Machine Learning for HOMO–LUMO Gap Prediction and Inverse Molecular Design

Authors:
Thibo Van Eeckhoorn, Yarno De Jaeger and Mattice Criel

Ghent University

Academic Year: 2025–2026

1 Introduction

Still need to add the references An important property of chemical systems, including organic molecules, is their HOMO-LUMO gap which is the energy difference between the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO). This is a fundamental descriptor of the electronic structure of a molecule. It governs key properties such as reactivity, optical absorption and charge transport characteristics. As a consequence, accurate determination of the HOMO-LUMO gap is essential for molecular design in chemistry, materials science, and molecular engineering.

Experimentally, the HOMO-LUMO gap can be approximated through optical spectroscopy. In UV-Vis absorption spectroscopy, electron excitation from the ground state (associated with the HOMO) to excited states (related to the LUMO) produces an absorption onset, or optical band edge, whose energy corresponds to the optical gap. Similarly, photoluminescence spectroscopy captures the energy released during radiative relaxation. However, experimental measurements require synthesized, purified materials and are not suitable for large-scale exploration of hypothetical chemical structures. An alternative way to get the HOMO-LUMO gap is with computational chemistry, such as Hartree-Fock and Density functional theory (DFT). Despite their accuracy, these quantum chemical methods are computationally expensive and scale poorly with molecular size and dataset volume. These factors are the motivation to search for faster predictive approaches to get insight in the electronic properties of molecules.

Machine learning (ML) provides a strategy to overcome these computational limitations. After training, ML models can predict electronic properties with negligible computational cost compared to traditional quantum chemistry. Recent work has demonstrated that methods such as gradient-boosted decision trees, kernel regressors, and graph neural networks (GNNs) achieve mean absolute errors well below 0.2 eV in HOMO-LUMO gap prediction tasks. Crucially, large-scale quantum-chemical datasets like QM9, which contains orbital energies for over 130,000 small organic molecules, enable the development and benchmarking of ML models with the data volume required for robust learning.

Beyond property prediction, ML also enables generative molecular design. Traditional experimental workflows require iterative synthesis and characterization to identify molecules with desired electronic properties. ML-based generative models, however, make it possible to explore hypothetical chemical structures, optimize properties in silico, and drastically reduce the number of costly experiments. This accelerates molecular discovery pipelines, supports safer design exploration, and opens chemical regions that would be impractical to investigate experimentally.

In this study, we pursue a two-fold objective. First, we build predictive models for HOMO-LUMO gap estimation using both Light Gradient Boostin Model (LightGBM), a high-performance gradient boosting framework and a graph neural network (GNN) that learns molecular representations directly from graph topology. Second, we integrate the predictive GNN into a conditional variational autoencoder (CVAE) to generate novel molecular structures with user-specified target HOMO-LUMO gaps. This combined framework enables both accurate property prediction and inverse molecular design, supporting accelerated discovery workflows in molecular design.

2 Methodology

2.1 Dataset

For all models in this study, we used the QM9 dataset, a widely used quantum chemistry dataset containing small organic molecules. Only molecules with valid SMILES representations were considered, as SMILES (Simplified Molecular Input Line Entry System) is a text-based language that encodes chemical structures as ASCII strings, allowing computers to process molecular structures efficiently. For example, benzene is represented as `C1=CC=CC=C1` in SMILES notation.Here you can include how many molecules are in the original QM9 dataset and how many remained after filtering invalid or duplicate SMILES. Although SMILES strings can be read by computers, they are not directly suitable for machine learning models, as most models require numerical input. Therefore, preprocessing steps are necessary to convert molecular structures into informative, fixed-length numerical representations.

2.2 Molecular Representations and Preprocessing

In this study, we focus on two different models for HOMO-LUMO gap prediction: LightGBM, a classical machine learning model and GNNs, a deep learning model, which require different types of molecular representations. Therefore, two types of molecular representations were generated from the SMILES strings.

2.2.1 Descriptor-based features for LightGBM

Using RDKit, a set of 217 molecular descriptors was generated for each molecule from the SMILES representation.[1] These descriptors capture various aspects of molecular structure and properties, including molecular size, shape, electronic characteristics, surface area, and the presence of functional groups. Together, they provide a fixed-length numerical representation of each molecule suitable for machine learning models such as LightGBM.

2.2.2 Graph-based representation for GNNs:

For graph neural networks, molecules were represented as undirected graphs where atoms are nodes and bonds are edges. Node features included atom types, degree, hybridization, aromaticity, and formal charge. Edge features included bond type and conjugation status. This representation allows the GNN to learn structural features directly from molecular graphs without relying on precomputed descriptors. This part is maybe writen better by Yarno because you made GNNs

2.3 Data Splitting

To ensure unbiased evaluation, the dataset was split into training, validation, and test sets. For LightGBM, nested cross-validation was used:

- **Outer loop:** 5-fold cross-validation to estimate generalization performance
- **Inner loop:** 10-fold cross-validation for hyperparameter tuning using Optuna

The splits were stratified with 30 bins to preserve the distribution of HOMO-LUMO gaps across the different folds for better results.

2.4 LightGBM

LightGBM is a gradient-boosted decision tree algorithm optimized for speed and high-dimensional data. It grows trees in a leaf-wise rather than level-wise manner, allowing it to capture complex, non-linear relationships more efficiently than many traditional boosting methods. This makes it well-suited for tabular molecular descriptor data, as it is robust to feature scaling, multicollinearity, and missing values.

Hyperparameter optimization for the LightGBM model was carried out using Optuna. The search space included parameters controlling model complexity and regularization: the number of leaves (num_leaves, 16–256), the learning rate (learning_rate, 1×10^{-3}–0.1, log-scaled), feature subsampling (feature_fraction, 0.7–1.0), sample bagging (bagging_fraction, 0.7–1.0; bagging_freq, 1–5), and the minimum number of samples per leaf (min_data_in_leaf, 10–100). Other LightGBM settings, such as the boosting type (gbdt) and objective function (regression_l1), were kept fixed throughout. Optuna selected the optimal hyperparameter configuration by minimizing the mean absolute error (MAE) obtained from the inner loop of the nested cross-validation, using a total of 50 trials.

After nested cross-validation, the best-performing hyperparameters were selected (either by averaging over outer folds or by majority vote). The final LightGBM model was then trained on the entire dataset to produce predictions for comparison with the GNN model.

References

[1] RDKit: Open-Source Cheminformatics RDKit Documentation and Python Tutorials. https://www.rdkit.org/docs/GettingStartedInPython.html, 2025; Accessed December 6, 2025.