doi.org/10.26434/chemrxiv.13087769.v1

Multi-Label Classification Models for the Prediction of Cross-Coupling Reaction Conditions

Michael Maser, Alexander Cui, Serim Ryou, Travis DeLano, Yisong Yue, Sarah Reisman

Submitted date: 14/10/2020 • Posted date: 15/10/2020

Citation information: Maser, Michael; Cui, Alexander; Ryou, Serim; DeLano, Travis; Yue, Yisong; Reisman, Sarah (2020): Multi-Label Classification Models for the Prediction of Cross-Coupling Reaction Conditions. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.13087769.v1

Machine-learned ranking models have been developed for the prediction of substrate-specific cross-coupling reaction conditions. Datasets of published reactions were curated for Suzuki, Negishi, and C–N couplings, as well as Pauson–Khand reactions. String, descriptor, and graph encodings were tested as input representations, and models were trained to predict the set of conditions used in a reaction as a binary vector. Unique reagent dictionaries categorized by expert-crafted reaction roles were constructed for each dataset, leading to context-aware predictions. We find that relational graph convolutional networks and gradient-boosting machines are very effective for this learning task, and we disclose a novel reaction-level graph-attention operation in the top-performing model.

Multi-Label Classification Models for the Prediction of Cross-Coupling Reaction Conditions

Michael R. Maser,[†,§] Alexander Y. Cui,[‡,§] Serim Ryou,[¶,§] Travis J. DeLano,[†] Yisong Yue,[‡] and Sarah E. Reisman[*,†]

†*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California, USA*

‡*Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA*

¶*Computational Vision Lab, California Institute of Technology, Pasadena, California, USA*

§*Equal contribution.*

E-mail: reisman@caltech.edu

Abstract

Machine-learned ranking models have been developed for the prediction of substrate-specific cross-coupling reaction conditions. Datasets of published reactions were curated for Suzuki, Negishi, and C–N couplings, as well as Pauson–Khand reactions. String, descriptor, and graph encodings were tested as input representations, and models were trained to predict the set of conditions used in a reaction as a binary vector. Unique reagent dictionaries categorized by expert-crafted reaction roles were constructed for each dataset, leading to context-aware predictions. We find that relational graph convolutional networks and gradient-boosting machines are very effective for this learning task, and we disclose a novel reaction-level graph-attention operation in the top-performing model.

1 Introduction

A common roadblock encountered in organic synthesis occurs when canonical conditions for a given reaction type fail in complex molecule settings.[1] Optimizing these reactions frequently requires iterative experimentation that can slow progress, waste material, and add significant costs to research.[2] This is especially prevalent in catalysis, where the substrate-specific nature of reported conditions is often deemed a major drawback, leading to the slow adoption of new methods.[1–3] If, however, a transformation's structure-reactivity relationships (SRRs) were well-known or predictable, this roadblock could be avoided and new reactions could see much broader use in the field.[4]

Machine learning (ML) algorithms have demonstrated great promise as predictive tools for chemistry domain tasks.[5] Strong approaches to molecular property prediction[6–9] and generative design[10–13] have been developed, particularly in the field of medicinal chemistry.[14] Some applications have emerged in organic synthesis, geared mainly towards predicting reaction products,[15,16] yield,[17–20] and selectivity.[21–25] Significant effort has also been invested in computer-aided synthesis planning (CASP)[26] and the development of retrosynthetic design algorithms.[27–30]

To supplement these tools, initial attempts have been made to predict reaction conditions in the forward direction based on the substrates and products involved.[31] Thus far, studies have focused on global datasets with millions of data points of mixed reaction types. Advantages of this approach include ample training data and the ability to query any transformation with a single model. However, the sparse representation of individual reactions is a major drawback, in that reliable predictions can likely only be expected for the most common reactions and conditions within. This precludes the ability to distinguish subtle variations in substrate structures that lead to different condition requirements, which is critical for SRR modeling.

In recent years, it has become a goal of ours to develop predictive tools to overcome challenges in selecting substrate-specific reaction conditions. Towards this end, we recently reported a preliminary study of graph neural networks (GNNs) as multi-label classification (MLC) models for this task.[32] We selected four high-value reaction types from the cross-coupling literature as testing grounds: Suzuki, C–N, and Negishi couplings, as well as Pauson-Khand reactions (PKRs).[33] Modeling studies indicated relational graph convolutional networks (R-GCNs)[34] as uniquely suited for our learning problem. We herein report the full scope of our studies, including improvements to the R-GCN architecture and an alternative tree-based learning approach using gradient-boosting machines (GBMs).[35]

2 Approach and Methods

A schematic representation of the overall approach is included in Figure 1. We direct the reader to our initial report[32] for additional procedural explanations.[i]

2.1 Data acquisition and preprocessing

A summary of the datasets studied here is shown in Table 1. Each dataset was manually preprocessed using the following procedure:

1. Reaction data was exported from Reaxys® query results (Figure 1A).[33,36]

2. SMILES strings[37] of coupling partners and major products were identified for each reaction entry (i.e., data point).

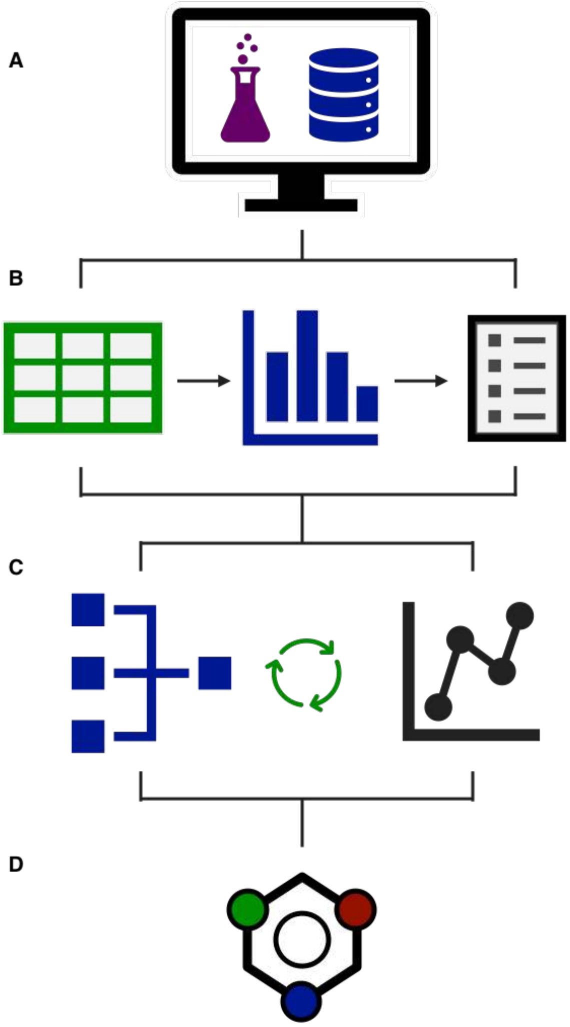


Figure 1: Schematic modeling workflow. A) Data gathering. B) Tabulation and dictionary construction. C) Iterative model optimization. D) Inference and interpretation.

3. Condition labels including reagents, catalysts, solvents, temperatures, etc. were extracted for each data point (Figure 1B).

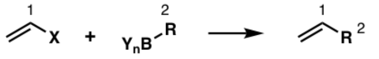
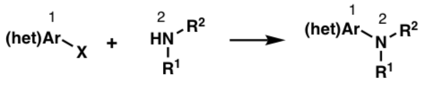
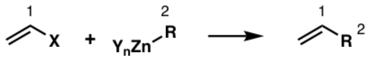
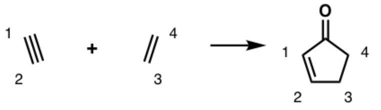
4. All unique labels were enumerated into a dataset dictionary, which was sorted by reaction role and trimmed at a threshold frequency to avoid sparsity.

5. Labels were re-indexed within categories and applied to the raw data to construct binary condition vectors for each reaction. We refer to this process as binning.

The reactions studied here were chosen for their ubiquity and value in synthesis, breadth

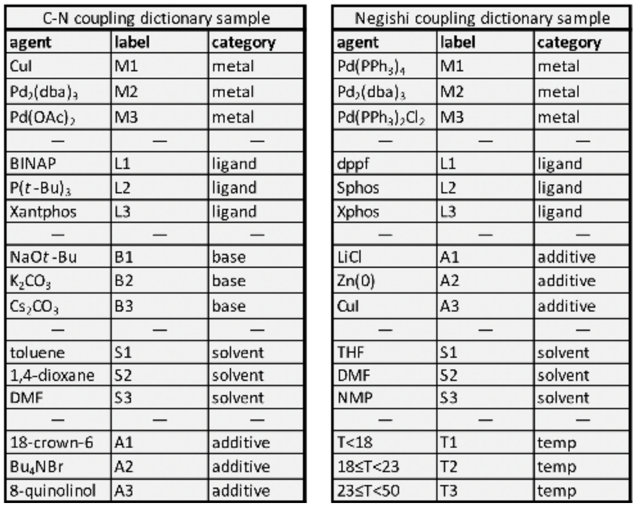
[i] We make our full modeling and data processing code freely available at `https://github.com/slryou41/reaction-gcnn`.

Table 1: Statistical summary of reaction datasets with Reaxys® queries.

name	depiction	reactions	raw labels	label bins	categories
Suzuki		145,413	3,315	118	5
C–N		36,519	1,528	205	5
Negishi		6,391	492	105	5
PKR		2,749	335	83	8

of known conditions, and range of dataset size and chemical space.[ii] It should be noted that certain parameters (e.g. temperature, pressure, etc.) were more fully recorded in some datasets than others. In cases where this data was well-represented, reactions with missing values were simply removed, or in the case of temperature and pressure were assumed to occur ambiently. However, when appropriate, these parameters were dropped from the prediction space to avoid discarding large portions of data.

The Suzuki dataset (Table 1, line 1) was obtained from a search of C–C bond-forming reactions between C(sp^2) halides or pseudohalides and organoboron species. Data processing returned 145k reactions with 118 label bins in 5 categories. Similarly, the C–N coupling dataset (line 2) details reactions between aryl (pseudo)halides and amines, with 37k reactions and 205 bins in 5 categories. The Negishi dataset (line 3) contains C–C bond-forming reactions between organozinc compounds and C(sp^2) (pseudo)halides. After processing, this dataset gave 6.4k reactions with 105 bins in 5 categories. The PKR dataset (line 4) describes couplings of C–C double bonds with C–C triple bonds to form the corresponding cyclopentenones, containing 2.7k reactions with 83 bins in 8 categories. For all datasets, atom mapping was used as depicted in Table 1 to ensure only the desired transformation type was obtained.[iii] Samples of the C–N and Negishi label dictionaries are included in Figure 2, and full dictionaries for all reactions are provided in the SI.



C-N coupling dictionary sample		
agent	label	category
CuI	M1	metal
$Pd_2(dba)_3$	M2	metal
$Pd(OAc)_2$	M3	metal
—	—	—
BINAP	L1	ligand
$P(t\text{-}Bu)_3$	L2	ligand
Xantphos	L3	ligand
—	—	—
NaOt-Bu	B1	base
K_2CO_3	B2	base
Cs_2CO_3	B3	base
—	—	—
toluene	S1	solvent
1,4-dioxane	S2	solvent
DMF	S3	solvent
—	—	—
18-crown-6	A1	additive
Bu_4NBr	A2	additive
8-quinolinol	A3	additive

Negishi coupling dictionary sample		
agent	label	category
$Pd(PPh_3)_4$	M1	metal
$Pd_2(dba)_3$	M2	metal
$Pd(PPh_3)_2Cl_2$	M3	metal
—	—	—
dppf	L1	ligand
Sphos	L2	ligand
Xphos	L3	ligand
—	—	—
LiCl	A1	additive
Zn(0)	A2	additive
CuI	A3	additive
—	—	—
THF	S1	solvent
DMF	S2	solvent
NMP	S3	solvent
—	—	—
$T<18$	T1	temp
$18 \leq T<23$	T2	temp
$23 \leq T<50$	T3	temp

Figure 2: Samples of categorized reaction dictionaries for C-N and Negishi datasets.

2.2 Model setup

For each dataset, an 80/10/10 train/validation/test split was used in modeling. Training and test sets were kept consistent between model types for sake of comparability. Model inputs were prepared as reactant/product structure tuples, with encodings tailored to each learning method. Models were trained using binary

[ii] Detailed molecular property distributions for each dataset can be found with our previous studies.[32]

[iii] Given their relative frequency and to maintain consistent formatting, intramolecular couplings were dropped from the first three reactions but were retained for the PKR dataset.

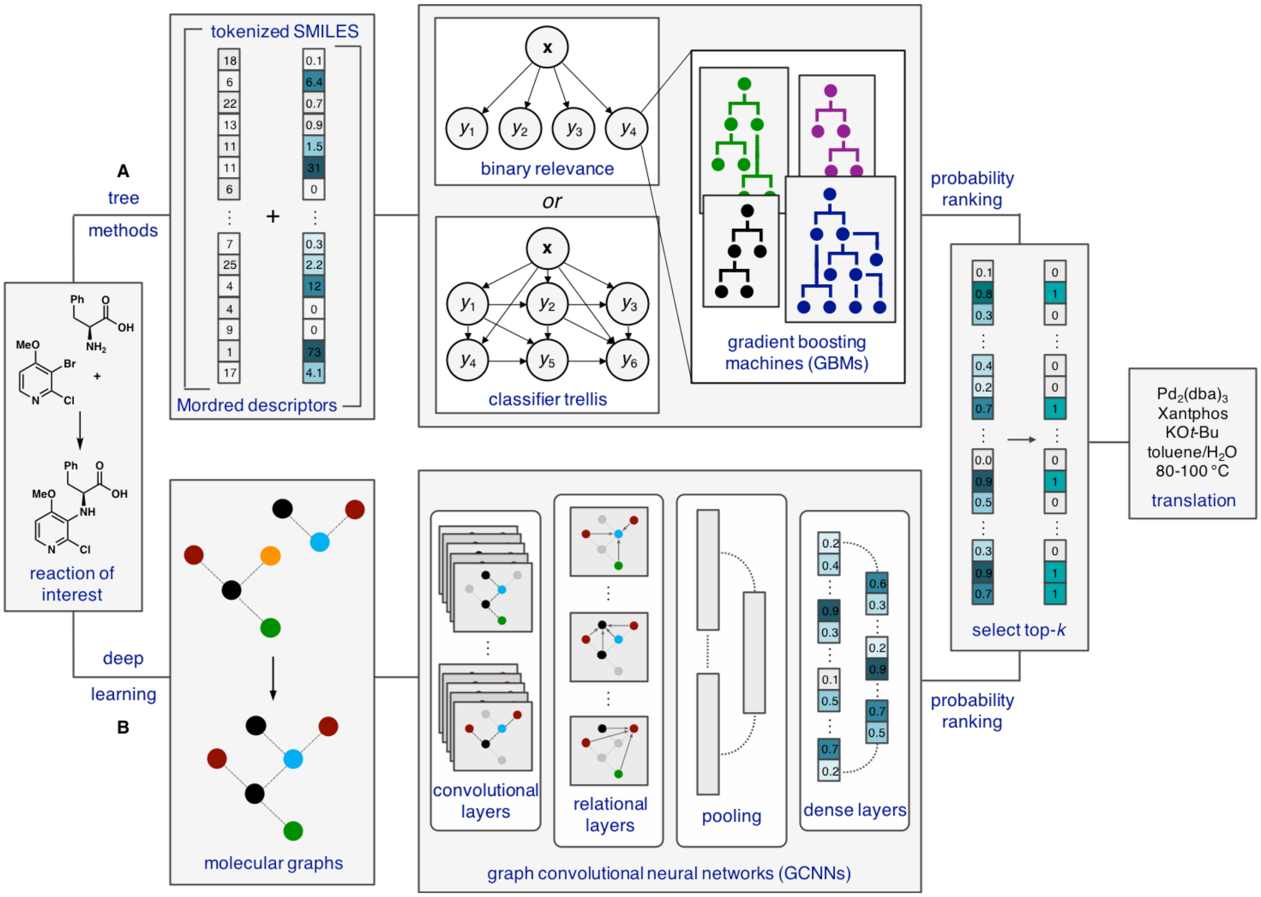


Figure 3: Schematic modeling workflow. A) Tree-based methods. String and descriptor vectors for each molecule in a reaction are concatenated and used as inputs to gradient-boosting machines (GBMs). B) Deep learning methods. Molecular graphs are constructed for each molecule in a reaction, which are passed as inputs to a graph convolutional neural network (GCNN). Both model types predict probability rankings for the full reaction dictionary, which are sorted by reaction role and translated to the final output.

cross-entropy loss to output probability scores for all reagent/condition labels in the reaction dictionary (Figure 1C). The top-k ranked labels in each dictionary category were selected as the final prediction, where k is user-determined.

We define an accurate prediction as one where the ground-truth label appears in the top-k predicted labels. Given the variable class-imbalance in each dictionary category,[32,38] accuracy is evaluated at the categorical level as follows:

$$A_c = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{Y}_i \cap Y_i]\,, \qquad (1)$$

where $\hat{Y}_i$ and Y_i are the sets of top-k predicted and ground truth labels for the i-th sample in category c, respectively. The correct instances are summed and divided by the number of samples in the test set, N, to give the overall test accuracy in the category, or A_c.[39]

As a general measure of a model's performance, we calculate its average error reduction (AER) from a baseline predictor (**dummy**) that always predicts the top-k most frequently occurring dataset labels in each category:

$$\text{AER} = \frac{1}{C}\sum_{c=1}^{C} \frac{A_c^g - A_c^d}{1 - A_c^d}\,, \qquad (2)$$

where A_c^g and A_c^d are the accuracies of the GNN and dummy model in the c-th category, respectively, and C is the number of categories in the dataset dictionary. AER represents a model's average improvement over the naive approach that one might use as a starting point for experimental optimization. In other words, AER is the percent of the gap closed between the naive model and a perfect predictor of accuracy 1.

2.3 Model construction

Both tree- and deep learning methods were explored for this MLC task (Figure 3), and their individual development is discussed below.

2.3.1 Gradient-boosting machines

GBMs are decision-tree-based learning algorithms that are popular in the ML literature for their performance in modeling numerical data.[40] We explored several string and descriptor-based encodings as numerical inputs (see SI) and found that a hybrid encoding scheme provided the greatest learnability (Figure 3A).[iv] The hybrid inputs are a concatenation of tokenized SMILES strings for each molecule in a reaction (coupling partners and products), further concatenated with molecular property vectors obtained from the Mordred descriptor calculator.[42] GBMs consistently outperformed other tree-based learners such as random forests (RFs),[43] perhaps owing to their use of sequential ensembling to improve in poor-performance regions.[40]

In our GBM experiments, a separate classifier was trained for all bins in a dataset dictionary, predicting whether or not they should be present in each reaction. Two general strategies have been developed for related MLC tasks, known as the binary relevance method (BM) and classifier chaining (CC).[44] The BM approach considers each classifier as an independent model, predicting the label of its bin irrespective of the others. Conversely, CCs make predictions sequentially, taking the output of each label as an additional input for the next one, where the optimal order of chaining is a learned parameter.[45] While the BM approach is significantly simpler from a computational perspective, CCs offer the potential for higher accuracy by modeling interdependencies between labels.[44]

We saw modeling reagent correlations as prudent in our studies since they are frequently observed in synthesis. Some examples relevant to this work include using a polar protic solvent with an inorganic base, excluding exogenous ligand when using a pre-ligated metal source, setting the temperature below the boiling point of the solvent, etc. We decided to explore both methods, testing BM against a modern update to CCs introduced by Read and coworkers known as classifier trellises (CTs).[46] In the CT method, instead of fully sequential propagation, models are fit in a pre-defined grid structure (the "trellis"), where the output of each prediction is passed to multiple downstream classifiers at once (Figure 3A, center). This eliminates the cost of chain structure discovery, while still benefiting from nesting predictions.[44]

The ordering of a CT is enforced algorithmically starting from a seed label, chosen randomly or by expert intervention. From Read et al.,[46] the trellis is populated by maximizing the mutual information (MI) between source and target labels (s_ℓ) at each step (ℓ) as follows:

$$s_\ell = \text{argmax}_{k \in S} \sum_{j \in \mathsf{pa}(\ell)} I(y_j; y_k) \ , \qquad (3)$$

where S and $\mathsf{pa}(\ell)$ are the set of remaining labels and the available trellis structure at the current step, respectively, and y_j and y_k are the j-th and k-th target labels, respectively. Here, $I(y_j; y_k)$ represents the MI between labels j and k based on their co-occurrences in the dataset. The matrix of *all* pairwise label dependencies $I(Y_j; Y_k)$ is constructed as below:

$$I(Y_j; Y_k) = \sum_{y_j \in \mathcal{Y}_j} \sum_{y_k \in \mathcal{Y}_k} p(y_j, y_k) \log \left(\frac{p(y_j, y_k)}{p(y_j)p(y_k)} \right) , \qquad (4)$$

where $p(y_j, y_k)$, and $p(y_j)$ and $p(y_k)$ are the joint and marginal probability mass functions of y_j and y_k, respectively. $\mathcal{Y}_j$ and $\mathcal{Y}_k$ represent the possible values y_j and y_k can each assume, which for our task of binary classification are both $\{0,1\}$. Full MI matrices and optimized trellises for each dataset are included in the SI, and an example is discussed with the results.

[iv] Gradient boosting was implemented using Microsoft's LightGBM.[41]

2.3.2 Relational graph convolutional networks

Originally reported by Schlichtkrull et al.,[34] R-GCNs are a subclass of message passing neural networks (MPNNs)[47] that explicitly model relational data such as molecular graphs. This is achieved by constructing sets of *relation* operations, where each relation $r \in \mathcal{R}$ is specific to a type and direction of edge between connected nodes. In our setting, the relations operate on atom-bond-atom triples using a learned, sparse weight matrix $\mathbf{W}_r^{(l)}$ in each layer l.[34] In a propagation step, each current node representation $h_i^{(l)}$ is transformed with all relation-specific neighboring nodes $h_j^{(l)}$ and summed over all relations such that:

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}h_j^{(l)} + \mathbf{W}_0^{(l)}h_i^{(l)}\right), \tag{5}$$

where $\mathcal{N}_i^r$ is the set of applicable neighbors and σ is an element-wise non-linearity, for us the tanh. The self-relation term $\mathbf{W}_0^{(l)}h_i^{(l)}$ is added to preserve local node information, and $c_{i,r}$ is a normalization constant.[34] Unlike traditional GCNs, R-GCNs intuitively model edge-based messages in local sub-graph transformations.[34] This is potentially very powerful for reaction learning in that information on edge types (i.e., single, double, triple, aromatic, and cyclic bonds) is crucial for modeling reactivity.

Here, we extend the R-GCN architecture with an additional graph attention layer (GAL) at the final readout step inspired by graph attention networks (GATs) from Veličković[48] and Busbridge.[49] As described by Veličković et al.,[48] GALs compute pair-wise node attention coefficients α_{ij} for each node h_i in a graph and its neighbors h_j. Two nodes' features are first transformed *via* a shared weight matrix $\mathbf{W}$, the results of which are concatenated before applying a learned weight vector and softmax normalization. The final update rule is simply a linear combination of α_{ij} with the newly transformed node vectors ($\mathbf{W}h_j$), summed over all neighboring nodes and averaged over a set of parallel attention mechanisms.[48]

In our recent studies,[32] we observed that existing relational GATs (R-GATs)[49] using atom-level attention layers were less effective for our task than simple R-GCNs.[v] Inspired nonetheless by the chemical intuition of graph attention, we adapted existing GALs to construct a *reaction-level* attention mechanism. Instead of pair-wise α_{ij}, we construct self-attention coefficients α_i^m for all nodes h_i^m in a molecular graph $\boldsymbol{h}^m = \{h_0^m, h_1^m, ..., h_L^m\}$. As in GATs, we take a linear combination of α_i^m for all L nodes in $\boldsymbol{h}^m$ after further transformation by matrix $\mathbf{W}^g$:

$$\alpha_i^m = \sigma\left(\mathbf{W}^s h_i^m\right), \forall\, i \in \{1, 2, ..., L\}, \tag{6}$$

$$h_i^a = \alpha_i^m \mathbf{W}^g h_i^m\ , \tag{7}$$

where $\mathbf{W}^s$ is the learned attention weight matrix, σ is the sigmoid activation function, and h_i^a is the updated node representation. The convolved graphs $\boldsymbol{h}^a = \{h_0^a, h_1^a, ..., h_L^a\}$ for each molecule m are then concatenated on the node feature axis to give an overall reaction representation $\boldsymbol{h}^r$ that we term the attended reaction graph (ARG):

$$\text{ARG} = \boldsymbol{h}^r = \left[\mathop{\Big\|}_{m=1}^{M} \boldsymbol{h}_{m^a}\right], \tag{8}$$

where M is the number of molecules in the reaction (reactants and products) and $\|$ denotes concatenation. Similar to the attention mechanism above, reaction-level attention coefficients α_i^r are then constructed and linearly combined with the ARG nodes h_i^r after transformation with $\mathbf{W}^v$. The final readout vector $\boldsymbol{v}_r$ is obtained from the attention layer by summative pooling over the nodes:

$$\alpha_i^r = \sigma\left(\mathbf{W}^r h_i^r\right), \forall\, i \in \{1, 2, ..., H\}, \tag{9}$$

$$\boldsymbol{v}_r = \sum_{i=1}^{H} \alpha_i^r \mathbf{W}^v h_i^r\ , \tag{10}$$

where H is the total number of nodes and $\mathbf{W}^r$ is the reaction attention weight matrix. This con-

[v] We found it necessary to reduce the hidden dimension of R-GATs to avoid excessive memory requirements relative to other GCNs,[48] and thus do not make a direct comparison of their performance.

Table 2: Prediction accuracy for all model types on the Suzuki dataset.

dataset	top-k	category	dummy	BM-GBM	CT-GBM	R-GCN	AR-GCN
Suzuki	top-1	**AER**	-	–0.0263[a]	–0.0554[b]	0.2767	**0.3115**
		metal	0.3777	0.5732	0.5629	0.6306	**0.6499**
		ligand	0.8722	0.8390	0.8408	0.9036	**0.9081**
		base	0.3361	0.4908	0.4777	0.5455	**0.5896**
		solvent	0.6377	0.6729	0.6751	0.7049	**0.7217**
		additive	0.9511	0.9259	0.9196	**0.9624**	0.9621
	top-3	**AER**	-	0.4088	0.3774	0.4936	**0.5246**
		metal	0.6744	0.8516	0.8475	0.8482	**0.8597**
		ligand	0.9269	0.9635	0.9606	0.9644	**0.9676**
		base	0.7344	**0.8338**	0.8250	0.8123	0.8285
		solvent	0.8013	0.8637	0.8577	0.8836	**0.8897**
		additive	0.9771	0.9842	0.9832	**0.9934**	0.9931

[a] AER excluding *additive*: 0.0962. [b] AER excluding *additive*: 0.0922.

struction differs from standard R-GCNs, which output readout vectors for individual molecules and concatenate them to form the ultimate reaction representation. Altogether, we term our hybrid architecture as an *attended relational graph convolutional network*, or AR-GCN.

In all deep learning experiments, with or without attention, the reaction vector readouts were passed to a multi-layer perceptron (MLP) of depth = 2.[vi] The final prediction is made as a single output vector with one entry for each label in the reaction dictionary, and the result is translated as described in Section 2.2.

3 Results and discussion

3.1 Model performance

Our modeling pipeline was first tested on the Suzuki coupling dataset, the largest of the four. Table 2 summarizes top-1 and top-3 categorical accuracies (Equation 1) and AERs (Equation 2) for the following models: GBMs with no trellising (**BM-GBM**), GBMs with trellising (**CT-GBM**), standard R-GCNs as reported by Schlichtkrull et al. (**R-GCN**),[32,34] our AR-GCNs developed here (**AR-GCN**), and the dummy predictor as a baseline (**dummy**).

For this dataset, GCN models significantly outperformed GBMs across categories for both top-1 and top-3 predictions. While GBMs actually gave negative top-1 AERs over baseline, these scores were dominated by the *additive* contribution; excluding this category the BM- and CT-GBMs gave modest 10% and 9% AERs, respectively. Despite struggling with top-1 predictions, GBMs gave significant AERs for top-3, with BM-GBMs at 41% and CT-GBMs at 38%. The AR-GCNs gave the best accuracy of all models, providing 31% and 52% top-1 and top-3 AERs, respectively. AR-GCNs gave roughly 3% AER gain over the R-GCN in both top-1 and top-3 predictions, demonstrating the value of the added attention layer.

A few interesting categorical trends can be seen across model types. For instance, models provide the best error reduction (ER = $\frac{A_c^g - A_c^d}{1 - A_c^d}$, see Equation 2) in the *metal* category, with the AR-GCN at 44% and 57% for top-1 and top-3, respectively. Similarly, models perform well in the *base* category, where the AR-GCN gave the best top-1 ER and BM-GBMs gave the best top-3 ER. Less consistent ERs between top-1 and top-3 predictions were obtained for the remaining three categories. For example, with *solvent*s, the AR-GCN improved baseline by 23% in top-1 predictions, but 44% in top-3. Likewise, for AR-GCN *ligand* predictions, a 28% ER was obtained for top-1 versus a 56% gain

[vi] All NN models were implemented using the Chainer Chemistry (ChainerChem) deep learning library.[50]

Table 3: Prediction accuracy for all model types on the C–N, Negishi, and PKR datasets.

dataset	top-k	category	dummy	BM-GBM	CT-GBM	R-GCN	AR-GCN
C–N	top-1	**AER**	-	–0.0413[a]	–0.1593[b]	0.3453	**0.3604**
		metal	0.2452	0.4825	0.4582	0.5989	**0.6162**
		ligand	0.5219	0.5538	0.5710	0.6981	**0.7068**
		base	0.2479	0.5028	0.5003	0.5932	**0.6066**
		solvent	0.3219	0.4582	0.4524	0.5647	**0.5674**
		additive	0.8904	0.7669	0.7031	0.8984	**0.8997**
	top-3	**AER**	-	0.3568	0.3131	0.5391	**0.5471**
		metal	0.6526	0.7928	0.7772	0.8479	**0.8490**
		ligand	0.6647	0.7933	0.7928	0.8605	**0.8688**
		base	0.6400	0.8008	0.7916	**0.8452**	0.8370
		solvent	0.5677	0.7370	0.7281	0.7973	**0.7997**
		additive	0.9156	0.9290	0.9184	0.9534	**0.9559**
Negishi	top-1	**AER**	-	0.3510	0.2773	0.4439	**0.4565**
		metal	0.2887	0.5444	0.5218	0.6555	**0.6730**
		ligand	0.7879	0.8174	0.7900	0.8724	**0.8772**
		temperature	0.3317	**0.6656**	0.6527	0.6188	0.6507
		solvent	0.6938	0.8562	0.8514	0.8868	**0.8915**
		additive	0.8309	0.8691	0.8401	**0.8724**	0.8644
	top-3	**AER**	-	0.5947	0.5199	0.6590	**0.6833**
		metal	0.5008	0.7771	0.7674	0.8086	**0.8517**
		ligand	0.8549	0.9548	0.9321	0.9522	**0.9553**
		temperature	0.5885	**0.9031**	0.8772	0.8517	0.8708
		solvent	0.8788	0.9321	0.9402	**0.9537**	**0.9537**
		additive	0.9043	0.9548	0.9354	**0.9761**	0.9729
PKR	top-1	**AER**	-	**0.4396**	0.4010	0.3973	0.4199
		metal	0.4302	**0.7901**	0.7786	0.7132	0.7057
		ligand	0.8792	**0.9351**	0.9237	0.9057	0.9094
		temperature	0.2830	0.5954	0.5649	0.6528	**0.6642**
		solvent	0.3321	0.6183	0.6260	0.6792	**0.6981**
		activator	0.6906	0.8244	0.8015	0.8415	**0.8491**
		CO (g)	0.7245	0.8855	0.8855	0.8717	**0.8868**
		additive	0.9057	**0.9008**	0.8893	0.8906	0.8491
		pressure	0.6528	0.8588	**0.8702**	0.8491	0.8491
	top-3	**AER**[c]	-	0.6987	0.6740	0.6844	**0.7145**
		metal	0.7132	**0.9351**	0.9313	0.9057	0.8906
		ligand	0.9019	**0.9962**	0.9924	0.9849	**0.9962**
		temperature	0.5962	**0.8740**	0.8321	0.8528	0.8604
		solvent	0.5925	0.8779	0.8550	0.8679	**0.8981**
		activator	0.8830	0.9466	0.9275	**0.9774**	**0.9774**
		CO (g)	1.0000	1.0000	1.0000	1.0000	1.0000
		additive	0.9321	**0.9885**	**0.9885**	0.9698	0.9736
		pressure	0.9623	0.9771	0.9847	**0.9849**	**0.9849**

[a] AER excluding *additive*: 0.2302. [b] AER excluding *additive*: 0.2282. [c] Excludes *CO(g)*.

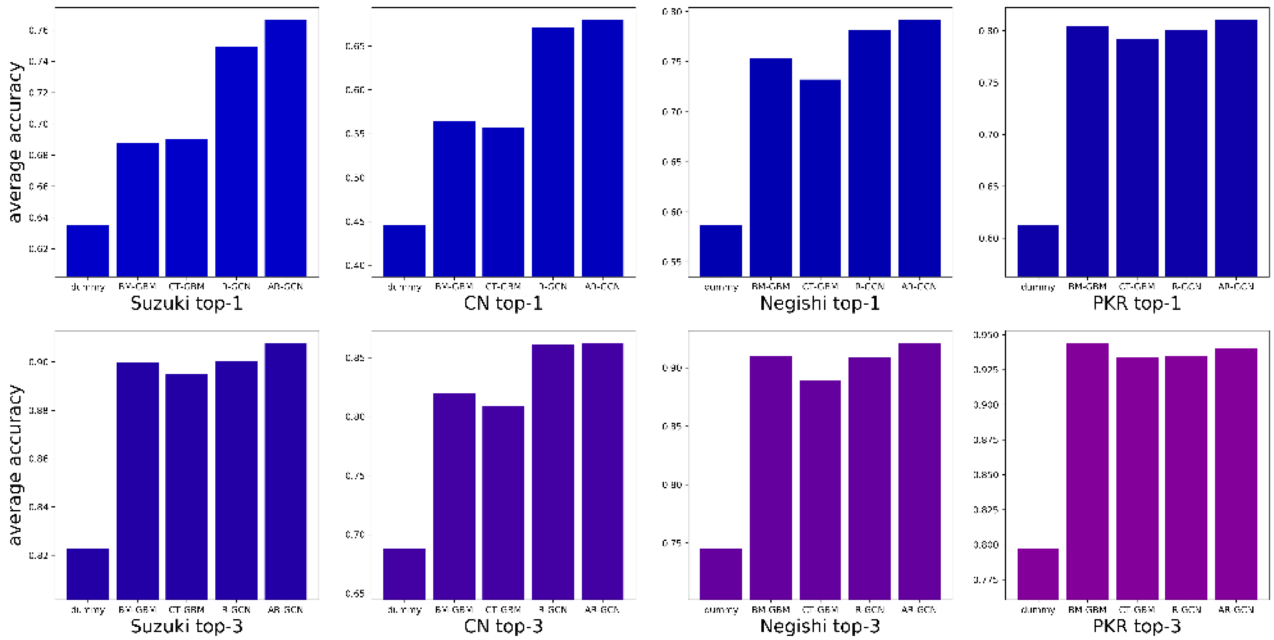


Figure 4: Average top-1 and top-3 categorical accuracies for each model across the four datasets.

in top-3. Finally, although the baseline *additive* accuracy is high as the majority of reactions are `null` in this category, the AR-GCN still gave a 23% top-1 ER and a 70% top-3 ER.

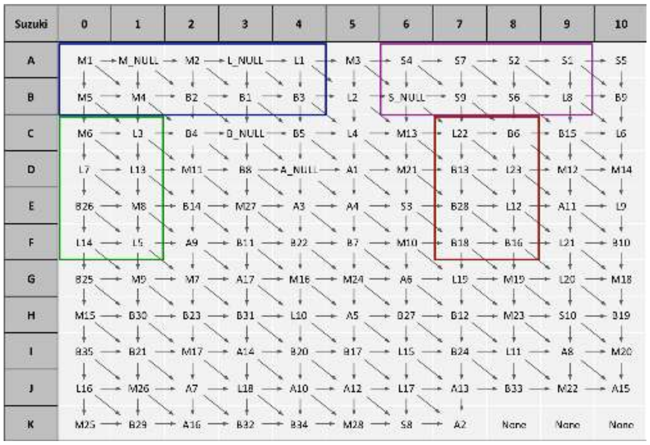
The trends and differences between top-1 and top-3 performance gains are reflective of the frequency distributions in each label category.[32] These intuitively resemble long-tail or Pareto-type distributions,[51] with the bulk of the cumulative density contained in a small number of bins and the remaining bins supporting smaller frequencies. The distribution shapes are likely to influence the relative top-1 and top-3 AERs, where the highly skewed distributions could be more difficult to improve over baseline.

Having demonstrated the utility of our predictive framework, we turned to the remaining datasets to assess its scope. Modeling results for C–N, Negishi, and PKRs are detailed in Table 3 and Figure 4. Notable observations for each dataset are discussed below.

C–N coupling. Similar to the Suzuki results, the AR-GCN was the top performer for C–N couplings in almost all categories, and slightly higher AERs were observed overall. The AR-GCN afforded 36% and 55% top-1 and top-3 AERs, respectively, again providing slight gains over R-GCNs at 35% and 54%. As above, GBMs struggled with this relatively large dataset (36,519 reactions) due to difficulties with the *additive* category. Models again made strong improvements in the *metal* and *base* categories, but also gave consistently strong gains for *ligand*s and *solvent*s, especially for top-3 predictions. For example, the AR-GCN returned top-3 ERs of 57% for *metal*s, 61% for *ligand*s, 55% for *base*s, and 54% for *solvent*s. Note that these ERs correspond to very high accuracies (A_c) of 85%, 87%, 84%, and 80%, respectively.

Negishi coupling. The highest AERs of all modeling experiments came with the Negishi dataset. The AR-GCN again gave the strongest performance, with top-1 and top-3 AERs of 46% and 68%, respectively. However, the R-GCN and even GBM models gave the highest accuracies in some categories. Interestingly, BM- and CT-GBMs performed significantly better than the GCNs for *temperature* predictions, though the strongest ER for most models came from the *solvent* category.

PKR. For the PKR dataset—the smallest of the four—simple BM-GBMs gave the best top-1 AER at 44%, followed closely by the AR-GCN at 42%. Similarly for top-3 predictions, these models gave AERs of 70% and 71%, respectively. Compared to the other reactions, GCNs are perhaps more prone to overfitting this small of a dataset,[52] making tree-based modeling more



Suzuki	0	1	2	3	4	5	6	7	8	9	10
A	M1	M_NULL	M2	L_NULL	L1	M3	S4	S7	S2	S1	S5
B	M5	M4	B2	B1	B3	L2	S_NULL	S9	S6	L8	B9
C	M6	L3	B4	B_NULL	B5	L4	M13	L22	B6	B15	L6
D	L7	L13	M11	B8	A_NULL	A1	M21	B13	L23	M12	M14
E	B26	M8	B14	M27	A3	A4	S3	B28	L12	A11	L9
F	L14	L5	A9	B11	B22	B7	M10	B18	B16	L21	B10
G	B25	M9	M7	A17	M16	M24	A6	L19	M19	L20	M18
H	M15	B30	B23	B31	L10	A5	B27	B12	M23	S10	B19
I	B35	B21	M17	A14	B20	B17	L15	B24	L11	A8	M20
J	L16	M26	A7	L18	A10	A12	L17	A13	B33	M22	A15
K	M25	B29	A16	B32	B34	M28	S8	A2	None	None	None

Figure 5: Optimized prediction trellis for the Suzuki dataset.

suitable. It is interesting to note that in general for PKRs, the GCN models were better at predicting physical parameters like *temperature*, *solvent*, and *CO(g)* atmosphere, whereas GBMs gave better performance for reaction components such as *metal*, *ligand*, and *additive*.

3.2 Interpretability

3.2.1 Tree methods

Given the results described above, we sought an understanding of the chemical features informing our predictions. Tree-based learning is often favored in this regard in that feature importances (FIs) can be directly extracted from models. We found that FIs for our GBMs were roughly uniform across the SMILES regions of the encodings. The most informative physical descriptors from the Mordred vectors pertained to two classes: topological charge distributions[53] correlated with local molecular dipoles; and Moreau–Broto autocorrelations[54] weighted by polarizability, ionization potential, and valence electrons (see SI for detailed rankings). The latter class is particularly intriguing as they are calculated from molecular graphs in what have been described as atom-pair convolutions,[55] not unlike the GCN models used here.[34]

An advantage to using CTs is the ability to extract their MI matrices and trellis structures for interpretation.[46] The optimized trellis for the Suzuki CT-GBMs is included in Figure 5, where several chemically intuitive features and category blocks can be noted:

1. Block A0–B4 (blue): The result of `M1` ($Pd(PPh_3)_4$) is used to predict three more metals: `M2` ($Pd(OAc)_2$), `M4` ($Pd(dppf)Cl_2 \cdot$ DCM), and `M5` ($Pd(PPh_3)_2Cl_2$). Based on these metal complexes, the probability of using exogenous ligand (L_NULL) and `L1` (PPh_3) is then predicted.

2. Block C0–F2 (green): The use of unligated `M6` ($Pd_2(dba)_3$) informs the predictions of ligands `L3` (XPhos), `L7` ($[(t\text{-Bu})_3PH]BF_4$), and `L13` (^{Me}CgPPh). These in turn feed the model of unligated `M8` ($Pd(dba)_2$), which then informs `L5` (P(*o*-tolyl)$_3$).

3. Block A6–B9 (purple): Several solvents are connected, where the predictions of `S4` (1,4-dioxane) and `S7` (PhMe) propagate through `S9` (H_2O), `S2` (EtOH), and `S6` (MeCN). These additionally feed classifiers of `S1` (THF) and `S_NULL` (neat).

4. Block C7–F8 (red): Four different classes of base are interwoven, including `B6` (CsF) and `B13` (KO*t*-Bu). This informs the prediction of `B28` ($LiOH \cdot H_2O$), which then goes on to feed models of `B18` (DIPEA) and `B16` (NaO*t*-Bu).

As a control experiment,[vii] we withheld the propagated predictions from the CT-GBMs to test whether the MI was actually being used.[56] Indeed, model accuracy dropped off markedly, even below baseline in some categories. While this suggests that CT-GBMs do learn reagent correlations, the sharp performance loss may also indicate overfitting to this information.[46] Further studies are necessary to uncover the optimal molecule featurization in combination with CTs, though the results here suggest their promise in modeling structured reaction data.

3.2.2 Deep learning methods

For AR-GCNs, a valuable interpretability feature lies in the learned feature weights α_i^r (Equation 9). Intuitively, the weights represent the

[vii] Detailed adversarial control studies for all GBM models are included in the SI.[56]

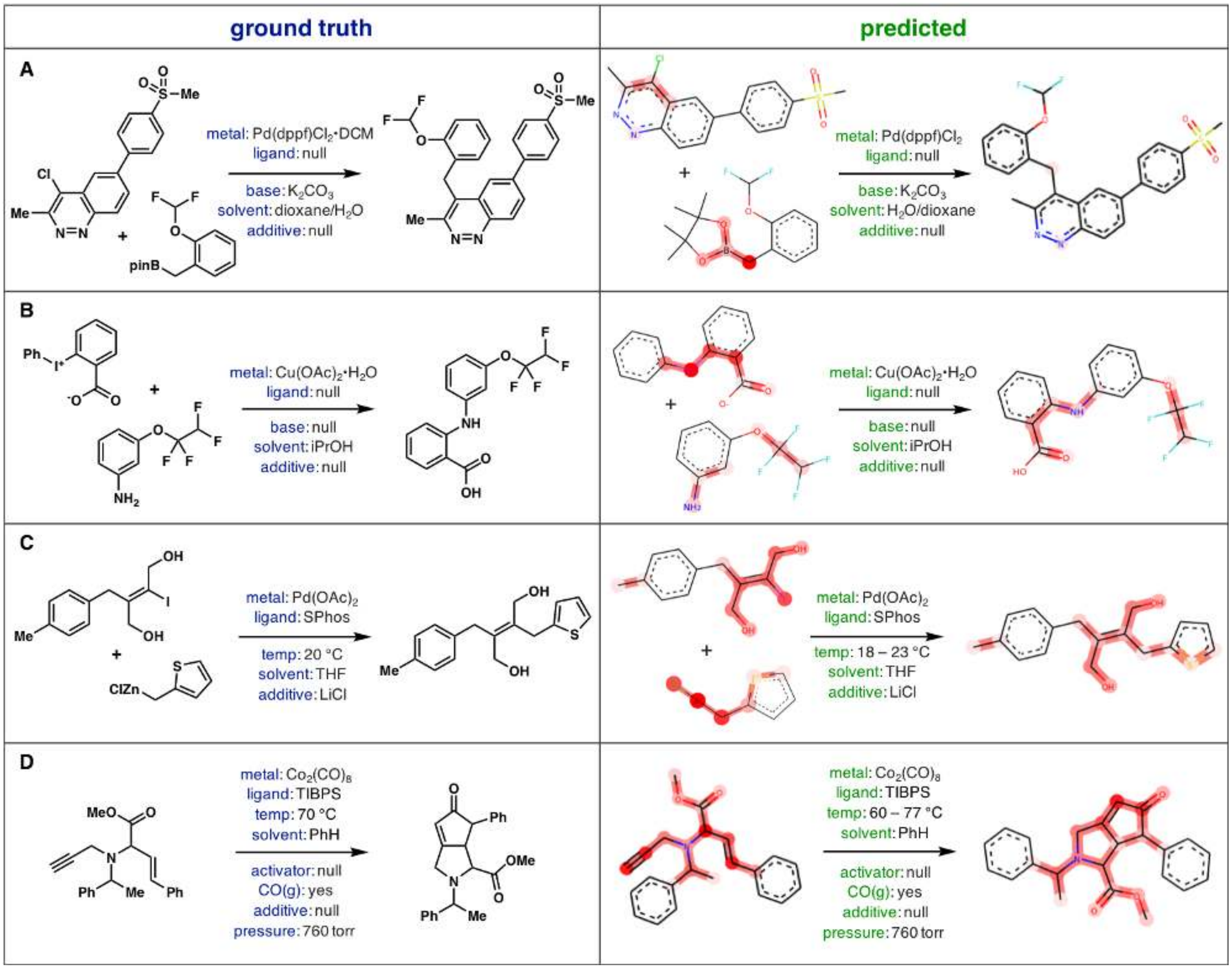


Figure 6: AR-GCN attention weight visualization and prediction examples from randomly chosen reactions in each dataset. Darker highlighting indicates higher attention.

model's assignment of importance on an atom, as they re-scale node features in the final graph layer before inference. When extracted, the weights can be mapped back onto a molecule's atoms and displayed by color scale using RDKit (Figure 1D).[57] This gives a visual interpretation of the functional groups most heavily informing the predictions. Example visualizations from a random reaction in each dataset and their AR-GCN predictions are included in Figure 6, and several additional random examples for each reaction type can be found in the SI.

In the Suzuki example (Figure 6A), the attention is dominated by the sp^3 carbon bearing the Bpin group, with additional contributions from the bis-*o*-substituted heteroaryl-chloride and its cinnoline nitrogen, all of which could be reasonably expected to influence reactivity. It is interesting that weights on the *o*-difluoromethoxy group, the sulfone, and the majority of the product are suppressed, perhaps indicating that an alkyl nucleophile is sufficient to predict the required conditions. The AR-GCN predictions are correct in each category besides the *metal*, where the model erroneously identifies the metal source Pd(dppf)Cl_2 instead of its ground truth DCM adduct Pd(dppf)$Cl_2 \cdot$ DCM.

Conversely, the weights in the C–N coupling example are more evenly distributed (Figure 6B). Intuitively, the chemically active iodonium benzoate is given strong attention in the electrophile, as is the nucleophilic aniline nitrogen. Here, the *m*-tetrafluoroethoxy group is also weighted significantly and these groups are given similar attention in the product. All categories are predicted correctly in this example, though three of them are `null`.

The Negishi example (Figure 6C) is an interesting C(sp^3)–C(sp^2) coupling of a fully substituted alkenyl-iodide and thiophenyl-methylzinc chloride. Like with A, the strongest weights correspond to the sp^3 nucleophilic carbon, though similarly strong attention is distributed over the electrophilic alkene including the pendant alcohols. These weights are again reflected in the product and all five condition categories are predicted correctly, including *temperature* and use of a LiCl *additive*.

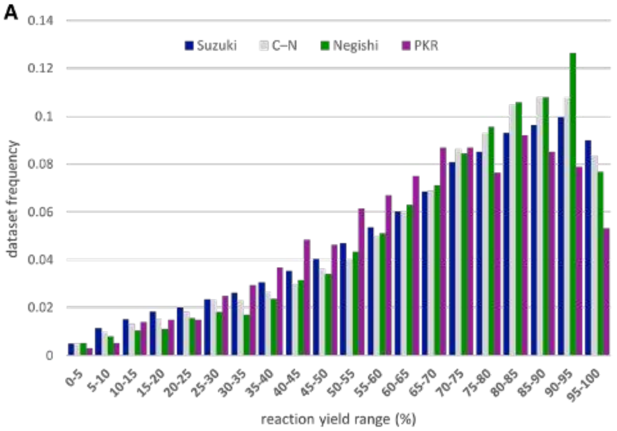
Lastly, an intramolecular PKR (Figure 6D) showed the most uniformly distributed attention of the four examples. Still, the strongest weights are given to the participating alkyne and alkene, with additional emphasis on the amino ester bridging group. Weights are similarly distributed in the product, though strongest attention is intuitively assigned to the newly formed enone. Here, all 8 categories are predicted correctly including the use of an ambient carbon monoxide atmosphere (*CO(g)* and *pressure*).

3.3 Yield Analysis

Having explored our models' chemical feature learning, we lastly investigated the effect of reaction yield, as it is a critical feature of synthesis data. Unsurprisingly, plotting the distribution of reaction yields in each dataset showed a uniformly strong bias towards high-yielding reactions (Figure 7A). Given the skewness of the data in this regard, we hypothesized that models would perform best at predicting conditions for high-yielding reactions.

We divided the dataset into quartiles by reaction yield and re-trained the AR-GCN with each sub-set, subsequently testing in each region and on the full test set (Figure 7B). Intuitively, models trained in any yield range tended to give highest accuracy when tested in the same range, occupying the confusion matrix diagonal in Figure 7B (top). To our surprise, however, the standard model trained on the full dataset gave consistently high accuracies, regardless of the test set (bottom row).

Since the yield bins contain varying amounts of data, we re-split the dataset, again ordered by yield but with equal sub-set sizes (Figure 7B bottom). A similar trend was observed where the highest accuracies were found on the diagonal and bottom row of the confusion matrix. Interestingly, the worst performing model was that trained in the highest yield range and tested in the lowest. We recognize that making "inaccurate" predictions on low-yielding reactions offers an avenue for predictive reaction optimization and future studies will explore this objective.



B

	yield range				
Train\Test	0-25	25-50	50-75	75-100	0-100
0-25	0.7275	0.6701	0.615	0.6072	0.6274
25-50	0.7217	0.7507	0.6781	0.6515	0.6795
50-75	0.6858	0.7309	0.7683	0.7094	0.7285
75-100	0.6552	0.6851	0.723	0.7649	0.7319
0-100	0.7481	0.773	0.7737	0.7586	0.7641

	data quartile				
Train\Test	0-25	25-50	50-75	75-100	0-100
0-25	0.7645	0.7059	0.6685	0.6713	0.7022
25-50	0.7057	0.7589	0.7084	0.6882	0.7153
50-75	0.6684	0.7138	0.7546	0.7224	0.7157
75-100	0.6417	0.6756	0.698	0.7559	0.693
0-100	0.7669	0.7729	0.7643	0.752	0.7641

Figure 7: Performance dependence on reaction yield. A) Distribution of reaction yields for the four datasets. B) AR-GCN average top-1 A_c values for Suzuki predictions when trained and tested in different yield ranges (top) and dataset quartiles arranged by yield (bottom).

4 Conclusion and Outlook

In summary, we present a multi-label classification approach to predicting experimental reaction conditions for organic synthesis. We successfully model four high-value reaction types using expert-crafted label dictionaries: Suzuki, C–N, and Negishi couplings, and Pauson–Khand reactions. We explore and optimize two model classes: gradient boosting machines and graph convolutional networks. We find that GCN models perform very well in larger datasets, while GBMs show success for smaller datasets.

We report the first use of classifier trellises in molecular machine learning, and find that they are able to incorporate label correlations in modeling. We introduce a novel reaction-level graph attention mechanism that provides significant accuracy gains when coupled with relational GCNs, and construct a hybrid GCN architecture called *attended relational GCNs*, or AR-GCNs. We further provide an analytical framework for the chemical interpretation of our models, extracting the trellis structures and mutual information matrices of the CT-GBMs, and visualizing the attention weights assigned in AR-GCN predictions.

Experimental studies are currently underway assessing the feasibility of model predictions on novel reactions. Additionally, efforts to apply our modeling framework to less-structured reaction types such as oxidations and reductions are ongoing. Future studies will address the interplay between structure representation and classifier chaining, as well as the extension of our reaction attention mechanism to other tasks. We expect the work herein to be very informative for future condition prediction studies, a highly valuable but underexplored learning task.

Acknowledgement We thank Prof Pietro Perona for mentorship guidance and helpful project discussions, and Chase Blagden for help structuring the GBM experiments. Fellowship support was provided by the NSF (M.R.M., T.J.D Grant No. DGE- 1144469). S.E.R. is a Heritage Medical Research Institute Investigator. Financial support from Research Corporation is warmly acknowledged.

Supporting Information Available

This will usually read something like: "Experimental procedures and characterization data for all new compounds. The class will automatically add a sentence pointing to the information on-line:

References

(1) Dreher, S. D. Catalysis in medicinal chemistry. *Reaction Chemistry & Engineering* **2019**, *4*, 1530–1535.

(2) Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; Wood, A. Organic synthesis provides opportunities to transform drug discovery. *Nature Chemistry* **2018**, *10*, 383–394.

(3) Mahatthananchai, J.; Dumas, A. M.; Bode, J. W. Catalytic Selective Synthesis. *Angewandte Chemie International Edition* **2012**, *51*, 10954–10990.

(4) Reid, J. P.; Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nature Reviews Chemistry* **2018**, *2*, 290–305.

(5) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

(6) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **2018**, *9*, 513–530.

(7) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **2019**, *59*, 3370–3388.

(8) Withnall, M.; Lindelöf, E.; Engkvist, O.; Chen, H. Building attention and edge message passing neural networks for bioactivity

and physical–chemical property prediction. *Journal of Cheminformatics* **2020**, *12*.

(9) Stokes, J. M. et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688–702.e13.

(10) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Molecular Informatics* **2018**, *37*, 1700123.

(11) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering* **2019**, *4*, 828–849.

(12) Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics* **2019**, *11*, 74.

(13) Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generating Customized Compound Libraries for Drug Discovery with Machine Intelligence. **2019**,

(14) Panteleev, J.; Gao, H.; Jia, L. Recent applications of machine learning in medicinal chemistry. *Bioorganic & Medicinal Chemistry Letters* **2018**, *28*, 2807–2815.

(15) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Scientific Reports* **2017**, *7*.

(16) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Central Science* **2017**, *3*, 434–443.

(17) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018**, *360*, 186–190.

(18) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *Journal of the American Chemical Society* **2018**, *140*, 5004–5008.

(19) Simón-Vidal, L.; García-Calvo, O.; Oteo, U.; Arrasate, S.; Lete, E.; Sotomayor, N.; González-Díaz, H. Perturbation-Theory and Machine Learning (PTML) Model for High-Throughput Screening of Parham Reactions: Experimental and Theoretical Studies. *Journal of Chemical Information and Modeling* **2018**, *58*, 1384–1396.

(20) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559*, 377–381.

(21) Hughes, T. B.; Miller, G. P.; Swamidass, S. J. Modeling Epoxidation of Druglike Molecules with a Deep Machine Learning Network. *ACS Central Science* **2015**, *1*, 168–180.

(22) Peng, Q.; Duarte, F.; Paton, R. S. Computing organic stereoselectivity – from concepts to quantitative calculations and predictions. *Chemical Society Reviews* **2016**, *45*, 6093–6107.

(23) Banerjee, S.; Sreenithya, A.; Sunoj, R. B. Machine learning for predicting product distributions in catalytic regioselective reactions. *Physical Chemistry Chemical Physics* **2018**, *20*, 18311–18318.

(24) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels–Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angewandte Chemie International Edition* **2019**, *58*, 4515–4519.

(25) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019**, *363*, eaau5631.

(26) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research* **2018**, *51*, 1281–1289.

(27) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.

(28) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *Journal of Chemical Information and Modeling* **2019**, *59*, 2529–2537.

(29) Badowski, T.; Gajewska, E. P.; Molga, K.; Grzybowski, B. A. Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angewandte Chemie International Edition* **2020**, *59*, 725–730.

(30) Nicolaou, C. A.; Watson, I. A.; LeMasters, M.; Masquelin, T.; Wang, J. Context Aware Data-Driven Retrosynthetic Analysis. *Journal of Chemical Information and Modeling* **2020**,

(31) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Central Science* **2018**, *4*, 1465–1476.

(32) Ryou*, S.; Maser*, M. R.; Cui*, A. Y.; DeLano, T. J.; Yue, Y.; Reisman, S. E. Graph Neural Networks for the Prediction of Substrate-Specific Organic Reaction Conditions. *arXiv:2007.04275 [cs, LG]* **2020**,

(33) Huerta, F.; Hallinder, S.; Minidis, A. *Machine Learning to Reduce Reaction Optimization Lead Time – Proof of Concept with Suzuki, Negishi and Buchwald-Hartwig Cross-Coupling Reactions*; preprint ChemRxiv.12613214, 2020.

(34) Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Berg, R. v. d.; Titov, I.; Welling, M. Modeling Relational Data with Graph Convolutional Networks. *arXiv:1703.06103 [cs, stat]* **2017**,

(35) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **2001**, *29*, 1189–1232.

(36) Reaxys. `https://new.reaxys.com/`, (accessed on May 13, 2019).

(37) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **1988**, *28*, 31–36.

(38) Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; Belongie, S. Class-Balanced Loss Based on Effective Number of Samples. *arXiv:1901.05555 [cs]* **2019**,

(39) Wu, X.-Z.; Zhou, Z.-H. A Unified View of Multi-Label Performance Measures. *arXiv:1609.00288 [cs]* **2017**,

(40) Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics* **2013**, *7*.

(41) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 3146–3154.

(42) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics* **2018**, *10*, 4.

(43) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(44) Zhang, M.-L.; Zhou, Z.-H. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering* **2014**, *26*, 1819–1837.

(45) Read, J.; Pfahringer, B.; Holmes, G.; Frank, E. Classifier Chains for Multi-label Classification. Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg, 2009; pp 254–269.

(46) Read, J.; Martino, L.; Olmos, P.; Luengo, D. Scalable Multi-Output Label Prediction: From Classifier Chains to Classifier Trellises. *Pattern Recognition* **2015**, *48*, 2096–2109.

(47) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *arXiv:1704.01212 [cs]* **2017**,

(48) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv:1710.10903 [cs, stat]* **2018**,

(49) Busbridge, D.; Sherburn, D.; Cavallo, P.; Hammerla, N. Y. Relational Graph Attention Networks. *arXiv:1904.05811 [cs, stat]* **2019**,

(50) Tokui, S.; Oono, K.; Hido, S.; Clayton, J. Chainer: a Next-Generation Open Source Framework for Deep Learning. **2015**,

(51) Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* **2005**, *46*, 323–351.

(52) Zhou, K.; Dong, Y.; Lee, W. S.; Hooi, B.; Xu, H.; Feng, J. Effective Training Strategies for Deep Graph Neural Networks. *arXiv:2006.07107 [cs, stat]* **2020**,

(53) Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *Journal of Chemical Information and Modeling* **1994**, *34*, 520–525.

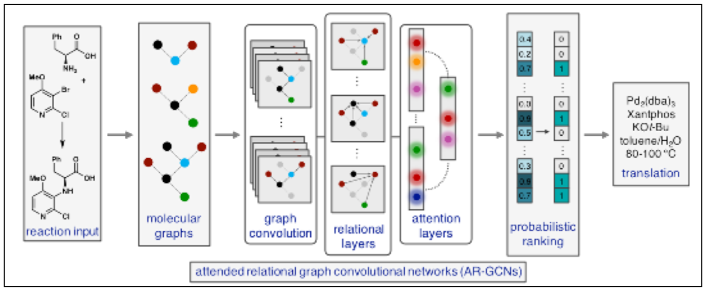
(54) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *New Journal of Chemistry* **1980**, *4*, 359–360.

(55) Hollas, B. An Analysis of the Autocorrelation Descriptor for Molecules. *Journal of Mathematical Chemistry* **2003**, *33*, 91–101.

(56) Chuang, K. V.; Keiser, M. J. Adversarial Controls for Scientific Machine Learning. *ACS Chemical Biology* **2018**, *13*, 2819–2821.

(57) Landrum, G. A. RDKit: Open-Source Cheminformatics Software. (accessed Nov 20, 2016).

Graphical TOC Entry



Supporting Information: Multi-Label Classification Models for the Prediction of Cross-Coupling Reaction Conditions

Michael R. Maser,[†,§] Alexander Y. Cui,[‡,§] Serim Ryou,[¶,§] Travis J. DeLano,[†] Yisong Yue,[‡] and Sarah E. Reisman[*,†]

†*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California, USA*

‡*Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA*

¶*Computational Vision Lab, California Institute of Technology, Pasadena, California, USA*

§*Equal contribution.*

E-mail: reisman@caltech.edu

S1 Data preparation and reaction dictionaries

Full procedures for data processing are outlined in our previous preprint.[S1] An example protocol with full code is included in the associated github repository: `https://github.com/slryou41/reaction-gcnn.git` in the path: `data/data_processing_example.ipynb`. The worked example includes procedures for sorting reagents into categories by reaction role and aggregating into a full reaction dictionary. Final dictionaries for all four datasets as .csv files can be found in the repository path: `data/all_dictionaries/`, and are tabulated below.

Table S1: Suzuki dataset dictionary.

category	bin label	dataset name	instances
metal	M1	tetrakis(triphenylphosphine) palladium(0)	55829
	M2	palladium diacetate	16927
	M3	(1,1'-bis(diphenylphosphino)ferrocene)palladium(II) dichloride	13723
	M4	dichloro(1,1'-bis(diphenylphosphanyl)ferrocene)palladium(II)*CH2Cl2	8918
	M5	bis-triphenylphosphine-palladium(II) chloride	8761
	M6	tris-(dibenzylideneacetone)dipalladium(0)	5241
	M7	palladium dichloride	1512
	M8	bis(dibenzylideneacetone)-palladium(0)	1013
	M9	dichloro[1,1'-bis(di-t-butylphosphino)ferrocene]palladium(II)	1074
	M10	bis(tri-t-butylphosphine)palladium(0)	736
	M11	chloro(2-dicyclohexylphosphino-2?,4?,6?-triisopropyl-1,1?-biphenyl)[2-(2?-amino-1,1?-biphenyl?)]palladium(II)	729
	M12	bis(di-tert-?butyl(4-?dimethylaminophenyl)?phosphine)?dichloropalladium(II)	711
	M13	bis(eta3-allyl-mu-chloropalladium(II))	559
	M14	tris(dibenzylideneacetone)dipalladium(0) chloroform complex	509
	M15	palladium 10% on activated carbon	861
	M16	sodium tetrachloropalladate(II)	283
	M17	palladium	280
	M18	(2-dicyclohexylphosphino-2?,4?,6?-triisopropyl-1,1?-biphenyl)[2-(2?-amino-1,1?-biphenyl)]palladium(II) methanesulfonate	191
	M19	bis(benzonitrile)palladium(II) dichloride	179
	M20	(1,2-dimethoxyethane)dichloronickel(II)	158
	M21	bis(1,5-cyclooctadiene)nickel (0)	155
	M22	[1,3-bis(2,6-diisopropylphenyl)imidazol-2-ylidene](3-chloropyridyl)palladium(ll) dichloride	151
	M23	(bis(tricyclohexyl)phosphine)palladium(II) dichloride	148
	M24	dichloro bis(acetonitrile) palladium(II)	143
	M25	Pd EnCat-30TM	137
	M26	nickel(II) nitrate hexahydrate	106
	M27	palladium(II) trifluoroacetate	106

Continued on next page

Table S1 – continued from previous page

category	bin label	dataset name	instances
	M28	dichlorobis[1-(dicyclohexylphosphanyl)piperidine]palladium(II)	102
ligand	L1	triphenylphosphine	4489
	L2	dicyclohexyl-(2',6'-dimethoxybiphenyl-2-yl)-phosphane	3163
	L3	XPhos	2100
	L4	tricyclohexylphosphine	1808
	L5	tris-(o-tolyl)phosphine	902
	L6	tri-tert-butyl phosphine	694
	L7	tri tert-butylphosphoniumtetrafluoroborate	616
	L8	trisodium tris(3-sulfophenyl)phosphine	556
	L9	1,1'-bis-(diphenylphosphino)ferrocene	486
	L10	4,5-bis(diphenylphos4,5-bis(diphenylphosphino)-9,9-dimethylxanthenephino)-9,9-dimethylxanthene	424
	L11	CyJohnPhos	370
	L12	ruphos	293
	L13	1,3,5,7-tetramethyl-8-phenyl-2,4,6-trioxa-8-phosphatricyclo[3.3.1.13,7]decane	279
	L14	tricyclohexylphosphine tetrafluoroborate	240
	L15	johnphos	223
	L16	4,4'-di-tert-butyl-2,2'-bipyridine	216
	L17	catacxium A	192
	L18	trifuran-2-yl-phosphane	183
	L19	triphenyl-arsane	182
	L20	1,1'-bis(di-tertbutylphosphino)ferrocene	142
	L21	2,2'-bis-(diphenylphosphino)-1,1'-binaphthyl	129
	L22	Tedicyp	218
	L23	bis[2-(diphenylphosphino)phenyl] ether	108
	B1	potassium carbonate	48981
	B2	sodium carbonate	39769
	B3	potassium phosphate	17799
	B4	caesium carbonate	13345
	B5	sodium hydrogencarbonate	3722
	B6	cesium fluoride	2810
	B7	sodium hydroxide	2156
	B8	potassium hydroxide	2155
	B9	potassium fluoride	2097
	B10	triethylamine	1370
	B11	potassium phosphate tribasic trihydrate	1016
	B12	potassium acetate	931
	B13	potassium tert-butylate	912

Continued on next page

Table S1 – continued from previous page

category	bin label	dataset name	instances
base	B14	potassium phosphate monohydrate	826
	B15	sodium acetate	418
	B16	sodium t-butanolate	392
	B17	barium dihydroxide	374
	B18	N-ethyl-N,N-diisopropylamine	336
	B19	lithium hydroxide	321
	B20	potassium phosphate tribasic heptahydrate	317
	B21	diisopropylamine	209
	B22	sodium methylate	175
	B23	tetrabutyl ammonium fluoride	173
	B24	barium hydroxide octahydrate	171
	B25	potassium dihydrogenphosphate	166
	B26	potassium fluoride dihydrate	156
	B27	1,4-diaza-bicyclo[2.2.2]octane	154
	B28	lithium hydroxide monohydrate	143
	B29	tetra-butylammonium acetate	137
	B30	sodium phosphate	133
	B31	potassium hydrogencarbonate	131
	B32	dipotassium hydrogenphosphate	127
	B33	tripotassium phosphate n hydrate	123
	B34	cesiumhydroxide monohydrate	112
	B35	sodium phosphate dodecahydrate	103
solvent	S1	tetrahydrofuran	18113
	S2	ethanol	24836
	S3	methanol	4374
	S4	1,4-dioxane	39107
	S5	1,2-dimethoxyethane	19131
	S6	acetonitrile	4366
	S7	toluene	28304
	S8	N,N-dimethyl formamide	15110
	S9	water	92175
	S10	1-methyl-pyrrolidin-2-one	472
additive	A1	tetrabutylammomium bromide	3003
	A2	water	1606
	A3	lithium chloride	819
	A4	hydrogenchloride	780
	A5	copper(l) iodide	546
	A6	silver(l) oxide	405
	A7	copper diacetate	183
	A8	dmap	181
	A9	Aliquat 336	169

Continued on next page

Table S1 – continued from previous page

category	bin label	dataset name	instances
	A10	cetyltrimethylammonim bromide	167
	A11	copper(l) chloride	164
	A12	potassium bromide	157
	A13	trifluoroacetic acid	151
	A14	oxygen	148
	A15	air	112
	A16	18-crown-6 ether	127
	A17	sodium dodecyl-sulfate	113

Table S2: C–N dataset dictionary.

category	bin label	dataset name	instances
	M1	copper(l) iodide	8180
	M2	tris-(dibenzylideneacetone)dipalladium(0)	6995
	M3	palladium diacetate	4668
	M4	copper	1875
	M5	bis(dibenzylideneacetone)-palladium(0)	1292
	M6	copper(I) oxide	932
	M7	copper(II) oxide	402
	M8	copper(l) chloride	386
	M9	copper(I) bromide	348
	M10	bis(eta3-allyl-mu-chloropalladium(II))	433
	M11	copper(II) acetate monohydrate	352
	M12	(1,1'-bis(diphenylphosphino)ferrocene)palladium(II) dichloride	159
	M13	bis(tri-t-butylphosphine)palladium(0)	181
	M14	iron(III) chloride	116
	M15	copper(II) bis(trifluoromethanesulfonate)	91
	M16	copper(ll) bromide	88
	M17	bis-triphenylphosphine-palladium(II) chloride	82
	M18	copper(II) sulfate	154
	M19	bis(acetylacetonate)nickel(II)	78
	M20	palladium 10% on activated carbon	71
	M21	tetrakis(triphenylphosphine) palladium(0)	68
	M22	dichlorobis(tri-O-tolylphosphine)palladium	67
	M23	(1,2-dimethoxyethane)dichloronickel(II)	66
	M24	palladium dichloride	63
	M25	copper(I) thiophene-2-carboxylate	58
	M26	cobalt(II) oxalate dihydrate	56

Continued on next page

Table S2 – continued from previous page

category	bin label	dataset name	instances
metal	M27	copper dichloride	52
	M28	dichloro(1,3-bis(2,6-bis(3-pentyl)phenyl)imidazolin-2-ylidene)(3-chloropyridyl)palladium(II)	49
	M29	chloro[2-(dicyclohexylphosphino)-3 ,6-dimethoxy-2?,4?, 6?-triisopropyl- 1,1?-biphenyl] [2-(2-aminoethyl)phenyl]palladium(II)	97
	M30	[2-(di-tert-butylphosphino)-2?,4?,6?-triisopropyl-1,1?-biphenyl][2-((2-aminoethyl)phenyl)]palladium(II) chloride	49
	M31	iron(III) oxide	48
	M32	C36H45Cl2N3OPd	46
	M33	nickel(II) bromide trihydrate	45
	M34	copper acetylacetonate	45
	M35	C36H43Cl2N3Pd	45
	M36	C30H43O2P*C13H12N(1-)*CH3O3S(1-)*Pd(2+)	45
	M37	bis(1,5-cyclooctadiene)nickel (0)	45
	M38	CuPy2Cl2	42
	M39	dichloro(3-chloropyridinyl)(1,3-(diisopropylphenyl)-4,5-bis(dimethylamino)imidazol-2-ylidene)palladium(II)	41
	M40	Al2O3*Cu(2+)	40
	M41	C33H40ClN3O2Pd	38
	M42	dichloro(1,1'-bis(diphenylphosphanyl)ferrocene)palladium(II)*CH2Cl2	36
	M43	(1,3-bis(2,6-diisopropylphenyl)-3,4,5,6-tetrahydropyrimidin-2-ylidene)Pd(cinnamyl, 3-phenylallyl)Cl	36
	M44	copper(II)iodide	35
	L1	2,2'-bis-(diphenylphosphino)-1,1'-binaphthyl	3014
	L2	tri-tert-butyl phosphine	2137
	L3	4,5-bis(diphenylphos4,5-bis(diphenylphosphino)-9,9-dimethylxanthenephino)-9,9-dimethylxanthene	1995
	L4	N,N‘-dimethylethylenediamine	1543
	L5	XPhos	830
	L6	1,10-Phenanthroline	703
	L7	L-proline	620
	L8	1,1'-bis-(diphenylphosphino)ferrocene	653
	L9	johnphos	444
	L10	DavePhos	374

Continued on next page

Table S2 – continued from previous page

category	bin label	dataset name	instances
ligand	L11	triphenylphosphine	275
	L12	ruphos	266
	L13	tri tert-butylphosphoniumtetrafluoroborate	265
	L14	tert-butyl XPhos	242
	L15	dicyclohexyl-(2',6'-dimethoxybiphenyl-2-yl)-phosphane	261
	L16	trans-1,2-Diaminocyclohexane	724
	L17	8-quinolinol	206
	L18	CyJohnPhos	192
	L19	trans-N,N'-dimethylcyclohexane-1,2-diamine	535
	L20	ethylenediamine	175
	L21	dimethylaminoacetic acid	167
	L22	dicyclohexyl[3,6-dimethoxy-2?,4?,6?-tris(1-methylethyl)[1,1?-biphenyl]-2-yl]phosphine	165
	L23	2,2,6,6-tetramethylheptane-3,5-dione	163
	L24	1,1'-bi-2-naphthol	162
	L25	bis[2-(diphenylphosphino)phenyl] ether	170
	L26	1-dicyclohexylphosphino-2-di-tert-butylphosphinoethylferrocene	142
	L27	P(i-BuNCH2)3CMe	110
	L28	di-tert-butyl2?-isopropoxy-[1,1?-binaphthalen]-2-ylphosphane	108
	L29	di-tert-butyl(2,2-diphenyl-1-methyl-1-cyclopropyl)phosphine	104
	L30	P(i-BuNCH2CH2)3N	98
	L31	N,N-dimethylglycine hydrochoride	96
	L32	N-[2-(di(1-adamantyl)phosphino)phenyl]morpholine	92
	L33	5-(di-tert-butylphosphino)-1?, 3?, 5?-triphenyl-1?H-[1,4?]bipyrazole	91
	L34	2-[2-(dicyclohexylphosphino)-phenyl]-1-methyl-1H-indole	86
	L35	4,4'-di-tert-butyl-2,2'-bipyridine	85
	L36	tris-(o-tolyl)phosphine	77
	L37	2,8,9-tris(2-methylpropyl)-2,5,8,9-tetraaza-1-phosphabicyclo[3.3.3]undecane	75
	L38	cis-N,N'-dimethyl-1,2-diaminocyclohexane	74
	L39	monophosphine 1,2,3,4,5-pentaphenyl-1'-(di-tert-butylphosphino)ferrocene	55
	L40	5-(di(adamantan-1-yl)phosphino)-1?,3?,5?-triphenyl-1?H-1,4?-bipyrazole	55
	L41	t-BuBrettPhos	53

Continued on next page

Table S2 – continued from previous page

category	bin label	dataset name	instances
	L42	2-(N,N-dimethylamino)athanol	53
	L43	tricyclohexylphosphine	46
	L44	(E)-3-(dimethylamino)-1-(2-hydroxyphenyl)prop-2-en-1-one	46
	L45	di-tert-butylneopentylphosphonium tetrafluoroborate	38
	L46	2-di-tertbutylphosphino-3,4,5,6-tetramethyl-2',4',6'-triisopropyl-1,1'-biphenyl	37
	L47	N,N,N,N,-tetramethylethylenediamine	26
base	B1	sodium t-butanolate	9103
	B2	potassium carbonate	7129
	B3	caesium carbonate	6957
	B4	potassium phosphate	3274
	B5	potassium tert-butylate	2167
	B6	potassium hydroxide	1420
	B7	triethylamine	500
	B8	lithium hexamethyldisilazane	432
	B9	sodium hydroxide	430
	B10	sodium hydride	228
	B11	sodium carbonate	200
	B12	potassium phosphate monohydrate	130
	B13	sodium hydrogencarbonate	128
solvent	S1	toluene	11970
	S2	1,4-dioxane	5273
	S3	N,N-dimethyl-formamide	4246
	S4	dimethyl sulfoxide	3790
	S5	water	2464
	S6	tetrahydrofuran	1457
	S7	1,2-dimethoxyethane	878
	S8	tert-butyl alcohol	841
	S9	acetonitrile	780
	S10	ethanol	549
	S11	5,5-dimethyl-1,3-cyclohexadiene	497
	S12	isopropyl alcohol	316
	S13	nitrobenzene	315
	S14	1-methyl-pyrrolidin-2-one	292
	S15	hexane	286
	S16	N,N-dimethyl acetamide	281
	S17	1,2-dichloro-benzene	254
	S18	neat (no solvent)	240
	S19	o-xylene	219

Continued on next page

Table S2 – continued from previous page

category	bin label	dataset name	instances
	S20	xylene	208
	S21	methanol	180
	S22	ethyl acetate	163
	A1	18-crown-6 ether	455
	A2	tetrabutylammomium bromide	372
	A3	8-quinolinol	206
	A4	dimethylaminoacetic acid	167
	A5	1,1'-bi-2-naphthol	162
	A6	water	160
	A7	sodium sulfate	132
	A8	2-(2-methyl-1-oxopropyl)cyclohexanone	121
	A9	phenylboronic acid	120
	A10	1,3-bis[(2,6-diisopropyl)phenyl]imidazolinium chloride	109
	A11	potassium iodide	108
	A12	hydrogenchloride	107
	A13	ethylene glycol	102
	A14	N,N-dimethylglycine hydrochoride	96
	A15	1,3-bis[2,6-diisopropylphenyl]imidazolium chloride	95
	A16	N-ethylmorpholine	93
	A17	tert-butyl alcohol	87
	A18	aluminum oxide	84
	A19	D-glucose	83
	A20	cetyltrimethylammonim bromide	71
	A21	1,3-dimethyl-3,4,5,6-tetrahydro-2(1H)-pyrimidinone	68
	A22	N',N'-diphenyl-1H-pyrrole-2-carbohydrazide	63
	A23	manganese(II) fluoride	63
	A24	dimethyl sulfoxide	55
	A25	2-(N,N-dimethylamino)athanol	53
	A26	air	48
	A27	iron(III) oxide	48
	A28	(E)-3-(dimethylamino)-1-(2-hydroxyphenyl)prop-2-en-1-one	46
	A29	lithium bromide	44
	A30	6,7-dihydro-5H-quinolin-8-one oxime	43
	A31	CVT-2537	42
	A32	ammonium chloride	42
	A33	1-methyl-pyrrolidin-2-one	42
	A34	tetra(n-butyl)ammonium hydroxide	40
	A35	salicylaldehyde-oxime	39
	A36	potassium fluoride on basic alumina	39

Continued on next page

Table S2 – continued from previous page

category	bin label	dataset name	instances
additive	A37	toluene-4-sulfonic acid	38
	A38	lithium chloride	38
	A39	pipecolic Acid	37
	A40	oxygen	37
	A41	metformin hydrochloride	37
	A42	8-Hydroxyquinoline-N-oxide	37
	A43	1-(5,6,7,8-tetrahydroquinolin-8-yl)ethan-1-one	36
	A44	tetrabutyl ammonium fluoride	36
	A45	N1,N2-bis(thiophen-2-ylmethyl)oxalamide	36
	A46	N-phenyl-2-pyridincarboxamide-1-oxide	35
	A47	N-((1-oxy-pyridin-2-yl)methyl)oxalamic acid	35
	A48	C19H19N5O	35
	A49	manganese(II) chloride tetrahydrate	34
	A50	1-tetralone oxime	32
	A51	N1,N2-bis(2,4,6-trimethoxyphenyl)oxalamide	31
	A52	N-methoxy-1H-pyrrole-2-carboxamide	29
	A53	ammonia	29
	A54	1,2,3-Benzotriazole	29
	A55	dimethylenecyclourethane	28
	A56	isopropylmagnesium chloride	27
	A57	N-(2-cyanophenyl)pyridine-2-carboxamide	27
	A58	C20H18N2O2	27
	A59	2-acetylcyclohexanone	27
	A60	2,6-di-tert-butyl-4-methyl-phenol	26
	A61	2-hydroxy-pyridine N-oxide	26
	A62	TPGS-750-M	25
	A63	N?-phenyl-1H-pyrrole-2-carbohydrazide	25
	A64	lanthanum(III) oxide	25
	A65	ethylmagnesium bromide	25
	A66	ethyl 2-oxocyclohexane carboxylate	25
	A67	1,4-dimethyl-1,2,3,4-tetrahydro-5H-benzo[e][1,4]diazepin-5-one	25
	A68	tetraethoxy orthosilicate	24
	A69	N,N,N',N'-tetramethylguanidine	24
	A70	C20H26N4O4	24
	A71	2-methyl-8-quinolinol	24
	A72	2-carbomethoxy-3-hydroxyquinoxaline-di-N-oxide	24
	A73	1,3-diisopropyl-1H-imidazol-3-ium chloride	24
	A74	MOF-199	24

Table S3: Negishi dataset dictionary.

category	bin label	dataset name	instances
metal	M1	tetrakis(triphenylphosphine) palladium(0)	1902
	M2	tris-(dibenzylideneacetone)dipalladium(0)	572
	M3	bis-triphenylphosphine-palladium(II) chloride	418
	M4	palladium diacetate	370
	M5	bis(dibenzylideneacetone)-palladium(0)	344
	M6	(1,1’-bis(diphenylphosphino)ferrocene)palladium(II) dichloride	334
	M7	bis(tri-t-butylphosphine)palladium(0)	273
	M8	dichloro(1,1’-bis(diphenylphosphanyl)ferrocene)palladium(II)*CH2Cl2	248
	M9	dichlorobis[1-(dicyclohexylphosphanyl)piperidine]palladium(II)	168
	M10	palladium(l) tri-tert-butylphosphine iodide dimer	101
	M11	bis(tricyclohexylphosphine)nickel(II) dichloride	99
	M12	[(C10H13-1,3-(CH2P(C6H11)2)2)Pd(Cl)]	87
	M13	1,3-bis[(diphenylphosphino)propane]dichloronickel(II)	63
	M14	bis(1,5-cyclooctadiene)nickel (0)	56
	M15	nickel dichloride	56
	M16	tris(dibenzylideneacetone)dipalladium(0) chloroform complex	46
	M17	dichlorobis(tri-O-tolylphosphine)palladium	46
	M18	palladium	44
	M19	[1,3-bis(2,6-diisopropylphenyl)imidazol-2-ylidene](3chloro-pyridyl)palladium(II) dichloride	136
	M20	C20H20ClN3Ni	42
	M21	dichloro(1,3-bis(2,6-bis(3-pentyl)phenyl)imidazolin-2-ylidene)(3-chloropyridyl)palladium(II)	39
	M22	bis(triphenylphosphine)nickel(II) chloride	38
	M23	C26H24ClN2NiP*0.1C7H8	35
	M24	cobalt(II) chloride	34
	M25	copper(I) bromide	31
	M26	C40H55Cl5N3Pd	30
	M27	[1,3-bis(2,6-diisoheptylphenyl)-4,5-dichloroimidazol-2-ylidene](3-chloropyridyl)palladium(II) dichloride	29
	M28	dichloro bis(acetonitrile) palladium(II)	29

Continued on next page

Table S3 – continued from previous page

category	bin label	dataset name	instances
	M29	palladium(II) trifluoroacetate	27
	M30	1,2-bis(diphenylphosphino)ethane nickel(II) chloride	27
	M31	C27H22Cl2N3NiP	24
	M32	C38H34Br2N4Ni2P2	23
ligand	L1	1,1’-bis-(diphenylphosphino)ferrocene	233
	L2	dicyclohexyl-(2’,6’-dimethoxybiphenyl-2-yl)-phosphane	196
	L3	XPhos	187
	L4	triphenylphosphine	161
	L5	trifuran-2-yl-phosphane	128
	L6	monophosphine 1,2,3,4,5-pentaphenyl-1’-(di-tert-butylphosphino)ferrocene	95
	L7	tris-(o-tolyl)phosphine	70
	L8	Ruphos	61
	L9	2?-(dicyclohexylphophanyl)-N2,N2,N6,N6-tetramethyl[1,1?-biphenyl]-2,6-diamine	37
	L10	tripiperidino-phosphine	37
	L11	tri tert-butylphosphoniumtetrafluoroborate	35
	L12	1,2-bis-(dicyclohexylphosphino)ethane	33
	L13	4,5-bis(diphenylphos4,5-bis(diphenylphosphino)-9,9-dimethylxanthenephino)-9,9-dimethylxanthene	31
	L14	N,N,N,N,-tetramethylethylenediamine	24
	L15	[2,2]bipyridinyl	22
	L16	4,4’-di-tert-butyl-2,2’-bipyridine	21
	L17	1,2-Ph2-3,4-bis(2,4,6-(t-Bu)3-phenylphophinidene)cyclobutene	20
	L18	johnphos	20
	L19	tri-tert-butyl phosphine	19
	L20	tricyclohexylphosphine	18
temperature	T1	-163 - 18	101
	T2	18 - 23	2313
	T3	23 - 50	643
	T4	50 - 61	975
	T5	61 - 80	658
	T6	80 - 100	673
	T7	100 - 120	696
	T8	120 - 220	479
	S1	tetrahydrofuran	4525
	S2	N,N-dimethyl-formamide	1003
	S3	1-methyl-pyrrolidin-2-one	674

Continued on next page

Table S3 – continued from previous page

category	bin label	dataset name	instances
solvent	S4	toluene	541
	S5	1,4-dioxane	335
	S6	N,N-dimethyl acetamide	247
	S7	hexane	219
	S8	diethyl ether	203
	S9	water	122
	S10	1,2-dimethoxyethane	67
additive	A1	lithium chloride	243
	A2	zinc	207
	A3	copper(l) iodide	154
	A4	water	62
	A5	diisobutylaluminium hydride	59
	A6	tetrabutylammomium bromide	52
	A7	ammonium chloride	51
	A8	n-butyllithium	46
	A9	1-Methylpyrrolidine	42
	A10	Li2CoCl4	42
	A11	sodium formate	42
	A12	hydrogenchloride	36
	A13	caesium carbonate	36
	A14	zinc diacetate	32
	A15	potassium carbonate	30
	A16	norborn-2-ene	30
	A17	lithium bromide	28
	A18	1,3-dimethyl-3,4,5,6-tetrahydro-2(1H)-pyrimidinone	23
	A19	methylzinc chloride	22
	A20	1-methyl-pyrrolidin-2-one	21
	A21	zinc(II) chloride	21
	A22	isoquinoline	20
	A23	sodium carbonate	19
	A24	1-ethyl-2-pyrrolidinone	18
	A25	sodium	16
	A26	1-methyl-1H-imidazole	15
	A27	oxovanadium(V) ethoxydichloride	12
	A28	2-(N,N-dimethylamino)athanol	11
	A29	[bdmim][BF4]	11
	A30	1-butyl-2-(diphenylphosphanyl)-3-methylimidazolium hexafluorophosphate	11

Table S4: PKR dataset dictionary.

category	**bin label**	**dataset name**	**instances**
metal	M1	dicobalt octacarbonyl	614
	M2	di(rhodium)tetracarbonyl dichloride	333
	M3	chloro(1,5-cyclooctadiene)rhodium(I) dimer	140
	M4	[RhCl(CO)dppp]2	92
	M5	cobalt(II) bromide	44
	M6	palladium dichloride	33
	M7	dodecacarbonyl-triangulo-triruthenium	32
	M8	Co2Rh2 nanoparticles immobilized on charcoal	50
	M9	tetracobaltdodecacarbonyl	44
	M10	molybdenum hexacarbonyl	23
	M11	Rh(dppp)2Cl	19
	M12	cobalt nanoparticles on charcoal	36
	M13	methylidynetricobalt nonacarbonyl	25
	M14	bis(triphenylphosphine)(carbonyl)rhodium chloride	11
	M15	PdCl(OHNCCH3C6H4)(C5H5N)	10
	M16	bis(1,5-cyclooctadiene)diiridium(I) dichloride	9
	M17	diiron nonacarbonyl	9
	M18	iron(II) bis(trimethylsilyl)amide	9
ligand	L1	1,1,3,3-tetramethyl-2-thiourea	128
	L2	1,3-bis-(diphenylphosphino)propane	93
	L3	2,2'-bis-(diphenylphosphino)-1,1'-binaphthyl	31
	L4	triphenylphosphine	16
	L5	tri-n-butylphosphine sulfide	15
	L6	(S)-3,5-di-tert-butyl-4-methoxyphenyl-(6,6?-dimethoxybiphenyl-2,2?-diyl)-bis(diphenylphosphine)	12
temperature	T1	-98 - 20	83
	T2	20	961
	T3	20 - 60	299
	T4	60 - 77	370
	T5	77 - 94	338
	T6	94 - 120	395
	T7	120 - 180	303
	S1	toluene	966
	S2	dichloromethane	601
	S3	tetrahydrofuran	318
	S4	1,2-dichloro-ethane	171
	S5	1,2-dimethoxyethane	145
	S6	acetonitrile	141
	S7	not listed	102

Continued on next page

Table S4 – continued from previous page

category	bin label	dataset name	instances
solvent	S8	water	71
	S9	benzene	76
	S10	para-xylene	136
	S11	hexane	43
	S12	dimethyl sulfoxide	39
	S13	1,4-dioxane	33
	S14	dibutyl ether	33
	S15	diethyl ether	22
activator	A1	4-methylmorpholine N-oxide	420
	A2	trimethylamine-N-oxide	212
	A3	dimethyl sulfoxide	137
	A4	cyclohexylamine	68
	A5	n-butyl methyl sulfide	27
	A6	silver trifluoromethanesulfonate	23
	A7	silver tetrafluoroborate	18
	A8	silver hexafluoroantimonate	19
	A9	(4-fluorobenzyl)(methyl)sulfide	14
	A10	dinitrogen monoxide	14
	A11	4-methylmorpholine 4-oxide monohydrate	13
CO (g)	G1	carbon monoxide	1169
	G2	none	1580
additive	O1	4 A molecular sieve	84
	O2	zinc	50
	O3	hydrogen	40
	O4	ethylene glycol	30
	O5	cetyltrimethylammonim bromide	22
	O6	Celite	17
	O7	Triton X(R)-100	37
	O8	acetic anhydride	15
	O9	lithium chloride	15
	O10	water	11
	O11	oxygen	10
	O12	potassium carbonate	8
	O13	triethylsilane	8
pressure	P1	37 - 760	35
	P2	760	2392
	P3	760 - 7600	169
	P4	7600 - 7500600	153

S2 Computational details and hyperparameters

S2.1 Gradient-boosting machines (GBMs)

Numerical inputs for GBM models were constructed by tokenizing SMILES strings for each molecule in a reaction with character–to–number mappings, and calculating chemical descriptor vectors using Mordred.[S2] Code examples for these processing protocols are provided in the associated github repository at the path `data/gbm{_}inputs/parsing-cols.ipynb`. All GBM classifiers were implemented using Microsoft's lightGBM.[S3] Specific non-default parameter settings are included in Table S5.

Table S5: Computational details and general parameters used for GBM models.

parameter	value	description
train/valid/test	81/9/10	data splitting[a]
max_depth	7	maximum tree depth for base learners
tree_method	'gpu_hist'	split continuous features into discrete bins
eval_metric	'aucpr'	evaluation metric

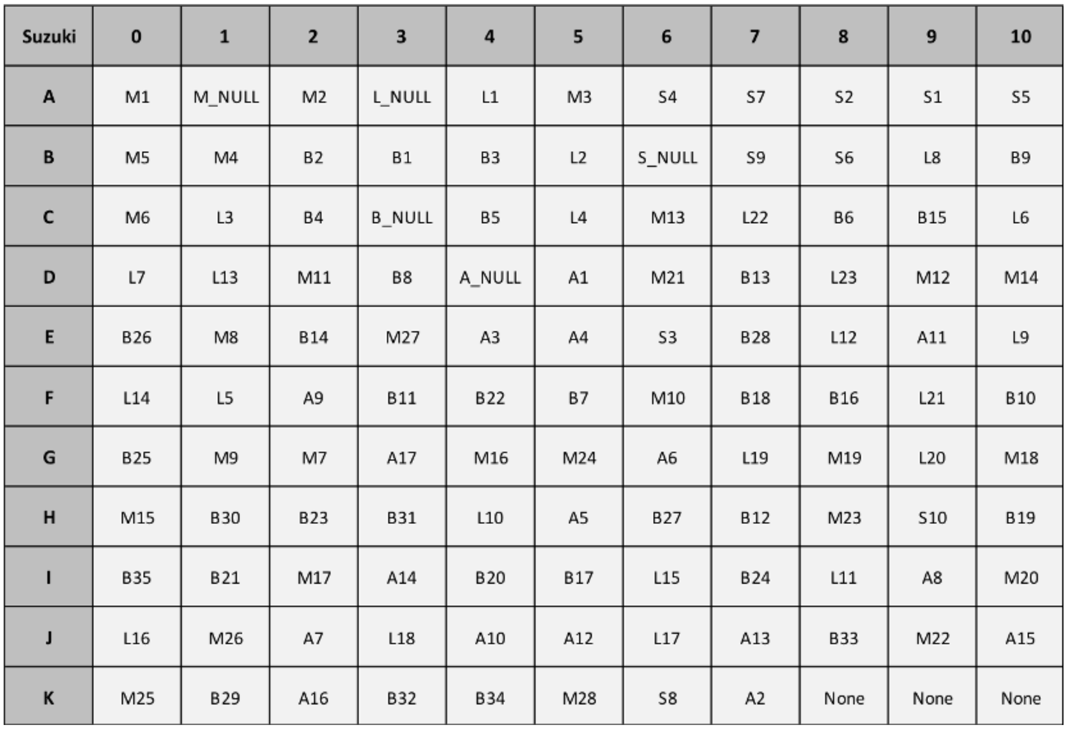
[a] Training, validation, and test sets were identical to those in GCNs.

S2.1.1 Binary relevance method (BM)

In BM experiments, an independent `lightgbm.LGBMClassifier` was fit for each label bin in a dataset's dictionary using the full input representation.

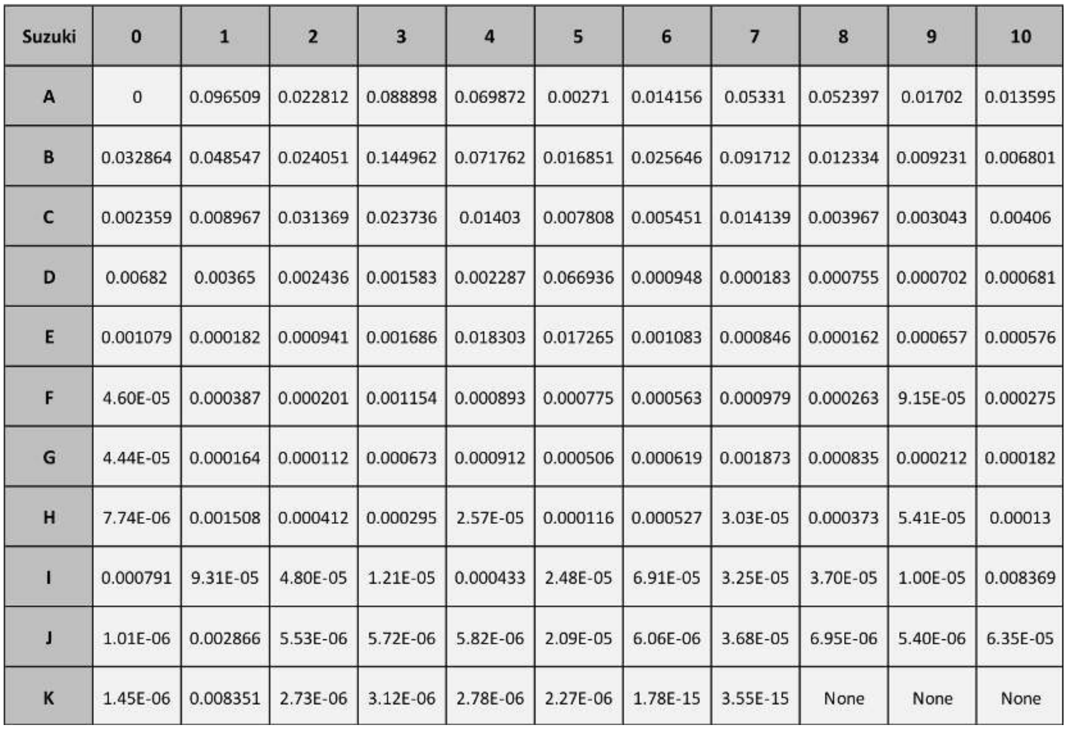
S2.1.2 Classifier trellises (CTs)

In CT experiments, `lightgbm.LGBMClassifier`s were fit for each label bin in a dataset's dictionary as part of a grid structure in which predictions are made sequentially and are passed to downstream models as additional inputs (see main text for explanation). Mutual information (MI) matrices were constructed for each dataset's label dictionary using scikit learn's `sklearn.metrics.mutual_info_score` module.[S4] Classifier trellises were then constructed following the algorithm reported by Read et al. (see main text and associated code for details).[S5] As shown in the example in the main text, each model takes additional input from the bins in directions `north`, `west`, and `northwest` of it. Models on the edges of the trellis take input only from those bins in the available directions (i.e., propagation does not wrap between rows). Here each trellis was initialized using the label `M1`, the most commonly used metal in each dataset. This can be chosen by user preference, expert intuition, or at random. Full MI matrices and trellis structures for all four datasets are provided below.



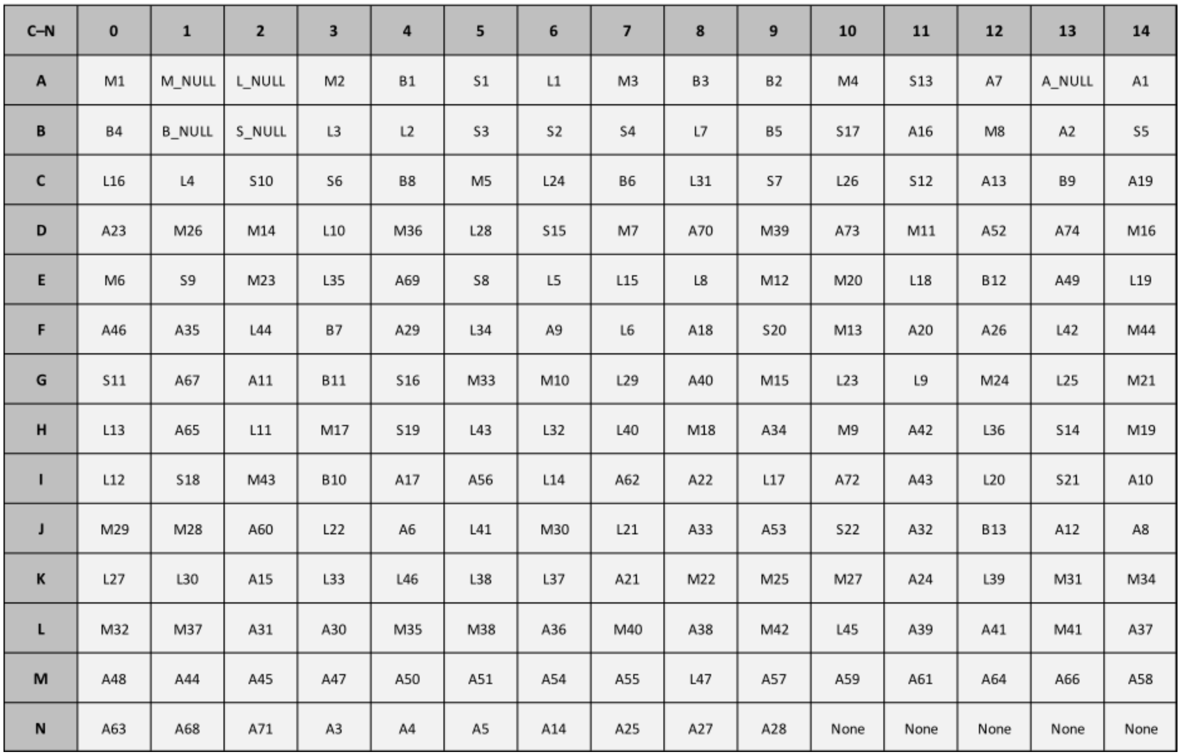
Suzuki	0	1	2	3	4	5	6	7	8	9	10
A	M1	M_NULL	M2	L_NULL	L1	M3	S4	S7	S2	S1	S5
B	M5	M4	B2	B1	B3	L2	S_NULL	S9	S6	L8	B9
C	M6	L3	B4	B_NULL	B5	L4	M13	L22	B6	B15	L6
D	L7	L13	M11	B8	A_NULL	A1	M21	B13	L23	M12	M14
E	B26	M8	B14	M27	A3	A4	S3	B28	L12	A11	L9
F	L14	L5	A9	B11	B22	B7	M10	B18	B16	L21	B10
G	B25	M9	M7	A17	M16	M24	A6	L19	M19	L20	M18
H	M15	B30	B23	B31	L10	A5	B27	B12	M23	S10	B19
I	B35	B21	M17	A14	B20	B17	L15	B24	L11	A8	M20
J	L16	M26	A7	L18	A10	A12	L17	A13	B33	M22	A15
K	M25	B29	A16	B32	B34	M28	S8	A2	None	None	None

Figure S1: Optimized classifier trellis for the Suzuki dataset.



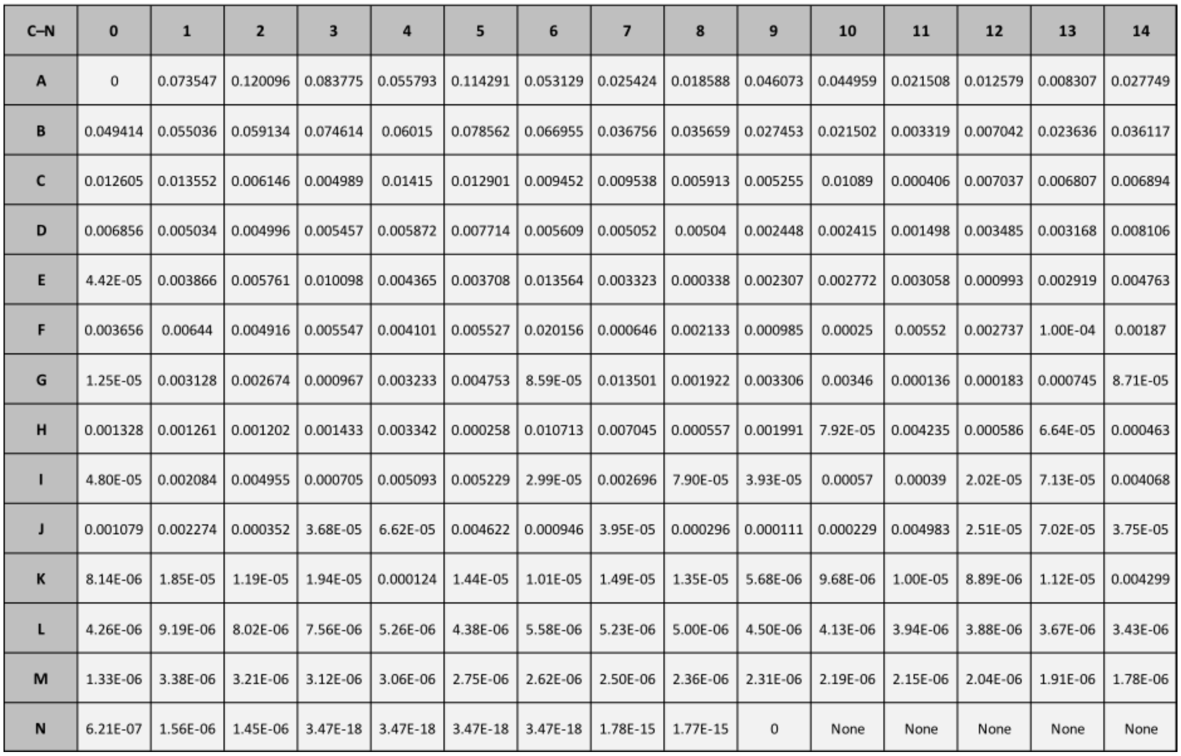
Suzuki	0	1	2	3	4	5	6	7	8	9	10
A	0	0.096509	0.022812	0.088898	0.069872	0.00271	0.014156	0.05331	0.052397	0.01702	0.013595
B	0.032864	0.048547	0.024051	0.144962	0.071762	0.016851	0.025646	0.091712	0.012334	0.009231	0.006801
C	0.002359	0.008967	0.031369	0.023736	0.01403	0.007808	0.005451	0.014139	0.003967	0.003043	0.00406
D	0.00682	0.00365	0.002436	0.001583	0.002287	0.066936	0.000948	0.000183	0.000755	0.000702	0.000681
E	0.001079	0.000182	0.000941	0.001686	0.018303	0.017265	0.001083	0.000846	0.000162	0.000657	0.000576
F	4.60E-05	0.000387	0.000201	0.001154	0.000893	0.000775	0.000563	0.000979	0.000263	9.15E-05	0.000275
G	4.44E-05	0.000164	0.000112	0.000673	0.000912	0.000506	0.000619	0.001873	0.000835	0.000212	0.000182
H	7.74E-06	0.001508	0.000412	0.000295	2.57E-05	0.000116	0.000527	3.03E-05	0.000373	5.41E-05	0.00013
I	0.000791	9.31E-05	4.80E-05	1.21E-05	0.000433	2.48E-05	6.91E-05	3.25E-05	3.70E-05	1.00E-05	0.008369
J	1.01E-06	0.002866	5.53E-06	5.72E-06	5.82E-06	2.09E-05	6.06E-06	3.68E-05	6.95E-06	5.40E-06	6.35E-05
K	1.45E-06	0.008351	2.73E-06	3.12E-06	2.78E-06	2.27E-06	1.78E-15	3.55E-15	None	None	None

Figure S2: Mutual information matrix for the Suzuki dataset.



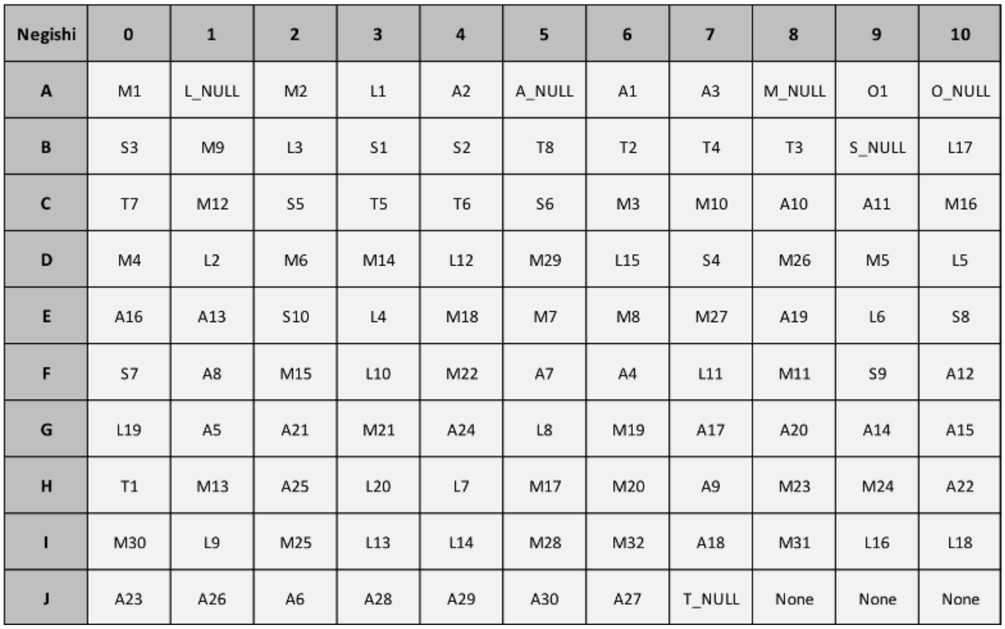
C–N	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	M1	M_NULL	L_NULL	M2	B1	S1	L1	M3	B3	B2	M4	S13	A7	A_NULL	A1
B	B4	B_NULL	S_NULL	L3	L2	S3	S2	S4	L7	B5	S17	A16	M8	A2	S5
C	L16	L4	S10	S6	B8	M5	L24	B6	L31	S7	L26	S12	A13	B9	A19
D	A23	M26	M14	L10	M36	L28	S15	M7	A70	M39	A73	M11	A52	A74	M16
E	M6	S9	M23	L35	A69	S8	L5	L15	L8	M12	M20	L18	B12	A49	L19
F	A46	A35	L44	B7	A29	L34	A9	L6	A18	S20	M13	A20	A26	L42	M44
G	S11	A67	A11	B11	S16	M33	M10	L29	A40	M15	L23	L9	M24	L25	M21
H	L13	A65	L11	M17	S19	L43	L32	L40	M18	A34	M9	A42	L36	S14	M19
I	L12	S18	M43	B10	A17	A56	L14	A62	A22	L17	A72	A43	L20	S21	A10
J	M29	M28	A60	L22	A6	L41	M30	L21	A33	A53	S22	A32	B13	A12	A8
K	L27	L30	A15	L33	L46	L38	L37	A21	M22	M25	M27	A24	L39	M31	M34
L	M32	M37	A31	A30	M35	M38	A36	M40	A38	M42	L45	A39	A41	M41	A37
M	A48	A44	A45	A47	A50	A51	A54	A55	L47	A57	A59	A61	A64	A66	A58
N	A63	A68	A71	A3	A4	A5	A14	A25	A27	A28	None	None	None	None	None

Figure S3: Optimized classifier trellis for the C–N dataset.



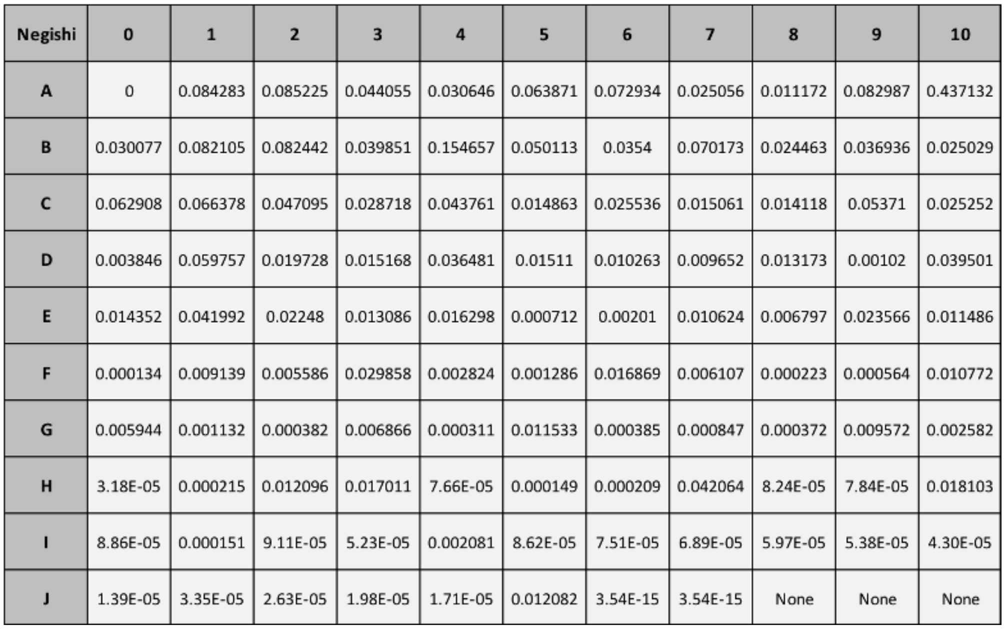
C–N	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	0.073547	0.120096	0.083775	0.055793	0.114291	0.053129	0.025424	0.018588	0.046073	0.044959	0.021508	0.012579	0.008307	0.027749
B	0.049414	0.055036	0.059134	0.074614	0.06015	0.078562	0.066955	0.036756	0.035659	0.027453	0.021502	0.003319	0.007042	0.023636	0.036117
C	0.012605	0.013552	0.006146	0.004989	0.01415	0.012901	0.009452	0.009538	0.005913	0.005255	0.01089	0.000406	0.007037	0.006807	0.006894
D	0.006856	0.005034	0.004996	0.005457	0.005872	0.007714	0.005609	0.005052	0.00504	0.002448	0.002415	0.001498	0.003485	0.003168	0.008106
E	4.42E-05	0.003866	0.005761	0.010098	0.004365	0.003708	0.013564	0.003323	0.000338	0.002307	0.002772	0.003058	0.000993	0.002919	0.004763
F	0.003656	0.00644	0.004916	0.005547	0.004101	0.005527	0.020156	0.000646	0.002133	0.000985	0.00025	0.00552	0.002737	1.00E-04	0.00187
G	1.25E-05	0.003128	0.002674	0.000967	0.003233	0.004753	8.59E-05	0.013501	0.001922	0.003306	0.00346	0.000136	0.000183	0.000745	8.71E-05
H	0.001328	0.001261	0.001202	0.001433	0.003342	0.000258	0.010713	0.007045	0.000557	0.001991	7.92E-05	0.004235	0.000586	6.64E-05	0.000463
I	4.80E-05	0.002084	0.004955	0.000705	0.005093	0.005229	2.99E-05	0.002696	7.90E-05	3.93E-05	0.00057	0.00039	2.02E-05	7.13E-05	0.004068
J	0.001079	0.002274	0.000352	3.68E-05	6.62E-05	0.004622	0.000946	3.95E-05	0.000296	0.000111	0.000229	0.004983	2.51E-05	7.02E-05	3.75E-05
K	8.14E-06	1.85E-05	1.19E-05	1.94E-05	0.000124	1.44E-05	1.01E-05	1.49E-05	1.35E-05	5.68E-06	9.68E-06	1.00E-05	8.89E-06	1.12E-05	0.004299
L	4.26E-06	9.19E-06	8.02E-06	7.56E-06	5.26E-06	4.38E-06	5.58E-06	5.23E-06	5.00E-06	4.50E-06	4.13E-06	3.94E-06	3.88E-06	3.67E-06	3.43E-06
M	1.33E-06	3.38E-06	3.21E-06	3.12E-06	3.06E-06	2.75E-06	2.62E-06	2.50E-06	2.36E-06	2.31E-06	2.19E-06	2.15E-06	2.04E-06	1.91E-06	1.78E-06
N	6.21E-07	1.56E-06	1.45E-06	3.47E-18	3.47E-18	3.47E-18	3.47E-18	1.78E-15	1.77E-15	0	None	None	None	None	None

Figure S4: Mutual information matrix for the C–N dataset.



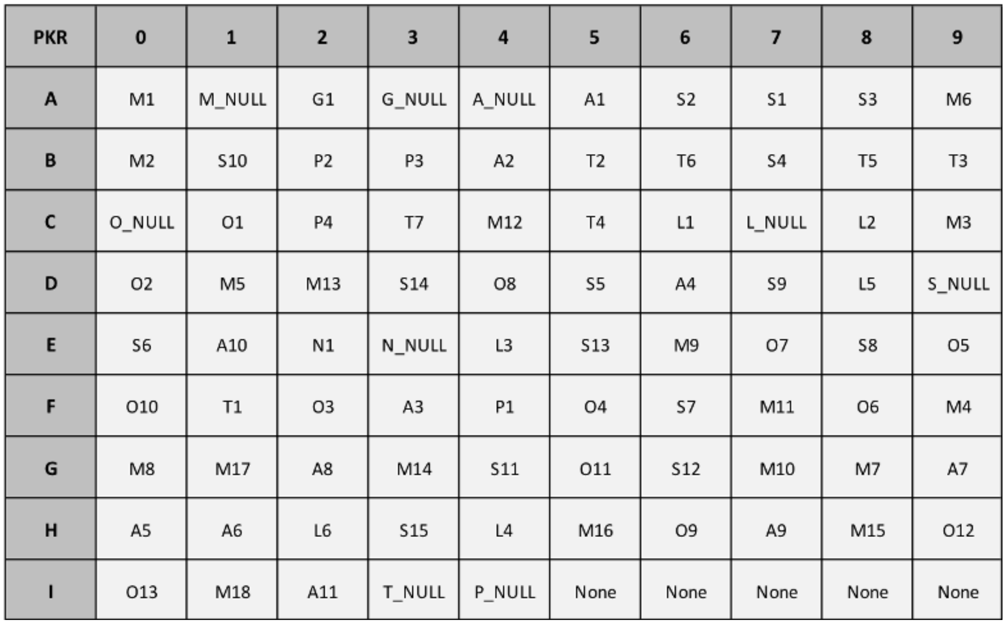
Negishi	0	1	2	3	4	5	6	7	8	9	10
A	M1	L_NULL	M2	L1	A2	A_NULL	A1	A3	M_NULL	O1	O_NULL
B	S3	M9	L3	S1	S2	T8	T2	T4	T3	S_NULL	L17
C	T7	M12	S5	T5	T6	S6	M3	M10	A10	A11	M16
D	M4	L2	M6	M14	L12	M29	L15	S4	M26	M5	L5
E	A16	A13	S10	L4	M18	M7	M8	M27	A19	L6	S8
F	S7	A8	M15	L10	M22	A7	A4	L11	M11	S9	A12
G	L19	A5	A21	M21	A24	L8	M19	A17	A20	A14	A15
H	T1	M13	A25	L20	L7	M17	M20	A9	M23	M24	A22
I	M30	L9	M25	L13	L14	M28	M32	A18	M31	L16	L18
J	A23	A26	A6	A28	A29	A30	A27	T_NULL	None	None	None

Figure S5: Optimized classifier trellis for the Negishi dataset.



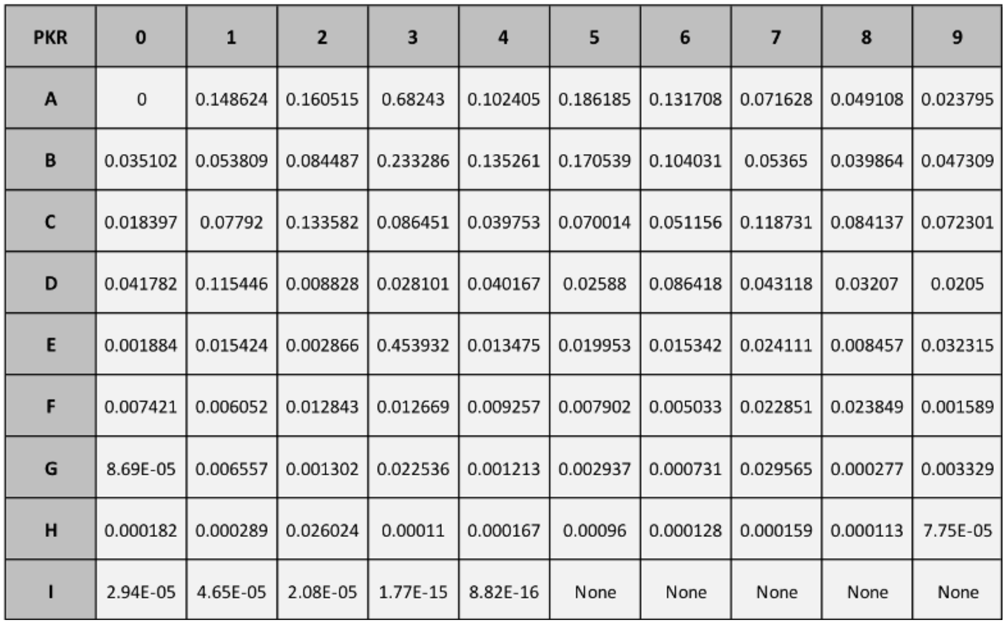
Negishi	0	1	2	3	4	5	6	7	8	9	10
A	0	0.084283	0.085225	0.044055	0.030646	0.063871	0.072934	0.025056	0.011172	0.082987	0.437132
B	0.030077	0.082105	0.082442	0.039851	0.154657	0.050113	0.0354	0.070173	0.024463	0.036936	0.025029
C	0.062908	0.066378	0.047095	0.028718	0.043761	0.014863	0.025536	0.015061	0.014118	0.05371	0.025252
D	0.003846	0.059757	0.019728	0.015168	0.036481	0.01511	0.010263	0.009652	0.013173	0.00102	0.039501
E	0.014352	0.041992	0.02248	0.013086	0.016298	0.000712	0.00201	0.010624	0.006797	0.023566	0.011486
F	0.000134	0.009139	0.005586	0.029858	0.002824	0.001286	0.016869	0.006107	0.000223	0.000564	0.010772
G	0.005944	0.001132	0.000382	0.006866	0.000311	0.011533	0.000385	0.000847	0.000372	0.009572	0.002582
H	3.18E-05	0.000215	0.012096	0.017011	7.66E-05	0.000149	0.000209	0.042064	8.24E-05	7.84E-05	0.018103
I	8.86E-05	0.000151	9.11E-05	5.23E-05	0.002081	8.62E-05	7.51E-05	6.89E-05	5.97E-05	5.38E-05	4.30E-05
J	1.39E-05	3.35E-05	2.63E-05	1.98E-05	1.71E-05	0.012082	3.54E-15	3.54E-15	None	None	None

Figure S6: Mutual information matrix for the Negishi dataset.



PKR	0	1	2	3	4	5	6	7	8	9
A	M1	M_NULL	G1	G_NULL	A_NULL	A1	S2	S1	S3	M6
B	M2	S10	P2	P3	A2	T2	T6	S4	T5	T3
C	O_NULL	O1	P4	T7	M12	T4	L1	L_NULL	L2	M3
D	O2	M5	M13	S14	O8	S5	A4	S9	L5	S_NULL
E	S6	A10	N1	N_NULL	L3	S13	M9	O7	S8	O5
F	O10	T1	O3	A3	P1	O4	S7	M11	O6	M4
G	M8	M17	A8	M14	S11	O11	S12	M10	M7	A7
H	A5	A6	L6	S15	L4	M16	O9	A9	M15	O12
I	O13	M18	A11	T_NULL	P_NULL	None	None	None	None	None

Figure S7: Optimized classifier trellis for the PKR dataset.



PKR	0	1	2	3	4	5	6	7	8	9
A	0	0.148624	0.160515	0.68243	0.102405	0.186185	0.131708	0.071628	0.049108	0.023795
B	0.035102	0.053809	0.084487	0.233286	0.135261	0.170539	0.104031	0.05365	0.039864	0.047309
C	0.018397	0.07792	0.133582	0.086451	0.039753	0.070014	0.051156	0.118731	0.084137	0.072301
D	0.041782	0.115446	0.008828	0.028101	0.040167	0.02588	0.086418	0.043118	0.03207	0.0205
E	0.001884	0.015424	0.002866	0.453932	0.013475	0.019953	0.015342	0.024111	0.008457	0.032315
F	0.007421	0.006052	0.012843	0.012669	0.009257	0.007902	0.005033	0.022851	0.023849	0.001589
G	8.69E-05	0.006557	0.001302	0.022536	0.001213	0.002937	0.000731	0.029565	0.000277	0.003329
H	0.000182	0.000289	0.026024	0.00011	0.000167	0.00096	0.000128	0.000159	0.000113	7.75E-05
I	2.94E-05	4.65E-05	2.08E-05	1.77E-15	8.82E-16	None	None	None	None	None

Figure S8: Mutual information matrix for the PKR dataset.

S2.2 Graph convolutional networks (GCNs)

Molecular graph calculations and all neural network (NN) architectures tested herein were implemented using the Chainer Chemistry (ChainerChem) library[S6]. Our previous study details their general construction.[S1] In all cases, a graph processing network (GPN) was constructed and combined with a dense multi-layer perceptron (MLP), which were trained together as a joint network. All models were trained for 100 epochs on 1 NVIDIA K80 GPU device, unless otherwise specified. Training and test sets were held consistent between models for each reaction dataset. This was done by first splitting each dataset into 90/10 train/test, then splitting the training set into 90/10 train/validation, resulting in a final split of 81/9/10 train/validation/test overall. A dummy predictor that always predicts the most frequent bin in each label category was also created for each dataset as a baseline performance reference.

General parameters and hyperparameter settings are summarized in Table S6, which are held constant between R-GCN and AR-GCN models for all datasets.

Table S6: Computational details and general parameters used for GCN models.

parameter	value	description
loss	sigmoid cross entropy	loss function used for training
optimizer	Adam	model optimization algorithm
train/valid/test	81/9/10	data splitting
batch size	32	batch size used for gradient calculations
epochs	100	number of training epochs
out_dim	128	number of units in the readout
hidden_dim	128	number of units in the hidden layers
n_layers	4	number of convolutional layers[a]
n_atom_types	117	number of allowed atom types
concat_hidden	False	readouts concatenated at each layer
ch_list	None	channels in update layers
input_type	‘int’	input vector type
scale_adj	True	normalize adjacency matrix

[a] AR-GCNs have two additional attention layers with hidden_dim=128 and out_dim= 128.

S3 Expanded results

S3.1 Average accuracy

Expanded modeling results separating accuracies (A_c) and error reductions (ER) are included in Tables S7-S10 along with their averages AA_c and AER, respectively (see main text for equations describing their calculation). It should be noted that since the "CO (g)" category in the PKR dataset is a binary class (either yes or no), the top-3 accuracy will always be 1. This category is therefore excluded from AER and AA_c calculations for this section.

Table S7: Top-1 A_c and AA_c for all model types on all four datasets.

dataset	category	dummy	BM-GBM	CT-GBM	R-GCN	AR-GCN
Suzuki	$\boldsymbol{AA}_c$	0.6350	0.7004	0.6952	0.7494	**0.7663**
	metal	0.3777	0.5732	0.5629	0.6306	**0.6499**
	ligand	0.8722	0.8390	0.8408	0.9036	**0.9081**
	base	0.3361	0.4908	0.4777	0.5455	**0.5896**
	solvent	0.6377	0.6729	0.6751	0.7049	**0.7217**
	additive	0.9511	0.9259	0.9196	**0.9624**	0.9621
C–N	$\boldsymbol{AA}_c$	0.4455	0.5528	0.5370	0.6706	**0.6793**
	metal	0.2452	0.4825	0.4582	0.5989	**0.6162**
	ligand	0.5219	0.5538	0.5710	0.6981	**0.7068**
	base	0.2479	0.5028	0.5003	0.5932	**0.6066**
	solvent	0.3219	0.4582	0.4524	0.5647	**0.5674**
	additive	0.8904	0.7669	0.7031	0.8984	**0.8997**
Negishi	$\boldsymbol{AA}_c$	0.5866	0.7506	0.7312	0.7812	**0.7914**
	metal	0.2887	0.5444	0.5218	0.6555	**0.6730**
	ligand	0.7879	0.8174	0.7900	0.8724	**0.8772**
	temperature	0.3317	**0.6656**	0.6527	0.6188	0.6507
	solvent	0.6938	0.8562	0.8514	0.8868	**0.8915**
	additive	0.8309	0.8691	0.8401	**0.8724**	0.8644
PKR	$\boldsymbol{AA}_c$	0.6123	0.8010	0.7925	0.8005	**0.8101**
	metal	0.4302	**0.7901**	0.7786	0.7132	0.7057
	ligand	0.8792	**0.9351**	0.9237	0.9057	0.9094
	temperature	0.2830	0.5954	0.5649	0.6528	**0.6642**
	solvent	0.3321	0.6183	0.6260	0.6792	**0.6981**
	activator	0.6906	0.8244	0.8015	0.8415	**0.8491**
	CO (g)	0.7245	0.8855	0.8855	0.8717	**0.8868**
	additive	**0.9057**	0.9008	0.8893	0.8906	0.8491
	pressure	0.6528	0.8588	**0.8702**	0.8491	0.8491

Table S8: Top-3 A_c and AA_c for all model types on all four datasets.

dataset	category	dummy	BM-GBM	CT-GBM	R-GCN	AR-GCN
Suzuki	$\boldsymbol{AA}_c$	0.8228	0.8994	0.8949	0.9004	**0.9077**
	metal	0.6744	0.8516	0.8475	0.8482	**0.8597**
	ligand	0.9269	0.9635	0.9606	0.9644	**0.9676**
	base	0.7344	**0.8338**	0.8250	0.8123	0.8285
	solvent	0.8013	0.8637	0.8577	0.8836	**0.8897**
	additive	0.9771	0.9842	0.9832	**0.9934**	0.9931
C–N	$\boldsymbol{AA}_c$	0.6881	0.8106	0.8016	0.8609	**0.8621**
	metal	0.6526	0.7928	0.7772	0.8479	**0.8490**
	ligand	0.6647	0.7933	0.7928	0.8605	**0.8688**
	base	0.6400	0.8008	0.7916	**0.8452**	0.8370
	solvent	0.5677	0.7370	0.7281	0.7973	**0.7997**
	additive	0.9156	0.9290	0.9184	0.9534	**0.9559**
Negishi	$\boldsymbol{AA}_c$	0.7455	0.9044	0.8905	0.9085	**0.9209**
	metal	0.5008	0.7771	0.7674	0.8086	**0.8517**
	ligand	0.8549	0.9548	0.9321	0.9522	**0.9553**
	temperature	0.5885	**0.9031**	0.8772	0.8517	0.8708
	solvent	0.8788	0.9321	0.9402	**0.9537**	**0.9537**
	additive	0.9043	0.9548	0.9354	**0.9761**	0.9729
PKR	$\boldsymbol{AA}_c$[a]	0.7973	0.9364	0.9302	0.9348	**0.9402**
	metal	0.7132	**0.9351**	0.9313	0.9057	0.8906
	ligand	0.9019	**0.9962**	0.9924	0.9849	**0.9962**
	temperature	0.5962	**0.8740**	0.8321	0.8528	0.8604
	solvent	0.5925	0.8779	0.8550	0.8679	**0.8981**
	activator	0.8830	0.9466	0.9275	**0.9774**	**0.9774**
	CO (g)	1.0000	1.0000	1.0000	1.0000	1.0000
	additive	0.9321	**0.9885**	**0.9885**	0.9698	0.9736
	pressure	0.9623	0.9771	0.9847	**0.9849**	**0.9849**

[a] Excludes *CO(g)*.

S3.2 Error reduction

Table S9: Top-1 ER and AER for all model types on all four datasets.

dataset	category	dummy	BM-GBM	CT-GBM	R-GCN	AR-GCN
Suzuki	**AER**	-	–0.0263^a	–0.0554^b	0.2767	**0.3115**
	metal	-	0.3142	0.2977	0.4064	**0.4374**
	ligand	-	–0.2595	–0.2455	0.2462	**0.2809**
	base	-	0.2331	0.2134	0.3155	**2934**
	solvent	-	0.0972	0.1032	0.1854	**0.2319**
	additive	-	–0.5164	–0.6456	**0.2298**	0.2255
C–N	**AER**	-	–0.0413^c	–0.1593^d	0.3453	**0.3604**
	metal	-	0.3143	0.2822	0.4686	**0.4915**
	ligand	-	0.0666	0.1027	0.3685	**0.3868**
	base	-	0.3389	0.3355	0.4590	**0.4769**
	solvent	-	0.2010	0.1924	0.3580	**0.3620**
	additive	-	–1.1275	–1.7095	0.0725	**0.0850**
Negishi	**AER**	-	0.3510	0.2773	0.4439	**0.4565**
	metal	-	0.3595	0.3277	0.5157	**0.5404**
	ligand	-	0.1394	0.0099	0.3985	**0.4211**
	temperature	-	**0.4996**	0.4802	0.4296	0.4773
	solvent	-	0.5305	0.5146	0.6302	**0.6458**
	additive	-	0.2260	0.0540	**0.2453**	0.1981
PKR	**AER**	-	**0.4396**	0.4010	0.3973	0.4199
	metal	-	**0.6316**	0.6115	0.4967	0.4834
	ligand	-	**0.4627**	0.3678	0.2188	0.2500
	temperature	-	0.4357	0.3931	0.5158	**0.5316**
	solvent	-	0.4286	0.4400	0.5198	**0.5480**
	activator	-	0.4326	0.3586	0.4878	**0.5122**
	CO (g)	-	0.5843	0.5843	0.5342	**0.5890**
	additive	-	**–0.0519**	–0.1733	–0.1600	–0.1200
	pressure	-	0.5932	**0.6262**	0.5652	0.5652

AER excluding *additive*: ^{a} 0.0962. ^{b} 0.0922. ^{c} 0.2302. ^{d} 0.2282.

Table S10: Top-3 ER and AER for all model types on all four datasets.

dataset	category	dummy	BM-GBM	CT-GBM	R-GCN	AR-GCN
Suzuki	**AER**	-	0.4088	0.3774	0.4936	**0.5246**
	metal	-	0.5442	0.5315	0.5336	**0.5692**
	ligand	-	0.5013	0.4608	0.5137	**0.5564**
	base	-	**0.3741**	0.3409	0.2934	0.3545
	solvent	-	0.3140	0.2838	0.4142	**0.4449**
	additive	-	0.3106	0.2698	**0.7130**	0.6979
C–N	**AER**	-	0.3568	0.3131	0.5391	**0.5471**
	metal	-	0.4034	0.3585	0.5623	**0.5655**
	ligand	-	0.3837	0.3820	0.5842	**0.6087**
	base	-	0.4468	0.4212	**0.5700**	0.5472
	solvent	-	0.3918	0.3712	0.5311	**0.5368**
	additive	-	0.1582	0.0328	0.4481	**0.4773**
Negishi	**AER**	-	0.5947	0.5199	0.6590	**0.6833**
	metal	-	0.5534	0.5340	0.6166	**0.7029**
	ligand	-	0.6883	0.5325	0.6703	**0.6923**
	temperature	-	**0.7644**	0.7016	0.6395	0.6860
	solvent	-	0.4402	0.5069	**0.6184**	**0.6184**
	additive	-	0.5273	0.3247	**0.7500**	0.7167
PKR	**AER**[a]	-	0.6987	0.6740	0.6844	**0.7145**
	metal	-	**0.7738**	0.7604	0.6711	0.6184
	ligand	-	0.9611	0.9222	0.8462	**0.9615**
	temperature	-	**0.6881**	0.5841	0.6355	0.6542
	solvent	-	0.7003	0.6441	0.6759	**0.7500**
	activator	-	0.5432	0.3801	**0.8065**	**0.8065**
	CO (g)	-	1.0000	1.0000	1.0000	1.0000
	additive	-	**0.8314**	**0.8314**	0.5556	0.6111
	pressure	-	0.3931	0.5954	**0.6000**	**0.6000**

[a] Excludes *CO(g)*

S4 Adversarial controls

Several control studies were conducted to ensure the validity of chemical feature learning.[S7] The main text and results tables above describe the reference model known as the "dummy" predictor. This model simply returns the most frequently occurring top-k labels in each dataset category as its prediction for each task. This gives a baseline accuracy to compare the actual learning our models achieve beyond simply fitting the dictionary frequency distributions. Additional controls conducted for the GBM models included:

1. Shuffling full inputs, leaving outputs in place (shuffle inputs)
2. Ablating the SMILES region of the inputs (Mordred only)
3. Substituting Mordred vectors with random unique numerical vectors (random Mordred)
4. Shuffling Mordred inputs, leaving outputs in place (shuffle Mordred)
5. Ablating the descriptor regions of the inputs (SMILES only)
6. Substituting SMILES vectors with random unique numerical vectors (random SMILES)
7. Shuffling SMILES inputs, leaving outputs in place (shuffle SMILES)

Each of these control models are compared to the standard BM-GBM model as well as the dummy predictor for all four datasets in Tables S11 – S14.

The results of these studies have very important implications, as they indicate that SMILES tokens are insufficient chemical representations for structure learning. On its own, that the SMILES-only models perform similarly to the hybrid (full input, BM-GBM) models in many cases could simply indicate the reduced dimensionality (300 vs 2304) helps feature learning and/or prevents underfitting. However, since some of the random SMILES control models retain similar accuracy, it must be concluded that little chemical feature learning is taking place in the SMILES models. Rather, SMILES are likely acting as unique "barcodes" for reaction components,[S7] and models are presumably able to glean enough information from the presence of specific molecules in reactions to establish some accuracy greater than baseline. This is perhaps unsurprising in datasets such as the Negishi, where the chemical space is limited by dataset size and less-accessible organometallic nucleophiles than in datasets like Suzuki, C–N, and PKR. This results in fewer unique molecules and thus the ability to fit models to their presence or absence.

We strongly recommend caution in reaction learning when using SMILES representations. In line with recent commentary highlighting these control experiments,[S7,S8] we do not suggest that SMILES modeling is entirely futile, but that in our case models trained on SMILES only will likely be unable to generalize to reactions containing new molecules not in the current dataset. That said, these models could perhaps be used to predict reaction conditions with high accuracy for yet untested substrate pairs whose individual components have been previously reported. With the goal of creating generalizable models, we do not explore SMILES-only models any further.

On the other hand, the full-input and Mordred-only models pass the described adversarial controls in most all cases, leading to the conclusion that these input types are capable of representing our chemical space. Replacing the hybrid or Mordred representations with randomized vectors led to a large breakdown of model accuracy (random inputs and random Mordred), as did shuffling their input-output pairs (shuffle inputs and shuffle Mordred). To aid in interpreting these results, we compare AERs between the control models ("straw", AER_s) and their featurized reference models ("feature", AER_f). We calculate the difference in AER between these two models, and take the ratio of this difference to the reference model's AER to give AER_d, which allows for comparison between model types as below. This translates to measuring the percent change in AER observed when true chemical features are replaced with random variables, and thus significant negative numbers should be expected.

$$\text{AER}_d = \frac{\text{AER}_s - \text{AER}_f}{\text{AER}_f} \times 100\% \tag{1}$$

Table S11: Top-1 modeling results (A_c and **AER**s) of control models run on the Suzuki dataset.

top-k	category	dummy	BM-GBM	shuffle inputs	random inputs	Mordred only	shuffle Mordred	random Mordred	SMILES only	shuffle SMILES	random SMILES
top-1	**AER**[i]	-	0.0962	-0.0403	-0.0469	0.0453	-0.0036	-0.0102	**0.1076**	-0.0052	0.0262
	AER$_d$[i]	-	-	–142%	–149%	-	–108%	–122%	-	–105%	–76%
	metal	0.3777	**0.5732**	0.3798	0.5202	0.4749	0.3785	0.4354	0.5459	0.3841	0.5298
	ligand	0.8722	0.8390	0.8536	0.7887	0.8466	**0.8724**	0.8352	0.8556	0.8668	0.8211
	base	0.3361	**0.4908**	0.3071	0.4529	0.4490	0.3093	0.4145	0.4649	0.3276	0.4497
	solvent	0.6377	**0.6729**	0.6463	0.6595	0.6575	0.6459	0.6512	0.6724	0.6462	0.6699
	additive	**0.9511**	0.9259	0.9097	0.7589	0.8977	0.9203	0.8905	0.9161	0.9285	0.8979
top-3	**AER**	-	**0.4088**	-0.0488	0.3252	0.3008	-0.0480	0.1887	0.3736	-0.0492	0.3307
	AER$_d$	-	-	–112%	–20%	-	–116%	–37%	-	–113%	–11%
	metal	0.6744	**0.8516**	0.6718	0.8297	0.7996	0.6748	0.7680	0.8264	0.6747	0.8183
	ligand	0.9269	**0.9635**	0.9234	0.9546	0.9533	0.9216	0.9461	0.9606	0.9221	0.9522
	base	0.7344	**0.8338**	0.7223	0.8092	0.8138	0.7277	0.7856	0.8078	0.7286	0.7960
	solvent	0.8013	**0.8637**	0.7996	0.8624	0.8375	0.8008	0.8306	0.8525	0.8015	0.8620
	additive	0.9771	0.9842	0.9740	0.9812	0.9834	0.9738	0.9783	**0.9863**	0.9733	0.9846

[i] Excludes *additive*.

Table S12: Top-1 modeling results (A_c and **AER**s) of control models run on the C–N dataset.

top-k	category	dummy	BM-GBM	shuffle inputs	random inputs	Mordred only	shuffle Mordred	random Mordred	SMILES only	shuffle SMILES	random SMILES
top-1	**AER**[i]	-	0.2302	-0.0311	0.1009	0.1019	-0.0367	0.0717	**0.3095**	-0.0279	0.2703
	AER$_d$	-	-	–114%	–56%	-	–136%	–30%	-	–109%	–13%
	metal	0.2452	0.4825	0.2259	0.4758	0.3602	0.2281	0.3114	**0.5217**	0.2281	0.4928
	ligand	0.5219	0.5538	0.5139	0.3164	0.5351	0.5067	0.5212	**0.6351**	0.5095	0.5825
	base	0.2479	0.5028	0.2167	0.4877	0.3777	0.2162	0.3546	**0.5298**	0.2240	0.5212
	solvent	0.3219	0.4582	0.2944	0.4639	0.3593	0.2880	0.3618	0.4983	0.3008	**0.5006**
	additive	**0.8904**	0.7669	0.7939	0.6978	0.7660	0.7981	0.7880	0.8061	0.8379	0.7053
top-3	**AER**	-	0.3568	-0.0789	0.2951	0.1662	-0.0937	0.1158	**0.4428**	-0.0917	0.3553
	AER$_d$	-	-	–122%	–17%	-	–156%	–30%	-	–121%	–20%
	metal	0.6526	0.7928	0.6287	0.7702	0.7078	0.6248	0.6822	**0.8203**	0.6248	0.7791
	ligand	0.6647	0.7933	0.6454	0.7616	0.7401	0.6496	0.7181	**0.8407**	0.6379	0.8042
	base	0.6400	0.8008	0.6078	0.7877	0.7106	0.6022	0.6944	**0.8109**	0.6103	0.7955
	solvent	0.5677	0.7370	0.5404	0.7284	0.6393	0.5331	0.6228	**0.7741**	0.5370	0.7532
	additive	0.9156	0.9290	0.9058	0.9212	0.9228	0.9022	0.9203	**0.9370**	0.9033	0.9270

[i] Excludes *additive*.

Table S13: Top-1 modeling results (A_c and **AER**) of control models run on the Negishi dataset.

top-k	category	dummy	BM-GBM	shuffle inputs	random inputs	Mordred only	shuffle Mordred	random Mordred	SMILES only	shuffle SMILES	random SMILES
top-1	**AER**	-	0.3510	-0.0451	0.2669	0.1528	-0.0039	0.0663	**0.3574**	-0.1049	0.2596
	AER$_d$	-	-	–113%	–24%	-	–103%	–57%	-	–129%	–27%
	metal	0.2887	0.5444	0.2294	0.4927	0.4249	0.2342	0.3328	**0.6284**	0.2068	0.5541
	ligand	0.7879	0.8174	0.7738	0.8045	0.7835	0.7528	0.7722	**0.8320**	0.7577	0.8174
	temperature	0.3317	**0.6656**	0.2924	0.6107	0.6204	0.2698	0.5331	0.5654	0.2763	0.5396
	solvent	0.6938	0.8562	0.6866	0.8433	0.7383	0.8352	0.7076	**0.8724**	0.6737	0.8401
	additive	0.8309	**0.8691**	0.8320	0.8417	0.8336	0.8061	0.8304	0.8595	0.8110	0.8304
top-3	**AER**	-	**0.5947**	-0.0869	0.5043	0.3418	-0.1138	0.2127	0.5851	-0.0900	0.5404
	AER$_d$	-	-	–115%	–15%	-	–133%	–38%	-	–115%	–7.6%
	metal	0.5008	0.7771	0.4814	0.7528	0.6737	0.4556	0.6123	**0.8045**	0.4330	0.7900
	ligand	0.8549	**0.9548**	0.8304	0.9370	0.9079	0.8352	0.8740	0.9451	0.8449	0.9354
	temperature	0.5885	**0.9031**	0.5735	0.8885	0.8643	0.5283	0.8078	0.8498	0.5574	0.8401
	solvent	0.8788	0.9321	0.8675	0.9208	0.8934	0.8627	0.8853	**0.9435**	0.8659	0.9370
	additive	0.9043	**0.9548**	0.8950	0.9402	0.9241	0.8982	0.9160	**0.9548**	0.8982	0.9499

Table S14: Top-1 modeling results (A_c and **AER**) of control models run on the PKR dataset.

top-k	category	dummy	BM-GBM	shuffle inputs	random inputs	Mordred only	shuffle Mordred	random Mordred	SMILES only	shuffle SMILES	random SMILES
top-1	**AER**	-	**0.4396**	-0.0737	0.1869	0.3412	-0.1679	0.1016	0.2882	-0.1533	0.1462
	AER$_d$	-	-	–117%	–57%	-	–149%	–70%	-	–153%	–49%
	metal	0.4302	**0.7901**	0.3969	0.6221	0.7405	0.3359	0.5763	0.6183	0.3321	0.4809
	ligand	0.8792	**0.9351**	0.8740	0.9008	0.9160	0.8626	0.9008	0.8969	0.8702	0.8893
	temperature	0.2830	**0.5954**	0.2939	0.4580	0.4695	0.2481	0.3740	0.5687	0.2443	0.5115
	solvent	0.3321	**0.6183**	0.2863	0.4466	0.5687	0.2366	0.3893	**0.6183**	0.3321	0.5076
	activator	0.6906	**0.8244**	0.6412	0.7061	0.7634	0.6374	0.6756	0.8053	0.6565	0.7290
	CO(g)	0.7245	**0.8855**	0.5573	0.7405	0.8817	0.5115	0.7061	0.7939	0.5076	0.6870
	additive	**0.9057**	0.9008	0.8893	0.8969	0.8931	0.8702	0.8855	0.8855	0.8626	0.8931
	pressure	0.6528	**0.8588**	0.8282	0.8435	**0.8588**	0.8168	0.8244	**0.8588**	0.8015	0.8473
top-3	**AER**^{i}	-	**0.6987**	0.1710	0.4857	0.5960	0.1466	0.3295	0.6012	0.1631	0.4645
	AER$_d{}^i$	-	-	–76%	–30%	-	–75%	–45%	-	–73%	–23%
	metal	0.7132	**0.9351**	0.7672	0.8588	0.9046	0.7481	0.8130	0.8588	0.7634	0.8359
	ligand	0.9019	**0.9962**	0.9733	0.9885	0.9885	0.9695	0.9847	0.9885	0.9618	0.9847
	temperature	0.5962	**0.8740**	0.6031	0.7481	0.7863	0.6221	0.6756	0.8359	0.5954	0.7519
	solvent	0.5925	**0.8779**	0.6412	0.7672	0.7977	0.5992	0.7061	0.8626	0.6374	0.7748
	activator	0.8830	0.9466	0.8855	0.9122	0.9351	0.8779	0.8855	**0.9618**	0.8779	0.9198
	CO(g)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	additive	0.9321	**0.9885**	0.9275	0.9771	0.9733	0.9313	0.9542	0.9656	0.9389	0.9618
	pressure	0.9623	0.9771	0.9695	0.9733	**0.9847**	0.9695	0.9733	0.9771	0.9695	0.9771

[i] Excludes *CO(g)*.

S4.1 CT-GBMs

Additional controls were conducted for the classifier trellis (CT) models where the propagated label predictions were withheld from downstream models. As above, AER_d values were recorded for these straw models relative to their featurized CT reference models. Top-1 and top-3 results for these experiments are included in Table S15 and S16, respectively. These results indicate that models heavily rely on upstream label information, and perhaps overfit to these input features.

It is interesting to note that the strongest drop in performance on holdout comes from heavily skewed categories (see C–N *ligand* and *additive*, for example). This is perhaps sensible in that the highest frequency bin in these categories is the `NULL` label. As such, it appears that interdependent models for these categories base their predictions largely on whether or not a reagent is expected to be used in their category at all. This seems to overtake most structural information in the inputs and when propagated predictions are removed, the models break down.

Table S15: Top-1 A_c, AER, and AER$_d$ values for CT-GBMs and CT-control models on all four datasets.

dataset	category	dummy	CT-GBM	CT-holdout
Suzuki	**AER**^{i}	-	**0.0922**	-2.0013
	AER$_d$^{i}	-	-	–2271%
	metal	0.3777	**0.5629**	0.3967
	ligand	**0.8722**	0.8408	0.0302
	base	0.3361	**0.4777**	0.3329
	solvent	0.6377	**0.6751**	0.1143
	additive	**0.9511**	0.9196	0.0217
C–N	**AER**^{i}	-	**0.2282**	-0.2362
	AER$_d$^{i}	-	-	–204%
	metal	0.2452	**0.4582**	0.2284
	ligand	0.5219	**0.5710**	0.0869
	base	0.2479	**0.5003**	0.2435
	solvent	0.3219	**0.4524**	0.3173
	additive	**0.8904**	0.7031	0.0136
Negishi	**AER**	-	**0.2773**	-0.9997^{i}
	AER$_d$	-	-	–461%
	metal	0.2887	**0.5218**	0.2827
	ligand	0.7879	**0.7900**	0.0420
	temperature	0.3317	**0.6527**	0.0129
	solvent	0.6938	**0.8514**	0.6947
	additive	0.8309	**0.8401**	0.0452
PKR	**AER**	-	**0.4010**	-1.7313^{i}
	AER$_d$	-	-	-532%
	metal	0.4302	**0.7786**	0.2137
	ligand	0.8792	**0.9237**	0.0496
	temperature	0.2830	**0.5649**	0.0191
	solvent	0.3321	**0.6260**	0.3626
	activator	0.6906	**0.8015**	0.1412
	CO (g)	0.7245	0.8855	0.4580
	additive	**0.9057**	0.8893	0.0267
	pressure	0.6528	**0.8702**	0.0267

^{i} Excludes additive

Table S16: Top-3 A_c, AER, and AER$_d$ values for CT-GBMs and CT-control models on all four datasets.

dataset	category	dummy	CT-GBM	CT-holdout
Suzuki	$\mathbf{AER}^{i}$	-	**0.3774**	-4.2619
	$\boldsymbol{AER_d}^{i}$	-	-	–1229%
	metal	0.6744	**0.8475**	0.3981
	ligand	0.9269	**0.9606**	0.0325
	base	0.7344	**0.8250**	0.3350
	solvent	0.8013	**0.8577**	0.3110
	additive	0.9771	**0.9832**	0.0238
C–N	$\mathbf{AER}^{i}$	-	**0.3832**	-0.8693
	$\boldsymbol{AER_d}^{i}$	-	-	–327%
	metal	0.6526	**0.7772**	0.2320
	ligand	0.6647	**0.7928**	0.0922
	base	0.6400	**0.7916**	0.6373
	solvent	0.5677	**0.7281**	0.3290
	additive	0.9156	**0.9184**	0.0148
Negishi	**AER**	-	**0.5199**	-0.6714^{i}
	$\boldsymbol{AER_d}$	-	-	–229%
	metal	0.5008	**0.7674**	0.2924
	ligand	0.8549	**0.9321**	0.0468
	temperature	0.5885	**0.8772**	0.4572
	solvent	0.8789	**0.9402**	0.8821
	additive	0.9043	**0.9354**	0.0501
PKR	$\mathbf{AER}^{ii}$	-	**0.6740**	-2.6600^{ii}
	$\boldsymbol{AER_d}^{i}$	-	-	-495%
	metal	0.7132	**0.9313**	0.2137
	ligand	0.9019	**0.9924**	0.1107
	temperature	0.5962	**0.8321**	0.4580
	solvent	0.5925	**0.8550**	0.6756
	activator	0.8830	**0.9275**	0.2519
	CO (g)	1.0000	1.0000	1.0000
	additive	0.9321	**0.9885**	0.0611
	pressure	0.9623	**0.9847**	0.9389

[i] Excludes additive

[ii] Excludes CO (g)

S5 Interpretability

S5.1 GBM feature importances (FIs)

FIs were calculated and averaged over all classifiers in all four BM-GBM models and the randomized control models (random inputs). The FIs plotted over the full input space are shown below. These show uniform decay in the SMILES region, where the token vectors for the three molecules (two reactants + one product) were padded to length 100 each, and thus the frequency of character presence decays with the vector position.

The FIs from the Mordred[S2] vector region were isolated and analyzed for their chemical significance (see main text for discussion). The top 20 FIs from this region are summarized in the tables below. Full length FI rankings and FI charts for all four models and controls are included in the code repository.

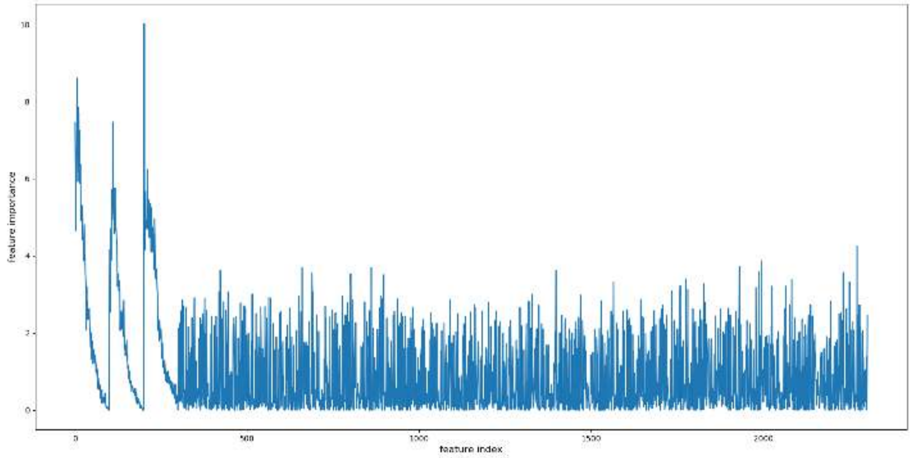


Figure S9: Relative feature importances for the full vector inputs averaged over the Suzuki BM-GBM classifiers.

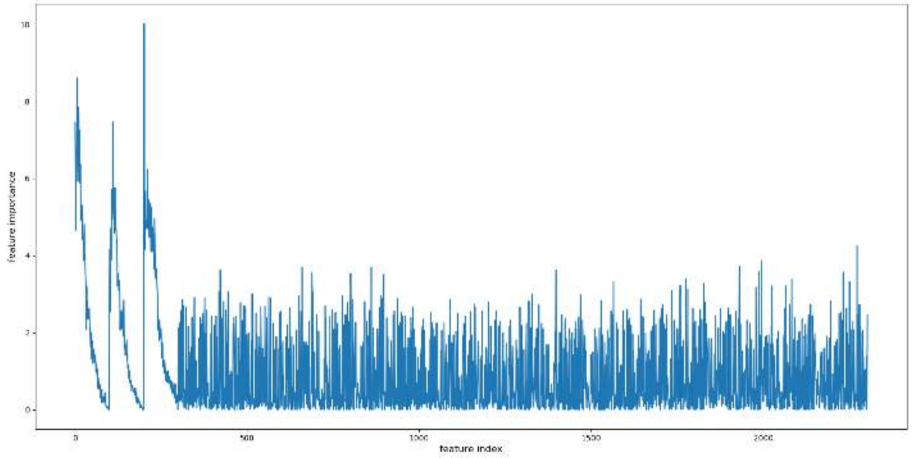


Figure S10: Relative feature importances for randomized vector inputs averaged over the Suzuki BM-GBM classifiers.

Table S17: Top-20 Mordred descriptor FIs for Suzuki BM-GBMs with chemical explanations.

rank	descriptor	species	FI score	description
1	JGI6	product	4.2500	6-ordered mean topological charge[a]
2	JGI3	product	3.8707	3-ordered mean topological charge[a]
3	JGI5	product	3.7241	5-ordered mean topological charge[a]
4	JGI3	reactant 1	3.6983	3-ordered mean topological charge[a]
5	AATSC0p	reactant 1	3.6897	averaged and centered Moreau–Broto autocorrelation of lag 0 weighted by polarizability[b]
6	SsOH	reactant 2	3.6207	sum of sOH[c]
7	IC1	reactant 1	3.6207	1-ordered neighborhood information content[d]
8	JGI4	product	3.5948	4-ordered mean topological charge[a]
9	SdssC	product	3.5690	sum of dssC[e]
10	ATSC3m	reactant 1	3.5603	centered Moreau–Broto autocorrelation of lag 3 weighted by mass[b]
11	EState_VSA6	reactant 1	3.5345	EState VSA Descriptor 6 (<= x <)[f]
12	SdssC	reactant 1	3.5086	sum of dssC[e]
13	JGI9	product	3.3966	9-ordered mean topological charge[a]
14	ATSC4i	product	3.3879	centered Moreau–Broto autocorrelation of lag 4 weighted by ionization potential[b]
15	SdssC	reactant 2	3.3276	sum of dssC[e]
16	SlogP_VSA8	product	3.3190	MOE logP VSA Descriptor 8 (0.25 <= x < 0.30)[g]
17	JGI2	product	3.2759	2-ordered mean topological charge[a]
18	JGI8	product	3.2328	8-ordered mean topological charge[a]
19	PEOE_VSA8	product	3.2069	MOE charge VSA Descriptor 8 (0.00 <= x < 0.05)[h]
20	SsOH	product	3.2069	sum of sOH[c]

[a] n-ordered mean topological charge describes sum of atom-pair charge-transfer terms up to edge-distance n, averaged over all atoms in a molecule.[S9]

[b] Moreau–Broto autocorrelation of lag n weighted by property p describes the distribution of p values over all atom pairs of edge-distance n.[S10,S11]

[c] Sum of electrotopological state of free-alcohol oxygens.[S12,S13]

[d] Measures graph complexity by summing local symmetry over nodes with unique neighborhoods at edge-distance 1.[S14]

[e] Sum of electrotopological state of disubstituted sp^2 carbons.[S12,S13]

[f] Describes the sum of the van der Waals surface area (VSA) with electrotopological state in the range 1.81-2.05.[S12,S14,S15]

[g] Describes the sum of the VSA with SlogP (hybrid atomistic logP) in the range 0.25-0.30.[S12,S16]

[h] Describes the sum of the VSA with partial charge in the range 0.00-0.05.[S12]

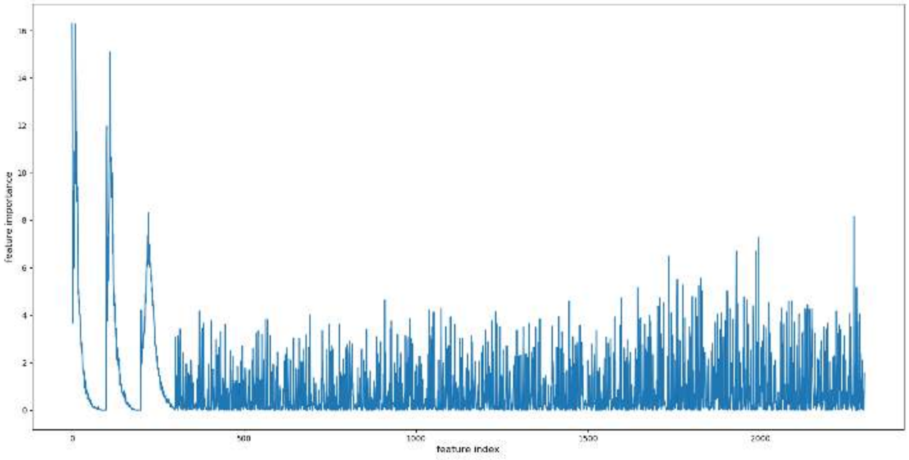


Figure S11: Relative feature importances for the full vector inputs averaged over the C–N BM-GBM classifiers.

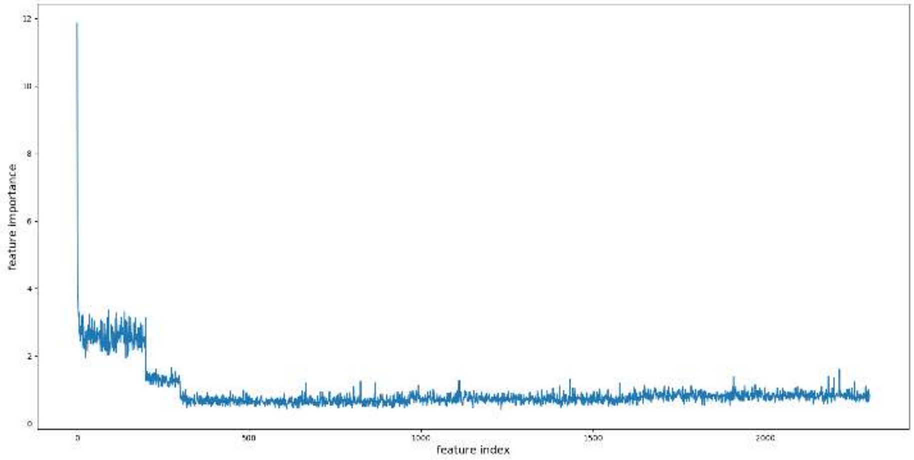


Figure S12: Relative feature importances for randomized vector inputs averaged over the C–N BM-GBM classifiers.

Table S18: Top-20 Mordred descriptor FIs for C–N BM-GBMs with chemical explanations.

rank	descriptor	species	FI score	description
1	JGI6	product	8.1667	6-ordered mean topological charge[a]
2	JGI3	product	7.2778	3-ordered mean topological charge[a]
3	JGI5	product	6.7071	5-ordered mean topological charge[a]
4	JGI4	product	6.6970	4-ordered mean topological charge[a]
5	JGI7	product	6.4899	7-ordered mean topological charge[a]
6	JGI2	product	5.5707	2-ordered mean topological charge[a]
7	JGI8	product	5.5152	8-ordered mean topological charge[a]
8	JGI9	product	5.2727	9-ordered mean topological charge[a]
9	JGI10	product	5.2424	9-ordered mean topological charge[a]
10	ATSC8d	product	5.1717	centered Moreau–Broto autocorrelation of lag 8 weighted by sigma electrons[b]
11	CIC3	product	5.1667	3-ordered complementary information content[c]
12	ATSC2dv	product	5.0303	centered Moreau–Broto autocorrelation of lag 2 weighted by valence electrons[b]
13	ATSC5i	product	5.0202	centered Moreau–Broto autocorrelation of lag 5 weighted by ionization potential[b]
14	ATSC7i	product	4.7929	centered Moreau–Broto autocorrelation of lag 5 weighted by ionization potential[b]
15	MIC1	product	4.7778	1-ordered modified information content weighted by mass[c]
16	ATSC5p	reactant 2	4.7323	centered Moreau–Broto autocorrelation of lag 5 weighted by polarizability[b]
17	EState_VSA7	product	4.7323	EState VSA Descriptor 7 (1.81 <= x < 2.05)[d]
18	CIC4	product	4.7273	4-ordered complementary information content[c]
19	ATSC4dv	product	4.6465	centered Moreau–Broto autocorrelation of lag 4 weighted by valence electrons[b]
20	ATSC5se	reactant 1	4.6414	centered Moreau–Broto autocorrelation of lag 5 weighted by Sanderson electronegativity[b]

[a] n-ordered mean topological charge describes sum of atom-pair charge-transfer terms up to edge-distance n, averaged over all atoms in a molecule.[S9]

[b] Moreau–Broto autocorrelation of lag n weighted by property p describes the distribution of p values over all atom pairs of edge-distance n.[S10,S11]

[c] Difference between actual and maximum possible graph complexity as sum of local symmetry over nodes with unique neighborhoods at edge-distance 3.[S14]

[d] Describes the sum of the van der Waals surface area (VSA) with electrotopological state in the range 1.81-2.05.[S12,S14,S15]

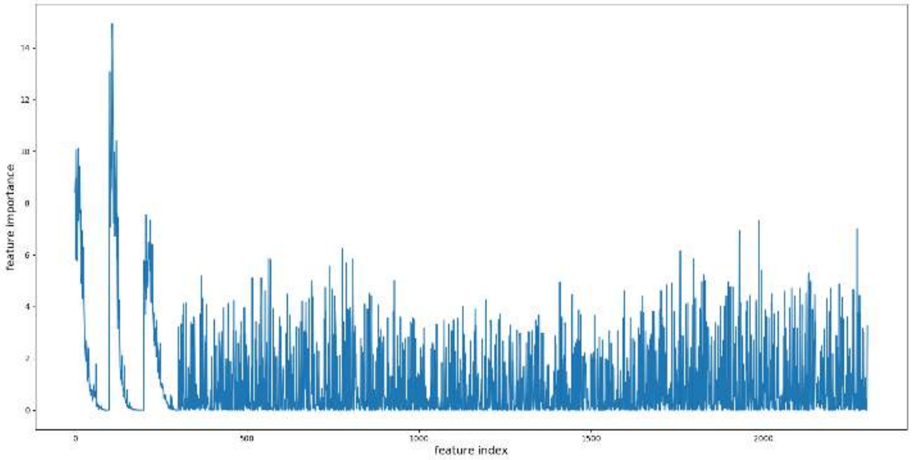


Figure S13: Relative feature importances for the full vector inputs averaged over the Negishi BM-GBM classifiers.

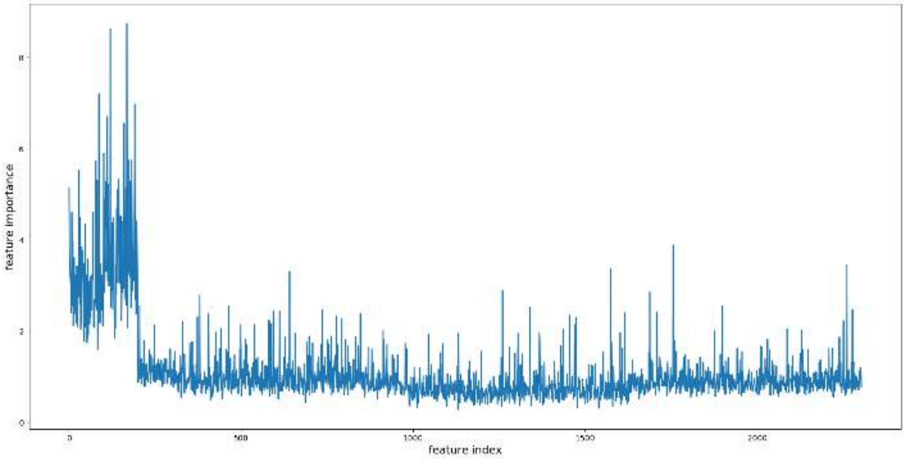


Figure S14: Relative feature importances for randomized vector inputs averaged over the Negishi BM-GBM classifiers.

Table S19: Top-20 Mordred descriptor FIs for Negishi BM-GBMs with chemical explanations.

rank	descriptor	species	FI score	description
1	JGI4	product	7.3238	4-ordered mean topological charge[a]
2	JGI6	product	7.0095	6-ordered mean topological charge[a]
3	JGI5	product	6.9429	5-ordered mean topological charge[a]
4	ATSC4p	reactant 1	6.2286	centered Moreau–Broto autocorrelation of lag 4 weighted by polarizability[b]
5	JGI8	product	6.1619	8-ordered mean topological charge[a]
6	AATSC0i	reactant 1	5.8667	averaged and centered Moreau–Broto autocorrelation of lag 0 weighted by ionization potential[b]
7	SaasC	product	5.8381	sum of aasC[c]
8	SsBr	reactant 1	5.8286	sum of sBr[d]
9	ATSC5i	reactant 1	5.8190	centered Moreau–Broto autocorrelation of lag 5 weighted by ionization potential[b]
10	ATSC6se	reactant 1	5.6762	centered Moreau–Broto autocorrelation of lag 6 weighted by Sanderson electronegativity[b]
11	ATSC1se	reactant 1	5.5619	centered Moreau–Broto autocorrelation of lag 1 weighted by Sanderson electronegativity[b]
12	JGI3	product	5.3905	3-ordered mean topological charge[a]
13	ATSC4d	product	5.2857	centered Moreau–Broto autocorrelation of lag 4 weighted by sigma electrons[b]
14	JGI2	product	5.2286	2-ordered mean topological charge[a]
15	ZMIC2	reactant 1	5.1810	2-ordered Z-modified information content weighted by atomic number[e]
16	IC0	reactant 1	5.1238	0-ordered information content[e]
17	bpol	reactant 1	5.1048	bond polarizability[f]
18	ATSC2dv	product	5.0095	centered Moreau–Broto autocorrelation of lag 2 weighted by valence electrons[b]
19	ATSC5p	reactant 1	5.0095	centered Moreau–Broto autocorrelation of lag 5 weighted by polarizability[b]
20	ATSC3m	reactant 1	5.0000	centered Moreau–Broto autocorrelation of lag 3 weighted by mass[b]

[a] n-ordered mean topological charge describes sum of atom-pair charge-transfer terms up to edge-distance n, averaged over all atoms in a molecule.[S9]

[b] Moreau–Broto autocorrelation of lag n weighted by property p describes the distribution of p values over all atom pairs of edge-distance n.[S10,S11]

[c] Sum of electrotopological state of substituted aromatic carbons.[S12]

[d] Sum of electrotopological state of organobromides.[S12]

[e] Measures graph complexity by summing local symmetry over nodes with unique neighborhoods at edge-distance 2.[S14]

[f] Sum of absolute value of polarizability differences between bound atom pairs.[S2]

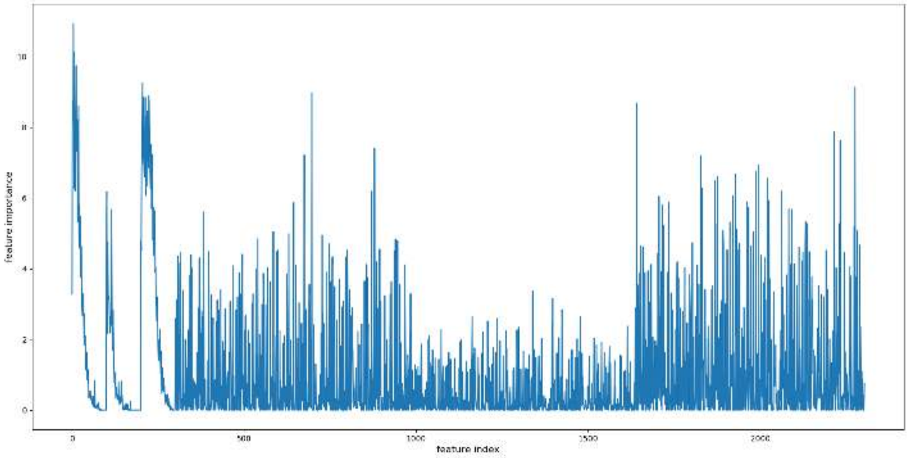


Figure S15: Relative feature importances for the full vector inputs averaged over the PKR BM-GBM classifiers.

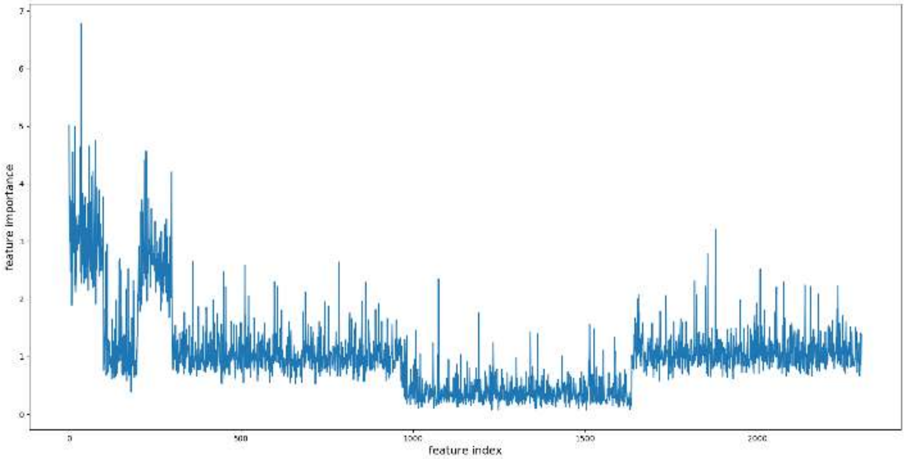


Figure S16: Relative feature importances for randomized vector inputs averaged over the PKR BM-GBM classifiers.

Table S20: Top-20 Mordred descriptor FIs for PKR BM-GBMs with chemical explanations.

rank	descriptor	species	FI score	description
1	SsssCH	product	9.1325	sum of sssCH[a]
2	SddC	reactant 1	8.9759	sum of ddC[a]
3	EState_VSA3	product	8.6747	EState VSA Descriptor 3 (0.29 <= x < 0.72)[b]
4	SdsCH	product	7.8675	sum of dsCH[a]
5	SdssC	product	7.6265	sum of dssC[a]
6	SdsCH	reactant 1	7.3976	sum of dsCH[a]
7	StsC	reactant 1	7.2169	sum of tsC[a]
8	JGI2	product	7.1928	2-ordered mean topological charge[c]
9	JGI3	product	6.9277	3-ordered mean topological charge[c]
10	JGI4	product	6.7590	4-ordered mean topological charge[c]
11	Xch-6dv	product	6.6747	6-ordered Chi chain weighted by valence electrons[d]
12	ATSC3d	product	6.6024	centered Moreau–Broto autocorrelation of lag 3 weighted by sigma electrons[e]
13	Xch-6d	product	6.5542	6-ordered Chi chain weighted by sigma electrons[d]
14	Xch-5dv	product	6.4940	5-ordered Chi chain weighted by valence electrons[d]
15	ATSC2dv	product	6.2771	centered Moreau–Broto autocorrelation of lag 2 weighted by valence electrons[e]
16	SdCH2	reactant 1	6.1928	sum of dCH2[a]
17	ATSC3dv	product	6.1928	centered Moreau–Broto autocorrelation of lag 3 weighted by valence electrons[e]
18	PEOE_VSA7	product	6.0602	MOE charge VSA Descriptor 7 (-0.05 <= x < 0.00)[f]
19	ATSC4v	product	6.0482	centered Moreau–Broto autocorrelation of lag 4 weighted by van der Waals volume[e]
20	PEOE_VSA8	product	5.9277	MOE charge VSA Descriptor 8 (0.00 <= x < 0.05)[f]

[a] Sum of electrotopological state of: sssCH = tertiary aliphatic carbons; ddC = central allenic carbons; dsCH = monosubstituted sp^2 carbons; dssC = disubstituted sp^2 carbons; tsC = internal alkyne carbons; dCH2 = terminal alkene carbons.[S12]

[b] Describes the fraction of the van der Waals surface area (VSA) with electrotopological state in the range listed.[S12,S14,S15]

[c] n-ordered mean topological charge describes sum of atom-pair charge-transfer terms up to edge-distance n, averaged over all atoms in a molecule.[S9]

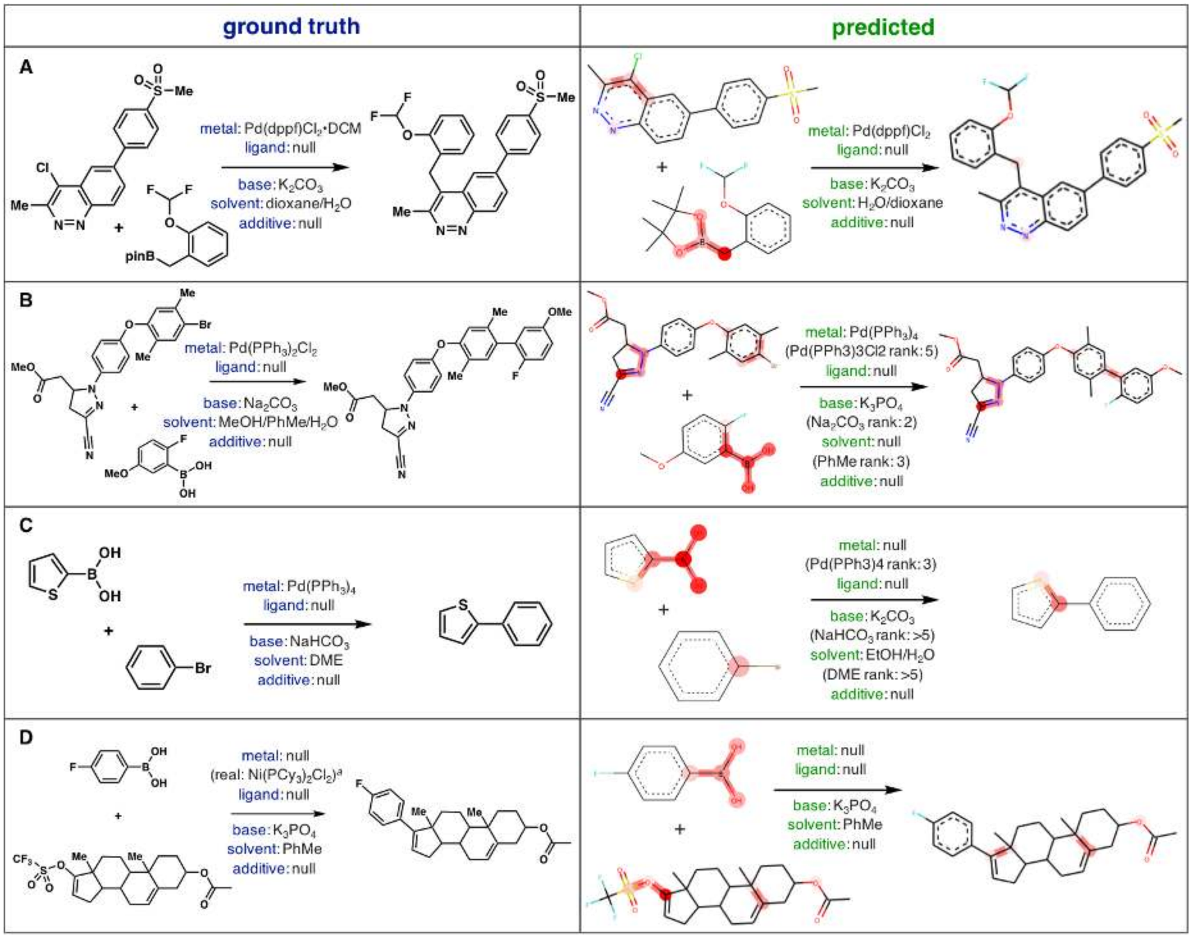
[d] Sum of the products of connectivity degrees of atoms in edge-distance of n-order weighted by property p.[S14]

[e] Moreau–Broto autocorrelation of lag n weighted by property p describes the distribution of p values over all atom pairs of edge-distance n.[S10,S11]

[f] Sum of the VSA with partial charge in the range specified.[S12]

S5.2 AR-GCN attention visualizations

Four random reactions were chosen from each dataset for AR-GCN attention visualization (see main text for explanation). The ground truth category labels and AR-GCN predictions are included in each example.



[a] Ground truth reagent below threshold dictionary frequency; read as null.

Figure S17: Attention visualizations for randomly chosen Suzuki couplings.

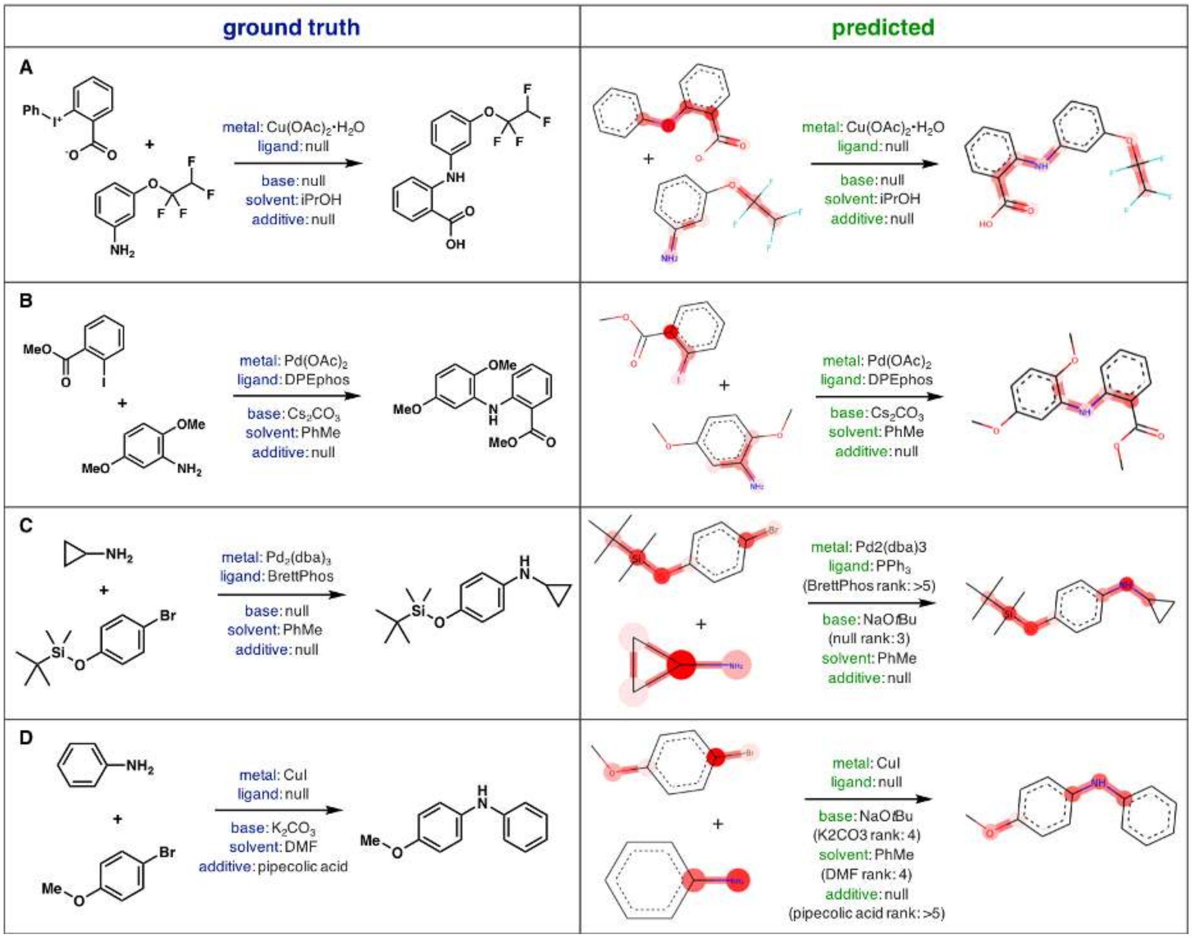
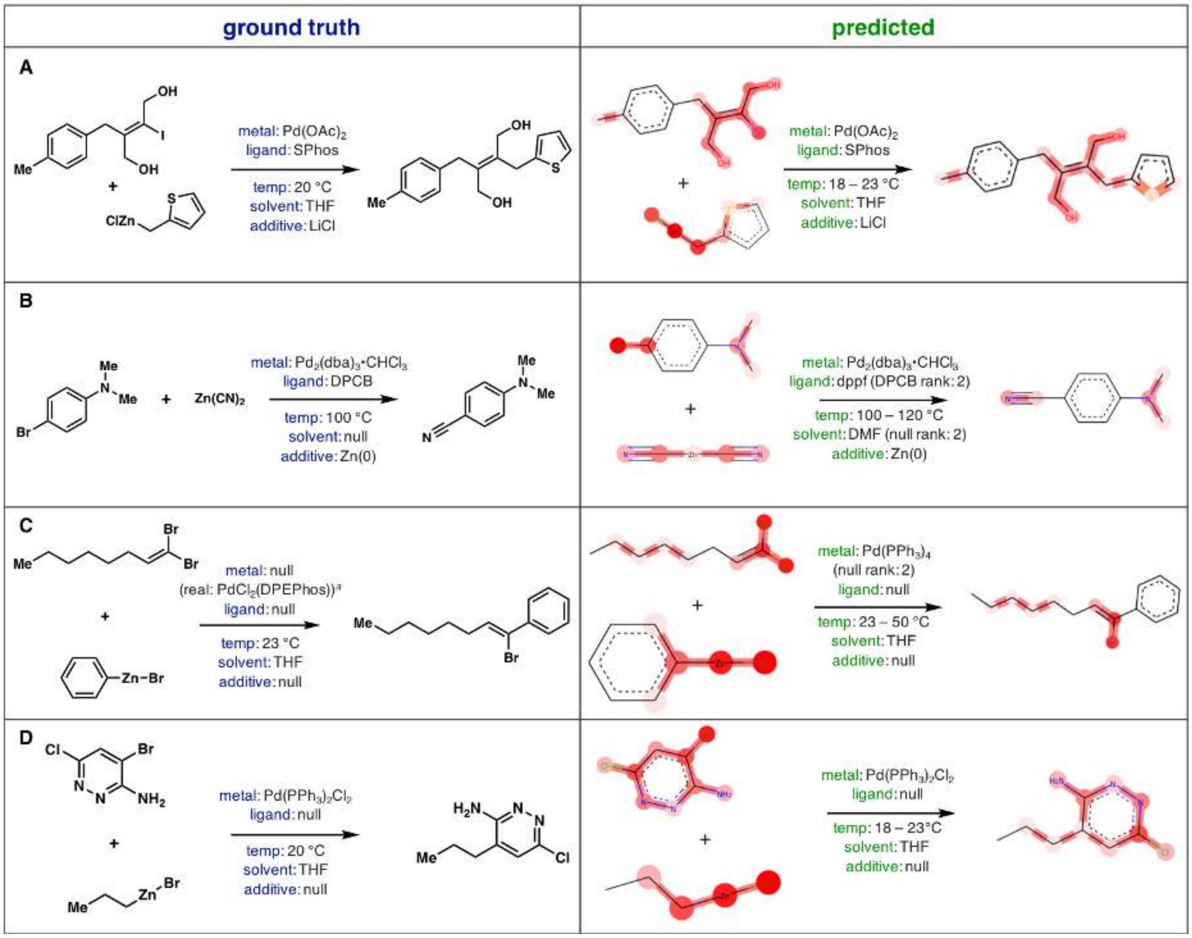


Figure S18: Attention visualizations for randomly chosen C–N couplings.



[a] Ground truth reagent below threshold dictionary frequency; read as null.

Figure S19: Attention visualizations for randomly chosen Negishi couplings.

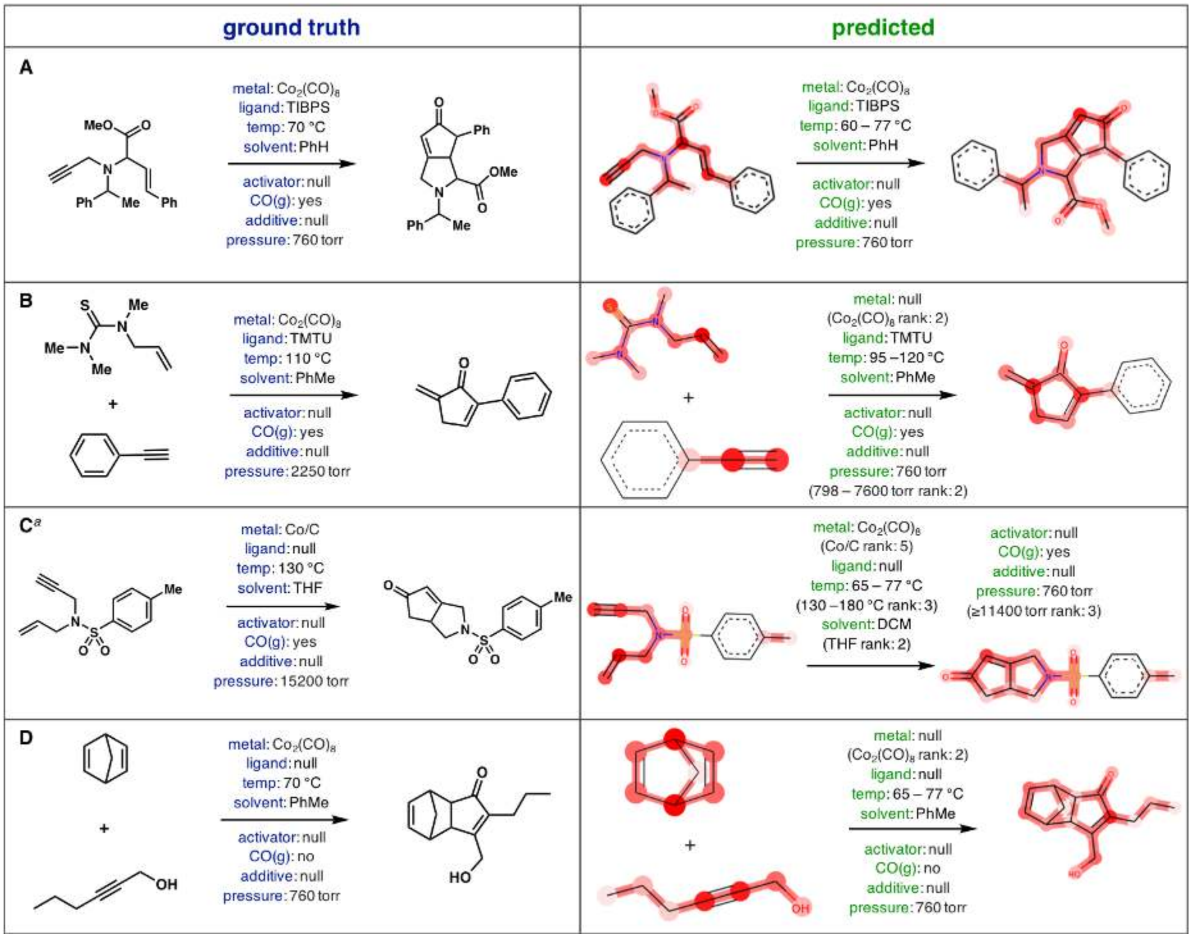


Figure S20: Attention visualizations for randomly chosen PKRs.

References

(S1) Ryou*, S.; Maser*, M. R.; Cui*, A. Y.; DeLano, T. J.; Yue, Y.; Reisman, S. E. Graph Neural Networks for the Prediction of Substrate-Specific Organic Reaction Conditions. *arXiv:2007.04275 [cs, LG]* **2020**,

(S2) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics* **2018**, *10*, 4.

(S3) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 3146–3154.

(S4) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(S5) Read, J.; Martino, L.; Olmos, P.; Luengo, D. Scalable Multi-Output Label Prediction: From Classifier Chains to Classifier Trellises. *Pattern Recognition* **2015**, *48*, 2096–2109.

(S6) Tokui, S.; Oono, K.; Hido, S.; Clayton, J. Chainer: a Next-Generation Open Source Framework for Deep Learning. **2015**,

(S7) Chuang, K. V.; Keiser, M. J. Adversarial Controls for Scientific Machine Learning. *ACS Chemical Biology* **2018**, *13*, 2819–2821.

(S8) Chuang, K. V.; Keiser, M. J. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science* **2018**, *362*, eaat8603.

(S9) Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *Journal of Chemical Information and Modeling* **1994**, *34*, 520–525.

(S10) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *New Journal of Chemistry* **1980**, *4*, 359–360.

(S11) Hollas, B. An Analysis of the Autocorrelation Descriptor for Molecules. *Journal of Mathematical Chemistry* **2003**, *33*, 91–101.

(S12) Hall, L. H.; Mohney, B.; Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *Journal of Chemical Information and Modeling* **1991**, *31*, 76–82.

(S13) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Modeling* **1995**, *35*, 1039–1045.

(S14) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References*, 1st ed.; Methods and Principles in Medicinal Chemistry; Wiley, 2009; Vol. 41.

(S15) Labute, P. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling* **2000**, *18*, 464–477.

(S16) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 868–873.