Predicting Palladium-Catalyzed Buchwald–Hartwig C–N Cross-Coupling Reaction Yields from DFT Descriptors:

A Reproducible Machine-Learning Study with Interpolation and Additive-Extrapolation Validation

K-Dense Web

July 10, 2026

Abstract

The 2018 *Science* study of Ahneman, Estrada, Lin, Dreher, and Doyle established a high-throughput experimentation (HTE) benchmark for palladium-catalyzed Buchwald–Hartwig C–N cross-coupling, in which four reaction components—an aryl or heteroaryl halide, a Buchwald biaryl phosphine ligand, an organic base, and an isoxazole additive—were combined combinatorially and each reaction assigned a measured percent yield. Working from the publicly released descriptor matrix, we reconstructed the canonical core set of **3,955 reactions** (the full $22 \times 15 \times 3 \times 4 = 3{,}960$ grid minus five missing-yield entries), each represented by **120 density-functional-theory (DFT) descriptors** (64 ligand, 27 aryl halide, 19 additive, and 10 base features). We trained regularized-linear, random-forest, and histogram gradient-boosting regressors to predict yield and evaluated them two complementary ways: **(a)** a random 70/30 train/test split (`random_state` = 42) probing *interpolation*, and **(b)** an *out-of-sample additive hold-out* in which every reaction containing three chemically distinct additives—*3,5-dimethylisoxazole*, *5-phenylisoxazole*, and *benzo[c]isoxazole*—was withheld entirely from training. The histogram gradient-boosting (HGB) model was selected on cross-validated training RMSE for both splits and achieved **RMSE** = 6.07 **%,** $R^2 = 0.953$ on the random test set and **RMSE** = 17.60 **%,** $R^2 = 0.576$ on the additive extrapolation set, quantifying the substantial but not catastrophic penalty of predicting yields for never-before-seen additives. Aggregating permutation importances by component, the aryl halide is the dominant driver by total weight (44.90 % of summed importance), followed by the additive (21.50 %), ligand (16.90 %), and base (16.80 %); on a per-descriptor basis the base and aryl halide are essentially tied for the highest mean importance. The single strongest descriptor is the aryl-halide C3 NMR shift. We release per-reaction predicted-versus-observed yields for both test sets as a machine-readable file to support independent verification.

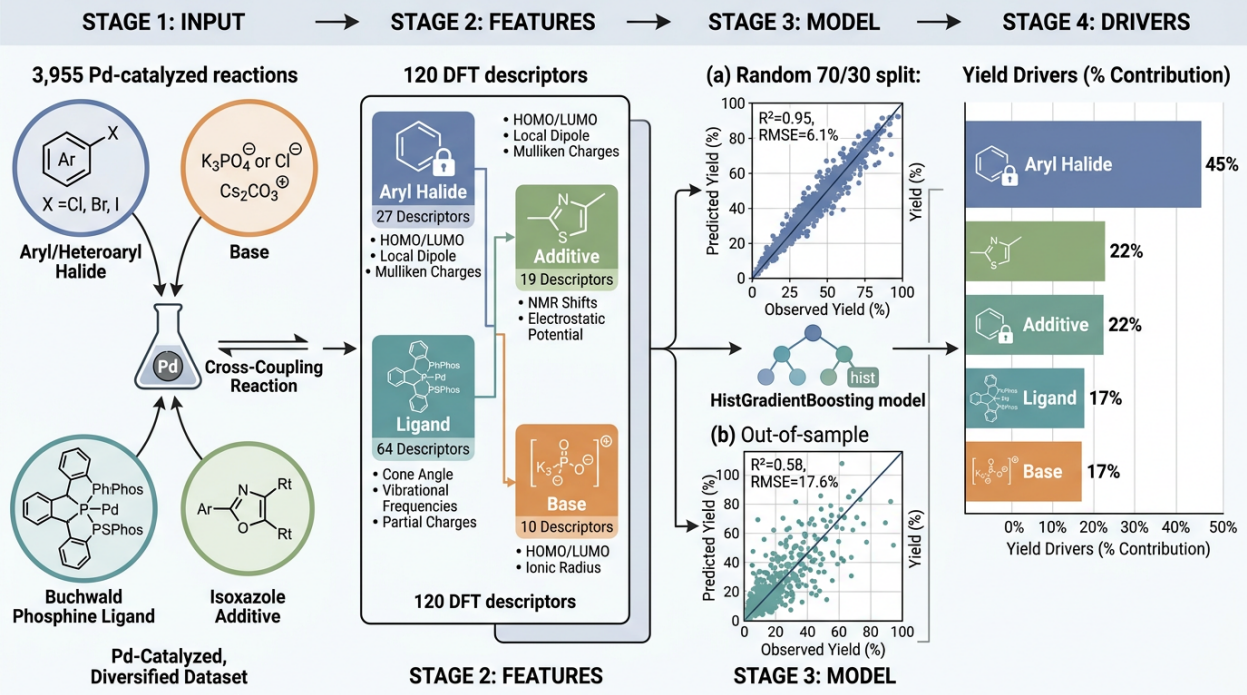

Figure 1: Graphical abstract. From 3,955 palladium-catalyzed Buchwald–Hartwig C–N coupling reactions we assemble a 120-dimensional DFT descriptor representation (64 ligand, 27 aryl halide, 19 additive, 10 base), fit a histogram gradient-boosting model, and validate it under two regimes: a random 70/30 split (interpolation; $R^2 = 0.95$, RMSE = 6.1) and an out-of-sample additive hold-out (extrapolation; $R^2 = 0.58$, RMSE = 17.6). Component-level importance analysis ranks the aryl halide as the leading yield driver.

Contents

1 Introduction **4**

2 Dataset **5**
2.1 Provenance and the 3,955-reaction core 5
2.2 Yield distribution 5

3 Methods **7**
3.1 Feature assembly 7
3.2 Validation splits 9
3.3 Models, tuning, and selection 9

4 Results **10**
4.1 Model comparison and selection 10
4.2 Split (a): random 70/30 interpolation 11
4.3 Split (b): out-of-sample additive extrapolation 11
4.4 Per-reaction predictions 12

5 Yield Driver Analysis **12**
5.1 Which components drive yield? 12
5.2 Top descriptors and their direction 14

5.3 Interpretation 16

6 Discussion **16**

7 Conclusion **17**

Data and Code Availability **17**

Reproducibility Checklist **17**

1 Introduction

Palladium-catalyzed Buchwald–Hartwig amination—the cross-coupling of an aryl (or heteroaryl) halide with an amine to forge a carbon–nitrogen bond—is one of the most heavily used transformations in pharmaceutical and agrochemical synthesis, underpinning the construction of anilines, azoles, and nitrogen-containing heterocycles at scale [1]. Despite its ubiquity, reaction yield remains notoriously difficult to predict *a priori*: subtle changes in the electronic and steric properties of the ligand, base, substrate, or trace additives can shift a coupling from near quantitative to complete failure. This sensitivity, combined with a combinatorially vast condition space, makes empirical optimization expensive and motivates data-driven approaches that learn structure–yield relationships directly from experiment.

The landmark 2018 study of Ahneman, Estrada, Lin, Dreher, and Doyle brought high-throughput experimentation (HTE) and machine learning to bear on precisely this problem [2]. Using nanomole-scale HTE platforms of the kind pioneered by Santanilla, Dreher, and co-workers [3], they assembled a densely populated dataset in which four reaction components were varied combinatorially and each combination assigned a measured percent yield by ultra-performance liquid chromatography–mass spectrometry. Crucially, every molecular component was characterized by a suite of density-functional-theory (DFT) descriptors—atomic NMR shifts, electrostatic charges, frontier-orbital energies, vibrational frequencies and intensities, dipole moments, and molecular size metrics—so that reactions could be represented as fixed-length numerical vectors amenable to supervised regression. A random forest [4] trained on these descriptors predicted held-out yields with high fidelity, and the study became a widely adopted benchmark for reaction-informatics methods [5, 6]. It also sparked methodological debate over feature attribution and the design of train/test splits [7, 8], underscoring that *how* a model is validated is as important as *whether* it fits.

The distinction between *interpolation* and *extrapolation* is central to that debate and to the practical value of any yield model. A model evaluated on a random hold-out is asked to predict reactions whose components it has already seen in other combinations—an interpolation task that random splits tend to make easy because closely related reactions appear on both sides of the partition. A model deployed to prioritize genuinely new chemistry, by contrast, must extrapolate to components it has never encountered. The additive dimension of the Buchwald–Hartwig dataset offers a clean stress test: the isoxazole additives were included specifically to probe catalyst inhibition, they span a wide range of mean yields, and holding an additive out entirely simulates the realistic scenario of forecasting how an unstudied heterocyclic contaminant will suppress (or spare) a coupling.

Objectives. This report documents a reproducible, end-to-end predictive-modeling pipeline built on the 2018 Buchwald–Hartwig core dataset. Our specific aims are to:

1. reconstruct and verify the canonical core set of 3,955 reactions and their 120 DFT descriptors;
2. build and tune yield-prediction regressors, selecting a final model on cross-validated training performance without test-set peeking;
3. report RMSE and R^2 under two validation regimes—a random 70/30 split (seed 42) and an out-of-sample additive hold-out—and release per-reaction predictions for both;
4. identify, at both the component and descriptor level, which reaction features most strongly drive predicted yield.

We emphasize transparency throughout: every split is leakage-verified, the final model is chosen by

nested cross-validation, and all predicted-versus-observed yields are provided as a machine-readable file.

2 Dataset

2.1 Provenance and the 3,955-reaction core

The modeling data derive from the descriptor tables and reaction matrix released alongside the 2018 *Science* study [2] and mirrored in the public `doylelab/rxnpredict` and `rxn4chemistry/rxn_yields` repositories. The experimental design is a complete four-factor grid:

$$\underbrace{22}_{\text{additives}} \times \underbrace{15}_{\text{aryl halides}} \times \underbrace{3}_{\text{bases}} \times \underbrace{4}_{\text{ligands}} = 3{,}960 \text{ reactions.}$$

Exactly five grid cells lack a measured yield; removing them yields the **3,955-reaction core** analyzed here. We recovered the identity of every component in each reaction by exact matching of its per-component descriptor sub-block against the named descriptor tables—no values were imputed or guessed—and attached canonical SMILES and CAS numbers from the accompanying name lists. A 23-point automated integrity audit confirmed the row count (3,955 core; 3,960 full grid; 5 missing), the absence of any missing descriptor or yield values in the core, the expected component counts (4 ligands, 3 bases, 15 aryl halides, 22 additives), and the uniqueness of all reaction combinations (all 23 checks passed).

Table 1: Composition of the reconstructed Buchwald–Hartwig core dataset. The four Buchwald ligands, three organic bases, and representative aryl/heteroaryl halides span the electronic and steric space probed by the original HTE campaign.

Component	#	Members
Ligand	4	XPhos, *t*-BuXPhos, *t*-BuBrettPhos, AdBrettPhos
Base	3	P2Et, BTMG, MTBD
Aryl/heteroaryl halide	15	4-substituted (CF_3, OMe, Et) chloro/bromo/iodo benzenes; 2- and 3-halopyridines (Cl/Br/I)
Additive (isoxazole)	22	e.g. 3,5-dimethylisoxazole, 5-phenylisoxazole, benzo[c]isoxazole, *N,N*-dibenzylisoxazol-3-amine, ethyl-isoxazole-4-carboxylate, and 17 further isoxazole/oxadiazole variants

2.2 Yield distribution

Percent yields span the full 0–100% range with mean 33.1% (median 28.8%, standard deviation 27.3%) and a pronounced right skew: a large cluster of failed or near-zero reactions coexists with a long tail of high-performing conditions (Figure 2, top-left). This heavy mass at zero yield is chemically meaningful—many additive/substrate combinations poison the catalyst—and it makes the prediction task genuinely bimodal rather than a smooth regression around a central tendency. At the component level, mean yield rises monotonically from XPhos to *t*-BuXPhos among ligands, is highest for the base MTBD, and varies more than fourfold across aryl halides (iodopyridines being most productive; chlorobenzene substrates least, Figure 2). The additive dimension is the widest lever of all: mean yield ranges from 46.7% for *N,N*-dibenzylisoxazol-3-amine down to 9.0% for ethyl-isoxazole-4-carboxylate (Figure 3), confirming that additive identity is a first-order determinant of reaction success and motivating the additive-based extrapolation test of Section 3.

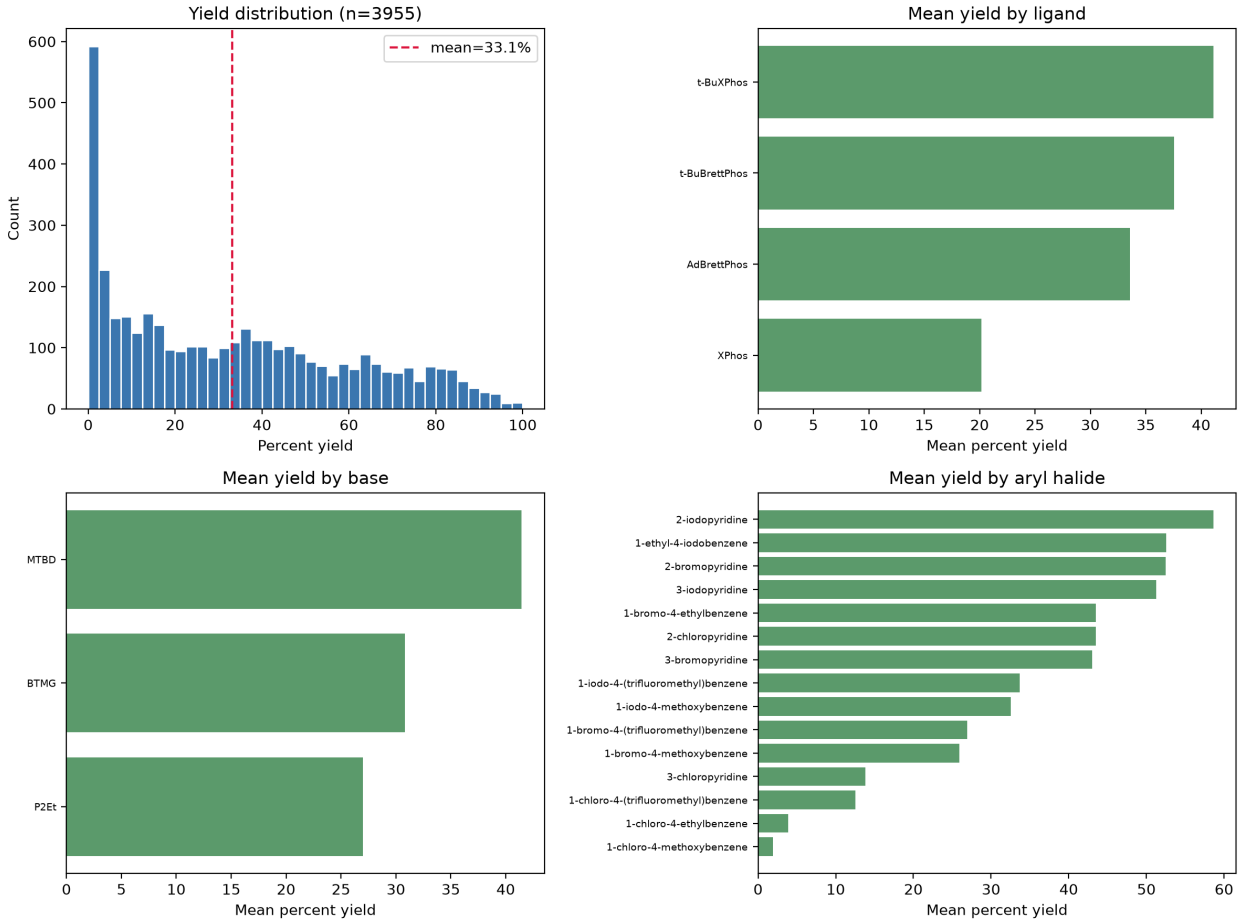

Figure 2: Exploratory analysis of the 3,955-reaction core. Top-left: distribution of percent yield (n = 3,955, mean 33.1%, dashed line), showing the characteristic spike at low yield and a long high-yield tail. Remaining panels: mean yield stratified by ligand, base, and aryl/heteroaryl halide. Iodinated and pyridyl substrates and the base MTBD are associated with the highest average yields.

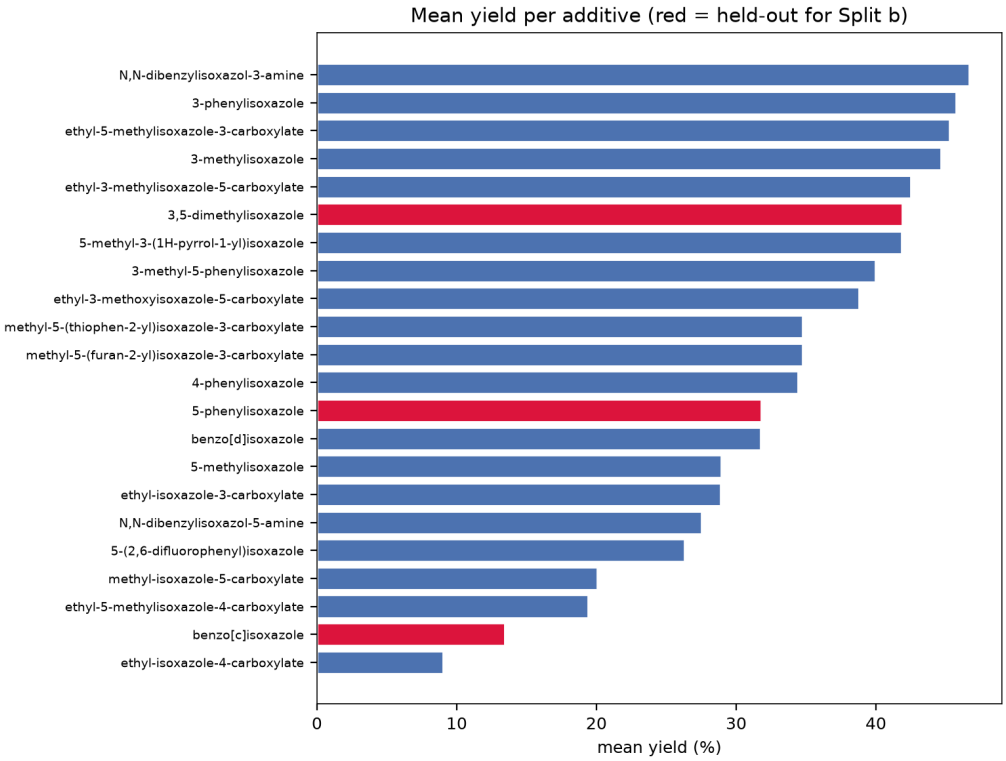

Figure 3: Mean yield by isoxazole additive. Additives are ordered by mean yield; the three highlighted (red) additives—*3,5-dimethylisoxazole* (high, 41.9%), *5-phenylisoxazole* (medium, 31.8%), and *benzo[c]isoxazole* (low, 13.4%)—were selected as the held-out test set for the out-of-sample extrapolation split (Section 3) because they jointly span the high, medium, and low ends of the additive yield range.

3 Methods

3.1 Feature assembly

Each reaction is represented by the **120 DFT descriptors** supplied with the original dataset, partitioned by component as summarized in Table 2. The descriptors are physically interpretable quantities computed for the isolated molecular components: frontier-orbital energies (E_{HOMO}, E_{LUMO}), atom-resolved NMR chemical shifts, atom-resolved electrostatic charges, vibrational frequencies and intensities, dipole moment, electronegativity, hardness, molecular volume, surface area, and ovality. Because these descriptors depend only on component identity, the representation encodes reactions through the physical properties of their parts rather than through categorical one-hot labels—the design choice that allows a model to generalize to component combinations, and in principle to components, not seen during training.

Table 2: Feature block sizes by reaction component. All 120 descriptors are DFT-derived, continuous quantities; none encodes component identity directly.

Component	# descriptors	Representative descriptors
Ligand	64	atom NMR shifts (C10/C11…), electrostatic charges, $E_{\mathrm{HOMO}}/E_{\mathrm{LUMO}}$, vibrational modes
Aryl halide	27	C1–C3 NMR shifts, H2 electrostatic charge, V1–V3 vibrational frequencies/intensities, E_{LUMO}, ovality
Additive	19	C3–C5 NMR shifts, O1 electrostatic charge, E_{HOMO}, dipole moment, molecular volume
Base	10	N1 electrostatic charge, E_{HOMO}, dipole, size metrics
Total	**120**	

Scaling and collinearity. Descriptors were standardized with scikit-learn's `StandardScaler` [9]; for the leakage-free modeling runs the scaler is fit *inside* a `Pipeline` on each training fold only. The core matrix contains no missing values and no zero-variance columns. A full 120×120 Pearson correlation analysis revealed substantial redundancy—403 descriptor pairs exceed $|r| > 0.90$, involving 95 of the 120 descriptors (Figure 4)—which is expected because descriptors within a component block are physically coupled. Tree-based ensembles are robust to such collinearity for prediction, so no descriptors were removed; the collinearity report is retained to inform the interpretation of feature importances (Section 5), where correlated descriptors can share or mask attributed importance.

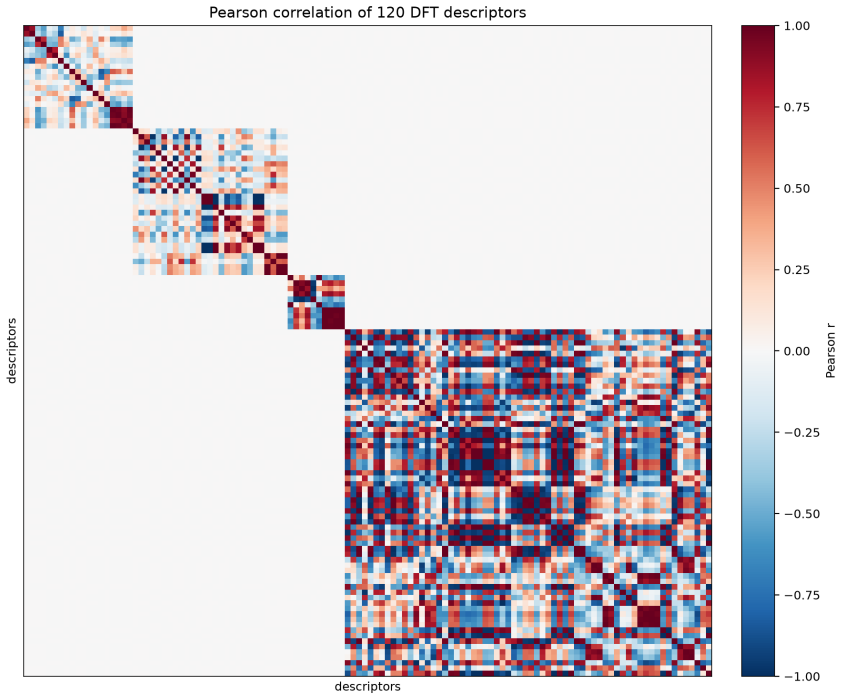

Figure 4: Pearson correlation of the 120 DFT descriptors, block-ordered by component (aryl halide, additive, base, ligand). Strong within-block correlation (deep red/blue) reflects the physical coupling of descriptors computed for the same molecule; 403 pairs exceed $|r| > 0.90$.

3.2 Validation splits

We evaluate every model under two complementary partitions of the 3,955-reaction core, illustrated in Figure 5.

(a) Random 70/30 split — interpolation. A single `train_test_split` with **`random_state`** = 42 assigns **2,768 reactions to training** and **1,187 to test** (train fraction 0.6999). Because the four factors are combinatorially crossed, essentially all component values appear on both sides of the partition; this split therefore measures the model's ability to *interpolate* within known chemical space. Train and test yield distributions are statistically indistinguishable (train mean 33.0%, test mean 33.3%).

(b) Out-of-sample additive hold-out — extrapolation. We remove *all* reactions containing three held-out additives from training and place them exclusively in the test set:

Held-out additive	Mean yield (%)	Regime
3,5-dimethylisoxazole	41.9	high
5-phenylisoxazole	31.8	medium
benzo[c]isoxazole	13.4	low

This yields **3,415 training reactions** (the remaining 19 additives) and **540 test reactions** (3 × 180). The three additives were chosen to be structurally distinct and to span the high, medium, and low ends of the additive yield range (Figure 3), so that success cannot be achieved by memorizing a single yield regime. Automated leakage checks confirm that none of the three appears in training, that training contains exactly 19 additives, and that the test set contains only the three held-out additives (all 9 split checks passed). This split measures *extrapolation* to never-before-seen additives—the realistic deployment scenario for triaging new chemistry.

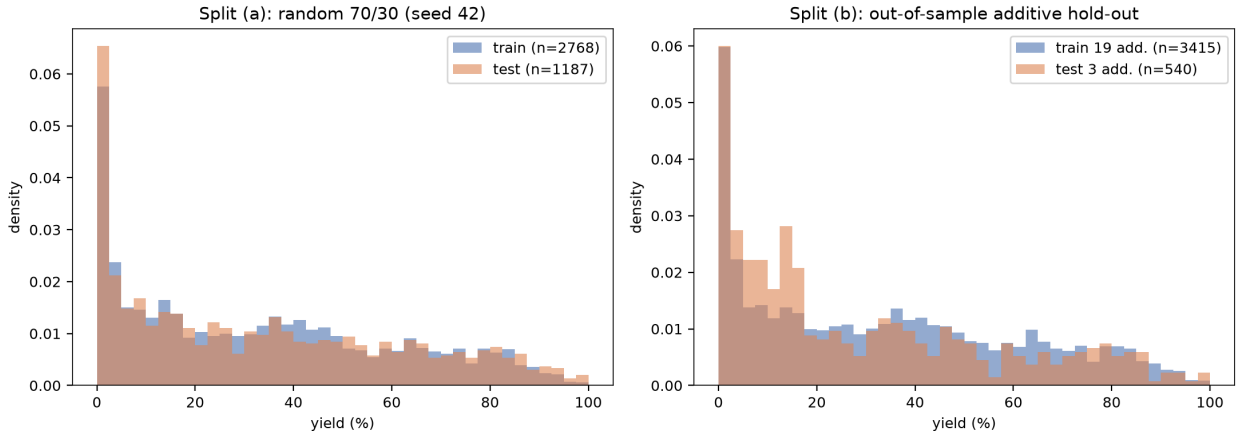

Figure 5: Train/test yield distributions for both splits. Left: the random 70/30 split (seed 42) yields near-identical train and test distributions (interpolation). Right: the additive hold-out shifts the test distribution toward lower yields and introduces the additive-specific structure absent from training (extrapolation).

3.3 Models, tuning, and selection

Three regressor families were evaluated, spanning the bias–variance spectrum:

- **Ridge regression** — a regularized linear baseline (`alpha` tuned over $\{10^{-3}, \ldots, 10^{2}\}$).
- **Random forest** [4] — bagged decision trees (tuned over `n_estimators`, `max_features`, `min_samples_leaf`, `max_depth`).
- **Histogram gradient boosting (HGB)** — scikit-learn's `HistGradientBoostingRegressor`, a fast, sklearn-native realization of gradient-boosted trees [10] conceptually equivalent to LightGBM [11] (tuned over `learning_rate`, `max_iter`, `max_leaf_nodes`, `min_samples_leaf`, `l2_regularization`).

All preprocessing and estimation steps are encapsulated in a scikit-learn `Pipeline`, and hyperparameters are tuned by **5-fold cross-validation on the training set only**. To avoid any test-set peeking, the *final* model reported for each split is chosen by the lowest mean 5-fold cross-validated RMSE among the three families—never by test performance. To guarantee a strictly honest featurization, the models use only the 120 DFT descriptors and never the additive identity as a categorical input; the additive's effect must be learned entirely through its DFT descriptors, which is what makes split (b) a true extrapolation test. Performance is reported as root mean squared error (RMSE, in percent-yield units) and the coefficient of determination R^2 on each held-out test set.

Software. Analyses used Python 3.12 with scikit-learn [9], NumPy [12], pandas, and Matplotlib [13]. Component-level importance is obtained by permutation importance on the test set (20 repeats, negative-RMSE scoring), and descriptor–yield directionality by Pearson and Spearman correlation with Benjamini–Hochberg false-discovery-rate correction [14].

4 Results

4.1 Model comparison and selection

Table 3 reports cross-validated and test performance for all three model families on both splits. On both splits the **histogram gradient-boosting model attained the lowest cross-validated training RMSE** and was therefore selected as the final model, with the random forest a close second and Ridge a distant third. The linear baseline is instructive: it reaches $R^2 \approx 0.71$ on the random split—confirming that a substantial fraction of yield variance is linearly encoded in the descriptors—but *collapses* on the additive hold-out ($R^2 = -10.3$, RMSE = 90.9), extrapolating wildly outside the trained yield range. The tree ensembles, which cannot extrapolate beyond the convex hull of their training targets, degrade far more gracefully. We note that on split (b) the random forest attains a marginally lower *test* RMSE (17.10 vs. 17.60) than HGB; because model selection is by cross-validated *training* RMSE—where HGB wins on both splits—HGB remains the reported model, and the two ensembles are effectively within noise of one another on the extrapolation set.

Table 3: Test-set performance by model and split. RMSE is in percent-yield units; the best (selected) model per split is bold. The final model was chosen by lowest 5-fold cross-validated training RMSE, not by test performance.

Split	Model	CV RMSE	Train RMSE	Test RMSE	Train R^2	Test R^2
(a) Random 70/30	Ridge	14.87	14.67	15.24	0.703	0.707
	Random forest	7.81	2.71	7.17	0.990	0.935
	HGB	**6.78**	**0.80**	**6.07**	**0.999**	**0.953**
(b) Additive hold-out	Ridge	14.92	14.75	90.85	0.707	−10.30
	Random forest	7.06	2.31	17.10	0.993	0.600
	HGB	**5.87**	**0.80**	**17.60**	**0.999**	**0.576**

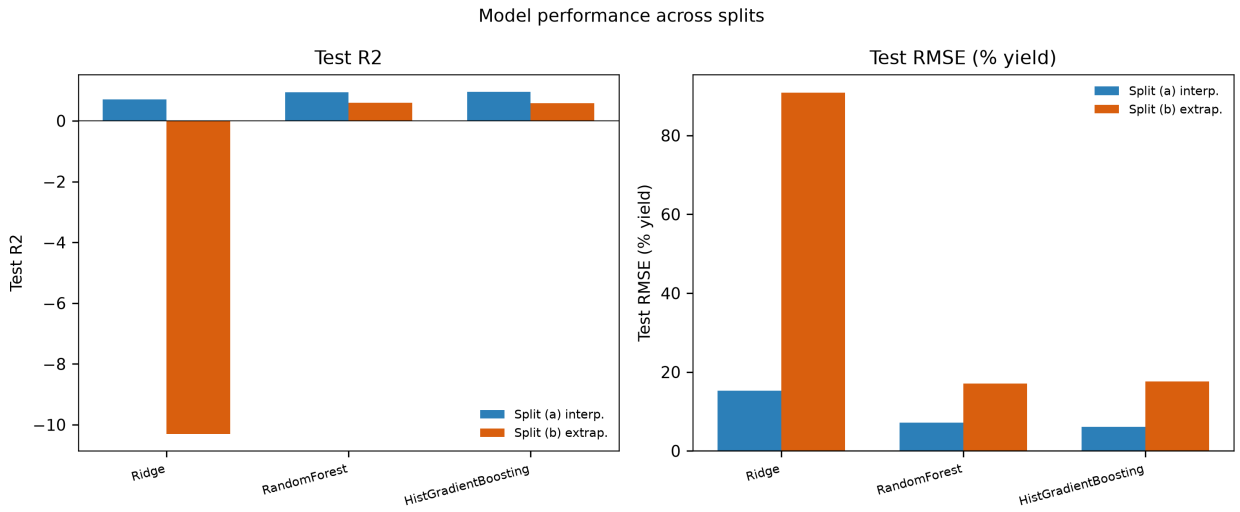

Figure 6: Model performance across splits. Test R^2 (left) and test RMSE (right) for Ridge, random forest, and HGB on the random 70/30 split (blue) and additive hold-out (orange). All models lose accuracy under extrapolation, but the linear Ridge baseline fails catastrophically ($R^2 < -10$), whereas the tree ensembles remain useful ($R^2 \approx 0.58$–0.60).

4.2 Split (a): random 70/30 interpolation

On the random test set of 1,187 reactions, the selected HGB model predicts yield with **RMSE** = 6.07% and $R^2 = 0.953$—i.e. it explains 95.3% of yield variance with a typical error of about six percentage points, comparable to or below the experimental reproducibility of HTE yield measurements. The predicted-versus-observed scatter (Figure 7, left) hugs the parity line tightly across the entire 0–100% range, with no systematic bias at either the low-yield spike or the high-yield tail. This confirms that, for reactions assembled from components the model has already encountered in other combinations, DFT descriptors plus gradient-boosted trees deliver highly accurate yield prediction, reproducing the strong interpolation performance that made this dataset a benchmark [2].

4.3 Split (b): out-of-sample additive extrapolation

When the model must instead predict the 540 reactions of three entirely unseen additives, performance drops to **RMSE** = 17.60% and $R^2 = 0.576$. The near-threefold increase in RMSE and the fall in R^2 from 0.95 to 0.58 quantify the real cost of extrapolation: the model retains

genuine, useful signal—explaining well over half of the yield variance for additives it has never seen—but is far less reliable than the interpolation figure would suggest. The predicted-versus-observed plot (Figure 7, right) reveals *additive-specific* error structure. Reactions of the low-yielding *benzo[c]isoxazole* (green) are systematically over-predicted—the model, having learned the general behavior of the training additives, does not fully anticipate this additive's stronger catalyst inhibition—while the higher-yielding *3,5-dimethylisoxazole* and *5-phenylisoxazole* points scatter more symmetrically about parity. This is the expected signature of extrapolation error and precisely the failure mode that a random split conceals.

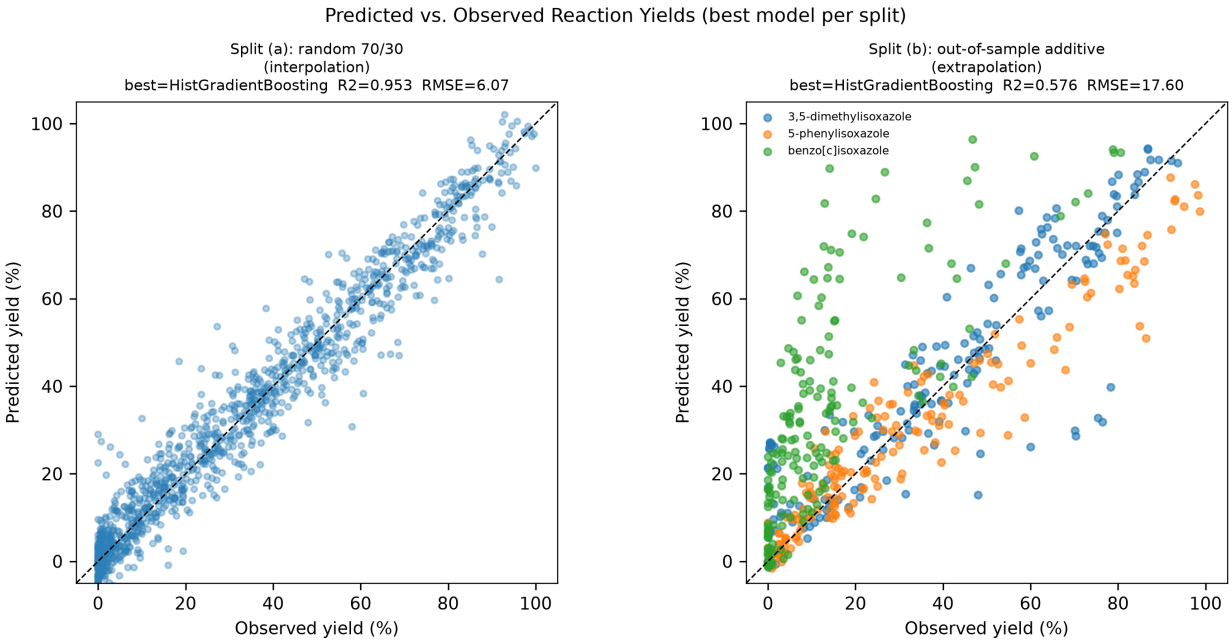

Figure 7: Predicted versus observed yields for the selected HGB model. Left — random 70/30 test set (interpolation): points cluster tightly on the parity line ($R^2 = 0.953$, RMSE = 6.07). Right — additive hold-out test set (extrapolation): predictions are noisier and show additive-specific bias ($R^2 = 0.576$, RMSE = 17.60); the low-yielding *benzo[c]isoxazole* (green) is systematically over-predicted. Dashed line is $y = x$.

4.4 Per-reaction predictions

Complete per-reaction predicted-versus-observed yields for both test sets are released as the machine-readable file `results/predictions_combined.csv` (1,727 rows: 1,187 from split (a) and 540 from split (b)). Each row carries the split label, the reaction identifier, the additive, the observed yield, the model-predicted yield, the residual, and the model name. Split-specific files (`predictions_split_a.csv`, `predictions_split_b.csv`) are also provided.

5 Yield Driver Analysis

5.1 Which components drive yield?

To attribute predicted yield to reaction components we computed permutation importance for every one of the 120 descriptors on the random-split test set (20 repeats), then aggregated the importances into the four component blocks two ways—by *sum* (total predictive weight carried by

a component) and by *mean* (average weight per descriptor). The two views are complementary and tell a nuanced story (Figure 8, Table 4).

Table 4: Component-level permutation importance, aggregated by sum and by mean. The aryl halide dominates total predictive weight, while on a per-descriptor basis the base and aryl halide are essentially tied for the highest mean importance. The ligand's 64 descriptors dilute its mean despite a moderate total.

Component	# desc.	Sum imp.	% of total	Rank (sum)	Mean imp.	Rank (mean)
Aryl halide	27	27.35	44.9%	1	1.013	2
Additive	19	13.12	21.5%	2	0.691	3
Ligand	64	10.28	16.9%	3	0.161	4
Base	10	10.22	16.8%	4	1.022	1

By **summed importance**, the ranking is

Aryl halide (44.9%) > Additive (21.5%) > Ligand (16.9%) ≈ Base (16.8%).

The aryl halide is unambiguously the leading driver of predicted yield, carrying nearly half of all attributed importance—consistent with the chemistry, in which the identity of the halide leaving group (I > Br > Cl) and the aromatic/heteroaromatic core govern oxidative addition and therefore turnover. The additive is second, reflecting its role as a deliberate catalyst inhibitor. The base and ligand carry similar total weight.

By **mean (per-descriptor) importance**, the ranking reorders to

Base (1.022) > Aryl halide (1.013) > Additive (0.691) > Ligand (0.161).

Here the base rises to the top: although it contributes only 10 descriptors, each one is on average as informative as an aryl-halide descriptor. The ligand, conversely, falls to last—its 64 descriptors, many mutually collinear (Figure 4), individually contribute little because the predictive signal is spread thinly and shared across the block. The apparent tension between the two rankings is therefore not a contradiction but a consequence of block size: *the aryl halide dominates total importance, while the base is the most efficient predictor per descriptor.* Both statements are reported because each answers a different question—"which component should I characterize most carefully?" (sum) versus "which single descriptor is most worth measuring?" (mean).

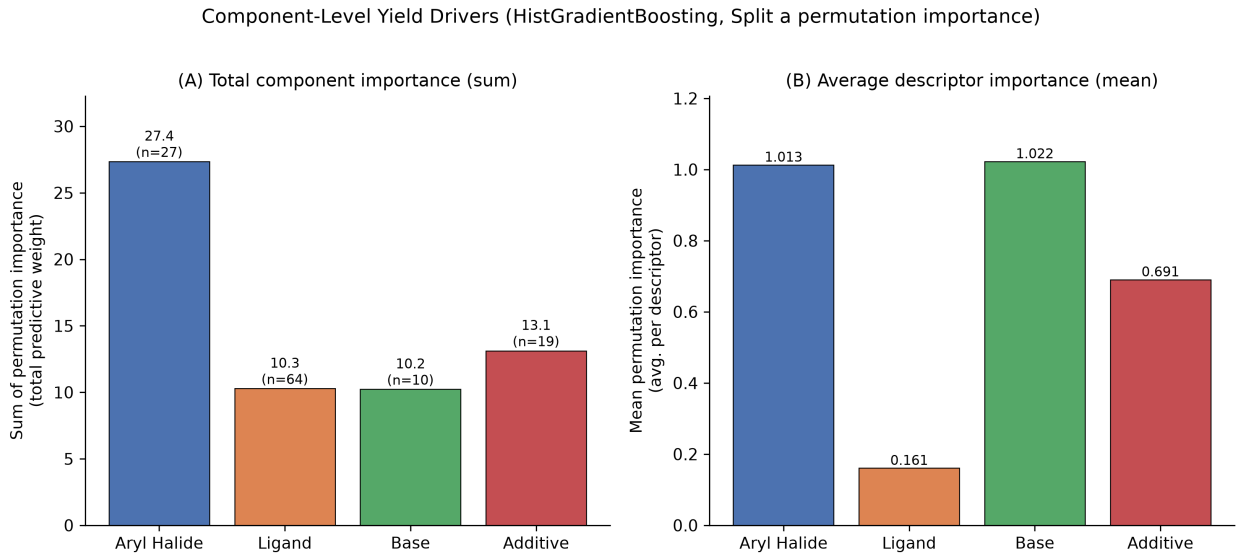

Figure 8: Component-level yield drivers from permutation importance of the HGB model on the random split. (A) Total importance (sum) per component: the aryl halide dominates (27.4), followed by the additive (13.1), ligand (10.3), and base (10.2); descriptor counts (n) are annotated. (B) Mean importance per descriptor: the base (1.022) and aryl halide (1.013) lead, the additive is intermediate (0.691), and the ligand trails (0.161) because its signal is diluted across 64 largely collinear descriptors.

5.2 Top descriptors and their direction

Descending to the descriptor level (Table 5, Figure 9), the single most important feature by a wide margin is the **aryl-halide C3 NMR shift** (permutation importance 12.1, more than 50% larger than the next feature), which correlates *positively* with yield (Pearson $r = +0.35$, Spearman $\rho = +0.43$; FDR $q \approx 10^{-113}$). The next tier comprises the ligand C11 NMR shift (+), the base N1 electrostatic charge (+), the additive C3 NMR shift (+), and the aryl-halide H2 electrostatic charge (−). Notably, the two aryl-halide vibrational-frequency descriptors V2 and V3 are strong *negative* drivers (Spearman $\rho \approx -0.46$ and -0.43), and the additive O1 electrostatic charge and C4 NMR shift are negative as well. Overall, 9 of the top-15 descriptors correlate positively with yield and 6 negatively, indicating that the model exploits a balanced mixture of yield-promoting and yield-suppressing structural signals rather than a single dominant axis.

Table 5: Top 15 descriptors by permutation importance, with Pearson and Spearman correlation to yield and Benjamini–Hochberg FDR-corrected significance. Nearly all associations are highly significant; direction indicates whether increasing the descriptor raises (+) or lowers (−) yield.

#	Descriptor	Component	Perm. imp.	Pearson r	Spearman ρ	Dir.
1	aryl_halide_C3_NMR_shift	Aryl halide	12.15	+0.350	+0.431	+
2	ligand_C11_NMR_shift	Ligand	8.17	+0.174	+0.258	+
3	base_N1_electrostatic_charge	Base	6.39	+0.205	+0.141	+
4	additive_C3_NMR_shift	Additive	5.91	+0.234	+0.239	+
5	aryl_halide_H2_electrostatic_charge	Aryl halide	5.05	−0.184	−0.237	−
6	base_E_HOMO	Base	3.61	−0.051	−0.057	−
7	aryl_halide_V2_frequency	Aryl halide	3.16	−0.267	−0.458	−
8	aryl_halide_V3_frequency	Aryl halide	1.94	−0.320	−0.433	−
9	ligand_C10_NMR_shift	Ligand	1.69	+0.042	+0.032	+
10	additive_O1_electrostatic_charge	Additive	1.38	−0.183	−0.212	−
11	additive_E_HOMO	Additive	1.13	+0.096	+0.105	+
12	additive_dipole_moment	Additive	0.82	+0.007	+0.008	+
13	additive_C4_NMR_shift	Additive	0.77	−0.158	−0.199	−
14	aryl_halide_E_LUMO	Aryl halide	0.71	+0.033	−0.156	+
15	additive_C5_electrostatic_charge	Additive	0.69	+0.055	+0.067	+

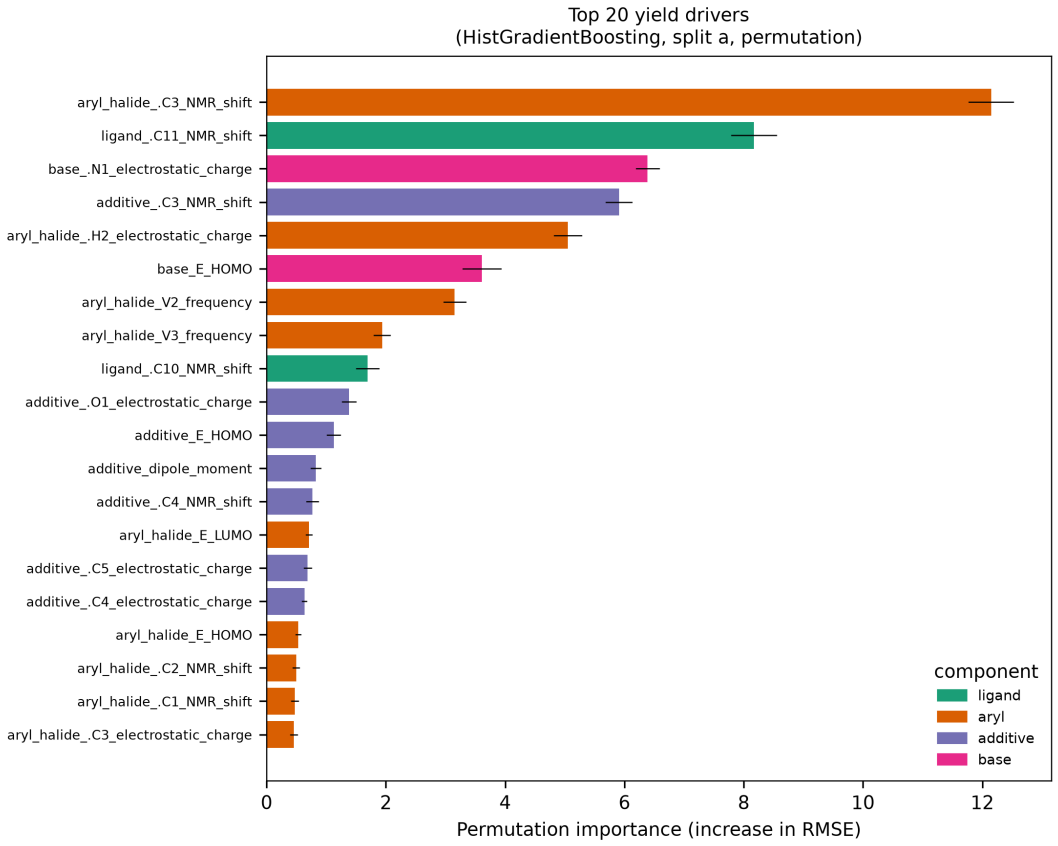

Figure 9: Top 20 individual yield drivers by permutation importance (increase in test RMSE when the descriptor is shuffled), colored by component. The aryl-halide C3 NMR shift dominates, and aryl-halide and additive descriptors populate much of the ranking, consistent with the component-level analysis.

5.3 Interpretation

The convergence of two independent lines of evidence—permutation importance (a model-based measure) and univariate correlation (a model-free measure)—lends confidence to the finding that the aryl halide is the dominant, and the additive a strong secondary, driver of predicted yield. The prominence of NMR-shift and electrostatic-charge descriptors over frontier-orbital energies is consistent with the catalytic mechanism: local electronic environment at reactive positions (captured by NMR shift and atomic charge) modulates oxidative addition and the competing additive-binding equilibria more directly than global HOMO/LUMO energies. The additive descriptors' generally negative correlations (O1 charge, C4 shift) echo the additives' designed role as catalyst poisons [2]. That said, permutation importance in the presence of the heavy collinearity documented in Figure 4 should be read as identifying informative *blocks and representative descriptors* rather than uniquely causal features; correlated descriptors can trade attributed importance among themselves.

6 Discussion

Interpolation vs. extrapolation. The headline contrast of this report is the gap between $R^2 = 0.953$ (random split) and $R^2 = 0.576$ (additive hold-out). Both numbers are "correct" for the questions they answer, and reporting only the first would materially overstate real-world reliability. The random split rewards a model for recognizing familiar components in new arrangements; the additive hold-out demands genuine generalization to unseen chemistry. The threefold RMSE inflation we observe (6.1 to 17.6 percentage points) is a concrete, actionable estimate of the uncertainty a practitioner should attach to yield predictions for a truly novel additive—and it directly echoes the community's concern that random splits can flatter reaction-yield models [5, 7, 8].

Graceful vs. catastrophic failure. The extrapolation results also argue for model choice. The linear Ridge baseline, adequate under interpolation ($R^2 = 0.71$), produces physically absurd predictions under extrapolation ($R^2 = -10.3$, RMSE = 90.9), because a linear model freely extrapolates outside the observed yield range. The tree ensembles, bounded by their training targets, degrade smoothly to a still-useful $R^2 \approx 0.58$. For deployment where out-of-domain queries are the norm, this bounded-error behavior is a decisive advantage.

Practical guidance. The driver analysis offers concrete advice for experimental design and future descriptor curation. Because the aryl halide carries the greatest total importance, accurate characterization of the substrate—especially descriptors of its C3 position and its low-frequency vibrational modes—yields the largest predictive return. Because the base is the most *efficient* predictor per descriptor, its small feature block is disproportionately valuable and should not be pruned. The additive, second by total importance and the widest experimental lever on mean yield, is exactly the dimension where extrapolation is hardest and where additional training diversity would most improve a deployed model.

Limitations. (i) The additive hold-out uses three additives; a leave-one-additive-out cross-validation across all 22 would give a distribution of extrapolation errors rather than a single estimate, at higher compute cost. (ii) Permutation importance under strong collinearity attributes importance to blocks and representatives rather than to uniquely causal descriptors; SHAP interaction analysis or a collinearity-pruned feature set would sharpen mechanistic claims. (iii) We deliberately restricted features to the 120 DFT descriptors and excluded categorical identity, which is the honest

choice for extrapolation but forgoes the extra interpolation accuracy that identity one-hots can provide. (iv) All conclusions pertain to this specific HTE chemical space (4 ligands, 3 bases, 15 halides, 22 isoxazole additives) and should be generalized to other C–N couplings only with caution.

7 Conclusion

We reconstructed the canonical 3,955-reaction Buchwald–Hartwig core dataset of the 2018 *Science* study, represented each reaction by 120 DFT descriptors, and built a histogram gradient-boosting model that predicts percent yield with $R^2 = 0.953$ (RMSE = 6.07) under a random 70/30 split (seed 42) and $R^2 = 0.576$ (RMSE = 17.60) under an out-of-sample hold-out of three additives (*3,5-dimethylisoxazole*, *5-phenylisoxazole*, and *benzo[c]isoxazole*). The contrast quantifies the interpolation–extrapolation gap that governs the practical trustworthiness of reaction-yield models. Component-level attribution identifies the aryl halide as the dominant driver by total importance (44.9%) with the additive second (21.5%), while the base is the most informative component per descriptor; the aryl-halide C3 NMR shift is the single strongest predictor. All per-reaction predictions for both test sets are released for independent verification.

Data and Code Availability

The reconstructed core dataset, standardized descriptor matrix, both validation splits, trained model artifacts, per-reaction predictions (`predictions_combined.csv`, `predictions_split_a.csv`, `predictions_split_b.csv`), feature-importance and correlation tables, and all figure-generation scripts accompany this report. Machine-readable performance and driver summaries are provided as `model_performance_summary.json` and `component_yield_drivers.json`. The source data derive from the public `doylelab/rxnpredict` and `rxn4chemistry/rxn_yields` repositories associated with Ref. [2].

Reproducibility Checklist

- **Dataset:** 3,955 core reactions verified by a 23-point integrity audit (all passed).
- **Random seed:** `random_state` = 42 for the 70/30 split.
- **Held-out additives (split b):** 3,5-dimethylisoxazole; 5-phenylisoxazole; benzo[c]isoxazole.
- **Leakage control:** scaler fit on train folds only; features restricted to 120 DFT descriptors; 9-point split audit passed.
- **Model selection:** lowest mean 5-fold CV training RMSE (no test-set peeking).
- **Metrics:** RMSE (% yield) and R^2 on each held-out test set.

References

[1] Paula Ruiz-Castillo and Stephen L. Buchwald. Applications of palladium-catalyzed C–N cross-coupling reactions. *Chemical Reviews*, 116(22):12564–12649, 2016. doi: 10.1021/acs.chemrev.6b00512.

[2] Derek T. Ahneman, Jesús G. Estrada, Shishi Lin, Spencer D. Dreher, and Abigail G. Doyle. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360 (6385):186–190, 2018. doi: 10.1126/science.aar5169.

[3] Alexander Buitrago Santanilla, Erik L. Regalado, Tony Pereira, Michael Shevlin, Kevin Bateman, Louis-Charles Campeau, Jonathan Schneeweis, Simon Berritt, Zhi-Cai Shi, Philippe Nantermet, Yong Liu, Roy Helmy, Christopher J. Welch, Petr Vachal, Ian W. Davies, Tim Cernak, and Spencer D. Dreher. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347(6217):49–53, 2015. doi: 10.1126/science.1259203.

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

[5] Frederik Sandfort, Felix Strieth-Kalthoff, Marius Kühnemund, Christian Beecks, and Frank Glorius. A structure-based platform for predicting chemical reactivity. *Chem*, 6(6):1379–1390, 2020. doi: 10.1016/j.chempr.2020.02.017.

[6] Philippe Schwaller, Alain C. Vaucher, Teodoro Laino, and Jean-Louis Reymond. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology*, 2 (1):015016, 2021. doi: 10.1088/2632-2153/abc81d.

[7] Kevin V. Chuang and Michael J. Keiser. Comment on "predicting reaction performance in C–N cross-coupling using machine learning". *Science*, 362(6416):eaat8603, 2018. doi: 10.1126/science.aat8603.

[8] Jesús G. Estrada, Derek T. Ahneman, Robert P. Sheridan, Spencer D. Dreher, and Abigail G. Doyle. Response to comment on "predicting reaction performance in C–N cross-coupling using machine learning". *Science*, 362(6416):eaat8763, 2018. doi: 10.1126/science.aat8763.

[9] 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.

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

[11] 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.

[12] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[13] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[14] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.