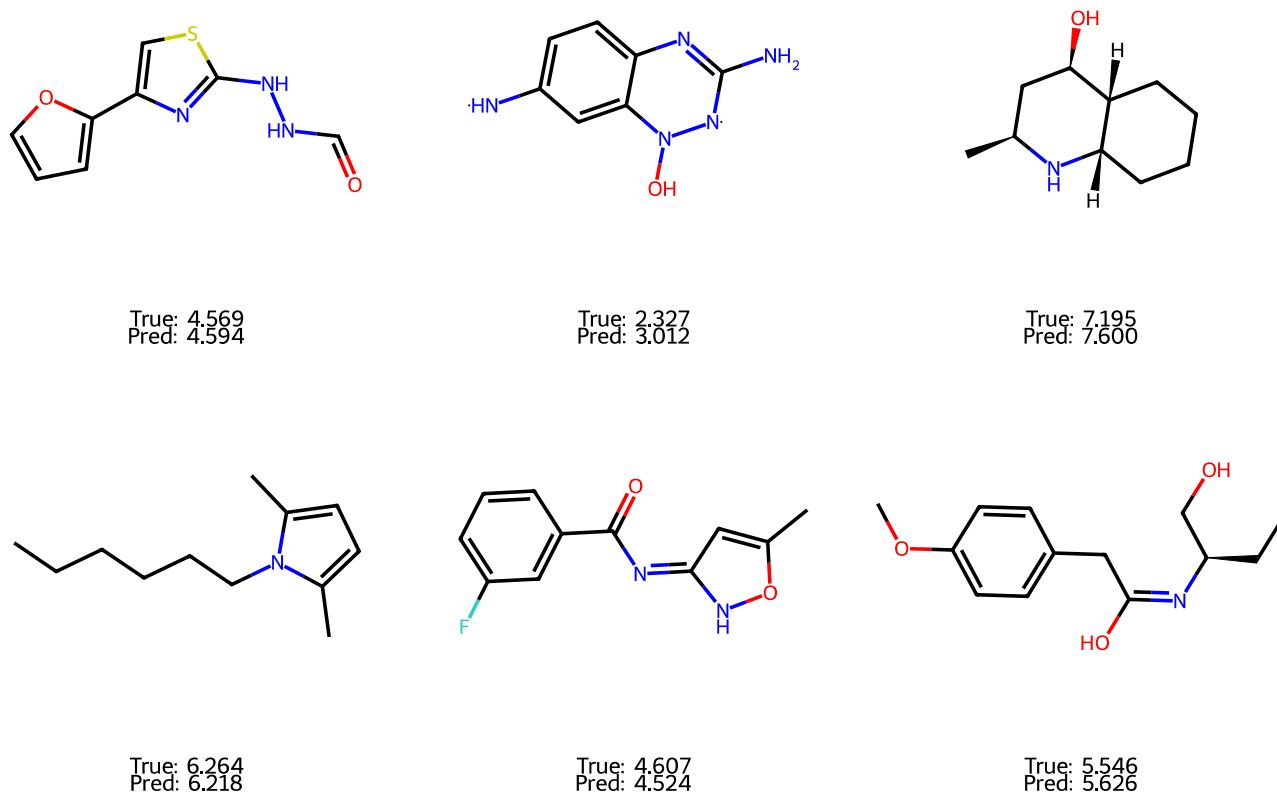


# Predicting Quantum Energy Gaps from Molecular Graphs Using Machine Learning

Matthew Graham\*  
magr2482@colorado.edu  
University of Colorado Boulder  
Boulder, Colorado, USA



**Figure 1: Predicted vs. Ground Truth HOMO-LUMO Gaps for Six Randomly Selected Validation Molecules from PCQM4Mv2.**

## Abstract

The energy difference between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) is the HOMO-LUMO gap. It is a critical quantum chemical property that influences molecular reactivity, stability, and electronic properties, with significant implications for the design of functional materials and drug molecules. While Density Functional Theory (DFT) provides highly accurate estimations, its high computational cost makes it impractical for large-scale virtual screening. To address this, I present a machine learning (ML) framework for rapid and accurate prediction of HOMO-LUMO gaps directly from molecular representations. Using the PCQM4Mv2 Open Graph Benchmark dataset, containing more than 3,746,620 molecules with DFT-calculated gaps, I trained and evaluated various regression models

using molecular encodings, including one dimensional SMILES, two dimensional graph-based representations (ChemML and OGB), Coulomb Matrices, Morgan Fingerprints, and novel approaches incorporating three dimensional molecular coordinate data and global molecular features from RDKit. My best performing model, based on [add best performing representation (likely hybrid 3D GNN)], achieved a mean absolute error (MAE) of X.XX eV on the held-out validation set. This demonstrates its ability to predict HOMO-LUMO gaps with near-DFT accuracy at a fraction of the computational cost.

To improve accessibility and facilitate practical application of my findings, I developed an interactive dashboard that allows users to visualize predicted HOMO-LUMO gaps alongside corresponding molecular structures. This tool enables intuitive exploration of chemical space and rapid identification of promising candidates for specific applications. The reduction in computational time offered by my ML approach, coupled with the friendly user interface,

provides a powerful platform for high-throughput screening and accelerated discovery of novel molecules with desired electronic properties.

## CCS Concepts

• **Computing methodologies** → *Feature selection; Regularization; Quantum mechanic simulation; Molecular simulation; Neural networks; Classification and regression trees*; • **Applied computing** → *Chemistry*.

## Keywords

HOMO-LUMO gap, molecular property prediction, machine learning, quantum chemistry, graph neural networks, 2D descriptors, PCQM4Mv2, virtual screening, interactive dashboard, chemical informatics

### ACM Reference Format:

Matthew Graham. 2025. Predicting Quantum Energy Gaps from Molecular Graphs Using Machine Learning. In *Proceedings of Data Mining Final Project (Data Mining Project '25)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/N/A>

## 1 Introduction

### 1.1 Problem Definition

Accurate prediction of molecular properties plays an important role in modern computational chemistry because it allows for virtual screening and design of novel compounds in materials science, pharmaceuticals, biopharmaceuticals, chemical engineering, and more applications [5, 22, 23]. One of the most fundamental properties is the HOMO-LUMO gap, which refers to the energy difference (measured in electron volts, eV) between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) in a molecule. The HOMO is the highest energy molecular orbital that is full of electrons, whereas the LUMO is the empty orbital right beneath it. The HOMO contains the outermost electrons that have the weakest binding to the nucleus and are most likely to participate in chemical reactions, while the LUMO contains no electrons and is the next available orbital to accept an electron in a chemical reaction. The LUMO is important because it helps define the likelihood of a molecule to accept an electron, a property defined as reduction potential. Similarly, the HOMO helps define the likelihood to donate an electron, a property defined by oxidation potential. A smaller HOMO-LUMO gap means that less energy is required to excite an electron from the HOMO to the LUMO; therefore, it is more reactive and less stable. This gap impacts chemical behaviors from reaction thresholds to photochemical activity to fluorescence properties and electrical conductivity in organic semiconductors [6, 10, 13, 19].

### 1.2 Importance of Molecular Property Prediction

Reliable predictions of electronic properties like the HOMO-LUMO gap are necessary for rational molecule design, where researchers try to tune properties like reactivity, stability, or optical absorption for a target application. For example, if a researcher were trying to create organic solar cell materials, they'd look for molecules with

small gaps for better charge transport [10]. On the other hand, if someone were looking to design a chemically inert molecule or if they wanted to create a highly insulating material, larger gaps are preferred.

## 1.3 Background and Motivation

The traditional method for quantum chemical property prediction has been Density Functional Theory (DFT), a class of computational methods that approximates the behavior of electrons within molecules. Despite DFT's high accuracy, it is computationally expensive and can scale cubically or worse with the number of electrons or atoms. This makes it infeasible to scale to large chemical libraries with millions of candidate molecules. For applications of drug discovery or materials screening, this computational bottleneck makes DFT workflows impractical.

## 1.4 ML-Based Approximation

To overcome this limitation, machine learning (ML) offers a reasonable alternative. By training on large datasets of DFT-calculated properties, ML models can learn to approximate quantum properties such as the HOMO-LUMO gap using only two dimensional or three dimensional molecular representations as input [12]. These models offer faster prediction times with minimal reductions in accuracy, allowing the ability to screen massive compound libraries in seconds rather than weeks. Herein, I explore the efficacy of several machine learning architectures, including kernel ridge regression, random forest, and graph neural networks trained on various molecular representations: SMILES strings, graph-based features, Coulomb matrices, Morgan fingerprints, and 3D coordinate-based representations.

## 1.5 Possible Impact

A successful ML-based predictor of HOMO-LUMO gaps could significantly accelerate innovation in fields such as organic electronics, fluorescent probe design, drug discovery, and more where electronic properties guide molecule selection. In this project, I leverage the PCQM4Mv2 dataset, a large and publicly available benchmark containing millions of molecules with precomputed DFT HOMO-LUMO gaps, to build regression models that generalize well across chemical space [15, 24]. I also plan to deploy an interactive visualization dashboard that will allow users to explore molecular predictions and structures, making the tool useful for researchers, educators, and students.

## 2 Related Work

### 2.1 Quantum Chemical Methods

Density Functional Theory (DFT) remains the most widely used method for predicting quantum mechanical (QM) properties of molecules. It enables accurate estimation of frontier orbital energies (HOMO-LUMO gap), which is important for understanding reactivity, photoexcitation, and charge transfer in molecules. Although DFT-based methods like B3LYP / 6-31G \* have been applied to massive datasets like PubChemQC, covering over 85 million molecules, they are computationally expensive, making them impractical for larger scale molecular screening tasks [14].

## 2.2 ML for Quantum Property Prediction

To address the limitations of DFT, various machine learning (ML) models have been developed to approximate quantum properties directly from molecular structures. Toolkits like ChemML [7] provide access to a range of representations (Coulomb matrices, fingerprints, learned embeddings) and models. Many models have shown success for predictions of quantum properties, such as kernel ridge regression, random forests, and deep neural networks [19, 21]. More recently, graph neural networks (GNNs) and transformer models have achieved state-of-the-art (SOTA) performance by learning from either 2D graphs or 3D conformations of molecules [3, 9, 12].

## 2.3 Benchmark Datasets

Several public datasets have driven progress in this field. The QM7b dataset contains HOMO and LUMO energies calculated at various levels of theory for 7,211 small organic molecules [2, 20]. The PCQM4Mv2 dataset is currently the largest, with 3,746,620 molecules and DFT-calculated HOMO-LUMO gaps, along with both 2D SMILES and optional 3D equilibrium structures [15, 24]. It is the current benchmark for evaluating large ML based quantum property prediction models.

## 2.4 Challenges of Predicting HOMO-LUMO Gaps

While benchmark datasets like PCQM4Mv2 have enabled advances, accurately predicting HOMO-LUMO gaps remains a complex challenge. Molecules in PCQM4Mv2 span a wide range of structural diversity, including simple organic compounds, highly conjugated systems, heteroatom-containing structures, and flexible macrocycles. This diversity results in highly non-linear relationships between molecular structure and electronic properties, making it difficult for simple models to generalize. Furthermore, small structural modifications, such as slight changes in bond lengths or the introduction of polar functional groups, can cause large shifts in the HOMO-LUMO gap. Molecules vary greatly in size and complexity, further complicating prediction tasks. In addition, many ML models rely on 2D graph representations, yet electronic properties are fundamentally three dimensional in nature. Small conformational changes can impact energy gaps significantly. These factors motivate the development of advanced machine learning methods capable of capturing both local atomic environments and global molecular topology to achieve accurate predictions.

## 2.5 State of Current Models

Recent advances have led to significant improvements in predictive performance. The Triplet Graph Transformer (TGT) model introduced by Hussain et al. [9] (2023) combines a novel triplet attention mechanism with interatomic distance prediction that achieves mean absolute errors (MAE) as low as 0.0683 eV on PCQM4Mv2 test-dev sets [9]. Similarly, Uni-Mol + uses refined coarse 3D structures with a neural network to reach MAE 0.0705 eV [11].

However, these models require hundreds of millions of parameters (200M+) and training on multiple high-performance GPU clusters, which may not be feasible in short-term or resource limited projects. More accessible models, like kernel ridge regression (KRR) using fingerprint or coulomb matrix representations have

achieved MAEs between 0.10–0.15 eV on QM9 and QM7b datasets with far fewer resources [13, 21]. These lower complexity baselines offer a practical starting point for building and benchmarking initial models within a limited timeline.

## 3 Proposed Work

This project aims to develop and evaluate machine learning models for fast and accurate prediction of HOMO-LUMO energy gaps, using publicly available quantum chemistry datasets and chemical featurization tools. My pipeline is designed to be modular, with good baseline performance and extensions into newer models if time and compute resources allow.

### 3.1 Data Sources

I will use one dataset: PCQM4Mv2: >3.7M molecules with DFT calculated HOMO-LUMO gaps and 2D SMILES and optional 3D (SDF) [15, 24].

Code will be implemented in Python version 3.8.20 [18]. Molecules will be parsed using RDKit, with both SMILES-based and 3D coordinate representations generated [4]. ML models will be created using scikit-learn [17].

### 3.2 Preprocessing and Feature Engineering

Initial molecular representations will include:

- Coulomb Matrix
- RDKit Morgan fingerprints
- Global RDKit descriptors (MolWt, NumRotatableBonds, TPSA, NumHAcceptors, NumHDonors, RingCount)
- Neural graph tensors from ChemML's `tensorise_molecules` (atom, bond, edge matrices) [7] and OGB's `smiles2graph`.

My first GNN model used ChemML's `tensorise_molecules`, which is based on RDKit internal atom/bond descriptors (atom number, degree, hybridization, etc.). The format is NumPy tensors (`X_atoms`, `X_bonds`, `X_edges`). It is very easy to implement and provides a good baseline, but it lacks customization and does not match the OGB leader board models. Therefore, I switched to OGB's `smiles2graph`, which uses predefined features from PCQM4Mv2. It is formatted as PyTorch Geometric Data objects and provides high customization, while matching OGB leader board. `smiles2graph` gives graph structure and feature encoding that matches the OGB evaluation setup. `AtomEncoder` and `BondEncoder` from `ogb.graphpropred.mol` turn categorical features into learnable embeddings, enabling more expressive models. It is more compatible with PyTorch Geometric, DGL, and OGB evaluation tools. Therefore, I will use `smiles2graph` from `ogb.utils` and also concatenate the graph-based fingerprints/embeddings (from GNNs) with SMILES-based descriptors from RDKit. I will feed into an attention block or possibly use a MLP head.

### 3.3 Modeling Plan

**Phase 1 Baseline Models with 2D Features (Day 1):** Train and evaluate simple regression models using 2D molecular representations via ChemML

- Kernel Ridge Regression
- Random Forest Regressor
- Multilayer Perceptron (MLP) via ChemML

Evaluate performance using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared ( $R^2$ ) using ChemML’s `regression_metrics`.

**Phase 2 (Day 2): Hyperparameter Optimization**

- Kernel Ridge Regression (alpha)
- Random Forest Regressor (number of trees, )
- Multilayer Perceptron (neurons, layers, learning rate)

**Phase 3 (Day 3): Graph Neural Network (GNN) Modeling**

Implement and train a lightweight GNN using ChemML’s graph-based modeling stack:

- Generate atom, bond, and edge feature tensors
- Utilize `NeuralGraphHidden` and `NeuralGraphOutput` layers (convolution width=8, fingerprint length=128)

Explore graph-based modeling using the OGB library:

- Load PCQM4Mv2 dataset using `PygPCQM4Mv2Dataset`
- Train a basic GNN (with `GINEConv`, global mean pool) on the OGB graph data, potentially incorporating RDKit features

**Phase 4 (Day 4): Integration of 3D Molecular Geometry**

- Process 3D molecular structures from SDF files
- Develop a 3D-aware GNN architecture incorporating: Atom features, edge features (including distances), message passing layers (extended `GINEConv`), global attention, CLS token pooling, feature fusion, and an MLP prediction head.
- Evaluate the performance of the 3D-aware GNN model.
- Time permitting, test 3D approaches inspired by Uni-Mol+ [11]

**Phase 5 (Day 5): Evaluation and Interactive Dashboard**

- Final evaluation of the best performing models
- Develop a Streamlit dashboard for model demonstration: Accepts SMILES strings as input, Outputs the predicted HOMO-LUMO gap and a 3D molecular visualization, Supports local predictions and querying models via SQLite or CSV files
- Save model weights, configurations, and predictions in JSON/pickle format for reproducibility.

**Goals and Deliverables**

- Must-do: At least one baseline ML model trained and evaluated and interactive dashboard.
- Stretch goal: Graph based model with 3D featurization if time and compute allow.
- Target performance: MAE around 0.15 eV.

## 4 Evaluation Plan

To assess model performance, I will use standard regression metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Coefficient of Determination ( $R^2$ ) (for interpretation only)

For validation, I will apply either an 80/20 train-test split. I will also create:

- Parity plots with Matplotlib of predicted vs. actual values to visualize predictive alignment [8]
- Residual histograms to assess distribution of errors and potential bias

If I have time, I will compare the inference time of ML models to approximate DFT computational times reported in literature,

emphasizing the potential speedup enabled by ML models. This comparison will highlight the practical value of ML approaches in screening large molecular libraries.

## 5 Discussion

### 5.1 Timeline and Status

**Table 1: Timeline and Status**

Day	Task
1	Download data, parse SMILES, perform EDA
2	Train baseline models and tune hyperparameters
3	Train GNN models and tune hyperparameters
4	Develop Streamlit dashboard
5	Finalize analysis, complete report

Current Status: The project is in the fine-tuning stage. Baseline models have been trained and tuned with Optuna [1]. The final 2D Hybrid GNN is trained and tuned, achieving a .159 eV MAE. The barebones for the Streamlit dashboard are created. 3D model is in development.

### 5.2 Challenges / Backup Plans

Several challenges arose and fallback strategies were/are as follows:

- Large dataset size (3.7M molecules): Use a small subset (5k molecules for baselines) to reduce memory and training time.
- Model convergence issues: Begin with simple models like Kernel Ridge using precomputed fingerprints.
- GNN implementation complexity: If PyTorch Geometric setup or ChemML’s GNN modules prove time-consuming, fallback to fingerprint-based MLPs with neural fingerprints. [16]
- Dashboard responsiveness for large models: Limit dashboard scope to local hosting with pretrained model loading and capped prediction sizes for fast response time.

These fallbacks make sure that even if advanced modeling components encounter roadblocks, a complete and functional baseline system and interactive application can still be delivered on time.

## 6 Conclusion

Predicting HOMO-LUMO energy gaps using machine learning offers a scalable alternative to computationally expensive quantum mechanical methods like DFT. By learning from large public datasets, ML models can approximate electronic properties with significant speedups, supporting faster exploration in materials science and drug discovery.

This project aims to deliver both a trained regression model and an interactive Streamlit dashboard that enables users to input molecular SMILES and receive rapid HOMO-LUMO gap predictions with visual feedback. While this proposal outlines the baseline modeling and deployment strategy, future extensions may include incorporating 3D molecular structures, refining prediction accuracy, and deploying the dashboard for public cloud access.

## References

- [1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- [2] L. C. Blum and J.-L. Reymond. 2009. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* 131 (2009), 8732.
- [3] Chi Chen, Weiye Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. 2019. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials* 31, 9 (April 2019), 3564–3572. doi:10.1021/acs.chemmater.9b01294
- [4] RDKit Development Community. 2024. RDKit: Open-source cheminformatics. <https://www.rdkit.org> Version 2024.3.5.
- [5] Pavlo O. Dral, O. Anatole von Lilienfeld, and Walter Thiel. 2015. Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations. *Journal of Chemical Theory and Computation* 11, 5 (05 2015), 2120–2125. doi:10.1021/acs.jctc.5b00141
- [6] Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, David Duvenaud, Dougal Maclaurin, Martin A Blood-Forsythe, Hyun Sik Chae, Markus Einzinger, Dong-Gwang Ha, Tony Wu, Georgios Markopoulos, Soonok Jeon, Hosuk Kang, Hiroshi Miyazaki, Masaki Numata, Sunghun Kim, Wenliang Huang, Seong Ik Hong, Marc Baldo, Ryan P Adams, and Alán Aspuru-Guzik. 2016. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat Mater* 15, 10 (Oct 2016), 1120–1127. doi:10.1038/nmat4717
- [7] Mojtaba Haghighatdari, Gaurav Vishwakarma, Doaa Altarawy, Ramachandran Subramanian, Bhargava U Kota, Aditya Sonpal, Srirangaraj Setlur, and Johannes Hachmann. 2020. ChemML: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 10 (2020), e1458. doi:10.1002/wcms.1458
- [8] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95. doi:10.1109/MCSE.2007.55
- [9] Md Shamim Hussain, Mohammed J. Zaki, and Dharmashankar Subramanian. 2024. Triplet Interaction Improves Graph Transformers: Accurate Molecular Graph Learning with Triplet Graph Transformers. arXiv:2402.04538 [cs.LG] <https://arxiv.org/abs/2402.04538>
- [10] Peter Bjørn Jørgensen, Murat Mesta, Suranjan Shil, Juan Maria García Lastra, Karsten Wedel Jacobsen, Kristian Sommer Thygesen, and Mikkel N Schmidt. 2018. Machine learning-based screening of complex molecules for polymer solar cells. *J Chem Phys* 148, 24 (Jun 2018), 241735. doi:10.1063/1.5023563
- [11] Shuqi Lu, Zhifeng Gao, Di He, Linfeng Zhang, and Guolin Ke. 2023. Highly Accurate Quantum Chemical Property Prediction with Uni-Mol+. arXiv:2303.16982 [physics.chem-ph] <https://arxiv.org/abs/2303.16982>
- [12] Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. 2023. One Transformer Can Understand Both 2D 3D Molecular Data. arXiv:2210.01765 [cs.LG] <https://arxiv.org/abs/2210.01765>
- [13] Bernard Mazouin, Alexandre Alain Schöpfer, and O. Anatole von Lilienfeld. 2021. Selected Machine Learning of HOMO-LUMO gaps with Improved Data-Efficiency. arXiv:2110.02596 [physics.chem-ph] <https://arxiv.org/abs/2110.02596>
- [14] Maho Nakata and Toshiyuki Maeda. 2023. PubChemQC B3LYP/6-31G\*\*/PM6 Data Set: The Electronic Structures of 86 Million Molecules Using B3LYP/6-31G\* Calculations. *Journal of Chemical Information and Modeling* 63, 18 (09 2023), 5734–5754. doi:10.1021/acs.jcim.3c00899
- [15] Maho Nakata and Tomomi Shimazaki. 2017. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling* 57, 6 (06 2017), 1300–1308. doi:10.1021/acs.jcim.7b00083
- [16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimeshine, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejwani, Vladimir Chilikurthy, Benoit Steiner, Lu Fang, Junjie Bai, Jitendra Chaudhry, Myles Wu, Roman Chien, Peng Xiao, Edward Bailie, Yuchen Garay, Jeremy Zhang, Michael Johnson, Hugo Touvron, Naman Khan, Ruifei Han, Soumith Pettehnofer, Dušan David, and Zachary and Lip-ton. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* 32. 8024–8035. <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [17] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Courneau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [18] Python Software Foundation. 2019. Python Language Reference, version 3.8.20. <https://www.python.org/>
- [19] Edward O. Pyzer-Knapp, Kewei Li, and Alan Aspuru-Guzik. 2015. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Advanced Functional Materials* 25, 41 (2025/04/22 2015), 6495–6502. doi:10.1002/adfm.201501919
- [20] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. 2012. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters* 108 (2012), 058301.
- [21] Annika Stuke, Milica Todorović, Matthias Rupp, Christian Kunkel, Kunal Ghosh, Lauri Himanen, and Patrick Rinke. 2019. Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *The Journal of Chemical Physics* 150, 20 (4/22/2025 2019), 204121. doi:10.1063/1.5086105
- [22] O. Anatole von Lilienfeld and Kieron Burke. 2020. Retrospective on a decade of machine learning for chemical discovery. *Nature Communications* 11, 1 (2020), 4895. doi:10.1038/s41467-020-18556-9
- [23] O Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. 2020. Exploring chemical compound space with quantum-based machine learning. *Nat Rev Chem* 4, 7 (Jul 2020), 347–358. doi:10.1038/s41570-020-0189-9
- [24] 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.