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]

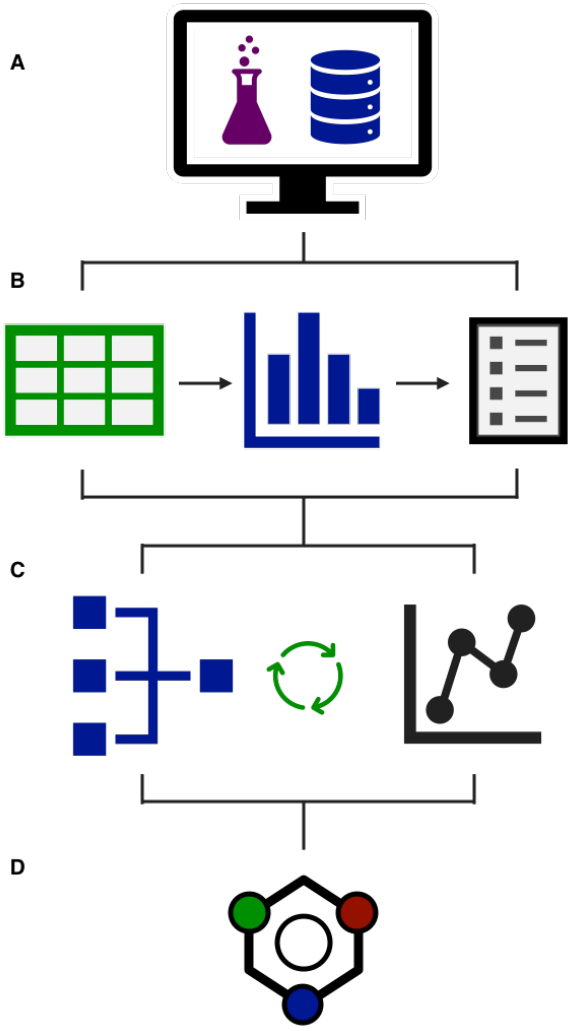


Figure 1: Schematic modeling workflow. A) Data gathering. B) Tabulation and dictionary construction. C) Iterative model optimization. D) Inference and interpretation.

2.1 Data acquisition and preprocessing

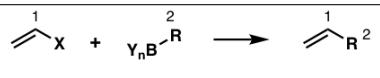
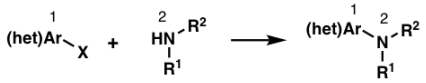
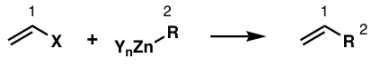
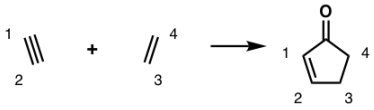
A summary of the datasets studied here is shown in Table 1. Each dataset was manually preprocessed using the following procedure:

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).
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 pseudo-halides 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

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*-Bu)$_3$	L2	ligand
Xantphos	L3	ligand
—	—	—
NaO*t*-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≤T<23	T2	temp
23≤T<50	T3	temp

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

included in Figure 2, and full dictionaries for all reactions are provided in the SI.

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.

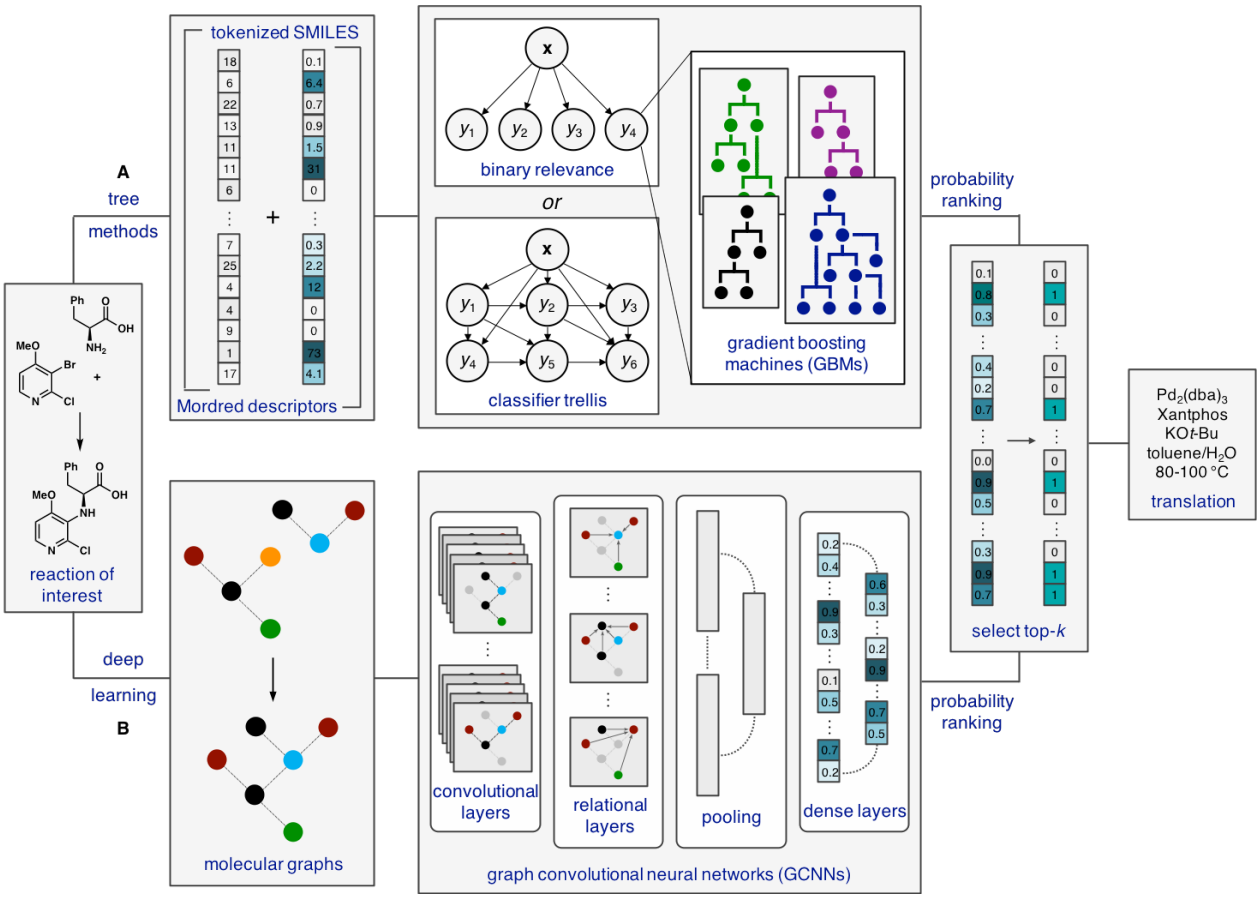


Figure 3: Schematic modeling workflow. A) Tree-based methods. String and descriptor vectors for each molecule in a reaction are concatenated and used as inputs to gradient-boosting machines (GBMs). B) Deep learning methods. Molecular graphs are constructed for each molecule in a reaction, which are passed 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:

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

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:

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

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]

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

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.

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):

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

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**).

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

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 ($\mathrm{ER} = \frac{A_c^g - A_c^d}{1 - A_c^d}$, 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

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)*.

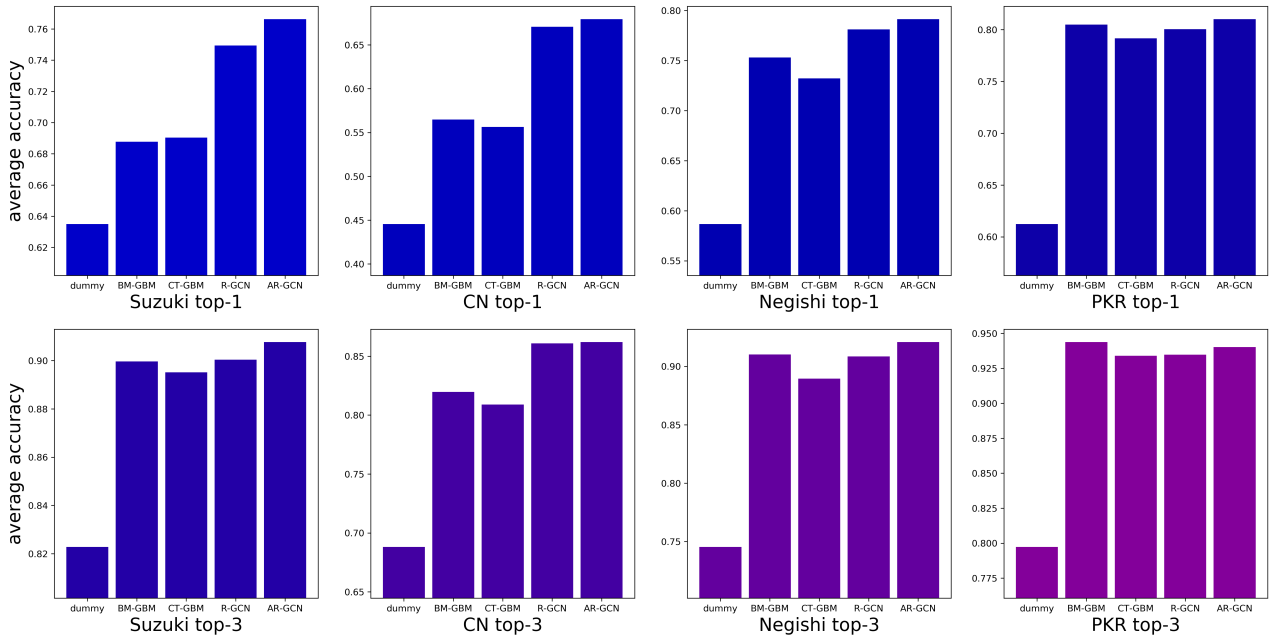


Figure 4: Average top-1 and top-3 categorical accuracies for each model across the four datasets.

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

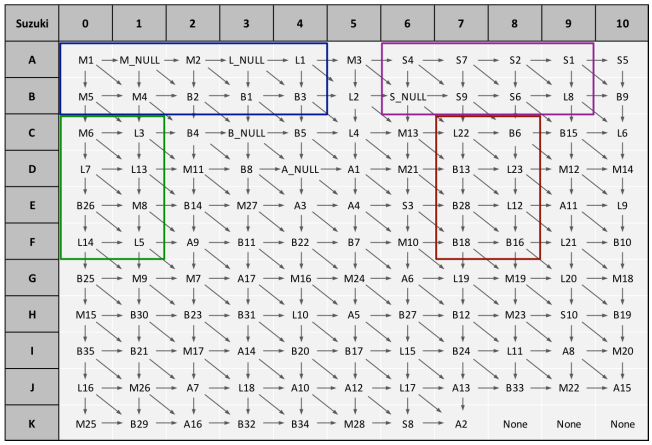


Figure 5: Optimized prediction trellis for the Suzuki dataset.

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` ($P(o\text{-tolyl})_3$).

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` (KOt-Bu). This informs the prediction of `B28` ($LiOH \cdot H_2O$), which then goes on to feed models of `B18` (DIPEA) and `B16` (NaOt-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]

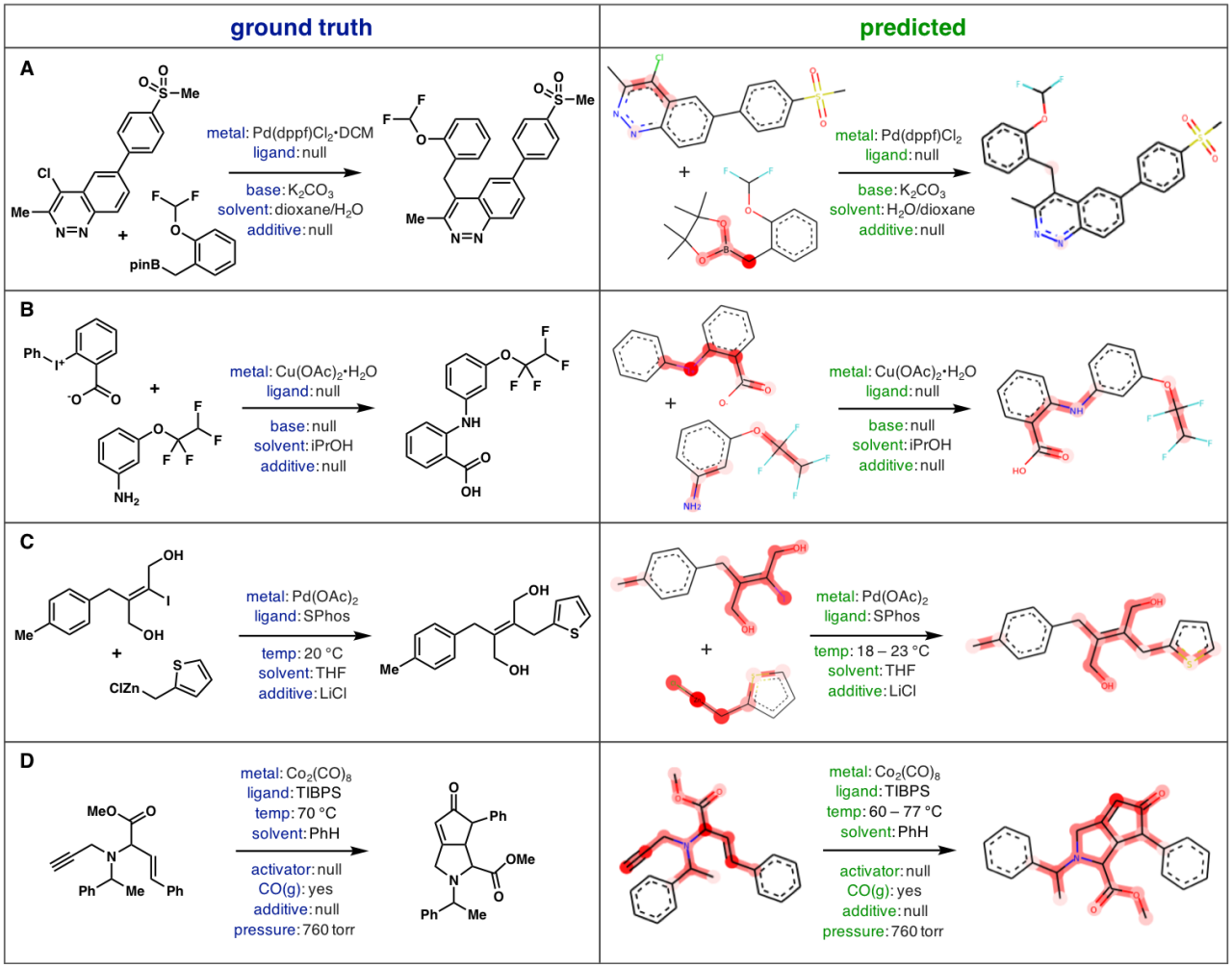


Figure 6: AR-GCN attention weight visualization and prediction examples from randomly chosen reactions in each dataset. Darker highlighting indicates higher attention.

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 inter-

esting 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*.

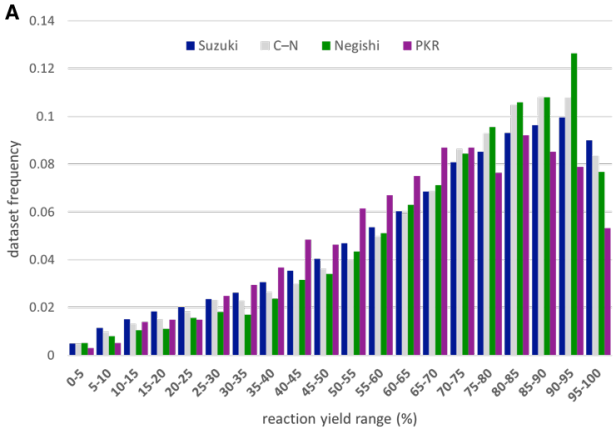
Lastly, an intramolecular PKR (Figure 6D) showed the most uniformly distributed attention of the four examples. Still, the strongest weights are given to the participating alkyne and alkene, with additional emphasis on the amino ester bridging group. Weights are similarly distributed in the product, 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



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).

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.

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 Drug-like 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