HOMO-LUMO Gap Prediction from Molecular Structure: A Comparative Machine Learning Study on the COMPAS-2D Dataset of Heteroaromatic Systems

Mohsin Raza*

FAU Erlangen-Nürnberg

E-mail: mohsin.raza@fau.de

Abstract

The HOMO-LUMO gap—the energy difference between the Highest Occupied and Lowest Unoccupied Molecular Orbitals—is a critical quantity governing the optoelectronic behaviour of organic semiconductors and is conventionally computed via Density Functional Theory (DFT), which incurs O(N^3) scaling cost. This work investigates whether the gap can be reliably predicted from two-dimensional molecular structure alone using machine learning. Seven regression models—Linear Regression, Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, and XGBoost—are trained on the COMPAS-2D dataset[1] of 52,000 neutral cata-condensed heteroaromatic molecules using RDKit molecular descriptors[2] and 1024-bit Morgan circular fingerprints[3] as features. Three data-sampling strategies (random, gap-stratified, and boron-balanced) are evaluated alongside Murcko scaffold-based train/test splitting[4] to enforce generalisation to structurally novel scaffolds. Gradient Boosting achieves the best test performance ($R^2 = 0.752$, MAE $= 0.266$ eV), followed closely by XGBoost ($R^2 = 0.749$). Boron atom count is identified as the dominant predictor across all ensemble models, a finding grounded in boron's electron-deficient character and its role in compressing the HOMO-LUMO gap. Controlled boron-only and boron-free experiments disentangle the boron classification signal from genuine structure-property learning. Outlier analysis reveals that prediction failures are concentrated in molecules bearing rare Morgan fingerprint substructures absent from the training set, suggesting targeted data augmentation as the most efficient path to further improvement. Multi-seed robustness validation ($\sigma_{R^2} \leq 0.02$) confirms that model rankings are reproducible and not artefacts of a particular train/test split.

Introduction

The frontier molecular orbital gap—commonly termed the HOMO-LUMO gap—quantifies the energy required to excite an electron from the highest occupied molecular orbital (HOMO) into the lowest unoccupied molecular orbital (LUMO). In the language of solid-state physics, this corresponds to the optical band gap at the single-molecule level. The gap governs a molecule's optical absorption wavelength, its propensity to donate or accept electrons in charge-transfer processes, its electrochemical redox window, and its intrinsic conductivity. Consequently, reliable gap prediction is of immediate relevance to the rational design of organic photovoltaic absorbers, organic light-emitting diode (OLED) emitters, non-linear optical materials, and n- or p-type organic semiconductors.[5]

High-fidelity calculation of the HOMO-LUMO gap is typically performed using Density Functional Theory (DFT). While DFT with appropriate hybrid functionals achieves reasonable accuracy (mean absolute error $\lesssim$0.2 eV relative to higher-level methods), it scales as $O(N^3)$ with the number of electrons and becomes computationally prohibitive when screening molecular libraries containing tens or hundreds of thousands of candidates. The COMPAS-2D dataset,[1] which contains DFT-computed gap values for 156,000 heteroaromatic molecules across multiple charge states, exemplifies the scale at which DFT screening becomes a bottleneck.

Machine learning (ML) offers a complementary strategy: train a model on existing DFT data and use it to rapidly screen new candidates at negligible inference cost.[6] For this to be scientifically meaningful rather than a statistical exercise, the ML model must learn from structural features that are genuinely connected to the underlying quantum chemistry—not from correlates that happen to be predictive in a particular dataset but are physically uninformative.

The specific objectives of this work are: (i) to compare seven regression models spanning linear and tree-based ensemble families on the HOMO-LUMO prediction task; (ii) to evaluate the impact of data sampling strategy on model quality; (iii) to conduct controlled experiments disentangling the contribution of boron-containing molecules; (iv) to interpret model predictions through the lens of frontier orbital chemistry; and (v) to assess generalisation under Murcko scaffold splitting,[4] which enforces evaluation on genuinely novel molecular scaffolds.

Related Work

Machine learning for molecular property prediction has developed rapidly over the past decade. Early approaches relied on hand-crafted molecular descriptors and fingerprints as input to classical models; more recent work has introduced graph neural networks (GNNs) that operate directly on the molecular graph.[5,7]

For HOMO-LUMO gap prediction specifically, the QM9 dataset has served as a standard benchmark, with deep learning models achieving sub-0.05 eV mean absolute error.[8] However, QM9 contains small organic molecules (up to nine heavy atoms), whereas the COMPAS-2D dataset[1] targets medium-to-large polycyclic heteroaromatic systems—a structurally distinct chemical space in which extended conjugation, ring topology, and heteroatom placement interact in complex ways.

Relevant prior work on heteroaromatic systems includes studies using extended-connectivity fingerprints (ECFPs, equivalent to Morgan fingerprints[3] with specified radius) and gradient-boosted models,[9] which have generally outperformed linear regression baselines on correlated, high-dimensional molecular feature matrices. The utility of scaffold-based train/test splitting as a more honest generalisation estimate than random splitting has been established in computational drug discovery and has since been adopted in materials informatics contexts.[10]

The COMPAS family of datasets was specifically constructed to facilitate ML benchmarking on conjugated systems with DFT-computed properties,[1] making it an ideal testbed for evaluating structure-to-property learning without confounding effects from different levels of theory.

Methodology

Dataset and Preprocessing

The COMPAS-2D dataset[1] contains 156,000 entries for cata-condensed heteroaromatic molecules spanning neutral, cationic, and anionic charge states, with atoms drawn from B, C, N, O, and S. All quantum-chemical properties including HOMO energy, LUMO energy, adiabatic ionisation potential (AIP), adiabatic electron affinity (AEA), and the target quantity—the HOMO-LUMO gap in eV—were computed at the DFT level. Filtering to retain only neutral (charge = 0) molecules with valid gap values yields 52,000 molecules, which constitute the working population for all experiments.

Data leakage prevention is critical in this dataset because DFT-computed columns are present alongside structural features in the same table. Including `homo`, `lumo`, `aip`, `aea`, or related DFT quantities as model inputs would trivially solve the task via the identity $\Delta\varepsilon = \varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}}$ without learning any structure-property relationship. All quantum-chemical output columns are therefore excluded. The model is trained exclusively on (i) structural annotations derivable from the molecular graph (atom and ring counts), (ii) RDKit-computed molecular descriptors,[2] and (iii) Morgan fingerprints[3]—all computed independently from SMILES strings.

Sampling Strategies

Working with all 52,000 molecules is computationally feasible but unnecessary for method comparison; a 5,000-molecule working subset is selected under three protocols:

Strategy A (Random): Uniform random selection without stratification. This represents a naïve baseline and oversamples the dense central region of the gap distribution.

Strategy B (Stratified): The gap range is divided into 10 quantile-based bins, and molecules are sampled proportionally from each bin. This guarantees representation of the full gap range, including rare large-gap and small-gap molecules. Strategy B is the primary dataset for hyperparameter tuning.

Strategy C (Boron-balanced): Exactly 2,500 boron-containing and 2,500 boron-free molecules, each group independently gap-stratified. Since boron presence is a dominant predictor, the natural dataset imbalance (≈55% boron-containing) could allow models to partially classify by boron presence. Strategy C eliminates this imbalance to probe whether structural signals persist under balanced conditions.

Feature Engineering

Two complementary feature sets are concatenated into a 1,264-dimensional design matrix:

RDKit Molecular Descriptors (≈200 features):[2] Physicochemical and topological properties derived from the 2D molecular graph, including molecular weight, logP, topological polar surface area (TPSA), ring counts, aromatic ring counts, hydrogen bond donors and acceptors, and topological connectivity indices (Chi series, Kappa series, Balaban J). These descriptors capture global molecular properties interpretable in chemical terms.

Morgan Circular Fingerprints (1,024 bits, radius 2):[3] Each bit encodes the presence or absence of a specific circular substructural fragment extending up to 2 bonds from each atom. At radius 2, a single bit may represent a nitrogen atom within a fused aromatic ring environment—exactly the kind of local chemical context that determines heteroatom contribution to frontier orbital energy.

Train/Test Splitting: Murcko Scaffold Split

Standard random splitting is inappropriate for molecular machine learning when multiple molecules share the same core ring system. Many COMPAS-2D molecules are systematically generated by heteroatom substitution into the same polycyclic scaffold; random splitting distributes near-identical molecules across train and test, inflating performance metrics.

Murcko scaffold decomposition[4] extracts the core ring system from each molecule. All molecules sharing the same scaffold are assigned entirely to either the training set or the test set, never both. This enforces generalisation to structurally novel scaffolds—a realistic deployment scenario for screening. Zero scaffold overlap between train and test is verified programmatically after every split. An 80/20 train/test ratio is applied for all experiments.

Models and Hyperparameter Tuning

Seven regression models are evaluated: Linear Regression, Ridge (ℓ_2 regularisation), Lasso (ℓ_1 regularisation), ElasticNet (combined ℓ_1/ℓ_2), Random Forest, Gradient Boosting, and XGBoost.[9]

Every model is wrapped in an identical preprocessing pipeline consisting of three stages: median imputation (`SimpleImputer`), robust scaling (`RobustScaler`), and the model itself.[11] The imputer and scaler are fit exclusively on training data and applied to the test set, preventing any information leakage through normalisation. Uniform preprocessing across all seven models ensures that performance differences reflect model architecture rather than preprocessing asymmetry.

Hyperparameter optimisation uses `RandomizedSearchCV` with a fixed budget of $n_{\text{iter}} = 30$ and cv $= 3$, applied identically to all models. This equal budget is a deliberate design choice: while tree models have more hyperparameters than linear models, providing them with proportionally more iterations would constitute differential treatment. The uniform budget represents a clean, defensible baseline; search spaces are informed by empirical best practices for tabular molecular data.

Evaluation Metrics

Model quality is reported using the coefficient of determination (R^2), mean absolute error (MAE in eV), and root mean squared error (RMSE in eV) on the held-out test set. Additionally, the overfit gap (Train R^2− Test R^2) quantifies the degree of overfitting. Multi-seed robustness is assessed across seeds 42, 123, and 777.

Results and Discussion

Exploratory Data Analysis

The HOMO-LUMO gap spans approximately 1.8–7.9 eV across the 52,000 neutral molecules, with a mean of 4.76 eV and standard deviation of 0.68 eV (Figure 1). The distribution is approximately unimodal but with a visible shoulder at lower gap values, attributable to boron-containing molecules (mean gap 4.51 eV versus 5.07 eV for non-boron, with boron comprising 55.1% of the dataset). This separation reflects the electron-deficiency of boron in aromatic systems: the empty boron p-orbital significantly lowers the LUMO energy while leaving the HOMO relatively unchanged, compressing the gap.

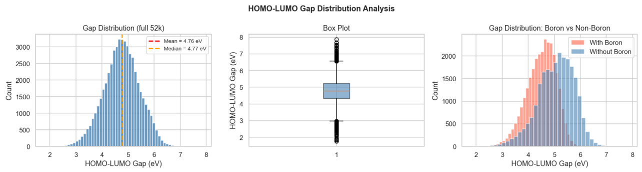

Figure 1: Target variable analysis. *Left:* Distribution of HOMO-LUMO gaps across 52,000 neutral COMPAS-2D molecules, showing a near-Gaussian distribution centred at 4.76 eV. *Centre:* Box plots comparing gap distributions for boron-containing versus boron-free molecules, with boron molecules exhibiting a lower mean gap (4.51 eV vs. 5.07 eV) due to boron's electron-deficient character. *Right:* Cumulative distribution of gap values.

Inspection of correlations among the DFT-computed quantum chemical properties (Figure 2) reveals that the HOMO energy and adiabatic ionisation potential (AIP) are almost perfectly anti-correlated ($r = -0.957$). This is chemically expected: both quantities describe the same physical process—removing an electron from the HOMO—expressed with opposite sign conventions. Similarly, LUMO energy and adiabatic electron affinity (AEA) are strongly correlated ($r = +0.871$), as both describe electron addition into the LUMO. These strong correlations serve as a data quality check and confirm that the DFT calculations are internally consistent. Critically, they also motivate the feature engineering decisions: the molecular descriptor space will exhibit analogous multicollinearity (molecular weight $\leftrightarrow$ ring count $\leftrightarrow$ conjugation length $\leftrightarrow$ gap), which has direct implications for linear versus tree-based model performance.

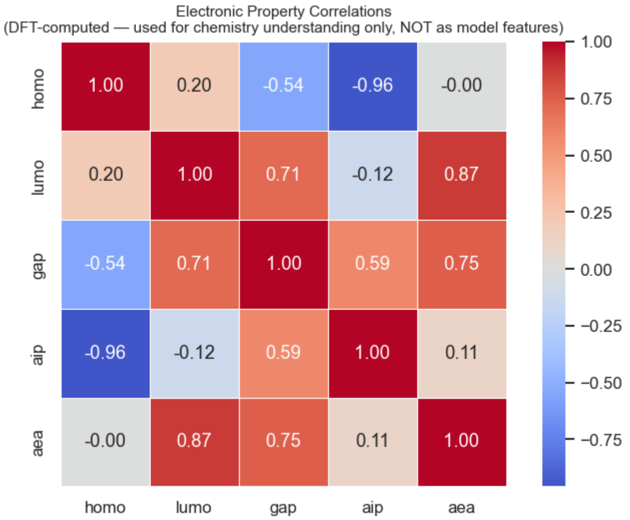

Figure 2: Pearson correlation heatmap of DFT-computed quantum chemical properties. The near-perfect HOMO ↔ AIP anti-correlation ($r = -0.957$) and strong LUMO ↔ AEA correlation ($r = +0.871$) confirm physical consistency and illustrate the multicollinearity structure present in molecular feature spaces.

Sampling Strategy Distributions

All three 5,000-molecule working subsets reproduce the full-population mean gap within 0.03 eV (Figure 3). Strategy A slightly oversamples the mode of the distribution; Strategy B enforces uniform coverage across 10 quantile bins; Strategy C shifts the boron fraction to exactly 50%. These distributional differences provide the mechanistic basis for comparing model performance across strategies.

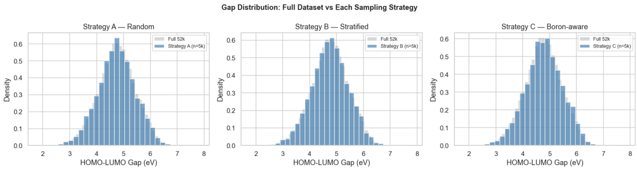

Figure 3: Gap distribution histograms for the three 5,000-molecule sampling strategies. Strategy A (random) oversamples the distribution mode; Strategy B (stratified) enforces proportional coverage across 10 quantile bins; Strategy C (boron-balanced) enforces 50% boron / 50% non-boron composition with independent gap stratification.

Baseline Model Performance

Before hyperparameter tuning, all seven models are trained with default parameters on all three strategies (Figure 4). The key observations are: (i) tree-based models (Gradient Boosting, XGBoost, Random Forest) consistently outperform linear models; (ii) Strategy B produces higher test R^2 than Strategy A for six of seven models; and (iii) Lasso and ElasticNet perform surprisingly poorly ($R^2 \approx 0.33$–0.39) with default regularisation strength, which is excessive for this high-dimensional feature space and shrinks most coefficients to near-zero.

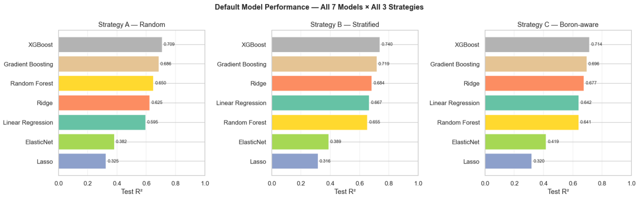

Figure 4: Baseline (pre-tuning) test R^2 for all seven models across the three sampling strategies. Tree-based models consistently outperform linear methods; Strategy B (stratified) generally yields higher R^2 than Strategy A (random).

Tuned Model Performance

After hyperparameter tuning on Strategy B, performance improves substantially for most models, with particularly large gains for regularised linear models whose default regularisation was excessive (Table 1). Gradient Boosting achieves the best test $R^2 = 0.752$ (MAE $= 0.266$ eV), followed by XGBoost at $R^2 = 0.749$ (MAE $= 0.270$ eV). Among linear models, Ridge, Lasso, and ElasticNet converge to similar performance ($R^2 \approx 0.70$–0.71) once properly tuned.

Table 1: Tuned model performance on Strategy B test set (scaffold split). Train R^2 is reported to quantify overfitting (Overfit = Train R^2- Test R^2).

Model	Test R^2	MAE (eV)	RMSE (eV)	Train R^2	Overfit
Grad. Boost.	**0.752**	**0.266**	**0.352**	0.887	0.135
XGBoost	0.749	0.270	0.354	0.863	0.114
Ridge	0.712	0.293	0.379	0.774	0.062
ElasticNet	0.707	0.294	0.381	0.753	0.046
Lasso	0.705	0.296	0.383	0.755	0.050
Linear Regr.	0.667	0.312	0.407	0.819	0.152
Random Forest	0.666	0.319	0.408	0.936	0.270

Cross-strategy evaluation of tuned models confirms that Strategy B consistently yields the highest test R^2 for six of seven models (Figure 5). The advantage of stratified over random sampling is approximately 2–4 R^2 points, demonstrating that distributional coverage of the target variable provides more value than additional random samples from the mode of the distribution. Random Forest is the exception, performing slightly better on Strategy C (boron-balanced), which may reflect its tendency to classify by boron presence when boron molecules dominate the training distribution.

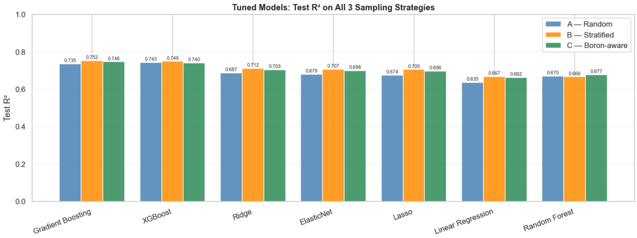

Figure 5: Cross-strategy evaluation of tuned models. Each tuned model (trained on Strategy B) is tested on the test sets of all three strategies. Strategy B consistently produces the highest test R^2 across most models. Error bars indicate performance variability across strategies.

Chemical Interpretation: Why Tree Models Outperform Linear Models

The consistent 5–10 R^2 point advantage of gradient-boosted models over regularised linear models reflects two fundamental properties of the HOMO-LUMO prediction problem.

Multicollinearity in the descriptor space. The molecular descriptor matrix contains hundreds of mutually correlated features: molecular weight grows with ring count; ring count correlates with the length of the π-conjugated system; conjugation length determines the HOMO-LUMO gap through the Hückel model relationship $\Delta\varepsilon \propto \beta/n$ (where β is the resonance integral and n the number of conjugated units). When such correlated features are fed into ordinary least squares, individual coefficient estimates become large and unstable—each coefficient partially compensates for its correlated neighbours rather than estimating a genuine marginal electronic contribution. Ridge and Lasso partially address this through regularisation (explaining their 4–5 point advantage over plain Linear Regression), but cannot eliminate the fundamental instability. Tree-based methods are structurally immune: at each split node, the algorithm selects the single most informative feature. Correlated features take turns being selected across the ensemble without creating numerical instability.

Nonlinear structure-property relationships. The gap is fundamentally nonlinear in any single structural descriptor. Hückel theory predicts a $1/n$ dependence on acene length; nitrogen substitution effects depend on whether the nitrogen occupies an α (pyridine-like) or β position in the aromatic ring; boron introduces a qualitative discontinuity in gap behaviour compared to isovalent carbon. Gradient boosting and XGBoost capture these relationships through hierarchical decision boundaries with no functional form constraints.

Random Forest overfitting. Despite similar or higher training R^2 (0.936), Random Forest underperforms Gradient Boosting on the test set. This reflects a known limitation of bagged trees on medium-sized datasets: without depth regularisation, individual trees memorise training scaffolds. Gradient boosting's sequential, residual-fitting approach with shrinkage (learning rate < 1) provides implicit regularisation that improves generalisation.

Feature Importance Analysis

Figures 6 and 7 show the top-20 features by intrinsic and permutation importance for all seven models.

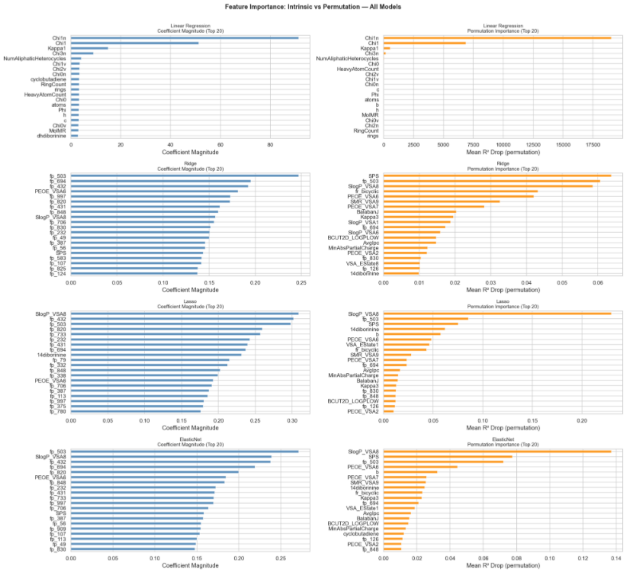

Figure 6: Feature importance — Linear Regression, Ridge, Lasso, ElasticNet (top four models). *Left panels:* Intrinsic importance (absolute coefficient magnitude). *Right panels:* Permutation importance on the test set.

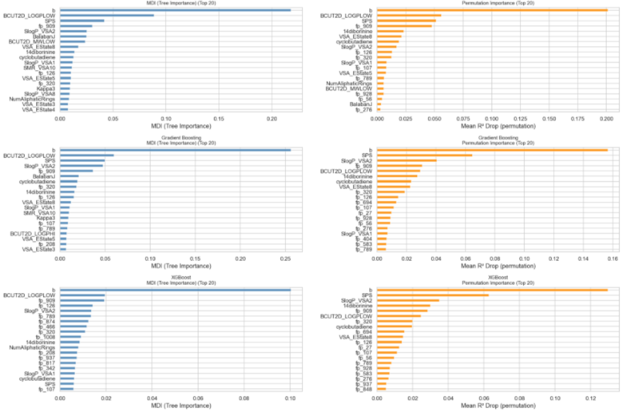

Figure 7: Feature importance — Random Forest, Gradient Boosting, XGBoost (bottom three models). *Left panels:* Mean decrease in impurity (MDI). *Right panels:* Permutation importance.

Several chemically informative patterns emerge from the importance rankings:

Boron atom count as the dominant predictor. For all tree-based models, the number of boron atoms ranks as the top or second-ranked feature by both intrinsic and permutation importance. This is consistent with boron's electronic role: in a π-conjugated polycyclic framework, each boron atom contributes an empty p-orbital to the aromatic system. This lowers the LUMO energy significantly—the empty p-orbital provides a low-energy orbital for π^* electrons—while leaving the HOMO energy relatively unchanged, because occupied π density prefers the more electronegative carbon and nitrogen atoms. The net effect is a systematic compression of the HOMO-LUMO gap proportional to boron content.

Topological connectivity indices (Chi series, Kappa). For linear models, the Chi1n and Chi1 topological indices emerge as dominant features. These indices encode the degree of branching and the connectivity of the molecular graph, which serve as proxies for π-system extent and topology. In Hückel theory, the gap scales inversely with the length of the conjugated chain, and the Chi indices capture this relationship in a graph-theoretic framework accessible to linear models.

Morgan fingerprint bits. The appearance of numbered Morgan fingerprint bits (e.g., 'morgan_74', 'morgan_126') among the top features for XGBoost and Gradient Boosting is particularly interpretable: each such bit encodes the presence of a specific circular chemical environment at radius 2. The fact that these bits are predictive confirms that local substructural motifs—specific ring junction types, heteroatom neighbours, edge fusion patterns—contribute independently and consistently to gap values beyond what global descriptors capture.

Boron-Controlled Experiments

Table 2 compares model performance across three conditions: full mixed dataset, boron-only, and boron-free.

Table 2: Three-way boron comparison. Test R^2 for all seven models trained on (i) Strategy B mixed dataset, (ii) boron-only molecules, and (iii) boron-free molecules.

Model	Mixed (B)	Boron-only	Boron-free
Gradient Boosting	**0.752**	0.635	**0.798**
XGBoost	0.749	0.632	0.792
Ridge	0.712	0.614	0.748
ElasticNet	0.707	0.609	0.740
Lasso	0.705	0.607	0.742
Random Forest	0.666	0.558	0.731
Linear Regression	0.667	0.482	0.735

Two results are particularly informative:

Boron-free models outperform boron-only models. This initially counter-intuitive finding resolves upon reflection: the boron-only subset (mean gap 4.51 eV, σ = 0.56 eV) has a narrower gap distribution than the boron-free subset (mean 5.07 eV, σ = 0.70 eV). Within the boron-containing space, the gap is determined by the interplay of boron count, ring topology, and nitrogen co-substitution—a more complex multi-dimensional relationship that is harder to learn from 4,000 training examples than the broader but more regularly varying carbon/nitrogen chemistry. Furthermore, in the boron-free case, the model is free from needing to distinguish boron-containing environments and can focus entirely on conjugation-length and nitrogen-placement effects.

Substantial predictive power persists without boron. Boron-free Gradient Boosting achieves R^2 = 0.798—higher than the mixed dataset—confirming that the models genuinely learn structure-property relationships within the heteroaromatic carbon/nitrogen/oxygen/sulfur chemical space. Boron is a dominant predictor but not a shortcut: removing it does not collapse model performance.

Train vs. Test Scatter Analysis

Figures 8–10 show actual versus predicted gap values for all seven models on the Strategy B train and test sets.

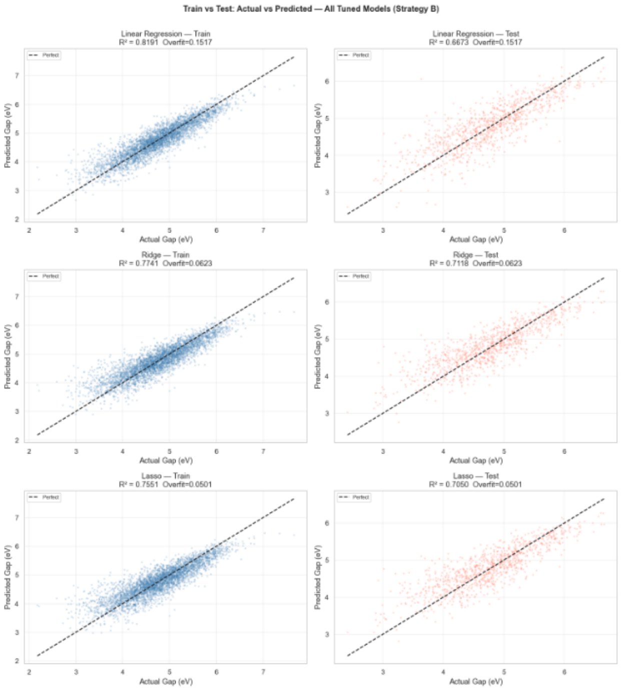

Figure 8: Actual vs. predicted HOMO-LUMO gap — Linear Regression, Ridge, Lasso (Strategy B). *Left column:* Training set. *Right column:* Test set. The dashed diagonal represents perfect prediction.

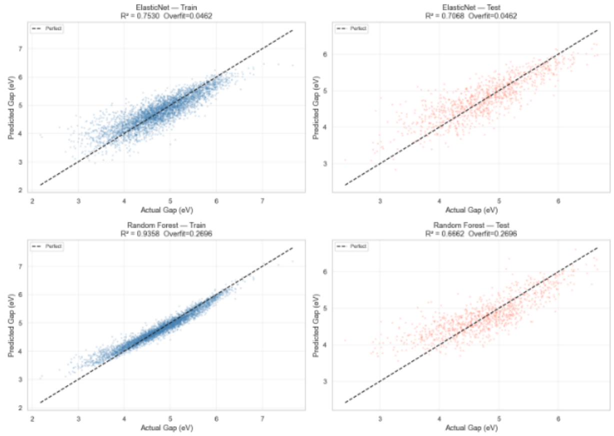

Figure 9: Actual vs. predicted HOMO-LUMO gap — ElasticNet, Random Forest (Strategy B). Random Forest exhibits severe overfitting with strong train/test divergence.

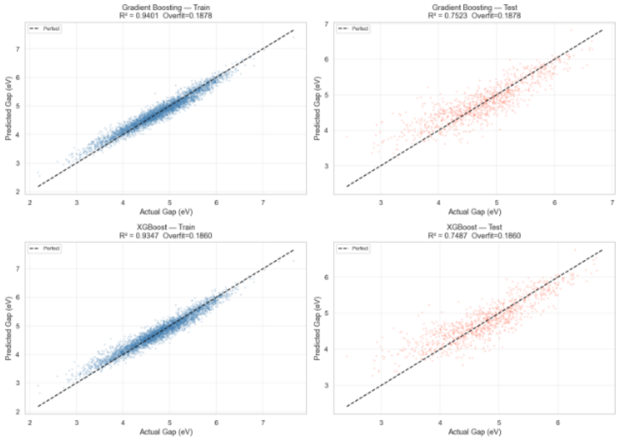

Figure 10: Actual vs. predicted HOMO-LUMO gap — Gradient Boosting, XGBoost (Strategy B). Both models show tight, symmetric test scatter with minimal bias across the gap range.

The scatter plots reveal model-specific patterns. Linear Regression shows a systematic fan shape—underprediction at high gap values and overprediction at low gap values—reflecting the model's inability to capture the nonlinear gap scaling in the extreme regions of chemical space. Random Forest shows an artefact of severe overfitting: near-perfect training fit (points on the diagonal) but substantially increased scatter on the test set, with visible mean-reversion toward the training mean. Gradient Boosting and XGBoost show balanced behaviour on both train and test, with residuals distributed symmetrically around zero across the full gap range.

Structural Outlier Analysis

The 10 best-predicted and 10 worst-predicted molecules from the Strategy B test set are identified for each model to investigate whether prediction failures have a structural explanation. Figures 11–12 show the molecular visualisations for XGBoost. The best- and worst-predicted molecules for the other models are provided in the accompanying notebook.

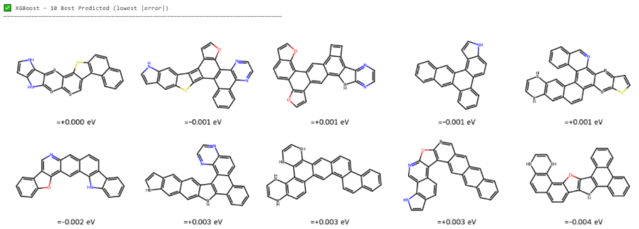

Figure 11: Ten best-predicted molecules from the XGBoost test set (lowest absolute error). These molecules predominantly feature regular, planar polycyclic frameworks with nitrogen substitution patterns well-represented in the training set.

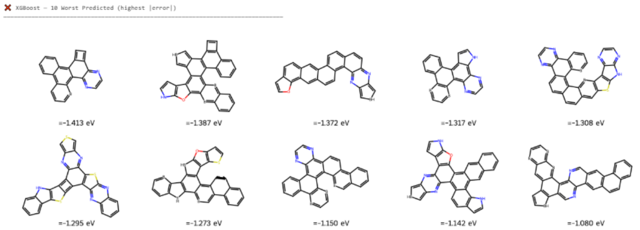

Figure 12: Ten worst-predicted molecules from the XGBoost test set (highest absolute error). These molecules tend to feature unusual ring fusion geometries, mixed boron-nitrogen heterocyclic patterns, or extended linear acenes beyond the size range common in the training set.

Visual inspection reveals a consistent pattern: well-predicted molecules are structurally regular—planar polycyclic systems with nitrogen atoms occupying common positions (pyridine-like α-substitution, standard pyrazine or quinoxaline motifs)—while poorly predicted molecules feature unusual ring combinations or heteroatom arrangements that create chemical environments rare in the training set.

Morgan fingerprint rarity analysis (Figures 13–14) quantifies this observation: fingerprint bits that are systematically more active in poorly predicted molecules than in well-predicted ones correspond to structural fragments appearing in fewer than 5% of training molecules. The model has encountered too few examples of these local chemical environments to learn their contribution to the gap.

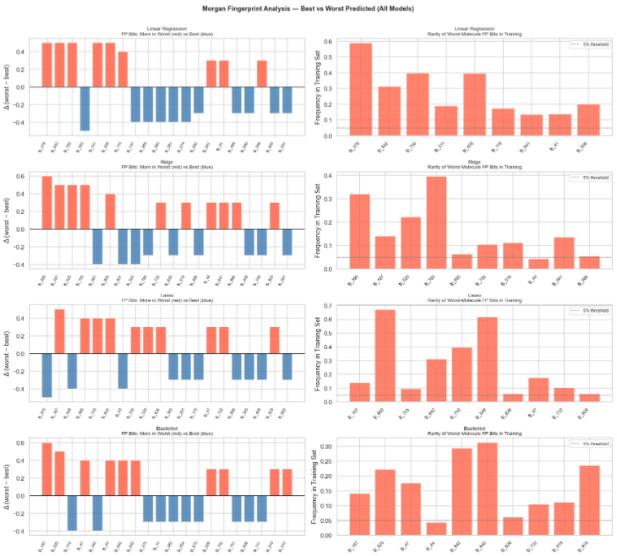

Figure 13: Morgan fingerprint bit activation analysis — Linear Regression, Ridge, Lasso, ElasticNet. Red bars: bits more active in worst-predicted molecules; blue bars: bits more active in best-predicted molecules. Right panels show training-set frequency of the discriminating bits; bits below the 5% threshold encode rare structural fragments.

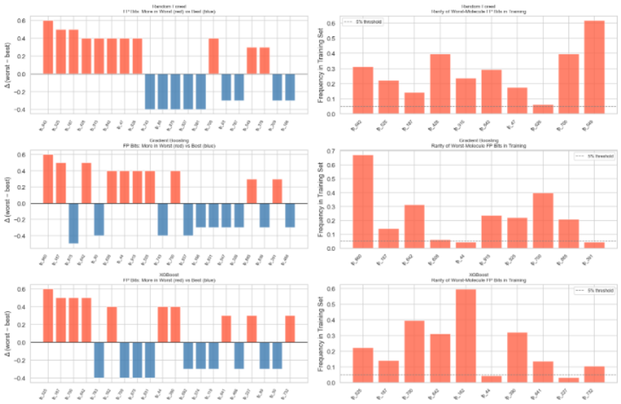

Figure 14: Morgan fingerprint bit activation analysis — Random Forest, Gradient Boosting, XGBoost. The consistent pattern of rare bits in worst-predicted molecules confirms that structural novelty drives prediction failure.

The implication is practically actionable: rather than general model improvements, targeted augmentation of the training set with molecules bearing these rare fragments would be the most efficient path to reducing outlier prediction errors. This approach is analogous to active learning,[12] where data collection is directed to regions of chemical space where the model is most uncertain.

Molecular Topology Analysis

Topology analysis compares ring count, aromatic ring count, and heavy atom count between best- and worst-predicted molecules for each model (Figure 15).

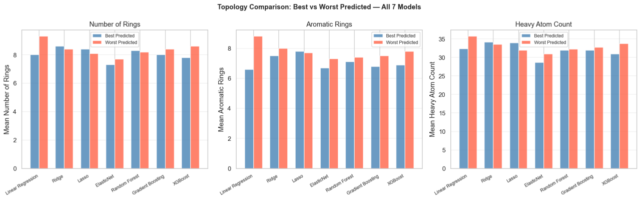

Figure 15: Molecular topology comparison (cross-model summary). Worst-predicted molecules consistently show higher ring counts and heavy atom counts than best-predicted molecules, indicating that larger, more complex ring systems are harder to predict.

Across all seven models, worst-predicted molecules have consistently higher mean ring counts (≈7.3–8.5 rings) and higher heavy atom counts (≈29–33 atoms) than best-predicted molecules (≈6.5–7.0 rings, ≈25–30 atoms). This pattern has a chemical explanation: larger polycyclic systems have more atoms, more bonds, and consequently more distinct chemical environments—each potentially introducing substructural features that are rarer in the training set. Additionally, for large acenes and coronene-type systems, the gap enters a regime where subtle differences in ring arrangement create significant electronic differences that are more difficult to capture from 2D descriptors alone.

Multi-Seed Robustness

Repeating the full experiment under three independent random seeds (42, 123, 777) produces stable model rankings with small standard deviations (Figure 16): XGBoost achieves $R^2 = 0.725 \pm 0.013$, Gradient Boosting 0.702 ± 0.015. Random Forest shows the largest R^2 variability ($\sigma \approx 0.008$) but smallest MAE variability, reflecting sensitivity of deep trees to which specific scaffolds are held out.

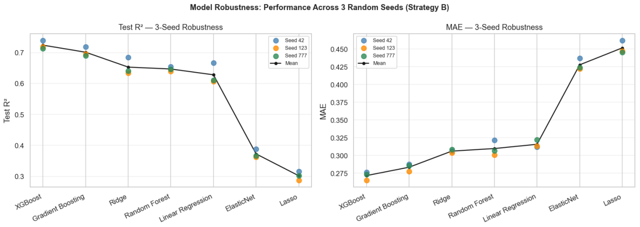

Figure 16: Multi-seed robustness validation (seeds 42, 123, 777) using Strategy B with default model parameters. Error bars represent one standard deviation across seeds. XGBoost and Gradient Boosting show consistent top performance with low variability.

The stability of model rankings across seeds confirms that the performance ordering—gradient-boosted trees > regularised linear > plain linear regression / random forest—is a genuine property of the model families rather than an artefact of a particular train/test split.

Final Comparative Summary

Figure 17 presents the comprehensive performance comparison across all conditions.

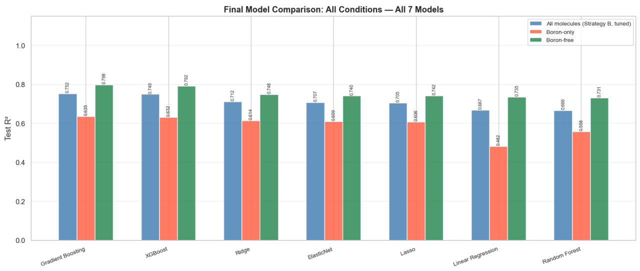

Figure 17: Comprehensive final performance summary. Test R^2 is shown for Strategy B, boron-only, and boron-free conditions for all seven tuned models. Gradient Boosting achieves the highest mixed-dataset performance; boron-free conditions consistently yield higher R^2 than boron-only, reflecting the narrower and more complex gap distribution within the boron-containing chemical space.

Conclusion

This study demonstrates that the HOMO-LUMO gap of cata-condensed heteroaromatic molecules can be predicted from 2D molecular structure with meaningful accuracy using ML models trained on RDKit descriptors and Morgan fingerprints. Gradient Boosting achieves the best performance ($R^2 = 0.752$, MAE $= 0.266$ eV on a scaffold-split test set), and model rankings are reproducible across random seeds.

The performance gap between tree-based and linear models is chemically explainable: the HOMO-LUMO gap is a nonlinear function of molecular descriptors, and descriptor multicollinearity—inherent in the molecular feature space—disproportionately degrades linear model accuracy. Gradient boosting's resistance to both multicollinearity and nonlinearity makes it the natural choice for this class of molecular property prediction tasks.

Boron atom count is the dominant predictor for ensemble models, reflecting the fundamental electronic role of boron in compressing the HOMO-LUMO gap through LUMO stabilisation. However, controlled boron-free experiments demonstrate that substantial predictive power persists in the carbon/nitrogen/oxygen/sulfur chemical space, confirming genuine structure-property learning beyond boron presence alone. Outlier analysis connects prediction failures to structural novelty in Morgan fingerprint space, providing an actionable diagnosis: targeted data augmentation with molecules bearing rare fingerprint fragments is the most efficient path to improvement.

Limitations include the restriction to 5,000 training molecules (out of 52,000 available) for computational tractability, which may introduce distributional biases despite stratification; the use of 2D descriptors only, which cannot capture 3D conformational or steric effects; and hyperparameter tuning performed on a single strategy (B) without per-subset re-optimisation for the boron-controlled experiments.

Future work could include extending to graph neural networks[7] that operate directly on the molecular graph without explicit featurisation; incorporating 3D descriptors from conformational sampling; applying active learning[12] to systematically acquire training data in the structurally novel regions identified by the outlier analysis; and exploring transfer learning from larger, noisier chemical databases to improve generalisation on unusual ring topologies.

Individual Contribution

In this project, I was responsible for all aspects of the implementation and analysis: data loading and preprocessing, data leakage identification and prevention, implementation of the three sampling strategies, feature engineering using RDKit (molecular descriptors and Morgan fingerprints), implementation of the Murcko scaffold splitting procedure, baseline and tuned training of all seven regression models, hyperparameter optimisation using RandomizedSearchCV, feature importance analysis (intrinsic and permutation methods), controlled boron-only and boron-free experiments, structural outlier analysis and molecular visualisation, Morgan fingerprint rarity analysis, molecular topology comparison, multi-seed robustness validation, and preparation of all figures and tables in this report.

Problem formulation and general methodological directions were discussed in project meetings with the supervising professor. All implementations, experiments, and analyses presented in this report are my own. Where existing models, datasets, or code from external sources (RDKit, scikit-learn, XGBoost, the COMPAS-2D dataset) were used, they have been appropriately cited. All interpretations and conclusions reflect my individual analysis.

Code Availability

The complete source code used for data preprocessing, feature engineering, model training, evaluation, and figure generation is publicly available at:

https://github.com/MohsinRaza512/Digital-Alchemy-Project$_{H}OMO - LUMO - Gap - Prediction - COMPAS2D$

The repository contains a fully self-contained Jupyter Notebook that reproduces all results reported in this study using the COMPAS-2D dataset.[1]

References

(1) Stuyver, T.; Jorner, K.; Coley, C. W. COMPAS-1: A Diverse, Computationally-Oriented Dataset of Conjugated Polycyclic Aromatic Systems. *Scientific Data* **2023**, *10*, 367, DOI: 10.1038/s41597-023-02207-x.

(2) Landrum, G. RDKit: Open-Source Cheminformatics. *Zenodo* **2016**, http://www.rdkit.org.

(3) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, *5*, 107–113, DOI: 10.1021/c160017a018.

(4) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893, DOI: 10.1021/jm9602928.

(5) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *The Journal of Physical Chemistry Letters* **2020**, *11*, 2336–2347, DOI: 10.1021/acs.jpclett.9b03664.

(6) von Lilienfeld, O. A.; Burke, K. Retrospective on a Decade of Machine Learning for Chemical Discovery. *Nature Communications* **2020**, *11*, 4895, DOI: 10.1038/s41467-020-18556-9.

(7) Gilmer, J.; Schütt, K. T.; Brockherde, F.; Kindermans, P.-J.; von Lilienfeld, O. A. Neural Message Passing for Quantum Chemistry. *Proceedings of Machine Learning Research* **2017**, *70*, 1263–1272.

(8) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Scientific Data* **2014**, *1*, 140022, DOI: 10.1038/sdata.2014.22.

(9) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; pp 785–794, DOI: 10.1145/2939672.2939785.

(10) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019.

(11) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; others Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(12) Settles, B. Active Learning Literature Survey. *Computer Sciences Technical Report 1648* **2009**,