Predicting Catalyst Selectivity in Chiral Phosphoric Acid–Catalyzed Thiol Additions to *N*-Acylimines: A Machine-Learning Model with Interpolative and Extrapolative Validation

July 10, 2026

Abstract

The 2019 *Science* study of Zahrt, Denmark, and co-workers established a now-canonical benchmark for data-driven asymmetric catalysis: the exhaustive combination of five *N*-acylimines, five thiols, and 43 BINOL-derived chiral phosphoric acid (CPA) catalysts, giving 1,075 enantioselective additions that form *N,S*-acetals, each with a measured enantioselectivity. We reproduce this dataset and build regression models that predict the differential activation free energy $\Delta\Delta G^{\ddagger}$ ($\mathrm{kcal\,mol^{-1}}$), the physically meaningful selectivity target obtained from the measured enantiomeric ratio through the transition-state relation $\Delta\Delta G^{\ddagger} = -RT\ln(\mathrm{er})$. Each reaction component is featurized with Morgan (extended-connectivity) fingerprints and a compact set of RDKit two-dimensional physicochemical descriptors, yielding a 3,129-dimensional representation. Four model families—Ridge regression, support vector regression (SVR), random forests (RF), and histogram-based gradient boosting (HGB)—were tuned by five-fold cross-validation and evaluated under two complementary regimes. Under a random 80/20 split (interpolation), the random forest achieved a mean absolute error (MAE) of $0.132\,\mathrm{kcal\,mol^{-1}}$ and $R^2 = 0.915$ on 215 held-out reactions. Under a deliberately harder catalyst-based holdout (extrapolation)—in which all 25 reactions of each of eight randomly chosen catalysts (IDs {4, 12, 24, 25, 34, 36, 37, 39}; seed 42) are removed from training—histogram-based gradient boosting generalized best, reaching an MAE of $0.189\,\mathrm{kcal\,mol^{-1}}$ and $R^2 = 0.895$ on the 200 unseen-catalyst reactions. Both best models attain the widely used "chemical accuracy" threshold of $< 0.2\,\mathrm{kcal\,mol^{-1}}$, and both retain strong rank ordering even for catalyst scaffolds never seen in training, while the systematic $\approx$43% increase in error from interpolation to extrapolation quantifies the intrinsic difficulty of generalizing across catalyst chemical space. We provide the complete per-reaction predicted-versus-observed $\Delta\Delta G^{\ddagger}$ tables for both test sets.

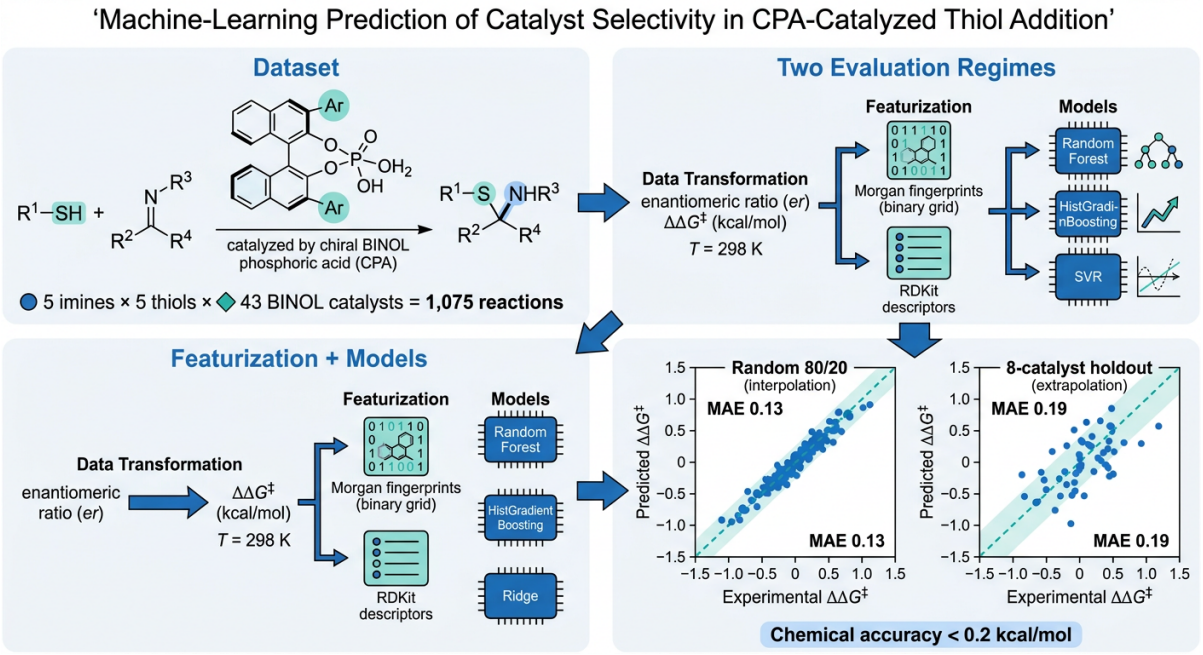


Figure 1: Graphical abstract. From the 1,075-reaction CPA thiol-addition dataset, measured enantioselectivities are converted to $\Delta\Delta G^{\ddagger}$ (kcal mol^{-1}), each reaction is featurized with Morgan fingerprints and RDKit descriptors, and four regression models are trained and evaluated under two regimes: a random 80/20 split (interpolation) and an eight-catalyst holdout (extrapolation). Both best models achieve < 0.2 kcal mol^{-1}.

1 Introduction

The rational discovery of highly enantioselective catalysts remains one of the central challenges of synthetic organic chemistry. For decades, catalyst optimization proceeded largely through structural intuition and iterative experimentation, a process that is effective but slow and difficult to transfer between reaction classes. The recognition that catalyst performance can be treated as a quantitative function of measurable molecular properties—a quantitative structure–selectivity relationship (QSSR)—has reframed catalyst design as a prediction problem amenable to statistical modeling [1–3]. Pioneering multivariate free-energy-relationship studies from the Sigman group demonstrated that steric and electronic descriptors, correlated against selectivity through linear models, could not only rationalize existing data but also guide the design of improved catalysts [1, 2]. In parallel, the broader emergence of machine learning across the molecular sciences [4, 5] created both the algorithms and the expectation that predictive models could accelerate reaction development.

A landmark demonstration of this paradigm for asymmetric catalysis is the 2019 *Science* report of Zahrt, Henle, Rose, Wang, Darrow, and Denmark, "Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning" [6]. The authors chose the chiral phosphoric acid (CPA)–catalyzed addition of thiols to *N*-acylimines—a Brønsted-acid-catalyzed reaction that forges synthetically valuable chiral *N,S*-acetals (thioaminals) [7, 8]—as a proof-of-principle system. CPA catalysis itself was inaugurated by the independent 2004 discoveries of Akiyama and of Terada, who showed that BINOL-derived phosphoric acids act as powerful, tunable chiral Brønsted acids for enantioselective additions to imines [9, 10]; the BINOL-phosphate scaffold has since become one of the most versatile platforms in organocatalysis [11]. The distinctive contribution of the Denmark study was methodological: rather than relying on chemically intuitive descriptors chosen reaction-by-reaction, the authors constructed a scaffold-agnostic, three-dimensional descriptor framework—the average steric occupancy (ASO) representation—designed to capture the conformational steric profile of each catalyst, and combined it with support vector machines and deep feed-forward neural networks to predict enantioselectivity across a very large virtual catalyst library. Crucially, the models were used prospectively to identify catalysts more selective than any in the training set, providing an early and influential demonstration that machine learning can extrapolate usefully in catalyst space [3, 6].

At the heart of that study is a compact but complete experimental dataset that has become a widely used benchmark in its own right: the exhaustive $5 \times 5 \times 43$ matrix of five imines, five thiols, and 43 BINOL-derived CPA catalysts, giving 1,075 reactions, each with a measured enantioselectivity. The completeness of this design—every catalyst is measured against every substrate pair—makes it especially valuable for methodological studies of predictive modeling, because it cleanly separates the question of *substrate* generalization from the question of *catalyst* generalization. The natural target for such modeling is not the enantiomeric excess itself but the differential activation free energy $\Delta\Delta G^{\ddagger}$ between the diastereomeric, enantiodetermining transition states. This quantity, related to the enantiomeric ratio through transition-state theory [12], is the physically meaningful, approximately additive descriptor of selectivity: a $\Delta\Delta G^{\ddagger}$ of only $\sim 2\,\mathrm{kcal\,mol^{-1}}$ already corresponds to roughly 95% enantiomeric excess, so accurate models must resolve differences well below $1\,\mathrm{kcal\,mol^{-1}}$ to be synthetically useful.

The predictive-modeling literature has repeatedly emphasized that the difficulty of a chemical machine-learning task depends critically on how the data are split. Models evaluated on random splits, where test examples closely resemble training examples, tend to report optimistic performance; the same models can degrade substantially when required to generalize to genuinely new reagents,

substrates, or scaffolds [13–15]. This distinction between *interpolation* and *extrapolation* is not academic—it determines whether a model can actually propose new catalysts rather than merely interpolate among known ones. In this report we therefore evaluate our models under both regimes and quantify the gap between them.

The present work has three objectives. First, we reconstruct the 1,075-reaction CPA thiol-addition dataset and convert the measured selectivities to $\Delta\Delta G^{\ddagger}$ (Section 2.1). Second, we build and tune a panel of four regression models on an interpretable, reproducible feature set combining Morgan fingerprints and RDKit physicochemical descriptors (Sections 2.2–2.3). Third, and most importantly, we benchmark these models under a standard random 80/20 split and under a stringent catalyst-based holdout in which eight entire catalysts are withheld from training, so that the test catalysts are truly unseen (Section 2.4). We report MAE and R^2 for every model under both regimes, identify the best model in each, analyze where and why extrapolation is harder, and supply the complete per-reaction predicted-versus-observed $\Delta\Delta G^{\ddagger}$ tables for both held-out sets. Our aim is not to surpass the original neural-network models but to provide a transparent, fully reproducible baseline and an explicit, quantitative characterization of the interpolation–extrapolation gap on this benchmark.

2 Methods

2.1 Dataset and target variable

The dataset comprises the complete set of $5 \times 5 \times 43 = 1{,}075$ CPA-catalyzed thiol additions to *N*-acylimines reported in the 2019 *Science* study [6]: five *N*-acyl (benzamide-derived) aryl imines, five thiols spanning aromatic and aliphatic classes, and 43 BINOL-derived chiral phosphoric acid catalysts bearing diverse 3,3′-substituents and, in several cases, partially or fully hydrogenated (H_8) binaphthyl backbones. Each of the three reaction components and the reaction product is represented as a canonical SMILES string. Every SMILES was parsed and sanitized with RDKit; all 5 imines, 5 thiols, 43 catalysts, and 25 products were confirmed to be valid and unique, and no reaction rows contained missing components. The full factorial design contains no duplicate component combinations, and all 1,075 selectivity values are present.

The modeling target is the differential activation free energy between the two enantiodetermining transition states,

$$\Delta\Delta G^{\ddagger} \;=\; -RT\,\ln(\mathrm{er}), \tag{1}$$

where er is the enantiomeric ratio of the two product enantiomers, R is the gas constant, and $T = 298.15$ K, giving $RT = 0.5925\,\mathrm{kcal\,mol^{-1}}$. Equation (1) follows directly from transition-state theory [12]: because the two enantiomeric products arise from transition states differing only in free energy, the logarithm of their ratio is proportional to $\Delta\Delta G^{\ddagger}$. The sign convention references a fixed product configuration, so positive $\Delta\Delta G^{\ddagger}$ indicates preferential formation of the reference enantiomer and negative $\Delta\Delta G^{\ddagger}$ indicates the opposite sense of induction. Across the dataset, $\Delta\Delta G^{\ddagger}$ ranges from -0.419 to $+3.135\,\mathrm{kcal\,mol^{-1}}$ (mean 0.988, median 1.051, standard deviation $0.700\,\mathrm{kcal\,mol^{-1}}$), corresponding to enantiomeric excesses from 0 to $\approx$99% (mean $\approx$59%). Modeling $\Delta\Delta G^{\ddagger}$ rather than the enantiomeric excess is preferable because $\Delta\Delta G^{\ddagger}$ is the additive, unbounded quantity that varies smoothly with catalyst and substrate structure, whereas the enantiomeric excess is a bounded, nonlinear transform that compresses precisely the high-selectivity region of greatest practical interest.

2.2 Feature engineering

Each reaction was represented by concatenating independent descriptor blocks for its imine, thiol, and catalyst, so that the model sees the full identity of all three components. For every molecule two complementary feature types were computed with RDKit. The first is a Morgan (extended-connectivity) fingerprint of radius 2 folded to 1,024 bits; these circular fingerprints, introduced by Rogers and Hahn [16], encode the presence of local atomic environments and are a standard, interpretable substructure representation for structure–property modeling. The second is a compact vector of 19 two-dimensional physicochemical descriptors capturing size, polarity, lipophilicity, hydrogen-bonding capacity, aromaticity, flexibility, and topological complexity: molecular weight, MolLogP, molar refractivity, topological polar surface area, counts of hydrogen-bond donors and acceptors, rotatable bonds, aromatic and aliphatic rings, total ring count, fraction of sp^3 carbon, heteroatom count, heavy-atom count, valence-electron count, N–H/O–H and N/O counts, Labute accessible surface area, the Balaban J index, and the Bertz complexity index. Such physicochemical descriptors, computed with the open-source RDKit toolkit, are widely used in cheminformatics pipelines and QSAR modeling [17]. The 19 descriptors of each component were standardized (zero mean, unit variance) using statistics fit only on the set of unique molecules of that component to avoid inflating the influence of repeated components.

Concatenating a 1,024-bit fingerprint and 19 descriptors gives 1,043 features per component, and stacking the imine, thiol, and catalyst blocks yields a final design matrix of $1{,}075 \times 3{,}129$. The matrix contained no missing, infinite, or undefined values. This deliberately generic, computationally inexpensive representation—derived from two-dimensional structure alone, without the three-dimensional conformer ensembles or quantum-chemical calculations used in the original ASO framework [6]—was chosen to provide a transparent and easily reproduced baseline.

2.3 Model families and hyperparameter tuning

Four regression algorithms spanning linear, kernel, and tree-ensemble paradigms were evaluated, all implemented in scikit-learn [18]. *Ridge regression* is an ℓ_2-regularized linear model that serves as an interpretable, high-bias baseline. *Support vector regression* (SVR) with a radial-basis-function kernel captures smooth nonlinear structure through the kernel trick [19]. *Random forests* (RF) average many decorrelated regression trees grown on bootstrap samples and random feature subsets, a bagging ensemble that is robust, resistant to overfitting, and a long-standing standard for descriptor- and fingerprint-based chemical modeling [20, 21]. *Histogram-based gradient boosting* (HGB) builds an additive ensemble of shallow trees in a stagewise fashion, following the gradient-boosting framework [22]; its histogram-binned implementation mirrors the algorithmic advances of modern boosted-tree libraries [23, 24] and is well suited to large tabular feature sets. Ridge and SVR were preceded by feature standardization inside a pipeline so that scaling parameters were fit only on training folds, preventing information leakage; the two tree ensembles are scale-invariant and used the raw features.

Each model's hyperparameters were tuned by grid search with five-fold cross-validation on the training partition, optimizing the negative mean absolute error. For Ridge, the regularization strength $\alpha \in \{0.1, 1, 10, 100, 300\}$ was searched. For SVR, the penalty $C \in \{1, 10, 100\}$ and tube width $\epsilon \in \{0.05, 0.1\}$ were searched with the scale-based γ heuristic. For the random forest, 400 trees were grown while searching maximum depth (unlimited or 20), the fraction of features considered per split ($\sqrt{p}$ or $0.3p$), and the minimum leaf size (1 or 2). For gradient boosting, the learning rate (0.05 or 0.1), maximum depth (unlimited or 6), and ℓ_2 regularization (0 or 1) were searched over

400 boosting iterations. All randomized operations—the cross-validation folds, the tree ensembles, and the data splits—were seeded with the fixed value **42** to guarantee exact reproducibility.

2.4 Validation strategies

Two evaluation regimes probe fundamentally different notions of generalization (Figure 4). The first is a **random 80/20 split**: the 1,075 reactions were shuffled (seed 42) and partitioned into 860 training and 215 test reactions. Because catalysts, imines, and thiols recur throughout the dataset, essentially every test reaction shares its catalyst—and both of its substrates—with many training reactions. This regime therefore measures *interpolation*: filling in missing entries of a partially observed reaction matrix.

The second is a **catalyst-based holdout** designed to measure *extrapolation* to unseen catalysts. Eight of the 43 catalysts were selected uniformly at random (NumPy `RandomState`, seed 42), yielding catalyst IDs **{4, 12, 24, 25, 34, 36, 37, 39}**. Every one of the $8 \times 25 = 200$ reactions involving these catalysts was withheld as the test set, and the model was trained on the remaining 875 reactions spanning the other 35 catalysts. We verified that no held-out catalyst appears anywhere in the training partition, so the test catalysts are genuinely unseen: the model must predict selectivity for catalyst scaffolds it has never encountered, using only their molecular features. This is a substantially harder and more realistic test of whether a model could guide the discovery of new catalysts. The eight held-out catalysts span a chemically diverse cross-section of the library, including electron-rich (e.g., *para*-methoxyphenyl, catalyst 12) and strongly electron-poor bis(trifluoromethyl)phenyl variants (catalysts 24, 25, 39), extended-aromatic and biaryl 3,3′-groups (catalysts 4, 34), and small-substituent bromo catalysts on both fully aromatic and H_8 backbones (catalysts 36, 37).

For each regime we tuned all four models on the training partition, evaluated the best configuration of each on the untouched test set, and report the mean absolute error (MAE, in $\mathrm{kcal\,mol^{-1}}$) and the coefficient of determination (R^2). The single best model per regime (lowest test MAE) was used to export per-reaction predictions and to draw parity plots. All computations used NumPy 2.5, pandas 3.0, scikit-learn 1.9, RDKit 2026.03, and Matplotlib 3.11.

3 Results

3.1 Comparative model performance

Table 1 collects the test-set MAE and R^2 for all four models under both regimes, and Figure 3 visualizes the same results. Three consistent patterns emerge. First, the two tree ensembles—random forests and histogram-based gradient boosting—clearly outperform the linear and kernel baselines in both regimes, confirming that the selectivity surface contains substantial nonlinear and interaction structure that a regularized linear model cannot capture and that RBF-SVR captures only partially. Second, every model degrades when moving from the random split to the catalyst holdout, but by markedly different amounts: the robust tree ensembles lose relatively little accuracy, whereas Ridge and especially SVR degrade sharply. Third, both winning models comfortably meet the "chemical accuracy" benchmark of $< 0.2\,\mathrm{kcal\,mol^{-1}}$ on their respective test sets.

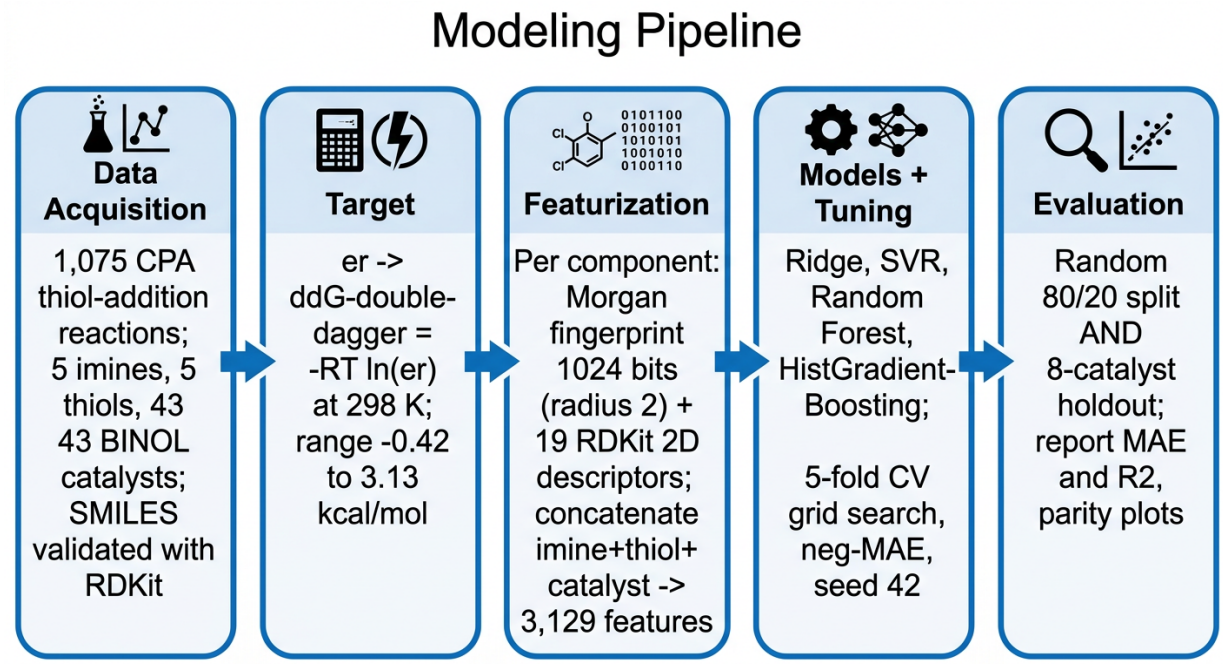


Figure 2: Modeling pipeline. Data acquisition and SMILES validation, conversion of the measured enantiomeric ratio to the target $\Delta\Delta G^{\ddagger}$, per-component featurization into a 3,129-dimensional vector, cross-validated tuning of four model families, and evaluation under two split strategies.

Table 1: Test-set performance of the four model families under the random 80/20 split (interpolation) and the eight-catalyst holdout (extrapolation). MAE is in kcal mol^{-1}; best value in each column is in **bold**. All models tuned by five-fold cross-validation with seed 42.

	Random 80/20 (interpolation)		**8-catalyst holdout** (extrapolation)	
Model	MAE ↓	R^2 ↑	MAE ↓	R^2 ↑
Ridge	0.183	0.854	0.273	0.744
Support vector regression	0.146	0.890	0.324	0.639
Random forest	**0.132**	**0.915**	0.195	0.894
Histogram gradient boosting	0.139	0.907	**0.189**	**0.895**
n (train / test)	860 / 215		875 / 200	

3.2 Interpolation: the random 80/20 split

On the random split, the random forest was the strongest model, achieving an MAE of **0.132** kcal mol^{-1} and R^2 = **0.915** across the 215 test reactions; histogram gradient boosting was a close second (0.139 kcal mol^{-1}, $R^2 = 0.907$). The corresponding parity plot (Figure 5a) shows predictions tracking the ideal $y = x$ line tightly across the full $\Delta\Delta G^{\ddagger}$ range, from near-racemic reactions at the origin to the most selective reactions above 3 kcal mol^{-1}. The error distribution is sharply peaked: the median absolute error is only 0.084 kcal mol^{-1}, 79.5% of test reactions are predicted within 0.2 kcal mol^{-1}, and 97.2% within 0.5 kcal mol^{-1}. The largest deviations occur in the sparsely populated high-selectivity tail ($\Delta\Delta G^{\ddagger} > 2.5$ kcal mol^{-1}), where a handful of extreme reactions are slightly underpredicted—a mild regression-to-the-mean effect expected of tree ensembles at the edge of the

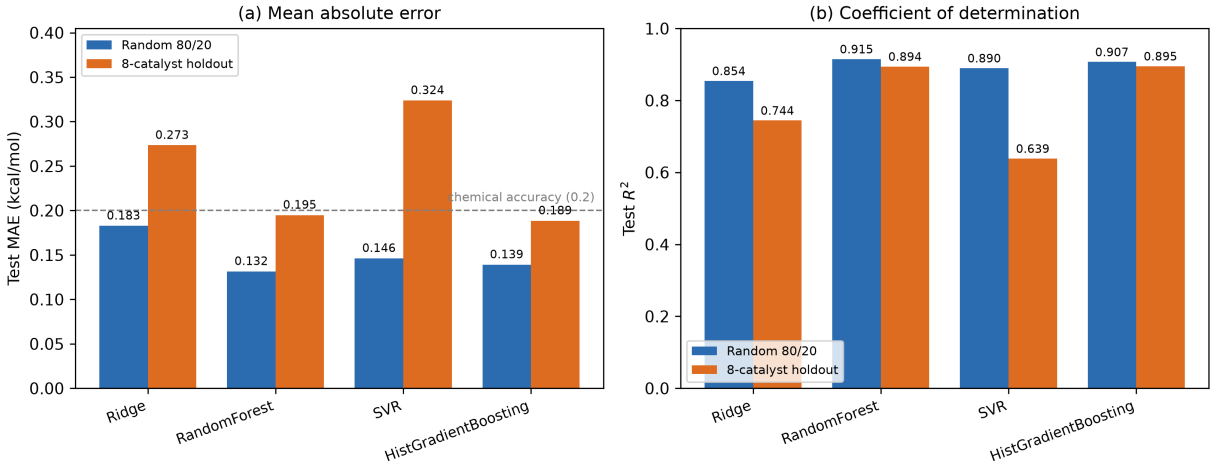


Figure 3: Comparative model performance. (a) Test mean absolute error and (b) test R^2 for the four model families under the random 80/20 split (blue) and the eight-catalyst holdout (orange). The dashed line in (a) marks the 0.2 kcal mol^{-1} chemical-accuracy threshold. Tree ensembles are both the most accurate and the most robust to the interpolation→extrapolation shift.

target range. For context, 0.132 kcal mol^{-1} corresponds to less than a ∼10% error in enantiomeric ratio for a typical reaction, comfortably within experimental reproducibility.

3.3 Extrapolation: the eight-catalyst holdout

When the eight test catalysts were removed entirely from training, histogram gradient boosting generalized best, achieving an MAE of **0.189** kcal mol^{-1} and R^2 = **0.895** on the 200 unseen-catalyst reactions, narrowly ahead of the random forest (0.195 kcal mol^{-1}, $R^2 = 0.894$). Both tree ensembles thus remained below the 0.2 kcal mol^{-1} chemical-accuracy threshold even for catalysts never encountered during training—a strong result given that selectivity in CPA catalysis is governed by subtle three-dimensional interactions in the chiral pocket. The parity plot for this regime (Figure 5b) again shows good correlation along $y = x$, with visibly more scatter than the random split and a mild compression of the most selective predictions. Quantitatively, the median absolute error rises to 0.142 kcal mol^{-1}, the 90th percentile to 0.43 kcal mol^{-1}, and the fraction of reactions predicted within 0.2 kcal mol^{-1} falls to 62.5% (though 96.0% remain within 0.5 kcal mol^{-1}). In other words, the model reliably recovers the correct sense and approximate magnitude of induction for new catalysts, with occasional larger misses.

The aggregate holdout error masks substantial catalyst-to-catalyst variation, which Figure 6 resolves. Four of the eight held-out catalysts are predicted very well—catalyst 24 (bis(3,5-(CF$_3$)$_2$-phenyl)-substituted) at 0.088 kcal mol^{-1}, catalyst 12 (*para*-methoxyphenyl) at 0.107 kcal mol^{-1}, and catalysts 25 and 36 at ≈0.14 kcal mol^{-1}—all well inside chemical accuracy. Three more sit near or just above the threshold (catalysts 37, 34, 39 at 0.20–0.24 kcal mol^{-1}). The single clear outlier is catalyst 4 at 0.369 kcal mol^{-1}, roughly double the overall holdout MAE. Catalyst 4 carries an unusual, sterically elaborate 3,5-diisopropyl-4′-*tert*-butylbiphenyl 3,3′-substituent whose particular combination of bulk and shape is poorly represented among the 35 training catalysts; with no structurally close analogue to interpolate from, the model has little basis for an accurate estimate. This behavior—accurate prediction for catalysts resembling the training set and degraded prediction for structurally isolated ones—is the expected signature of an applicability-domain limit, and it

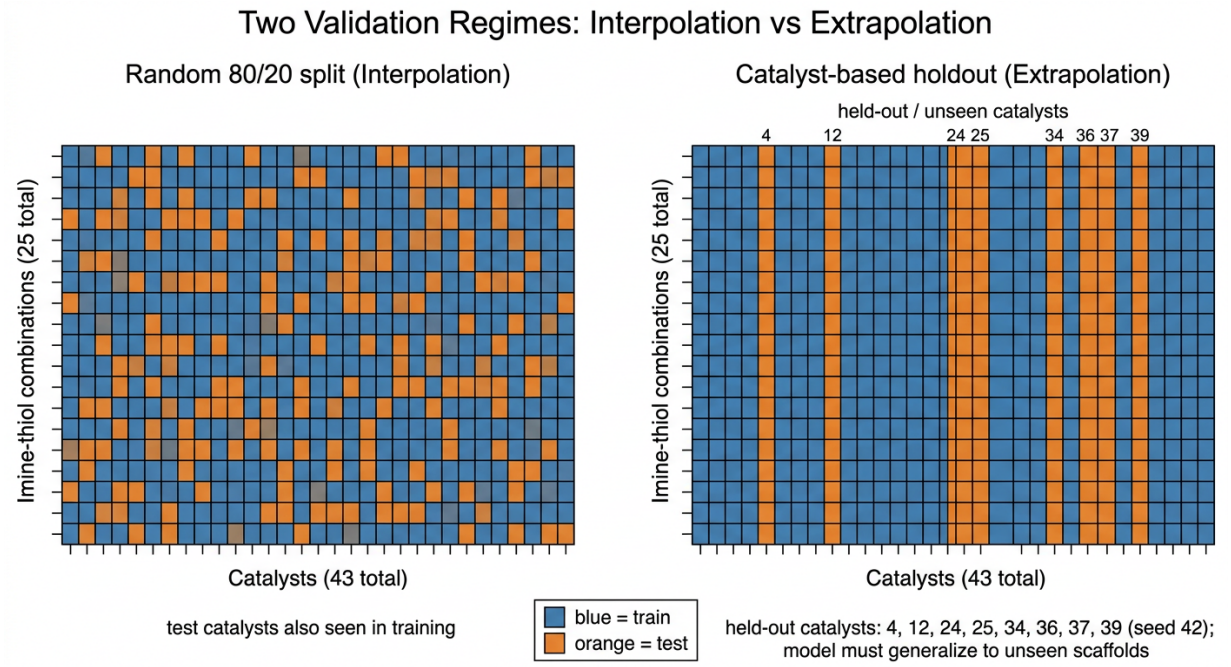


Figure 4: Two validation regimes. Schematic of the reaction matrix (columns: 43 catalysts; rows: 25 imine–thiol pairs). Left: the random 80/20 split scatters test reactions (orange) across all catalysts, so test catalysts are also seen in training (interpolation). Right: the catalyst-based holdout removes eight entire catalyst columns (IDs 4, 12, 24, 25, 34, 36, 37, 39; seed 42) as the test set, forcing generalization to unseen scaffolds (extrapolation).

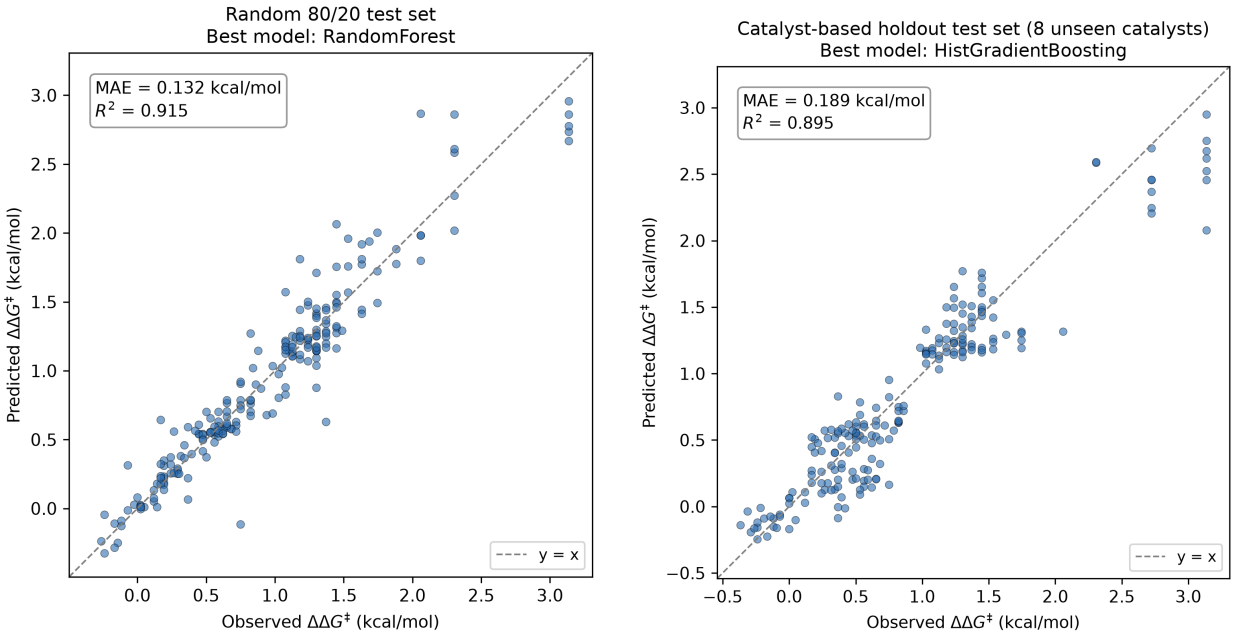
identifies exactly which new scaffolds would most benefit from additional experimental data.

3.4 The interpolation–extrapolation gap

Comparing the two regimes directly quantifies the cost of generalizing across catalyst space. For the best model in each case, the MAE rises from 0.132 to 0.189 kcal mol^{-1}—a 43% increase—while R^2 falls modestly from 0.915 to 0.895. That the R^2 penalty is so small is itself informative: even when every test catalyst is unseen, the model still explains nearly 90% of the variance in $\Delta\Delta G^{\ddagger}$, meaning it has learned genuinely transferable structure–selectivity relationships rather than merely memorizing catalyst identities. The gap is concentrated in absolute error and in the tails of the error distribution rather than in a collapse of rank ordering. Notably, the *ranking* of model robustness differs from the ranking of raw interpolation accuracy: the random forest is marginally best at interpolation, but histogram gradient boosting is both best and most stable under extrapolation, making it the model of choice when the intended use is prediction for new catalysts.

4 Discussion

The central finding of this study is quantitative and, we believe, general: on the Denmark 1,075-reaction benchmark, a transparent model built from two-dimensional fingerprints and physicochemical descriptors predicts $\Delta\Delta G^{\ddagger}$ to within chemical accuracy under interpolation (0.132 kcal mol^{-1}) and remains within chemical accuracy even under strict catalyst extrapolation (0.189 kcal mol^{-1}),



(a) Random 80/20 split (interpolation)

(b) Eight-catalyst holdout (extrapolation)

Figure 5: Predicted versus observed $\Delta\Delta G^{\ddagger}$ for the two test sets. (a) Random-split test set (215 reactions, random forest): MAE $= 0.132\,\mathrm{kcal\,mol^{-1}}$, $R^2 = 0.915$. (b) Catalyst-holdout test set (200 reactions, histogram gradient boosting): MAE $= 0.189\,\mathrm{kcal\,mol^{-1}}$, $R^2 = 0.895$. Dashed line is $y = x$. Extrapolation to unseen catalysts adds scatter but preserves strong overall correlation.

with a well-characterized $\sim$43% error penalty for the harder task. Each of these observations merits interpretation.

Why extrapolation is harder than interpolation. The difference between the two regimes is not an artifact of the algorithm but a direct consequence of what each task demands. Under the random split, the model must estimate the selectivity of a known catalyst on a substrate pair for which that catalyst has already been measured many times in other combinations; the relevant catalyst "fingerprint in selectivity space" is effectively pinned by the training data, and the remaining task—interpolating across the small 5×5 substrate grid—is easy. Under the catalyst holdout, the model has never seen the test catalyst at all and must infer its entire selectivity behavior purely from molecular features, relying on the assumption that catalysts with similar descriptors induce similar selectivity. That assumption holds well on average but fails locally wherever a test catalyst occupies a sparsely sampled region of chemical space, as catalyst 4 vividly illustrates. This is the familiar problem of the *applicability domain*: predictions are trustworthy inside the region spanned by the training set and progressively less so outside it. The broader machine-learning-for-chemistry literature has repeatedly cautioned that random splits can substantially overstate real-world performance, and that reagent- or scaffold-based splits give a more honest estimate of how a model will behave on genuinely new chemistry [13–15]. Our results make that cautionary principle concrete on a clean, fully factorial catalysis benchmark and show that—at least for tree ensembles on this system—the honest estimate remains encouragingly good.

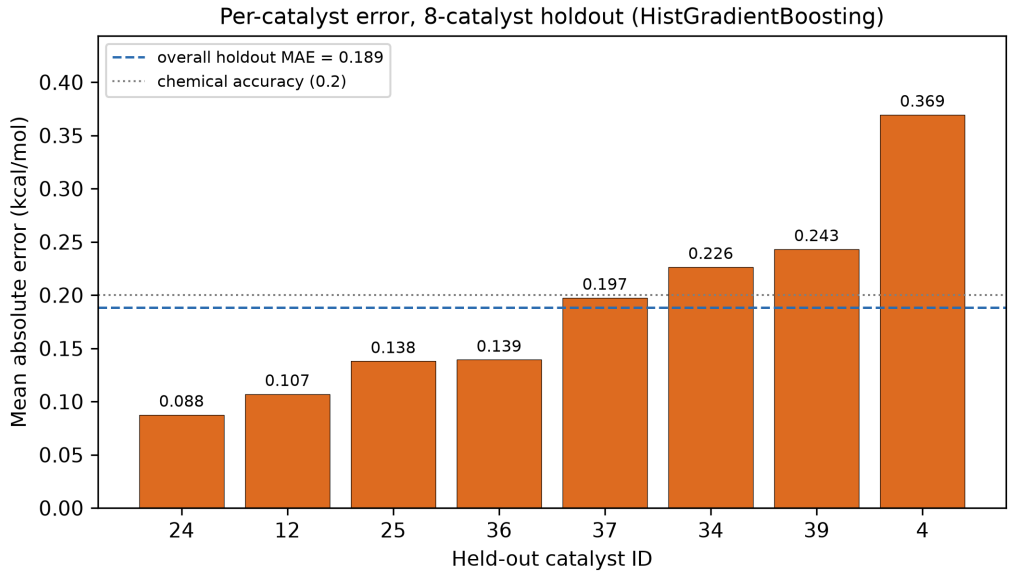


Figure 6: Per-catalyst error on the holdout set. Mean absolute error for each of the eight held-out catalysts (histogram gradient boosting, 25 reactions each). Five catalysts fall at or below chemical accuracy; catalyst 4, bearing a sterically distinctive biphenyl substituent with no close training analogue, is the clear outlier. The dashed line is the overall holdout MAE (0.189 kcal mol^{-1}); the dotted line marks 0.2 kcal mol^{-1}.

Why tree ensembles win. That random forests and gradient boosting dominate Ridge and SVR is consistent with the structure of the problem and with long experience in cheminformatics [20, 21]. The selectivity of a CPA catalyst is a strongly nonlinear, interaction-rich function of local structural motifs—the identity of the 3,3′-substituents, the electronic character of the aryl groups, the presence of a hydrogenated backbone—and these motifs are exactly what Morgan-fingerprint bits encode. Decision-tree ensembles excel at capturing conjunctions of such binary substructural features and at ignoring the many irrelevant bits, whereas a linear model must assign a single additive weight to each bit and an RBF kernel must find a single global length scale. The greater robustness of gradient boosting under extrapolation likely reflects its stagewise, regularized additive construction, which tends to produce smoother extrapolations beyond the training envelope than the piecewise-constant averaging of a random forest.

Chemical accuracy and its practical meaning. The 0.2 kcal mol^{-1} threshold is a useful yardstick because it corresponds to the scale of energetic differences that separate a synthetically "good" catalyst from a "great" one. An error of 0.19 kcal mol^{-1} is comparable to the run-to-run experimental uncertainty in measuring small $\Delta\Delta G^{\ddagger}$ values and is small relative to the 3.5 kcal mol^{-1} span of the dataset. In practical terms, a model at this accuracy can reliably triage a virtual library of candidate catalysts—confidently discarding poor performers and nominating a short list of promising ones for synthesis—even when those candidates are new scaffolds, which is precisely the prospective use that motivated the original study [6]. The per-catalyst analysis further shows how to deploy such a model responsibly: predictions should be trusted most for candidates that are structurally close to the training set and flagged for experimental confirmation when, like catalyst 4, they venture into under-sampled regions.

Relationship to the original study and to descriptor design. It is worth emphasizing what this baseline does *not* do. It does not use the three-dimensional average-steric-occupancy descriptors or the conformer ensembles that were the intellectual core of the Denmark workflow [6], nor the DFT-derived electronic descriptors common in physical-organic modeling [2, 3]. That a purely two-dimensional, seconds-to-compute representation already reaches chemical accuracy on interpolation and near-chemical accuracy on extrapolation is a testament to how much selectivity information is latent in connectivity alone for this reaction family, and it establishes a strong, cheap reference against which richer three-dimensional or quantum-chemical descriptors—and modern graph-neural-network approaches to stereoselectivity—can be judged. We would expect the largest gains from such richer descriptors to appear precisely in the extrapolation regime and specifically for structurally isolated catalysts like catalyst 4, where two-dimensional similarity is least informative about the shape of the chiral pocket.

Limitations and future directions. Several limitations frame these results. The benchmark is deliberately narrow—a single reaction class, five imines, and five thiols—so the substrate-generalization question is not stressed here; a complementary substrate-holdout analysis would round out the picture. The eight held-out catalysts, though randomly chosen, represent one particular draw; averaging performance over repeated random catalyst holdouts, or over a systematic leave-one-catalyst-out protocol, would yield a more complete distribution of extrapolation errors and tighter uncertainty on the reported gap. Feature attribution (for example, permutation importance or SHAP analysis) would help identify which substructural and physicochemical features drive the predictions and would connect the statistical model back to mechanistic understanding of the enantiodetermining transition state. Finally, coupling the model to an uncertainty estimate and an explicit applicability-domain criterion would let it not only predict $\Delta\Delta G^{\ddagger}$ but also report when a prediction should be trusted—an essential capability for closing the loop between prediction and experiment in catalyst discovery.

5 Conclusion

We built and rigorously validated machine-learning models for the selectivity of CPA-catalyzed thiol additions to *N*-acylimines, using the complete 1,075-reaction dataset from the 2019 *Science* benchmark and predicting the physically meaningful target $\Delta\Delta G^{\ddagger} = -RT\ln(\mathrm{er})$. With an inexpensive, fully reproducible feature set of Morgan fingerprints and RDKit descriptors, tree-ensemble models reached chemical accuracy: a random forest predicted $\Delta\Delta G^{\ddagger}$ with an MAE of $0.132\,\mathrm{kcal\,mol^{-1}}$ ($R^2 = 0.915$) under a random 80/20 split, and histogram-based gradient boosting predicted $\Delta\Delta G^{\ddagger}$ with an MAE of $0.189\,\mathrm{kcal\,mol^{-1}}$ ($R^2 = 0.895$) under a stringent eight-catalyst holdout (IDs {4, 12, 24, 25, 34, 36, 37, 39}, seed 42) in which the test catalysts were never seen during training. The 43% increase in error from interpolation to extrapolation—and its concentration in a single structurally isolated catalyst—quantifies both the promise and the boundary of data-driven catalyst prediction: models learn genuinely transferable structure–selectivity relationships, but their reliability tracks proximity to the training set's chemical space. The complete per-reaction predicted-versus-observed $\Delta\Delta G^{\ddagger}$ tables for both test sets are provided in Appendices A and B.

Data and Reproducibility

The source dataset is the experimental selectivity data of Zahrt et al. [6]. All targets were computed via Equation (1) at 298.15 K. Featurization used RDKit 2026.03; modeling used scikit-learn 1.9 with all random operations seeded at 42. Complete per-reaction predictions for both test sets are tabulated in the appendices and provided as accompanying CSV files (`predictions_random_split.csv`, `predictions_catalyst_holdout.csv`).

References

[1] Kaid C. Harper and Matthew S. Sigman. Three-dimensional correlation of steric and electronic free energy relationships guides asymmetric propargylation. *Science*, 333(6051):1875–1878, 2011. doi: 10.1126/science.1206997.

[2] Jolene P. Reid and Matthew S. Sigman. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature*, 571(7765):343–348, 2019. doi: 10.1038/s41586-019-1384-z.

[3] Andrew F. Zahrt, Soumitra V. Athavale, and Scott E. Denmark. Quantitative structure-selectivity relationships in enantioselective catalysis: Past, present, and future. *Chemical Reviews*, 120(3):1620–1689, 2020. doi: 10.1021/acs.chemrev.9b00425.

[4] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018. doi: 10.1038/s41586-018-0337-2.

[5] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018. doi: 10.1021/acscentsci.7b00572.

[6] Andrew F. Zahrt, Jeremy J. Henle, Brennan T. Rose, Yang Wang, William T. Darrow, and Scott E. Denmark. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science*, 363(6424):eaau5631, 2019. doi: 10.1126/science.aau5631.

[7] Gaurav K. Ingle, Michael G. Mormino, Lukasz Wojtas, and Jon C. Antilla. Chiral phosphoric acid-catalyzed addition of thiols to N-acyl imines: Access to chiral N,S-acetals. *Organic Letters*, 13(18):4822–4825, 2011. doi: 10.1021/ol201899c.

[8] Jin-Sheng Yu, Wen-Biao Wu, and Feng Zhou. The first catalytic asymmetric thioacetalization of imines by chiral phosphoric acid catalysis. *Organic and Biomolecular Chemistry*, 14(7): 2205–2209, 2016. doi: 10.1039/C5OB02495A.

[9] Takahiko Akiyama, Junji Itoh, Kazuhiro Yokota, and Kohei Fuchibe. Enantioselective Mannich-type reaction catalyzed by a chiral Brønsted acid. *Angewandte Chemie International Edition*, 43(12):1566–1568, 2004. doi: 10.1002/anie.200353240.

[10] Daisuke Uraguchi and Masahiro Terada. Chiral Brønsted acid-catalyzed direct Mannich reactions via electrophilic activation. *Journal of the American Chemical Society*, 126(17): 5356–5357, 2004. doi: 10.1021/ja0491533.

[11] Dixit Parmar, Erli Sugiono, Sadiya Raja, and Magnus Rueping. Complete field guide to asymmetric BINOL-phosphate derived Brønsted acid and metal catalysis: History and classification by mode of activation. *Chemical Reviews*, 114(18):9047–9153, 2014. doi: 10.1021/cr5001496.

[12] Henry Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115, 1935. doi: 10.1063/1.1749604.

[13] 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.

[14] Kangway 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.

[15] 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.

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

[17] A. Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J. Bellis, Marleen De Veij, and Andrew R. Leach. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, 12(1):51, 2020. doi: 10.1186/s13321-020-00456-1.

[18] 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.

[19] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995. doi: 10.1007/BF00994018.

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

[21] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003. doi: 10.1021/ci034160g.

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

[23] 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.

[24] 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 30 (NeurIPS 2017)*, pages 3146–3154, 2017.

A Per-reaction predictions: random 80/20 test set

Table 2 lists the observed and predicted $\Delta\Delta G^{\ddagger}$ (kcal mol^{-1}) and absolute error for all 215 reactions in the random-split test set (best model: random forest), ordered by imine, thiol, and catalyst index.

Table 2: Per-reaction observed vs. predicted $\Delta\Delta G^{\ddagger}$ (kcal/mol) for the random 80/20 test set (215 reactions; best model: Random Forest). Im, Th, Cat are imine, thiol, and catalyst indices.

#	Im	Th	Cat	Obs. $\Delta\Delta G^{\ddagger}$	Pred. $\Delta\Delta G^{\ddagger}$	|err|
1	0	0	14	1.051	1.024	0.027
2	0	0	17	0.750	0.749	0.000
3	0	0	19	0.982	0.689	0.293
4	0	0	24	0.821	0.702	0.119
5	0	0	30	0.446	0.543	0.096
6	0	0	31	0.167	0.645	0.478
7	0	0	33	0.474	0.534	0.061
8	0	0	35	0.651	0.602	0.049
9	0	0	37	-0.071	0.314	0.385
10	0	1	0	1.238	1.188	0.050
11	0	1	2	1.370	1.287	0.083
12	0	1	5	2.058	2.868	0.809
13	0	1	7	0.290	0.253	0.036
14	0	1	15	0.821	0.790	0.031
15	0	1	16	1.301	1.275	0.026
16	0	1	22	2.305	2.584	0.279
17	0	1	31	1.238	1.476	0.238
18	0	1	33	1.126	1.173	0.047
19	0	1	37	0.240	0.257	0.017
20	0	1	38	0.651	0.705	0.055
21	0	2	7	0.191	0.232	0.041
22	0	2	13	0.191	0.222	0.031
23	0	2	16	1.126	1.253	0.128
24	0	2	19	1.126	1.129	0.003
25	0	2	24	1.301	1.401	0.100
26	0	2	35	1.301	1.416	0.115
27	0	2	40	0.167	0.221	0.054
28	0	2	42	0.191	0.187	0.004
29	0	3	0	0.716	0.629	0.086
30	0	3	8	0.716	0.604	0.112
31	0	3	13	0.750	-0.114	0.863
32	0	3	21	1.370	1.197	0.173
33	0	3	25	0.367	0.593	0.226
34	0	3	29	0.651	0.668	0.017
35	0	4	0	1.370	1.275	0.095
36	0	4	2	0.821	1.271	0.450
37	0	4	9	1.446	1.325	0.122
38	0	4	12	1.301	1.169	0.132
39	0	4	23	1.446	1.755	0.309
40	0	4	27	1.238	1.238	0.000
41	0	4	32	1.301	1.094	0.207
42	0	4	36	0.167	0.234	0.067
43	1	0	1	1.370	0.631	0.739
44	1	0	3	0.024	0.019	0.004
45	1	0	7	-0.119	-0.128	0.009
46	1	0	8	0.530	0.554	0.024

continued on next page

Table 2 – *continued*

#	Im	Th	Cat	Obs. $\Delta\Delta G^{\ddagger}$	Pred. $\Delta\Delta G^{\ddagger}$	|err|
47	1	0	13	-0.265	-0.239	0.026
48	1	0	20	1.532	1.959	0.427
49	1	0	24	0.750	0.922	0.173
50	1	0	29	0.530	0.654	0.124
51	1	0	30	0.530	0.549	0.019
52	1	0	31	0.651	0.786	0.136
53	1	1	7	0.265	0.258	0.006
54	1	1	17	1.532	1.568	0.036
55	1	1	22	2.305	2.271	0.033
56	1	1	23	1.446	2.065	0.618
57	1	1	24	1.370	1.459	0.089
58	1	1	40	0.240	0.321	0.081
59	1	2	9	1.370	1.369	0.002
60	1	2	14	1.744	2.002	0.258
61	1	2	18	1.629	1.919	0.289
62	1	2	19	1.301	1.150	0.151
63	1	2	21	1.684	1.941	0.256
64	1	2	29	1.370	1.168	0.202
65	1	2	30	1.301	1.146	0.155
66	1	2	36	0.290	0.292	0.002
67	1	2	37	0.290	0.278	0.012
68	1	3	7	0.024	-0.003	0.027
69	1	3	9	0.683	0.580	0.103
70	1	3	11	0.821	0.760	0.061
71	1	3	21	1.488	1.293	0.195
72	1	3	22	1.301	1.296	0.005
73	1	3	23	1.238	1.069	0.169
74	1	3	28	0.716	0.558	0.157
75	1	3	30	0.589	0.569	0.020
76	1	3	32	0.419	0.565	0.146
77	1	3	35	0.750	0.723	0.027
78	1	3	36	0.119	0.133	0.014
79	1	3	41	0.024	0.025	0.001
80	1	3	42	0.024	0.010	0.013
81	1	4	0	1.301	1.254	0.047
82	1	4	7	0.240	0.372	0.132
83	1	4	8	1.238	1.169	0.070
84	1	4	9	1.180	1.290	0.110
85	1	4	10	1.446	1.499	0.052
86	1	4	13	0.502	0.372	0.130
87	1	4	14	1.629	1.810	0.181
88	1	4	36	0.341	0.367	0.027
89	2	0	4	1.180	1.219	0.039
90	2	0	8	0.474	0.498	0.024
91	2	0	9	0.474	0.508	0.034
92	2	0	20	1.180	1.810	0.630
93	2	0	23	1.075	0.828	0.247
94	2	0	25	0.341	0.460	0.119
95	2	0	29	0.619	0.541	0.079
96	2	0	35	0.474	0.543	0.069
97	2	0	39	-0.240	-0.325	0.084
98	2	1	7	0.191	0.349	0.158
99	2	1	10	1.446	1.462	0.015
100	2	1	15	0.898	0.872	0.026
101	2	1	17	1.629	1.445	0.185

continued on next page

Table 2 – *continued*

#	Im	Th	Cat	Obs. $\Delta\Delta G^{\ddagger}$	Pred. $\Delta\Delta G^{\ddagger}$	|err|
102	2	1	21	2.058	1.983	0.075
103	2	1	26	0.878	1.146	0.268
104	2	1	27	1.126	1.105	0.021
105	2	1	36	0.474	0.416	0.057
106	2	1	38	0.589	0.702	0.113
107	2	2	0	1.370	1.248	0.122
108	2	2	3	0.119	0.074	0.045
109	2	2	6	1.301	1.452	0.150
110	2	2	16	1.075	1.176	0.101
111	2	2	17	1.301	1.384	0.083
112	2	2	21	1.882	1.884	0.002
113	2	2	28	1.180	1.242	0.062
114	2	2	30	1.180	1.085	0.095
115	2	2	32	1.126	1.194	0.068
116	2	3	2	0.683	0.578	0.105
117	2	3	8	0.559	0.558	0.002
118	2	3	16	0.530	0.557	0.027
119	2	3	40	-0.167	-0.110	0.057
120	2	3	41	-0.024	0.027	0.051
121	2	4	6	1.629	1.414	0.216
122	2	4	12	1.075	1.151	0.076
123	2	4	15	1.027	0.805	0.222
124	2	4	16	1.075	1.207	0.132
125	2	4	26	1.301	1.143	0.158
126	2	4	31	1.744	1.493	0.251
127	2	4	34	1.446	1.491	0.045
128	2	4	36	0.315	0.382	0.067
129	2	4	42	0.191	0.175	0.017
130	3	0	3	-0.071	-0.011	0.060
131	3	0	10	0.859	0.900	0.041
132	3	0	17	0.589	0.631	0.042
133	3	0	20	1.301	1.711	0.409
134	3	0	21	0.840	1.021	0.181
135	3	0	23	1.301	0.876	0.425
136	3	0	28	0.559	0.481	0.078
137	3	0	37	-0.240	-0.044	0.196
138	3	0	39	-0.167	-0.283	0.116
139	3	0	41	0.024	0.012	0.012
140	3	1	12	1.126	1.107	0.019
141	3	1	13	0.191	0.311	0.120
142	3	1	21	2.058	1.983	0.076
143	3	1	22	3.135	2.670	0.465
144	3	1	24	1.370	1.440	0.069
145	3	1	25	1.075	1.171	0.096
146	3	1	36	0.167	0.322	0.155
147	3	1	42	0.167	0.178	0.011
148	3	2	2	1.238	1.226	0.012
149	3	2	5	2.305	2.612	0.307
150	3	2	6	1.075	1.571	0.496
151	3	2	20	3.135	2.861	0.274
152	3	2	23	2.305	2.019	0.286
153	3	2	26	1.100	1.144	0.044
154	3	2	30	1.126	1.140	0.015
155	3	2	39	0.265	0.558	0.294
156	3	3	0	0.619	0.579	0.041

continued on next page

Table 2 – *continued*

#	Im	Th	Cat	Obs. $\Delta\Delta G^{\ddagger}$	Pred. $\Delta\Delta G^{\ddagger}$	|err|
157	3	3	1	0.619	0.546	0.073
158	3	3	2	0.446	0.543	0.096
159	3	3	4	1.446	1.275	0.171
160	3	3	7	0.143	0.010	0.133
161	3	3	9	0.589	0.548	0.041
162	3	3	10	1.027	0.977	0.051
163	3	3	11	0.939	0.677	0.262
164	3	3	16	0.559	0.535	0.025
165	3	3	18	1.075	1.126	0.051
166	3	3	30	0.589	0.524	0.065
167	3	3	35	0.502	0.703	0.201
168	3	3	36	0.000	0.080	0.080
169	3	3	42	0.047	0.011	0.036
170	3	4	6	1.446	1.551	0.105
171	3	4	15	0.651	0.767	0.116
172	3	4	20	3.135	2.958	0.177
173	3	4	25	1.301	1.252	0.049
174	3	4	32	1.301	1.177	0.124
175	3	4	41	0.191	0.135	0.056
176	4	0	4	1.446	1.313	0.134
177	4	0	6	0.750	0.908	0.158
178	4	0	7	-0.119	-0.090	0.029
179	4	0	9	0.446	0.608	0.162
180	4	0	10	0.982	1.034	0.052
181	4	0	13	-0.143	-0.250	0.107
182	4	0	27	0.619	0.566	0.054
183	4	0	29	0.651	0.590	0.061
184	4	0	36	0.367	0.067	0.300
185	4	0	37	0.367	0.220	0.147
186	4	1	3	0.119	0.052	0.067
187	4	1	5	3.135	2.736	0.399
188	4	1	15	0.750	0.786	0.037
189	4	1	28	1.180	1.252	0.072
190	4	1	30	1.180	1.115	0.065
191	4	1	33	1.238	1.225	0.013
192	4	2	4	2.305	2.861	0.556
193	4	2	15	0.821	0.784	0.037
194	4	2	18	2.058	1.801	0.258
195	4	2	19	1.446	1.163	0.284
196	4	2	21	1.882	1.777	0.105
197	4	2	31	1.532	1.758	0.226
198	4	2	32	1.075	1.218	0.143
199	4	2	33	1.301	1.153	0.148
200	4	2	34	1.180	1.443	0.263
201	4	2	35	1.238	1.501	0.263
202	4	2	40	0.303	0.254	0.049
203	4	2	42	0.143	0.179	0.037
204	4	3	0	0.559	0.595	0.036
205	4	3	14	1.152	1.246	0.093
206	4	3	23	1.301	1.037	0.264
207	4	3	31	1.075	0.881	0.194
208	4	3	32	0.589	0.596	0.007
209	4	3	33	0.651	0.618	0.032
210	4	3	35	0.821	0.672	0.149
211	4	4	2	1.370	1.341	0.029

continued on next page

Table 2 – *continued*

#	Im	Th	Cat	Obs. $\Delta\Delta G^{\ddagger}$	Pred. $\Delta\Delta G^{\ddagger}$	|err|
212	4	4	5	3.135	2.778	0.357
213	4	4	14	1.629	1.774	0.145
214	4	4	18	1.744	1.722	0.021
215	4	4	36	0.393	0.395	0.002

B Per-reaction predictions: eight-catalyst holdout test set

Table 3 lists the observed and predicted $\Delta\Delta G^{\ddagger}$ (kcal mol^{-1}) and absolute error for all 200 reactions in the catalyst-holdout test set (best model: histogram gradient boosting), grouped by held-out catalyst ID.

Table 3: Per-reaction observed vs. predicted $\Delta\Delta G^{\ddagger}$ (kcal/mol) for the 8-catalyst holdout test set (200 reactions; best model: HistGradientBoosting). Grouped by held-out catalyst ID.

#	Im	Th	Cat	Obs. $\Delta\Delta G^{\ddagger}$	Pred. $\Delta\Delta G^{\ddagger}$	|err|
1	0	0	4	1.238	1.652	0.414
2	0	1	4	3.135	2.525	0.610
3	0	2	4	3.135	2.456	0.678
4	0	3	4	1.301	1.771	0.470
5	0	4	4	3.135	2.619	0.516
6	1	0	4	1.180	1.498	0.318
7	1	1	4	2.721	2.206	0.515
8	1	2	4	3.135	2.077	1.057
9	1	3	4	1.238	1.497	0.259
10	1	4	4	2.721	2.246	0.475
11	2	0	4	1.180	1.376	0.196
12	2	1	4	2.305	2.593	0.289
13	2	2	4	2.721	2.368	0.353
14	2	3	4	1.238	1.567	0.329
15	2	4	4	2.721	2.460	0.262
16	3	0	4	1.301	1.453	0.152
17	3	1	4	3.135	2.675	0.459
18	3	2	4	2.721	2.457	0.264
19	3	3	4	1.446	1.605	0.159
20	3	4	4	3.135	2.752	0.383
21	4	0	4	1.446	1.717	0.271
22	4	1	4	2.721	2.695	0.027
23	4	2	4	2.305	2.587	0.282
24	4	3	4	1.446	1.760	0.313
25	4	4	4	3.135	2.948	0.186
26	0	0	12	0.446	0.618	0.172
27	0	1	12	1.238	1.229	0.009
28	0	2	12	1.126	1.267	0.142
29	0	3	12	0.530	0.785	0.255
30	0	4	12	1.301	1.161	0.140
31	1	0	12	0.393	0.458	0.065
32	1	1	12	1.301	1.173	0.128
33	1	2	12	1.301	1.220	0.081
34	1	3	12	0.393	0.588	0.196
35	1	4	12	1.301	1.123	0.178
36	2	0	12	0.341	0.407	0.067
37	2	1	12	1.027	1.171	0.143
38	2	2	12	1.075	1.175	0.100
39	2	3	12	0.502	0.558	0.057
40	2	4	12	1.075	1.191	0.116
41	3	0	12	0.393	0.322	0.071
42	3	1	12	1.126	1.136	0.010
43	3	2	12	1.027	1.161	0.134
44	3	3	12	0.502	0.508	0.007
45	3	4	12	1.238	1.231	0.007

continued on next page

Table 3 – *continued*

#	Im	Th	Cat	Obs. $\Delta\Delta G^{\ddagger}$	Pred. $\Delta\Delta G^{\ddagger}$	|err|
46	4	0	12	0.651	0.527	0.123
47	4	1	12	1.238	1.244	0.006
48	4	2	12	1.027	1.331	0.303
49	4	3	12	0.559	0.621	0.061
50	4	4	12	1.126	1.230	0.104
51	0	0	24	0.821	0.720	0.101
52	0	1	24	1.446	1.436	0.010
53	0	2	24	1.301	1.521	0.220
54	0	3	24	0.859	0.719	0.139
55	0	4	24	1.532	1.424	0.108
56	1	0	24	0.750	0.953	0.203
57	1	1	24	1.370	1.343	0.027
58	1	2	24	1.532	1.552	0.020
59	1	3	24	0.821	0.751	0.070
60	1	4	24	1.446	1.500	0.054
61	2	0	24	0.619	0.536	0.083
62	2	1	24	1.370	1.388	0.018
63	2	2	24	1.370	1.428	0.058
64	2	3	24	0.716	0.612	0.104
65	2	4	24	1.238	1.377	0.139
66	3	0	24	0.446	0.530	0.084
67	3	1	24	1.370	1.509	0.139
68	3	2	24	1.446	1.463	0.017
69	3	3	24	0.821	0.631	0.190
70	3	4	24	1.446	1.652	0.206
71	4	0	24	0.502	0.601	0.100
72	4	1	24	1.301	1.285	0.016
73	4	2	24	1.301	1.348	0.047
74	4	3	24	0.530	0.535	0.004
75	4	4	24	1.446	1.479	0.032
76	0	0	25	0.474	0.571	0.097
77	0	1	25	1.238	1.143	0.095
78	0	2	25	1.027	1.152	0.125
79	0	3	25	0.367	0.567	0.200
80	0	4	25	1.180	1.163	0.017
81	1	0	25	0.367	0.829	0.462
82	1	1	25	1.126	1.110	0.016
83	1	2	25	1.370	1.200	0.170
84	1	3	25	0.502	0.635	0.134
85	1	4	25	1.180	1.258	0.078
86	2	0	25	0.341	0.404	0.063
87	2	1	25	1.126	1.033	0.093
88	2	2	25	1.027	1.077	0.050
89	2	3	25	0.619	0.478	0.141
90	2	4	25	1.370	1.179	0.192
91	3	0	25	0.446	0.408	0.038
92	3	1	25	1.075	1.146	0.071
93	3	2	25	1.027	1.146	0.119
94	3	3	25	0.683	0.497	0.185
95	3	4	25	1.301	1.257	0.044
96	4	0	25	0.651	0.611	0.039
97	4	1	25	0.982	1.196	0.214
98	4	2	25	1.532	1.237	0.295
99	4	3	25	0.750	0.506	0.244
100	4	4	25	1.532	1.264	0.268

continued on next page

Table 3 – *continued*

#	Im	Th	Cat	Obs. $\Delta\Delta G^{\ddagger}$	Pred. $\Delta\Delta G^{\ddagger}$	|err|
101	0	0	34	0.785	0.572	0.212
102	0	1	34	1.532	1.182	0.349
103	0	2	34	1.301	1.230	0.071
104	0	3	34	0.821	0.646	0.175
105	0	4	34	1.238	1.163	0.075
106	1	0	34	0.750	0.822	0.072
107	1	1	34	1.744	1.193	0.551
108	1	2	34	1.238	1.322	0.084
109	1	3	34	0.859	0.757	0.101
110	1	4	34	1.744	1.315	0.428
111	2	0	34	0.502	0.551	0.049
112	2	1	34	1.446	1.177	0.269
113	2	2	34	1.370	1.222	0.148
114	2	3	34	0.651	0.647	0.003
115	2	4	34	1.446	1.159	0.288
116	3	0	34	0.559	0.601	0.042
117	3	1	34	1.744	1.305	0.438
118	3	2	34	1.629	1.294	0.336
119	3	3	34	0.651	0.744	0.094
120	3	4	34	2.058	1.316	0.743
121	4	0	34	0.821	0.631	0.190
122	4	1	34	1.446	1.195	0.251
123	4	2	34	1.180	1.195	0.015
124	4	3	34	0.821	0.639	0.182
125	4	4	34	1.744	1.250	0.494
126	0	0	36	-0.095	-0.161	0.066
127	0	1	36	0.315	0.125	0.191
128	0	2	36	0.240	0.100	0.140
129	0	3	36	0.530	0.091	0.440
130	0	4	36	0.167	0.179	0.012
131	1	0	36	-0.119	-0.150	0.032
132	1	1	36	0.240	0.263	0.023
133	1	2	36	0.290	0.177	0.113
134	1	3	36	0.119	0.029	0.090
135	1	4	36	0.341	0.277	0.064
136	2	0	36	0.000	-0.169	0.169
137	2	1	36	0.474	0.261	0.213
138	2	2	36	0.240	0.178	0.062
139	2	3	36	0.000	0.022	0.022
140	2	4	36	0.315	0.311	0.004
141	3	0	36	-0.265	-0.165	0.100
142	3	1	36	0.167	0.271	0.104
143	3	2	36	0.167	0.239	0.073
144	3	3	36	0.000	0.064	0.064
145	3	4	36	0.502	0.444	0.058
146	4	0	36	0.367	-0.085	0.452
147	4	1	36	0.502	0.212	0.289
148	4	2	36	0.367	0.204	0.162
149	4	3	36	0.419	-0.013	0.433
150	4	4	36	0.393	0.288	0.105
151	0	0	37	-0.071	-0.076	0.005
152	0	1	37	0.240	0.417	0.177
153	0	2	37	0.191	0.407	0.216
154	0	3	37	0.750	0.164	0.585
155	0	4	37	0.167	0.448	0.282

continued on next page

Table 3 – *continued*

#	Im	Th	Cat	Obs. $\Delta\Delta G^{\ddagger}$	Pred. $\Delta\Delta G^{\ddagger}$	|err|
156	1	0	37	-0.143	-0.074	0.069
157	1	1	37	0.265	0.573	0.308
158	1	2	37	0.290	0.515	0.225
159	1	3	37	0.119	0.109	0.010
160	1	4	37	0.367	0.553	0.186
161	2	0	37	0.047	-0.102	0.149
162	2	1	37	0.502	0.535	0.033
163	2	2	37	0.216	0.478	0.262
164	2	3	37	0.000	0.064	0.064
165	2	4	37	0.315	0.578	0.263
166	3	0	37	-0.240	-0.118	0.122
167	3	1	37	0.191	0.506	0.315
168	3	2	37	0.167	0.523	0.356
169	3	3	37	0.024	0.109	0.085
170	3	4	37	0.530	0.691	0.160
171	4	0	37	0.367	-0.005	0.372
172	4	1	37	0.559	0.505	0.054
173	4	2	37	0.341	0.519	0.178
174	4	3	37	0.393	0.071	0.322
175	4	4	37	0.419	0.555	0.136
176	0	0	39	-0.290	-0.191	0.099
177	0	1	39	0.651	0.206	0.445
178	0	2	39	0.530	0.128	0.403
179	0	3	39	-0.216	-0.011	0.205
180	0	4	39	0.651	0.208	0.443
181	1	0	39	-0.240	-0.158	0.082
182	1	1	39	0.367	0.151	0.215
183	1	2	39	0.559	0.149	0.410
184	1	3	39	-0.071	-0.059	0.012
185	1	4	39	0.683	0.321	0.362
186	2	0	39	-0.240	-0.245	0.005
187	2	1	39	0.474	0.203	0.271
188	2	2	39	0.619	0.144	0.475
189	2	3	39	-0.191	-0.088	0.103
190	2	4	39	0.589	0.274	0.315
191	3	0	39	-0.167	-0.225	0.058
192	3	1	39	0.589	0.202	0.387
193	3	2	39	0.265	0.126	0.139
194	3	3	39	-0.315	-0.035	0.280
195	3	4	39	0.619	0.361	0.258
196	4	0	39	-0.367	-0.139	0.228
197	4	1	39	0.559	0.194	0.365
198	4	2	39	0.341	0.129	0.212
199	4	3	39	-0.119	-0.090	0.029
200	4	4	39	0.559	0.283	0.276