CHEMISTRY OF MATERIALS

pubs.acs.org/cm

Methods/Protocols

Fast Customization of Chemical Language Models to Out-of-Distribution Data Sets

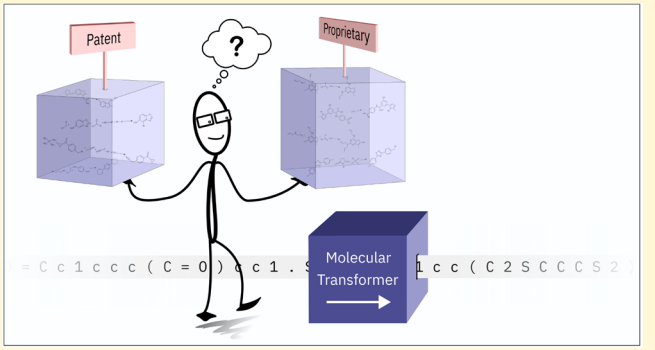
Alessandra Toniato,[§] Alain C. Vaucher,[§] Marzena Maria Lehmann, Torsten Luksch, Philippe Schwaller, Marco Stenta, and Teodoro Laino*

Cite This: *Chem. Mater.* 2023, 35, 8806–8815

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The world is on the verge of a new industrial revolution, and language models are poised to play a pivotal role in this transformative era. Their ability to offer intelligent insights and forecasts has made them a valuable asset for businesses seeking a competitive advantage. The chemical industry, in particular, can benefit significantly from harnessing their power. Since 2016 already, language models have been applied to tasks such as predicting reaction outcomes or retrosynthetic routes. While such models have demonstrated impressive abilities, the lack of publicly available data sets with universal coverage is often the limiting factor for achieving even higher accuracies. This makes it imperative for organizations to incorporate proprietary data sets into their model training processes to improve their performance. So far, however, these data sets frequently remain untapped as there are no established criteria for model customization. In this work, we report a successful methodology for retraining language models on reaction outcome prediction and single-step retrosynthesis tasks, using proprietary, nonpublic data sets. We report a considerable boost in accuracy by combining patent and proprietary data in a multidomain learning formulation. This exercise, inspired by a real-world use case, enables us to formulate guidelines that can be adopted in different corporate settings to customize chemical language models easily.



1. INTRODUCTION

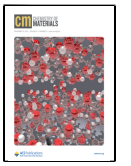
Language models have the potential to change business operations. In recent years, their size has grown exponentially and, with it, their performance in language tasks such as question answering, machine translation, or summarization. If this trend continues, many industries will be redefined and our society will undergo a major transformation within our lifetimes, including the manner in which we run data-driven research operations. This is especially true for the chemical industry.

Chemistry has already raised broad interest in the field of artificial intelligence (AI) for some time already. For instance, data-driven models trained on chemical knowledge have witnessed increased adoption to accelerate chemical discovery, by suggesting potential compounds of interest,[1,2] predicting their properties,[3] or recommending how to synthesize them.[4−10] Among the different architectures, language models demonstrated to be the most flexible architectures when dealing with the continuous flow of new chemical data.[5,10−22]

Many AI models for chemistry are trained on publicly available data. Public data sets, however, are frequently biased toward particular areas of chemical space or reaction classes. Consequently, they may be of little relevance to the chemistry of interest to organizations. To improve the accuracy of these models and increase their scope of applicability within an organization, one must leverage more specific data sets. Many large corporate databases offer vast volumes of data that, despite being considered of little value due to the lack of quality control, could be valuable for improving AI models' accuracy and applicability. In fact, despite reasonable concerns on data quality issues, data-driven architectures demonstrated the ability to learn from noisy data,[21] thus making such proprietary data sets a very appealing source of chemical knowledge in the customization of AI models to different areas of chemistry.

Recently, there have been a few examples from academia and industry reporting the use of proprietary data sets for exploring performance improvements of AI systems. For instance, Stanley et al. presented a few-shot learning data set to build

Received: June 6, 2023
Revised: October 9, 2023
Accepted: October 9, 2023
Published: October 27, 2023



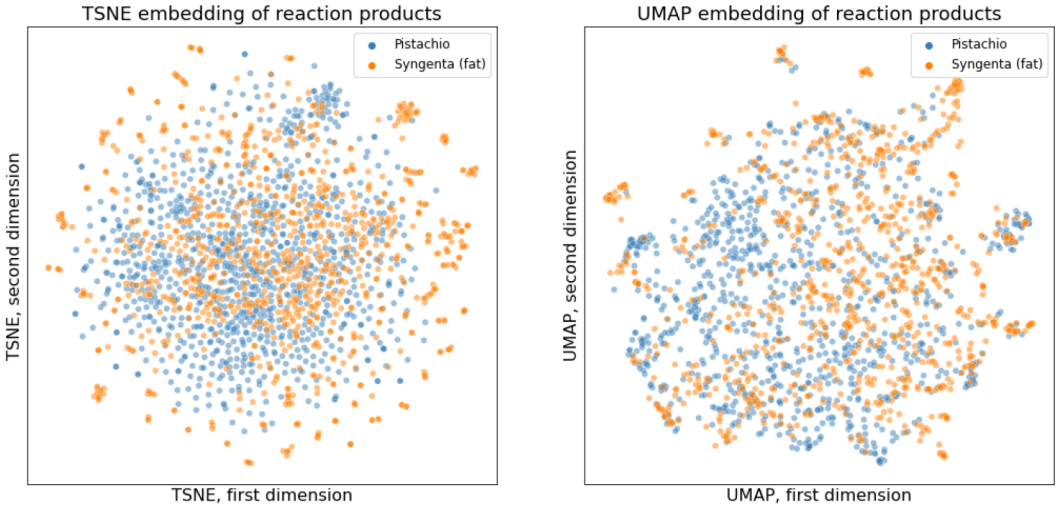


Figure 1. t-SNE (left) and UMAP (right) projections of the products for the Pistachio and Syngenta (fat) data sets. The projections were obtained after randomly sampling 1,000 products from each data set.

machine learning models in the low data regime.[23] In a similar spirit, there have been different efforts to provide unsupervised pretrained models as a basis for fine-tuning on property prediction tasks.[24,25] In de novo design, Zhavoronkov et al. combined multiple data sets and machine learning approaches to identify DDR1 kinase inhibitors.[26] It is also worth mentioning the MELLODDY (machine learning ledger orchestration for drug discovery) consortium, comprising ten pharmaceutical companies, and exploring the application of federated learning to collectively learn from each other's proprietary data.[27]

Adapting the domain of applicability of chemical reaction models by utilizing out-of-distribution data has challenges of its own, some of which have been addressed in recent studies. For instance, Pesciullesi et al. demonstrated that the Molecular Transformer architecture[16] (a language model) can be used to customize reaction outcome prediction models.[18] In that work, a data set of 25 000 reactions from the field of carbohydrate chemistry was combined with a data set of patent reactions to improve the accuracy of predicted carbohydrate products. The authors showed that a model trained in a transfer learning setting significantly outperformed both a model trained on patent data only and a model trained on carbohydrate reactions only. Similar approaches have been followed for other classes of reactions.[28−31]

For organizations working on diverse chemical compounds and processes, specializing models to a subset of chemical reaction space is not enough. It is critical for models to be accurate across the entire breadth of the chemistry of interest. This is one example in which proprietary data sets have substantial value. Still, the main barrier for organizations to exploit their data for retraining chemical reaction models is the lack of robust guidelines.

In this work, we use language models, specifically the Molecular Transformer,[16] to explore the chemical knowledge potential of proprietary data sets by retraining reaction outcome prediction and single-step retrosynthesis models. We compare models retrained following multidomain and fine-tuning strategies to baselines trained solely on patent or proprietary data. We then analyze and discuss the differences between the predictions of the different models before providing a list of best practices for retraining. We share this precompetitive experience to simplify the adoption of chemical language models across a wide range of industries.

2. RESULTS AND DISCUSSION

2.1. Data. The proprietary data set studied in this work consists of a random subset of 356,059 reactions from electronic lab notebooks collected at Syngenta. For the fine-tuning and multidomain experiments, over 2 million patent reaction records, obtained from Pistachio,[32] were also incorporated during model training. We represent the compounds in the simplified molecular-input line-entry system (SMILES) notation,[33,34] and combine them into strings in the reaction SMILES format to designate chemical transformations. While chemists often report reactions only in terms of reactants and products, it is common practice for chemical language models to consider all the reagents as well[5] (see Appendix A for a discussion of reaction roles). Therefore, special effort was paid during the collection of the ELN data to gather this information. To evaluate the effect of the presence or absence of reagents, we compiled two versions of the ELN data. The "slim" data set comprises reactants only, while the "fat" data set also keeps the reagents as precursors. The interested reader can find more information about the data sets and their processing in Appendix A.

To illustrate some of the differences between the patent and proprietary ELN data sets, we report in Figure 1 the chemistry covered by both data sets using t-SNE[35] and UMAP[36] projections of the reaction products for a random sample of 1000 records from both data sets. While some reactions from both sets occupy the same space, it is apparent that the ELN reactions cover compounds different from the patent data.

In addition to focusing on the reaction products, we analyzed the chemical transformations present in the data sets. First, we predicted[20] the RXNO[37] superclass of all reactions. Table 1 shows the frequencies of the different reaction categories in the data sets. The proprietary data set contains slightly more unrecognized reactions than the patent one. The proportion of reactions in some reaction categories varies considerably between the slim and fat data sets. This is caused in part by the reactions not included in the slim data set due to validity filters; for instance, reduction, oxidation, or deprotection reaction usually have a single precursor in the slim data

https://doi.org/10.1021/acs.chemmater.3c01406
Chem. Mater. 2023, 35, 8806−8815

Table 1. Reaction Class Frequency (in %) in Terms of RXNO Superclasses[37,38] for the Different Data Sets

Reaction category	Patent	Slim	Fat
Heteroatom alkylation and arylation	19.5	21.2	18.4
Acylation and related processes	16.8	26.5	21.1
Carbon−carbon bond formation	8.8	10.4	9.0
Heterocycle forming reactions	3.1	3.5	3.2
Protection reactions	1.4	0.9	1.0
Deprotection reactions	13.7	2.9	7.2
Reductions	5.5	1.7	3.6
Oxidations	2.3	1.8	2.7
Functional group interconversion	7.3	6.5	8.2
Functional group addition	2.3	2.1	2.7
Resolution reactions	0.9	0.1	0.5
Unrecognized	18.3	22.2	22.4

set (since the reagents and solvents are not included), and are discarded as they do not fulfill the minimum of two precursors required for model training. As a whole, the class distribution is similar in the patent and proprietary data sets. The proprietary data is characterized by fewer protection, deprotection and reduction reactions, while there is a larger proportion of acylations, oxidations, and functional group interconversions.

To obtain more insight into the actual transformations of the different data sets and evaluate the reaction diversity, we extracted the reaction templates for the train splits of all three data sets (see Appendix B for details on the procedure). The number of extracted templates for the different data sets and their overlap can be seen in Figure 2. While for the patent data set there are, on average, 14.5 reactions per template, this number is much lower for the proprietary data set, with 3.3 and 4.4 for the slim and fat data sets, respectively. The overlaps between the templates obtained from the patent and the proprietary data sets are relatively small, with 21% and 26% of the proprietary templates already present in the patent data set. This indicates that the majority of the proprietary reactions are not directly accounted for in the patent reactions. Accordingly, the proprietary data set brings substantial novelty and diversity, especially when considering the overall sizes of the data sets.

2.2. Models. We trained and compared multiple models for both data sets (slim and fat) and both tasks (reaction outcome prediction and single-step retrosynthesis). In the reaction outcome prediction task (also known as *forward reaction prediction*), a model predicts the reaction product for a set of precursors. The single-step retrosynthesis task addresses the opposite problem, where a set of precursors is predicted for a given product. As a baseline, we consider models[21] trained on patent data obtained from the Pistachio database.[32] We then study three types of models that take the proprietary reaction data into account. First, we trained a model from scratch on the proprietary reaction data only. Second, we fine-tuned models on the proprietary reaction data, starting from the patent model. Different learning rates were applied, leading to multiple fine-tuned models. Third, we trained models following a multidomain approach (sometimes also called "multitask" in this context[18]), where the patent and proprietary data sets are used simultaneously to train a model from scratch.[18] We selected different sets of weights for multidomain training, leading to multiple multidomain models.

Details on the model architecture and training process can be found in Appendix C.

2.3. Metrics. *2.3.1. Reaction Outcome Prediction.* We analyzed the reaction outcome prediction models in terms of top-*N* accuracy (see Appendix D for a formal definition). We show the results for the validation split of the proprietary data set for a selection of models in Tables 2 and 3. We also show the top-1 accuracy on the validation split of the patent data set. The results for all of the trained models can be found in Tables 9 and 10 in Appendix E.

Table 2. Reaction Outcome Prediction Metrics on the Models Trained with the Slim Data Set[a]

Model	Accuracy (%)			
	top-1	top-2	top-5	top-1 (patents)
Proprietary data only	79.9	83.8	86.3	29.2
Fine-tuning	84.9	88.8	91.0	66.2
Multidomain	**85.1**	**89.2**	**91.2**	68.6
Patent data only	70.8	75.5	78.2	**68.8**

[a]The results for all of the experiments can be found in Appendix E (Table 9).

Table 3. Reaction Outcome Prediction Metrics on the Models Trained with the Fat Data Set[a]

Model	Accuracy (%)			
	top-1	top-2	top-5	top-1 (patents)
Proprietary data only	80.6	84.9	87.2	35.8
Fine-tuning	**85.8**	**90.1**	**92.0**	65.5
Multidomain	84.7	89.0	91.0	68.6
Patent data only	73.0	77.8	80.5	**68.8**

[a]The results for all the experiments can be found in Appendix E (Table 10).

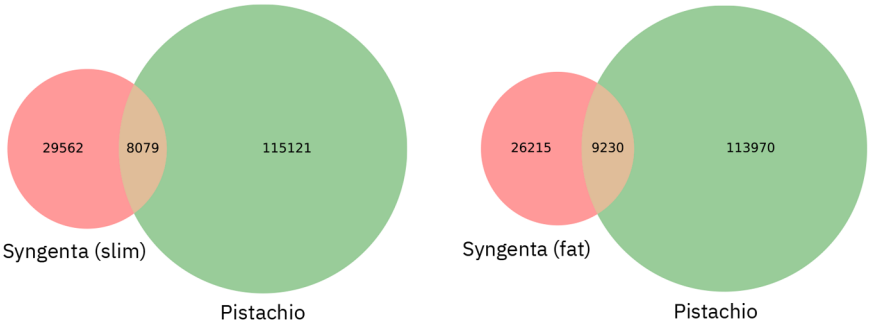


Figure 2. Venn diagrams for the number of reaction templates obtained from the slim (left) and fat (right) proprietary data sets and the patent data set.

https://doi.org/10.1021/acs.chemmater.3c01406
Chem. Mater. 2023, 35, 8806−8815

The trends in model performance are similar for models trained on both data set versions (slim and fat). The patent model reaches the lowest accuracy, followed by the model trained on proprietary data only. It is, in all cases, beneficial to combine patent and proprietary data: the fine-tuning and multidomain approaches consistently reach higher accuracies. Both approaches perform equally well on the proprietary data set but differ when assessed on patent data. With the fine-tuning approach, the accuracy on patent data is very sensitive to the learning rate, while multidomain models are comparatively robust with respect to changes in the weights of the underlying data sets (see Tables 9 and 10 in Appendix E). The best multidomain model nearly reaches the same accuracy on patent data as the model trained exclusively on patent data. In other words, both strategies lead to similar performances on the target proprietary data set, but the models trained in a multidomain setting will generalize better to other reactions.

In order to grasp the significance of the absolute accuracies, we contextualize them by comparing them to typical levels of accuracy reported in the existing literature. In the case of noisy patent data sets, it is common to achieve top-1 accuracies ranging from 60% to 70%.[16,39] Conversely, when dealing with cleaner reaction data sets like USPTO-MIT,[40] models tend to achieve higher top-1 accuracies, often reaching around 90%.[16,39,40] Consequently, the accuracy results attained by the models trained in this study align with the anticipated range of values for both relatively noisy data (patents) and clean data (proprietary).

2.3.2. Single-Step Retrosynthesis. We analyzed the single-step retrosynthesis models in terms of top-*N* and round-trip accuracy (see Appendix D for the definition of these metrics). In short, the top-*N* accuracy relates to the ability of the model to predict an exact match to the ground truth among the first *N* predictions, while the round-trip accuracy assesses whether the predicted precursors are valid (not necessarily matching the ground truth) for the target compound. We show the results for the validation split of the proprietary data set for a selection of models in Tables 4 and 5. The results for all models can be found in Tables 11 and 12 in Appendix E.

Table 4. Single-Step Retrosynthesis Metrics on the Models Trained with the Slim Data Set[a]

	Accuracy (%)			Round-trip (%)
Model	top-1	top-5	top-1 (patents)	top-1
Proprietary data only	43.6	59.1	1.3	83.2
Fine-tuning	43.6	60.3	3.2	90.6
Multidomain	**48.6**	**65.5**	10.0	**91.7**
Patent data only	10.3	24.1	**10.8**	87.7

[a]The results for all of the experiments can be found in Appendix E (Table 11).

The round-trip accuracy values rely on the ability to determine the reaction product for a given set of precursors. As a proxy for the (unknown) truth, we must rely on a reaction outcome prediction model to enable the calculation of this metric. For this purpose, we selected the best-performing reaction outcome prediction model trained with the multidomain approach (one for the slim and one for the fat data set) and retrained it on a training set comprising the original training and validation splits.

Table 5. Single-Step Retrosynthesis Metrics on the Models Trained with the Fat Data Set[a]

	Accuracy (%)			Round-trip (%)
Model	top-1	top-5	top-1 (patents)	top-1
Proprietary data only	32.2	44.4	0.8	82.9
Fine-tuning	33.9	47.9	2.0	90.7
Multidomain	**35.6**	**50.7**	9.2	**92.1**
Patent data only	5.0	12.1	**10.8**	86.1

[a]The results for all the experiments can be found in Appendix E (Table 12).

The trends for the single-step retrosynthesis models are slightly different from those for the reaction outcome prediction models. As expected, the model trained on patents reaches the lowest accuracy of all the models. The model trained from scratch on the proprietary data set and the best fine-tuning models reach similar top-*N* accuracy values, while the fine-tuned models are better in terms of round-trip accuracy. The multidomain models reach significantly higher top-*N* accuracies than the other models, also when assessed on patent data. In addition, they are more robust in terms of changes in the hyperparameters (Tables 11 and 12 in Appendix E).

In all cases, the accuracy reached by the models trained on the slim data sets is higher than that for the models trained on the fat data sets. This is consistent with the fact that the underlying data set contains fewer precursors. Therefore these models have a higher chance of predicting an exact match with the ground truth. In practice, the round-trip accuracy is more relevant.[5] There, we observed very similar values for the slim and fat versions of the models when assessed on the proprietary data set. The round-trip accuracy obtained on the patent data set (Tables 11 and 12) is higher for the fat models, indicating a better generalization ability for these models than for the slim models.

We hypothesized that the better generalization ability of the fat models might be the consequence of underrepresented reaction classes in the slim data sets caused by the validity filters applied during data preprocessing (see Appendix A). To support this hypothesis, we examined the set of reactions that were discarded in the slim, but not in the fat data set. In Table 13 in Appendix F, we inspect the accuracies of the selected models on reaction products that are present only in the validation split of the fat data set. While the far greater top-1 accuracy of the fat models is expected, the round-trip accuracy is more than 15% smaller for the slim models than for the fat models. This evidence demonstrates the importance of training reactivity models on reaction data listing all precursors, including solvents, reagents, and catalysts.

2.4. Analysis of Errors by the Reaction Outcome Prediction Models. We analyzed the mistakes (i.e., predictions different from the ground truth) made by the reaction outcome prediction models listed in Section 2.3. While most of the mistakes cannot be classified easily, evaluating how many cases fall into specific categories is helpful. A summary is shown in Tables 6 and 7. There, we arranged the incorrect predictions into the following categories: products with invalid SMILES syntax, with incorrect stereochemistry, with a tautomeric representation different from the ground truth (as determined from the InChI representation), with different regiochemistry (approximated

https://doi.org/10.1021/acs.chemmater.3c01406
Chem. Mater. 2023, 35, 8806−8815

Table 6. Forward Prediction Mistakes of the Models on the Slim Data Set (in % of the Full Data Set)

Error	Patent data only	Proprietary data only	Fine-tuning	Multidomain
Total incorrect	29.2	20.1	15.1	14.9
Invalid SMILES	1.1	1.3	0.1	0.2
Stereochemistry	5.1	2.4	2.5	2.5
Tautomers	1.1	0.6	0.6	0.7
Regiochemistry	3.1	1.9	1.4	1.3
No transformation	1.2	1.3	0.2	0.2

Table 7. Forward Prediction Mistakes of the Models on the Fat Data Set (in % of the Full Data Set)

Error	Patent data only	Proprietary data only	Fine-tuning	Multidomain
Total incorrect	27.0	19.4	14.2	15.3
Invalid SMILES	0.8	0.9	0.1	0.3
Stereochemistry	4.6	2.4	2.3	2.5
Tautomers	1.2	0.7	1.0	1.0
Regiochemistry	2.8	1.8	1.1	1.4
No transformation	0.9	1.2	0.3	0.3

by comparison of the molecular formula), or with no transformation (product identical to one of the precursors).

For the best models (fine-tuning and multidomain), very few predictions contain invalid SMILES, with well under 1% of all the predictions. Selectivity mistakes (stereochemistry and regiochemistry) account for over a quarter of the incorrect predictions. It is noteworthy that considering patent data in addition to the proprietary data leads to an important decrease in invalid SMILES or predictions identical to those of the precursors. This highlights the value of the additional data for learning the syntax rules of the SMILES notation, as well as an initial understanding of chemical reactions.

Some of the incorrect predictions refer to predicted products represented in a tautomeric form different from the ground truth. These cases account for 0.6% to 1.3% of the predictions, depending on the model. Given that tautomers are interconverted into each other in solution, the model predictions are actually chemically sound in such cases despite being different from the ground truth.

When the models predict an incorrect product as the top-1 prediction, one can inspect how often the second model prediction is identical with the ground truth for the error categories shown above. These values are shown in Tables 14 and 15 in Appendix G. On average, the second model prediction will be correct in up to 30% of the cases, this value being higher for models combining patent and proprietary data. Among the different error categories, the models are more likely to predict the correct product as the second suggestion if the error was related to stereochemistry, tautomery, or regiochemistry.

2.5. Insights and Guidelines. The experiments carried out in this study provide useful insights into tailoring chemical language models trained on publicly available data to out-of-distribution data sets. Importantly, they allow us to formulate guidelines that can expedite the customization process for new data sets.

First, whenever possible, one should mix proprietary data with publicly available data. Combining proprietary and public data can help models learn the syntax of the chemical language and the chemical transformations that occur more effectively.

Second, retraining efforts should be concentrated on multidomain training approaches, as they lead to more robust models than fine-tuning. While both approaches delivered similar results when evaluated on the proprietary data set, multidomain learning is less sensitive to the choice of hyperparameters. Also, it generalizes better to other reactions, such as the ones present in the patent data set. We expect the fine-tuning approach to perform better only in the low-data regime. In either case, it is advisable to train multiple models with different hyperparameters.

Third, we recommend collecting and reporting all of the precursors in the training data sets. This may require special effort to ensure that reagents, solvents, and catalysts are included when reporting reactions in ELNs. While doing so does not lead to major differences for the reaction outcome prediction model, we observed that models trained on complete reactions generalize better for the single-step retrosynthesis domain.

We have made available a GitHub repository[41] with code to facilitate model training based on these recommendations.

3. CONCLUSIONS

In recent years, language models for chemical reactivity have demonstrated the potential to unlock hidden knowledge in chemical industries. Their usefulness is beyond discussion, having been adopted in several academic and industrial applications. In this work, we studied models for reaction outcome prediction and single-step retrosynthesis and how they can be tailored to proprietary data sets. This enabled us to provide guidelines for the fast domain adaptation of such models to out-of-distribution data sets.

For instance, we highlighted the usefulness of integrating public data sets in the training pipeline. This makes the models more robust and improves their generalization ability, even if overlap with the reaction space of interest is limited. We also showed the advantages of multidomain training as compared to fine-tuning approaches, including, for instance, a lower sensitivity to the choice of hyperparameters. Last, we highlighted the importance of capturing the full reaction specification, even when chemists may deem the indication of solvents and reagents superfluous or unnecessary.

In summary, the current work and the guidelines derived therein show that proprietary reaction data sets can easily be exploited for customizing the reaction models. We are confident these guidelines will be useful to reach a higher level of performance with chemical language model customization while reducing the number of training strategies experiments and boost the adoption of such models in academia and industry.

APPENDIX A: DATA

Reaction Roles

When working with digitally stored reactions, it is useful to differentiate between the reactants and the reagents being used. In this work, we consider reactants to be precursors that contribute at least one heavy (non-hydrogen) atom to the product. All other precursors (including solvents and catalysts) are considered to be reagents. It is also helpful to distinguish between products, which refer to the desired compound

https://doi.org/10.1021/acs.chemmater.3c01406
Chem. Mater. 2023, 35, 8806−8815

produced in a reaction, byproducts, such as condensation water, and side products.

Proprietary Data Set

As a data set, we selected a random subset of 356,059 reactions from electronic lab notebooks (ELNs) collected at Syngenta. We extracted the compounds in the simplified molecular-input line-entry system (SMILES) notation[33,34] and combined them into reaction SMILES strings for processing.

Data Preprocessing

Both proprietary data sets (slim and fat) underwent a series of transformations and validity filters. The transformations comprise canonicalization of all the compounds (with RDKit[42]), alphabetic sorting of the precursors, removing duplicate molecules, as well as removing duplicate reaction SMILES. The validity checks discard reactions that do not fulfill a set of criteria based on the number of compounds, the formal charges, or the presence of elements missing from the precursors (see below for more details). After the corresponding processing, 136,791 and 175,762 were obtained for the slim and fat data sets, respectively. The difference in size between the two versions is due to the larger number of slim reactions failing the validity checks. The reactions were split into training, validation, and test sets of relative sizes 90%, 5%, and 5%. The splitting procedure ensured that no product SMILES was present in more than one split.

Filtering of Reactions

During data preprocessing, the series of filters listed in Table 8 was applied to the reaction SMILES present in the data set.

Table 8. Constraints Placed on Reaction SMILES during Preprocessing

Constraint	Value
Minimum number of precursors	2
Maximum number of precursors	10
Maximum number of precursor tokens	300
Minimum number of products	1
Maximum number of products	1
Maximum number of product tokens	200
Maximum absolute formal charge	2

The reactions not passing all the filters were discarded. Reactions for which the product contains an atom type not present in the precursors were also discarded.

Patent Data Set

For the fine-tuning and multidomain experiments, reaction records from patents were incorporated during model training. These reactions were obtained from Pistachio,[32] which provides reactants, reagents, and products in SMILES format. The data set was processed according to the procedure described in ref 21 to provide sets of 1.78M, 0.13M, and 0.13M reaction SMILES strings for training, validation, and testing, respectively. Note that the splitting procedure ensured that reactions (from any of the considered data sets) associated with the same product SMILES ended up in the same split (i.e., no information leakage).

APPENDIX B: REACTION TEMPLATE EXTRACTION

The reaction templates for both proprietary data sets (slim and fat) and for the patent data set were extracted with RDChiral.[43] As the template extraction requires knowledge of the atom-to-atom mapping, we first applied NameRXN[44] to all the reactions. For the reactions for which NameRxn could not determine the atom-to-atom mapping, we applied RXNMapper instead.[45] The reactions for which RDChiral could not identify a template were ignored.

APPENDIX C: METHODS

Model Architecture

The forward and single-step retrosynthesis models are based on the Molecular Transformer architecture.[5,16] They follow an encoder−decoder, transformer-based, sequence-to-sequence formulation. The input and output of the model are tokenized versions of SMILES strings describing the precursors or products. Forward models convert precursors SMILES into products SMILES, while single-step retrosynthesis models convert products SMILES into precursors SMILES.

The transformer model is implemented with the `OpenNMT-py` library.[46,47] The standard transformer implementation is applied with the following changes: the parameter `layers` is set to 4, `rnn_size` to 384, `word_vec_size` to 384, `max_generator_batches` to 32, `accum_count` to 4, and `label_smoothing` to 0.

Patent-Only Model

The baseline model (trained on patent data from Pistachio) was obtained from ref 21.

Proprietary-Only Model

For the model trained from scratch on proprietary data, an initial learning rate of 2.0 is used, with 250,000 training steps.

Fine-Tuning

For the fine-tuning approach, the model weights are initialized to the ones from the patent-only model. We then train the model for 100,000 steps on the proprietary data set, varying the starting learning rates (see Appendix E).

Multidomain Learning

For the models trained in a multidomain fashion, both data sets are seen during training. The relative frequency at which the samples from each data set are taken into account is specified by the respective data set weight (a hyperparameter). An initial learning rate of 2.0 is used, with 250,000 or 500,000 training steps (see Appendix E).

APPENDIX D: METRICS

Top-N Accuracy

The top-N accuracy a_N quantifies the fraction of reactions in the test set for which the ground truth value is included in the top N predictions of the model. It is given by

$$a_N = \frac{1}{M}\sum_{i}^{M} \text{included}_i \tag{1}$$

with

$$\text{included}_i = \max\left(1, \sum_{j}^{N} v_{ij}\right) \tag{2}$$

$$v_{ij} = \begin{cases} 1 & \text{if can}(\text{pred}_{ij}) = \text{can}(\text{gt}_i) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where M is the number of reactions in the test set, gt_i is the ground truth value for reaction i of the test set, gt_{ij} is the jth prediction (ordered by confidence) for the reaction i of the test

https://doi.org/10.1021/acs.chemmater.3c01406
Chem. Mater. 2023, 35, 8806−8815

set, and can(x) is the canonical representation of a SMILES string x calculated with RDKit.[42]

The top-N accuracy can be calculated for both forward prediction and single-step retrosynthesis models.

Round-Trip Accuracy

The round-trip accuracy b_N quantifies the percentage of valid retrosynthetic suggestions when considering the top N predictions for each molecule. It is given by

$$b_N = \frac{1}{NM} \sum_{i}^{M} \sum_{j}^{N} r_{ij} \tag{4}$$

with

$$r_{ij} = \begin{cases} 1 & \text{if can(forward(pred}_{ij})) = \text{can(product}_i) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where M is the number of reactions in the test set, forward(x) is a function predicting the product for a given set of precursors x with a previously trained reaction outcome prediction model, product$_i$ is the ground truth product of reaction i in the test set, and pred$_{ij}$ represent the precursors for jth prediction (ordered by confidence) by the retrosynthesis model to obtain product$_i$.

APPENDIX E: RESULTS FOR ALL THE EXPERIMENTS

Here, we provide the full results for all the trained models. The results for the reaction outcome prediction models can be found in Tables 9 (slim) and 10 (fat), and the results for the single-step retrosynthesis models are presented in Tables 11 (slim) and 12 (fat). The models prepended by an asterisk are the ones included in the tables of the main text.

Table 9. Reaction Outcome Prediction Metrics on the Models Trained with the Slim Data Set[a]

Model	Accuracy (%)			
	top-1	top-2	top-5	top-1 (patents)
*Proprietary data only (250k steps)	79.9	83.8	86.3	29.2
Fine-tuning (LR = 0.02)	84.6	88.6	90.7	68.6
*Fine-tuning (LR = 0.06)	84.9	88.8	91.0	66.2
Fine-tuning (LR = 0.2)	84.9	88.9	90.7	63.0
Fine-tuning (LR = 0.6)	84.2	88.1	90.2	57.8
Multidomain (w_p = 1, w_s = 1, 250k steps)	84.9	89.0	90.9	67.4
Multidomain (w_p = 1, w_s = 1, 500k steps)	85.2	89.0	91.2	68.3
Multidomain (w_p = 1, w_s = 2, 250k steps)	84.9	89.0	90.9	66.8
Multidomain (w_p = 1, w_s = 2, 500k steps)	85.1	89.1	90.8	67.5
Multidomain (w_p = 2, w_s = 1, 250k steps)	85.2	89.3	91.2	68.2
*Multidomain (w_p = 2, w_s = 1, 500k steps)	85.1	89.2	91.2	68.6
*Patent data only	70.8	75.5	78.2	68.8

[a]LR indicates the value of the learning rate for the fine-tuning experiments. For the multidomain experiments, w_p and w_s indicate the weight of the patent and of the proprietary data sets, respectively.

Table 10. Reaction outcome prediction metrics on the models trained with the fat data set[a]

Model	Accuracy (%)			
	top-1	top-2	top-5	top-1 (patents)
*Proprietary data only (250k steps)	80.6	84.9	87.2	35.8
Fine-tuning (LR = 0.006)	82.6	87.5	90.1	69.4
Fine-tuning (LR = 0.02)	84.6	89.2	91.4	67.7
*Fine-tuning (LR = 0.06)	85.8	90.1	92.0	65.5
Fine-tuning (LR = 0.2)	85.8	89.7	91.9	62.6
Fine-tuning (LR = 0.6)	84.9	89.3	91.5	58.2
Multidomain (w_p = 1, w_s = 1, 250k steps)	85.0	89.1	91.3	67.1
Multidomain (w_p = 1, w_s = 1, 500k steps)	85.1	88.9	91.0	68.1
Multidomain (w_p = 1, w_s = 2, 250k steps)	84.8	89.0	91.1	65.7
Multidomain (w_p = 1, w_s = 2, 500k steps)	84.8	89.2	91.1	67.5
Multidomain (w_p = 2, w_s = 1, 250k steps)	85.0	88.8	91.2	67.9
*Multidomain (w_p = 2, w_s = 1, 500k steps)	84.7	89.0	91.0	68.6
*Patent data only	73.0	77.8	80.5	68.8

[a]LR indicates the value of the learning rate for the fine-tuning experiments. For the multi-domain experiments, w_p and w_s indicate the weight of the patent and of the proprietary data sets, respectively.

Table 11. Single-Step Retrosynthesis Metrics on the Models Trained with the Slim Data Set[a]

Model	Accuracy (%)				Round-trip (%)	
	top-1	top-2	top-5	top-1 (patents)	top-1	top-1 (patents)
*Proprietary data only (250k steps)	43.6	51.8	59.1	1.3	83.2	46.7
Fine-tuning (LR = 0.02)	36.6	46.4	54.9	5.0	90.4	84.0
Fine-tuning (LR = 0.06)	41.5	51.0	58.8	4.2	91.2	80.9
*Fine-tuning (LR = 0.2)	43.6	53.5	60.3	3.2	90.6	76.2
Fine-tuning (LR = 0.6)	43.3	52.0	58.8	2.2	89.3	67.7
Multidomain (w_p = 1, w_s = 1, 250k steps)	48.3	58.0	66.2	8.9	92.0	84.5
*Multidomain (w_p = 1, w_s = 1, 500k steps)	48.6	57.8	65.5	10.0	91.7	84.9
Multidomain (w_p = 1, w_s = 2, 250k steps)	48.3	57.9	64.9	7.7	91.7	86.2
Multidomain (w_p = 1, w_s = 2, 500k steps)	48.8	58.1	65.6	8.9	91.6	82.7
Multidomain (w_p = 2, w_s = 1, 250k steps)	46.4	56.8	65.6	9.6	92.3	83.8
*Patent data only	10.3	15.9	24.1	10.8	87.7	90.1

[a]LR indicates the value of the learning rate for the fine-tuning experiments. For the multidomain experiments, w_p and w_s indicate the weight of the patent and of the proprietary data sets, respectively.

APPENDIX F: EVALUATION OF SINGLE-STEP RETROSYNTHESIS MODELS ON THE DIFFERENCE DATA SET

Table 13 shows the accuracy and round-trip accuracy of the slim and fat model on the difference data set from the validation split. The difference data set was obtained by

https://doi.org/10.1021/acs.chemmater.3c01406
Chem. Mater. 2023, 35, 8806−8815

Table 12. Single-Step Retrosynthesis Metrics on the Models Trained with the Fat Data Set[a]

Model	Accuracy (%)				Round-trip (%)	
	top-1	top-2	top-5	top-1 (patents)	top-1	top-1 (patents)
*Proprietary data only (250k steps)	32.2	38.6	44.4	0.8	82.9	47.9
Fine-tuning (LR = 0.006)	12.5	18.2	25.7	6.8	89.2	90.1
Fine-tuning (LR = 0.02)	18.2	24.5	32.7	6.0	90.8	89.1
Fine-tuning (LR = 0.06)	24.7	31.8	39.8	4.9	90.8	87.5
Fine-tuning (LR = 0.2)	31.0	38.9	46.4	3.1	91.6	83.6
*Fine-tuning (LR = 0.6)	33.9	41.4	47.9	2.0	90.7	78.2
Multidomain (w_p = 1, w_s = 1, 250k steps)	34.7	42.2	49.6	8.3	92.2	86.8
*Multidomain (w_p = 1, w_s = 1, 500k steps)	35.6	43.6	50.7	9.2	92.1	87.0
Multidomain (w_p = 1, w_s = 2, 250k steps)	34.6	42.4	49.1	7.2	90.8	84.7
Multidomain (w_p = 1, w_s = 2, 500k steps)	35.5	42.6	49.4	8.1	90.2	82.8
Multidomain (w_p = 2, w_s = 1, 250k steps)	31.8	39.8	47.7	9.4	92.8	87.8
*Patent data only	5.0	7.7	12.1	10.8	86.1	89.9

[a]LR indicates the value of the learning rate for the fine-tuning experiments. For the multidomain experiments, w_p and w_s indicate the weight of the patent and of the proprietary data sets, respectively.

Table 13. Top-1 Accuracy and Top-1 Round-Trip Accuracy of the Selected Models (Slim and Fat Versions) Applied to Products from the Validation Split of the Fat Data Set That Are Not Present in the Slim Data Set

Model	Accuracy (%)		Round-trip (%)	
	top-1 (slim)	top-1 (fat)	top-1 (slim)	top-1 (fat)
Proprietary data only	1.9	28.4	56.0	75.6
Fine-tuning	3.1	32.3	67.5	86.5
Multidomain	3.7	33.4	71.1	88.1
Patent data only	5.6	5.6	74.5	77.6

Table 14. Percentage of Correct Second Product Predictions When the First Prediction Is Incorrect, by Error Type, for the Slim Data Set

Error	Patent data only	Proprietary data only	Fine-tuning	Multidomain
All error categories (average)	16.1	19.4	26.3	27.9
Invalid SMILES	20.5	13.3	16.7	15.4
Stereochemistry	16.6	38.3	55.9	50.9
Tautomers	34.6	23.1	41.5	33.3
Regiochemistry	34.0	31.6	42.6	39.6
No transformation	8.1	4.5	0.0	6.7

keeping reactions of the validation split of the fat data set that are not present in the slim data set.

APPENDIX G: CORRECT TOP-2 PREDICTIONS

Tables 14 and 15 show, for the incorrectly-predicted reactions, how frequently the second model prediction is correct, by error type.

Table 15. Percentage of Correct Second Product Predictions When the First Prediction Is Incorrect, by Error Type, for the Fat Data Set

Error	Patent data only	Proprietary data only	Fine-tuning	Multidomain
All error categories (average)	17.8	22.2	30.0	27.9
Invalid SMILES	21.9	11.1	0.0	9.7
Stereochemistry	19.6	49.8	52.6	42.7
Tautomers	45.2	28.1	44.6	53.5
Regiochemistry	32.8	43.0	39.6	40.7
No transformation	12.7	13.1	0.0	14.8

ASSOCIATED CONTENT

Data Availability Statement

The data used in this study cannot be made publicly available as it is proprietary and confidential. The code for training the reaction outcome prediction and single-step retrosynthesis models is available at https://github.com/rxn4chemistry/rxn-onmt-models. The code for computing the metrics can be found under https://github.com/rxn4chemistry/rxn-metrics.

AUTHOR INFORMATION

Corresponding Author

Teodoro Laino − *IBM Research Europe, Rüschlikon 8803, Switzerland; National Center for Competence in Research-Catalysis (NCCR-Catalysis), 8093 Zürich, Switzerland;* orcid.org/0000-0001-8717-0456; Email: teo@zurich.ibm.com

Authors

Alessandra Toniato − *IBM Research Europe, Rüschlikon 8803, Switzerland; National Center for Competence in Research-Catalysis (NCCR-Catalysis), 8093 Zürich, Switzerland;* orcid.org/0000-0002-5218-8653

Alain C. Vaucher − *IBM Research Europe, Rüschlikon 8803, Switzerland; National Center for Competence in Research-Catalysis (NCCR-Catalysis), 8093 Zürich, Switzerland;* orcid.org/0000-0001-7554-0288

Marzena Maria Lehmann − *Syngenta Crop Protection AG, Stein 4332, Switzerland*

Torsten Luksch − *Syngenta Crop Protection AG, Stein 4332, Switzerland*

Philippe Schwaller − *IBM Research Europe, Rüschlikon 8803, Switzerland; National Center for Competence in Research-Catalysis (NCCR-Catalysis), 8093 Zürich, Switzerland*

Marco Stenta − *Syngenta Crop Protection AG, Stein 4332, Switzerland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemmater.3c01406

Author Contributions

[§]A.T. and A.C.V. contributed equally to this paper.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This publication was created as part of NCCR Catalysis (Grant Number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

https://doi.org/10.1021/acs.chemmater.3c01406
Chem. Mater. 2023, 35, 8806−8815

REFERENCES

(1) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design — a Review of the State of the Art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828−849.

(2) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative Models for Molecular Discovery: Recent Advances and Challenges. *WIREs Comput. Mol. Sci.* **2022**, *12*, e1608.

(3) 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. *Chem. Sci.* **2018**, *9*, 513−530.

(4) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.

(5) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways using Transformer-based Models and a Hyper-graph Exploration Strategy. *Chem. Sci.* **2020**, *11*, 3316−3325.

(6) Molga, K.; Szymkuć, S.; Grzybowski, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54*, 1094−1106.

(7) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a Fast, Robust and Flexible Open-source Software for Retrosynthetic Planning. *J. Cheminf.* **2020**, *12*, 70.

(8) Coley, C. W.; et al. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365*, eaax1566.

(9) 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 Cent. Sci.* **2018**, *4*, 1465−1476.

(10) Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T. Inferring Experimental Procedures from Text-based Representations of Chemical Reactions. *Nat. Commun.* **2021**, *12*, 2573.

(11) Nam, J.; Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. *arXiv*, 2016-12-29, DOI: 10.48550/arXiv.1612.09529 (accessed 2023-07-11).

(12) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.

(13) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction using Neural Sequence-to-sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103−1113.

(14) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-novo Design through Deep Reinforcement Learning. *J. Cheminf.* **2017**, *9*, 48.

(15) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(16) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572−1583.

(17) Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. *J. Chem. Inf. Model.* **2020**, *60*, 1175−1183.

(18) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. Transfer Learning Enables the Molecular Transformer to Predict Regio- and Stereoselective Reactions on Carbohydrates. *Nat. Commun.* **2020**, *11*, 4874.

(19) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields using Deep Learning. *Mach. Learn.: Sci. Technol.* **2021**, *2*, No. 015016.

(20) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the Space of Chemical Reactions using Attention-based Neural Networks. *Nat. Mach. Intell.* **2021**, *3*, 144−152.

(21) Toniato, A.; Schwaller, P.; Cardinale, A.; Geluykens, J.; Laino, T. Unassisted Noise Reduction of Chemical Reaction Datasets. *Nat. Mach. Intell.* **2021**, *3*, 485−494.

(22) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale Chemical Language Representations Capture Molecular Structure and Properties. *Nat. Mach. Intell.* **2022**, *4*, 1256−1264.

(23) Stanley, M.; Bronskill, J.; Maziarz, K.; Misztela, H.; Lanini, J.; Segler, M.; Schneider, N.; Brockschmidt, M. FS-Mol: A Few-Shot Learning Dataset of Molecules, In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*; Online[December 6-14, 2021; Vanschoren, J., Yeung, S., Eds.; Curran Associates, Inc.: Red Hook, USA, 2021; Vol. *1*.

(24) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv*, 2020-10-19, DOI: 10.48550/arXiv.2010.09885 (accessed 2023-07-11).

(25) Fabian, B.; Edlich, T.; Gaspar, H.; Segler, M. H. S.; Meyers, J.; Fiscato, M.; Ahmed, M. Molecular Representation Learning with Language Models and Domain-relevant Auxiliary Tasks. *arXiv*, 2020-11-26, DOI: 10.48550/arXiv.2011.13230 (accessed 2023-07-11).

(26) Zhavoronkov, A.; et al. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038−1040.

(27) *MELLODDY consortium*. https://www.melloddy.eu (accessed 2022-07-26).

(28) Wang, L.; Zhang, C.; Bai, R.; Li, J.; Duan, H. Heck Reaction Prediction using a Transformer Model Based on a Transfer Learning Strategy. *Chem. Commun.* **2020**, *56*, 9368−9371.

(29) Wu, Y.; Zhang, C.; Wang, L.; Duan, H. A Graph-convolutional Neural Network for Addressing Small-scale Reaction Prediction. *Chem. Commun.* **2021**, *57*, 4114−4117.

(30) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting Enzymatic Reactions with a Molecular Transformer. *Chem. Sci.* **2021**, *12*, 8648−8659.

(31) Probst, D.; Manica, M.; Nana Teukam, Y. G.; Castrogiovanni, A.; Paratore, F.; Laino, T. Biocatalysed Synthesis Planning using Data-driven Learning. *Nat. Commun.* **2022**, *13*, 964.

(32) *Nextmove Software Pistachio*. http://www.nextmovesoftware.com/pistachio.html (accessed 2023-07-11).

(33) Weininger, D. SMILES a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(34) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(35) Van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(36) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, 2018-02-09, DOI: 10.48550/arXiv.1802.03426 (accessed 2023-07-11).

(37) *RSC's RXNO Ontology*. http://www.rsc.org/ontologies/RXNO/index.asp (accessed 2022-03-22).

(38) *Wikipedia: RXNO Ontology*. https://en.wikipedia.org/wiki/RXNO_Ontology (accessed 2022-08-05).

(39) Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat. Commun.* **2020**, *11*, 5575.

(40) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler−Lehman Network. In *Advances in Neural Information Processing Systems*; Long Beach, USA, December 4−9, 2017; Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, USA, 2017; Vol. *30*.

(41) *rxn-onmt-models library, version 1.0.0*. https://github.com/rxn4chemistry/rxn-onmt-models (accessed 2023-03-03).

https://doi.org/10.1021/acs.chemmater.3c01406
Chem. Mater. 2023, 35, 8806−8815

(42) Landrum, G. *RDKit: Open-source cheminformatics*, version Release 2021.03.3, Release 2021.03.3, 2021, DOI: 10.5281/zenodo.4973812. https://www.rdkit.org.
(43) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *J. Chem. Inf. Model.* **2019**, *59*, 2529−2537.
(44) *Nextmove Software NameRxn*. https://www.nextmovesoftware.com/namerxn.html (accessed 2022-08-08).
(45) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of Organic Chemistry Grammar from Unsupervised Learning of Chemical Reactions. *Sci. Adv.* **2021**, *7*, eabe4166.
(46) Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*; Vancouver, Canada, July 30−August 4, 2017; Bansal, M., Ji, H., Eds.; Association for Computational Linguistics: Vancouver, Canada, 2017; pp 67−72.
(47) *OpenNMT-py library*, version 1.2.0. https://github.com/OpenNMT/OpenNMT-py (accessed 2023-07-11).

https://doi.org/10.1021/acs.chemmater.3c01406
Chem. Mater. 2023, 35, 8806−8815