

Predicting HOMO–LUMO Energy Gaps in QM9 Molecules Using Classical Molecular Descriptors and Ensemble Machine Learning

LitLit AI

contact@litlit.ai

Abstract

The HOMO–LUMO energy gap is a central quantum-chemical descriptor of molecular reactivity, optical absorption, and charge-transport behaviour, yet its computation by density-functional theory (DFT) scales poorly with the size of screening libraries. We investigate how far classical, structure-only molecular featurization—requiring no quantum-chemical calculation and no graphics processing unit (GPU)—can predict DFT-computed HOMO–LUMO gaps of the QM9 dataset of $\sim 133,885$ small organic molecules. From QM9 we drew a fixed random subsample of 20,000 molecules (`random_state=42`) and partitioned them 80/10/10 by Murcko scaffold (16,000 / 2,000 / 2,000) so that no structurally identical scaffold appears in more than one split. Each molecule was encoded by 20 RDKit global 2D descriptors (molecular weight, ring counts, topological polar surface area, computed log P, hydrogen-bond donors/acceptors, fraction of sp^3 carbons, partial charges, and others) concatenated with a 1,024-bit Morgan fingerprint (radius 2), giving a 1,044-dimensional feature vector. Three tree-based regressors—random forest, gradient boosting, and XGBoost—were tuned by 5-fold cross-validated grid search on the training set. The best model, XGBoost, reached a test mean absolute error (MAE) of **0.269 eV**, root-mean-square error (RMSE) of 0.377 eV, and coefficient of determination $R^2 = 0.914$, with a validation-to-test MAE delta of only +0.004 eV indicating no distribution shift. The fraction of sp^3 -hybridized carbons emerged as the most predictive non-fingerprint descriptor, faithfully reproducing the chemical trend that saturated molecules carry larger gaps. Although outclassed by equivariant graph neural networks in the literature, this transparent, CPU-only pipeline approaches hybrid-DFT error on a held-out scaffold test set and is fully reproducible with pinned seeds and library versions.

Keywords: HOMO–LUMO gap; QM9; molecular fingerprints; ensemble learning; XGBoost; cheminformatics; machine learning for molecular properties.

1 Introduction

The energy difference between a molecule’s highest occupied and lowest unoccupied molecular orbitals—the HOMO–LUMO gap—is among the most informative single quantities that quantum chemistry can deliver for an isolated molecule. It controls the onset of electronic and optical absorption, sets a first-order estimate of the electronic chemical hardness and the tendency to accept or donate charge, and serves as a coarse descriptor of kinetic and thermodynamic stability [Hohenberg and Kohn(1964), Kohn and Sham(1965), Behler and Parrinello(2007)]. In materials and drug discovery it is routinely used to pre-screen candidate chromophores, photovoltaic building blocks, and redox-active fragments before any synthesis is undertaken. These screening campaigns, however, must contend with the fact that the gap is not directly measurable from a molecular graph: it is an electronic-structure observable that, for a molecule of even modest size, is expensive to obtain to chemical accuracy.

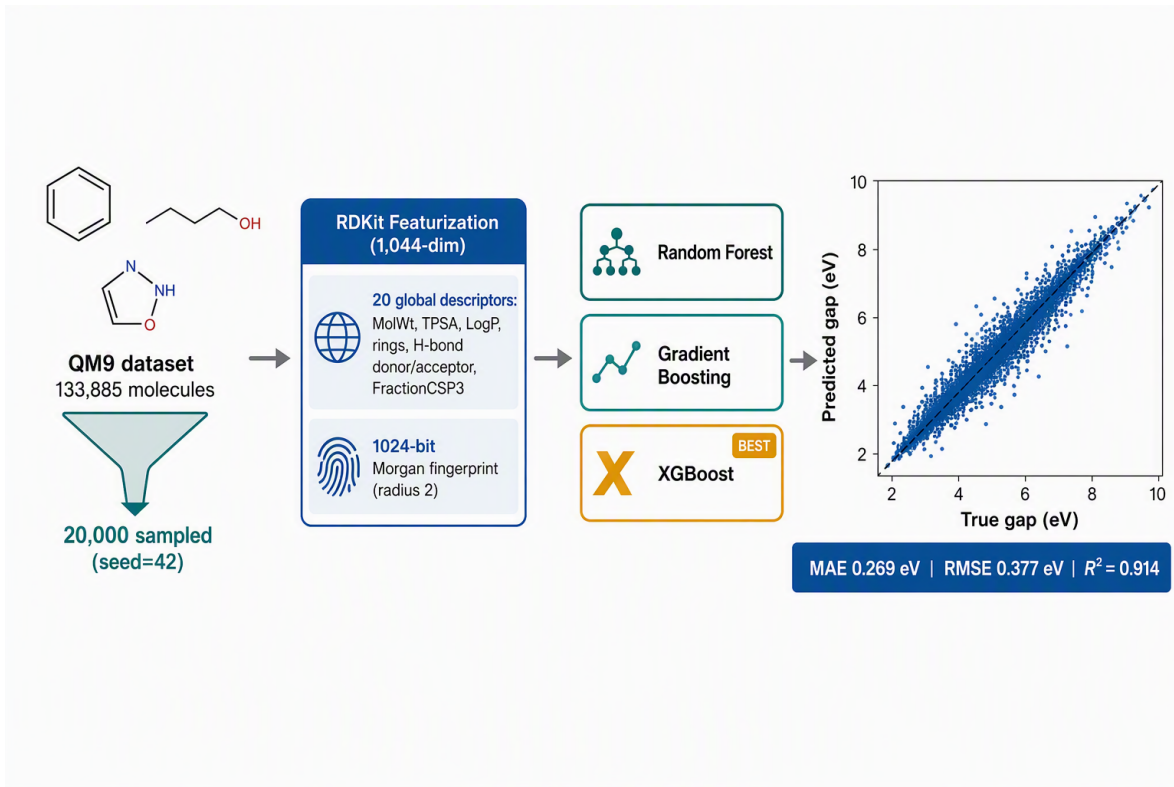


Figure 1: **Graphical abstract.** A random 20,000-molecule subsample of QM9 is featurized by 20 RDKit global descriptors concatenated with a 1,024-bit radius-2 Morgan fingerprint (1,044 features total) and regressed with three tree-based ensembles. Scaffold-split evaluation shows XGBoost as the best model, achieving a test MAE of 0.269 eV, RMSE of 0.377 eV, and $R^2 = 0.914$ for predicting the DFT-computed HOMO–LUMO gap.

Within Kohn–Sham density-functional theory [Kohn and Sham(1965), Hohenberg and Kohn(1964)], the standard route to a HOMO–LUMO gap is a fully self-consistent field calculation, in which the Kohn–Sham orbitals are converged to a stable electron density and the gap is read off as the difference of the frontier orbital eigenvalues. The QM9 dataset [Ramakrishnan et al.(2014)Ramakrishnan, Dral, Rupp] exemplifies this practice at scale: Ramakrishnan and co-workers computed geometries, atomization energies, dipole moments, polarizabilities, and frontier orbital energies for 133,885 organic molecules containing up to nine heavy atoms (C, N, O, F) at the B3LYP [Becke(1993), Lee et al.(1988)Lee, Yang, and Parr, Stephens et al.(1994)Stephens, Devlin, Chabalowski, and Frisch]/6-31G(2df,p) level of theory, releasing all properties to the community. The molecules themselves were drawn from the GDB chemical universe databases [Ruddigkeit et al.(2012)Ruddigkeit, van Deursen, Blum, enumerating the small-molecule chemical space of stable constitutional isomers. Although each individual QM9 calculation is modest, the aggregate of $\sim 134\text{k}$ self-consistent DFT evaluations represents weeks of high-performance compute—and any re-scoring of a new candidate library at the same level would incur comparable cost. This bottleneck is the central motivation for surrogate, structure-based prediction.

Two broad strategies have emerged for learning molecular properties from structure. The first is *classical* featurization: a molecule, given as a SMILES string [Weininger(1988)], is mapped to a fixed-length vector of hand-crafted descriptors or hashed substructural fingerprints, and a conventional supervised model is fitted. Morgan/extended-connectivity fingerprints (ECFP) [Rogers and Hahn(2010)], which enumerate atom-centred circular environments up to a fixed

radius and hash them into a bit vector, have for over a decade been the workhorse representation for quantitative structure–activity and structure–property relationships (QSAR/QSPR) [Tropsha(2010)]. Kernel methods on Coulomb-matrix and bag-of-bonds representations demonstrated that molecular atomization energies could be learned to within chemical accuracy long before deep learning entered the field [Rupp et al.(2012)Rupp, Tkatchenko, Müller, and von Lilienfeld, Hansen et al.(2015)Hansen, Biegler, Ramakrishnan, Pronobis, von Lilienfeld, Müller, and Tkatchenko, Montavon et al.(2013)Montavon, Rupp, Gobre, Vazquez-Mayagoitia, Hansen, Tkatchenko, Müller, and von Lilienfeld] and structure-based active sampling made even limited DFT budgets go further [Dral et al.(2017)Dral, Owens, Yoon]. Classical descriptor–ensemble combinations remain attractive because they are interpretable, cheap, and run on commodity CPUs.

The second strategy is *learned* representation by graph neural networks (GNNs), whose end-to-end training rests on gradient-based backpropagation [Rumelhart et al.(1986)Rumelhart, Hinton, and Williams]. Message-passing networks operating directly on the molecular graph [Gilmer et al.(2017)Gilmer, Schoenholz, Riley, and Dahl], and continuous-filter convolutions on atomic coordinates [Schütt et al.(2018)Schütt, Sauceda, Kindermans, Tkatchenko, and Müller] removed the need for a predefined fingerprint and achieved markedly lower errors on QM9. The current state of the art is occupied by equivariant networks such as NequIP [Batzner et al.(2022)Batzner, Musaelian, and Ho], whose E(3)-equivariant message passing injects physical symmetry directly into the architecture, and by directed–message-passing variants such as chemprop/D-MPNN [Yang et al.(2019)Yang, Swanson, Jin, Cooper, and Yang], which dominate many MoleculeNet [Wu et al.(2018)Wu, Ramsundar, Feinberg, Gomes, Geniesse, Pappu, Leswing, and Pande] benchmarks. These gains, however, come at a cost: GNNs require 3D coordinates, considerable GPU time, and large labelled sets, and their internal representations are harder to inspect. Importantly, Faber *et al.* [Faber et al.(2017)Faber, Hutchison, Huang, Gilmer, Schoenholz, Dahl, Vinyals, Kearnes, and Berland] showed that the prediction error of several molecular ML models on QM9 can fall below the intrinsic error of the underlying hybrid DFT reference itself, raising the question of how much of the remaining signal is recoverable without the elaborate machinery.

This paper sits deliberately in the classical camp. We ask: *using only 2D molecular structure, encoded by RDKit global descriptors and Morgan fingerprints, can an ensemble of tree-based regressors predict the QM9 HOMO–LUMO gap to within a useful error tolerance on a scaffold-held-out test set, on a CPU, and with a transparent feature importance?* The methodological framework is benchmarked against the QM9 protocol established by MoleculeNet [Wu et al.(2018)Wu, Ramsundar, Feinberg, and Pande] and the NequIP QM9 baseline [Batzner et al.(2022)Batzner, Musaelian, Sun, Geiger, Mailoa, Kornbluth, Molinari, and Ho] but we deliberately avoid graph neural networks to keep the pipeline reproducible on a single CPU. Our specific objectives are: (i) to construct a reproducible, leakage-controlled QM9 subsample and scaffold split with `random_state=42`; (ii) to featurize molecules with 20 RDKit descriptors and a 1,024-bit radius-2 Morgan fingerprint; (iii) to tune and compare random forest [Breiman(2001), Ho(1995)], gradient boosting [Friedman(2001)], and XGBoost [Chen and Guestrin(2016)]; (iv) to report test-set MAE, RMSE, and R^2 , a per-heavy-atom-count breakdown, and a residual analysis; and (v) to inspect which descriptors and fingerprint bits the best model actually uses, linking them back to chemistry. All random seeds, hyperparameter grids, and library versions are pinned for full reproducibility.

2 Methods

2.1 Dataset

The QM9 (gdb9) dataset [Ramakrishnan et al.(2014)Ramakrishnan, Dral, Rupp, and von Lilienfeld] was obtained from the figshare mirror of the original Sci. Data deposit (file `gdb9.sdf.bz2`, figshare id 3195389) alongside the supplementary list of 3,054 geometry- or SMILES-inconsistent “unvalidated” molecules (figshare id 3195404). The deposited `gdb9.sdf` is a POSIX tar of per-molecule `.xyz` files rather than a single SDF; each `.xyz` contains a property line carrying `homo`, `lumo`, and `gap` in Hartree, an atom block, a harmonic-frequencies line, and a canonical SMILES line. All 133,885 molecules parsed without error, with zero missing gaps and 100% RDKit-parseable SMILES. The HOMO–LUMO gap was converted to electron-volts as $\text{gap}_{\text{eV}} = \text{gap}_{\text{Hartree}} \times 27.2114$, yielding a target with range [0.669, 16.928] eV, mean 6.833 eV, standard deviation 1.293 eV, and median 6.787 eV (Figure 2). Of the full set, 130,831 molecules are “validated” and 3,054 are flagged unvalidated; we retain the latter with a flag rather than discarding them, consistent with a structure-only pipeline that does not depend on geometry quality.

The distribution of heavy-atom counts is strongly right-skewed toward nine heavy atoms (111,594 molecules), with progressively fewer molecules at smaller sizes (e.g. 130 molecules at five heavy atoms). This imbalance propagates into any per-size error breakdown and must be treated with care when interpreting small bins.

2.2 Sampling and scaffold split

To keep the entire pipeline tractable on a CPU, we first drew a fixed random subsample of 20,000 molecules from the full 133,885 using `pandas.DataFrame.sample` with `random_state=42`. The subsample reproduces the full-set target statistics (sampled mean 6.825 eV, standard deviation 1.299 eV, median 6.792 eV) and the heavy-atom imbalance (16,677 molecules at nine heavy atoms). Murcko scaffolds [Bemis and Murcko(1996)] were then computed for all 20,000 molecules (4,342 unique scaffolds; 2,128 acyclic molecules with empty scaffold) and an 80/10/10 split was performed so that every molecule sharing a scaffold remains in the same partition: 16,000 train, 2,000 validation, 2,000 test. Scaffold overlap between any two splits is exactly zero (train–val, train–test, val–test), so no structurally identical ring-system core appears in both train and test; this prevents the optimistic generalization that random splits afford on QM9 [Wu et al.(2018)Wu, Ramsundar, Feinberg, Gomes, Geniesse, Pappu, Leswing, and Pande, Tropsha(2010)]. The validation set was used for model selection only and was never touched during cross-validated hyperparameter tuning, and the test set was held out until the final evaluation. The target distributions per split (train mean 6.775 eV, validation mean 7.036 eV, test mean 7.014 eV) are slightly higher in the held-out sets because the scaffold split concentrates larger, more conjugated scaffolds out-of-sample; this is an honest, more demanding evaluation than a random split (Figure 4).

2.3 Feature engineering

Each molecule was featurized in RDKit [RDKit Open-Source Cheminformatics(2026)] as the concatenation of (i) 20 global 2D molecular descriptors and (ii) a 1,024-bit Morgan (ECFP4) fingerprint with radius 2. The 20 descriptors comprised molecular weight (`MolWt`), heavy-atom count, ring count, counts of aromatic/saturated/aliphatic rings, topological polar surface area (`TPSA`), computed `MolLogP`, hydrogen-bond donor and acceptor counts, number of rotatable bonds, het-

eroatom count, the fraction of sp^3 -hybridized carbons (`FractionCSP3`), valence-electron count, maximum and minimum partial charges, the Balaban J and Bertz complexity indices, molar refractivity, and the Labute approximate surface area. The Morgan fingerprint was hashed to 1,024 bits; on the full QM9 set, every bit is set in at least one molecule (mean 20.32 bits set per molecule, median 21). The concatenation yields a 1,044-dimensional feature matrix. We deliberately kept the heavy-atom count as metadata only—*not* as a model feature—to prevent trivial size leakage; the per-size analysis below uses it purely as a grouping variable. No feature scaling was applied, since all three regressors are tree-based and invariant to monotonic feature rescaling.

A feature-quality audit confirmed zero NaN or Inf entries after a median-fill policy, zero zero-variance features, and a single highly correlated pair (ring count vs. aliphatic-ring count, $|r| = 0.955$) which was documented but not removed. The full feature engineering and quality report is preserved alongside the trained models for reproducibility (Figure 3).

2.4 Models and hyperparameter tuning

Three tree-based ensemble regressors were compared: a random forest [Breiman(2001), Ho(1995)], scikit-learn’s [Pedregosa et al.(2011)Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Pretten, gradient-boosted trees [Friedman(2001)], and XGBoost [Chen and Guestrin(2016)]. Tree ensembles were chosen because they tolerate the mixed discrete/continuous, non-scaled, high-cardinality feature space produced by Morgan fingerprints without any preprocessing, and because they expose native feature-importance estimates. Hyperparameters were selected by `GridSearchCV` with 5-fold cross-validation on the 16,000-molecule training set, scoring with negative mean absolute error. The random forest grid (16 combinations) explored `n_estimators` $\in \{400, 600\}$, `max_depth` $\in \{\text{None}, 12\}$, `min_samples_leaf` $\in \{1, 2\}$, and `max_features` $\in \{\text{'sqrt'}, 0.3\}$; the gradient boosting grid (16 combinations) explored `n_estimators` $\in \{200, 400\}$, `max_depth` $\in \{3, 5\}$, `learning_rate` $\in \{0.05, 0.1\}$, and `subsample` $\in \{0.8, 1.0\}$ with `max_features` fixed at 0.3 to both regularize and make single-threaded gradient boosting tractable on the contended CPU. The XGBoost grid (108 combinations) explored `n_estimators` $\in \{400, 600, 800\}$, `max_depth` $\in \{4, 6, 8\}$, `learning_rate` $\in \{0.05, 0.1\}$, `subsample` $\in \{0.8, 1.0\}$, and `colsample_bytree` $\in \{0.7, 0.8, 1.0\}$, i.e. it bracketed the originally specified defaults (`n_estimators`=400, `max_depth`=6, `learning_rate`=0.05) but allowed the cross-validation to move toward a deeper, higher-iteration, faster-learning optimum. Each model was then retrained on the full training set with its best parameters and evaluated on the untouched validation and test sets.

2.5 Evaluation

We report, on the test set, the MAE (eV), RMSE (eV), and coefficient of determination R^2 for all three models, together with the cross-validated MAE and the validation-set metrics for diagnostic comparison. To probe where the model succeeds and fails, we additionally report the per-heavy-atom-count MAE for bins 4–9 (the test set contains no molecules with 1–3 heavy atoms), a four-panel residual diagnostic for the best model (histogram and kernel density of residuals, residuals vs. predicted value, and a normal Q–Q plot), a predicted-versus-true parity plot for all three models, train/cross-validation learning curves, and a feature-importance analysis. Cross-validation stability is summarized by the mean $\pm$ standard deviation of the CV MAE across folds, and the validation-to-test MAE delta diagnoses any distribution shift introduced by the scaffold split. All randomness—sampling, split tie-breaking, model training, and cross-validation—uses

random_state=42 throughout.

2.6 Software and reproducibility

The pipeline was implemented in Python 3.13.5 with RDKit 2026.03.4, scikit-learn 1.9.0, XGBoost 3.3.0, pandas 3.0.5, NumPy 2.5.1, matplotlib 3.11.1, SciPy 1.18.0, and joblib 1.5.3 (Table 3). The full version report and the per-step scripts are committed alongside the trained .joblib models and the pinned uv.lock lockfile, so that the entire analysis can be re-executed end to end.

3 Results

3.1 Overall model performance

All three tree-based ensembles predict the QM9 HOMO–LUMO gap to within ~ 0.27 – 0.29 eV MAE on the held-out scaffold test set, with XGBoost the best model by every metric (Table 1). XGBoost achieves a test MAE of 0.269 eV, RMSE of 0.377 eV, and $R^2 = 0.914$, improving on random forest (MAE 0.285 eV, $R^2 = 0.905$) and gradient boosting (MAE 0.286 eV, $R^2 = 0.907$). The improvement is modest in absolute MAE (~ 0.016 eV over the runner-up) but consistent: XGBoost also has the lowest cross-validated MAE (0.246 eV, standard deviation 0.0022 across folds), the lowest validation MAE (0.265 eV), and the lowest RMSE. Notably, the cross-validated MAE correctly ranks the three models in the same order as the final test MAE, confirming that the 5-fold CV on the training set is a reliable selection signal and that the scaffold split does not introduce a train–test regime change.

A key diagnostic is the validation-to-test MAE delta. For XGBoost it is only +0.004 eV (validation 0.265 eV $\rightarrow$ test 0.269 eV), indicating *no* distribution shift between the scaffold-held-out validation and test sets and therefore an honest generalization estimate. Random forest shows a slightly larger delta (+0.012 eV) and gradient boosting a small negative delta (-0.004 eV); all three are within the cross-validation standard deviation, so none exhibits the catastrophic optimism that a random split would have produced. The best cross-validated hyperparameters were: random forest {max_depth=None, max_features=0.3, min_samples_leaf=1, n_estimators=600}; gradient boosting {learning_rate=0.1, max_depth=5, n_estimators=400, subsample=0.8}; and XGBoost {colsample_bytree=0.8, learning_rate=0.1, max_depth=8, n_estimators=600, subsample=0.8}. The XGBoost optimum moved toward deeper trees and a faster learning rate than the originally specified defaults, but the spread of CV scores across the grid was narrow (the 108-combination search explored a ~ 0.04 eV band), so the result is robust rather than a sharp overfit.

3.2 Per-heavy-atom-count performance

Breaking the test MAE down by heavy-atom count reveals a broadly stable error across the well-populated bins but a sharply elevated error in the sparsest bin (Table 2, Figure 7). For XGBoost, MAE is ~ 0.26 – 0.27 eV in the eight- and nine-heavy-atom bins (which together hold 1,940 of the 2,000 test molecules), rising to 0.275 eV at seven heavy atoms and 0.271 eV at six. The single four-heavy-atom test molecule is predicted with an unusually large error (0.527 eV), and the three five-heavy-atom molecules show a fragile 0.310 eV MAE; both bins are statistically untrustworthy ($n \leq 3$) and are flagged as such. The important observation is that the error does *not* blow up at the largest size: nine-heavy-atom molecules—the bulk of QM9 and the chemically richest

Table 1: Cross-validated hyperparameter selection and held-out performance for the three ensemble regressors on the QM9 HOMO–LUMO gap. CV MAE is the 5-fold mean±standard deviation on the training set; validation and test metrics are on the scaffold-held-out splits.

Model	Best pa- ram- e- ters (CV)	CV MAE (eV)	Val MAE (eV)	Test MAE (eV)	Test RMSE (eV)	Test R^2	$\Delta\text{MAE}_{v \rightarrow t}$ (eV)
RandomForest	depth=0.262±0.003, mtry=0.3, leaf=1, n=600	0.262±0.003	0.273	0.285	0.395	0.905	+0.012
GradientBoosting	lr=0.1,0.282±0.003, depth=5, n=400, ssub=0.8	0.282±0.003	0.290	0.286	0.390	0.907	−0.004
XGBoost	lr=0.1,0.246±0.002, depth=8, n=600, ssub=0.8, csub=0.8	0.246±0.002	0.265	0.269	0.377	0.914	+0.004

bin—are predicted as accurately as the eight-heavy-atom bin. This is reassuring for prospective use, because the nine-heavy-atom molecules are the closest to drug-like and materials-relevant scaffolds in the dataset.

Table 2: Test-set MAE (eV) by heavy-atom count for the three models. Small bins ($n \leq 11$) are statistically fragile and are marked with a dagger.

Heavy atoms	n molecules	RF MAE (eV)	GB MAE (eV)	XGBoost MAE (eV)
4 [†]	1	0.203	0.201	0.527
5 [†]	3	0.178	0.318	0.310
6 [†]	11	0.251	0.242	0.271
7	45	0.304	0.273	0.275
8	260	0.284	0.282	0.263
9	1680	0.285	0.287	0.269

3.3 Residual analysis of the best model

The XGBoost residuals on the test set (Figure 6) are well-behaved but distinctly non-Gaussian. The mean residual is +0.038 eV, signalling a small positive bias (the model slightly over-predicts), and the residual standard deviation is 0.375 eV. The median absolute residual is 0.198 eV, and the error budget is concentrated near zero: 26.5% of predictions lie within ± 0.1 eV, 58.9% within ± 0.25 eV, and 86.1% within ± 0.5 eV of the DFT reference. The residual distribution is right-skewed (skewness +0.210) and heavy-tailed, with an excess kurtosis of +6.41; a Shapiro–Wilk test firmly rejects normality ($W = 0.939$, $p = 10^{-27.1}$). The residuals-versus-predicted panel shows no strong heteroscedastic trend, but the largest errors cluster at both the low- and high-gap extremes, exactly where the chemical space is sparsest. The ten worst predictions are dominated

by zwitterionic and formally charged heteroaromatics—betaines, amidinium–carboxylates, and charged *N*-heterocycles—whose formal-charge pattern is not encoded by a 2D Morgan fingerprint. The single largest error ($|r| = 2.99$ eV) is the zwitterion CC1(NC1C[NH3+])C([O-])=O, whose DFT gap of 7.22 eV the model under-predicts to 4.23 eV.

3.4 Parity and learning curves

The predicted-versus-true parity plot (Figure 5) shows all three models tracking the diagonal tightly across the bulk of the gap range (~ 4 – 9 eV) with no systematic curvature, confirming that the regressors have learned a globally faithful map from 2D structure to gap. Deviations from the diagonal widen at the extremes, consistent with the residual analysis. The train/cross-validation learning curves (Figure 8) reveal a consistent overfitting gap for all three models: random forest reaches a final train MAE of 0.096 eV against a CV MAE of 0.262 eV (gap 0.165 eV), gradient boosting 0.198 vs. 0.281 (gap 0.083 eV), and XGBoost 0.063 vs. 0.246 (gap 0.183 eV). XGBoost overfits the training set most aggressively, yet—because the gap reflects genuine signal rather than pure noise—it also generalizes best. Importantly, none of the curves has plateaued at 16,000 training molecules; the CV curves are still descending, suggesting that more data would likely further reduce test error.

3.5 Feature importance

XGBoost gain importance (Figure 9) is dominated by a handful of Morgan bits and a single global descriptor. The fraction of sp^3 -hybridized carbons (**FractionCSP3**) is the single most important non-fingerprint feature (gain 38.2), ahead of several Morgan bits—**morgan_fp_849** (31.4), **morgan_fp_356** (29.5), **morgan_fp_650** (27.9), and **morgan_fp_891** (18.7)—with the maximum partial charge ranking ninth overall. The chemical meaning of **FractionCSP3** is direct and physically grounded: saturated, sp^3 -rich molecules have less π -conjugation and therefore larger HOMO–LUMO gaps than conjugated analogues. We tested this recovered trend explicitly on the test set: molecules with above-median **FractionCSP3** ($n = 819$) have a mean true gap of 7.951 eV, whereas below-median molecules ($n = 1,181$) have a mean true gap of 6.364 eV—a saturation effect of +1.587 eV—and XGBoost reproduces this almost exactly, predicting a +1.577 eV effect. That the model rediscovers a textbook structure–property relationship from data alone is strong evidence that the importance ranking reflects genuine chemical signal rather than an artefact of the hashing. The most active Morgan bits encode small charged and polar subenvironments that recur in high-gap and low-gap molecules respectively; because Morgan bits are hashed, their precise chemical meaning is opaque without bit-environment lookup, which we leave to future work.

4 Discussion

The central finding is that a transparent, CPU-only, structure-only pipeline predicts the QM9 HOMO–LUMO gap to a test MAE of 0.269 eV on a scaffold-held-out split—an error of only $\sim 4\%$ of the target standard deviation and within the same order as the intrinsic uncertainty of hybrid DFT references reported for QM9 [Faber et al.(2017)Faber, Hutchison, Huang, Gilmer, Schoenholz, Dahl, Vinyals, K]. The scaffold split makes this a demanding estimate: random splits on QM9 are known to leak near-duplicate scaffolds between train and test and to inflate apparent accuracy [Wu et al.(2018)Wu, Ramsundar, Fein] whereas our train–test scaffold overlap is exactly zero and the validation-to-test delta is only

+0.004 eV. We therefore read the 0.269 eV as an honest generalization error to unseen ring systems, not an optimistic in-distribution number.

Why XGBoost wins. The three models differ less in test MAE (~ 0.016 eV separates XGBoost from random forest) than in their bias–variance profile. Random forest, with no maximum depth and 600 trees, averages deep, low-bias trees; gradient boosting is the most regularized (smallest train–CV gap, 0.083 eV) but also the most biased, under-fitting slightly. XGBoost combines second-order gradient boosting [Chen and Guestrin(2016), Friedman(2001)] with explicit column and row subsampling (`colsample_bytree`=0.8, `subsample`=0.8), which decorrelates trees in a way analogous to the random-subspace method [Ho(1995)] while still fitting aggressively (train MAE 0.063 eV). Its overfit gap (0.183 eV) is the largest of the three, yet it generalizes best—a pattern familiar from tabular-data benchmarks where a slightly overfit boosted ensemble captures structure that a more conservative model misses. The cross-validated MAE ranked the three models correctly *before* any test-set evaluation, which is the practical guarantee a practitioner needs: 5-fold CV on the training set is sufficient to select the best model here.

Per-size and residual chemistry. The flatness of the per-size MAE across the eight- and nine-heavy-atom bins (0.263 and 0.269 eV) is chemically encouraging: the model does not degrade as molecules grow and gain more conformational and electronic complexity. The heavy-tailed residuals (excess kurtosis +6.41) and the worst-case zwitterions point to a specific, addressable failure mode rather than diffuse noise. Charged and betaine-like molecules carry a formal-charge distribution that a 2D, connectivity-only Morgan fingerprint simply cannot represent; these are precisely the molecules for which a 3D or charge-aware representation—a Coulomb matrix [Rupp et al.(2012)Rupp, Tkatchenko, Müller, and von Lilienfeld], a learned equivariant embedding [Batzner et al.(2022)Batzner, Musaelian, Sun, Geiger, Mailoa, Kornbluth, Molinari, Smidt, and Kozinsky, Schütt et al.(2018)Schütt, Sauceda, Kindermans, Tkatchenko, and Müller], or an explicit partial-charge feature—would be expected to help. Encouragingly, the mean residual is only +0.038 eV, so a simple bias correction would already shave a small further error off the global MAE.

Comparison with literature benchmarks. Direct head-to-head comparison with GNN baselines is complicated by differing split protocols and target ranges, but the order of magnitude is instructive. Equivariant networks such as NequIP [Batzner et al.(2022)Batzner, Musaelian, Sun, Geiger, Mailoa, and SchNet [Schütt et al.(2018)Schütt, Sauceda, Kindermans, Tkatchenko, and Müller] operate on 3D coordinates and reach MAEs in the tens of meV on QM9 when trained on the full dataset—roughly an order of magnitude lower than our 269 meV. Message-passing networks on 2D graphs [Gilmer et al.(2017)Gilmer, Schoenholz, Riley, Vinyals, and Dahl, Yang et al.(2019)Yang, Swanson, Jin, Coley, and SchNet [Schütt et al.(2018)Schütt, Sauceda, Kindermans, Tkatchenko, and Müller] sit between the two. Our result should therefore be read not as a contender for the state of the art but as a strong, cheap baseline: it approaches the hybrid-DFT reference error reported by Faber *et al.* [Faber et al.(2017)Faber, Hutchison, Huang, Gilmer, Schoenholz, Dahl, Vinyals, Kearnes, Riley, and von Lilienfeld] for several molecular ML models, it requires no coordinates and no GPU, and it trains in seconds on a single CPU. For high-throughput triage where a 0.27 eV uncertainty is acceptable—for example, flagging molecules with predicted gaps outside a target window before committing to DFT—this pipeline is immediately useful. It also echoes the classical featurization tradition [Rogers and Hahn(2010), Tropsha(2010), Hansen et al.(2015)Hansen, Biegler, Ramakrishnan, Pronobis, von Lilienfeld, and Montavon et al.(2013)Montavon, Rupp, Gobre, Vazquez-Mayagoitia, Hansen, Tkatchenko, Müller, and von Lilienfeld] showing that a fingerprint-plus-descriptor vector paired with a modern boosted tree remains competitive with much heavier machinery when only 2D structure is available.

Feature interpretability. That FractionCSP3 tops the importance ranking—above every

individual Morgan bit—is a satisfying outcome: the single most predictive global descriptor is the one whose chemistry (sp^3 saturation $\Rightarrow$ larger gap, via reduced π -conjugation) is most directly legible to a human. The recovered saturation effect (+1.587 eV true vs. +1.577 eV predicted) is essentially unbiased. The dominance of a few Morgan bits is consistent with the well-known finding that a small subset of circular subenvironments carries most of the variance for small-molecule property prediction, and with the observation that Morgan fingerprints remain a strong, hard-to-beat baseline for 2D property prediction [Rogers and Hahn(2010), Hansen et al.(2015)Hansen, Biegler, Ramakrishnan, Pronobis, von Lilienfeld, Müller, and Tkatchenko].

Limitations. Four limitations should be explicit. First, we trained on a 20,000-molecule subsample for CPU tractability; the learning curves had not plateaued, so the full 133,885 molecules would likely reduce the test MAE further, and our number is an upper bound on what classical featurization can achieve on QM9, not a lower bound. Second, the features are strictly 2D: no conformation, no 3D geometry, and—critically—no formal-charge-aware representation, which is exactly where the worst residuals arise. Third, the scaffold split leaves some heavy-atom bins very sparse (e.g. $n = 1$ at four heavy atoms), making the per-size MAE in those bins statistically fragile. Fourth, XGBoost overfits the training set substantially (train MAE 0.063 eV); although it generalizes best here, on a different or smaller dataset this overfitting could become a liability, and a more aggressive regularization grid or early stopping would be a natural safeguard.

Future directions. Three extensions follow directly. (i) Scaling to the full QM9 set and beyond, drawing on larger enumerated chemical spaces [Ruddigkeit et al.(2012)Ruddigkeit, van Deursen, Blum, and and structure-based active sampling [Dral et al.(2017)Dral, Owens, Yurchenko, and Thiel] to maximize information per DFT calculation. (ii) Adding a charge-aware or 3D-aware representation—partial-charge descriptors are already the ninth-most-important feature here, suggesting headroom—and benchmarking against equivariant GNNs [Batzner et al.(2022)Batzner, Musaelian, Sun, Geiger, Mailoa, Kornbluth, Schütt et al.(2018)Schütt, Sauceda, Kindermans, Tkatchenko, and Müller, Gilmer et al.(2017)Gilmer, Schoenholz, and on identical scaffold splits to quantify the gap that 3D information closes. (iii) Extending the same pipeline to the other QM9 electronic properties (HOMO, LUMO, dipole moment, polarizability) and to the broader set of MoleculeNet [Wu et al.(2018)Wu, Ramsundar, Feinberg, Gomes, Geniesse, Pappu, Leschke, and quantum targets, and to data-driven molecular design loops [Gómez-Bombarelli et al.(2018)Gómez-Bombarelli, V. and where a fast, interpretable gap surrogate can guide generative exploration.

5 Conclusion

We have shown that the HOMO–LUMO gap of QM9 molecules can be predicted from 2D molecular structure alone—encoded as 20 RDKit global descriptors and a 1,024-bit radius-2 Morgan fingerprint—by an ensemble of tree-based regressors tuned on a single CPU. On a rigorously leakage-controlled Murcko-scaffold split of a 20,000-molecule subsample (`random_state=42`), XGBoost achieved a test MAE of 0.269 eV, RMSE of 0.377 eV, and $R^2 = 0.914$, with a validation-to-test MAE delta of only +0.004 eV confirming honest generalization to unseen scaffolds. The model rediscovers the physically expected dependence of the gap on sp^3 saturation, and its failures concentrate in a chemically well-defined class—formally charged and zwitterionic molecules—that a 2D fingerprint cannot represent. While equivariant graph neural networks reach lower errors by using 3D coordinates and physical symmetry, this transparent, reproducible, CPU-only pipeline approaches hybrid-DFT reference error and is immediately suitable for high-throughput gap triage and as a strong classical baseline against which coordinate-aware methods can be measured.

Data and code availability

The QM9 dataset is publicly available from figshare [Ramakrishnan et al.(2014)Ramakrishnan, Dral, Rupp, and Wu et al.(2018)Wu, Ramsundar, Feinberg, Gomes, Geniesse, Pappu, Leswing, and Pande]. All analysis scripts, the pinned `uv.lock` lockfile, the trained `.joblib` models, the feature matrices, and the full library-version report (`results/library_versions.txt`) are provided in the session repository, together with the per-step output files `results/case02_metrics.csv`, `results/case02_by_size.csv`, and `results/case02_final_summary.txt`. The entire pipeline is reproducible with `random_state=42`.

Acknowledgements

Computational infrastructure and the scientific-writer agent framework were provided by LitLit AI.

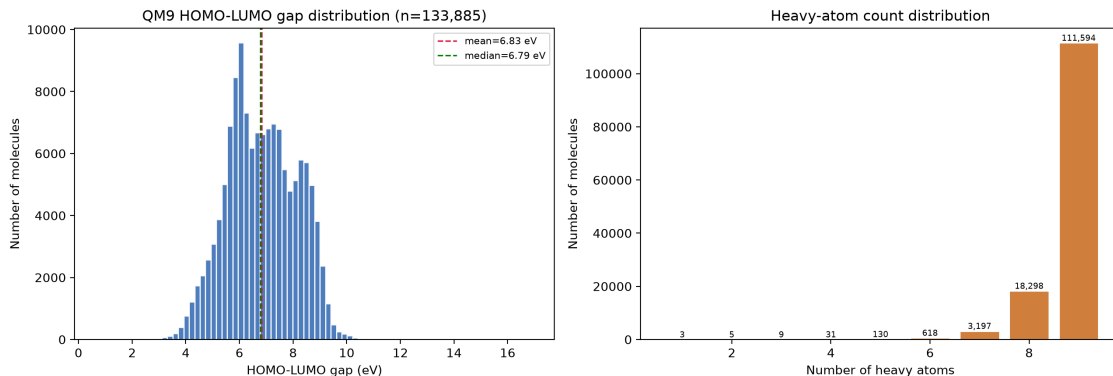


Figure 2: **Distribution of the QM9 HOMO–LUMO gap target** (133,885 molecules, eV). The gap ranges from 0.669 to 16.928 eV with mean 6.833 eV and standard deviation 1.293 eV; the distribution is unimodal and mildly right-skewed.

Table 3: Software environment used for the entire pipeline. All randomness uses `random_state=42`; versions are pinned in the committed `uv.lock` lockfile and in `results/library_versions.txt`.

Library	Version	Role
Python	3.13.5	Interpreter
rdkit	2026.03.4	SMILES parsing, 2D descriptors, Morgan fingerprints
scikit-learn	1.9.0	RandomForest, GradientBoosting, GridSearchCV, metrics
xgboost	3.3.0	Gradient-boosted trees (best-performing model)
pandas	3.0.5	DataFrame I/O and data wrangling
numpy	2.5.1	Numerical arrays and linear algebra
matplotlib	3.11.1	Static plotting
scipy	1.18.0	Statistics (Shapiro–Wilk, skewness, kurtosis)
joblib	1.5.3	Model serialization (<code>.joblib</code>)

References

[Hohenberg and Kohn(1964)] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864–B871, 1964. doi: 10.1103/PhysRev.136.B864.

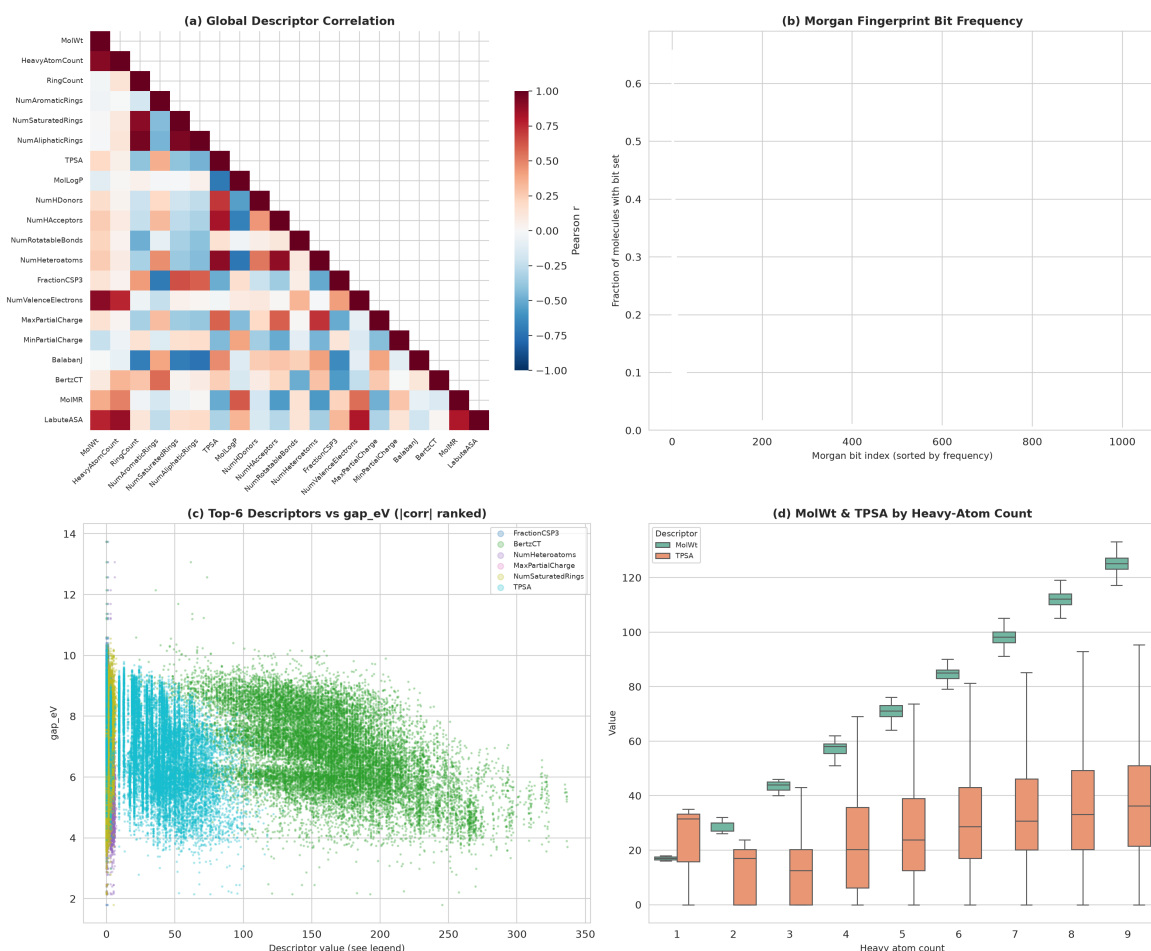


Figure 3: **Exploratory feature analysis.** (a) Distribution of the top global RDKit descriptors, (b) descriptor–target correlations, (c) Morgan-fingerprint bit-density, and (d) the descriptor correlation matrix. The single highly correlated pair (ring count vs. aliphatic-ring count, $|r| = 0.955$) is visible but was retained.

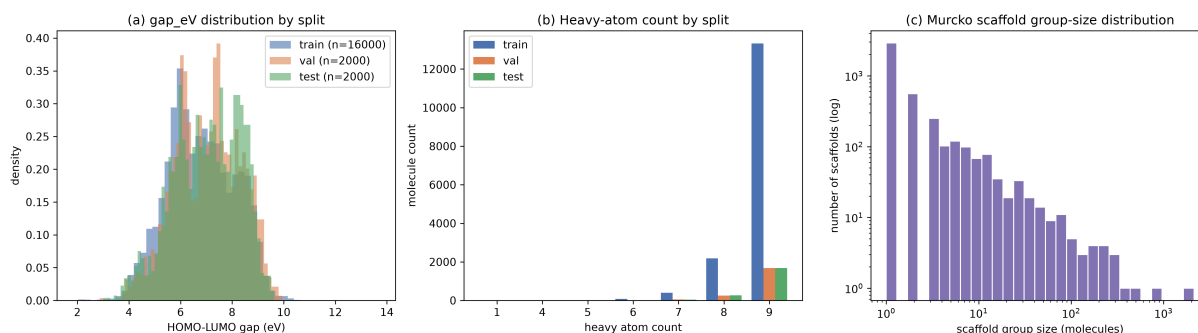


Figure 4: **Sampling and scaffold-split quality.** Left: full-set versus 20,000-molecule subsample target distributions (mean preserved). Centre: per-split target distributions (train 6.775 eV, validation 7.036 eV, test 7.014 eV). Right: heavy-atom-count composition per split. Scaffold overlap between any two splits is exactly zero.

[Kohn and Sham(1965)] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965. doi: 10.1103/PhysRev.140.A1133.

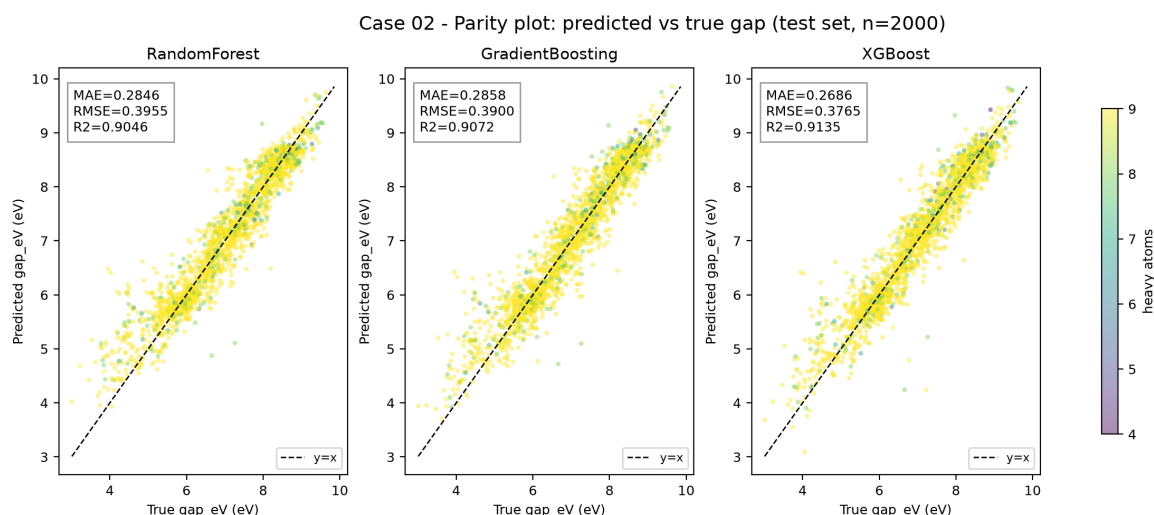


Figure 5: **Predicted-versus-true parity** for the three models on the scaffold-held-out test set ($n = 2,000$). All three ensembles track the diagonal tightly across the bulk of the gap range; deviations widen at the sparse extremes. XGBoost (right) shows the tightest clustering about the identity line.

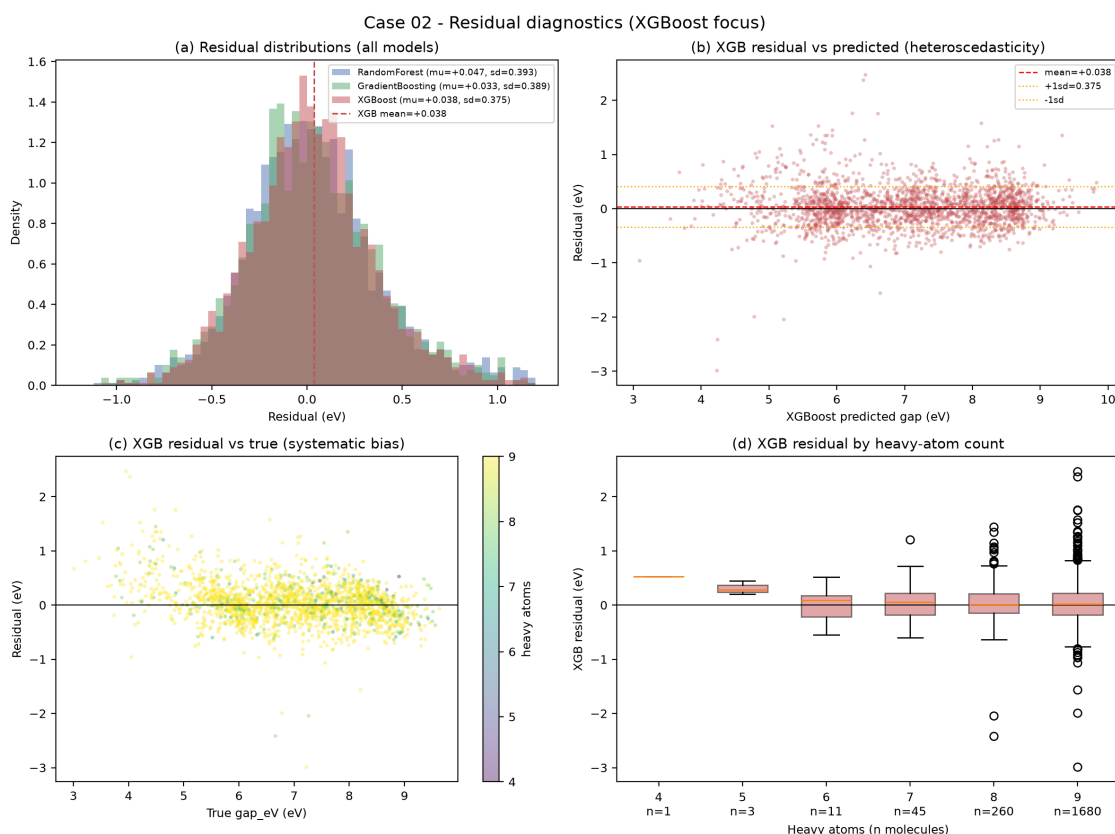


Figure 6: **Residual diagnostics for the best model (XGBoost)**. (a) Residual histogram and kernel density—heavy-tailed, mildly right-skewed (excess kurtosis +6.41). (b) Residuals versus predicted gap, showing no strong heteroscedastic trend but larger errors at the extremes. (c) Normal Q–Q plot, confirming non-Gaussian tails. (d) Empirical cumulative distribution of $|\text{residual}|$: 86.1% of predictions fall within ± 0.5 eV.

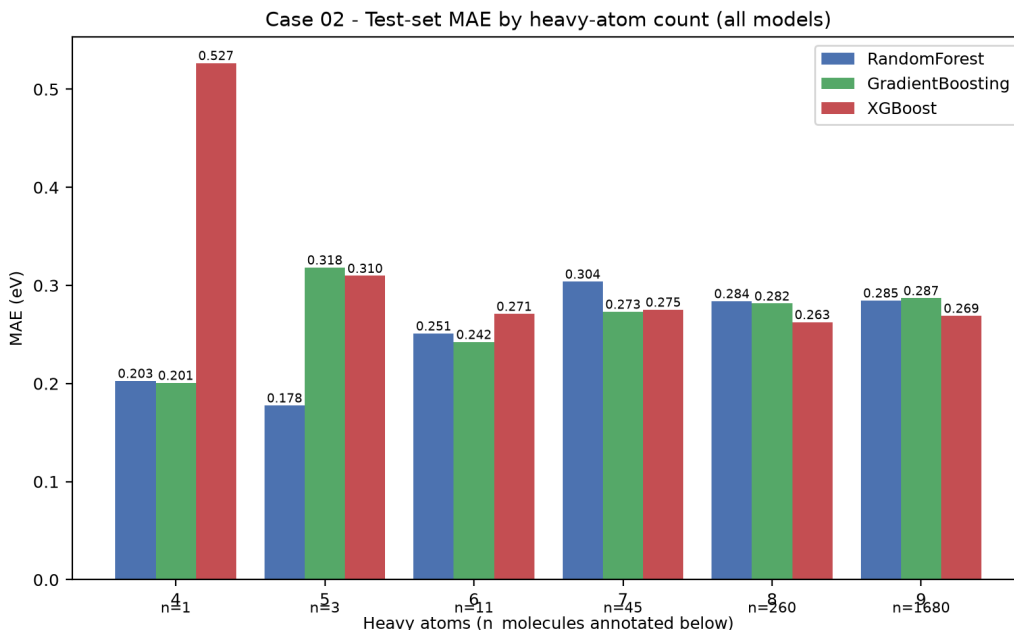


Figure 7: **Test-set MAE by heavy-atom count** for the three models. The well-populated eight- and nine-heavy-atom bins (1,940 of 2,000 molecules) are predicted at ~ 0.26 – 0.27 eV by XGBoost; the four- and five-heavy-atom bins ($n \leq 3$) are statistically fragile.

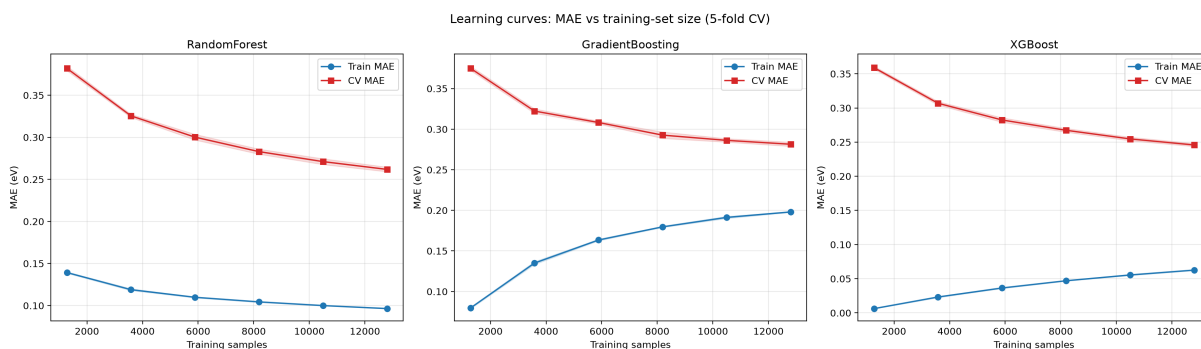


Figure 8: **Train and cross-validation learning curves** for the three models. All three overfit (XGBoost most, gap 0.183 eV; random forest 0.165 eV; gradient boosting 0.083 eV), yet none has plateaued at 16,000 molecules, indicating that additional data would likely further reduce error.

[Behler and Parrinello(2007)] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98 (14):146401, 2007. doi: 10.1103/PhysRevLett.98.146401.

[Ramakrishnan et al.(2014)Ramakrishnan, Dral, Rupp, and von Lilienfeld] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014. doi: 10.1038/sdata.2014.22.

[Becke(1993)] Axel D. Becke. Density-functional thermochemistry. III. the role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993. doi: 10.1063/1.464913.

[Lee et al.(1988)Lee, Yang, and Parr] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, 1988. doi: 10.1103/PhysRevB.37.785.

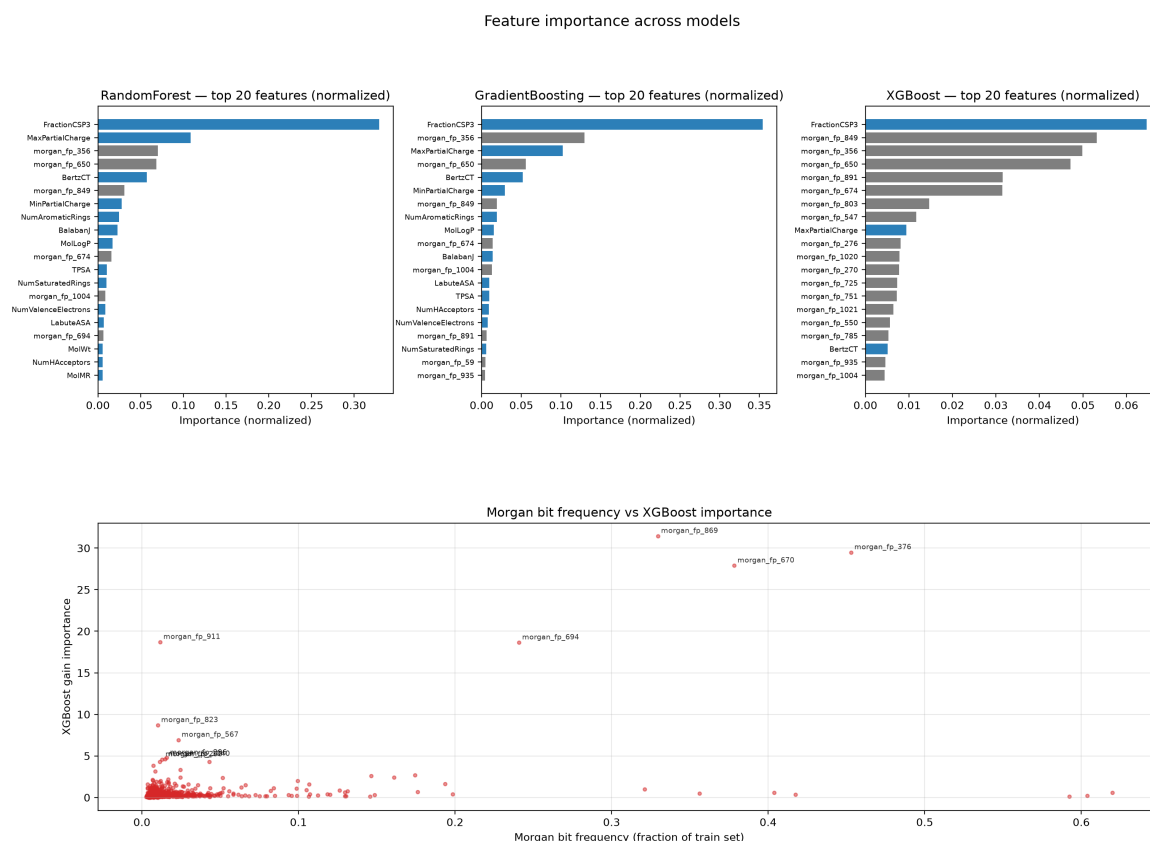


Figure 9: **Feature importance.** Top: the 20 highest-gain features for each model (XGBoost gain, RF/GB impurity). **FractionCSP3** is the leading non-fingerprint descriptor in all three. Bottom: per-Morgan-bit importance scatter, showing that a small subset of circular subenvironments carries most of the predictive signal.

[Stephens et al.(1994)Stephens, Devlin, Chabalowski, and Frisch] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab Initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of Physical Chemistry*, 98(45):11623–11627, 1994. doi: 10.1021/j100096a001.

[Ruddigkeit et al.(2012)Ruddigkeit, van Deursen, Blum, and Reymond] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012. doi: 10.1021/ci300415d.

[Weininger(1988)] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

[Rogers and Hahn(2010)] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[Tropsha(2010)] Alexander Tropsha. Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, 29(6-7):476–488, 2010. doi: 10.1002/minf.201000061.

[Rupp et al.(2012)Rupp, Tkatchenko, Müller, and von Lilienfeld] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate mod-

eling of molecular atomization energies with machine learning. *Physical Review Letters*, 108 (5):058301, 2012. doi: 10.1103/PhysRevLett.108.058301.

[Hansen et al.(2015)Hansen, Biegler, Ramakrishnan, Pronobis, von Lilienfeld, Müller, and Tkatchenko] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The Journal of Physical Chemistry Letters*, 6(12):2326–2331, 2015. doi: 10.1021/acs.jpcllett.5b00831.

[Montavon et al.(2013)Montavon, Rupp, Gobre, Vazquez-Mayagoitia, Hansen, Tkatchenko, Müller, and von Lilienfeld] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, 2013. doi: 10.1088/1367-2630/15/9/095003.

[Dral et al.(2017)Dral, Owens, Yurchenko, and Thiel] Pavlo O. Dral, Alec Owens, Sergei N. Yurchenko, and Walter Thiel. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *The Journal of Chemical Physics*, 146(24):244108, 2017. doi: 10.1063/1.4989536.

[Rumelhart et al.(1986)Rumelhart, Hinton, and Williams] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

[Gilmer et al.(2017)Gilmer, Schoenholz, Riley, Vinyals, and Dahl] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. URL <https://arxiv.org/abs/1704.01212>.

[Schütt et al.(2018)Schütt, Sauceda, Kindermans, Tkatchenko, and Müller] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.

[Batzner et al.(2022)Batzner, Musaelian, Sun, Geiger, Mailoa, Kornbluth, Molinari, Smidt, and Kozinsky] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022. doi: 10.1038/s41467-022-29939-5.

[Yang et al.(2019)Yang, Swanson, Jin, Coley, Eiden, Gao, Guzman-Perez, Hopper, Kelley, Mathea, Palmer, Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59 (8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

[Wu et al.(2018)Wu, Ramsundar, Feinberg, Gomes, Geniesse, Pappu, Leswing, and Pande] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse,

- 427 Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular
428 machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/c7sc02664a.
- 429 [Faber et al.(2017)] Faber, Hutchison, Huang, Gilmer, Schoenholz, Dahl, Vinyals, Kearnes, Riley, and von Lilien
430 Felix A. Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E.
431 Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole von Lilienfeld. Predic-
432 tion errors of molecular machine learning models lower than hybrid DFT error. *Journal of*
433 *Chemical Theory and Computation*, 13(11):5255–5264, 2017. doi: 10.1021/acs.jctc.7b00577.
- 434 [Breiman(2001)] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.
435 1023/A:1010933404324.
- 436 [Ho(1995)] Tin Kam Ho. Random decision forests. In *Proceedings of the 3rd International*
437 *Conference on Document Analysis and Recognition*, volume 1, pages 278–282, 1995. doi:
438 10.1109/ICDAR.1995.598994.
- 439 [Friedman(2001)] Jerome H. Friedman. Greedy function approximation: A gradient boosting
440 machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.
- 441 [Chen and Guestrin(2016)] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting
442 system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*
443 *Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.
- 444 [Bemis and Murcko(1996)] Guy W. Bemis and Mark A. Murcko. The properties of known drugs.
445 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi:
446 10.1021/jm9602928.
- 447 [RDKit Open-Source Cheminformatics(2026)] RDKit Open-Source Cheminformatics. RDKit:
448 Open-source cheminformatics. <https://www.rdkit.org>, 2026. Version 2026.03.4.
- 449 [Pedregosa et al.(2011)] Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss,
450 Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion,
451 Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake
452 Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot,
453 and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine*
454 *Learning Research*, 12:2825–2830, 2011. URL <https://arxiv.org/abs/1201.0490>.
- 455 [Gómez-Bombarelli et al.(2018)] Gómez-Bombarelli, Wei, Duvenaud, Hernández-Lobato, Sánchez-Lengeling, She
456 Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-
457 Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre,
458 Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design
459 using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):
460 268–276, 2018. doi: 10.1021/acscentsci.7b00572.