ORDerly: Datasets and benchmarks for chemical reaction data

Daniel S. Wigh
Department of Chemical
Engineering and Biotechnology
University of Cambridge, UK
dsw46@cam.ac.uk

Joe Arrowsmith
Department of Chemical
Engineering and Biotechnology
University of Cambridge, UK
ja799@cam.ac.uk

Alexander Pomberger
Department of Chemical
Engineering and Biotechnology
University of Cambridge, UK
ap2153@cam.ac.uk

Kobi C. Felton
Department of Chemical
Engineering and Biotechnology
University of Cambridge, UK
kcmf2@cam.ac.uk

Alexei A. Lapkin
Department of Chemical
Engineering and Biotechnology
University of Cambridge, UK
aal35@cam.ac.uk

Abstract

Machine learning has the potential to provide tremendous value to the chemical and material sciences by providing models that promise to save time, energy, and starting material. Model training requires large amounts of clean high-quality data, and the methodology for transforming raw data to machine learning-ready data should be robust, adaptable, and accessible. However, data is often cleaned differently for different projects using proprietary code, making it difficult to compare approaches and creating additional effort for other researchers who want to work with literature-mined data. Herein, we present ORDerly, an open-source Python package with a novel benchmark for reaction data and a highly customizable pipeline for cleaning chemical reaction data stored in accordance with the Open Reaction Database (ORD) schema. ORDerly contains standard cleaning operations, such as frequency filtering and canonicalization checks, in addition to chemically-informed assignment of reaction roles using atom mapping, bespoke name resolution, and reproducible open-source benchmark generation. We use ORDerly to generate a machine learning-ready benchmark dataset for the prediction of reaction conditions, and through extensive analysis, we find the aforementioned cleaning steps to be essential to provide a high quality dataset for machine learning. In particular, we show that datasets missing key cleaning steps can lead to silently overinflated performance metrics. We then demonstrate that ORDerly can be used in an end-to-end pipeline that goes from raw data to a reaction condition prediction model in less than a day. With this customizable open-source solution for cleaning and preparing chemical reaction data, ORDerly is poised to push forward the boundaries of artificial intelligence applications in chemistry by providing a novel benchmark for chemical reaction conditions, and a data pipeline for researchers in the chemical sciences to leverage large reaction datasets.

Preprint. Under review.

1 Introduction

Advancements in chemistry and material sciences hinge on the availability of high-quality chemical reaction data, and the advent of machine learning (ML) for science has highlighted the value that data can bring to chemistry. One important application is in the pharmaceutical industry, where figuring out *how* to make novel molecules remains a significant bottleneck, causing delays in the "make" step of the "design, make, test" cycle [11]. Making a molecule (product) includes predicting the reaction pathway (retrosynthesis) and the reaction conditions (e.g. solvents and reagents). ML approaches can provide tools to assist in all of these tasks, and for reaction condition prediction (the main intended application of the benchmark presented here), there already are a number of ML tools available [8, 22, 23, 19].

However, building reaction prediction tools requires access to large datasets to train predictive models. Historically, this data has been acquired through commercial databases such as Reaxys™[6]. The advantage of these databases is both the scale of the datasets available (often millions of reactions) and the annotation already completed by the publishers. Yet, these datasets are not freely available to ML practitioners, stymieing advances in reaction condition prediction in both academia and industry.

Recently, efforts have been made to create openly-accessible databases for chemistry data. In particular, the Open Reaction Database (ORD) [14] is promising due to its exhaustive schema for describing chemical reaction data and breadth of data already incorporated. Yet, many of the datasets in ORD (license: Creative Commons Attribution Share Alike 4.0 International) require further processing before they can be used in ML pipelines, preventing practical use. This is especially true for the largest dataset in ORD extracted from the US patent literature (the "USPTO dataset"). In this work, we endeavor to close this gap.

We present ORDerly, a new framework for extracting and cleaning data from ORD and an accompanying benchmark for reaction condition prediction. The motivation behind ORDerly is threefold: (1) To create a pipline for the creation of ML-ready datasets from the growing number of chemical reactions stored in the ORD format, both publicly and privately; (2) to determine a set of evidence-based recommendations for the hyperarameters of this cleaning pipeline; and (3) to provide a benchmark for reaction condition prediction. By offering an open-source and customizable solution for cleaning chemical reaction data, ORDerly aims to contribute to the development of advanced ML models in chemistry and materials science. In this work, we focus on cleaning data extracted from the US Patent literature [18] (license: CC0), but the methods proposed could be equally valuable for any ORD dataset. Furthermore, while our focus is primarily on reaction conditions, the tools here could be also used for retrosynthesis[24] and forward prediction tasks[4].

The remainder of the paper proceeds as follows. After reviewing related work in section 2, we present the extraction and cleaning methodology of ORDerly in section 3. This is followed by experimental validation of the cleaned benchmark dataset using a previously published neural network architecture [8]. We demonstrate how missing key cleaning steps results in a dataset with contamination of the

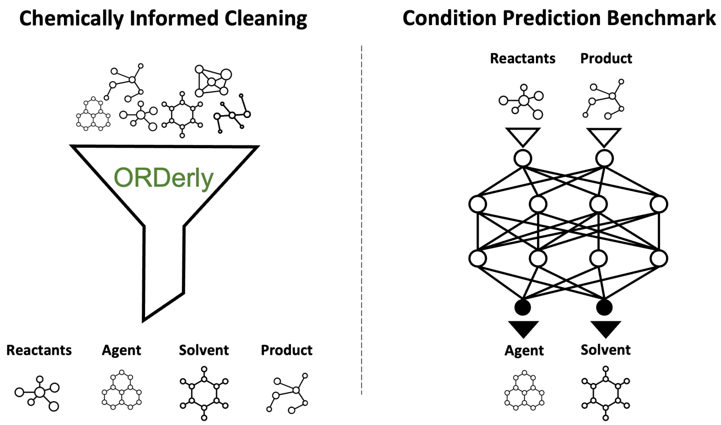


Figure 1: Overview of ORDerly.

inputs with prediction targets and thus inflates the performance metrics on the test set. We finally discuss the technical limitations of ORDerly in section 6 and present our conclusions.

2 Related Work

2.1 Relation to chemical reaction cleaning tools

Existing tools for cleaning reaction data are primarily targeted at retrosynthesis and forward prediction tasks [13, 1, 28, 9] and have somewhat limited extensibility, given that they are built to take inputs as csv files or the stationary XML files of the USPTO dataset. Furthermore, there is little to no discussion of how decisions made during cleaning (e.g. restricting the number of components in a reaction or the minimum frequency of occurrence) impact the datasets being cleaned or performance of models trained on the datasets. We believe that this is in part due to data cleaning historically being viewed as a 'low value' task, and therefore not adequately discussed and published on. In this work, we make recommendations for the available parameters in ORDerly, and thus this work also serves as a blueprint for good cleaning practices for chemical reaction data.

Common cleaning operations on chemical reaction data include: (1) Restricting reactions to being single-product and single-step (2) Ensuring that all molecules can be sanitized and canonicalized by RDKit [15] (3) Restricting the maximum number of unique catalysts, solvents, and reagents in a reaction (4) Enforcing a minimum frequency of components in the reaction to avoid rare chemicals which would cause sparsity in the one-hot encoding (OHE) (5) Sanity checking the yield ($0\% \leq yield \leq 100\%$), temperature, and pressure, (6) Removing duplicates, and finally (7) Applying a random split to create training/validation/test sets, carefully ensuring that any inputs present in the train set (i.e. reactants and products for reaction condition prediction) are not also present in the test set.

2.2 Relation to chemical reaction data benchmarks

USPTO, being the largest open-source chemical reaction dataset, has been cleaned a number of times for different learning tasks. For example, the Therapeutics Data Commons (TDC) [10] contains a number of benchmarks for chemical reaction prediction based on USPTO dataset. However, the code used to generate all of these benchmarks was not published, and furthermore, none of the benchmarks are suitable for the learning task of reaction condition prediction. The reaction condition prediction benchmark presented in this work could therefore be integrated into frameworks such as TDC as a novel benchmark.

2.3 Machine learning for reaction prediction

The outcome of a ML pipeline for reaction prediction is typically either a string representing a molecule, a vector (e.g. a one-hot encoding) or a float (e.g. yield). The SMILES string is arguably the most common representation of molecules used in ML [30], and reactions in ORD are represented either as structured inputs or with reaction SMILES. While ORDerly is focused on condition prediction, it can also prepare datasets for other tasks.

Catalyst prediction. Identification of a potent catalyst which balances selectivity and conversion is critical for most of today's organic reactions towards drug products. For the prediction task, both reactant molecules and product molecules of the reaction were used as input with the aim to predict the catalyst. Gao *et al.* [8] presented a strategy for data processing and created a catalyst prediction benchmark, though the code for cleaning was not made public.

Molecule generation. There are a near infinite number of small molecules of potential pharmaceutical relevance, and navigating this space is difficult. Algorithmic generation of novel molecules with desirable properties is therefore of significant interest, and can be achieved by training ML models on molecular characteristics from a large dataset. A few datasets and benchmarks exist, including ZINC [25], MOSES [21], and ChEMBL[20].

Retrosynthesis. The process of tracing complex molecular compounds back to their initial commercially available starting materials, retrosynthesis, is an extremely important strategy for assessing synthesizability of promising drug molecules prior to experiments. Within the retrosynthetic prediction task a complex molecule of interest is used as input and simple starting materials as well as the identified synthetic routes are obtained as output. The USPTO dataset has been used for such a prediction task using either the full [18] or partial dataset (50k datapoints) consisting of extracted, atom mapped reactions of 10 reaction types [32, 17].

Reaction outcome prediction Identification of potential reaction products (outcomes) given the reactants and conditions (input) is a fundamental challenge and can be seen as the inverse process to retrosynthesis. While pure experimental evaluation requires expert chemists and is a time intense task, reaction outcome prediction can help as a tool to improve synthesis planning. Reaction outcome prediction machine learning models can either be template-based [4] or template-free [12].

Condition prediction algorithms

While retrosynthetic prediction tools enable route planning towards a target compound as well as identification of the commercially relevant starting materials, experimental evaluation requires detailed condition information. Reaction condition recommendation using predictive machine learning is a powerful tool for bench chemists to help them with defining starting points for experimental campaigns. Typical variables to be predicted include the reagent(s), catalyst(s), solvent(s) as well as reaction temperature. Gao *et al.* [8] used approx. 10 million reactions mined from Reaxys using a combination of fingerprint and OHE representations to sequentially predict catalyst, reagent, solvent and temperature. For training and evaluating the neural network-based model the data was split randomly into train/validate/test 80/10/10 partitions, and a 50.1% top-3 accuracy of predicted components was achieved. While a detailed data cleaning pipeline was not published, their model architecture was, so it is this architecture that we use for experimental validation of the datasets generated by ORDerly. It is worth noting that conditions suggested by a reaction condition algorithm are likely to only result in a reaction with detectable yield; an optimisation campaign will often be required to achieve a high conversion [27].

3 Extraction and cleaning methodology

Data is stored in ORD in accordance with the ORD schema. This is a structured data format, and this labeling of information in the ORD files ostensibly makes it easier to extract data. However, we found that further effort is required to transform a labeled dataset into a ML-ready dataset. ORDerly is centered around a data extraction script and a data cleaning script, both of which take numerous arguments that customize the operations.

We experiment with two approaches to extracting molecule roles from ORD: trusting the labeling of the data in the input/outcome labeling in ORD (e.g., whether a molecule is a reactant) or applying chemical reaction logic to identify the role of different molecules from the reaction string. Our reaction logic identifies reactants (molecules that contribute atoms to the product(s)) and spectator molecules (molecules that do not contribute atoms to the product(s)) based on their position in the reaction SMILES string. Solvents are identified in the list of spectator molecules by cross checking against a list of solvents compiled from prior research [2, 5, 7], while all other spectator molecules are marked as agents. Note that, to ensure high data quality, only molecules with SMILES that are canonicalizable by RDKit [15] are kept.

Most chemical reactions have only two reactants and a small number of agents, so reactions with large number of participating molecules are likely due to transcription errors. Therefore, we set thresholds for the maximum number of different molecules within each reaction role. Removing reactions with excess components is done sequentially (reactants → products → solvents → agents. as in Table 1), so the threshold chosen in one filtering step will impact all downstream filtering steps. For all of the datasets, reactions with up to two reactants, one product, two solvents, and three agents were kept (see SI F.1). Finally, the sensitivity to the frequency threshold for solvents and agents to be considered rare was investigated (see SI F.2), with the minimum frequency of occurrence being set to 100.

The distribution of solvent and agent frequency is long-tailed, so strategies are needed to avoid sparsity in the OHE of solvents and agents used in the model architecture of Gao *et al.* [8] We explore two different approaches to removing rare solvents and agents: deleting reactions with rare molecules (rare→delete rxn) or keeping the reactions but mapping the rare molecules to an "other" category (rare→"other") (see Figure 2). We conduct experiments with both the rare→delete rxn and rare→"other" strategies. Deleting reactions with rare molecules may create a more cohesive

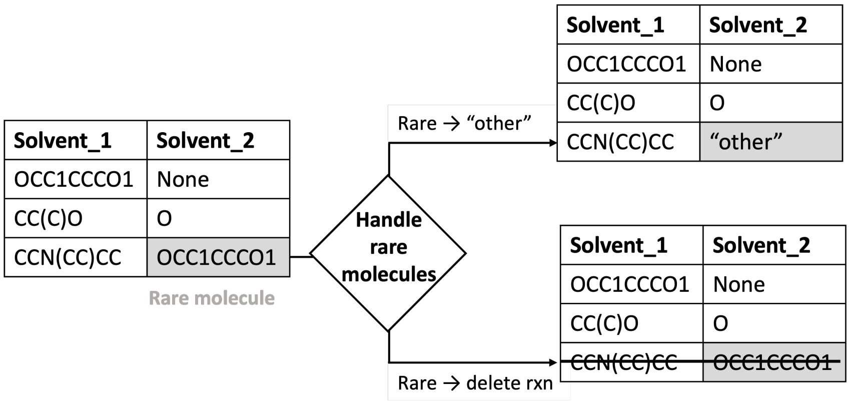


Figure 2: We present two different approaches for handling rare molecules. Rare → "other" is investigated as a strategy to avoid deleting reactions with rare molecules.

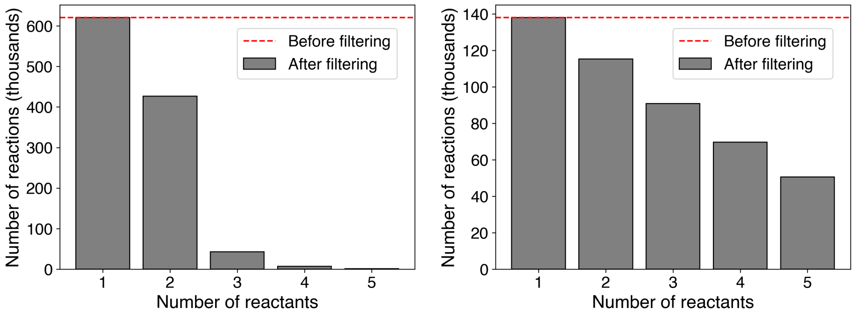
dataset by removing outliers, while renaming rare molecules "other" allows more reactions to be kept, offering more training data for the model. In total, we create four datasets for benchmarking based on two options: extracting molecules from the labeling or from the reaction string; rare→delete rxn or rare→"other" (2x2 options).

All extraction/cleaning operations described in this section were performed using a 2022 Mac Studio with an Apple M1 Max chip and 32GB memory. 1.7 million reactions from ORD belonging to the USPTO dataset were extracted to create condition prediction datasets, though of course additional value of ORDerly is unlocked as ORD grows significantly beyond USPTO (currently USPTO makes up 99% of ORD). Extracting and sanitizing the reaction components using the ORD labeling of components was slightly faster than using our custom logic applied to the reaction string, taking 28 minutes and 48 minutes, respectively, with the cleaning steps taking 6-8 minutes.

4 Dataset composition

We used ORDerly to create four datasets, to which a random train/test split was applied. The four datasets are signified by reaction role extraction method (using the reaction string or the labeling) and handling of rare molecules (rare→delete rxn or rare→"other"):

- Dataset A: Labeling, rare→"other" (76,634 reactions)
- Dataset B: Labeling, rare→delete rxn (75,033 reactions)



(a) Reaction string dataset, reactant filtering

(b) Labeling dataset, reactant filtering

Figure 3: Distribution of the number of reactants between the reaction string and labeling datasets. The labeling dataset contains more reactants per reaction on average; this may be due to agents being mislabeled as reactants.

Table 1: Number of reactions left in the USPTO dataset after each filtering step using the labeling in ORD and the reaction string. Reactions with rare molecules can be handled in two ways, as per Figure 2: either deleted (rare→delete rxn) as shown in the table below, or renamed to other (rare→"other"), in which case the dataset size would not change during the final cleaning step.

	labeling	reaction string
	rare→delete rxn	rare→delete rxn
Full dataset	1,771,032	1,771,032
Too many reactants	1,470,060	1,631,394
Too many products	1,329,399	1,593,196
Too many solvents	1,222,381	1,388,312
Too many agents	1,202,790	1,279,833
No reactants or no products	1,202,758	1,262,333
No solvents	870,888	950,189
No agents	135,139	690,234
Inconsistent yields	126,948	658,071
Dropping duplicates	76,634	392,996
Removing rare molecules	75,033	356,906

- Dataset C: Reaction string, rare→"other" (392,996 reactions)
- Dataset D: Reaction string, rare→delete rxn (356,906 reactions)

Each dataset was saved in Apache Parquet format, and has the following column groups:

- Reaction SMILES (string), is_mapped (bool)
- Reactants & products (SMILES strings)
- Solvents and agents (rxn string data), or solvents, catalysts, and reagents (labeling data) (SMILES strings)
- Temperature, reaction time, yield (floats)
- Procedure details (string)
- Grant date (datetime), date of experiment (datetime), file name (string)

Finally, a CSV file is also created to keep track of the frequency of non-SMILES names used to represent molecules, so the most common molecule names could be added to the manual name resolution dictionary. The four datasets are freely available and can be downloaded immediately from FigShare or regenerated using the code in the ORDerly Github repository.

5 Experiments validating ORDerly

In this section, we compare the performance of the models trained on each of the four datasets described above. The model architecture used for experimentally validating the ORDerly datasets is from Gao *et al.*[8] with only minor adjustments, including the order in which components are being predicted, and not including prediction of reaction temperature (see D) All models were trained on an A10G cloud GPU instance provided by lightning.ai for 100 epochs to minimize cross entropy loss for each reaction component. The best model by validation loss was chosen for evaluation.

5.1 Condition prediction accuracy across datasets

The reaction condition prediction model used in this work predicts five categorical variables: two solvents and three agents. These five molecules form a set (order invariant), though the loss function in the model used to predict the molecules considers them sequentially (with order) since this was found to work better in practice [8]. The metric used to evaluate the accuracy of the model should be order invariant, since the problem is order invariant, and for this reason the accuracy metrics used are top-1 (see D) and top-3 (see table 2) exact match combination accuracy for each type of component (i.e., solvent, agent). Beam search was used to identify the top-3 highest probability sets of reaction

Table 2: Top-3 exact match combination accuracy (%): frequency informed guess // model prediction // AIB%.

Dataset	A labeling rare→"other"	B labeling rare→delete rxn	C reaction string rare→"other"	D reaction string rare→delete rxn
Solvents	47 // 58 // 21%	50 // 61 // 22%	23 // 42 // 26%	24 // 45 // 28%
Agents	54 // 70 // 35%	58 // 72 // 32%	19 // 39 // 25%	21 // 42 // 27%
Solvents & Agents	31 // 44 // 19%	33 // 47 // 21%	4 // 21 // 18%	5 // 24 // 21%

conditions. The top-3 accuracy was compared to the baseline predictive accuracy of simply predicting on the test set the most common molecules found in the train set.

Additionally, we define a metric inspired by Maser *et al.* [19] called the average improvement over baseline (AIB%):

$$AIB\% = \frac{A_m - A_b}{1 - A_b} * 100 \tag{1}$$

where A_m is the exact match combination accuracy of the model and A_b is the exact match combination accuracy of choosing the top 3 most common values of a component in the respective train set.

Table 2 shows the predictive performance on the test set using each of the four datasets. All models show an improvement over the frequency informed baseline. The performance of the labeling datasets at first appears to be better than those that use our custom logic to extract reaction components from the reaction string. However, as shown in Figure 3, many of the reactions in datasets where we trust the labeling in ORD have more than three reactants, while most reactions in organic chemistry only have two reactants. Upon manual inspection, we found that many agents were mislabeled as reactants and, therefore, the prediction problem was made significantly easier by only requiring a single catalyst to be predicted. In contrast, our custom cleaning pipeline that defines components using the reaction string avoided contamination of the desired prediction targets (i.e., the agents) in the inputs, and therefore, better represents the downstream application of reaction condition prediction models. This insight is confirmed in Table 3; there are fewer unique solvents and agents and a higher density of null components when using the ORD labeling instead versus the reaction string indicating that many components might be mislabeled as reactants. This discrepancy demonstrates that naive creation of datasets based on ORD can lead to inflated performance metrics.

For the datasets that extract the components from the reaction string, overall top-3 accuracy is less than 25% across solvents and agents. While not directly comparable, our overall accuracy is lower than what Gao *et al.* [8] achieved with 50.1% top-3 accuracy across catalysts, solvents and agents. However, Gao *et al.* trained on approximately ten million reactions, while we train on less than four percent of that (∼350k). As shown in Figure 4, we see consistent increases in AIB (%) with the number of data points for the dataset which uses reaction strings and deletes rare reactions, and this scaling performance indicates that as ORD grows, better performance could be achieved, even with potentially fewer data points than used in the paper by Gao *et al.*

Finally, the approach to dealing with rare values is investigated. The reaction string datasets would have more than 10,000 unique agents (see Table 1) with no frequency based filtering, which would cause excessive sparcity in the OHE. We initially hypothesised that the rare → "other" strategy would allow for better generalisation, since the edge case reactions would be kept in a way that also keeps the OHE at a reasonable size, however, in practice, the rare → delete rxn strategy had better performance across train set sizes, as seen in Figure 4.

5.2 The ORDerly benchmark

We therefore choose Dataset D (reaction string, rare→delete rxn) as the the ORDerly condition prediction benchmark to be used as an open source benchmark for reaction condition prediction. While it was possible to build predictive models from all four datasets (since $AIB\% > 0$, as per table 2), Dataset D most accurately reflects the practical downstream of chemists and the scaling behavior we desire.

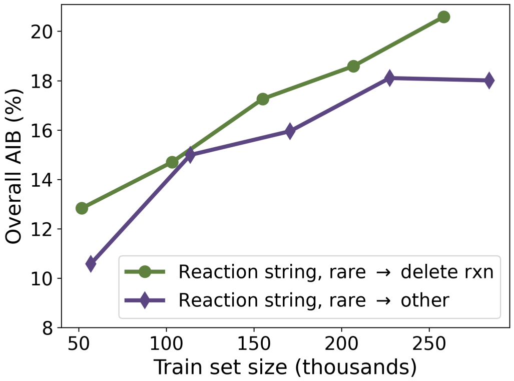


Figure 4: Scaling behaviour of different datasets with respect to overall top-3 AIB for all solvents and agents (third row from table 2.)

6 Technical limitations

6.1 Component labeling

There are two ways of assigning reaction roles to molecules found in ORD files, either relying on the labeling, or identifying reaction roles by considering the atom mapping of a reaction SMILES string. Identifying the role of molecules in a reaction provides crucial context to machine learning models, adding domain knowledge to the data thereby improving performance. Atom mapping the reactions with the newest algorithm may allow for greater accuracy in identifying reaction roles [16], however, an atom mapping algorithm was not integrated into ORDerly to keep ORDerly lightweight. With the existing atom mapping in ORD molecules contributing atoms to the product could readily be bundled together and labeled as reactants. However, subdividing spectator molecules into different categories (e.g. agents, reagents, solvents, catalysts, precatalysts, ligands, acids/bases) is a difficult task. The difficulty is compounded by the fact that the same molecule can play different roles depending on the context. The role that a molecule plays in a reaction may more easily identified when only considering one reaction class at [19], since this allows the mechanistic details of the reaction class [29, 26, 31] to be considered. Handling large and diverse datasets inevitably requires generalizeations that may result in contradictions upon a more fine-grained inspection. In this work, solvents were separated from the other spectator molecules, because these can somewhat reliably be identified. Catalysts were not separated into their own category, since identifying catalysts is more subtle (especially with organocatalysis), and few reactions in the reaction string datasets contained transition metals.

6.2 Condition prediction

While there are a number of different approaches for reaction condition prediction, this work is focused reaction cleaning methodology and the introduction of a new benchmark dataset, as opposed to being a benchmarking case study comparing various model architectures - this would be an interesting area of further study. Furthermore, correctly considering the redundancy of the ordering

Table 3: Sparsity in the datasets. Frequency filtering is necessary for the solvents and agents to avoid sparcity in the one-hot encoding. Columns: Number of unique molecules with a frequency above the threshold; number of unique molecules with a frequency below the threshold ; percentage of the dataset that is None.

	labeling			reaction string		
Reactants	40,020	0	25.7%	317,184	0	18.4%
Products	38,816	0	0.0%	382,850	0	0.0%
Solvents	29	204	40.0%	85	313	28.0%
Agents	48	447	56.2%	255	11,945	37.0%

of molecules may be a way to improve reaction condition prediction algorithms further. The model used for benchmark evaluation is trained to predict one molecule at at time, and it is penalized for predicting the correct two solvents if the solvents are predicted in the wrong order. However, our chosen evaluation metric, whether all five molecules were correctly predicted, does take this redundancy into account, and we implore researchers using this benchmark to also use this evaluation metric as it better represents the goal of condition prediction.

6.3 Useful condition prediction models: data split and diversity

Processing and cleaning a list of chemical reactions is a necessary step for generating a benchmark dataset to train models on, but the test set created from a random train/test split of the benchmark does not fully capture how these algorithms may be used. It is important for the evaluation data and metrics to be reflective of the real world task, and using an order invariant evaluation metric coupled with beam search to identify spectator molecule combinations is not fully sufficient. A test dataset constructed in this way may indicate the performance of a model within its domain of applicability, but does little to investigate the domain of applicability itself (i.e., for which reactant and product molecule combinations can we expect good accuracy?); consideration of reaction classes may be one way to better understand the domain of applicability, and would need to be viewed in a broader context of dataset diversity. Table 2 shows that the frequency informed baseline accuracy is much higher on the labelling datasets than on the reaction string datasets, implying that the dataset is less diverse, but an outstanding question is how should this lack of diversity be coupled with the test performance to understand generalizeability of the model? In general, a truly representative test performance is likely only achievable when coupled with wet lab chemistry.

7 Conclusions

In this work, we presented ORDerly, an open-source framework for preparing chemical reaction data stored in the Open Reaction Database (ORD) for machine learning applications. In order to create a new reaction condition benchmark, we generated four datasets with varied methods for reaction role identification and rare molecule handling. ORD stores molecular data with labeling, in addition to a reaction string, and we found that using chemically informed cleaning and categorization of molecules in the reaction string was preferable. Predictive models could be trained using all four datasets to outperform a frequency informed baseline, but the performance on their respective test sets should be viewed in the context of dataset quality. In particular, datasets that used the labeling in ORD had contamination of the inputs (reactants) with the outputs (agents), resulting in a problem that was unrealistically easy. In contrast, the datasets that used our custom logic to extract reaction roles provided a more realistic performance estimate. These datasets also contained more data, with more diversity in the molecules encountered, implying that it would generalize better, and be more useful in a wet-lab setting. Thus, we present the *ORDerly benchmark*, a dataset for reaction condition prediction formed from reaction string data, where rare reactions are also deleted. All datasets experimented with in this work, as well as the code used to generate them, are freely available online, and we hope the benchmark will make the task of reaction condition prediction more accessible to ML practitioners with limited domain knowledge. ORDerly presents a fully open-source pipeline to go from raw ORD data to a fully trained condition prediction model, allowing for an avenue to leverage the growing contributions to open source chemistry.

8 Acknowledgements

This work is co-funded by UCB Pharma and Engineering and Physical Sciences Research Council via project EP/S024220/1 EPSRC Centre for Doctoral Training in Automated Chemical Synthesis Enabled by Digital Molecular Technologies. This project was co-funded by European Regional Development Fund via the project "Innovation Centre in Digital Molecular Technologies".

References

[1] Alain Vaucher and Hélder Lopes. RXN reaction preprocessing, May 2023. https://github.com/rxn4chemistry/rxn-reaction-preprocessing.

[2] Yehia Amar, Artur M. Schweidtmann, Paul Deutsch, Liwei Cao, and Alexei Lapkin. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chemical Science*, 10(27):6697–6706, July 2019.

[3] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, May 2016.

[4] Connor W. Coley, Regina Barzilay, Tommi S. Jaakkola, William H. Green, and Klavs F. Jensen. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Central Science*, 3(5):434–443, May 2017.

[5] Louis J. Diorazio, David R. J. Hose, and Neil K. Adlington. Toward a More Holistic Framework for Solvent Selection. *Organic Process Research & Development*, 20(4):760–773, April 2016.

[6] Elsevier. Reaxys, January 2009. https://www.reaxys.com.

[7] Kobi C. Felton, Jan G. Rittig, and Alexei A. Lapkin. Summit: Benchmarking Machine Learning Methods for Reaction Optimisation. *Chemistry–Methods*, 1(2):116–122, 2021.

[8] Hanyu Gao, Thomas J. Struble, Connor W. Coley, Yuran Wang, William H. Green, and Klavs F. Jensen. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Central Science*, 4(11):1465–1476, November 2018.

[9] Samuel Genheden, Per-Ola Norrby, and Ola Engkvist. AiZynthTrain: Robust, Reproducible, and Extensible Pipelines for Training Synthesis Prediction Models. *Journal of Chemical Information and Modeling*, 63(7):1841–1846, April 2023.

[10] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, December 2021.

[11] Klavs F. Jensen, Connor W Coley, and Natalie S Eyke. Autonomous discovery in the chemical sciences part I: Progress. *Angewandte Chemie International Edition*, September 2019.

[12] Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[13] Christos Kannas, Amol Thakkar, Esben Bjerrum, and Samuel Genheden. rxnutils – A Cheminformatics Python Library for Manipulating Chemical Reaction Data, August 2022. ChemRxiv, 2022. DOI: 10.26434/CHEMRXIV-2022-WT440-V2.

[14] Steven M. Kearnes, Michael R. Maser, Michael Wleklinski, Anton Kast, Abigail G. Doyle, Spencer D. Dreher, Joel M. Hawkins, Klavs F. Jensen, and Connor W. Coley. The Open Reaction Database. *Journal of the American Chemical Society*, 143(45):18820–18826, November 2021.

[15] Greg Landrum. The RDKit Documentation (accessed January 10 2020), 2006. https://www.rdkit.org/docs/.

[16] Arkadii Lin, Natalia Dyubankova, Timur I. Madzhidov, Ramil I. Nugmanov, Jonas Verhoeven, Timur R. Gimadiev, Valentina A. Afonina, Zarina Ibragimova, Assima Rakhimbekova, Pavel Sidorov, Andrei Gedich, Rail Suleymanov, Ravil Mukhametgaleev, Joerg Wegner, Hugo Ceulemans, and Alexandre Varnek. Atom-to-atom Mapping: A Benchmarking Study of Popular Mapping Algorithms and Consensus Strategies. *Molecular Informatics*, 41(4):2100138, 2022.

[17] Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS central science*, 3(10):1103–1113, October 2017.

[18] Daniel Lowe. Chemical reactions from US patents (1976-Sep2016), June 2017.

[19] Michael R. Maser, Alexander Y. Cui, Serim Ryou, Travis J. DeLano, Yisong Yue, and Sarah E. Reisman. Multilabel Classification Models for the Prediction of Cross-Coupling Reaction Conditions. *Journal of Chemical Information and Modeling*, 61(1):156–166, January 2021.

[20] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, January 2019.

[21] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alán Aspuru-Guzik, and Alex Zhavoronkov. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology*, 11, 2020.

[22] Philippe Schwaller, Théophile Gaudin, Dávid Lányi, Costas Bekas, and Teodoro Laino. “Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science*, 9(28):6091–6098, July 2018.

[23] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583, September 2019.

[24] Marwin H. S. Segler, Mike Preuss, and Mark P. Waller. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610, March 2018.

[25] Teague Sterling and John J. Irwin. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, November 2015.

[26] Bo Sun, Lulu Ning, and Hua Chun Zeng. Confirmation of Suzuki–Miyaura Cross-Coupling Reaction Mechanism through Synthetic Architecture of Nanocatalysts. *Journal of the American Chemical Society*, 142(32):13823–13832, August 2020.

[27] Connor J. Taylor, Kobi C. Felton, Daniel Wigh, Mohammed I. Jeraal, Rachel Grainger, Gianni Chessari, Christopher N. Johnson, and Alexei A. Lapkin. Accelerated Chemical Reaction Optimization Using Multi-Task Learning. *ACS Central Science*, April 2023.

[28] Amol Thakkar, Thierry Kogej, Jean-Louis Reymond, Ola Engkvist, and Esben Jannik Bjerrum. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical Science*, 11(1):154–168, 2020.

[29] A. A. Thomas and S. E. Denmark. Pre-transmetalation intermediates in the Suzuki-Miyaura reaction revealed: The missing link. *Science*, 352(6283):329–332, April 2016.

[30] Daniel S. Wigh, Jonathan M. Goodman, and Alexei A. Lapkin. A review of molecular representation in the age of machine learning. *WIREs Computational Molecular Science*, 12(5):e1603, 2022.

[31] Daniel S. Wigh, Matthieu Tissot, Patrick Pasau, Jonathan M. Goodman, and Alexei A. Lapkin. Quantitative In Silico Prediction of the Rate of Protodeboronation by a Mechanistic Density Functional Theory-Aided Algorithm. *The Journal of Physical Chemistry A*, March 2023.

[32] Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *Journal of Chemical Information and Modeling*, 60(1):47–55, January 2020.