Active Causal Machine Learning for Molecular Property Prediction

Zachary R Fox
Computational Sciences and Engineering Division
Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37831, USA
foxzr@ornl.gov

Ayana Ghosh
Computational Sciences and Engineering Division
Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37831, USA
ghosha@ornl.gov

Abstract

Predicting properties from molecular structures is paramount to design tasks in medicine, materials science, and environmental management. However, design rules derived from the structure-property relationships using correlative data-driven methods fail to elucidate underlying causal mechanisms controlling chemical phenomena. This preliminary work proposes a workflow to actively learn robust cause-effect relations between structural features and molecular property for a broad chemical space utilizing smaller subsets, entailing partial information.

1 Introduction

Understanding structure-property relationships within broad chemical space is essential for chemical discovery. The growth in public repositories (e.g., PubChem [1], ZINC [2], ChEMBL [3], QM9 [4, 5] ANI-1x,[6] and QM7-X [7]) containing structural and physiochemical properties (computed with quantum mechanical calculations or observed with experiments) on thousands to millions of molecules along with applications of machine learning/deep learning (ML/DL) workflows have skyrocketed in the recent years. Such developments have paved the path forward for modeling molecular interactions, chemical bonding, reaction energy pathways, docking, inverse design of molecules for targets, synthesis, gaining novel insights into mechanisms with numerous applications such as drug discovery, [8, 9, 10, 11] antibiotics [12], catalysts [13, 14], photovoltaics [15], organic electronics [16], and redox-flow batteries [17]. The quantitative structure-activity/property relationships (QSAR/QSPR)-type models [11, 18, 8] and more recently the generative models have largely contributed to the *in silico* molecular design efforts.

Modern molecular generative models have transformed standard string representations of molecules towards embedded spaces [10] with information on the entire molecular scaffold. However, the latent embeddings of most generative models are neither smooth nor carry ton of useful information, limiting their utility in direct gradient-based optimization methods for targeted design. The standard Gaussian processes (GPs) within Bayesian optimization (BO) methods, as combined with generative models for finding optimized solutions, fail to incorporate any prior information of physical or chemical behavior of the system in the process. Recent work lead by Ghosh at al. [19] has shown how a physics-augmented GP within a hypothesis-driven active learning workflow can be employed to reconstruct functional behavior over an unknown chemical space (for which prior data may not be available). However, most of these efforts still rely on the *in-built correlative relationships* acting behind the molecular representations and property, by drawing inferences purely from statistical dependencies.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

The fundamental *cause-effect relations* are practically missing in these purely data-driven approaches except a few recent studies [20, 21, 22, 23] in the domain of materials science.

In this study, we demonstrate how an active learning workflow, informed with causal discovery models, can successfully learn structure-property relationship from subsets of data, sampled from any part of the chemical space of interest, compare predictive accuracy, all combined together to actively learn the relations for the entire dataset. The target property is dipole moment for the QM9 dataset. For simplicity, we utilize easily-computable molecular features to represent each molecule. It is important to note that the choice of subsets is more or less arbitrary, meaning the information we start with is partial, catering to the adaptive needs for real-time workflows for AI-guided design, automated synthesis, automated characterization while deriving fundamental understandings behind a molecular property.

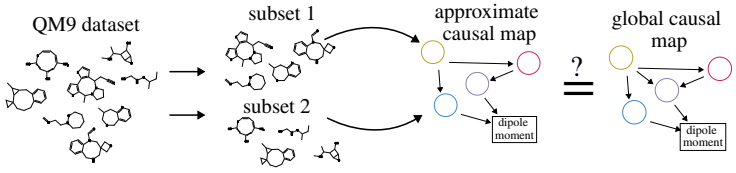

Figure 1: Overview of workflow

2 Results & Discussion

We first demonstrate that the prediction of dipole moment and associated causal relations exhibit large variability across different regions of chemical space. The models trained to predict in one region of chemical space do not generalize well to other regions as represented in Fig. A2. Therefore, causal relations derived from one region may not be robust enough to capture structure-functionality relations for another subset or even the entire dataset. To address this challenge, we introduce an innovative active learning approach designed to recover comprehensive causal relationships using a minimal dataset, as represented in Fig. 1.

2.1 Generating molecular data subsets

We begin our analysis by characterizing each molecule within the QM9 dataset [4, 5] through a vector consisting of twenty descriptors, computed using RDKit. The atomistic mechanisms behind dipole moment leading to polarization may depend on several factors such as electronegativity, bonding, presence or absence of specific functional groups etc. We have summarized some of these mechanisms (as noted in the Introduction section) in our related works investigated by in-depth first-principles computations. The features computed by RDkit as considered in this study accounts for these type of factors at a rudimentary level, computed using the SMILE representations which is the motivation behind choosing these specific features. Instead of employing fingerprint or latent representations of molecules, we choose to work directly with these molecular features, enabling the use of straightforward causal approaches. Furthermore, different regions of the molecular feature space contribute to the creation of distinct causal maps and predictive models. We have used a Gaussian Mixture Model to create three subsets, $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ of the QM9 dataset by clustering based on three pivotal features: `MolLogP` (lipophilicity), `TPSA` (topological polar surface area), and `MolMR` (molar refractivity) as shown in Fig. 2. After clustering, each data subset retained twenty chemical descriptors.

2.2 Feature selection based on predicting polarizability

We use polarizability as an intermediate target to down-select features for subsequent causal analyses and predictions of dipole moments, thereby capturing the control of an intrinsic property by extrinsic influences. Using the `LinGAM` causal discovery framework [24], we pick the top $k \leq 20$ features within the structure equation model[25]. This approach is valid when each feature has its own additive non-Gaussian noise term ϵ_i, and constructs a model with linear relationships between each variable. For each subset of the data $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$, we have used `LinGAM` to construct a weighted directed acyclic

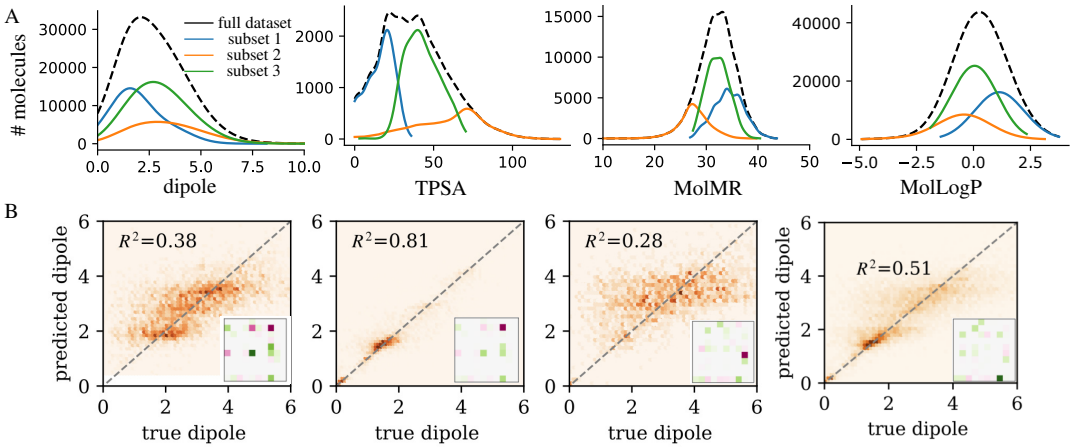

Figure 2: Causal discovery and property prediction in different regions of chemical space. (A) Feature distributions for different subsets. (B) Test R^2 and parity plots for a random forest model trained on $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$ and the full dataset (left to right). Insets show the corresponding adjacency matrix of the nine downselected features, where green colors signify positive causal relations and pink/purple signify negative relations. Full causal graphs are given in the Appendix.

graph (DAG), denoted by $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$. In each graph, the target variable (dipole moment) is a sink, i.e. it does not have any downstream variables. All 20 features are ranked by the strength of their relationships, and the top k are selected. For the numerical experiments below, we set $k = 9$. This causal analysis is performed for the full dataset, and the same 9 features are used for each data subset.

In each data regime, the prediction accuracy of dipole moment using a random forest model over the $k = 9$ features is significantly different, shown in Fig. 2(B). Furthermore, we find that each data subset results in distinct causal relationships (Fig. 2B and A3-A6) between the features. Based on these results, we next investigate how one can construct a minimal dataset that accurately builds causal maps representative of the full dataset.

2.3 Causally-informed active learning to build minimal molecular datasets

Active learning aims to optimize the training process by selecting the most informative data points for labeling, rather than relying on random or pre-defined data sampling. In this context, the goal is to reduce the annotation cost and resource requirements while improving model performance. Traditional active learning approaches choose sampling data by evaluating the uncertainty in a predicted value [26] or through constructing a dataset which samples the entire input space [27, 28]. Here, we build an active learning algorithm, detailed in Algorithm 2.3, to reconstruct a global causal map from a minimal dataset. A global causal model may be constructed from existing knowledge about how different molecular features contribute to a target property, such as dipole moment. The aim of this active learning algorithm is to reconstruct the global causal map, denoted by $\mathcal{G}_\rho$, from a minimal set of data. We note that the goal of this algorithm is distinct from traditional active learning; rather than building a minimal dataset to predict a property of interest, we aim to build a minimal dataset which recapitulates causal structure/property relationships. The algorithm uses a graph distance metric, $\mathcal{L}(\mathcal{G}_1, \mathcal{G}_2)$ to compare the global DAG, $\mathcal{G}_\rho$ to the DAG $\mathcal{G}_{\text{AL}}$ describing the actively learned dataset, $\mathcal{D}_{\text{AL}}$. $\mathcal{G}_{\text{AL}}$ is found using the `LinGAM` causal discovery framework [24]. During each iteration of the active learning scheme, we sample M points, uniformly, from each of the k data subsets described above, denoting the candidate dataset $\tilde{\mathcal{D}}^k_{\text{AL}}$. For each candidate dataset, we construct a causal graph, denoted $\tilde{\mathcal{G}}_k$, and compare it to the global graph $\mathcal{G}_\rho$ using the graph metric $\mathcal{L}(\mathcal{G}_1, \mathcal{G}_2)$. The graph loss function used is the adjacency spectral distance [29]:

$$\mathcal{L}(\mathcal{G}_1, \mathcal{G}_2) = \sqrt{\sum_{i=1}^{N} \left(\lambda_i^{A_1} - \lambda_i^{A_2}\right)^2}, \tag{1}$$

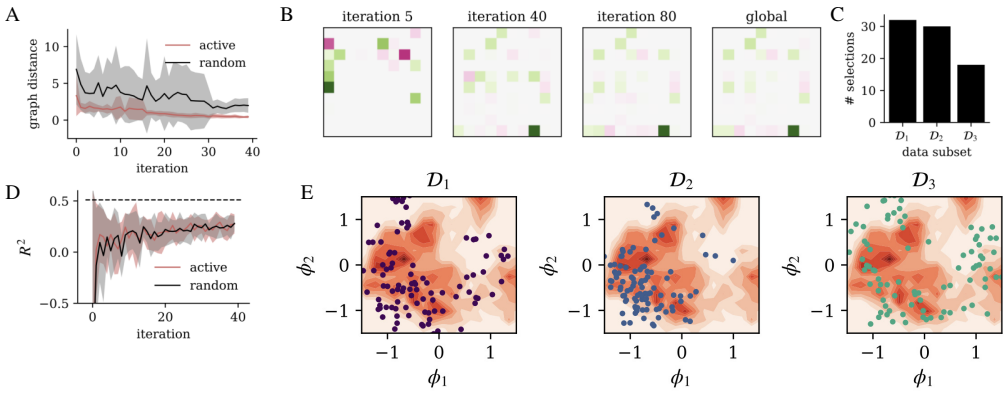

Figure 3: Active learning to recover causal relations. (A) Average and one standard deviation of the graph distance (upper) between the global graph, $\mathcal{G}_\rho$ and the graph of the candidate data set $\mathcal{G}_{\text{AL}}$ at each iteration of the active learning algorithm (red) and for randomly selected data (black). (B) Visualization of the adjacency matrices corresponding to $\mathcal{G}_{\text{AL}}$ at different iterations and the global DAG $\mathcal{G}_\rho$. (C) The number of times each data subset was selected during the active learning procedure. (D) R^2 on test data throughout the active learning experiment. The dashed line corresponds to the value when all data is used. (E) Densities of the ECFPs for the entire dataset projected onto its first two principle components and samples from each data subset (scatter plots).

which computes the ℓ_2 norm of the top N eigenvalues of adjacency matrix $\mathcal{A}$ for each graph.

Algorithm 1: Causally-informed active learning algorithm

$\mathcal{G}_\rho$, global causal graph ;
$\mathcal{D}_{\text{AL}} = \emptyset$;
N_s, number of data subsets ;
N_{iter}, number of iterations ;
while $n < N_{\text{iter}}$ **do**
  for $k \in (1, N_s)$ **do**
    $\tilde{\mathcal{D}}^k_{\text{AL}} = \mathcal{D}_{\text{AL}} \cup$ sample($\mathcal{D}_k$,M);
    $\tilde{\mathcal{G}}_k = \texttt{LinGAM}(\tilde{\mathcal{D}}_{\text{AL}})$;
    $s_k = \mathcal{L}(\tilde{\mathcal{G}}_k, \mathcal{G}_\rho)$
  $k^* = \arg\min(s_k)$;
  $\mathcal{D}_{\text{AL}} = \tilde{\mathcal{D}}^{k^*}_{\text{AL}}$

The dataset with the minimal graph distance to $\mathcal{G}_\rho$, denoted by $\tilde{\mathcal{D}}^*_{\text{AL}}$, is selected, and the algorithm continues. We compare the graph distance of the selected datasets with those chosen from random subsets in Fig. 3A. The actively generated dataset not only converges to the global graph more quickly than the random data (see Fig. 3A-B), but also does so with less noise, as demonstrated by the shaded regions in Fig. 3A. The shaded regions correspond to the mean $\pm$ one standard deviation over ten realizations of the algorithm. Interestingly, the test R^2 of the random forest model performs equally well on either the active or random dataset (Fig. 3(D)), indicating that optimizing for causal structure neither helps nor harms the regression accuracy. The extended connectivity fingerprints (ECFPs) [30] over the entire dataset are projected onto their first two principal components, denoted ϕ_1 and ϕ_2, to demonstrate that the space explored does not exactly fit into the typical diversity-uncertainty sampling paradigms (Figure 3C, E).

3 Summary

In summary, we have developed a causal active learning workflow for iterative identification of causal relations with corresponding prediction of dipole moment for a broad chemical space, from subsets of chemically diverse molecules. The actively-learnt causal relations also pertain to our chemical

understandings. For e.g., molecules containing NH, OH bonds are highly polar in nature (i.e., a bond dipole) whereas presence of more valence electrons will screen the long-range order, resulting in reduction of dipole moment. Tuneable features accounting for molecular weight, atomic charges, number of electrons, electronegativity, presence of NH, OH bonds, have the highest coefficients of cause-effect relations towards dipole moment which can be intervened to optimize the target.

This approach allows for two significant advances in enabling AI-guided design, synthesis and characterization of molecules. One lies in its potential to guide autonomous experiments where partial information from past measurements may be available based on which the model can adaptively learn causal relations for targeted molecular design and synthesis. More importantly, it is even effective to identify intrinsic and extrinsic features in real-time, providing scientists to derive understandings of the underlying mechansims controlling physical/chemical phenomena.

Acknowledgments and Disclosure of Funding

This research was sponsored by the Laboratory Directed Research and Development Program (AI Initiative and SEED program) of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U. S. Department of Energy. ORNL is managed by UT-Battelle, LLC, for DOE under Contract No. DE-AC05-00OR22725. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

[1] Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. PubChem: a public information system for analyzing bioactivities of small molecules.

[2] John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073, December 2020. Publisher: American Chemical Society.

[3] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(Database issue):D1100–D1107, January 2012.

[4] 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, November 2012. Publisher: American Chemical Society.

[5] Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *The Journal of Machine Learning Research*, 14(1):111–152, 2013.

[6] Justin S. Smith, Roman Zubatyuk, Benjamin Nebgen, Nicholas Lubbers, Kipton Barros, Adrian E. Roitberg, Olexandr Isayev, and Sergei Tretiak. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific Data*, 7(1):134, May 2020. Number: 1 Publisher: Nature Publishing Group.

[7] Johannes Hoja, Leonardo Medrano Sandonas, Brian G. Ernst, Alvaro Vazquez-Mayagoitia, Robert A. DiStasio Jr., and Alexandre Tkatchenko. QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Scientific Data*, 8(1):43, February 2021. Number: 1 Publisher: Nature Publishing Group.

[8] Paula Carracedo-Reboredo, Jose Liñares-Blanco, Nereida Rodríguez-Fernández, Francisco Cedrón, Francisco J. Novoa, Adrian Carballal, Victor Maojo, Alejandro Pazos, and Carlos Fernandez-Lozano. A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal*, 19:4538–4558, January 2021.

[9] J. Phillip Kennedy, Lyndsey Williams, Thomas M. Bridges, R. Nathan Daniels, David Weaver, and Craig W. Lindsley. Application of Combinatorial Chemistry Science on Modern Drug Discovery. *Journal of Combinatorial Chemistry*, 10(3):345–354, May 2008. Publisher: American Chemical Society.

[10] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12(1):56, September 2020.

[11] Eugene N. Muratov, Jürgen Bajorath, Robert P. Sheridan, Igor V. Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I. Oprea, Igor I. Baskin, Alexandre Varnek, Adrian Roitberg, Olexandr Isayev, Stefano Curtalolo, Denis Fourches, Yoram Cohen, Alan Aspuru-Guzik, David A. Winkler, Dimitris Agrafiotis, Artem Cherkasov, and Alexander Tropsha. QSAR without borders. *Chemical Society Reviews*, 49(11):3525–3564, June 2020. Publisher: The Royal Society of Chemistry.

[12] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A Deep Learning Approach to Antibiotic Discovery. *Cell*, 180(4):688–702.e13, February 2020.

[13] Takashi Toyao, Zen Maeno, Satoru Takakusagi, Takashi Kamachi, Ichigaku Takigawa, and Ken-ichi Shimizu. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catalysis*, 10(3):2260–2297, February 2020. Publisher: American Chemical Society.

[14] Wenhong Yang, Timothy Tizhe Fidelis, and Wen-Hua Sun. Machine Learning in Catalysis, From Proposal to Practicing. *ACS Omega*, 5(1):83–88, January 2020. Publisher: American Chemical Society.

[15] Wenbo Sun, Yujie Zheng, Ke Yang, Qi Zhang, Akeel A. Shah, Zhou Wu, Yuyang Sun, Liang Feng, Dongyang Chen, Zeyun Xiao, Shirong Lu, Yong Li, and Kuan Sun. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science Advances*, 5(11):eaay4275, November 2019. Publisher: American Association for the Advancement of Science.

[16] R. et al. Gómez-Bombarelli. Design of efficient molecular organic light- emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials*, 15, 2016.

[17] Süleyman Er, Changwon Suh, Michael P. Marshak, and Alán Aspuru-Guzik. Computational design of molecules for an all-quinone redox flow battery. *Chemical Science*, 6(2):885–893, January 2015. Publisher: The Royal Society of Chemistry.

[18] R. P. Sheridan. The relative importance of domain applicability metrics for estimating prediction errors in qsar varies with training set diversity. *J. Chem. Inf. Model*, pages 1098–1107, 2015.

[19] Ayana Ghosh, Sergei V. Kalinin, and Maxim A. Ziatdinov. Discovery of structure-property relations for molecules via hypothesis-driven active learning over the chemical space, May 2023. arXiv:2301.02665 [cs, q-bio].

[20] Dennis P Trujillo Monirul Shaikh Saurabh Ghosh Ayana Ghosh, Gayathri Palanichamy. Insights into cation ordering of double perovskite oxides from machine learning and causal relations. *Chemistry of Materials*, 34(16):7563–7578, 2022.

[21] Rama Vasudevan Maxim Ziatdinov Sergei V Kalinin, Ayana Ghosh. From atomically resolved imaging to generative and causal models. *Nature Physics*, 18(10):1152–1160, 2022.

[22] Xiaohang Zhang Rama K. Vasudevan Eugene Eliseev Anna N. Morozovska Ichiro Takeuchi Sergei V. Kalinin Maxim Ziatdinov, Christopher T. Nelson. Causal analysis of competing atomistic mechanisms in ferroelectric materials from high-resolution scanning transmission electron microscopy data. *npj Computational Materials*, 6(127), 2020.

[23] Maxim Ziatdinov Yongtao Liu and Sergei V. Kalinin. Exploring causal physical mechanisms via non-gaussian linear models and deep kernel learning: Applications for ferroelectric domain structures. *ACS nano*, 16(1):9, 2021.

[24] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[25] Kenneth A Bollen. *Structural equations with latent variables*, volume 210. John Wiley & Sons, 1989.

[26] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994.

[27] Pedro M. Ferreira. Unsupervised entropy-based selection of data sets for improved model fitting. pages 3330–3337, July 2016. ISSN: 2161-4407.

[28] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[29] Peter Wills and François G Meyer. Metrics for graph comparison: a practitioner's guide. *Plos one*, 15(2):e0228728, 2020.

[30] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. Publisher: American Chemical Society.