Towards Rational Pesticide Design with Graph Machine Learning Models for Ecotoxicology

Jakub Adamczyk
jadamczy@agh.edu.pl
Faculty of Computer Science, AGH University of Krakow
Cracow, Poland

Abstract

This research focuses on rational pesticide design, using graph machine learning to accelerate the development of safer, eco-friendly agrochemicals, inspired by *in silico* methods in drug discovery. With an emphasis on ecotoxicology, the initial contributions include the creation of ApisTox, the largest curated dataset on pesticide toxicity to honey bees. We conducted a broad evaluation of machine learning (ML) models for molecular graph classification, including molecular fingerprints, graph kernels, GNNs, and pretrained transformers. The results show that methods successful in medicinal chemistry often fail to generalize to agrochemicals, underscoring the need for domain-specific models and benchmarks. Future work will focus on developing a comprehensive benchmarking suite and designing ML models tailored to the unique challenges of pesticide discovery.

CCS Concepts

• **Applied computing** → **Agriculture**; **Bioinformatics**; • **Computing methodologies** → *Machine learning approaches.*

Keywords

machine learning, agrochemistry, ecotoxicology, chemoinformatics, molecular graphs, graph classification

ACM Reference Format:
Jakub Adamczyk. 2025. Towards Rational Pesticide Design with Graph Machine Learning Models for Ecotoxicology. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3746252.3761660

1 Introduction

Pesticides are a central part of agrochemistry, comprising herbicides, insecticides, fungicides, and more. They are indispensable for modern agriculture and crop efficiency. However, growing concern surrounds their safety, both for humans (e.g. carcinogenicity, bioaccumulation) and the environment, particularly toxicity to honey bees, fish, birds, and small mammals. Pesticides must be bioactive, often targeting proteins similarly to drugs but with toxic, pest-killing effects. There is often a tradeoff between effectiveness and selectivity, i.e. targeting only specific organism groups, such as parasitic fungi. Computationally, this results in a complex multimodal optimization problem: we seek high lethality to pests, low toxicity to many beneficial (and often evolutionarily distant) species, long-lasting crop protection, but low bioaccumulation, etc.

Agrochemicals, like drugs, are small compounds (typically less than 50 atoms), represented as molecular graphs consisting of atoms and bonds, i.e. attributed graphs. The idea of rational drug design [15], where compound development is guided by computational predictions, has transformed the pharmaceutical industry. *In silico* methods based on data mining and machine learning (ML) estimate key properties of candidate molecules, e.g. solubility, absorption, and toxicity, using prior experimental data. These methods help prioritize promising compounds for expensive wet lab testing, accelerating discovery and cutting costs.

In contrast, data science in agrochemistry remains nascent, and the field is still highly reliant on costly, time-consuming laboratory and field measurements. These deficiencies have consequences: registering a new active agrochemical costs over $300M and takes about 12 years [6]. Meanwhile, the EU Farm-to-Fork strategy aims to cut the use and risk of chemical pesticides by 50% by 2030 [9], driving the demand for faster development of safer alternatives. Laboratory studies are also ethically constrained, as ecotoxicity is commonly assessed using LD50, the dose that kills 50% test organisms, such as honey bees or rats.

In my PhD work, I propose the concept of **rational pesticide design**, leveraging advances in graph machine learning and molecular data mining for novel agrochemical design. From a data science viewpoint, pesticides can be treated similarly to small drug-like molecules, with many of the same chemoinformatics tools applicable. The major challenge is the scarcity of publicly available datasets and benchmarks. As a result, the real-world performance of most molecular graph classification algorithms outside of medicinal chemistry remains unknown. This hinders fair evaluation, as most models are tested on only a few datasets from standard benchmarks, such as MoleculeNet [23].

The research questions in my work are as follows:
RQ1: How can we design a reproducible data processing pipeline to generate high-quality datasets for pesticide property prediction?
RQ2: How well do molecular graph classification algorithms perform in the agrochemical domain, especially those achieving state-of-the-art (SOTA) results in medicinal chemistry?
RQ3: Are agrochemical datasets sufficiently distinct and challenging from medicinal chemistry to serve as meaningful benchmarks for molecular property prediction?

Initial experiments indicate that ML-based rational pesticide design is both feasible and effective, but introduces new challenges.

Supervised by Witold Dzwinel (dzwinel@agh.edu.pl) and Wojciech Czech (czech@agh.edu.pl), AGH University of Krakow.

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761660

The results also question the generalizability of methods that achieve SOTA performance on medicinal datasets, highlighting the need for agrochemistry-specific benchmarks.

2 State of the art

Agrochemistry. The use of ML in agrochemistry remains very limited. The few available datasets focus almost exclusively on ecotoxicology - the most regulated aspect of pesticide design - overseen by agencies such as the US EPA and EU EFSA. Honey bees are of particular interest, as both vital pollinators and economically significant organisms [4, 5, 10].

Existing datasets, such as CropCSM [19] and BeeTOX [22], are small, often just a few hundred molecules, and suffer from quality issues like invalid structures and duplicates [4]. Data curation in agrochemistry is especially challenging and requires domain-specific approaches. For instance, while salts, inorganics, mixtures, and organometallics are typically excluded in medicinal chemistry, they often carry crucial information in pesticides [18].

Measurement errors are significant in toxicology, especially in ecotoxicology - a challenge acknowledged by the US EPA [10]. Regulators define official LD50 toxicity thresholds, allowing the molecular graph regression task to be framed as binary classification. For example, pesticides with LD50 below 11 μg/organism for honey bees are classified as highly toxic (positive class). Similar thresholds apply to other species, such as rats [11] or fish [14].

Graph classification. Many data mining approaches to graph classification have been created over the years. Simple baselines, utilizing topological graph descriptors or distributions of atoms and bonds, have shown remarkable performance in many cases, e.g. LTP [1] and MOLTOP [2]. Graph kernels provide a flexible way to create a pairwise molecule similarity matrix, coupled with SVM classifier, e.g. Weisfeiler-Lehman (WL) or WL Optimal Assignment (WL-OA) kernels [13].

Molecular fingerprints are a staple of chemoinformatics, providing automated feature extraction from molecular graphs [3]. They fall into two main types: substructural fingerprints, which use predefined patterns designed by domain experts (e.g., MACCS, Laggner), and hashed fingerprints, which encode all subgraph occurrences of a given shape, such as circular neighborhoods (ECFP) or short paths (Topological Torsion). These representations are typically used with tree-based ensemble methods for classification.

Graph neural networks (GNNs) are neural networks with graph-based inductive biases, such as the permutation invariance of graph nodes and edges [20]. They follow a message-passing paradigm, where graph convolution combines the feature vector of each atom with those of its neighbors at each layer. GNNs vary in their aggregation mechanisms, e.g. GCN, GraphSAGE, GIN, and GAT. Several architectures are tailored for molecular graphs, such as AttentiveFP. These models learn task-specific features and are typically trained from scratch.

Many approaches have been proposed to pretrain molecular ML models. GNN-based graph transformers such as MAT [16], R-MAT [17], and GROVER [21] combine the inductive biases of GNNs with the high expressiveness of transformers. Alternatively, molecular graphs can be serialized as SMILES strings - an efficient format for datasets and text-based data mining. NLP-inspired models, such as the ChemBERTa transformer [7] and Mol2Vec [12] (which combines ECFP fingerprints with Word2Vec), operate directly on SMILES. Given the small size of molecular datasets, to avoid overfitting during finetuning, embeddings from these models are often used as pretrained feature extractors.

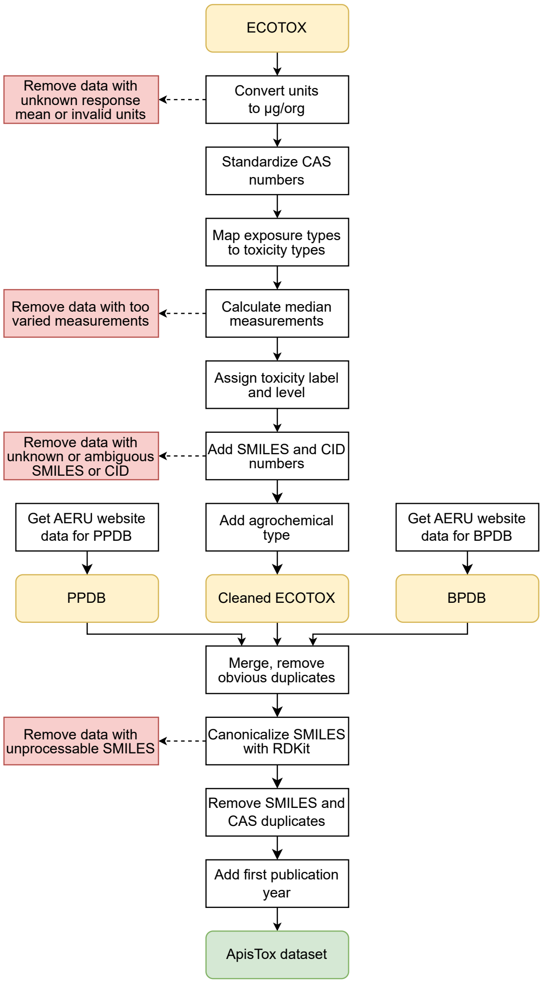
ECOTOX
Convert units to µg/org
Remove data with unknown response mean or invalid units
Standardize CAS numbers
Map exposure types to toxicity types
Calculate median measurements
Remove data with too varied measurements
Assign toxicity label and level
Add SMILES and CID numbers
Remove data with unknown or ambiguous SMILES or CID
Get AERU website data for PPDB
Add agrochemical type
Get AERU website data for BPDB
PPDB
Cleaned ECOTOX
BPDB
Merge, remove obvious duplicates
Canonicalize SMILES with RDKit
Remove data with unprocessable SMILES
Remove SMILES and CAS duplicates
Add first publication year
ApisTox dataset

Figure 1: ApisTox data processing workflow [4].

3 Approach and methodology

ApisTox dataset. As a practical first step in rational pesticide design, we focused on ecotoxicity, specifically for honey bees (*Apis mellifera*). Three major public databases cover this area: ECOTOX, PPDB, and BPDB. ECOTOX contains individual experimental measurements, but requires extensive curation. The pipeline shown in Figure 1 is designed to be general and applies to most ecotoxicology endpoints in ECOTOX. Due to space constraints, only a brief overview is provided here; full details are given in [4].

First, all units are standardized to μg/organism. For each pesticide, measurements are grouped by toxicity type - oral, contact, or other - and the median value per group is taken for. The lowest of the three medians (i.e. the strongest toxicity) is used as the overall LD50. SMILES strings are added using CAS numbers via the PubChem database.

PPDB and BPDB, which are manually curated and provide one record per pesticide, are merged with the preprocessed ECOTOX data. Molecular structures are standardized with RDKit, ensuring no structural duplicates (i.e. identical molecular graphs). Additional metadata include pesticide type (e.g. herbicide) and the first publication date of the literature, allowing for more detailed data analysis.

Importantly, this pipeline is reusable for other ecotoxicology endpoints, for e.g. for algae, fish, or birds. The user only needs to specify the toxicity threshold for the organism of interest.

Predictive ML models. To evaluate the usefulness of ApisTox, several molecular graph classification algorithms were implemented. Baselines included simple atom counts, LTP, and MOLTOP. Over 30 molecular fingerprints were generated using scikit-fingerprints [3], with Random Forest as a classifier. Graph kernels included top-performing options from the literature, particularly WL and WL-OA kernels. GNNs covered general-purpose models - GCN, GraphSAGE, GIN, and GAT - as well as chemistry-specific AttentiveFP. All were trained from scratch with extensive hyperparameter tuning. Pretrained models included graph- and SMILES-based architectures: MAT, R-MAT, GROVER, ChemBERTa, and Mol2Vec. These were used as feature extractors without finetuning, which caused immediate overfitting in preliminary tests. Logistic regression was chosen as the classifier for pretrained embeddings, as it yielded slightly better and more stable results than Random Forest.

Performance evaluation. Random train-test split in molecular data tend to overestimate model performance, as nearly identical molecules often appear in both sets [23]. In medicinal chemistry, scaffold split is used to address this, utilizing the internal graph structure, but it does not work for salts (disconnected graphs), which are common in agrochemistry. We propose two novel approaches: MaxMin split and time split. MaxMin split selects the test molecules to maximize their total distance, creating a diverse test set that uniformly covers the chemical space. The time split assigns the newest molecules to the test set, based on their publication dates in the literature, similar to the real-world pesticide design process.

Molecular diversity. A common measure of molecular diversity in chemoinformatics is the average pairwise Tanimoto similarity between ECFP4 fingerprints [8]. This metric quantifies both inter- and intra-dataset diversity, enabling meaningful comparisons between datasets and benchmarks. It was applied here to compare ApisTox with MoleculeNet classification datasets.

4 Results

Using the designed workflow, we created the ApisTox dataset [4]. It consists of 1035 pesticide molecules in SMILES format with binary toxic/non-toxic labels, following US EPA guidelines. It is moderately imbalanced, with 29% positive cases. Compared to previous datasets (e.g., CropCSM, BeeTox), it is the largest and the only one free of invalid entries or structural duplicates.

The initial classification results are presented in Table 1, with more details in the preprint [5]. Five top-performing molecular fingerprints were selected for brevity. Matthews correlation coefficient (MCC) was used as an evaluation metric, as it works well for imbalanced classification.

Table 1: Classification results. The best metric value for each split (column), in each group, is marked in bold.

Group	Method	MCC	
		MaxMin split	**Time split**
Fingerprints	Atom Pairs	0.45 ± 0.03	0.37 ± 0.03
	Avalon	**0.48 ± 0.03**	0.43 ± 0.02
	ECFP	0.42 ± 0.02	**0.48 ± 0.02**
	RDKit	0.43 ± 0.03	0.46 ± 0.02
	Laggner	0.46 ± 0.03	0.37 ± 0.03
Baselines	Atom counts	**0.36 ± 0.03**	0.29 ± 0.04
	LTP	0.18 ± 0.02	0.23 ± 0.01
	MOLTOP	**0.36 ± 0.03**	**0.33 ± 0.01**
Graph kernels	Propagation	0.32	0.36
	Shortest paths	0.29	0.31
	WL	0.42	0.41
	WL-OA	**0.49**	**0.43**
GNNs	GCN	0.25 ± 0.04	0.30 ± 0.04
	GraphSAGE	0.31 ± 0.05	**0.33 ± 0.04**
	GIN	0.24 ± 0.04	0.32 ± 0.06
	GAT	0.26 ± 0.03	0.26 ± 0.05
	AttentiveFP	**0.35 ± 0.04**	0.29 ± 0.06
Pretrained neural networks	MAT	0.36	0.25
	R-MAT	0.31	**0.35**
	GROVER	0.22	0.05
	ChemBERTa	**0.37**	0.27
	Mol2Vec	0.34	0.31

The main observation is that ApisTox is challenging, with the best method achieving an MCC of 0.48. In both splits, the best method is a molecular fingerprint, calling into question the effectiveness of GNNs on small datasets typical of agrochemistry. In fact, all GNNs and nearly all pretrained neural networks fail to outperform the MOLTOP baseline. Graph kernels, particularly the WL-OA kernel, perform strongly, achieving the best results on the MaxMin split. These findings differ significantly from medicinal chemistry, so how can they be explained?

First, pretrained models often rely on heavily filtered medicinal data. For example, MAT and R-MAT were pretrained on the ZINC dataset, further filtered by the conservative Lipinski Rule of 5. This inherently biases the models toward a limited data distribution. Additionally, a quantitative comparison of ApisTox and the commonly used MoleculeNet datasets (see Figure 2) shows that ApisTox is highly distinct from typical medicinal compounds. Meanwhile, MoleculeNet is internally homogeneous; for instance, BACE and BBBP datasets show high internal similarity. This limits its usefulness, as models are implicitly encouraged to overfit these common patterns to boost benchmark scores.

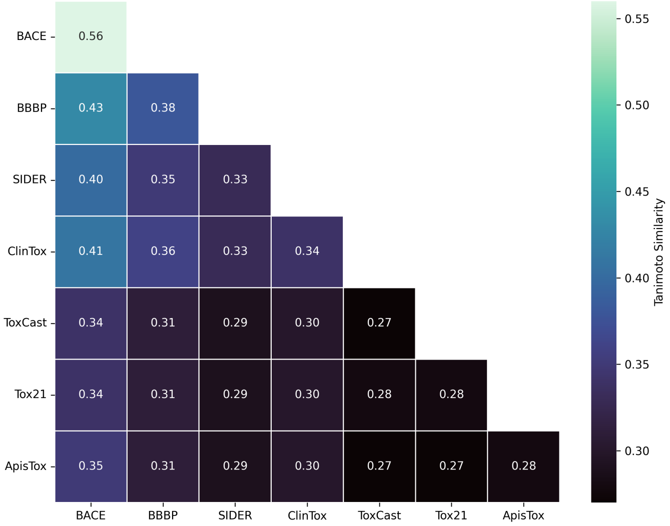

Figure 2: Average Tanimoto similarity between molecules from different datasets.

5 Conclusion and future work

Machine learning, graph mining, and predictive analytics are well established in pharmaceutical sciences and drug design, yet their application in agrochemistry remains nascent. This work introduces the concept of **rational pesticide design**, applying computational *in silico* approaches to aid the development of new agrochemicals. By mirroring pharmaceutical computational methods, we aim to reduce costs and accelerate novel agrochemical development.

Initial efforts include creating the ApisTox dataset [4] of pesticide toxicity to honey bees and training predictive ML models for molecular graph classification [5]. The results show that agrochemistry occupies a chemical space distinct from those of medicinal chemistry benchmarks such as MoleculeNet. Furthermore, models pretrained on pharmaceutical data appear overtuned, failing to generalize to novel agrochemical molecules. This highlights the need for new datasets and benchmarks beyond medicinal chemistry, providing initial answers to the research questions RQ1, RQ2, and RQ3, although further work is required for definitive conclusions.

Future work includes creating a comprehensive pesticides benchmark to fairly evaluate molecular graph classification models in this chemical space. We will also develop agrochemistry-specific predictive models, supported by large-scale pretraining datasets with reduced bias toward medicinal compounds.

Acknowledgments

I thank my supervisors, Prof. Witold Dzwinel and Dr. Wojciech Czech. I thank Jakub Poziemski and Paweł Siedlecki, coauthors of ApisTox and subsequent ML modeling for this dataset. I thank BIT Student Scientific Group for computational resources and aid for this project. This work was supported by the Ministry of Science and Higher Education funds assigned to AGH University in Krakow. I thank Aleksandra Elbakyan for her work and support for accessibility of science.

References

[1] Jakub Adamczyk and Wojciech Czech. 2023. Strengthening Structural Baselines for Graph Classification Using Local Topological Profile. In *Computational Science – ICCS 2023*, Jiří Mikyška, Clélia de Mulatier, Maciej Paszynski, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, and Peter M.A. Sloot (Eds.). Springer Nature Switzerland, Cham, 597–611.

[2] Jakub Adamczyk and Wojciech Czech. 2024. Molecular Topological Profile (MOLTOP)-Simple and Strong Baseline for Molecular Graph Classification. In *ECAI 2024*. IOS Press, 1575–1582.

[3] Jakub Adamczyk and Piotr Ludynia. 2024. Scikit-fingerprints: Easy and efficient computation of molecular fingerprints in Python. *SoftwareX* 28 (2024), 101944.

[4] Jakub Adamczyk, Jakub Poziemski, and Pawel Siedlecki. 2025. ApisTox: a new benchmark dataset for the classification of small molecules toxicity on honey bees. *Scientific Data* 12, 1 (02 Jan 2025), 5.

[5] Jakub Adamczyk, Jakub Poziemski, and Pawel Siedlecki. 2025. Evaluating machine learning models for predicting pesticides toxicity to honey bees. *arXiv preprint arXiv:2503.24305* (2025).

[6] AgbioInvestor. 2024. "Cost of New Agrochemical Product Discovery Development and Registration" Study Results. https://www.agribusinessglobal.com/agrochemicals/agbioinvestor-publishes-cost-of-new-agrochemical-product-discovery-development-and-registration-study-results/.

[7] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712* (2022).

[8] Dávid Bajusz, Anita Rácz, and Károly Héberger. 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* 7, 1 (20 May 2015), 20.

[9] European Commission. 2020. A Farm to Fork Strategy. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52020DC0381.

[10] US EPA. 2025. Pollinator Risk Assessment Guidance. https://www.epa.gov/pollinator-protection/pollinator-risk-assessment-guidance.

[11] Domenico Gadaleta, Kristijan Vuković, Cosimo Toma, Giovanna J Lavado, Agnes L Karmaus, Kamel Mansouri, Nicole C Kleinstreuer, Emilio Benfenati, and Alessandra Roncaglioni. 2019. SAR and QSAR modeling of a large collection of LD 50 rat acute oral toxicity data. *Journal of Cheminformatics* 11 (2019), 1–16.

[12] Sabrina Jaeger, Simone Fulle, and Samo Turk. 2018. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *Journal of Chemical Information and Modeling* 58, 1 (2018), 27–35. PMID: 29268609.

[13] Nils M. Kriege, Fredrik D. Johansson, and Christopher Morris. 2020. A survey on graph kernels. *Applied Network Science* 5, 1 (14 Jan 2020), 6.

[14] Thomas R Lane, Joshua Harris, Fabio Urbina, and Sean Ekins. 2023. Comparing LD50/LC50 machine learning models for multiple species. *ACS Chemical Health & Safety* 30, 2 (2023), 83–97.

[15] Soma Mandal, Mee'nal Moudgil, and Sanat K. Mandal. 2009. Rational drug design. *European Journal of Pharmacology* 625, 1 (2009), 90–100. New Vistas in Anti-Cancer Therapy.

[16] Lukasz Maziarka, Tomasz Danel, Slawomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanislaw Jastrzebski. 2020. Molecule Attention Transformer. *CoRR* abs/2002.08264 (2020). arXiv:2002.08264

[17] Łukasz Maziarka, Dawid Majchrowski, Tomasz Danel, Piotr Gaiński, Jacek Tabor, Igor Podolak, Paweł Morkisz, and Stanisław Jastrzębski. 2024. Relative molecule self-attention transformer. *Journal of Cheminformatics* 16, 1 (03 Jan 2024), 3.

[18] TWG NAFTA. 2012. Quantitative Structure Activity Relationship [(Q) SAR] Guidance Document. *North American Free Trade Agreement (NAFTA), Technical Working Group on Pesticides (TWG)* 186 (2012).

[19] Douglas E V Pires, Keith A Stubbs, Joshua S Mylne, and David B Ascher. 2022. cropCSM: designing safe and potent herbicides with graph-based signatures. *Briefings in Bioinformatics* 23, 2 (02 2022), bbac042.

[20] Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, et al. 2022. Graph neural networks for materials science and chemistry. *Communications Materials* 3, 1 (2022), 93.

[21] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying WEI, Wenbing Huang, and Junzhou Huang. 2020. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12559–12571.

[22] Fan Wang, Jing-Fang Yang, Meng-Yao Wang, Chen-Yang Jia, Xing-Xing Shi, Ge-Fei Hao, and Guang-Fu Yang. 2020. Graph attention convolutional neural network model for chemical poisoning of honey bees' prediction. *Science Bulletin* 65, 14 (2020), 1184–1191.

[23] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* 9, 2 (2018), 513–530.