# Machine Learned Prediction of Reaction Template Applicability for Data-Driven Retrosynthetic Predictions of Energetic Materials

Michael E. Fortunato,[1] Connor W. Coley,[1] Brian C. Barnes,[2, a)] and Klavs F. Jensen[1]

[1)] *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*
[2)] *Energetic Materials Science Branch, CCDC Army Research Laboratory, Aberdeen Proving Ground, MD, USA*

[a)] *Corresponding author: brian.c.barnes11.civ@mail.mil*

**Abstract.** State of the art computer-aided synthesis planning models are naturally biased toward commonly reported chemical reactions, thus reducing the usefulness of those models for the unusual chemistry relevant to shock physics. To address this problem, a neural network was trained to recognize reaction template applicability for small organic molecules to supplement the rare reaction examples of relevance to energetic materials. The training data for the neural network was generated by brute force determination of template subgraph matching for product molecules from a database of reactions in U.S. patent literature. This data generation strategy successfully augmented the information about template applicability for rare reaction mechanisms in the reaction database. The increased ability to recognize rare reaction templates was demonstrated for reaction templates of interest for energetic material synthesis such as heterocycle ring formation.

## INTRODUCTION

Recently there has been a strong push for accelerating the materials discovery process to shorten the time from concept to tangible product [1, 2]. The material discovery process often begins with a prediction of a chemical compound with some desirable property for a given application. These predictions may be produced through human intuition, or through some modeling effort such as *ab initio* calculations to determine electronic structure [3], classical molecular dynamics to determine long-range order [4], and more recently artificial intelligence search algorithms such as generative modeling [5, 6]. It is possible that these predictions result in compounds that have never been synthesized before, and it then becomes the responsibility of a chemist to plan a viable synthetic route to ultimately produce the compound in the lab to test properties of interest. There have been a number of efforts to develop computational tools to aid in the synthesis planning effort [7, 8, 9, 10, 11, 12, 13]. These methods have gained great traction in the pharmaceutical industry for small-molecule drug targets, and are primed to expand to all areas of chemical space. However, the large corpus of public data used to train models tested on pharmaceutical synthesis may only rarely represent the chemistry relevant for synthesis of molecules that undergo a shock-to-detonation transition. Thus, to advance the state of the art in energetic materials chemistry, we must address the problem of machine learning relevant, rare chemistries.

The goal of retrosynthetic analysis, formalized by E.J. Corey [14], is to predict possible reactions that can produce a given target molecule that use smaller, less complex, and easier to synthesize molecules as reactants. This task has usually been accomplished by trained expert chemists who leverage knowledge collected over their years of experience. Recently, efforts to teach computers to accomplish this task using machine learning algorithms have gained much interest. Segler and Waller formulated the problem as a classification task using a neural network (NN), where, given a target molecule, the computational algorithm is tasked to predict a probability distribution over types of molecular transformations (reaction templates) that could be applied to the target molecule [7]. A number of methods that do not rely on reaction templates have also been suggested using sequence-to-sequence [8], transformer [12], or graph based [13] methods. In all of these cases, the ultimate predictive power is limited by the data each model was trained on, and none of these methods escape the problems that arise from rare reaction mechanisms in the reaction databases on which these models were trained.

The database of reactions used in this work was an open-access set of roughly 1.5 million reactions that were extracted from U.S. patent literature (USPTO dataset) by Lowe [15]. Although this provides relatively diverse coverage of possible organic reactions, it can lack sufficient amounts of data in very specific areas of chemistry, namely energetically relevant reactions. An oxygen balance analysis of the product molecules showed that only 2% of the product molecules had an oxygen balance >= -100, and 0.5% of the product molecules had an oxygen balance >= -75. For a

point of reference, the oxygen balance of 2,4,6-trinitrotoluene (TNT) is -74 (where an oxygen balance closer to 0 is more desirable). Although this amounts to thousands of potentially energetic reaction examples, these ultimately get distributed sparsely among the reaction template transformations extracted from the reaction database. In this work, we expand upon data-driven methods for reaction pathway prediction, specifically aiming to improve the recognition of the rare reaction templates relevant for the synthesis of energetic materials. Two neural networks were trained: 1) a reaction relevance NN similar to the work of Segler and Waller and 2) a new template applicability NN. The ability of each neural network to recognize useful templates, specifically those that represent rare or exotic chemical reactions, will be compared.

## DATA SOURCE AND METHODS

The template-based method proposed by Segler and Waller was chosen to study further in this work due to its ability to more easily be supplemented by data augmentation. Reaction data was taken from the open-access USPTO dataset. The reactions were available as atom-mapped SMILES strings, which enabled the identification of reaction centers and automatic extraction of reaction templates using Coley's template extraction code, a part of the rdchiral Python package [16]. In total, 253,795 unique reaction templates were extracted from 1,571,440 valid reaction examples. Each product molecule from the reaction database was converted from its SMILES string format to a Morgan fingerprint using RDKit [17], with a radius of two and a fixed length folding into 2048 bits. These fingerprint bit vectors served as the input features for the neural network classifiers. It is common practice to only retain the reaction templates which show up more frequently in the reaction database, filtering out those that appear less than some threshold value. The reasoning for this decision can be understood by imagining, in the extreme, a model trained to recommend a template with only one reaction example. It will be nearly impossible for the model to generalize and recommend this template unless the exact product molecule is fed through the neural network due to the lack of diversity of training examples. Figure 1 shows the distribution of template popularity extracted from the USPTO reaction database. Here, template popularity means the number of example reactions in the database from which the same template was extracted. Although the majority of training examples (right) represent popular templates, the majority of templates (left) fall under the category of rare templates.

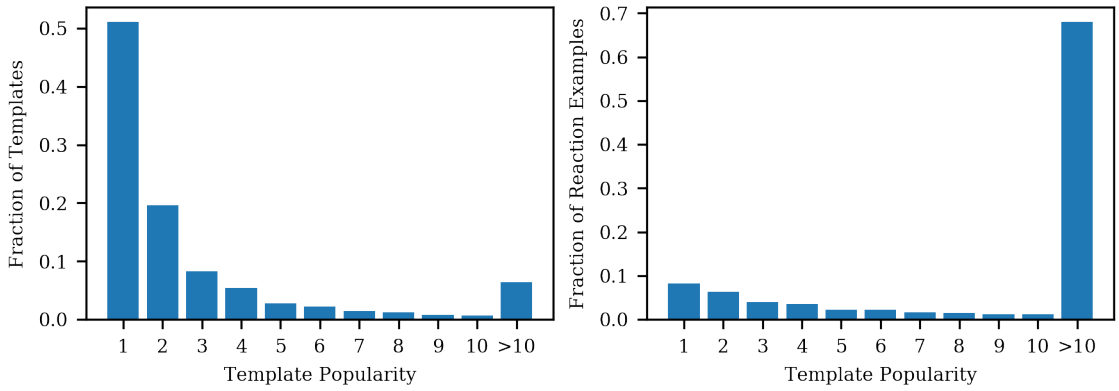


**FIGURE 1.** Distribution of USPTO templates (left) and training examples (right) by template popularity. Although the majority of training examples represent popular templates, the majority of templates fall under the category of rare templates.

One way to address the rare template problem is to collect a larger database of reaction examples to increase the number of so-called "popular" reaction templates. However, this leaves many potentially interesting molecular transformations out of model training. As explained in the remainder of this paper, the goal of this work is to understand to what extent these rare reaction templates can be included through the use of data augmentation.

In order to supplement template recommendation with more information, a neural network was trained to identify template applicability, or sub-graph matching between a given molecule and the entire set of templates (template applicability NN). This is an important distinction from the reaction relevance NN which is only trained to recognize a single reaction template when given a molecule as input. Broadly speaking, the reaction relevance NN is taught to learn the single best template, whereas the template applicability NN is taught to learn all of the templates that

can possibly generate precursors. The data augmentation strategy is important to provide more data from which the model can learn template applicability. Sub-graph isomorphism between a template and molecule can be determined exactly, however it is prohibitively expensive to perform this test at the time of inference for every template in the library. Instead, by generating synthetic data on template applicability ahead of time for a large dataset of molecules, a neural network can be trained to approximate exact sub-graph matching. For this work, the unique set of product molecules from the USPTO dataset of reactions was curated, and tested for template applicability. After removing duplicate products, 929,889 molecules remained. However, this method is extensible in that even larger databases of molecules could be included to further augment the training data and improve model performance.

The synthetic data was generated by enumerating pairs of molecules and templates and using RDKit to try and apply each template to each molecule. If the template was a sub-graph match of part of the molecule, a 1 was recorded in a sparse matrix with indices according to the molecule and template. Rows of this sparse matrix, with vector elements representing the templates that matched a given molecule, served as multi-class multi-label training labels to train the template applicability NN. Performing $2.36\text{x}10^{11}$ (929,889 molecules x 253,795 templates) individual template applications required $\sim$13,000 CPU hours. Figure 2 shows the distribution of supporting reaction examples following template applicability augmentation. Many new examples for rare templates were introduced, shifting the distribution towards more popular templates. The fraction of augmented reaction examples for popular templates (those with $\geq$ 10 examples), was greater than 99.9%. Additionally, a set of training labels that only contained the index of the “true" template class (the template that was extracted from the reaction to make a given product molecule) was also prepared to train a neural network similar to the work of Segler and Waller [7] (reaction relevance NN).

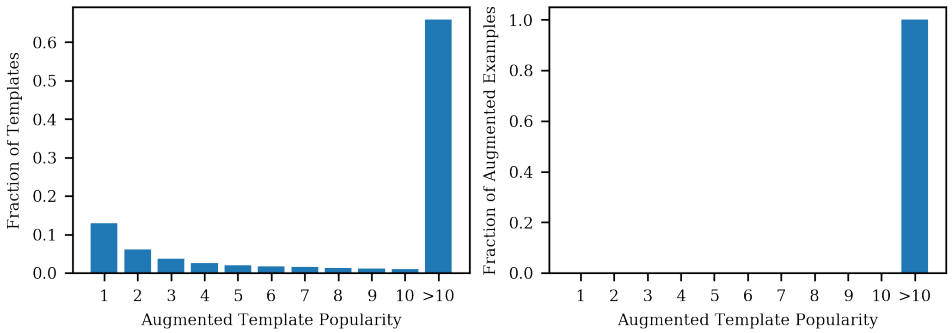


**FIGURE 2.** Distribution of USPTO templates (left) and training examples (right) by template popularity following data augmentation. Compared to Figure 1, the distribution has shifted towards more reaction examples per template.

To evaluate the performance of the reaction relevance NN on rare templates, the reaction dataset was split into train, validation, and test sets. To ensure appropriate evaluation on both rare and popular classes, the examples were split such that, where possible, at least one example for each class ended up in the test set. In general, an 80/10/10 split by template class was used for reaction examples of popular templates. If fewer than 10 reaction examples existed for a given template, 1 example was placed in the test set, 1 in the validation set, and the rest were placed in the training set. In the special case where only 2 examples existed, the examples were split among training and testing. Reaction examples corresponding to templates that only appeared once were divided randomly following an 80/10/10 split.

For both the reaction relevance and template applicability NNs, a fully connected feed forward neural network with a network size of 5 layers of 300 artificial neurons per layer was used, based on our previous, unpublished work. Although this served as a preliminary starting point that offered acceptable performance, it is recognized that a more thorough hyperparameter optimization should be carried out to interrogate the effect of network size on performance of both networks. Both models were trained using the Adam optimizer [18] with a learning rate of $1\text{x}10^{-3}$. Rectified linear unit (ReLU) activation functions were used at each layer except the final output layer, where a softmax activation was used for the reaction relevance NN and a sigmoid activation was used for the template applicability NN. The loss function in both cases was the appropriate implementation of categorical cross-entropy depending on the mutual exclusivity of training class labels. The data augmentation strategy, while increasing examples for rare templates, also greatly increased class imbalance between the remaining rare and popular templates. To handle this increased class imbalance, the contribution to the loss function was weighted inversely proportionally to the number of examples per class (only in the case of the template applicability NN). Models were built and trained using the Keras API [19] for TensorFlow [20].

## RESULTS

Figure 3 shows the reaction relevance NN top-k accuracy results for the test set across the entire test set (blue), popular reaction templates (green; greater than 10 reaction examples), rare reaction templates (orange; fewer than 10 reaction examples and greater than 5 reaction examples), and very rare reaction templates (red; fewer than 5 reaction examples). Test examples where an identical fingerprint was included in model training were omitted from evaluation. These duplicate entries occurred either due to reactions with identical product molecules and templates (but possibly different reactants or reaction conditions) or due to complete fingerprint bit collision. Relative to the average performance across the entire test set, there was a boost in performance for popular templates, however there was a drop in performance for rare and very rare templates. This leads to the conclusion that if these templates describe desirable molecular transformation and should be included in model training, more precedent reaction examples need to be included during model training.

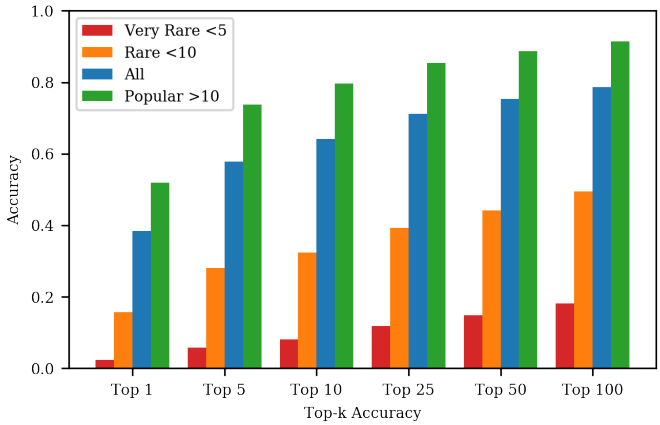


**FIGURE 3.** Top-k reaction relevance neural network accuracy for prediction of the “true" template for a given molecule. The true template is defined as the template that was extracted from the reaction which produced the given molecule. Model performance increases with the number of example reactions available during training.

Because the labels for the template applicability NN examples had multiple true values, recall and precision were used as metrics to evaluate the performance in place of top-k accuracy. These metrics evaluate the ability of the network to identify all of the applicable templates (recall) without overestimating applicability and predicting false positives (precision). Maximizing both of these metrics leads to a more performant model. Figure 4 shows the recall (left) and precision (right) during model training on the training and validation sets. Recall and precision on the test set were 0.830 and 0.834, respectively.

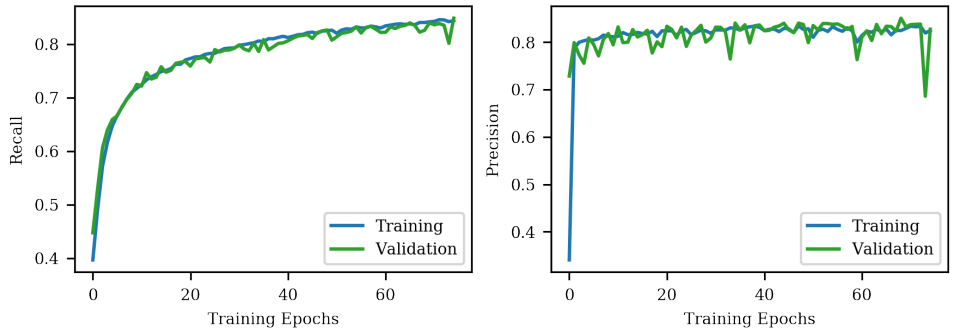


**FIGURE 4.** Recall (left) and precision (right) of the template applicability neural network during model training for training and validation data sets.

## Examples of Improved Template Recognition

Figure 5 shows a selection of cases where the template applicability NN was better able to recognize true templates (those extracted from reactions that resulted in the given product molecule). Specifically, these examples showcase the ability for this new template applicability NN to improve template recognition for rare reaction templates. In each case shown, the product molecule from the known reaction was passed through each neural network, and the score given for the known template is shown. It should be noted that although the score shown is higher, the predicted class from the template applicability NN is just one of many predicted templates that may generate precursors. The small scores output by the reaction relevance NN imply that these would not be selected as a candidate transformation in a synthetic planning tool powered by this neural network model. Conversely, the high score output from the template applicability NN implies that the model was successfully able to recognize the transformation as useful.

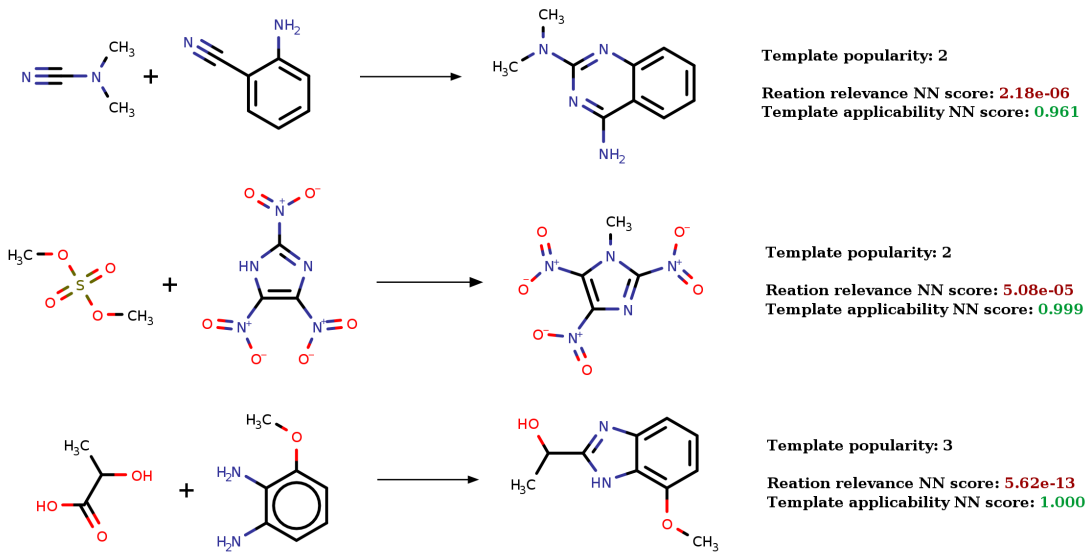


**FIGURE 5.** Example product molecules and reactions where the template applicability NN was better able to recognize the true template. Selected reactions were examples of heterocycle ring formation and product molecules heavily functionalized with nitro groups.

## CONCLUSIONS

A new neural network trained on synthetically generated data that represents template applicability showed improved performance to recommend rare templates of interest to energetic material synthesis compared to the reaction relevance NN. Because of the nature of the training data, a useful ranking between predicted templates is still missing and the model does not function well in isolation. However, in combination with the reaction relevance NN, results from the template applicability NN can be used to quickly filter out predicted templates that do not lead to precursors. This can speed up computation and lead to increased use of rare templates. A further study to optimize the hyperparameters during model training could lead to improved performance in the case of both models.

## ACKNOWLEDGMENTS

This research was supported with funding from the CCDC Army Research Laboratory via award W911NF-19-2-0034. Computational resources were provided by the High Performance Computation Modernization Program (HPCMP). We would like to thank Betsy Rice for providing helpful comments on the manuscript.

# REFERENCES


1. J. J. de Pablo, B. Jones, C. L. Kovacs, V. Ozolins, and A. P. Ramirez, Curr. Opin. Solid St. M. **18**, 99–117 (2014).
2. J. J. de Pablo, N. E. Jackson, M. A. Webb, L. Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, *et al.*, npj Comput. Mater. **5**, 41 (2019).
3. E. F. Byrd and B. M. Rice, J. Phys. Chem. A **110**, 1005–1013 (2006).
4. M. Baer, Thermochim. Acta **384**, 351–367 (2002).
5. B. Sanchez-Lengeling and A. Aspuru-Guzik, Science **361**, 360–365 (2018).
6. O. Prykhodko, S. Johansson, P. C. Kotsias, E. J. Bjerrum, O. Engkvist, and H. Chen, ChemRxiv (2019), 10.26434/chemrxiv.8299544.v3.
7. M. H. S. Segler, M. Preuss, and M. P. Waller, Nature **555**, 604–610 (2018).
8. B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender, and V. Pande, ACS Cent. Sci. **3**, 1103–1113 (2017).
9. J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade, and H. Y. Ando, J. Chem. Inf. Model. **49**, 593–602 (2009).
10. C. W. Coley, L. Rogers, W. H. Green, and K. F. Jensen, ACS Cent. Sci. **3**, 1237–1245 (2017).
11. T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, and A. Toutchkine, Chem **4**, 522–532 (2018).
12. P. Karpov, G. Godin, and I. Tetko, ChemRxiv (2019), 10.26434/chemrxiv.8058464.v1.
13. X. Liu, P. Li, and S. Song, bioRxiv:677849 (2019).
14. E. J. Corey, Angew. Chem. Int. Ed. Engl. **30**, 455–465 (1991).
15. D. M. Lowe, *Extraction of Chemical Structures and Reactions from the Literature*, Ph.D. thesis, University of Cambridge (2012).
16. C. W. Coley, W. H. Green, and K. F. Jensen, J. Chem. Inf. Model. (2019).
17. G. Landrum *et al.*, "Rdkit: Open-source cheminformatics," `https://www.rdkit.org/` (2006).
18. D. P. Kingma and J. Ba, arXiv:1412.6980 (2014).
19. F. Chollet *et al.*, "Keras," `https://keras.io` (2015).
20. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015), software available from tensorflow.org.