Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 203

High throughput tight binding calculation of electronic HOMO–LUMO gaps and its prediction for natural compounds

Sascha Thinius

This research investigates predicting the Highest Occupied Molecular Orbital and the Lowest Unoccupied Molecular Orbital (HOMO–LUMO; short HL) gap of natural compounds, a crucial property for understanding molecular electronic behavior relevant to cheminformatics and materials science. To address the high computational cost of traditional methods, this study develops a high-throughput, machine learning (ML)-based approach. Using 407 000 molecules from the COCONUT database, RDKit was employed to calculate and select molecular descriptors. The computational workflow, managed by Toil and CWL on a high-performance computing (HPC) Slurm cluster, utilized Geometry – Frequency – Noncovalent – eXtended Tight Binding (GFN2-xTB) for electronic structure calculations with Boltzmann weighting across multiple conformational states. Three ensemble methods, namely Gradient Boosting Regression (GBR), eXtreme Gradient Boosting Regression (XGBR), Random Forrest Regression (RFR) and a Multi-layer Perceptron Regressor (MLPR) were compared based on their ability to accurately predict HL-gaps in this chemical space. Key findings reveal molecular polarizability, particularly SMR_VSA descriptors, as crucial for HL-gap determination in all models. Aromatic rings and functional groups, such as ketones, also significantly influence the HL-gap prediction. While the MLPR model demonstrated good overall predictive performance, accuracy varied across molecular subsets. Challenges were observed in predicting HL-gaps for molecules containing aliphatic carboxylic acids, alcohols, and amines in molecular systems with complex electronic structure. This work emphasizes the importance of polarizability and structural features in HL-gap predictive modeling, showcasing the potential of machine learning while also highlighting limitations in handling specific structural motifs. These limitations point towards promising perspectives for further model improvements.

Received 8th May 2025
Accepted 30th October 2025

DOI: 10.1039/d5dd00186b

rsc.li/digitaldiscovery

1 Introduction

The HL-gap, a fundamental electronic property of molecules, plays a crucial role in understanding and predicting their reactivity, stability, and optical properties. Accurate prediction of the HL-gap is essential in diverse fields such as materials science, drug discovery, organic electronics and energy storage, among others. The design and development of novel functional materials and pharmaceuticals often rely on the ability to fine-tune the electronic properties of molecules, including the HL-gap. However, accurate HL-gap calculation is challenging due to the inherent complexities of electronic structure theory, as it requires careful selection of theoretical methods (*e.g.*, density functional theory (DFT) functionals, basis sets), exploration of molecular conformational space (*i.e.*, sampling) and consideration of environmental influences attributable to a solvent, all of which introduce approximations and computational expense. Especially for large datasets of complex molecules like natural products, this bottleneck hinders the rapid exploration of chemical space and the identification of promising candidates.

While machine learning models have shown promise in predicting molecular properties, their application to HL-gap prediction in large and diverse datasets of natural products remains unexplored. Furthermore, an accurate prediction of the HL-gap for natural products is particularly challenging due to their structural complexity and diversity.

Several studies have demonstrated the potential of ML models to accurately and efficiently estimate various molecular characteristics, including electronic properties crucial for understanding chemical behavior.[1–10] The application of ML in this domain has ranged from traditional models using curated feature sets[1,2,6,9,11–13] to sophisticated deep learning architectures that learn directly from molecular structures.[1,2,14–16] Early and contemporary studies have successfully used models like Random Forests,[1,8,9,11,17] Support Vector Machines,[1,12] and Gradient Boosting Regressors[4,17–19] with pre-calculated molecular descriptors and fingerprints[9,12,13,17] to achieve strong

Fraunhofer Institute for Manufacturing Technology and Advanced Materials, IFAM, Wiener Strasse 12, 28359 Bremen, Germany. E-mail: sascha.thinius@ifam.fraunhofer.de; sascha.thinius.87@gmail.com

predictive performance on diverse datasets.[2,7,9,12,17,18,20,21] Those linear models have shown success in specific chemical spaces like polycyclic aromatic hydrocarbons.[22,23] While powerful, these methods' performance is intrinsically tied to the quality and relevance of the features. For instance, Pereira *et al.*[9] explored random forest models for predicting HOMO and LUMO energies, achieving good accuracy with molecular descriptors combined with semi-empirical orbital energies. Schmidt *et al.*[24] explored various ML algorithms, including linear and kernel-based regression, decision trees, and neural networks, for predicting properties like crystal structure and thermal conductivity, emphasizing the trade-off between accelerated research and the challenges of interpretability and data quality.

Reiser *et al.*[25] reviewed the application of graph neural networks (GNNs) in materials science. The field has seen a significant shift towards deep learning, particularly with the advent of GNNs that can leverage complete atomic-level representations. These end-to-end models learn relevant features directly from the molecular graph, mitigating the need for manual feature engineering.[1,11,14–16,19,20,26] Seminal works on the QM9 benchmark dataset established the high performance of these following methods:

- Schrödinger Convolutional Neural Network (SchNet),[16] which operates on atomic types and Cartesian coordinates and has been successfully applied not only to QM9 (ref. 16) but also to complex systems like oligothiophenes,[1,15] with SchNet achieving the best performance among other GNNs, particularly for larger molecules.
- Message Passing Neural Networks (MPNNs),[27] a general framework for learning on graphs, with variants like deep (D) MPNN[8] also showing excellent performance.
- MatErials Graph Network (MEGNet),[2,26] a universal GNN framework for predicting properties of both molecules and crystals, incorporating global state variables and demonstrating transfer learning capabilities.
- Other advanced architectures like Deep Tensor Neural Networks (DTNNs)[11,15,21,28] have also proven effective for predicting electronic properties and designing novel molecules.
- Generative models as the Recurrent Neural Network (RNN) with transfer learning specifically employed by Yuan *et al.*[19] on electronic properties, to generate novel oligomers with targeted HL-gaps, demonstrating the potential of deep generative models but also the inherent trade-off between chemical space exploration and property optimization.
- Finally, Montavon *et al.*[6] introduced a deep multi-task neural network for predicting multiple electronic properties.

A key challenge in applying these data-hungry models is the scarcity of high-quality data for specific or complex chemical systems. To address this, transfer learning has emerged as a powerful strategy. By pre-training a model on a large, general dataset (*e.g.*, PubChemQC) and then fine-tuning it on a smaller, specific dataset, researchers have successfully predicted properties for conjugated oligomers,[3] porphyrins[8] and organic photovoltaics.[4,29]

While these studies highlight the remarkable potential of ML for predicting electronic properties, challenges remain in addressing data requirements, interpretability, and the accurate prediction of properties across vast and highly diverse chemical spaces, such as the natural products domain beyond the limited complexity of the QM9 dataset. This work aims to address this latter challenge by developing a high-throughput workflow and robust ML model for predicting the HOMO–LUMO gaps of over 400 000 natural products. This study aims to not only develop predictive models but also to gain insights into the key molecular features that influence the HL-gap, contributing to a deeper understanding of structure–property relationships.

2 Computational methods

2.1 Data and code preparation

The molecular structures for this study were sourced from the Collection of Open Natural Products (COCONUT) database.[30] This database was chosen as it represents one of the largest, precompiled, and open-access resources for natural products. COCONUT aggregates molecular collections from a multitude of sources, including subsets from other well-known repositories like the ZINC Natural Products database and the Universal Natural Products Database (UNPD). By providing a single, comprehensive, and curated collection, it eliminates the need to gather and harmonize data from various individual repositories, making it an ideal starting point for a large-scale analysis and for training a robust machine learning model. The database provides molecular data in the Structure-Data File (SDF) format, which was parsed for this work. From these initial structures, SMILES (Simplified Molecular-Input Line-Entry System) strings were generated for use in subsequent descriptor calculations. Beyond structural information, COCONUT collects and curates a variety of data on natural products, including calculated properties and descriptors. The provided structures do not contain explicit solvent information; therefore, all subsequent electronic structure calculations were performed assuming gas-phase conditions.

The Common Workflow Language[31,32] (CWL) is a highly flexible language widely used in the field of bioinformatics to create computational workflows in contrast to others[33–35] utilized in the field of computational chemistry. The only prerequisite for workflow integration is that the computational task must be executable on the command line. To ensure consistency and package isolation, the software packages were installed using python package managers into a virtual environment with a python–click interface. Specifically, the following essential packages were installed in this way: RDKit,[36] pandas,[37] Atomic Simulation Environment (ASE)[38] and xTB.[39]

In addition, a modest effort was required to integrate functions for the conformer generation, the Boltzmann weighting, the xTB-wrapper, I/O handling as well as the click interface into the virtual environment as a python package on **https://github.com/sthinius87/HL-gaps-pub**. The CWL-Input files are written YAML-format. All code developed is published at Zenodo (**https://zenodo.org/records/15113790**) *via* GitHub (**https://github.com/sthinius87/HL-gaps-pub**).[40]

2.2 Workflow and computational details

The provided flowchart in Fig. 1 outlines a computational workflow managed by the Toil workflow engine and orchestrated using the CWL. The workflow, designated by unique Database Identifiers (DB-IDs) within the COCONUT project, is executed on a high-performance computing cluster, with resource allocation optimized by the Slurm workload manager.

The workflow's core functionality is encapsulated within a virtual environment. This environment houses Python code that, triggered *via* a click interface, initiates a series of computational tasks to finally evaluate the molecule's HL-gap. These tasks involve: employing the RDKit cheminformatics toolkit, the workflow generates diverse molecular conformations, exploring the potential spatial arrangements of atoms within a molecule. To account for conformational flexibility, a set of 10 conformers was chosen to balance the need for adequate conformational space sampling with the computational cost inherent in a high-throughput study of this scale. This approach, combined with Boltzmann weighting, provides a thermodynamically averaged property that is often more representative than relying on a single lowest-energy conformer, whose ranking might be inaccurate or which may not be the sole contributor to the molecule's properties at room temperature. Following the initial generation of conformers with RDKit, each conformation was subjected to a geometry optimization to find its nearest local energy minimum. This optimization was performed at the GFN2-xTB level of theory, employing the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimizer. After optimization, the electronic properties for each of these now stable conformations were computed at the same GFN2-xTB level through self-consistent charges (SCC) to determine the HOMO–LUMO gap. More advanced methods, like DFT, would also be conceivable, but would go beyond the limit of our computational resources. The accuracy of the calculated HL-gaps with GFN2-xTB method against higher-level theoretical benchmarks (*e.g.*, DFT) or experimental data was not assessed in this study. Finally, a Boltzmann weighting scheme is applied to assess the relative stability and population of each conformation at a given temperature. The parameters for each task are transferred to the code *via* the click interface, which are as follows:

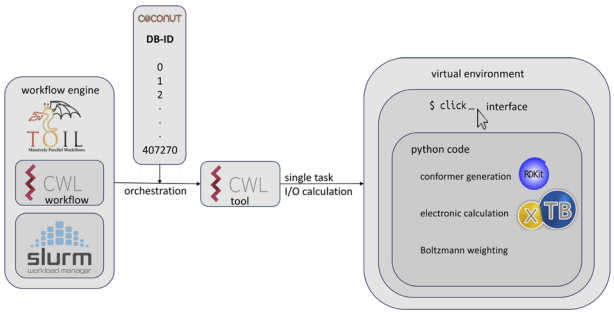

Fig. 1 This flowchart illustrates a Toil–CWL workflow for automated molecular electronic structure simulations. The workflow leverages RDKit for conformer generation, XTB for electronic structure calculations, and Slurm for efficient resource allocation on HPC clusters.

- Number of conformers (RDKit)
- Accuracy (xTB: SCC convergence criteria)
- Electronic temperature (xTB: fermi smearing)
- Calculation method (xTB: current code flavors).

This workflow finally was applied to ~407k SMILES strings, resulting in ~406k results after curation of the dataset.

2.3 Descriptor calculation and machine learning model

For developing the machine learning model, a dataset was constructed, comprising the molecules COCONUT-ID, the calculated HL-gap, and its SMILES string. It is essential to clarify that the L-gap is not a property provided by the COCONUT database. The HL-gaps used in this work were explicitly calculated for each molecule using the GFN2-xTB method as described in Section 2.2. These xTB-calculated HL-gaps served as the target property (or ground truth) for the machine learning models. Consequently, all reported prediction errors are calculated by comparing the models' predictions against these xTB-calculated values. Based on the molecule's SMILES string, 210 molecular descriptors were calculated for each molecule using RDKit. The number of descriptors was further reduced by the feature correlation with a threshold ≥ 0.75 in the correlation matrix and as a second condition descriptors with a threshold ≤ 0.15 in the variance were removed. This results in a set of 56 features for the machine learning model that can be accessed in the SI.

Four regression algorithms were selected for this study: three powerful tree-based ensembles—GBR, RFR, and XGB—and a MLPR. This selection allows for a robust comparison between two distinct and widely used classes of machine learning algorithms: tree-based ensembles and artificial neural networks (ANN). The literature confirms that both algorithmic classes are frequently employed and serve as strong baselines for predicting molecular properties. For example, GBR has been successfully applied, and found to be the best-performing model, for predicting properties of non-fullerene acceptors.[18] Similarly, MLPR and other neural network architectures are a common choice for modeling electronic properties in large molecular datasets, from early deep learning[6,7,9,11] demonstrations to more recent ANN studies.[18] This choice allows for a valuable comparison between these two established algorithmic approaches on a large-scale natural product dataset using pre-calculated molecular descriptors.

For the hyperparameter optimization a randomized search with 2000 iterations and cross-validation with a fold of 3 has been applied. Multiple parameters were involved in the randomized search. For the GBR the most critical parameters are the learning rate (learning_rate), the number of boosting stages (n_estimators), the fraction of samples to be used for fitting the individual base learners (subsample) and the number of nodes in the tree (max_depth) whereas for the MLPR the number neurons and layers (hidden_layer_sizes), the L2-regularization term (alpha) and the exponential decay rate for estimates of first moment vector in the Adam[41] solver (beta_1) parameters were considered in the optimization. For the RFR, the most critical parameters are n_estimators, max_depth, the

minimum samples required to split a node (min_samples_split), and the number of features to consider for a split (max_features), whereas for the XGB model, the learning_rate, n_estimators, max_depth, and the L1 (reg_alpha) and L2 (reg_lambda) regularization terms were considered in the optimization. The optimized set of parameters can be found in the SI. Further the train-test split was set to a ratio of 0.7 to 0.3, respectively. For initial transformation of the data the StandardScaler was applied.

To ensure the random data partitioning was representative, the statistical distributions of key molecular descriptors and the target property were analyzed across multiple splits and found to be virtually indistinguishable between the training and test sets (see SI, Fig. S3). This confirms the absence of systematic bias in the data split. However, this high-level statistical similarity masks the underlying structural novelty of the test set, which serves as the true measure of the model's generalization ability. The test set is composed of over 120 000 unique molecular structures that the model has not encountered during training. The critical test is whether the model can generalize beyond the specific examples it has seen to accurately predict properties for these new chemical entities. Additionally, the analysis of the MLPR model's performance on distinct structural subgroups, presented later in Section 2.4.3, provides strong evidence for this robust generalization. The model maintains high predictive accuracy across various challenging structural elements, including different numbers of aromatic rings and complex functional groups. This demonstrates that the model is not merely interpolating based on overall statistical similarity but has learned the fundamental relationships between molecular structure and the HOMO–LUMO gap.

When evaluating the GBR and the MLPR model, the metrics of both models were calculated using a 10-fold shuffle-split cross-validation strategy with a 0.7 to 0.3 train-test ratio as shown in Table 1. Comparing the models reveals a nuanced picture. A fair comparison requires establishing a baseline of predictive accuracy. All four models demonstrate strong absolute performance on the unseen test data, achieving R^2 scores above 0.94 and MAE values below 0.21 eV. This confirms their validity as powerful predictive tools for this chemical space. Notably, the XGB model emerged as the top-performing model in terms of absolute accuracy, yielding the highest test R^2 (0.958) and the lowest MAE (0.180 eV). The model's robustness was evaluated by analyzing the generalization gap—the difference in performance between the training and test sets. A comparison across all four models reveals significant differences, as detailed in Table 1. While all three tree-based ensemble models show a large performance drop from the training data to the test data, the Random Forest model shows the most substantial gap in the R^2 score ($\Delta R^2 \approx 0.048$). In contrast, the MLPR model displays the greatest generalization stability, with the smallest performance gap in its R^2 score ($\Delta R^2 \approx 0.022$). This conclusion is strongly corroborated when analyzing the absolute error metrics. The tree-based models exhibit a pronounced increase in error on the test set. For instance, the XGB model's MAE increases by over 150% (from 0.069 eV to 0.180 eV), and the Random Forest model's MAE shows a more than six-fold increase (from 0.030 eV to 0.194 eV). The MLPR model, however, shows a much smaller and more controlled relative increase in its MAE of only 24% (from 0.169 eV to 0.210 eV). A similar trend is observed for the MSE and RMSE. This expanded analysis reveals a clear trade-off. For applications where achieving the lowest possible prediction error is the sole priority, the XGB model is the superior choice based on its test set MAE. However, for the goal of this study—developing a reliable and robustly generalizable model—the MLPR's demonstrated stability across multiple metrics makes it the most suitable candidate for the subsequent in-depth feature importance and error analyses.

2.4 Learning outcomes

In this section, the learning outcomes derived from analyzing both the GBR and MLPR models are presented, focusing on feature importance, overall performance, subset analysis and an in-depth analysis of prediction errors. This final step aims to identify molecular features and subgroups that pose a challenge to the model, providing insights into its limitations and potential avenues for improvement.

2.4.1 Feature importance analysis. To understand the key molecular properties driving HL-gap predictions and to compare the learning strategies of the different models, a feature importance analysis was conducted. For the MLPR model, which was selected for in-depth analysis due to its superior generalization, permutation importance on the test set was used to identify features crucial for predicting on unseen data. For the tree-based models (GBR, RFR, and XGB), the built-in Gini importance was calculated.

Table 1 Metrics and standard deviation (±) of the GBR and MLPR models for train and test sets evaluated using a 10-fold shuffle split of the data set

Metrics	R^2-score	MSE [eV^2]	MAE [eV]	RMSE [eV]
MLPR-train	0.9688 ± 0.0009	0.0519 ± 0.0017	0.1686 ± 0.0025	0.2279 ± 0.0037
MLPR-test	0.9470 ± 0.0008	0.0886 ± 0.0010	0.2099 ± 0.0012	0.2976 ± 0.0016
GBR-train	0.9917 ± 0.0001	0.0138 ± 0.0001	0.0905 ± 0.0002	0.1173 ± 0.0003
GBR-test	0.9562 ± 0.0006	0.0732 ± 0.0007	0.1865 ± 0.0005	0.2706 ± 0.0013
XGB-train	0.9943 ± 0.0001	0.0094 ± 0.0001	0.0694 ± 0.0003	0.0972 ± 0.0004
XGB-test	0.9580 ± 0.0005	0.0702 ± 0.0005	0.1799 ± 0.0005	0.2650 ± 0.0010
RFR-train	0.9989 ± 0.0001	0.0019 ± 0.0001	0.0295 ± 0.0001	0.0432 ± 0.0001
RFR-test	0.9505 ± 0.0006	0.0828 ± 0.0005	0.1940 ± 0.0006	0.2878 ± 0.0009

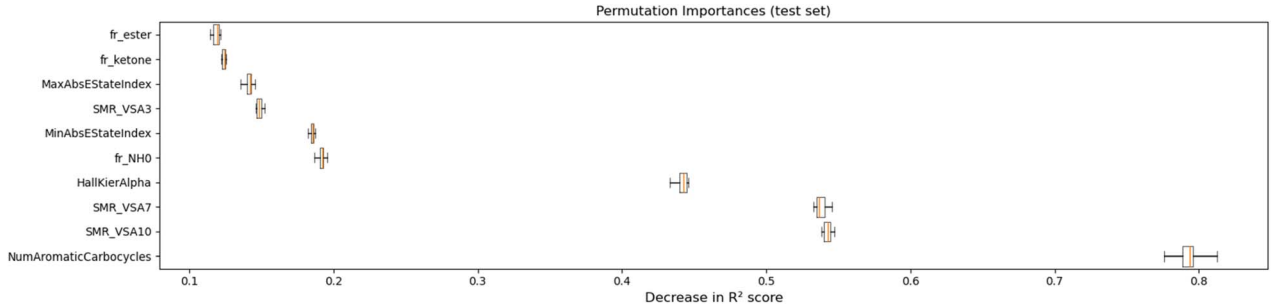

Fig. 2 Permutation importance of the 10 most important features on test set performance of MLPR model for HL-gap prediction. NumAromaticCarbocycles, SMR_VSA7, SMR_VSA10, and HallKierAlpha exhibit the greatest impact on model performance on the test set, as revealed by their high permutation importance scores.

Table 2 Top 5 most important features for each model, sorted by importance

Rank	MLPR (permutation)	GBR (Gini)	RFR (Gini)	XGB (Gini)
1	NumAromaticCarbocycles	SMR_VSA7	SMR_VSA7	SMR_VSA7
2	SMR_VSA7	SMR_VSA10	HallKierAlpha	NumRadicalElectrons
3	SMR_VSA10	fr_ketone	SMR_VSA10	fr_ketone
4	HallKierAlpha	SlogP_VSA8 (ref. 42 and 45)	MinAbsEStateIndex	SMR_VSA10
5	MinAbsEStateIndex	MinAbsEStateIndex	MaxAbsEStateIndex	SlogP_VSA12 (ref. 42 and 45)

The permutation importance of the MLPR model, shown in Fig. 2, reveals that a combination of structural, polarizability, and electronic descriptors governs the HL-gap prediction. By a significant margin, the most influential feature is NumAromaticCarbocycles, indicating that the presence and number of aromatic rings is a primary determinant. This is followed by descriptors related to molecular polarizability[42] (SMR_VSA7 and SMR_VSA10) and shape (HallKierAlpha[43]). MinAbsEStateIndex and MaxAbsEStateIndex, and the presence of specific fragments[44] like amides (fr_NH0), ketones (fr_ketone), and esters (fr_ester) also play a significant role. A clear consensus emerges across all models: molecular polarizability is a fundamental driver of the HL-gap (see Table 2). The SMR_VSA7 and SMR_VSA10 descriptors appear in the top features for every model. Similarly, molecular shape (HallKierAlpha) and the presence of specific functional groups like ketones (fr_ketone) are consistently identified as significant contributors. This agreement between methodologically distinct models provides strong confidence that these features have a true physical relationship with the HOMO–LUMO gap.

However, the models exhibit highly divergent strategies in how they weigh these features. The GBR model shows an extreme reliance on its top two polarizability descriptors, which together account for over 67% of its total feature importance. In contrast, the Random Forest model displays a more balanced approach, giving high importance to both polarizability and molecular shape. The top-performing XGBoost model reveals another unique strategy, identifying NumRadicalElectrons as its second most important feature—a descriptor not ranked highly by any other model. This suggests XGBoost successfully leveraged a feature that is critical for a specific, yet impactful, subset of molecules. In conclusion, this comparative analysis highlights that while all models correctly identify polarizability and shape as critical, their varied performance stems from different learning strategies. The tree models, particularly GBR, fixated on the strongest individual signals, while the MLPR learned a more holistic representation by balancing structural counts with electronic properties, which is key to its superior generalization stability.

2.4.2 General performance of the MLPR model. The trained model comes with strong correlation and reasonable accuracy. The heatmap plot in Fig. 3 shows a clear positive correlation between the predicted and actual HL-gap values. The bins are clustered relatively closely around the ideal prediction line, demonstrating that the model is generally capturing the trend well. The high R^2 value of 0.961 confirms this strong correlation, indicating that the model explains 96.1% of the variance in the true HL-gaps. The MSE (0.089 eV^2),

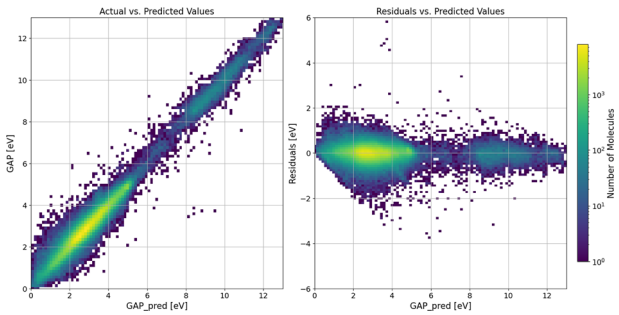

Fig. 3 The 2D histogram shows the correlation between predicted *vs.* actual HL-gaps (left) and residual plot (right) for the test set of the MLPR model. The residual plot appears to reveal heteroscedasticity, with larger errors observed for lower predicted gap values.

MAE (0.210 eV), and RMSE (0.298 eV) values are relatively low, suggesting that the model's predictions are reasonably accurate. The MAE indicates that, on average, the model's predictions are off by about 0.210 eV. The RMSE, being more sensitive to larger errors, is slightly higher at 0.298 eV, but still within a reasonable range.

Nevertheless, the model inherits weaknesses due to heteroscedasticity and potential for improvement in the lower gap region. The residual plot shows some evidence of heteroscedasticity, particularly at lower predicted values. This means that the variance of the errors is not constant across the range of predictions. The model tends to have larger errors for molecules with smaller HL-gaps. The points are more scattered in this region, indicating lower accuracy. To prove this, the metrics were re-evaluated for the HL-gap range below and above 6 eV, as it is possible the heteroscedasticity arises from the data distribution itself with 98% of the molecules having HL-gaps <6 eV. Even if the underlying error distribution is homoscedastic (constant variance), the sheer number of points in a dense region makes it more likely to observe larger errors. Table 3 clearly proves that the absolute errors (MSE, MAE, RMSE) are larger for the ≥6 eV subset. This directly contradicts the initial interpretation of the heatmap where we observed higher precision at higher HL-gaps.

This reinforces the point that was discussed above: the apparent higher precision at higher gaps in the heatmap was likely an artifact of the lower data density in that region. Even though the model makes larger absolute errors for higher gaps, there are fewer data points to show this spread, creating the illusion of tighter clustering around the diagonal. The higher R^2 for the ≥6 eV subset is misleading because R^2 is sensitive to the variance of the target variable. Since the ≥6 eV subset likely has a larger variance, it can lead to a higher R^2 even with larger absolute errors. This illustrates how R^2 does not directly reflect the accuracy of predictions but rather their relative performance in capturing the variance of the data. This analysis highlighted the importance of considering multiple metrics, particularly when interpreting visualizations like heatmaps and emphasized the need for caution when dealing with unbalanced data, where smaller data clusters can disproportionately influence visual trends.

2.4.3 MLPR performance by structural elements. In the following the correlation of specific molecular features with the prediction of HL-gaps is discussed. By creating subsets of molecules based on the presence of these features and then evaluating the metrics on those subsets, valuable insights are revealed. The metrics of the subsets will be compared to the metrics of the entire dataset (see Table 4). The associated plot can be found in the SI.

Table 3 Metrics of the GBR model split by the HL-gap at 6 eV

Range	Count	R^2	MSE [eV^2]	MAE [eV]	RMSE [eV]
<6 eV	398 003	0.917	0.064	0.183	0.254
≥6 eV	8200	0.934	0.135	0.257	0.368
All	406 203	0.947	0.089	0.210	0.298

All subsets perform worse than the full dataset, which is expected. The full model is trained on all molecules and learns to capture the combined effects of all features. Subsets, by focusing on a single feature, lose this comprehensive perspective. Some subsets perform surprisingly well. This indicates that those specific features are strong indicators of the HL-gap for molecules possessing them. NumRadicalElectrons is a clear outlier. It's very low R^2 (0.506) and high error metrics indicate it's not a good predictor of the HL-gap on its own. This is also expected as it is a very specific property not generally related to the HL-gap. The NumAliphaticHeterocycles and fr_bicyclic subsets show R^2 values very close to the full dataset (0.954 and 0.955 respectively). This suggests that the presence of aliphatic heterocycles or bicyclic structures is strongly correlated with the HL-gap, and the model captures this well. fr_NH2, fr_allylic_oxid and fr_piperdine subsets also perform relatively well (R^2 > 0.93), indicating that these functional groups also have a significant influence on the HL-gap. The subsets fr_NH and fr_esters have fair R^2 values (0.914), but the MSE, MAE, and RMSE are somewhat higher than the full dataset, suggesting that while the general trend is captured, the predictions are less precise. NumAromaticCarbocycles, fr_Al_COO and fr_ketone subsets show moderate performance (R^2 around 0.88–0.90). This indicates that while these features do influence the HL-gap, their effect is less pronounced or more complex compared to the features in the higher-performing subsets. The subsets NumAromaticHeterocycles, fr_Al_OH_noTert, fr_Ar_N, fr_para_hydroxylation, fr_aniline and fr_aryl_methyl, have the lowest R^2 values among the fragment counts (around 0.85–0.87). This suggests that these features have a weaker or more intricate relationship with the HL-gap, or that their effect is more context-dependent, meaning influenced by other parts of the molecule.

2.4.4 Analysis of prediction errors. In the following, the model's predictive accuracy is evaluated across different

Table 4 MLPR metrics of molecular subsets selected by structural units

Subset	R^2	MSE [eV^2]	MAE [eV]	RMSE [eV]
NumRadicalElectrons	0.506	0.319	0.434	0.565
NumAliphaticHeterocycles	0.954	0.068	0.190	0.262
NumAromaticCarbocycles	0.896	0.054	0.170	0.231
NumAromaticHeterocycles	0.868	0.051	0.166	0.227
fr_Al_COO	0.888	0.070	0.192	0.265
fr_Al_OH_noTert	0.852	0.101	0.229	0.318
fr_Ar_N	0.860	0.060	0.180	0.245
fr_NH0	0.914	0.072	0.194	0.268
fr_NH2	0.936	0.104	0.226	0.322
fr_allylic_oxid	0.934	0.068	0.189	0.260
fr_aniline	0.875	0.059	0.177	0.243
fr_aryl_methyl	0.868	0.046	0.156	0.214
fr_bicyclic	0.955	0.060	0.177	0.244
fr_ester	0.914	0.063	0.185	0.251
fr_ketone	0.885	0.063	0.185	0.251
fr_para_hydroxylation	0.871	0.057	0.175	0.238
fr_piperdine	0.941	0.089	0.213	0.298
Whole set	**0.947**	**0.089**	**0.210**	**0.298**

molecular subgroups by analyzing the distribution of prediction errors. The key question is to identify which molecules, particularly those containing certain functional groups, are predicted with lower accuracy and should therefore be interpreted with caution.

The heatmap (Fig. 4) displays the HL-gap error distribution mapped into ranges, providing a deeper insight into where and why larger prediction errors occur. Based on the suggestion that errors greater than 0.4 eV are likely unusable for rigorous scientific work, range quality assignments might be defined as follows.

■ Excellent precision 0.0–0.1 – errors in this range are exceptionally small and likely inconsequential for most rigorous scientific applications.

■ High precision (0.1–0.2) – errors in this range are still quite small and should be suitable for most scientific studies.

■ Acceptable precision (0.2–0.4) – errors in this range might introduce some uncertainty but are likely tolerable for many scientific investigations, particularly in complex systems and high throughput screening applications.

■ Marginal precision (0.4–0.8) – errors in this range are becoming substantial and may limit the reliability of conclusions drawn from the data. Careful consideration and potentially additional validation are necessary.

■ Low precision (0.8–1.2) – errors in this range are likely to compromise the accuracy of scientific results. This range is likely unsuitable for quantitative applications.

■ Poor precision (1.2–2.0) – errors in this range are likely to lead to unreliable or misleading results. Significant improvements in model accuracy are needed for this range to be useful.

■ Negligible precision (2.0–10.0) – errors in this range are so large that the data is essentially unusable for any scientific purpose.

For most molecular groups, the majority of molecules fall within the “Excellent Precision” and “High Precision” ranges. This suggests that the model performs reasonably well overall. A noticeable variation in the distribution of errors is observed across different molecular groups. Some groups have a higher proportion of molecules in the “Acceptable Precision” range and beyond, indicating potential challenges for specific chemical functionalities. The fr_Al_COO, fr_NH2 and fr_Al_OH_noTert groups appear to have a relatively higher proportion of molecules in the “Marginal Precision” range and beyond, signifying that predictions for molecules containing these groups might be less reliable. Analysis of the HL-gap prediction model revealed a notable trend. Molecular groups with smaller representation in the dataset tended to exhibit poorer predictive performance. This observation is not coincidental but rather reflects the influence of sample size on model accuracy and robustness. Smaller molecular groups suffer from reduced statistical power, limiting the model's ability to discern true relationships between specific chemical features and HL-gap values. This limitation arises from the increased susceptibility of smaller groups to noise, random variations, and the disproportionate impact of outliers. The poorer performance observed for smaller groups does not necessarily indicate a weaker or more complex relationship between their chemical features and HL-gap. Instead, it may reflect the model's inability to effectively capture these relationships due to data scarcity.

The subsets consistently performing the best, with combined percentages up to “Acceptable Precision” above 90%.

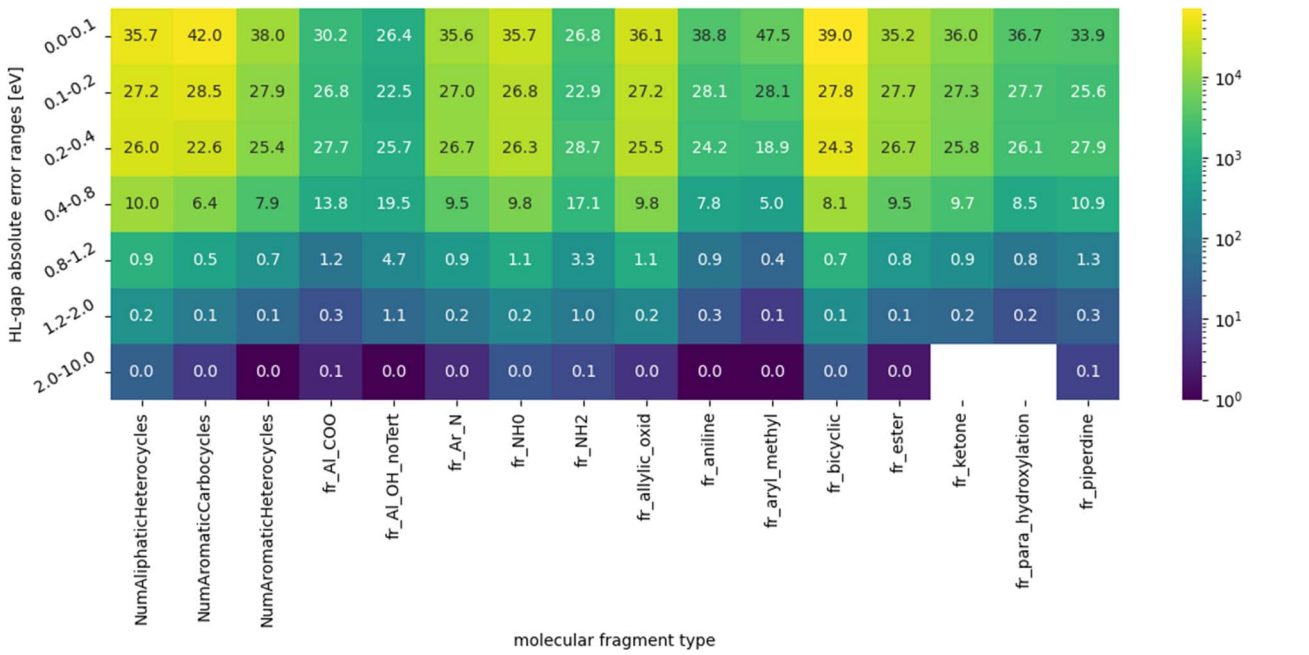

Fig. 4 The heatmap displays the distribution of prediction errors for different molecular subsets, defined by the presence of specific structural features (*x*-axis), across a range of absolute errors in the HL-gap (*y*-axis). The numbers within each cell represent the percentage of molecules from a given subset that fall within a specific error range, providing a normalized measure of the model's predictive accuracy for that subset. The color intensity reflects the total number of molecules in a certain range and subset.

With 94.55% fr_aryl_methyl is the best performing subgroup overall followed by NumAromaticCarbocycles (93.08%), implying excellent overall performance. Subsequently, NumAromaticHeterocycles (91.31%), fr_aniline (91.08%), fr_bicyclic (91.12%) and fr_para_hydroxylation (90.47%) demonstrate strong predictive capabilities. Subsets with combined percentages in the high 80 s, indicating good but slightly less precise predictions. Those include NumAliphaticHeterocycles (88.88%), fr_Ar_N (89.32%), fr_NH0 (88.88%), fr_allylic_oxid (88.84%), fr_ester (89.61%), fr_ketone (89.21%) and fr_piperdine (87.46%). While still a reasonable performance, with 84.69% fr_Al_COO it is noticeably lower than the top performers. The subgroup fr_NH2 (78.38%) shows a lower combined percentage compared to most other groups, suggesting potential challenges in accurate prediction. The fr_Al_OH_noTert (74.66%) group stands out as having the lowest combined percentage, indicating that the model might struggle with this specific functional group. However, the majority of molecules that is in the range of low and poor precision refers to molecules with a complicated electronic structure, like having ionic or radical character or having multiple functional groups, both donors and acceptors or multiple heteroatoms up 3rd and 4th period non-metals or metalloids. Example images of molecules can be found in the git repository.

2.4.5 Performance in the context of published work. To rigorously contextualize the contributions of this study, it is essential to benchmark our model's performance against the extensive body of published work in molecular property prediction. Our MLPR model, trained on the COCONUT database with GFN2-xTB calculated HOMO–LUMO gaps as the target property, achieved a test set MAE of 0.210 ± 0.001 eV and RMSE of 0.298 ± 0.002 eV. A systematic comparison of these results with the literature, focusing on methodology, feature representation, dataset characteristics, and target properties, reveals the specific contributions and positioning of our work.

The performance of our model is highly consistent with other studies that have utilized similar descriptor-based machine learning approaches on large-scale molecular datasets.[7,9,12,13,19,46] Notably, Pereira *et al.*[9] reported a very similar MAE of 0.21 eV and RMSE of 0.30 eV using Random Forest and MLPR models on over 111 000 organic molecules with DFT-calculated properties. Our MAE is also comparable to the 0.19 eV achieved by Xu *et al.*[22] using a linear model on Polycyclic Aromatic Hydrocarbons (PAHs). Furthermore, our RMSE is more favorable than the 0.36–0.43 eV range reported by Nakata *et al.*[12] using SVM and Ridge Regression on a subset of the PubChemQC[12] database. These comparable error metrics suggest that our model achieves a robust and expected level of performance for its methodological class.

The current state-of-the-art in this field, however, is dominated by deep learning models, particularly GNNs, that learn features directly from molecular topology and 3D coordinates. These models consistently achieve significantly lower prediction errors. Seminal works on the QM9 benchmark dataset established the high performance of these methods, with models such as SchNet,[1,15,16] MPNNs,[8,14,15] MEGNet,[2,26] and other Graph Convolutional Networks reporting MAEs for the HOMO–LUMO gap in the remarkably low range of 0.06–0.09 eV.[1,2,4,8,14,15] This superior accuracy has been replicated by advanced architectures like PaiNN, which reported an exceptional MAE of just 0.01 eV on the Harvard organic photovoltaic dataset[47] (HOPV) dataset.

The primary factors driving this performance disparity are the feature representation and the dataset characteristics. Our work employs traditional ML models that rely on pre-calculated 2D molecular descriptors. This methodological choice is shared by several studies reporting similar error magnitudes.[7,9,12] In contrast, the highest-performing models are overwhelmingly GNNs that learn richer, tailored feature representations directly from the 3D molecular graph, using atomic types and Cartesian coordinates as inputs. This end-to-end learning allows the model to capture more nuanced and relevant structural information than is possible with predefined 2D descriptors.

Furthermore, our study tackles the COCONUT database, a large-scale collection of over 400 000 structurally diverse and complex natural products. This presents a significant learning challenge compared to the benchmark QM9 dataset, which consists of ∼134 000 smaller, less complex organic molecules and is the basis for many of the lowest reported errors.[2,4,5,7,14–16,20] Many other high-performance models are trained on smaller, chemically homogeneous datasets focused on specific molecular classes.[1,11,17–19] The structural complexity and diversity inherent in our natural product dataset likely establish a higher error floor. The challenges of complex datasets are highlighted by Deng *et al.*,[3] where a GNN approach on conjugated oligomers still resulted in a high MAE of 0.54 eV.

2.4.6 Validation against public DFT benchmarks. To rigorously address the need for validation and to test the trained MLPR model's ability to generalize to unseen chemical structures, a comprehensive external validation was performed. The model, which was trained exclusively on GFN2-xTB gaps from the COCONUT dataset, was used to predict the HL-gaps for the unseen part of the QM9 (ref. 55 and 56) dataset (∼133 000 molecules). This dataset serves as a true "out-of-distribution" test set, as its molecules were not part of the training data. The model's predictions were then compared against two distinct, higher-precision quantum chemical benchmarks provided by the QM9 dataset. These benchmarks are: first, the DFT[55] gaps, calculated at the B3LYP/6-31G(2df,p) level of theory, and second, the higher-accuracy Quasi-Particle (QP) GW[56] gaps, which are considered a more rigorous "gold standard" for electronic gaps. The results of this external validation are presented in Fig. 5.

The comparison to the DFT reference (Fig. 5, left) demonstrates the model's successful generalization. The model's predictions show a strong correlation with the DFT values and, crucially, reproduce the distinct two-cluster structure of the data, which separates saturated (high-gap) from conjugated (low-gap) systems. This confirms that the model, using only 2D descriptors, has learned the fundamental structure–property relationships governing the HL-gap. As the MLPR model was trained on xTB data, its predictions carry the known systematic bias of that method, resulting in a mean underestimation of the DFT gaps by 3.94 eV.

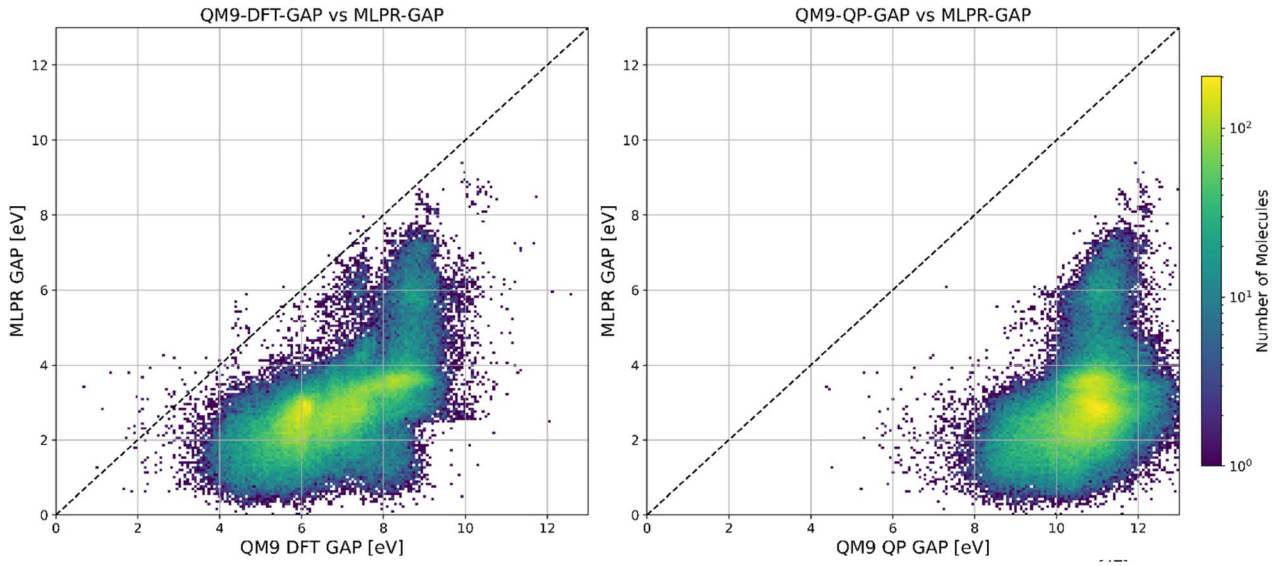

Fig. 5 Scatter plots demonstrating the generalization of the MLPR model on the external QM9 dataset. The model was trained only on GFN2-xTB data. The *y*-axis in both plots shows the MLPR's predictions. The *x*-axis shows the reference "true" gaps from QM9 at two different levels of theory: (left) DFT (B3LYP/6-31G(2df,p)) and (right) GW.

The comparison to the higher-level GW reference (Fig. 5, right) provides a more complete picture. As expected, the model's predictions show a larger systematic underestimation of 7.75 eV relative to the GW gaps. However, it is critical to note that it is a well-established fact that DFT itself (particularly with the B3LYP functional) significantly underestimates the more sophisticated GW gap. Therefore, this large deviation is not a failure of the model; rather, it correctly reflects the combined, systematic underestimation of both the GFN2-xTB training data and the DFT benchmark relative to the GW standard.

2.4.7 Model improvements. While this study demonstrates the successful application of descriptor-based machine learning, a promising avenue for future improvement is the implementation of GNN architectures. The current state-of-the-art in molecular property prediction is dominated by models such as SchNet,[1,15,16] MPNNs,[8,14,15] MEGNet[2,26] and PaiNN,[47] which learn features directly from the 3D molecular graph rather than relying on pre-calculated descriptors. These methods have achieved exceptionally low prediction errors (MAE < 0.1 eV) and represent the next logical step for enhancing predictive accuracy. Therefore, testing additional descriptor-based models like Random Forest or XGBoost is unlikely to yield fundamentally new insights, as they operate on the same feature space. Future work will focus on developing a GNN-based pipeline to investigate if this methodological shift can overcome the challenges our current models face with molecules possessing complex electronic structures, thereby pushing the boundaries of predictive accuracy for large-scale natural product databases.

In parallel with exploring new architectures, several refinements could enhance the current modeling framework. A deeper investigation into the chemical structures of molecules with high prediction errors—particularly for the challenging functional groups identified in this study, such as aliphatic carboxylic acids, alcohols, and amines—could reveal specific structural motifs or electronic interactions that the current descriptors fail to capture. The observation that smaller, under-represented molecular groups exhibited poorer predictive performance underscores the need to address data scarcity. Future work should prioritize strategies such as targeted data augmentation techniques, gathering more data for these groups, or employing weighting schemes during regression to account for potential biases in the training data. A primary strategy is the use of deep generative models,[19] such as Variational Autoencoders (VAEs)[48,49] or Generative Adversarial Networks (GANs),[49] which can learn to produce novel, yet chemically valid, molecular structures within a specific chemical domain. This approach would allow for the targeted generation of new molecules belonging to the poorly predicted classes, directly enriching and balancing the training set. Alternatively, data augmentation can be performed in the descriptor space. Techniques like the Synthetic Minority Oversampling Technique[50] (SMOTE) and its variants[50] have been successfully adapted for QSAR datasets, where they create synthetic minority class samples by interpolating between existing data points in the high-dimensional feature space.

To make data acquisition more efficient, an active learning[51–53] loop represents another promising direction. In this paradigm, the model's own uncertainty estimates are used to intelligently select the most informative molecules for which to perform expensive quantum chemical calculations. This ensures that computational resources are focused on the areas of chemical space where the model would benefit most from new information. These established strategies, from generative models to active learning, provide a clear and feasible path toward significantly improving model robustness and predictive accuracy for the challenging molecular subgroups identified in this work.

Other avenues for improvement include exploring novel descriptor selection strategies, incorporating domain knowledge through expert-curated features, and adjusting the computational workflow to incorporate higher-precision quantum chemical methods for the target property, which would enable enhanced reliability and practicability of the findings based on the semi-empirical xTB data. A small but structurally diverse subset of molecules, particularly those where the current model shows high error or those belonging to challenging chemical groups, could be re-evaluated using DFT or the recently developed general-purpose Extended Tight-Binding[54] (g-xTB). Comparing the ML model's predictions not only to the xTB target values but also to these more accurate DFT-level results would serve two key purposes. First, it would provide a valuable cross-check on the physical trends identified by the model. Second, it would help to disentangle the model's prediction error from the inherent error of the underlying semi-empirical method. This would provide a more robust assessment of the model's performance and its applicability for practical high-throughput screening campaigns.

3 Conclusions

This study successfully developed a high-throughput, machine learning-based approach for predicting the HL-gap of natural products, addressing the computational expense and time limitations of traditional quantum mechanical methods like DFT when applied to large datasets of molecules. Utilizing a curated dataset of over 400 000 molecules from the COCONUT database and a streamlined computational workflow, the efficacy of combining xTB calculations with advanced machine learning algorithms was demonstrated. The findings highlight the critical role of molecular polarizability, specifically SMR_VSA descriptors, in determining the HL-gap in both models. All tested machine learning models, including GBR, MLPR, XGB, and RFR, achieved good overall predictive performance, though the MLPR model showed a slight advantage in generalization ability. A comprehensive external validation confirmed this, as the MLPR model successfully predicted gaps for the, QM9 dataset, with its predictions faithfully capturing the underlying chemical trends and the known systematic biases of its training method when compared to DFT and GW benchmarks. Challenges remain in accurately predicting HL-gaps for molecules containing multiple functional groups, notably aliphatic carboxylic acids, alcohols, and amines. Analysis of feature importance and performance across molecular subsets revealed that aromatic carbocycles and polarizability are strong predictors of the HL-gap, while the presence of multiple interacting functional groups or complex electronic structures often leads to reduced accuracy. These observations underscore the importance of considering both electronic and structural features in HL-gap modeling and suggest that further model refinement, particularly in addressing the complexities of specific functional group interactions and complex electronic structures, holds significant promise for future improvements in the predictive accuracy. This study therefore contributes a reliable, high-throughput methodology and provides a quantitative performance baseline, paving the way for future large-scale screening of electronic properties in the vast and biomedically important chemical space of natural products.

Author contributions

Sascha Thinius: conceptualization, methodology, software, data curation, formal analysis, investigation, visualization, writing – original draft, writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

The source code developed for this study, including the Common Workflow Language (CWL) definitions and analysis scripts, along with the full dataset of calculated HOMO–LUMO gaps and molecular descriptors, are openly available. The specific version used in this publication is v0.2.1. The development repository is hosted on GitHub at: **https://github.com/sthinius87/HL-gaps-pub**. A persistent, archived version of the software and data (v0.2.1) is available on Zenodo under the Digital Object Identifier (DOI): **https://doi.org/10.5281/zenodo.15113790**. The original molecular structures were sourced from the publicly available COCONUT database. Further details, including the final feature list and optimized model hyperparameters, are available in the supplementary information (SI) document accompanying this article. All code and data are distributed under the MIT License.

Supplementary information: parameters for the ML-Model and additional plots referenced in the actual paper. See DOI: **https://doi.org/10.1039/d5dd00186b**.

Acknowledgements

I express my profound gratitude to the Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. for their financial contribution towards the publication of this manuscript.

References

1 C.-K. Lee, C. Lu, Y. Yu, Q. Sun, C.-Y. Hsieh, S. Zhang, Q. Liu and L. Shi, *J. Chem. Phys.*, 2021, **154**, 24906.
2 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564.
3 S. Deng, J. X. Ng and S. Li, *Mol. Syst. Des. Eng.*, 2025, **10**, 413.
4 T. Kirschbaum and A. Bande, *AIP Adv.*, 2024, **14**(10), 105119.
5 B. Mazouin, A. A. Schöpfer and O. A. von Lilienfeld, *Mater. Adv.*, 2022, **3**, 8306.
6 G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller and O. Anatole von Lilienfeld, *New J. Phys.*, 2013, **15**, 95003.
7 E. O. Pyzer-Knapp, K. Li and A. Aspuru-Guzik, *Adv. Funct. Mater.*, 2015, **25**, 6495.

8 A. Su, X. Zhang, C. Zhang, D. Ding, Y.-F. Yang, K. Wang and Y.-B. She, *Phys. Chem. Chem. Phys.*, 2023, **25**, 10536.
9 F. Pereira, K. Xiao, D. A. R. S. Latino, C. Wu, Q. Zhang and J. Aires-de-Sousa, *J. Chem. Inf. Model.*, 2017, **57**, 11.
10 J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller and A. Tkatchenko, *Chem. Rev.*, 2021, **121**, 9816.
11 P. B. Jørgensen, M. Mesta, S. Shil, J. M. García Lastra, K. W. Jacobsen, K. S. Thygesen and M. N. Schmidt, *J. Chem. Phys.*, 2018, **148**, 241735.
12 M. Nakata and T. Shimazaki, *J. Chem. Inf. Model.*, 2017, **57**, 1300.
13 B. Kang, C. Seok and J. Lee, *J. Chem. Inf. Model.*, 2020, **60**, 5984.
14 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, Neural Message Passing for Quantum Chemistry, *arXiv*, 2017, preprint, arXiv:1704.01212v2, DOI: **10.48550/arXiv.1704.01212**.
15 C. Lu, Q. Liu, Q. Sun, C.-Y. Hsieh, S. Zhang, L. Shi and C.-K. Lee, *J. Phys. Chem. C*, 2020, **124**, 7048.
16 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, **148**, 241722.
17 B. Liu, Y. Yan and M. Liu, *Nanoscale*, 2025, **17**, 7865.
18 T. Zhang, J. Yuk Lin Lai, M. Shi, Q. Li, C. Zhang and H. Yan, *Adv. Sci.*, 2024, **11**, e2308652.
19 Q. Yuan, A. Santana-Bonilla, M. A. Zwijnenburg and K. E. Jelfs, *Nanoscale*, 2020, **12**, 6744.
20 F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2017, **13**, 5255.
21 K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari and P. Rinke, *Adv. Sci.*, 2019, **6**, 1801367.
22 Y. Xu, Q. Chu, D. Chen and A. Fuentes, *Front. Mech. Eng.*, 2021, **7**, 744001.
23 A. Mohamed, D. P. Visco, K. Breimaier and D. M. Bastidas, *ACS Omega*, 2025, **10**, 2799.
24 J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, 2019, **5**, 83.
25 P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, *et al.*, *Commun. Mater.*, 2022, **3**, 93.
26 C. Chen, Y. Zuo, W. Ye, X. Li and S. P. Ong, *Nat. Comput. Sci.*, 2021, **1**, 46.
27 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl in *Proceedings of Machine Learning Research*, ed. D. Precup and Y. W. Teh, PMLR, 2017, pp. 1263–1272.
28 T. Luo, W. T. Tang, M. K. F. Lee, C. Qu, W.-F. Wong, R. Goh: *Energy-efficient Inference with Dendrite Tree Inspired Neural Networks for Edge Vision Applications*, 2021, **https://arxiv.org/abs/2105.11848**.
29 D. Padula, J. D. Simpson and A. Troisi, *Mater. Horiz.*, 2019, **6**, 343.
30 M. Sorokina and C. Steinbeck, *COCONUT: the COlleCtion of Open NatUral producTs*, Zenodo, 2021.
31 M. R. Crusoe, S. Abeln, A. Iosup, P. Amstutz, J. Chilton, N. Tijanić, H. Ménager, S. Soiland-Reyes, B. Gavrilović, C. Goble, *et al.*, *Commun. ACM*, 2022, **65**, 54.
32 P. Amstutz, M. R. Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer, A. Kartashov, D. Leehr, H. Ménager, M. Nedeljkovich, *et al.*, *Common Workflow Language, v1.0*, figshare, 2016.
33 F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, *et al.*, *F1000Research*, 2021, **10**, 33.
34 S. P. Huber, S. Zoupanos, M. Uhrin, L. Talirz, L. Kahle, R. Häuselmann, D. Gresch, T. Müller, A. V. Yakutovich, C. W. Andersen, *et al.*, *Sci. Data*, 2020, **7**, 300.
35 H. E. Bal, J. G. Steiner and A. S. Tanenbaum, *ACM Comput. Surv.*, 1989, **21**, 261.
36 G. Landrum, P. Tosco, B. Kelley, Ric, D. Cosgrove, sriniker, gedeck, R. Vianello, NadineSchneider, E. Kawashima, *et al.*, *rdkit/rdkit: 2023_09_3 (Q3 2023) Release*, Zenodo, 2023.
37 The pandas development team, *pandas-dev/pandas: Pandas*, Zenodo, 2023.
38 A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, *et al.*, *J. Phys.:Condens. Matter*, 2017, **29**, 273002.
39 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652.
40 sthinius87, *sthinius87/HL-gaps-pub: HL-gaps v0.2.1*, Zenodo, 2025.
41 D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 2014.
42 P. Labute, *J. Mol. Graphics Modell.*, 2000, **18**, 464.
43 L. H. Hall, L. B. Kier in *Reviews in Computational Chemistry*, ed. K. B. Lipkowitz and D. B. Boyd, Wiley, 1991, pp. 367–422.
44 T. M. Martin, P. Harten, R. Venkatapathy, S. Das and D. M. Young, *Toxicol. Mech. Methods*, 2008, **18**, 251.
45 *Reviews in Computational Chemistry*, ed. K. B. Lipkowitz and D. B. Boyd, Wiley, 1991.
46 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2015, **11**, 2087.
47 S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann and A. Aspuru-Guzik, *Sci. Data*, 2016, **3**, 160086.
48 D. Yoon, J. Lee and S. Lee, *Appl. Sci.*, 2025, **15**, 3640.
49 Y. Bian and X.-Q. Xie, *J. Mol. Model.*, 2021, **27**, 71.
50 J. Jiang, C. Zhang, L. Ke, N. Hayes, Y. Zhu, H. Qiu, B. Zhang, T. Zhou and G.-W. Wei, *Chem. Sci.*, 2025, **16**, 7637.
51 L. Wang, Z. Zhou, X. Yang, S. Shi, X. Zeng and D. Cao, *Drug Discovery Today*, 2024, **29**, 103985.
52 M. Bailey, S. Moayedpour, R. Li, A. Corrochano-Navarro, A. Kötter, L. Kogler-Anele, S. Riahi, C. Grebner, G. Hessler and H. Matter, *eLife*, 2023, **12**, RP89679.
53 M. A. Masood, S. Kaski and T. Cui, *J. Cheminf.*, 2025, **17**, 58.
54 T. Froitzheim, M. Müller, A. Hansen and S. Grimme, *ChemRxiv*, 2025, preprint, DOI: **10.26434/chemrxiv-2025-bjxvt**.
55 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
56 A. Fediai, P. Reiser, J. E. O. Peña, *et al.*, *Sci. Data*, 2023, **10**, 581.