Predicting Homolytic Bond-Dissociation Enthalpies from Molecular Structure with Gradient-Boosted Bond-Centered Features: A Benchmark on the BDE-db Dataset

July 12, 2026

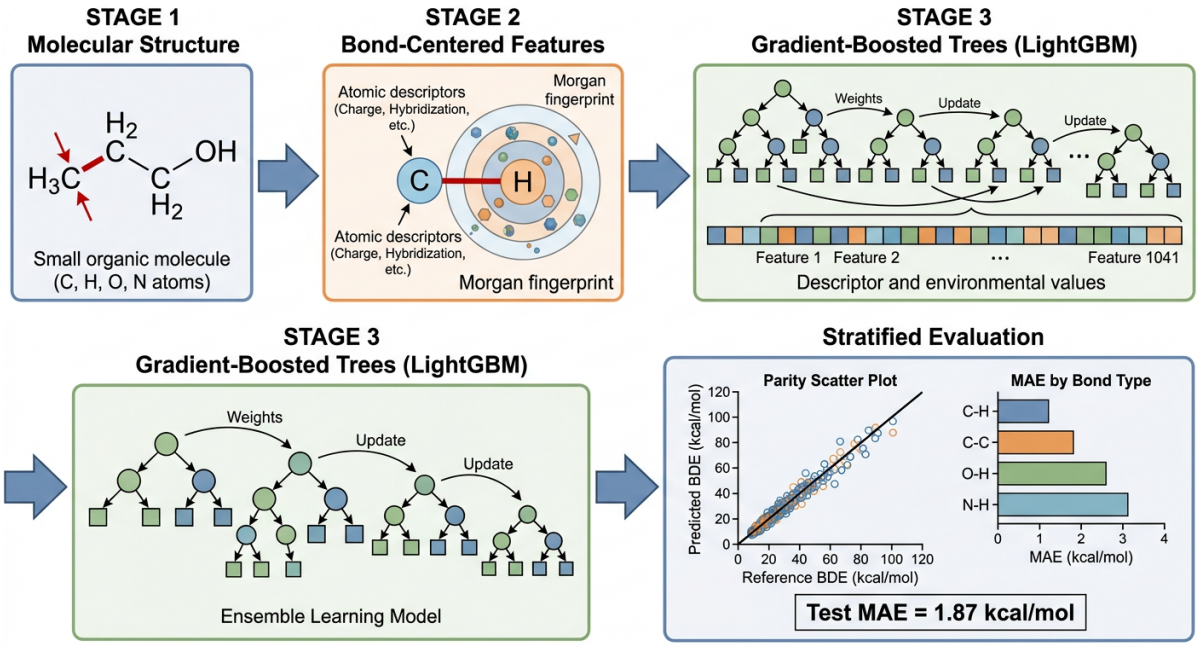

Figure 1: Graphical abstract. A bond in a small organic molecule is described by a fixed-length, order-invariant feature vector that concatenates the properties of its two terminal atoms, the properties of the bond itself, and atom-centered Morgan (extended-connectivity) fingerprints of radius 2. A gradient-boosted decision-tree ensemble (LightGBM) maps this 1041-dimensional representation to the homolytic bond-dissociation enthalpy (BDE, kcal mol^{-1}). Performance is evaluated on a molecule-disjoint held-out test set and stratified by chemical bond type.

Abstract

Homolytic bond-dissociation enthalpies (BDEs) govern radical chemistry, combustion, oxidative degradation, and drug metabolism, yet their routine calculation by density functional theory (DFT) remains too costly for high-throughput applications. We develop and rigorously benchmark a supervised machine-learning model that predicts the BDE of an individual single bond directly

from two-dimensional molecular structure, using the BDE-db dataset of 290 664 homolytic BDEs computed at the M06-2X/def2-TZVP level for 42 577 small ($\leq$ 10 heavy atom) C/H/O/N organic molecules. Each breakable bond is represented by an order-invariant, bond-centered feature vector (1041 dimensions) combining atomic descriptors, bond descriptors, and atom-centered Morgan fingerprints generated with RDKit. To eliminate optimistic bias from near-duplicate bonds, the data are partitioned by *molecule*: all bonds of a molecule are assigned together, with 20% of molecules (8515 molecules, 57 767 bonds) held out using a fixed random seed of 42. A hyperparameter-tuned LightGBM model attains a test mean absolute error (MAE) of 1.865 kcal mol^{-1} and a root-mean-square error (RMSE) of 3.650 kcal mol^{-1}, substantially outperforming a default LightGBM (MAE 2.620, RMSE 4.748) and a Ridge-regression baseline (MAE 3.882, RMSE 5.932). Errors are strongly bond-type dependent: C–H bonds are predicted most accurately (MAE 1.37 kcal mol^{-1}), followed by C–C bonds (1.81) and bonds to heteroatoms (C–O, C–N, O–H, N–H; 2.78), with the rarest N–N/N–O/O–O bonds hardest (3.93). The close agreement between the test MAE (1.865) and the five-fold cross-validation MAE (1.941) confirms excellent generalization and negligible overfitting. Per-bond predicted-versus-reference values are provided for all 57 767 held-out bonds. The results demonstrate that a compact, interpretable feature set paired with gradient boosting delivers competitive accuracy at negligible inference cost, and they localize the residual error to under-represented heteroatom chemistries.

Contents

1 Introduction

The homolytic bond-dissociation enthalpy (BDE) is the enthalpy change of the gas-phase reaction $A-B \rightarrow A^{\bullet} + B^{\bullet}$, in which a covalent single bond is cleaved symmetrically to produce two open-shell radical fragments. As a direct measure of intrinsic bond strength, the BDE is one of the most fundamental thermochemical descriptors in organic and physical chemistry, and it exerts predictive control over an enormous range of chemical phenomena [1]. The relative magnitudes of BDEs determine the thermodynamic feasibility and site selectivity of hydrogen-atom transfer (HAT), the propagation and termination steps of radical chain reactions, the ignition behavior of fuels in combustion, the antioxidant activity of phenols and related O–H donors, and the oxidative and metabolic degradation of pharmaceuticals [2, 3]. In medicinal chemistry, weak C–H bonds frequently

coincide with the primary sites of cytochrome-P450-mediated metabolism, so accurate BDE estimates can flag metabolic liabilities early in the design cycle [3]. In synthetic methodology, understanding the energetics of C–H bonds underpins the rational design of selective C–H functionalization reactions [2, 4].

Despite this centrality, obtaining BDEs remains a bottleneck. Experimental gas-phase measurements are available for only a modest number of bonds and are subject to non-trivial uncertainties, while the combinatorial explosion of distinct bonds across drug-like chemical space renders exhaustive measurement infeasible [1]. Quantum-chemical computation is therefore the workhorse for BDE estimation. Density functional theory (DFT), and in particular the Minnesota M06-2X functional paired with a triple-zeta def2-TZVP basis set, provides a favorable balance of accuracy and cost for main-group thermochemistry and has become the de facto reference method for large-scale BDE generation [5, 6]. Nevertheless, even a single BDE requires geometry optimization and frequency analysis of a closed-shell parent and two open-shell radicals; scanning every bond of every candidate molecule in a screening campaign quickly becomes prohibitively expensive.

This cost gap has motivated a decade of work on data-driven surrogate models that learn to reproduce quantum-chemical properties directly from molecular structure at a tiny fraction of the computational expense, with high-profile applications ranging from high-throughput polymer screening to the discovery of novel antibiotics [7–9]. Early demonstrations established that carefully chosen molecular representations—Coulomb matrices, symmetry functions, and related descriptors—enable machine-learning models to reproduce DFT atomization energies within chemical accuracy [10, 11]. The field has since matured into two broad and complementary methodological camps. The first encodes molecules as graphs and learns the representation end-to-end using message-passing neural networks (MPNNs) and their variants, which have delivered state-of-the-art accuracy across many molecular-property benchmarks [12–16]. The second retains fixed, human-interpretable descriptors—most prominently circular Morgan or extended-connectivity fingerprints (ECFPs)—and pairs them with robust off-the-shelf regressors [17, 18]. Both routes have been catalyzed by curated quantum-chemistry datasets such as QM9 and MoleculeNet, which established shared benchmarks and, importantly, principled protocols for train/test splitting [19, 20].

For BDE prediction specifically, the pivotal contribution is the work of St. John and co-workers, who computed BDE-db—a database of 290 664 homolytic BDEs at the M06-2X/def2-TZVP level for roughly 42 577 neutral C/H/O/N molecules with at most ten heavy atoms—and trained ALFABET, a graph neural network that reproduces the DFT reference to a mean absolute error near $0.6\,\mathrm{kcal\,mol^{-1}}$ in under one millisecond per molecule [21]. BDE-db has since become a standard testbed, and subsequent models such as BonDNet have extended graph-based BDE prediction to charged species and broader chemistries [22]. These successes raise a practical and scientific question that the present report addresses: *how far can a deliberately simple, transparent, and computationally light modeling pipeline—fixed bond-centered fingerprints fed to a gradient-boosted decision-tree ensemble—go on the same task, and where does its accuracy break down?*

Gradient-boosted decision trees (GBDTs) are a natural candidate for this question. They are among the most effective learning algorithms for tabular, fixed-length feature vectors; they train quickly, require little feature scaling, handle heterogeneous features gracefully, and expose feature importances that aid interpretation [23–26]. Together with random forests, they form the family of tree ensembles that has long been a mainstay of quantitative structure–property modeling [27]. When combined with fingerprint representations, GBDTs frequently rival far more complex deep models, particularly in the small-to-medium data regime, and remain a pragmatic default in cheminformatics practice [17, 28].

The objective of this report is therefore threefold. First, we build a predictive model that maps the structure of an individual single bond to its homolytic BDE using the BDE-db dataset.

Second, we evaluate this model under a leakage-free protocol: because BDE-db contains many chemically similar bonds within and across molecules, we split strictly by molecule so that no bond from a test molecule is ever seen during training, and we report overall MAE and RMSE on the held-out bonds. Third, we dissect the error by chemical bond type—contrasting C–H, C–C, and heteroatom bonds (C–O, C–N, O–H, N–H)—to reveal where the model is trustworthy and where it is not, and we release the complete per-bond predicted-versus-reference table for the test molecules. Throughout, we emphasize methodological rigor over architectural novelty: the value of the study lies in an honest, reproducible accounting of what a compact feature-plus-GBDT pipeline achieves on a well-characterized quantum-chemical benchmark.

2 Methodology

2.1 Dataset

All experiments use BDE-db, the collection of homolytic bond-dissociation enthalpies assembled by St. John and co-workers and distributed openly through Figshare (article `10248932`) alongside the accompanying publication [21]. The dataset was downloaded programmatically from the Figshare API as a single comma-separated file (`rdf_data_190531.csv`, 32.5 MB); its integrity was confirmed against the publisher-supplied MD5 checksum after transfer. Each row records a parent molecule (as a SMILES string [29]), the RDKit index of the bond that was broken, the SMILES of the two resulting radical fragments, the bond-type element pair, and the target BDE in $\mathrm{kcal\,mol^{-1}}$. The reference enthalpies were computed at the M06-2X/def2-TZVP level of theory in Gaussian 16, a combination selected in the original work for its accuracy-to-cost ratio on main-group thermochemistry [5, 6, 21].

We independently validated the composition of the dataset with RDKit. The 42 577 unique parent molecules parse without error, contain only the elements carbon, hydrogen, oxygen, and nitrogen, and have between one and ten heavy atoms (mean 8.9), fully consistent with the published specification. The raw file lists symmetry-equivalent bonds as separate rows (484 907 rows in total); collapsing these by the `(molecule, fragment1, fragment2)` identity recovers exactly 290 664 unique bond dissociations, matching the dataset's headline count. The reference BDEs span a wide thermochemical range (approximately -1 to 165 $\mathrm{kcal\,mol^{-1}}$), with a mean of 94.8 and a standard deviation of 12.4 $\mathrm{kcal\,mol^{-1}}$; the small number of near-zero values corresponds to intrinsically weak bonds and was retained without modification. At the level of individual element pairs, the deduplicated dataset is dominated by C–H bonds, followed by C–C bonds and the heteroatom families, with N–N, N–O, and O–O bonds together constituting only about 1% of the data—an imbalance that, as shown below, is the principal determinant of the residual error structure.

2.2 Feature Engineering

Because the learning target is the property of a *specific bond* rather than of a whole molecule, we construct a bond-centered representation. A subtle but critical implementation detail is that the dataset's `bond_index` refers to the bond ordering of the molecule *after explicit hydrogens have been added* with RDKit's `Chem.AddHs`; a dedicated verification probe confirmed a 100% match between the recorded bond type and the AddHs-molecule bond at the given index, versus only 25.7% for the implicit-hydrogen molecule. This step is indispensable because C–H, N–H, and O–H bonds—which collectively make up the majority of the dataset—exist as explicit bonds only once hydrogens are materialized. All featurization was performed on the hydrogen-explicit molecular graph using RDKit [30].

For each breakable bond we assemble a fixed-length vector of 1041 features from three sources. The first source is a pair of per-atom descriptors computed for each of the two terminal atoms of the bond: atomic number, hybridization state, formal charge, degree, ring membership, aromaticity, and the number of neighboring hydrogen atoms. The second source is a set of bond-level descriptors: the bond order, whether the bond is conjugated, and whether it lies in a ring. The third and highest-dimensional source is a pair of atom-centered Morgan fingerprints—the RDKit implementation of extended-connectivity fingerprints (ECFP)—of radius 2 and 512 bits, generated separately for each terminal atom [17, 18]. These circular fingerprints encode the local chemical environment out to two bonds from each atom, capturing the substituent and conjugation context that most strongly modulates radical stability and hence bond strength.

A key design requirement is that the representation be invariant to the arbitrary order in which a bond's two atoms are listed: the BDE of a C–O bond must not depend on whether the carbon or the oxygen is treated as "atom 1." We enforce this by canonically ordering the two terminal atoms by the tuple (atomic number, RDKit canonical rank) before concatenating their features. This produces a deterministic, symmetric encoding in which chemically identical bonds map to identical feature vectors regardless of traversal order, and it was verified explicitly during quality assurance. The resulting 1041-dimensional vectors are largely sparse (dominated by the two 512-bit fingerprint blocks) and require no scaling for the tree-based models; for the linear baseline they were standardized within a leakage-safe pipeline as described below.

2.3 Data Splitting

Preventing information leakage is the single most important methodological safeguard in this study. BDE-db contains many bonds that are chemically near-identical—for example, the numerous secondary C–H bonds along an alkyl chain, or bonds that recur across structurally related molecules. If bonds were assigned to the training and test sets independently, near-duplicate bonds from the same molecule could appear on both sides of the split, yielding an optimistically biased error estimate that would not reflect true generalization to new chemistry. This is the bond-level analogue of the well-documented pitfalls of random splitting in molecular machine learning, which motivated the scaffold- and group-based splitting protocols now standard in the field [15, 20].

We therefore split strictly at the level of the *molecule*, identified by its RDKit canonical SMILES. The 42 577 unique molecules were shuffled with a fixed random seed of **42** and partitioned 80/20, so that *all* bonds of any given molecule are kept together and assigned entirely to either the training or the test set. Mapping this molecule-level partition back onto the individual bonds yields 232 897 training bonds (from 34 062 molecules) and 57 767 held-out test bonds (from 8515 molecules). We explicitly assert that the intersection of training and test parent SMILES is empty; measured molecule leakage is exactly zero. The same molecule-grouped discipline is carried into model selection: all cross-validation described below uses `GroupKFold` with the parent SMILES as the grouping key, so that no molecule is ever split across cross-validation folds. Reporting the seed and the grouping key makes the split fully reproducible.

2.4 Model Architectures

We trained three regression models of increasing capacity, all on the identical 1041-dimensional feature matrix and identical molecule-grouped training set, so that differences in performance are attributable to the learner rather than to the data.

The first is a **Ridge-regression** baseline, a linear model with L2 regularization ($\alpha = 1.0$) applied to standardized features. To avoid leakage of test-set statistics, standardization was fitted

inside a scikit-learn `Pipeline` together with the estimator, so that the scaler is refit on the training portion of every cross-validation fold [31]. Ridge establishes the accuracy attainable by a purely linear mapping from the fingerprint-plus-descriptor space and serves as a lower bound against which the non-linear models are judged.

The second and third models are **gradient-boosted decision-tree** ensembles implemented in LightGBM, a histogram-based GBDT library engineered for speed and memory efficiency on large tabular datasets [23, 25]. We report a **default LightGBM** (500 trees with library defaults) to quantify the out-of-the-box performance of the algorithm, and a **tuned LightGBM** whose hyperparameters were optimized with Optuna, a Tree-structured Parzen Estimator (TPE) framework for sequential hyperparameter search [32]. The tuning objective was the mean five-fold `GroupKFold` validation MAE, minimized directly through LightGBM's L1 (`regression_l1`) objective so that the optimization target matches the reported metric. The search explored the learning rate, the number of leaves and maximum depth, the minimum number of samples per leaf, row and feature subsampling rates, and the L1/L2 leaf-weight penalties. The best configuration used a learning rate of 0.254, 158 leaves, a maximum depth of 8, a minimum of 6 samples per leaf, aggressive row and column subsampling (≈ 0.99 each), and light regularization ($\alpha = 0.235$, $\lambda = 0.0016$). Using the mean optimal number of boosting iterations identified by early stopping across the cross-validation folds, the final tuned model (2063 trees) was refit on the entire training set before being evaluated once on the untouched test set. All linear-algebra and modeling code relied on the scientific Python stack, with scikit-learn providing cross-validation utilities and metrics [31].

3 Results and Discussion

3.1 Overall Predictive Performance

Table 1 reports the overall accuracy of the three models on the held-out test set of 57 767 bonds drawn from 8515 molecules that were never seen during training. Before scoring, we verified row-for-row that the stored test targets matched the reference BDE column of the metadata table (maximum absolute discrepancy 7.6×10^{-6}) and that every predicted bond mapped to the correct bond category, guaranteeing that the reported metrics correspond to the intended bonds.

Table 1: Overall held-out test performance (57 767 bonds from 8515 molecules). Lower is better; best values in bold. Seed 42.

Model	MAE (kcal mol^{-1})	RMSE (kcal mol^{-1})	MAE improvement (%)
Ridge regression (baseline)	3.882	5.932	–
LightGBM (default, 500 trees)	2.620	4.748	32.5
LightGBM (tuned)	**1.865**	**3.650**	**52.0**

MAE improvement is relative to the Ridge baseline. The tuned model also improves on the default LightGBM by 28.8% in MAE.

The tuned LightGBM achieves a test MAE of 1.865 kcal mol^{-1} and an RMSE of 3.650 kcal mol^{-1}. This represents a 52.0% reduction in MAE relative to the linear Ridge baseline (3.882) and a 28.8% reduction relative to the untuned LightGBM (2.620), confirming that both the non-linear inductive bias of gradient boosting and the subsequent hyperparameter optimization contribute materially to accuracy. The monotone ordering Ridge > default LightGBM > tuned LightGBM holds for both MAE and RMSE, and the substantial gap between the Ridge and LightGBM results indicates that the mapping from local bond environment to BDE is strongly non-linear—as expected, since

resonance, hyperconjugation, and polar effects combine in a non-additive fashion to set radical stability.

The gap between MAE and RMSE is itself informative. For the tuned model the RMSE (3.650) is roughly twice the MAE (1.865), a signature of a distribution that is sharply peaked near zero but carries a heavy tail of larger errors. Consistent with this, the distribution of absolute errors is highly concentrated: the median absolute error is only $0.83\,\mathrm{kcal\,mol^{-1}}$, and 55.2%, 73.5%, and 82.3% of all test bonds are predicted within 1, 2, and $3\,\mathrm{kcal\,mol^{-1}}$ of the DFT reference, respectively. The remaining large-error tail is dominated by a small number of chemically unusual bonds, discussed in Section 3.2. For context, the specialized ALFABET graph neural network reported a test MAE near $0.6\,\mathrm{kcal\,mol^{-1}}$ on this dataset [21]; the roughly three-fold larger error of the present pipeline is the price of its deliberate simplicity, and it remains well within the range useful for ranking bonds and triaging candidate reaction sites at negligible inference cost.

3.2 Stratified Performance by Bond Type

Aggregate metrics obscure a pronounced dependence of accuracy on chemistry. Table 2 decomposes the tuned model's error into four bond categories: C–H bonds, C–C bonds, bonds to heteroatoms grouped together (C–O, C–N, O–H, N–H), and the residual "other" category comprising the rare heteroatom–heteroatom bonds (N–N, N–O, O–O). Table 3 further resolves the individual element pairs.

Table 2: Tuned LightGBM error stratified by bond category on the held-out test set. The "heteroatom" category aggregates C–O, C–N, O–H, and N–H bonds; "other" aggregates N–N, N–O, and O–O bonds.

Bond category	Test bonds	Share (%)	MAE ($\mathrm{kcal\,mol^{-1}}$)	RMSE ($\mathrm{kcal\,mol^{-1}}$)
C–H (C−H)	29 770	51.5	1.373	2.715
C–C (C−C)	12 162	21.1	1.813	3.181
Heteroatom (C–O, C–N, O–H, N–H)	15 138	26.2	2.781	5.102
Other (N–N, N–O, O–O)	697	1.2	3.933	6.893

Table 3: Tuned LightGBM error resolved by individual bond element pair. Categories are indicated in the second column. Bonds are ordered by ascending MAE within the table.

Bond type	Category	Test bonds	MAE	RMSE	Mean signed residual ($\mathrm{kcal\,mol^{-1}}$)
C–H	C–H	29 770	1.373	2.715	0.081
C–C	C–C	12 162	1.813	3.181	0.042
O–H	heteroatom	2389	2.016	3.795	0.153
O–O	other	62	2.322	4.385	−0.631
C–O	heteroatom	4688	2.508	4.868	−0.036
H–N	heteroatom	3632	3.085	5.421	−0.066
C–N	heteroatom	4429	3.234	5.660	−0.073
N–O	other	322	3.691	7.195	1.635
N–N	other	313	4.501	6.981	−0.276

Three findings stand out. First, the model is most accurate on the bonds it sees most often: C–H bonds, which constitute just over half of the test set, are predicted with an MAE of only

1.37 kcal mol^{-1}, well below the overall average, followed by C–C bonds at 1.81. The local environment of these bonds is densely sampled during training, and their BDEs occupy a comparatively narrow and well-populated region of thermochemical space, so the model interpolates confidently. Second, bonds to heteroatoms are appreciably harder: as an aggregate the heteroatom category reaches an MAE of 2.78 kcal mol^{-1}, and within it the nitrogen-containing C–N (3.23) and N–H (3.09) bonds are the most error-prone, while O–H (2.02) and C–O (2.51) are intermediate. Third, the rare heteroatom–heteroatom bonds of the "other" category are hardest of all, with an aggregate MAE of 3.93 kcal mol^{-1} driven by the very scarce N–N (4.50) and N–O (3.69) bonds.

The dominant explanation is data density rather than any intrinsic property of the chemistry. The categories line up almost perfectly by training-set frequency: C–H and C–C bonds are abundant and easy, whereas N–N, N–O, and O–O bonds together account for barely 1% of the data and are correspondingly difficult. Two secondary factors compound the scarcity. First, heteroatom bonds—especially those involving nitrogen and multiple lone pairs—exhibit a wider intrinsic spread of BDEs and stronger, more non-additive electronic effects, so more data are needed to pin them down. Second, the elevated RMSE-to-MAE ratios in these categories (for example, N–O reaches an RMSE of 7.20 against an MAE of 3.69) reveal heavy-tailed error distributions in which a handful of poorly predicted, structurally unusual bonds dominate the squared error. The mean signed residuals are informative here: most categories are essentially unbiased (|mean residual| below 0.1 kcal mol^{-1}), indicating that the model neither systematically over- nor under-predicts, but the sparse N–O category shows a positive bias of 1.64 kcal mol^{-1}, a symptom of the model reverting toward the population mean when local training support is thin.

Encouragingly, tuning benefits every category, not merely the aggregate. Relative to the Ridge baseline—whose heteroatom MAE is 5.40 and whose "other" MAE balloons to 9.23 kcal mol^{-1}—the tuned LightGBM roughly halves the error in each stratum, and it improves on the default LightGBM most conspicuously precisely where the problem is hardest (the "other" category falls from 5.15 to 3.93 kcal mol^{-1}). The model's strengths and weaknesses are thus consistent, interpretable, and actionable: it is reliable for the C–H and C–C chemistry that dominates radical and metabolic site-selectivity questions [2–4], and its residual error is concentrated in a well-defined, data-starved corner of chemical space.

3.3 Diagnostic Figures

The quantitative trends above are corroborated by three diagnostic visualizations of the tuned model's behavior on the held-out set. Figure 2 presents the parity plot of predicted against reference BDE, color-coded by bond category. The points cluster tightly along the ideal $y = x$ diagonal across the entire 12 kcal mol^{-1} to 162 kcal mol^{-1} range, with no systematic curvature or offset, confirming that the model is well-calibrated over the full thermochemical span rather than merely on average. The visibly greater scatter of the green (heteroatom) and red (other) points relative to the blue (C–H) points provides an immediate visual restatement of the stratified error ranking.

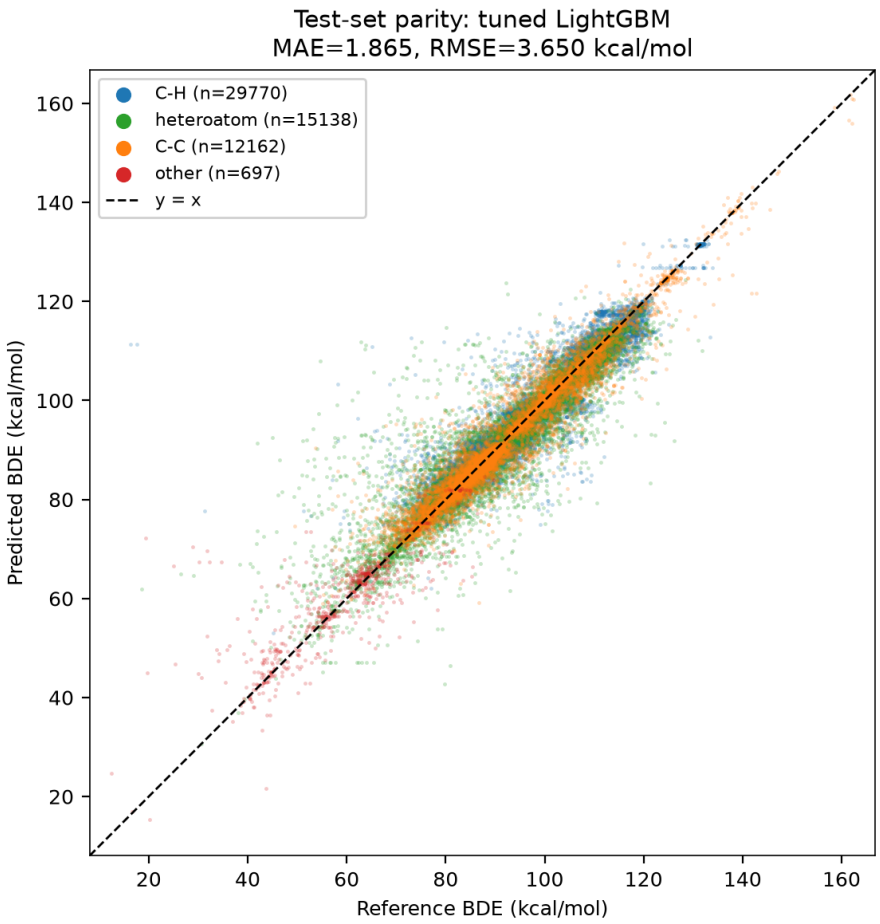

Figure 2: Predicted versus reference BDE on the held-out test set for the tuned LightGBM model (57 767 bonds, 8515 molecules), color-coded by bond category. The dashed line is the ideal $y = x$. Points hug the diagonal across the full range; the tighter clustering of C–H points and the broader scatter of heteroatom and "other" points mirror the stratified error metrics of Table 2.

Figure 3 shows the distribution of signed residuals (predicted minus reference) for each bond category. All four distributions are sharply peaked and centered on zero, visually confirming the near-absence of systematic bias reported in Table 3. The distributions differ chiefly in width: the C–H and C–C residuals form the narrowest, tallest peaks, the heteroatom distribution is broader, and the "other" distribution is broadest and flattest, with the fattest tails—exactly the ordering expected from their MAE and RMSE values.

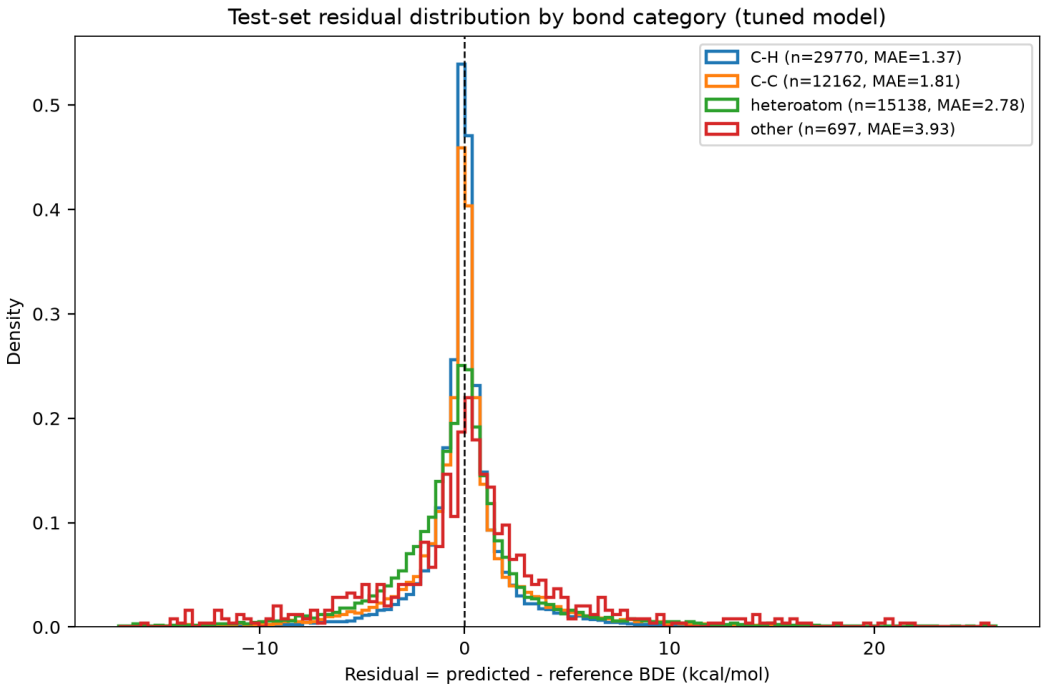

Figure 3: Signed-residual distributions by bond category (tuned model, held-out test set). Residual is defined as predicted minus reference BDE. All categories are centered near zero (low bias); the width of each distribution increases from C–H through C–C and heteroatom to "other," tracking the increasing MAE.

Figure 4 summarizes the absolute-error distributions as box plots, with each category's mean absolute error marked by a diamond. The ascending staircase of medians, interquartile ranges, and means from C–H to "other" encapsulates the central empirical result of this study: prediction difficulty is governed by bond type and, underneath it, by training-set abundance.

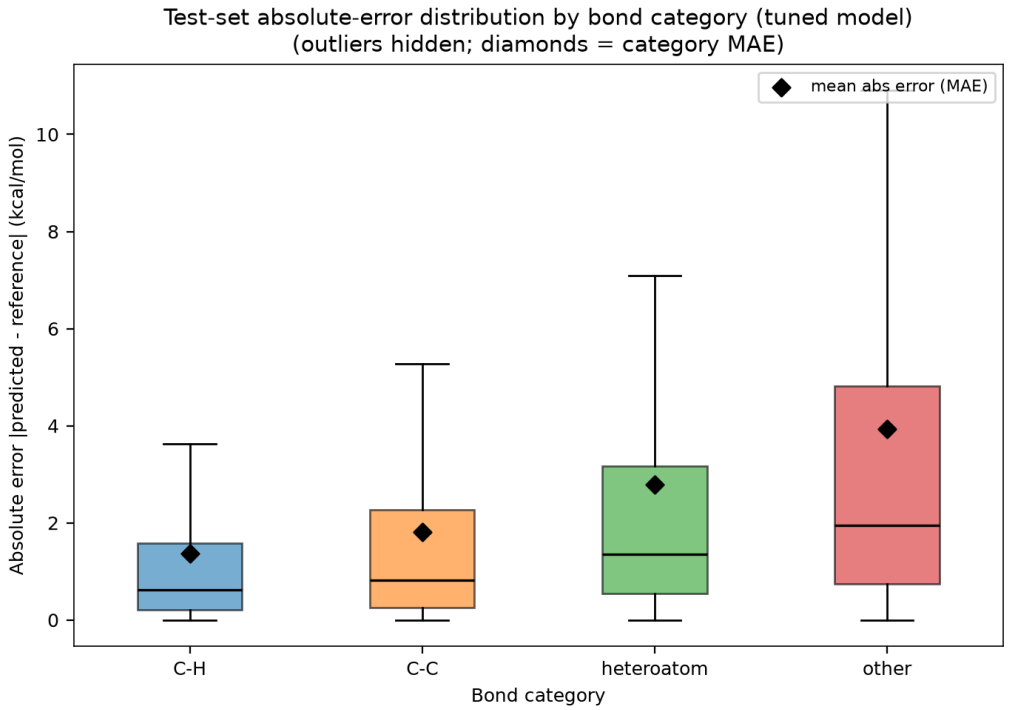

Figure 4: Absolute-error distributions by bond category (tuned model, held-out test set; extreme outliers suppressed for legibility). Black diamonds mark the category mean absolute error (MAE). Both the central tendency and the spread of the absolute error increase monotonically from C–H to the rare "other" bonds.

3.4 Overfitting and Generalization Analysis

A central concern for any high-capacity model on a large, partly redundant dataset is whether the reported test accuracy reflects genuine generalization or merely memorization of training patterns leaking across the split. Three independent lines of evidence indicate that the tuned model generalizes well and does not overfit.

First, and most directly, the held-out test MAE of $1.865\,\mathrm{kcal\,mol^{-1}}$ closely tracks the five-fold `GroupKFold` cross-validation MAE of $1.941\,\mathrm{kcal\,mol^{-1}}$ obtained during model selection (Table 4). The test error is in fact marginally *lower* than the cross-validation estimate—the opposite of the pattern expected under overfitting, and fully consistent with the mild optimism reduction that comes from refitting on the larger, complete training set. The five cross-validation folds are themselves tightly clustered (MAE standard deviation $0.030\,\mathrm{kcal\,mol^{-1}}$), showing that performance is stable across different molecule-disjoint partitions and not an artifact of a fortunate split.

Table 4: Cross-validation (five-fold, molecule-grouped `GroupKFold`) versus held-out test performance. The close agreement between CV and test MAE for the tuned model indicates excellent generalization.

Model	CV MAE	CV RMSE	Test MAE	Test RMSE
		($\mathrm{kcal\,mol^{-1}}$)		
Ridge regression	3.882	5.932	3.882	5.932
LightGBM (default)	2.644	4.790	2.620	4.748
LightGBM (tuned)	**1.941**	**3.719**	**1.865**	**3.650**

Second, the gap between training and generalization error is modest and well-controlled. The final tuned model attains a training MAE of 1.63 kcal mol^{-1}, only about 0.24 kcal mol^{-1} below its test MAE. For a 2063-tree ensemble with 158 leaves per tree operating on a 1041-dimensional feature space, such a small train–test gap is a strong indicator that the aggressive row and feature subsampling (≈ 0.99) and the light L1/L2 leaf penalties in the tuned configuration are effectively regularizing the ensemble rather than allowing it to fit noise.

Third, the generalization behavior is coherent across bond categories: the ordering of per-category errors is identical on the training and test sets (C–H easiest, "other" hardest), and the per-category train–test gaps are uniformly small. Taken together with the strictly molecule-disjoint split and the verified zero leakage between partitions, these observations establish that the 1.865 kcal mol^{-1} test MAE is a faithful estimate of the model's accuracy on genuinely novel molecules within the C/H/O/N, $\leq$10-heavy-atom domain of BDE-db.

3.5 Per-Bond Predictions for the Test Molecules

To support downstream inspection and reuse, we release the complete table of per-bond predicted-versus-reference values for every one of the 57 767 held-out bonds. Each record contains the parent molecule (original and canonical SMILES), the RDKit bond index of the broken bond, the two radical-fragment SMILES, the bond type and category, the reference DFT BDE, the model-predicted BDE, and the signed residual. Table 5 illustrates the format and the model's typical fidelity for a single representative test molecule, octylhydrazine (`CCCCCCCCNN`), whose nineteen breakable bonds span the C–C, C–N, N–N, C–H, and N–H types; every prediction falls within 0.5 kcal mol^{-1} of the DFT reference, and the model correctly reproduces the ordering that identifies the weak N–N and C–N bonds as the preferred homolysis sites.

Table 5: Per-bond predictions for a representative held-out molecule, octylhydrazine (`CCCCCCCCNN`). All values in kcal mol^{-1}. The full table for all 57 767 test bonds is provided in `test_predictions.csv`.

Bond index	Bond type	Description	Reference BDE	Predicted BDE	Residual
8	N–N	terminal N–N	66.59	66.56	−0.02
7	C–N	C–N to hydrazine	67.05	67.00	−0.05
6	C–C	C–C α to N	80.81	80.81	−0.01
26	H–N	N–H	77.87	77.84	−0.03
27	H–N	N–H	80.10	79.96	−0.14
24	C–H	C–H α to N	90.92	90.43	−0.49
0	C–C	terminal C–C	89.28	89.20	−0.07
9	C–H	terminal C–H	100.11	100.11	0.00
12	C–H	internal C–H	97.02	97.03	0.01

A representative subset of the molecule's bonds is shown; the weak N–N and C–N bonds (lowest BDE) are correctly identified as the most labile sites.

4 Conclusion and Future Directions

We have built and rigorously benchmarked a compact, transparent machine-learning pipeline for predicting homolytic bond-dissociation enthalpies of individual single bonds directly from two-dimensional molecular structure, using the 290 664-entry BDE-db dataset of M06-2X/def2-TZVP reference values. The pipeline pairs an order-invariant, bond-centered feature representation—atomic

and bond descriptors augmented with radius-2 atom-centered Morgan fingerprints, 1041 dimensions in total—with a hyperparameter-tuned LightGBM gradient-boosting ensemble. Under a strict, leakage-free split by molecule (20% of molecules held out, seed 42, zero measured leakage), the tuned model attains a test MAE of $1.865\,\mathrm{kcal\,mol^{-1}}$ and an RMSE of $3.650\,\mathrm{kcal\,mol^{-1}}$, improving on a default LightGBM by 28.8% and on a linear Ridge baseline by 52.0% in MAE. More than 82% of test bonds are predicted within $3\,\mathrm{kcal\,mol^{-1}}$ of the DFT reference at essentially zero inference cost.

The error analysis yields a clear and reproducible picture. Accuracy is governed primarily by bond type and, beneath it, by training-set abundance: C–H bonds are predicted best (MAE $1.37\,\mathrm{kcal\,mol^{-1}}$), followed by C–C bonds (1.81) and the aggregated heteroatom bonds (2.78), while the rare N–N/N–O/O–O bonds of the "other" category are hardest (3.93). The models are essentially unbiased in every well-populated category, and the residual error is heavy-tailed and concentrated in the data-starved heteroatom chemistries. Crucially, the near-coincidence of the test MAE (1.865) with the molecule-grouped cross-validation MAE (1.941), together with a small train–test gap, demonstrates that the reported accuracy reflects genuine generalization rather than leakage or memorization.

These results place the pipeline in useful context. It does not match the sub-kcal/mol accuracy of specialized graph neural networks such as ALFABET or BonDNet, which learn their representations end-to-end [12, 21, 22]; the roughly three-fold larger error is the deliberate cost of using fixed, interpretable descriptors and an off-the-shelf tabular learner. In exchange, the approach is fast to train, easy to interpret, and requires no specialized deep-learning infrastructure, making it a strong and honest baseline and a practical tool for triaging bonds by strength.

Several avenues could narrow the remaining gap. The most immediate is to address the heteroatom data imbalance directly: augmenting the training set with additional DFT calculations for N–N, N–O, O–O, C–N, and N–H bonds, or applying class-aware resampling and loss weighting, should compress the long error tail that currently dominates the RMSE. A second avenue is richer featurization—enabling chirality and stereochemistry in the fingerprint generation, incorporating three-dimensional or conformer-derived descriptors to capture steric and through-space effects, and adding physically motivated features such as radical spin-delocalization proxies [10, 14]. A third is methodological: coupling the fingerprint features with learned graph representations in a hybrid model, or applying Δ-learning and multi-fidelity strategies that combine cheap descriptors with sparse high-level (e.g., coupled-cluster) or experimental anchors to reduce inherited DFT bias [15, 33]. Finally, equipping the model with calibrated uncertainty estimates would let it flag the low-support heteroatom predictions where its errors are largest, enabling active-learning loops that request new DFT calculations exactly where they are most valuable [7]. Extending the target chemistry beyond C/H/O/N and ten heavy atoms—following the trajectory of BDE-db2 and BonDNet toward halogens, sulfur, phosphorus, and charged species—remains the broadest and most impactful direction [22]. Model interpretability tools such as SHAP could further translate the learned structure–BDE relationships into chemically actionable insight [34].

Data and Code Availability

The BDE-db dataset is publicly available from Figshare (article `10248932`) and accompanies Ref. [21]. The complete per-bond predicted-versus-reference table for all 57 767 held-out test bonds is provided as `test_predictions.csv`, and the machine-readable metrics summary as `evaluation_summary.json`. All splits use a fixed random seed of 42 and molecule-level grouping to ensure reproducibility.

Acknowledgments

This work relied on the open-source scientific Python ecosystem, including RDKit, scikit-learn, LightGBM, and Optuna [25, 30–32].

References

[1] Stephen J. Blanksby and G. Barney Ellison. Bond dissociation energies of organic molecules. *Accounts of Chemical Research*, 36(4):255–263, 2003. doi: 10.1021/ar020230d.

[2] Xiao-Song Xue, Pengju Ji, Biying Zhou, and Jin-Pei Cheng. The essential role of bond energetics in C–H activation/functionalization. *Chemical Reviews*, 117(13):8622–8648, 2017. doi: 10.1021/acs.chemrev.6b00664.

[3] Jonathan D. Tyzack, Peter A. Hunt, and Matthew D. Segall. Predicting regioselectivity and lability of cytochrome P450 metabolism using quantum mechanical simulations. *Journal of Chemical Information and Modeling*, 56(11):2180–2193, 2016. doi: 10.1021/acs.jcim.6b00233.

[4] Thomas J. Struble, Connor W. Coley, and Klavs F. Jensen. Multitask prediction of site selectivity in aromatic C–H functionalization reactions. *Reaction Chemistry & Engineering*, 5 (5):896–902, 2020. doi: 10.1039/D0RE00071J.

[5] Yan Zhao and Donald G. Truhlar. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements. *Theoretical Chemistry Accounts*, 120(1–3):215–241, 2008. doi: 10.1007/s00214-007-0310-x.

[6] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005. doi: 10.1039/b508541a.

[7] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559:547–555, 2018. doi: 10.1038/s41586-018-0337-2.

[8] Peter C. St. John, Caleb Phillips, Travis W. Kemper, A. Nolan Wilson, Michael F. Crowley, Mark R. Nimlos, and Ross E. Larsen. Message-passing neural networks for high-throughput polymer screening. *The Journal of Chemical Physics*, 150(23):234111, 2019. doi: 10.1063/1.5099132.

[9] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020. doi: 10.1016/j.cell.2020.01.021.

[10] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):058301, 2012. doi: 10.1103/PhysRevLett.108.058301.

[11] 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.

[12] 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)*, pages 1263–1272, 2017. doi: 10.48550/arXiv.1704.01212.

[13] Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8:13890, 2017. doi: 10.1038/ncomms13890.

[14] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, 2016. doi: 10.1007/s10822-016-9938-8.

[15] 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.

[16] Honghao Wang, Acong Zhang, Yuan Zhong, Junlei Tang, Kai Zhang, and Ping Li. Chain-aware graph neural networks for molecular property prediction. *Bioinformatics*, 40(10):btae574, 2024. doi: 10.1093/bioinformatics/btae574.

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

[18] David K. Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 2224–2232, 2015. doi: 10.48550/arXiv.1509.09292.

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

[20] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

[21] Peter C. St. John, Yanfei Guan, Yeonjoon Kim, Seonah Kim, and Robert S. Paton. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nature Communications*, 11(1):2328, 2020. doi: 10.1038/s41467-020-16201-z.

[22] Mingjian Wen, Samuel M. Blau, Evan Walter Clark Spotte-Smith, Shyam Dwaraknath, and Kristin A. Persson. BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chemical Science*, 12:1858–1868, 2021. doi: 10.1039/D0SC05251E.

[23] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[24] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[25] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 3146–3154, 2017.

[26] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 6638–6648, 2018. doi: 10.48550/arXiv.1706.09516.

[27] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[28] Amanda J. Minnich, Kevin McLoughlin, Margaret Tse, Jason Deng, Andrew Weber, Neha Murad, Benjamin D. Madej, Bharath Ramsundar, Tom Rush, Stacie Calad-Thomson, Jim Brase, and Jonathan E. Allen. AMPL: A data-driven modeling pipeline for drug discovery. *Journal of Chemical Information and Modeling*, 60(4):1955–1968, 2020. doi: 10.1021/acs.jcim.9b01053.

[29] 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.

[30] Greg Landrum and RDKit contributors. RDKit: Open-source cheminformatics software. `https://www.rdkit.org`, 2024. Version 2024.x; DOI: 10.5281/zenodo.591637.

[31] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, 2019. doi: 10.1145/3292500.3330701.

[33] Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data*, 7:137, 2020. doi: 10.1038/s41597-020-0460-4.

[34] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774, 2017. doi: 10.48550/arXiv.1705.07874.