecules despite their computational complexity? To this end, we evaluate the molecular fingerprint representation, 2D message-passing neural networks, and 3D equivariant attention-based graph neural networks. We evaluate their performance on four datasets, considering both low- and high-data regimes and different types of Bayesian regressors. Finally, we explore the transfer learning capabilities of 2D and 3D graph features by treating the graph networks as foundation models.

## 1 Introduction

The discovery of new materials is crucial for technological advancements, yet Density Functional Theory (DFT), the gold standard for predicting molecular properties, is limited by its high computational cost, especially for large datasets or complex molecules [39]. To overcome this, Bayesian optimization (BO) has emerged as a promising method for efficiently exploring the vast space of potential materials and guiding experimental efforts toward the most promising candidates for DFT calculations [30, 19, 21, 16]. BO typically relies on training probabilistic surrogate functions, such as Gaussian Processes (GPs) or Bayesian Neural Networks (BNNs), using vector representations of molecules. However, the importance of selecting an appropriate molecular representation is often overlooked. While transformers and graph neural networks (GNNs) have been used [19], they assume molecules are represented by 1D representations, like strings, or 2D graphs, which simplifies their inherent three-dimensional nature.

Recently, 3D GNNs, particularly equivariant attention-based GNNs, have shown promise in capturing the full geometric structure and the symmetries of molecules, potentially offering superior predictive performance [11, 22, 23, 3, 7]. This work revisits whether incorporating 3D molecular features into BO justifies the increased computational cost compared to 1D and 2D features. We evaluate

[*]Correspondence to: andresguzco@gmail.com, akristiadi@vectorinstitute.ai.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

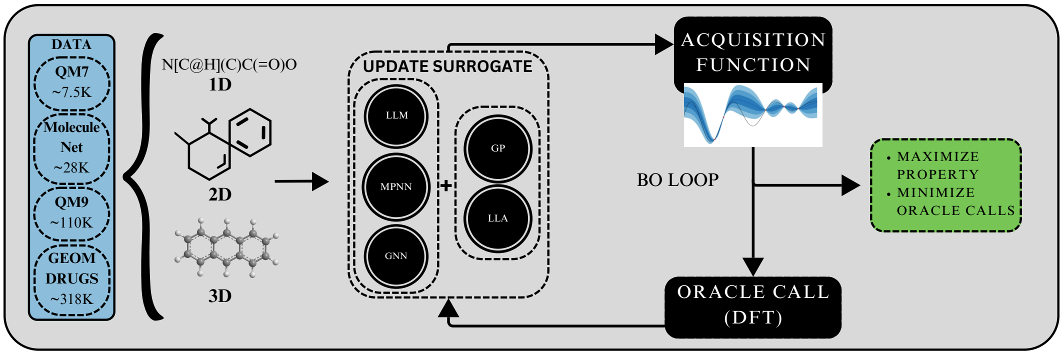

Figure 1: Comparison of the performance of 1D, 2D, and 3D representations in BO for molecular property optimization. The models are tested on four datasets across varying data sizes to assess the trade-offs between computational cost and predictive accuracy for each representation.

Gaussian Process regression and Large Language Model (LLM) prediction with 1D representations, 2D Message-Passing Neural Networks (MPNNs), and 3D equivariant attention-based GNNs across various data regimes, assessing predictive performance using Laplace-approximated BNNs and GPs, shown in Fig. 1. This study aims to determine if leveraging 3D structural information significantly enhances BO's effectiveness in materials discovery.

## 2 Preliminaries

**Bayesian Optimization** BO considers the problem of finding a global maximizer of an unknown objective function $f : \mathcal{X} \to \mathcal{Y}$ denoted $\mathbf{x}_* = \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$, over some $d$-dimensional search space $\mathcal{X}$. Although $f$ lacks a simple closed form, it can be evaluated at any point $\mathbf{x}$ in the domain, and the goal is to minimize the number of evaluations [37]. Key components of BO include (1) a surrogate function $g$ that approximates $f$ ; (2) a probabilistic belief $p(f \mid \mathcal{D})$ over the unknown function $f$; and (3) an acquisition function $\alpha : \mathcal{X} \to \mathbb{R}$, which guides where to evaluate $f$. The representational capacity of $g$ determines how closely we can approximate $f$, while the calibration of the posterior distribution $p(g_t \mid \Omega_t)$ at time $t$, given the previous observations $\Omega_t := \{(\mathbf{x}_i, f(\mathbf{x}_i))\}_{i=1}^{t-1}$, dictates the strategy for balancing exploration and exploitation within $\mathcal{X}$ [17]. This balance between exploration and exploitation is crucial for the success of BO in efficiently identifying the optimal $\mathbf{x}_*$ within a reasonable timeframe [18].

**Bayesian Regressors** The *de facto* choice for surrogate $p_t(g_t|\Omega_t)$ in BO is Gaussian Processes [37]. Furthermore, there is extensive literature demonstrating the use of other parametric models, such as Bayesian Neural Networks (BNNs), as surrogates [20]. We consider two different alternatives: Gaussian processes, and a BNN using Laplace Approximation. A Gaussian process can be thought of as a collection of normally distributed random variables that emulate the behavior of $h(\mathbf{x}) + \varepsilon$, where $h \sim \text{GP}(\mu(\mathbf{x}|\theta), K(\mathbf{x}, \mathbf{x}'|\theta))$, $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $\varepsilon \sim N(0, \sigma^2)$. Here, $\mu(\mathbf{x}|\theta)$ denotes the mean function and $K(\mathbf{x}, \mathbf{x}'|\theta)$ denotes the covariance or kernel function [31].

**Laplace Approximation** While standard NNs make point-wise predictions, BNNs provide point-wise predictive distributions, thus measuring the uncertainty of our estimates. Let $f : \mathcal{X} \times \Theta \to \mathcal{Y}$ defined by $(\mathbf{X}, \theta) \mapsto f_\theta(\mathbf{x})$ be a NN. Then, the frequentist point-wise estimate is $\theta_* = \arg\max_{\theta\in\Theta} \log p(\mathcal{D}|\theta) + \log p(\theta)$ [25]. The Laplace Approximation (LA) fits a Gaussian distribution $q(\theta|\mathcal{D}) := \mathcal{N}(\theta_*, \Sigma_*^{-1})$, this centered at the frequentist point-wise estimated $\theta_*$ with covariance given by the inverse of the Hessian $\Sigma_= - \nabla_\theta^2 \log p(\theta|\mathcal{D})|\theta = \theta_*$ [27]. We employ the Linearized LA (LLA) [8], using the Gauss-Newton matrix and the linearization $f_\theta^{\text{lin}}(\mathbf{x}) = f_\theta(\mathbf{x}) - \nabla_\theta f_\theta(\mathbf{x})|_{\theta=\theta*} \cdot (\theta - \theta_*)$ [17].

**Feature Extractors** Molecular representations play a crucial role in the success of BO, as molecules are naturally expressed as 3D graphs. However, the computational costs of generating 3D structures over a large candidate space often lead to a preference for simpler 1D or 2D representations [13]. While molecular fingerprints are efficiently calculated, they may lack structural detail for complex tasks. GNNs with 2D graph representation capture local structures in graph-structured data by passing messages between nodes [9, 11]. Equivariant attention-based GNNs [23], further enhance this by incorporating symmetries of the molecules in 3D space and attention mechanisms [6, 35, 28]. Recently, large language models (LLMs) specifically tailored for chemistry-related applications have gained significant traction, particularly in their ability to extract meaningful features from chemical data [36, 5]. These models are trained on large-scale chemical databases via the molecules' 1D string representations. This makes LLMs useful as molecular feature extractors and has been used for materials discovery [14, 19].

## 3 Experiments

We investigate BO performance using 1D, 2D, and 3D representations of molecules. This section outlines our experimental setup, including datasets, feature extractors, tasks, sample complexities, and evaluation methods. Implementation in `https://github.com/andresguzco/molecular-bayes`.

**Datasets** We use four datasets in our experiments: QM7 [26, 4], QM9 [34, 29], GEOM's MoleculeNet and DRUGS [41, 1]. QM7 includes 7165 molecules with atomization energies ($\Delta$E) in kcal/mol and up to seven heavy atoms (C, N, O, S), while QM9 contains 133 885 molecules with up to nine heavy atoms (C, N, O, F) and 12 properties. The MoleculeNet dataset consists of benchmark datasets designed for molecular machine learning tasks and includes 28 295 molecules, covering tasks like quantum mechanics, physical chemistry, biophysics, and physiology. GEOM provides an enhanced version of this dataset that includes conformers for each example. GEOM also contains the DRUGS dataset, which provides molecular geometries for drug-like molecules, with up to 91 heavy atoms and 317 928 molecules, which are useful for studies involving conformational flexibility and geometric properties. The models were trained on QM9, which was split into a training set and a virtual library serving as the search space. The virtual library and the other datasets were used to evaluate the models with no overlap with training observations.

**Feature Extractors** We implement three types of models: Molformer, an LLM trained on 1D molecular representations [33], an MPNN which inherently leverages 2D molecular information, and an equivariant attention-based GNN. All models serve as feature extractors up to their respective readout layers, encoding molecules into high-dimensional embeddings before making predictions on the target properties. To ensure consistency, the feature extractors are constrained to similar sizes, with each containing approximately 1.5 million parameters. The readout layer consists of two hidden layers, which are optimal for BO with LLA [21]. For the benchmark GP, we utilize the Tanimoto kernel with molecular fingerprints, where the kernel function is defined as $K(\mathbf{x}, \mathbf{x}'|\theta) = \langle \mathbf{x}, \mathbf{x}' \rangle \cdot (||\mathbf{x}||^2 + ||\mathbf{x}'||^2 - \langle \mathbf{x}, \mathbf{x}' \rangle)^{-1}$ [38].

**Tasks** Each model is trained for two tasks: target property prediction and transfer learning. In the target property prediction task, the model has a single readout layer trained to predict a specific property—HOMO-LUMO gap ($\Delta \mathrm{E}_{\mathrm{gap}}$ in eV) from QM9. In transfer learning, the model has $n-1$ readout layers, each trained on different tasks. We aim to assess whether a model trained on one set of properties can still provide accurate predictions on different experimental datasets by fine-tuning only the final layer, evaluating its potential as a foundation model [42].

**Sample Complexity** The models were trained on datasets with varying sample complexities to evaluate their performance based on the number of observations needed to effectively utilize 3D information. Previous research indicates that equivariant models generally require more samples compared to non-equivariant models to achieve similar performance levels [10]. For the feature extractors, we experimented with four different training set sizes: 500, 1000, 10 000, and 50 000 observations. This approach investigated how model performance scales with sample availability.

**Framework** In our experimental setup, each BO loop uses a pre-trained model to make predictions over 1000 iterations. The model is frozen except for the readout layer, which estimates uncertainty in two ways: by training the layer's weights with known data and fitting a linear Laplace approximation, or by using a GP for the embedding readout. The loop begins with an initial set of 10 observations, with the highest expected improvement per observation added to this set at each iteration.

**Evaluation** For each loop, we subsample 10 000 observations from our virtual library or use the entirety of the dataset if it's small enough (QM7), repeating the processes for 15 different seeds and averaging over the results to obtain an unbiased estimator for the performance with its mean and standard error. The models are compared with random search, which uniformly samples from the molecular space, and GPs utilizing a 1D molecular fingerprint representation. In total, 2100 distinct runs were perfromed, providing a comprehensive evaluation across settings.

## 4 Results

**QM7** For QM7, which features relatively simple molecules, highlights the surprising effectiveness of molecular fingerprints. Notably, the 1D GP method performed slightly worse than more complex models. In contrast, GP regression with binary encoded SMILES, serving as a baseline, demonstrated that even simple 1D representations can capture sufficient information to remain competitive. Although 2D models outperformed 3D models overall—particularly when combined with GP regressors—the performance gap was modest. While the 3D models saw slight improvements when paired with LLA, the gains were limited, suggesting that higher-dimensional representations may not be critical for simpler molecular structures like those in QM7. This is evident in Fig. 2, where the top-performing models for simpler tasks did not heavily rely on 3D data. The results of the LLM were striking, as it outperformed all other models by a significant margin. Its ability to leverage contextual information and generate accurate predictions, even for relatively simple molecules in the QM7 dataset, demonstrated its superior generalization capabilities.

**QM9** In contrast, the QM9 dataset, which features slightly more complex molecules, underscores the limitations of 1D representations. Here, 2D MPNNs achieve the highest performance, and consistently outperform 1D GP and RS methods, while 3D models only outperform RS. The differences become even more pronounced with increasing molecular complexity as described by size. While 2D models continue to demonstrate strong performance and stability, the 3D GNNs, particularly when enhanced with LLA, begin to close the gap, indicating that the extra structural information provided by 3D representations becomes more important as molecular complexity increases. Despite the overall advantage of the 2D models, the smaller margins between 2D and 3D performance suggest that for highly complex molecules, further optimization of 3D models may yield competitive results, as seen in the bottom row of Fig. 2. Contrary to all other datasets, LLMs performed worse than 2D and 3D models. This task may have been the most dependent on information not captured by 2D and 3D representations the specific, which could explain why it performed worse.

**MoleculeNet** As shown in Fig. 2, 2D models consistently outperformed 3D models across a wide range of tasks, which suggests that 2D representations efficiently capture the necessary structural information for accurate predictions without the computational overhead of 3D models. The slight improvements observed with 3D models when using techniques like LLA are insufficient to justify their use, as the performance gains are marginal and insufficient. The consistently strong performance of 2D models raises important questions about the value of incorporating 3D information. Even as molecular size increases, 3D models fail to offer significant advantages, and in many cases, they underperform compared to 2D approaches. Additionally, the competitive performance of 1D models, such as GPs with SMILES encoding, highlights the efficiency of simpler representations in certain scenarios. Although 1D models struggle with larger datasets and more complex molecular structures, their ability to remain competitive in simpler tasks emphasizes that higher-dimensional representations are not always necessary. The LLM, as with the QM7 and DRUGS results, outperformed all other models. Its superior performance across both simple and complex molecular datasets highlights its ability to generalize effectively, surpassing the limitations of both 2D and 3D models. This reinforces the trend observed before, where the LLM demonstrated remarkable versatility and accuracy.

**GEOM DRUGS** The DRUGS dataset emphasizes the importance of higher-dimensional features. However, despite the increased molecular size, both 1D and 2D representations manage to capture sufficient information to perform competitively. As shown in Fig. 2, LLMs and 2D models consistently outperform 3D models across most tasks, even for large drug-like molecules. Interestingly, simple

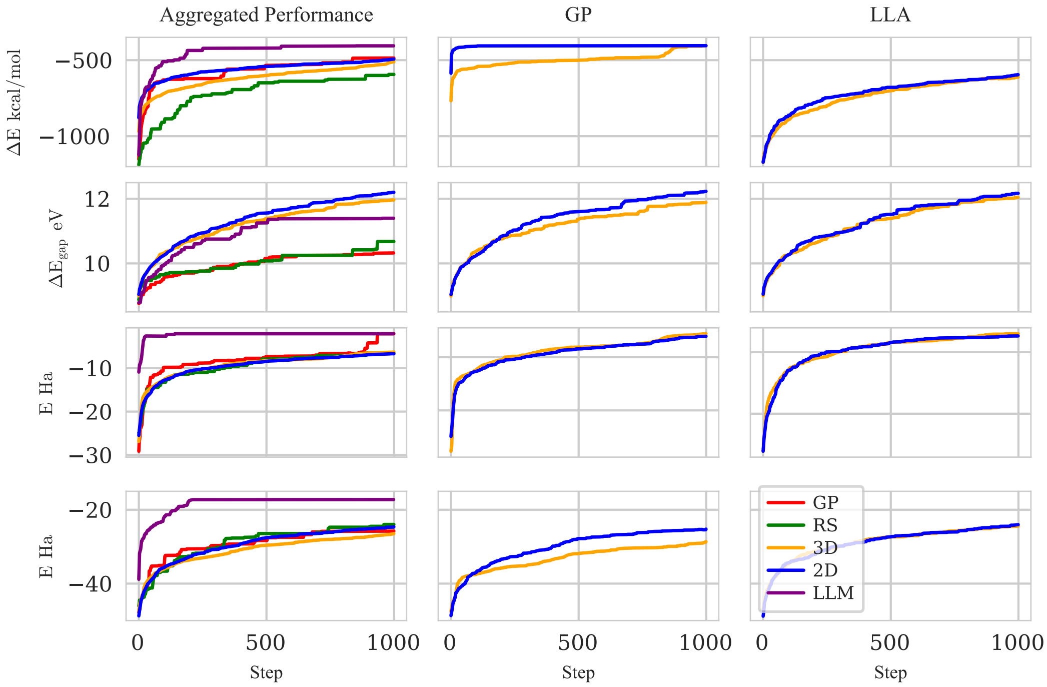

Figure 2: **Experimental Results**. Top row: QM7. Bottom row: QM9.

GP regression and random search performed similarly, suggesting that the models used may not be sufficiently complex to outperform these benchmark methods. This indicates that larger models or further optimization would be necessary to see significant improvements beyond the baseline methods. The 3D models, while more suited for capturing subtle geometric features, do not provide a substantial performance increase, reinforcing that for property optimization in such datasets, 3D features may not be necessary unless ultra-high precision is required. The LLM achieved the most substantial performance gap so far, outperforming all other models by a wide margin. This is particularly remarkable given the complexity and size of the molecules in this dataset. Despite the strong performance of 2D models, the LLM's ability to handle intricate molecular details allowed it to excel far beyond both 2D and 3D representations. This suggests that the LLM's contextual understanding is especially beneficial for larger, more complex molecular structures.

As molecular size increases, 1D models and 2D representations capture enough information to perform effectively, rendering 3D features generally unnecessary.

## 5 Conclusion

1D representations consistently outperformed both 2D and 3D models across all datasets, providing a stable and efficient solution for molecular property prediction, particularly in complex datasets like GEOM DRUGS, where LLMs showed a significant advantage even with large, intricate molecules. Notably, 2D representations also outperformed 3D models across many tasks, indicating that lower-dimensional models can achieve strong predictive accuracy with an optimal balance between efficiency and computational cost. This trend persisted even in datasets traditionally suited to 3D models, where 3D information offered only minor improvements at a greater computational expense. Future research should focus on tasks where 3D data might have a greater impact, such as protein docking, while exploring how BO performance scales with model size and complexity to optimize these methods. Additionally, investigating graph foundation models could open new paths for molecular data representation by combining the strengths of graph-based and foundational approaches.

## References

[1] Simon Axelrod and Rafael Gómez-Bombarelli. Geom, Energy-annotated Molecular Conformations for Property Prediction and Molecular Generation. *Scientific Data*, 9(1), April 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01288-4.

[2] M Balandat, B Karrer, D Jiang, S Daulton, B Letham, A.G. Wilson, and E Bakshy. Botorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[3] Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. In *NeurIPS*, 2022.

[4] L. C. Blum and J.-L. Reymond. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.*, 131:8732, 2009.

[5] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. In *NeurIPS*, 2020.

[6] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *ICLR*, 2018.

[7] Jordan E. Crivelli-Decker, Zane Beckwith, Gary Tom, Ly Le, Sheenam Khuttan, Romelia Salomon-Ferrer, Jackson Beall, Rafael Gómez-Bombarelli, and Andrea Bortolato. Machine Learning Guided AQFEP: A Fast and Efficient Absolute Free Energy Perturbation Solution for Virtual Screening. *Journal of Chemical Theory and Computation*, August 2024. ISSN 1549-9626. doi: 10.1021/acs.jctc.4c00399.

[8] Erik A. Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace Redux - Effortless Bayesian Deep Learning. In *NeurIPS*, 2021.

[9] David K. Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P. Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *NeurIPS*, 2015.

[10] Bryn Elesedy and Sheheryar Zaidi. Provably Strict Generalisation Benefit for Equivariant Models. In *ICML*, 2021.

[11] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. In *ICML*, 2017.

[12] Ryan-Rhys Griffiths, Leo Klarner, Henry Moss, Aditya Ravuri, Sang Truong, Samuel Stanton, Gary Tom, Bojana Rankovic, Yuanqi Du, Arian Jamasb, Aryan Deshwal, Julius Schwartz, Austin Tripp, Gregory Kell, Simon Frieder, Anthony Bourached, Alex J. Chan, Jacob Moss, Chengzhi Guo, Johannes Durholt, Saudamini Chaurasia, Ji Won Park, Felix Strieth-Kalthoff, Alpha A. Lee, Bingqing Cheng, Alán Aspuru-Guzik, Philippe Schwaller, and Jian Tang. Gauche: a Library for Gaussian Processes in Chemistry. In *NeurIPS*, 2024.

[13] Lingshen He, Yuxuan Chen, zhengyang shen, Yiming Dong, Yisen Wang, and Zhouchen Lin. Efficient Equivariant Network. In *NeurIPS*, 2021.

[14] Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D. Bocarsly, Andres M. Bran, Stefan Bringuier, L. Catherine Brinson, Kamal Choudhary, Defne Circi, Sam Cox, Wibe A. de Jong, Matthew L. Evans, Nicolas Gastellu, Jerome Genzling, María Victoria Gil, Ankur K. Gupta, Zhi Hong, Alishba Imran, Sabine Kruschwitz, Anne Labarre, Jakub Lála, Tao Liu, Steven Ma, Sauradeep Majumdar, Garrett W. Merz, Nicolas Moitessier, Elias Moubarak, Beatriz Mouriño, Brenden Pelkie, Michael Pieler, Mayk Caldas Ramos, Bojana Ranković, Samuel G. Rodriques, Jacob N. Sanders, Philippe Schwaller, Marcus Schwarting, Jiale Shi, Berend Smit, Ben E. Smith, Joren Van Herck, Christoph Völker, Logan Ward, Sean Warren, Benjamin Weiser, Sylvester Zhang, Xiaoqi Zhang, Ghezal Ahmad Zia, Aristana Scourtas, K. J. Schmidt, Ian Foster, Andrew D. White, and Ben Blaiszik. 14 Examples of How LLMs Can Transform Materials Science and Chemistry: a Reflection on a Large Language Model Hackathon. *Digital Discovery*, 2(5):1233–1250, 2023. ISSN 2635-098X. doi: 10.1039/d3dd00113j.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2014.

[16] 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 *AISTATS*, pages 3393–3403, 2020.

[17] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Learnable Uncertainty Under Laplace Approximations. In *Conference on Uncertainty in Artificial Intelligence*, 2021.

[18] Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, and Vincent Fortuin. Promises and Pitfalls of the Linearized Laplace in Bayesian Optimization. In *Symposium on Advances in Approximate Bayesian Inference*, 2023.

[19] Agustinus Kristiadi, Felix Strieth-Kalthoff, Marta Skreta, Pascal Poupart, Alán Aspuru-Guzik, and Geoff Pleiss. A Sober Look at LLMs for Material Discovery: Are They Actually Good for Bayesian Optimization Over Molecules? In *ICML*, 2024.

[20] George Lamb and Brooks Paige. Bayesian Graph Neural Networks for Molecular Property Prediction. *arXiv*, abs/2012.02089, 2020.

[21] Yucen Lily Li, Tim G. J. Rudner, and Andrew Gordon Wilson. A Study of Bayesian Neural Network Surrogates for Bayesian Optimization. In *ICLR*, 2024.

[22] Yi-Lun Liao and Tess E. Smidt. Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs. In *ICLR*, 2023.

[23] Yi-Lun Liao, Brandon M. Wood, Abhishek Das, and Tess E. Smidt. Equiformerv2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *ICLR*, 2024.

[24] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic Gradient Descent with Warm Restarts. *ICLR*, 2016.

[25] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press, 2003.

[26] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole von Lilienfeld. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New Journal of Physics*, 15(9):095003, 2013.

[27] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[28] Saro Passaro and C. Lawrence Zitnick. Reducing SO(3) Convolutions to SO(2) for Efficient Equivariant GNNs. In *ICML*, 2023.

[29] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Scientific Data*, 1, 2014.

[30] Bojana Ranković, Ryan-Rhys Griffiths, Henry B. Moss, and Philippe Schwaller. Bayesian Optimisation for Additive Screening and Yield Improvements – Beyond One-Hot Encoding. *Digital Discovery*, 3(4):654–666, 2024. ISSN 2635-098X. doi: 10.1039/d3dd00096f.

[31] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, November 2005. ISBN 9780262256834.

[32] Hippolyt Ritter, Aleksandar Botev, and David Barber. A Scalable Laplace Approximation for Neural Networks. In *ICLR*, 2018.

[33] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale Chemical Language Representations Capture Molecular Structure and Properties. *Nature Machine Intelligence*, 4(12), 2022.

[34] 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. doi: 10.1021/ci300415d.

[35] Kristof Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant Message Passing for the Prediction of Tensorial Properties and Molecular Spectra. In *ICML*, 2021.

[36] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583, August 2019. ISSN 2374-7951. doi: 10.1021/acscentsci.9b00576.

[37] 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. *Proc. IEEE*, 104(1):148–175, 2016. doi: 10.1109/JPROC.2015.2494218.

[38] Austin Tripp, Sergio Bacallado, Sukriti Singh, and José Miguel Hernández-Lobato. Tanimoto Random Features for Scalable Molecular Machine Learning. In *NeurIPS*, 2023.

[39] Tanja van Mourik, Michael Bühl, and Marie-Pierre Gaigeot. Density Functional Theory across Chemistry, Physics and Biology. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2011):20120488, March 2014. ISSN 1471-2962. doi: 10.1098/rsta.2012.0488.

[40] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's Transformers: State-of-the-art Natural Language Processing, 2020. URL https://arxiv.org/abs/1910.03771.

[41] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay S. Pande. Moleculenet: a Benchmark for Molecular Machine Learning. *Chemical Science*, 9:513 – 530, 2017.

[42] Shaolun Yao, Jie Song, Lingxiang Jia, Lechao Cheng, Zipeng Zhong, Mingli Song, and Zunlei Feng. Fast and Effective Molecular Property Prediction with Transferability Map. *Communications Chemistry*, 7(1), April 2024. ISSN 2399-3669. doi: 10.1038/s42004-024-01169-4.

# A Appendix A: Sample Complexity and Transfer Learning

## A.1 How many samples does each dimension need?

As illustrated in Fig. 3, 3D models consistently required a larger number of training samples to outperform or even match the performance of 2D models, particularly for simpler datasets such as QM7. In these lower-complexity tasks, the computational overhead introduced by 3D features did not translate into closes the performance gap until the sample size exceeded 10,000 observations. For example, while 3D models did show some improvement with more samples, their performance remained inferior to that of 2D models with smaller datasets.

In contrast, the 2D models were highly data-efficient across all datasets, capturing essential structural information with relatively few samples. Even with a modest dataset of 500 to 1,000 observations, 2D models achieved competitive performance, suggesting that the information content provided by 2D representations is generally sufficient for many molecular property prediction tasks. The gap between the performance of 2D and 3D models was most pronounced in smaller datasets, where 3D models often struggled to justify their computational expense. This finding aligns with earlier research [10], which highlights the difficulty of leveraging 3D information in data-scarce regimes.

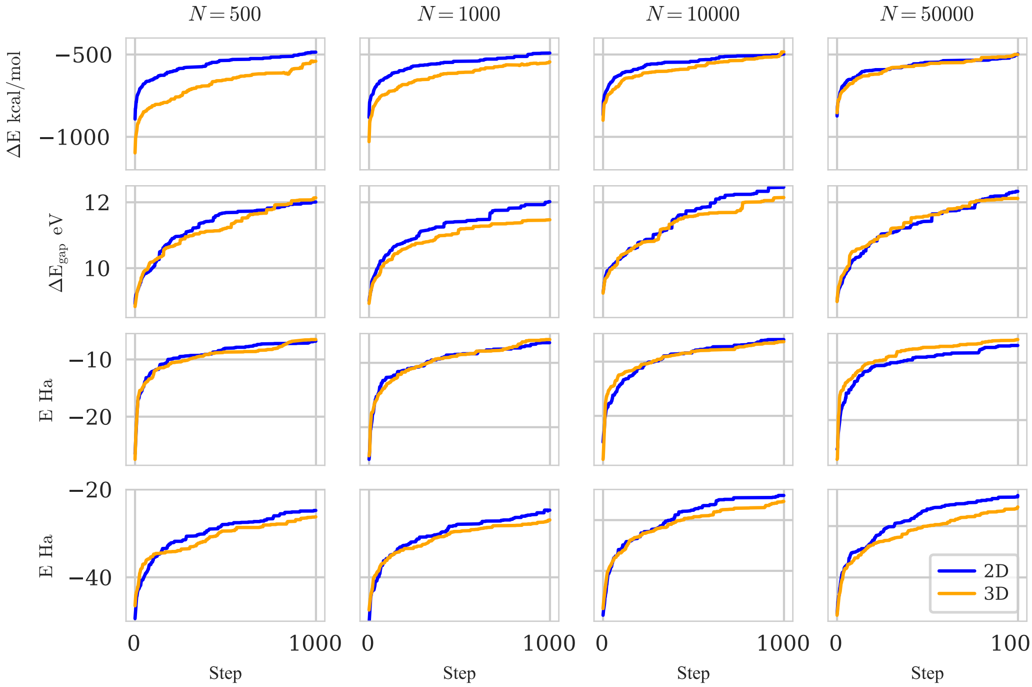

Figure 3: **Experimental Results per Sample Complexity**. Top row: QM7. Bottom row: QM9.

## A.2 Is transfer learning beneficial?

The results comparing 2D and 3D models across single-property prediction and transfer learning tasks, as shown in Fig. 4, reveal key differences in their effectiveness. The LLM used was trained in multiple tasks, thus offering only a transfer learning perspective. In single-property tasks, 2D models consistently outperform 3D models, particularly in datasets with limited data, like QM7 and QM9. This suggests that 2D representations capture essential structural information efficiently, without the computational cost of 3D models. Even in more complex datasets like GEOM DRUGS, where the performance gap between 2D and 3D models narrows, 2D models remain more competitive and effective for property prediction, offering a balance of simplicity and accuracy. However, in transfer learning—where models trained on one molecular property are fine-tuned to predict another—3D models show some improvement but still lag behind 2D models. The additional geometric detail provided by 3D representations enhances generalization across tasks but is not enough to outperform 2D models in terms of efficiency and accuracy.

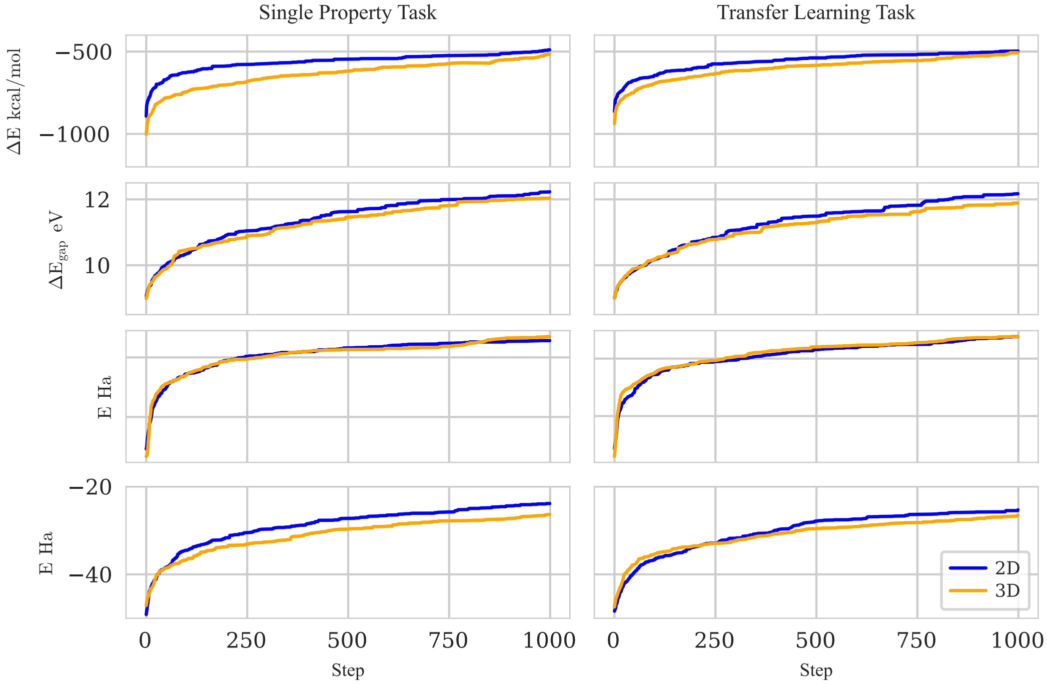

Figure 4: **Experimental Results per Task**. Top row: QM7. Bottom row: QM9.

# B Appendix B: Algorithms and Hyperparameters

## B.1 Pseudocodes

We present the pseudocode of the Bayesian Optimization (BO) loop and Section 5 in Algorithm 1.

**Algorithm 1** Using an NN as a feature extractor in BO.

```
Require: Pre-trained feature extractor $\phi_{W^*}$, mapping a molecular representation $c(x)$ to its embedding vector $h \in \mathbb{R}^H$; surrogate model $g_\theta : \mathbb{R}^H \to \mathbb{R}$; candidate molecules $D_{\text{cand}} = \{x_i\}_{i=1}^n$; initial dataset $D_1 = \{(x_i, f(x_i))\}_{i=1}^m$; time budget $T$.
for $t = 1, \ldots, T$ do
    $\Phi_t = \{(\phi_{W^*}(c(x)), f(x)) : (x, f(x)) \in D_t\}$
    $p(g_t|D_t) = \text{infer}(g_\theta, \Phi_t)$
    $x_t = \arg\max_{x \in D_{\text{cand}}} \alpha(p(g_t(c(x))|D_t))$
    $D_{t+1} = D_t \cup \{(x_t, f(x_t))\}$
    $D_{\text{cand}} = D_{\text{cand}} \setminus \{x_t\}$
end for
return $\arg\max_{(x,f(x)) \in D_{T+1}} f(x)$
```

## B.2 Training

### B.2.1 Fixed-Feature Surrogates

The following are the training details of the surrogates we used in Section 5. We used HuggingFace's transformers library [40] for MolFormer. For GPs, we use BoTorch [2] to construct the surrogate function. The Tanimoto kernel is taken from Gauche [12]. To optimize the marginal likelihood, we use Adam [15] with a learning rate of 0.01 for 500 epochs. We constrain the GNNs to $\tilde{1}$.5 million parameters, and further train Equifrormer v2 on noisy nodes. We optimize the GNNs with Adam with a learning rate of $1 \times 10^{-4}$ and weight decay of $5 \times 10^{-4}$ until convergence with early stopping at 20 epochs without improvement with a batch size of 64. We anneal the learning rate with the cosine annealing scheme [24]. On the other hand for LLA, our implementation is based on the laplace-bayesopt package. The neural net used is a 2-hidden-layer multilayer perceptron with 50 hidden units on each layer along with the ReLU activation function. The Laplace approximation is done post-hoc, and we tune the prior precision with the marginal likelihood for 100 iterations. The Hessian is approximated with a Kronecker structure [32]. At last, we show the parameters used to train Equiformer v2 [23], the 3D GNN, in Table 1.

Table 1: Hyper-parameters for the EquiformerV2 Model.

| **Hyper-parameters** | **Value or description** |
|---|---|
| Optimizer | AdamW |
| Learning rate scheduling | Cosine learning rate with linear warmup |
| Warmup epochs | 5 |
| Maximum learning rate | $5 \times 10^{-4}$ |
| Batch size | 64 |
| Number of epochs | 300 |
| Weight decay | $5 \times 10^{-3}$ |
| Dropout rate | 0.1 |
| Drop path rate | 0.05 |
| Project drop rate | 0.0 |
| Cutoff radius (Å) | 5.0 |
| Maximum number of neighbors | 500 |
| Maximum atomic number | 90 |
| Number of radial bases | 64 |
| Dimension of hidden scalar features in radial functions | 64 |
| Maximum degree $L_{\text{max}}$ | [3] |
| Maximum order $M_{\text{max}}$ | [3] |
| Number of Transformer blocks | 2 |
| Embedding dimension | 16 |
| Attention hidden dimension | 64 |
| Number of attention heads $h$ | 4 |
| Attention alpha channels | 32 |
| Attention value channels | 16 |
| Hidden dimension in feed forward networks | 128 |
| Resolution of point samples $R$ | 64 |
| Distance function | Gaussian |
| Attention activation | scaled silu |
| Attention renormalization | True |
| Noise standard deviation $\sigma_{\text{denoise}}$ | 0.02 |
| Denoising coefficient $\lambda_{\text{denoise}}$ | 0.1 |
| Denoising probability $p_{\text{denoise}}$ | 0.5 |
| Corrupt ratio $r_{\text{denoise}}$ | 0.25 |

### B.3 Prompting

Following the framework in Kristiadi et al. [19], we used the prompt "The estimated {objective str} of the molecule {smiles str} is:" in our experiments. The variable `smiles_str` equals the SMILES representation of the molecule at hand, e.g., "OS(=O)(=O)O" for sulfuric acid. The variable `obj_str` has the value of the textual description of the problem at hand: "HOMO-LUMO gap in eV" for QM9, " atomization energy in kcal/mol " for QM7, "total energy in Hartees" for GEOM's Molecule Net and DRUGS.