A graph-convolutional neural network model for the prediction of chemical reactivity

Connor W. Coley\textsuperscript{a}  Wengong Jin\textsuperscript{b}  Luke Rogers\textsuperscript{a}  Timothy F. Jamison\textsuperscript{c}

Tommi S. Jaakkola\textsuperscript{b}  William H. Green\textsuperscript{a}  Regina Barzilay\textsuperscript{b}  Klavs F. Jensen\textsuperscript{a}

\textsuperscript{a} Department of Chemical Engineering, Massachusetts Institute of Technology
\textsuperscript{b} Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
\textsuperscript{c} Department of Chemistry, Massachusetts Institute of Technology

Abstract

We present a supervised learning approach to predict the products of organic reactions given their reactants, reagents, and solvent(s). The prediction task is factored into two stages comparable to manual expert approaches: considering possible sites of reactivity and evaluating their relative likelihoods. By training on hundreds of thousands of reaction precedents covering a broad range of reaction types from the patent literature, the neural model makes informed predictions of chemical reactivity. The model predicts the major product correctly over 85\% of the time requiring around 100 ms per example, a significantly higher accuracy than achieved by previous machine learning approaches, and performs on par with expert chemists with years of formal training. We gain additional insight into predictions via the design of the neural model, revealing an understanding of chemistry qualitatively consistent with manual approaches.

Introduction

The prediction of reaction outcomes is a fundamental exercise in chemistry. The ability to anticipate reaction products correctly enables chemists to realize more quickly the chemical compounds that they desire and that form the basis of pharmaceutical, electronic, optical, and mechanical applications. A successful computational approach to this as-yet manual task could confirm chemist intuition about different modes of reactivity, enumerate potential side products preceding structural elucidation of a complex mixture, and validate retrosynthetic suggestions from a computer-aided synthesis planning program to increase the confidence of experimental success.

As a significant step toward this ultimate goal of reaction evaluation, we address the task of predicting the major products of organic reactions based on the reactant, reagent, and solvent species. The methodology we describe below is directly applicable to the automated identification of not just major products, but all species in a mixture of products. This fulfills an additional need for impurity identification and quantification during process development, particularly in the context of drug substance manufacturing where it is essential to understand the exact composition of crude product mixtures. Reaction evaluation also play central role in the hypothesize-make-measure iterative cycle underlying small molecule discovery. Small-scale high throughput experimentation has tremendously accelerated chemical synthesis [6, 27], but analysis and interpretation of results has lagged behind. A recent Merck study [23] employed MALDI-TOF MS for rapid online analysis or 1536 well plates in just 10 minutes. Yet with these high throughput analyses, one must specify which masses to quantify or inspect results manually – in the case of the Merck study, there were “nearly 400” distinct mass peaks to extract.

Preprint. Work in progress.
There is a rich history of computer-assistance in chemical synthesis \cite{10,38,12} including this task of reaction prediction. In 1980, Jorgensen and coworkers introduced Computer Assisted Mechanistic Evaluation of Organic Reactions (CAMEO) \cite{29}. This and other early approaches, including EROS \cite{15}, IGOR \cite{37}, SOPHIA \cite{30}, and Robia \cite{35} use expert heuristics to define possible mechanistic reactions. What most of these approaches have in common, advocated particularly strongly by Ugi \cite{37}, is the desire to enable predictions of novel chemistry, i.e., reactions that do not correspond to a previously known codifiable reaction template. However, none achieved broad use within the chemistry community.

Major developments in machine learning and data availability have enabled new approaches to this problem \cite{9}. For specific reaction families with sufficiently detailed reaction condition data, machine learning can be applied to the quantitative prediction of yield \cite{1}. Ideally, however, models could be trained on historical reaction data to predict a broad range of reaction families. One way to do so is by combining the traditional use of reaction templates in cheminformatics with machine learning, either by learning to select relevant reaction rules from a template library \cite{39,32} or by learning to rank template-generated products \cite{8}. Reaction templates are the classic approach to codifying the “rules” of chemistry \cite{20,7,4,56}, whose use dates back to Corey’s synthesis planning program Logic and Heuristics Applied to Synthetic Analysis (LHASA) \cite{11}. Presently, decades later, reaction template approaches continue to find extensive applications in computer-aided synthesis planning \cite{9,35,53}. However, while reaction rules are suitable for interpolating known chemistry to novel substrates, they leave no opportunity to describe reactions with even minor structural differences at the reaction center. Other approaches have involved learning to propose mechanistic or pseudo-mechanistic reaction steps \cite{18,17,13,5}, but these require human annotations or heuristically-generated mechanisms in addition to published experimental results. At the other extreme, one can neglect all chemical domain knowledge and use off-the-shelf machine translation models to generate products directly from reactants \cite{26,31}. Here, we describe a chemically-informed model that incorporates domain expertise through its architecture.

Our overall model structure (Fig. 1) is designed to reflect how expert chemists approach the same task. First, we learn to identify reactive sites that are most likely to undergo a change in connectivity – this parallels the identification of reactive functional groups and consideration of how they might react, but without codifying rigid rules about functional group decomposition (Fig. reffgr:approaches arrow 2). Next, we perform a focused enumeration of products that could result from those interactions subject to chemical valence rules (Fig. reffgr:approaches arrow 3). We learn to rank those candidates – determining what modes of reactivity are most likely, as would a chemist – to produce the final prediction of major products (Fig. reffgr:approaches arrow 4). By dividing the prediction task into these two stages of reactivity perception and outcome scoring, we can gain insight into the neural model’s suggestions and find qualitative alignment with how chemists analyze organic reactivity. The key to the success of our approach is learning a representation of molecules that captures properties relevant to reactivity. We use a graph-based representation of reactant species to propose changes in bond order, introduced in a recent conference publication \cite{16}. Graphs provide a natural way of describing molecular structure; nodes correspond to atoms, and edges to bonds. Indeed, graph theoretical approaches have been used to analyze various aspects of chemical systems \cite{3} and even for the representation of reactions themselves \cite{14}. As we show below, the formalization of predicting reaction outcomes as predicting graph edits—which bonds are broken, which are formed—enables the design and application of graph convolutional models that can begin to understand chemical reactivity.

A quantitative analysis of model performance shows an accuracy of over 85.6%. 5.3% higher than the previous state-of-the-art, and performance on par with human experts for this complex prediction task. Predictions are made on the order of 100 ms per example on a single consumer GPU, enabling its application to virtual screening pipelines and computer-assisted retrosynthesis workflows. More importantly however, the model provides the capacity for collaborative interaction with human chemists through its interpretability. Despite the utility of high-performing black box models, we argue that understanding predictions is equally important. Reaction prediction models that operate on the level of mechanistic steps offer a clear parallel to how human chemists rationalize how reactions proceed \cite{18,17,13,5}. The Baldi group’s ReactionPredictor learns mechanisms from expert-encoded rules as a supervised learning problem \cite{17} and Bradshaw et al.’s ELECTRO model reproduces pseudo-mechanistic steps as defined by an expert-encoded heuristic function \cite{5}. Neither of these approaches enables models to develop their own justifications for predictions, and neither demonstrates perception
We represent the (A) pool of reactant molecules as a (B) attributed graph. A graph convolutional neural network learns to calculate (C) likelihood scores for each bond change between each atom pair. The most likely changes are used to perform a focused, ranked enumeration of (D) candidate products, which are filtered by chemical valence rules. These candidates are then rescored by another graph convolutional network to yield (E) a probability distribution over predicted product species.

of reagent effects. Schwaller et al.’s translation model can illustrate which reactant tokens inform which product tokens [31], which is useful for predicting atom-to-atom mapping, but does not reveal chemical understanding and is not aligned with how humans describe chemical reactivity.

Results

Perceiving likely modes of reactivity

We describe a reaction as a set of changes in bond order in a collection of reactant molecules (Fig. 1A). More formally, we treat these reactant molecules as a single molecular graph where nodes and edges describe atoms and bonds, respectively (Fig. 1B). Reactions are thus a set of graph edits where the edges (or lack of edges) between two or more nodes are changed. One aspect of our approach is the fact that for most organic reactions in this data set, the reaction center – the set of nodes and edges undergoing a change in connectivity – consists of a relatively small number of atoms and typically only up to 5 bonds (Table S1). To be able to describe certain types of reactions not represented in this data set (e.g., cascade reactions involving many mechanistic steps occurring at many atoms throughout the molecule), this observation would need to be revisited.

As the first step in predicting reaction outcomes, we predict the most likely changes in connectivity: the sets of (atom, atom, new bond order) changes that describe the difference between the reactant molecules and the major product. We train a Weisfeiler-Lehman Network (WLN) [22], a type of graph convolutional neural network, to analyze the reactant graph and predict the likelihood of each (atom, atom) pair to change to each new bond order, including a 0th order bond, i.e., no bond (Fig. 1C). The WLN workflow is depicted in Fig. 2 and is described in the following paragraph. Mathematical details can be found in the Supplementary Materials.

The WLN starts with, as input, the reactant graph. Atoms are featurized by their atomic number, formal charge, degree of connectivity, explicit and implicit valence, and aromaticity. Bonds are
Figure 2: Weisfeiler-Lehman Network (WLN) model for predicting likely changes in bond order between every pair of atoms. Starting from an (A) attributed graph representation of molecules, we (B) iteratively update feature vectors describing each atom by incorporating neighboring atoms’ information. After multiple iterations of this embedding, (C) a local feature vector is calculated for each atom based on its updated representation and those of its neighbors. To account for the effects of disconnected atoms such as reagents, (D) a global attention mechanism produces a context vector for each atom as a learned, weighted combination of all other atoms’ local features. Finally, (E) a combination of local features and context vectors are used to predict the likelihood of bond changes for each pair of atoms.

featurized solely by their bond order and ring status. We forgo more complex atom- and bond-level descriptors (e.g., partial charge estimates, surface area contributions) because these can be learned implicitly from the molecular structure and, empirically, do not improve performance. A local embedding iteratively updates atom-level representations by incorporating information from adjacent atoms, processed by a parameterized neural network. To account for the effects of distant atoms, e.g.,activating reagents, a global attention mechanism is used whereby all atoms in the reactant graph attend to (“look at”) all other atoms; a global context vector for each atom is based on contributions from the representations of all atoms weighted by the strength of this attention. Pairwise sums of these learned atom-level representations, both from the local atomic environment and from the influence of all other species, are used to calculate the likelihood scores. The model is trained to score the true (recorded) graph edits highly using a sigmoid cross entropy loss function.

The combinatorial nature of candidate product enumeration is drastically simplified by restricting our enumeration to only draw from the most likely $K$ bond changes. By using up to 5 unique bond changes to generate each candidate outcome – a decision based on the empirically-low frequency of reactions involving more than 5 simultaneous bond changes (Table. S1) – the number of candidates is bounded by

$$\sum_{n=1}^{5} \binom{K}{n}$$

(1)
where \( \binom{n}{k} \) is the binomial coefficient. Valence and connectivity constraints substantially reduce the number of valid candidates produced by this enumeration (Fig. 1D).

Fig. 3 illustrates the efficacy of this approach. Using preprocessed training, validation, and testing sets of ca. 410k, 30k, and 40k reactions from the United States Patent and Trademark Office (USPTO) [24], we evaluate how frequently the true (recorded) product of a reaction is included among the enumerated candidates, approximately normalized to the number of candidates. While an exhaustive enumeration of all possible bond changes guarantees 100% coverage, the subsequent problem of selecting the most likely outcome would be intractable. In this work, the parameter \( K \) can be tuned to simultaneously increase the coverage and increase the number of candidates; similar tuning is possible for the comparative template-based [8] and graph-based [16] enumeration strategies. In a template-based approach, one generally truncates the template library by only including templates that were derived from a certain minimum number of precedent reactions; a higher minimum threshold produces a template set covering more common reaction types.

Evaluating candidate reaction outcomes

The set of valid product structures resulting from the focused enumeration requires additional evaluation and ranking to produce the final model prediction (Fig. 1E). We have previously treated the problem of ranking candidate products as an isolated task [8, 16], yet this fails to utilize a key aspect of the enumeration approach: candidate outcomes produced by combinations of more likely bond changes are themselves more likely to be the true outcome. The quantitative scores for each bond change when perceiving likely modes of reactivity provides an initial ranking of candidate reaction outcomes. The remaining evaluation task is akin to a residual- or \( \Delta \)-learning approach [28] and serves to refine these preliminary rankings.

The Weisfeiler-Lehman Difference Network (WLDN) used for ranking refinement is conceptually similar to the WLN (details in the supplemental information). Reactants and candidate outcomes
Table 1: Overall performance in reaction prediction. \( \theta \) denotes the approximate size of the model (number of trainable parameters). Because all methods produce a ranked list of candidate outcomes, performance is reported in terms of Top-1, 2, 3, and 5 accuracies. *estimated

<table>
<thead>
<tr>
<th>Method</th>
<th>( \theta )</th>
<th>Top-1 [%]</th>
<th>Top-2 [%]</th>
<th>Top-3 [%]</th>
<th>Top-5 [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>WLN/WLDN [16]</td>
<td>3.2 M</td>
<td>79.6</td>
<td>-</td>
<td>87.7</td>
<td>89.2</td>
</tr>
<tr>
<td>Sequence-to-sequence [31]</td>
<td>30 M*</td>
<td>80.3</td>
<td>84.7</td>
<td>86.2</td>
<td>87.5</td>
</tr>
<tr>
<td>This work</td>
<td>2.6 M</td>
<td>85.6</td>
<td>90.5</td>
<td>92.8</td>
<td>93.4</td>
</tr>
</tbody>
</table>

are embedded as attributed graphs to obtain local atom-level representations. For each candidate, a numerical reaction representation is calculated based on the differences in atom representations between that candidate’s atoms and the reactants’ atoms. The overall candidate reaction outcome score is produced by processing that reaction representation through a final neural network layer and adding it to the preliminary score obtained by summing the bond change likelihoods as perceived by the WLN. The model is trained to score the true candidate outcome most highly using a softmax crossentropy loss function. The bond changes corresponding to the top combinations are imposed on the reactant molecules to yield the predicted product molecules, which are then canonicalized in terms of their SMILES [40] representation using RDKit [19].

Quantitative evaluation is performed on the same atom-mapped data set of 410k/30k/40k reactions from the USPTO to enable comparison to 16 and 31. Statistics for this data set are shown in Fig. S5, S6, S7, S8, and S9. For consistency, we explicitly exclude reactant species not contributing heavy atoms to the recorded products in the enumeration; that is, the model is made aware of which species are non-reactants. An exact match in product SMILES is required for a prediction to be considered correct. The performance comparison is shown in Table 1.

The new model offers a substantial improvement in accuracy over the state of the art on this data set. In particular, the top-1, top-2, and top-3 accuracies are each over 5% higher, a significant reduction of the probability of mispredicting the major product. Comparisons to the data subset of Schwaller et al. [31] (Table S2) and Bradshaw et al. [5] (Table S3), can be found in the supplemental information. Predictions are made in ca. 100 ms per example using a single Titan X GPU; a more detailed description of computational cost in terms of wall times for training and testing can be found in the supplemental information.

Fig. 4 shows model prediction performance broken down by the rarity of each reaction in the test set as measured by the popularity of the corresponding reaction template extracted from the training set. More common chemistries are empirically easier to predict, as one would expect from a data-driven approach to model reactivity. Reactions for which the corresponding template has fewer than 5 precedents (or none at all) are still predicted with >60% top-1 accuracy by the graph-based model, demonstrating its ability to generalize to previously unseen structural transformations.

Human benchmarking

To evaluate the difficulty of this prediction task for human experts, we asked eleven human participants to write the likely major products for 80 reaction examples from the test set. The 80 total questions have been divided into 8 categories of 10 randomly-selected questions each based on the rarity of the reaction template that could have been used to recover the true outcome. Among the participants were chemistry and chemical engineering graduate students, postdocs, and professors. A performance comparison is shown in Fig. 4, illustrating that the model performs at the level of an expert chemist for this prediction task. Though we present the quantitative results of the benchmarking study, this is a small-scale qualitative comparison and should be treated as such. Details of this study can be found in the supplementary information.

Because we are making predictions across a wide range of reaction types, performing well on this task necessitates an esoteric knowledge of chemical reactivity. This highlights a key reason why our model is a useful addition to the synthetic chemists toolbox. Machine learning models are well-suited to aggregate massive amounts of prior knowledge and apply them to new molecules, whereas most humans may only be able to recall the most common reaction types and general reactivity trends.
Figure 4: Performance in reaction prediction by the model for the entire test set of 40,000 reactions (top) and for the human benchmarking set of 80 reactions only in comparison to eleven human participants (bottom) as a function of reaction rarity, where rarity refers to the popularity of the corresponding reaction template extracted from the training set.

Interpreting model understanding

In many deep learning applications, improved accuracy comes at the expense of interpretability. Some strategies for interpretation can be applied to models as black box mathematical functions \[25\], but the insight into model behavior is only indirect. Instead, the architecture of deep learning models should inherently enable some degree of interpretability as has been done in natural language processing \[21\].

The global attention mechanism in the WLN, in addition to improving accuracy by accounting for reagent and other long-range effects, is designed to offer interpretability. The predicted reactivity of each atom pair is informed by the atoms’ representations, which are in turn informed by their local and global environments. For example, a phenolic oxygen may not be inherently reactive, but in the presence of a strong base is amenable to various substitution or etherification reactions; in this scenario, we would expect oxygen to attend to a basic reagent in addition to its reacting partner. On the other hand, a diazo compound is of sufficient inherent reactivity that the local environment is all the model requires to predict the outcome accurately, and there is little information gained by attending to the global environment.

Fig. 5 depicts a series of correct predictions from the test set selected to showcase the diversity of correctly-predicted reaction types. The model is able to make accurate predictions for common alkyations (Fig. 5A-B), for cases of ambiguous regioselectivity (Fig. 5C-D), and for various methods of halogenation (Fig. 5E-G). It can distinguish between use cases of similar reagents (e.g., alkyl magnesium species in Fig. 5H-I) and recognize common metal-catalyzed (Fig. 5J-L) or other C-C bond forming reactions (Fig. 5M). It also is able to predict complex preparations of quartenary carbon centers (Fig. 5N), specialized preparations of difluoromethyl ethers (Fig. 5D), Schmidt ring expansions (Fig. 5P), amine nitrosations (Fig. 5Q), and Wittig reactions (Fig. 5R).

With each reaction, we are able to examine which aspects of the reactant species most influences the model’s perception of reactivity at the atom highlighted in green. A darker blue color indicates a larger attention score, which in turn indicates a stronger influence on its perceived reactivity. While these do not reveal information about reaction mechanisms directly, they are a significant step beyond a black box prediction of major products. The global attention mechanism reveals an explanation consistent with how we might justify the prediction in most cases by identifying suitable reaction partners and activating reagents.

For example, Fig. 5A indicates that the iodide carbon’s reactivity is influenced by the presence of suitable reaction partners, namely the three most likely reaction sites on the purine; Fig. 5D is similar. For the Suzuki coupling in Fig. 5L, the aryl boronic acid carbon attends to both the iodo- and bromo-
Figure 5: Correct predictions from the test set illustrating the neural model’s learned chemical intuition and ability to make accurate predictions across a wide range of reaction families. For each example (A-R), we select one atom (green highlight) and examine to what extent every other atom contributed to the perceived reactivity at that location (blue highlight, where a darker color indicates stronger influence). Note that these model-generated structures do not follow traditional drawing standards (e.g., use of abbreviations) in order to illustrate quantitatively attention weights for each atom individually.

sites of its potential coupling partners in addition to sodium carbonate; however, as the reaction database often does not include catalysts even when employed in the actual experiments, the model does not rely on Pd(P(Ph)3)4 as an indication of reactivity. The tertiary carbon in Fig. 5N attends to chloroiodomethane most strongly, but also attends to Cs2CO3 as an activating base. In cases where reactivity can be correctly predicted based on solely local feature vectors (Fig. 5G and Q), attention scores are low and do not reveal additional insight.
Although only the correct rank-1 products are shown above, the model always predicts a probability distribution over multiple product species. An example of impurity prediction showing the top six predictions can be found in Fig. S1. Several examples where the model predicts the recorded outcome as the second-most likely (rank-2) are shown in Fig. S2 and Fig. S3. These “near-misses” are primarily cases where the model predicts an intermediate or a plausible side product, discussed in the supplemental information. Fig. S4 shows additional examples where the recorded outcome is not predicted in the model’s top-10 suggestions. The full set of predictions for all 40,000 test reaction examples is available online in addition to the original code and data.

Conclusion

By designing a neural model to be aligned with how domain experts (chemists) might analyze a problem, we exceed state-of-the-art accuracy in reaction outcome prediction while simultaneously understanding how the model perceives chemical reactivity. Neural networks are therefore not resigned to be used as black box tools, nor are applications of machine learning techniques in chemistry restricted to off-the-shelf models. Model rationales provide insight to human experts and may support new human-machine collaborations for mechanism discovery. In predicting reaction outcomes, the model considers variables in a manner qualitatively similar to that of a human, namely the presence of suitable reaction partners and the effects of reagents and catalysts. The data set covers a broad range of common reaction types that would be found in medicinal and process chemistry settings. We believe that predictive models will have a large role to play in the future of automated experimentation, both to effectively use reaction data and to assist in the interpretation thereof.

Acknowledgements

We thank members of the MIT Department of Chemistry and Department of Chemical Engineering who participated in the human benchmarking study. This work was supported by the DARPA Make-It program under contract ARO W911NF-16-2-0023; CWC received additional funding from the NSF GRFP under Grant No. 1122374.

References


Supporting Information

A graph-convolutional neural network model for the prediction of chemical reactivity

Connor W. Coley, a Wengong Jin, b Luke Rogers, a Timothy F. Jamison, c Tommi S. Jaakkola, b William H. Green, a Regina Barzilay, a,b and Klavs F. Jensen a

E-mail: regina@csail.mit.edu; kfjensen@mit.edu

a Department of Chemical Engineering
b Computer Science and Artificial Intelligence Laboratory
c Department of Chemistry

Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139.
S1 Code and Data

All code used for model training can be found at https://github.com/connorcoley/rexgen_direct. The full data set of USPTO reactions used in this study can be found at the same link. We have included a “deployed” model that uses the trained weights of the model analyzed in detail in the manuscript.
S2 Methods

S2.1 Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u, v$</td>
<td>atoms</td>
</tr>
<tr>
<td>$N(v)$</td>
<td>Set of atoms adjacent to $v$</td>
</tr>
<tr>
<td>$\tau(\cdot)$</td>
<td>ReLU activation function</td>
</tr>
<tr>
<td>$\sigma(\cdot)$</td>
<td>Sigmoid function</td>
</tr>
<tr>
<td>$U, V, W$</td>
<td>matrices in WLN/WLDN</td>
</tr>
</tbody>
</table>

S2.2 Weisfeiler-Lehman Network (WLN)

Weisfeiler-Lehman Network[22] is a type of graph convolutional network derived from Weisfeiler-Lehman (WL) graph kernel[34]. The architecture is designed to embed the computations inherent in WL graph kernel to learn isomorphism invariant representation of atoms. The atom representation is computed by iteratively augmenting the representation of adjacent atoms. Specifically, each atom $v$ is initialized with a feature vector $f_v$ indicating its atomic number, formal charge, degree of connectivity, explicit and implicit valence, and aromaticity. Each bond $(u, v)$ is associated with a feature vector $f_{uv}$ indicating its bond order and ring status. In each iteration, we updated atom representations as follows:

$$f^l_v = \tau \left( U_1 f^{l-1}_v + U_2 \sum_{u \in N(v)} \tau(V_1 f^{l-1}_u + V_2 f_{uv}) \right) \quad (1 \leq l \leq L)$$

where $f^l_v$ is the atom representation at the $l$th iteration, initialized with $f^0_v = f_v$ atom features. $U_1, U_2, V_1, V_2$ are model parameters to be learned, shared across all $L$ iterations. The final local atom representations are computed as

$$c_v = \sum_{u \in N(v)} W_1 f^L_u \odot W_2 f_{uv} \odot W_3 f^L_v$$

We refer the reader to[22] for more details about the mathematical intuition and justification of the WLN.

S2.3 Attention Mechanism

The atom embedding $c_v$ only record local chemical environment, namely atoms and bonds accessible within $L$ steps from atom $v$. Even if $L$ were very large, $c_v$ could not encode any information about other reactant molecules, as information cannot be propagated between two reactant molecules that are disconnected. We argue that it is important to enable information to flow between distant or disconnected atoms. For example, the reaction center may be influenced by certain reagents that are disconnected from reactant molecules. In this case, it is necessary for atom representation $c_v$ to encode such distal chemical effects. Therefore, we propose to enhance the model in previous section with an attention mechanism[2].

Specifically, let $\alpha_{vz}$ be the attention score of atom $v$ upon atom $z$. The “global” atom representation $\tilde{c}_v$ of atom $v$ is calculated as the weighted sum of all reactant atoms where the weight comes from the attention module:

$$\alpha_{vz} = \sigma(u^T \tau(P_0 c_v + P_0 c_z + P_0 b_{vz}))$$

$$\tilde{c}_v = \sum_v \alpha_{vz} c_z$$

The attention score is computed based on “local” atom representations $c_v$ from WLN.
S2.4 Reaction Center Prediction

The WLN is trained to predict reaction center, a set of changes in graph connectivity that describe the difference between reactant molecules and major products. Mathematically, a reaction center is a set \( \{(u, v, b)\} \), where \( (u, v) \) is a pair of atoms whose connecting bond has changed to type \( b \). We predict the likelihood of \( (u, v, b) \) being in reaction center by passing atom representations from WLN through another neural network:

\[
s_{u,v,b} = \sigma(u^T \tau(M_a \bar{e}_u + M_b \bar{e}_v + P_a c_u + P_b c_v + M_b f_{uv}))
\]

The above neural network is jointly optimized with WLN to minimize the cross entropy loss:

\[
- \sum_{u,v,b: u \neq v} y_{u,v,b} \log s_{u,v,b} + (1 - y_{u,v,b}) \log(1 - s_{u,v,b})
\]

where \( y_{u,v,b} = 1 \) iff \( (u, v, b) \) is in the reaction center, and the above loss sweeps over every pair of atoms and bond types (including no bond).

S2.5 Candidate Ranking via Weisfeiler-Lehman Difference Network (WLDN)

At the stage of candidate reaction evaluation, we have a list of candidate products \( \{p_0, p_1, \ldots, p_m\} \) given a set of reactant molecules \( r \). The goal is to learn a scoring function that ranks the true product \( p_0 \) to be the highest. The challenge in ranking candidate products is again representational. We must learn to represent \( (r, p) \) in a manner that can focus on the key difference between the reactants \( r \) and products \( p \), while also incorporating the necessary chemical contexts surrounding the changes.

The architecture of WLDN is designed to highlight such differences. Specifically, it has two components. The first component is a Siamese WLN that learns atom representation of reactant \( r \) and candidate products \( p_i \). Let \( c_v^{(p_i)} \) be the learned atom representation of atom \( v \) in candidate product molecule \( p_i \). We define difference vector \( d_v^{(p_i)} \) pertaining to atom \( v \) as follows:

\[
d_v^{(p_i)} = c_v^{(p_i)} - c_v^{(r)}
\]

Because the reactants and products are atom-mapped, we can use \( v \) to refer to the same atom in different molecules. The second component of WLDN is another WLN that operates on the difference graph between reactants and products. A difference graph \( D(r, p_i) \) is defined as a molecular graph which has the graph structure as \( p_i \), with atom \( v \)'s feature vector replaced by \( d_v^{(p_i)} \). Operating on the difference graph has several benefits. First, in \( D(r, p_i) \), atom \( v \)'s feature vector deviates from zero only if it is close to the reaction center, thus focusing the processing on the reaction center and its immediate context. Second, \( D(r, p_i) \) explicates neighbor dependencies between difference vectors. The WLDN maps this graph-based representation into a fixed-length vector, by applying the second WLN on top of \( D(r, p_i) \):

\[
h_v^{(p_i,l)} = \tau \left( U_1 h_v^{(p_i,l-1)} + U_2 \sum_{u \in N(v)} \tau \left( V_1 h_u^{(p_i,l-1)} + V_2 f_{uv} \right) \right) \quad (1 \leq l \leq L)
\]

\[
y_v^{(p_i)} = \sum_{u \in N(v)} W_1 h_u^{(p_i,L)} \odot W_2 f_{uv} \odot W_3 h_v^{(p_i,L)}
\]

where \( h_v^{(p_i,0)} = d_v^{(p_i)} \). Note that though with the same notation, matrices \( U_s, V_s, W_s \) are distinct parameters from the WLN used in the reaction center prediction. We use the same character for notational convenience.

Let \( RC(p_i) = \{(u_i, v_i, b_i)\} \), the set of bonds that changed from the reactant \( r \) to product \( p_i \). The final score of candidate \( p_i \) is:
\[ s(p_i) = u^T \tau \left( M \sum_{v \in p_i} g_v^{(p_i)} \right) + \sum_{(u,v,b) \in RC(p_i)} s_{u,v,b} \]

Compared to [16], we augment the WLDN with the quantitative scores \( s_{u,v,b} \) for each bond change in reaction center prediction. This is beneficial as the candidate outcomes produced by combinations of more likely bond changes are themselves more likely to be the true outcome.
S3  Additional Results

S3.1  Number of bond changes per reaction

The combinatorics of the enumeration scales poorly with the number of bond changes allowed per reaction. As stated in the main text, the number of candidates per reaction is bounded by

\[ \sum_{n=1}^{5} \binom{K}{n} \]

Where we have allowed up to 5 simultaneous bond changes and \( K = 16 \) (i.e., select up to 5 bond changes from the 16 most likely bond changes) in our later evaluations. The choice of 5 was motivated by an analysis of the number of bond changes in training and validation examples shown in Table S1. We sacrifice 0.1-0.2% loss in maximum possible predictive accuracy through this limitation, but significantly restrict the number of candidates that must be ranked. To allow predictions of the remaining 0.1-0.2%, it would be possible to have a dynamic upper limit of the number of simultaneous bond changes that takes into account what those bond changes are (e.g., to allow complex pericyclic reactions that may have many bond rearrangements).

Table S1: Number of simultaneous bond changes for reaction examples in the USPTO dataset used in this study. Very few reactions involve 6 bond changes, so the candidate enumeration is limited to selecting only up to 5.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>1 [%]</th>
<th>2 [%]</th>
<th>3 [%]</th>
<th>4 [%]</th>
<th>5 [%]</th>
<th>6 [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>17.1</td>
<td>55.1</td>
<td>19.5</td>
<td>6.6</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Validation</td>
<td>17.4</td>
<td>54.9</td>
<td>19.7</td>
<td>6.5</td>
<td>1.4</td>
<td>0.1</td>
</tr>
<tr>
<td>Testing</td>
<td>17.4</td>
<td>55.0</td>
<td>19.8</td>
<td>6.4</td>
<td>1.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

S3.2  Computational cost of training and prediction

An important aspect of predictive model performance is speed or computational cost, particularly in cases where the model may be applied in a high throughput virtual screen. All experiments were run using a single NVIDIA Titan X graphics card. Training of the Reaction Center Prediction model (WLN) completed after 19 hours (140,000 minibatches of 20, an average of 24 ms per example). Training of the Candidate Ranking model (WLDN) completed after 72 hours (2,400,000 minibatches of a single reaction and its candidate outcomes, an average of 108 ms per example).

Prediction times for the 40,000 test examples were 28.5 minutes and 141 minutes respectively using a single Titan X GPU and a single data preprocessing thread. This translates to a throughput of 43 ms/example and 212 ms/example, respectively. It’s important to note that when making these predictions, preprocessing occurs on a single thread (i.e., converting a SMILES to the reactant graph, combinatorically enumerating candidate outcomes and determining their validity). The throughput for testing – as implemented – is currently limited by these CPU-related tasks, particularly for the candidate ranking model. These preprocessing steps could be trivially parallelized at the reaction level as was done during training and would enable inference times below that of training times (i.e., below 24 and 108 ms per example for each model).

S3.3  Reaction prediction performance compared to Schwaller et al.’s single product subset

In [S1], Schwaller et al. report their model performance on single product reactions in addition to those in the full data set. Only 3.4% of test examples in the Jin et al. data set have multiple species in the products (e.g., counterions, amine salts). However, their neural translation model is currently not designed to predict multiple species separated by a “.” SMILES token. Table [S2] shows the comparison using the subset of 38,648 single product examples. While the difference in performance is less significant than with the full test set, our graph-based model still achieves several percent higher accuracy.
Table S2: Performance in reaction prediction when only testing on the 38,648/40,000 reactions with a single reported product (i.e., no counterion or salt).

<table>
<thead>
<tr>
<th>Method</th>
<th>Top-1 [%]</th>
<th>Top-2 [%]</th>
<th>Top-3 [%]</th>
<th>Top-5 [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence-to-sequence (51)</td>
<td>83.2</td>
<td>87.7</td>
<td>89.2</td>
<td>-</td>
</tr>
<tr>
<td>This work</td>
<td><strong>86.4</strong></td>
<td><strong>91.3</strong></td>
<td><strong>92.9</strong></td>
<td><strong>94.2</strong></td>
</tr>
</tbody>
</table>

S3.4 Reaction prediction performance compared to Bradshaw et al’s “linear mechanism” subset

In [5], Bradshaw et al. formulate the task of forward prediction as a predicting a sequence of electron paths as a pseudo-mechanism. However, of the ca. 470k USPTO reactions representing the combined training, validation, and test set used in this study and used by Jin et al. and Schwaller et al. previously, only 73% of examples can be represented in this manner. That is, 27% of reactions from this data set are impossible to predict because they do not fit within this linear framework. The first five test reactions that are not in their subset are a reductive amination, a total deprotection of a tertiary amine to a primary amine, a thioether oxidation to a sulfoxide, thiourea addition to an alkyl iodide, and an alkene ozonolysis to an aldehyde. These are reaction types that one should expect these models to be able to predict. Our formulation of reactions as sets of bond order changes allows 98.6% of test examples to be represented and reconstructed (note: we consider the remaining 1.4% as failed predictions in all evaluations).

In their preprint, Bradshaw et al. show a comparison between ELECTRO and the previous models of Jin et al. and Schwaller et al. However, a very important point must be made: the focus on reactions that can be described as sequential electron movements represents a significant restriction on possible outcomes. This added problem structure simplifies the prediction task and, in the context of our model, would restrict the number of valid enumerated candidates. In general, one should expect it to be easier to perform well on a narrower task with a model that is tailored to that task’s scope.

We make a comparison using the test subset of 29,360 reactions provided by Bradshaw et al.. For the sake of this evaluation, because the exact training data is not available, we use the trained models designed for the broader prediction task over the whole data set. While it would be possible to restrict candidate enumeration to only include products consistent with the restricted reaction scope, we have not done so in this comparison. The results are shown in Table S3. Although the model described in this study is designed for a more general prediction task, it still outperforms the ELECTRO model in top-1 and top-2 accuracy by a small margin.

Table S3: Performance in reaction prediction when only testing on the 29,360/40,000 reactions able to be predicted by the ELECTRO model; our model was not designed to take advantage of the narrower prediction scope but still achieves slightly higher top-1 and top-2 accuracies.

<table>
<thead>
<tr>
<th>Method</th>
<th>Top-1 [%]</th>
<th>Top-2 [%]</th>
<th>Top-3 [%]</th>
<th>Top-5 [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELECTRO (5)</td>
<td>87.0</td>
<td>92.6</td>
<td><strong>94.5</strong></td>
<td><strong>95.9</strong></td>
</tr>
<tr>
<td>This work</td>
<td><strong>88.3</strong></td>
<td><strong>92.9</strong></td>
<td>94.2</td>
<td>95.3</td>
</tr>
</tbody>
</table>

S3.5 Human benchmarking

Supporting information file “Human benchmarking (80 examples) and answers” contains all 80 questions from the human benchmarking test and their answers. Recorded reactants and products are shown in black. For reactions where the model did not predict the true (recorded) product as its top prediction, the predicted outcome is shown in red. Reactions from the test set of 40,000 reactions were divided into 8 categories based on the rarity of the retrosynthetic reaction template required to reproduce the example using a template-based method. Ten reactions were randomly selected from each of these eight categories to cover both common and rare reactions. Human performers were given depictions of the reactant molecules and asked to draw or otherwise indicate the expected major product, which exactly matches the model’s prediction task; no explicit time limit was provided. To evaluate the model on the same data set of 80 test reactions, the model was not given explicit information about which species are known a priori to be reagents, to make it a fair comparison.
The comparison between the model and human performers serves to indicate that this is a nontrivial prediction task and that the model is able to perform at the level of an expert human. A more tightly controlled, larger-scale study would be required for a more rigorous comparison.

S3.6 Application to impurity prediction

Figure S1: Example of impurity prediction for the preparation of an isocyanate. The model correctly predicts the recorded product (rank 1, >99% confidence), but also suggests several minor side products (only ranks 2-6 shown) as potential outcomes.

S3.7 Near-miss predictions

Fourteen mispredictions are shown in Fig. S2 and Fig. S3 where the model predicts the recorded outcome with rank 2; the rank 1 prediction is shown in red. In all cases, the model makes reasonable predictions given the problem formulation and information it has access to. Fig. S2A shows an example of regioisomerism, where the model predicts an isopropylation at the 1 position of the indazole, but the recorded product is at the 2 position. Fig. S2B is an example where the recorded reaction is neutralization of a deprotonated carboxylic acid; under our formalization of reactions as changes in bond order, this would be seen as “no reaction” – an outcome that the model is not allowed to predict. Given that the model must make a prediction of a product with modified bond orders, it defaults to a fairly common reaction from the database, a deprotection, although Cbz is generally stable to base. Fig. S2C is another case of regioisomerism, where condensation can occur on either side of the methyl butyl ketone. Fig. S2D is a case of complex regioselectivity, where the model identifies the recorded iodination as likely, but believes a different site to be more so. One would expect that both products would be observed experimentally. The model prediction for Fig. S2E can be seen as the single-step intermediate prior to a second amidation to the recorded outcome; likewise, the model prediction for Fig. S2F is an intermediate prior to elimination to form the recorded product. Fig. S2G is a C-N bond forming reaction where the rank-1 and rank-2 predictions correspond to the two most likely sites of addition; the distinction between the two is subtle, resulting from a distant bromine that breaks the molecule’s symmetry. The site that the model predicts as more likely does not match what was recorded.

Fig. S3A shows the acylation of an alkene mispredicted as the esterification of an enol. Fig. S3B is a case of regioisomerism similar to Fig. S2G, where the asymmetry between the two reactive sites is subtle. The model rank-1 prediction for Fig. S3C is singly-alkylated intermediate to the recorded and rank-2 prediction, the double-alkylated ethanolamine. The presence of both HF and
Figure S2: Miscellaneous “near-miss” mispredictions from the test set where the model proposes the recorded outcome as its rank-2 prediction (black); the incorrect rank-1 prediction is shown in red. (A) mispredicted regioselectivity for indazole N-alkylation; (B) ester hydrolysis predicted, acid neutralization recorded; (C) kinetically-favored aldol condensation predicted, thermodynamically-favored aldol condensation recorded; (D) misprediction of iodination site selectivity; (E) ring-opening amidation predicted, double-amidation recorded; (F) misprediction of ester elimination upon -ene cyanation; (G) misprediction of N-alkylation selectivity.

pyridine for the epoxide opening in Fig. S3D leads to the prediction of both regioisomers, whereas only one is recorded. The rank-1 and rank-2 predictions of Fig. S3F are tautomers, highlighting the fact that while the SMILES strings of each species are sanitized and canonicalized by RDKit, there are additional standardization steps that might reveal some products to be equivalent. The recorded reaction of Fig. S3F suggests that the example may be missing a reagent; the model, required to make a prediction and in the absence of any better candidates, suggests an unlikely SNAr between chloropyridine and triethylamine. Indeed, the true reaction example includes potassium trifluoro(vinyl)borate (PFIZER LIMITED - US2007/197478, 2007, A1). The recorded Chan-Lam
coupling in Fig. S3G is mispredicted as a Suzuki coupling. Despite the absence of any Pd catalyst, the model has learned that this is a very likely outcome and scores it higher than the true product.

Figure S3: Miscellaneous “near-miss” mispredictions from the test set where the model proposes the recorded outcome as its rank-2 prediction (black); the incorrect rank-1 prediction is shown in red. (A) esterification predicted, alpha carbon acylation recorded; (B) mispredicted regioselectivity of N-alkylation (C) single N-alkylation predicted, double N-alkylation recorded; (D) misprediction of epoxide opening selectivity; (E) prediction of tautomer; (F) SNAr predicted, alkenation recorded; (G) Suzuki coupling predicted, N-alkylation recorded.

S3.8 Complete-miss predictions

Nine mispredictions are shown in Fig. S4 where the model does not predict the recorded outcome in its top ten predictions; the rank 1 prediction is shown in red. In most cases, the model makes reasonable predictions given the problem formulation and information it has access to. Fig. S4A shows a two-step azidation followed by a reduction with sodium borohydride, where the model only predicts the azide product and does not continue to the aniline. Fig. S4B shows a chlorination reaction where the recorded outcome is a substitution at the aryl nitro group; the model instead predicts that the phallic anhydride will open to form the acid chloride. Fig. S4C is a case where the recorded outcome does not make physical sense, due to the presence of an additional carbon atom unlikely to be contributed by any of the reagents. Fig. S4D is similar, where the n-propylation is challenging to explain based on the recorded reactant/reagent species. In Fig. S4E, it is unclear what the source of the carbonyl carbon and oxygen are in the recorded outcome species, but the predicted outcome
is quite reasonable. The recorded and predicted outcomes in Fig. S4F are tautomers, yet this is considered a misprediction due to having distinct SMILES representations. The misprediction in Fig. S4G is a legitimate one, where the model perhaps doesn’t understand the role of mercury oxide and falls back on predicting an alkyne reduction. The spiroether motif is not terribly common in the patent set, so it is likely that the model has not seen enough of these examples to confidently predict this type of reaction. The recorded outcome in Fig. S4H is essentially “no reaction”, and the model’s prediction of the acid chloride makes more physical sense. The final example shown in Fig. S4I is what appears to be a simple N-alkylation with an alkyl bromide, but the recorded product is the imine, rather than the amine.
Figure S4: Miscellaneous “complete-miss” mispredictions from the test set where the model does not propose the recorded outcome (black) in any of its top-10 predictions; the incorrect rank-1 prediction is shown in red. (A) Azidation predicted, two-step azidation and reduction recorded; (B) ring-opening chlorination of anhydride predicted, nitro substitution recorded; (C) alpha acylation predicted, unexplainable alpha arylation recorded with additional carbon; (D) ester cleavage predicted, unexplainable n-propylation recorded; (E) esterification predicted, unexplainably amidation recorded with additional carbonyl; (F) amidation predicted, equivalent amidation tautomer recorded; (G) alkyne reduction predicted, spiroether formation recorded; (H) chlorination of carboxylic acid predicted, disassociated chloride salt recorded; (I) N-alkylation predicted, unexplainable N-alkylation to imine recorded.
Figure S5: Analysis of the USPTO data set used in this study in terms of the number of reactant fragments present in each reaction SMILES as determined by the presence of the period ("." ) symbol.
Figure S6: Analysis of the USPTO data set used in this study in terms of the number of product fragments present in each reaction SMILES as determined by the presence of the period (\text{"."}) symbol. The relatively few examples with multiple product species are primarily salts (e.g., amine HCl salts) or counterions.

Figure S7: Analysis of the USPTO data set used in this study in terms of the diversity of atomic identities appearing in the reactant species. Note the log scale of the y axis.
Figure S8: Analysis of the USPTO data set used in this study in terms of the diversity of atomic identities appearing in the product species. Note the log scale of the y axis. Comparison to Fig. S7 shows that there are fewer distinct heavy atoms that appear in products as compared to reactants; many predominantly appear in reagents or catalysts.

Figure S9: Popularity of reaction templates extracted from the training data set of ca. 410k reactions showing an inverse power law relationship. Of the 30,762 distinct reaction templates, 18,725 have a single precedent reaction and 25,756 have fewer than five.