Holistic chemical evaluation reveals pitfalls in reaction prediction models

Victor Sabanza Gil[123*]   **Andres M. Bran**[12*]   **Malte Franke**[1*]
Rémi Schlama[1]   **Jeremy S. Luterbacher**[23]   **Philippe Schwaller**[12]
[1] Laboratory of Artificial Chemical Intelligence (LIAC), ISIC, EPFL
[2] National Centre of Competence in Research (NCCR) Catalysis, EPFL
[3] Laboratory of Sustainable and Catalytic Processing (LPDC), ISIC, EPFL
*Contributed equally.
philippe.schwaller@epfl.ch

Abstract

The prediction of chemical reactions has gained significant interest within the machine learning community in recent years, owing to its complexity and crucial applications in chemistry. However, model evaluation for this task has been mostly limited to simple metrics like top-k accuracy, which obfuscates fine details of a model's limitations. Inspired by progress in other fields, we propose a new assessment scheme that builds on top of current approaches, steering towards a more holistic evaluation. We introduce the following key components for this goal: CHORISO, a curated dataset along with multiple tailored splits to recreate chemically relevant scenarios, and a collection of metrics that provide a holistic view of a model's advantages and limitations. Application of this method to state-of-the-art models reveals important differences on sensitive fronts, especially stereoselectivity and chemical out-of-distribution generalization. Our work paves the way towards robust prediction models that can ultimately accelerate chemical discovery.

1 Introduction

In recent years, there has been a significant increase in the development and application of machine learning (ML) algorithms for solving various tasks in science-related fields[1–5]. Advances in these models have been greatly accelerated by model developments[6–8], acquisition of extensive training data[9,10], and the establishment of benchmarks[11–17], which have enabled researchers to evaluate and compare new models based on multiple aspects relevant to the task at hand. Chemistry has also experienced remarkable progress in problems such as retrosynthetic planning[2,18–23], reaction condition recommendation[24], reaction prediction[25–29], and others[30–36]. Among these, reaction prediction has gained considerable importance due to its broad applicability in areas such as waste material valorization[37], reaction network analysis[38], and even the evaluation of retrosynthesis prediction models[39]. Compared to the other tasks, reaction prediction benefits from having a less ambiguous objective, simplifying the evaluation process.

A wave of progress in this field has been further propelled by the publication of the USPTO reaction dataset[9,40,41], which has led to the emergence of benchmarks for various tasks, including USPTO_STEREO[42] and USPTO_480k[43] for reaction prediction. These consist of tailored subsets of the USPTO dataset, randomly split for training and evaluation. From the algorithmic side, transformer-based sequence-to-sequence models have emerged as the top-performing algorithms for reaction prediction[44], achieving top-1 accuracies of over 91%[29] on stereochemistry-free datasets. Other widely used model types include template-based[45] and graph-to-sequence models[46], each

leveraging different inductive biases derived from chemical expertise, achieving comparable top-1 accuracy.

Evaluation has received considerable attention in fields such as computer vision [47,48] and language models (LMs) [49,50]. Out-of-distribution (OOD) shifts have been thoroughly discussed [51,52], and the need for testing in this domain has been emphasized [53]. Broader evaluation schemes that aim to expose the strengths and failure modes of different models have also been proposed for LMs [54]. However, standardized model evaluation in the field of reaction prediction has been largely neglected, with most studies relying solely on top-k reaction outcome accuracies, a restricted measure of model performance that overlooks a diverse range of complexities inherent to reaction prediction. Some works perform additional analyses and comparisons, giving more insight into the model's performance, however they lack a standardized format and are constrained by the quality of the reaction data that is used [55,56].
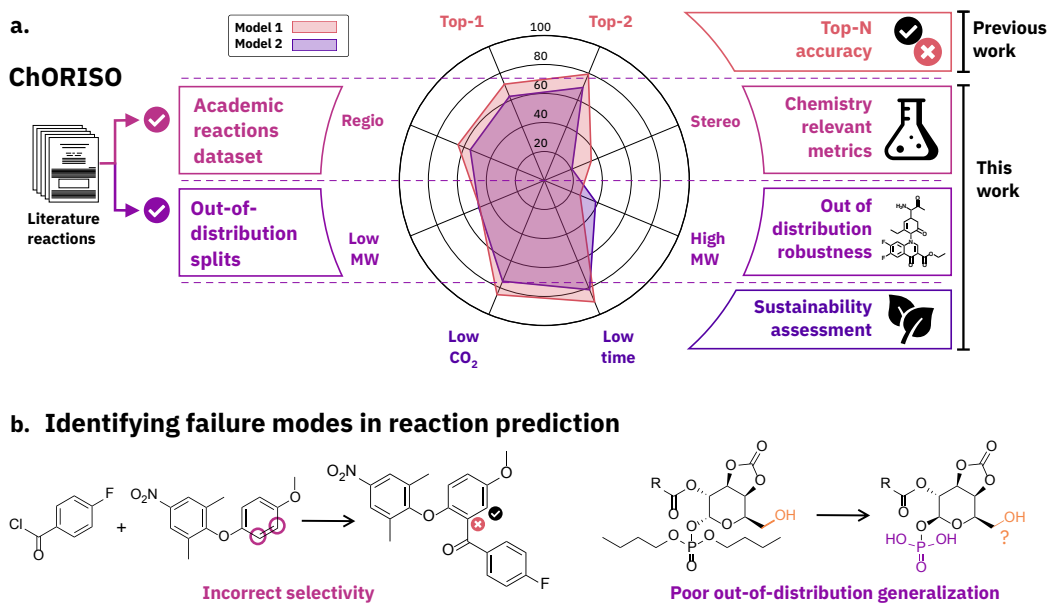

Figure 1: **Holistic evaluation of reaction prediction ML models. a.** Chemistry-relevant metrics, out-of-distribution (OOD) robustness tests and sustainability assessments are proposed. Evaluation is done by using the CHORISO dataset, which provides reaction entries and from which OOD splits are derived. **b.** Failure modes of current reaction prediction models include poor performance in reactions with potential selectivity issues (left), and OOD scenarios such as reactions involving species with higher molecular weight (right).

In this work, we propose a holistic evaluation pipeline for a better chemical assessment of reaction prediction models. For this purpose, we introduce CHORISO (**CH**emical **O**rganic **R**eact**I**on **S**MILES **O**mnibus), a curated dataset of academic chemical reactions, along with a suite of chemically relevant metrics for the standardized evaluation of these models. Addressing the limitations in existing evaluation methods, we propose multiple slices and splits of CHORISO, serving as distinct scenarios for testing, considering OOD scenarios [51]. Several chemically relevant desiderata have been implemented in the proposed standard metrics, including different types of chemical selectivity, along with measures of model efficiency and environmental impact, as shown in Figure 1a. Through a combination of standardized metrics, curated data, OOD testing, and a collection of models, we aim to facilitate the development new cutting-edge ML models for reaction prediction (Figure 1b). This effort could ultimately lead to more accurate and reliable predictions with applications in various fields.

2 Holistic Evaluation of Chemical Reaction Models

Prediction of chemical reaction outcomes is a key problem in organic chemistry. Achieving accurate, trustworthy, and scalable chemical reaction prediction could accelerate chemistry discovery [37,57].

Excellent performance in this task is thus crucial for the development of chemical models in general. These models are typically tested and compared in the literature using the top-k accuracy —with $k \in [1, 5]$, with the best models reaching top-1 accuracies higher than 91% on patent reaction benchmarks without stereochemical information[46]. However, the utility of these models is limited, as they tend to underperform in real-world use cases[28,55]. The disconnection between such high accuracies and poor practical performance thus highlights the need for better evaluation methods.

The recent work of Liang et al.[54] sets the basis for the Holistic Evaluation of Language Models (HELM). The authors argue that, given the flexibility and generality of LMs, it is desirable to establish an evaluation scheme that more transparently assesses the capacities and flaws of these types of models. Holistic evaluation requires the identification of potential scenarios —encoding use-cases— and metrics —encoding desiderata— that are of interest to LMs. In this setting, the authors can adequately test the models in terms of accuracy, robustness, and toxicity, among other metrics, across a wide range of scenarios, exposing the advantages and trade-offs of popular LMs. Building upon this work, our aim in this section is to identify relevant settings where chemical reaction models need to be tested and to develop metrics that align with key desired characteristics of these models.

2.1 Data

A critical piece on the road toward holistic evaluation is data. Data not only feeds the models with latent knowledge but also allows researchers to model scenarios for testing, ultimately allowing to gauge and compare the adequacy of reaction prediction models in such scenarios. Despite the wave of reaction prediction models fueled by the USPTO dataset[41,58], these models learn from regions of the chemical space that are not necessarily typical targets of academic researchers, mostly featuring well-established, industry-relevant reactions.

To alleviate this, we propose CHORISO, a curated reaction dataset of diverse academic reactions. CHORISO is a mix betweencleaned and processed versions of CJHIF, a dataset of reactions extracted from high-impact academic journals[59], and USPTO, a dataset of reactions extracted from patents[9] (see Appendix A.1). As shown in Figure 2, CHORISO features around 2.2M reactions, including a high ratio of C-C bond formation and functional group interconversion reactions, which are fundamental for strategic synthetic planning[60]. CHORISO additionally exhibits heavier-tailed distributions of molecular weight and number of stereocenters in products, covering a larger and more relevant portion of the chemical space. This is particularly important for applications where stereocontrol is fundamental[61], as well as for investigations regarding the scope of chemical reactions, where extrapolation to higher Product's Molecular Weight (PMW) —e.g. larger substituents— is desired. We use this dataset as the data source for our holistic evaluation.
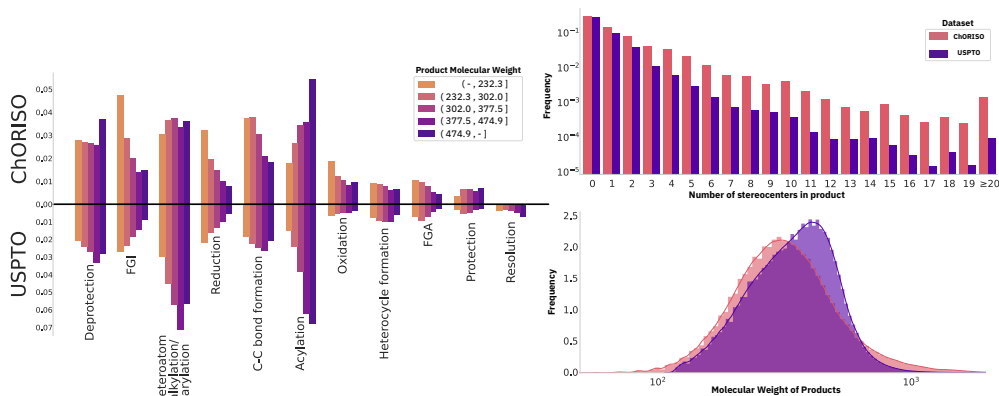

Figure 2: **Dataset distributions comparison.** Differences between the two datasets (CHORISO and USPTO) are highlighted. **a.** Reaction types across various bins of product molecular weight (PMW). The distribution of reaction types varies considerably across PMW for CHORISO. Unrecognized reactions in the datasets are not plotted for clarity. **b.** Distribution of number of stereocenters in products shows that products in CHORISO overall contain a higher number of chiral centers. **c.** PMW distribution. CHORISO's distribution makes it suitable for OOD splits on both ends of the PMW distribution.

2.2 Desiderata for models

A number of properties are expected from reaction prediction models. High accuracy is certainly one of them, but over-reliance on this metric can be misleading, especially in unbalanced datasets[55]. From the chemical side, good handling of stereo- and regiochemistry is desired, as well as a correct understanding of functional group effects, effects of catalysts, reagents and other additives, and conditions like temperature and pressure[55]. Proper modelling of these variables is necessary to have an accurate prediction system, and measuring how models perform in each variable is central to identifying trade-offs between them. Other, less field-specific properties are also desired, such as low inference speeds, low carbon footprint, and low energy consumption, which become important in high-demand applications like chemical networks space exploration[57].

In this work, we pave the way towards tackling the accurate evaluation of models in these regards, by implementing two chemically relevant and two performance-related metrics. The first two, regio- and stereo-accuracy, calculate the product prediction accuracy in subsets of reactions that are flagged to have regio- and stereo-selectivity issues. These reactions are commonly encountered in aromatic ring substitutions[62] or systems where chirality is affected during the reaction[28], among others[63]. These metrics thus highlight the models' capacities to predict the most reactive sites in a given context correctly, and to predict the dominant isomer in a potential isomer mixture. Furthermore, measurement of CO_2 emissions and training time are also considered, as per raising sustainability concerns in AI[64]. Addressing sustainability in models can also improve their accessibility due to reduced hardware requirements for model execution during inference.

2.3 Scenarios

Identifying and recreating scenarios relevant for testing models is one of the two key points highlighted by Liang et al.[54]. We pay particular attention to out-of-distribution (OOD) scenarios, where the train and test set distributions differ. Due to the inherent difficulties in characterizing feature relevance for this task, we focus mainly on marginal shifts, described by Teney et al.[51] as those where a distribution shift happens only across features irrelevant to the task. One easily measurable property, that to good approximation is irrelevant for reaction prediction, is the product molecular weight (PMW). A model will desirably perform well independent of the molecular size, as reactivity analysis is typically based on local molecular features such as functional groups, which PMW does not directly influence.

Following this reasoning, two types of data splits are proposed: low PMW and high PMW, where the test data corresponds to the lowest and highest end of the PMW distribution, respectively (Appendix A.4). In addition, a standard split by products is proposed, where the set of product molecules in the train set is disjoint from its counterpart in the test set. These sets are used to evaluate the model and to provide a wider picture of its capabilities. While we acknowledge the importance of other types of distribution shifts, and future research should focus on exploring these, our current approach already proves valuable in revealing several aspects of models, including their limitations and failure modes, as shown in Section 3. This demonstrates the importance and effectiveness of the scenarios and distribution shifts considered in our study.

3 Results and discussion

With the proposed holistic evaluation pipeline, two high-performing reaction prediction models from the literature were trained and evaluated. The models, Graph2SMILES[46] (G2S), and Molecular Transformer[27] (MT) both draw elements from the Transformer architecture[7]. This architecture relies on the attention mechanism[65] to infer inter-dependencies between tokens in token sequences. While the MT uses a string representation as input and output[44], G2S uses a graph encoder with a string decoder. Figure 1a provides a general comparison of the models, evaluated as proposed in this work using the CHORISO dataset. Contrary to previous reports[46], top-k accuracy indicates an advantage of MT over G2S. However, extending from this, our approach reveals important differences in models, namely their distinct performances in stereoselectivity and OOD generalization, both key for an appropriate evaluation at the chemical level. Figure 1 also illustrates how the implementation of the proposed holistic evaluation pipeline allows to dissect a model into multiple performance factors, providing a better picture of model limitations and trade-offs.

The models in question were trained and tested on multiple splits of the CHORISO dataset. These splits consisted of a random split, a split by products, and two splits by PMW (see Appendix A.4) to measure out-of-distribution performance. As displayed in Table 1, in-distribution splits show advantage of MT over G2S. Notably, results on the random split are systematically higher than those on the product split, highlighting the more challenging nature of the latter, and supporting previous claims of random splits leading to misleading results for generalization[55]. The product split is thus kept as the standard split for this work and all the top-k and selectivity comparisons are based on the predictions on this set. At a high level, top-k accuracy shows lower values compared to the reference reported values. The same training procedure and evaluation on the full USPTO dataset reveal a similar trend on this dataset (Appendix B). The source of the general performance decrease could be assigned to the inherent more difficult split by product and to the nature of both datasets (bigger and noisier than the original USPTO_480k that was used in the original work). Prior studies demonstrated that Transformer models typically achieve 60-70% top-1 accuracy on noisy datasets, consistent with the results for CHORISO and USPTO in this study[44,66]. Finally, a high top-1 accuracy difference (around 10%) between MT and G2S on the CHORISO dataset compared to USPTO also suggests that MT adapts better to the different chemical reactions contained in the new CHORISO dataset.

Split type	Model	Top-1 acc (%)	Top-2 acc (%)	Stereo-acc (%)	Regio-acc (%)
Product	MT	**71.9**	**79.4**	**35.0**	**64.4**
	G2S	62.8	69.4	20.6	55.3
Random	MT	**72.6**	**80.2**	**37.7**	**64.9**
	G2S	62.9	69.6	22.2	54.5
low PMW	MT	48.2	**57.0**	**23.5**	**34.8**
	G2S	**48.9**	56.6	20.3	31.7
high PMW	MT	26.7	29.8	12.8	24.7
	G2S	**33.6**	**37.4**	**17.0**	**32.8**

Table 1: **Model benchmarking results.** Performance of Molecular Transformer (MT) and Graph2SMILES (G2S) models for each split of the CHORISO dataset. The columns display the results of each model/split combination for each of the proposed metrics. Bold font highlights the best performing model in a metric, for a given data splitting method.

Chemistry-specific metrics enrich the analysis and reveal important references between both models. Notably, the performance of both models in selectivity metrics decreases relative to top-k accuracy, an expected behavior given that these scores are computed by selecting an inherently more challenging subset of the test data. Apart from representing a smaller subset of the main test data, regioselective and stereoselective reactions impose an extra layer of difficulty on the reaction prediction task (see Appendix A.3. In these cases, the model has to learn the preferred site reactivity or product chirality in addition to the main task of learning the reactive pattern based on molecule functional groups. MT outperforms G2S both in terms of regio and stereo-accuracy. In terms of regiochemistry, the performance difference between MT and G2S is similar the one for top-1 accuracy, suggesting that the source of advantage of MT for this type of reactions is similar than the one for the non-selective ones. However, stereo-accuracy reveals an extra performance of the MT over G2S, suggesting a special advantage of MT with respect to G2S when predicting stereoselective transformations. This fact has been noted and exploited in previous reports[28]. Likely due to the graph encoder used in G2S, this model's stereochemistry performance is substantially lower, with an over 15% loss in stereo-accuracy relative to MT in the product split. MT is thus a better performing model for reactions where product chirality plays a key role, such as those with stereocenter inversions and stereocenter formation as displayed in Figure 3. These results highlight the potential of the proposed evaluation scheme to reveal otherwise hidden differences between models, otherwise diluted in the top-k accuracy.

Thanks to our proposed metrics, we can locate and rationalize the prediction failure of a model using a chemical context. This and sources of challenge for , and not simply consider it as a merely wrong prediction. Focusing on reactions where MT performs better than G2S, we can observe some of the trends and explain the difference in performance between the models. Figure 3 displays some

reactions where MT provides the correct product and G2S fails from the selective reactions set. Reaction a) shows a stereoselective reduction using sodium borohydride where both models predict the correct molecule (in which the ketone has been reduced to an alcohol), but G2S misses the correct chirality of the generated stereocenter. Reaction b showcases a Curtius rearrangement that preserves the original chirality of the reacting center after the transformation. Here G2S also predicts the correct reacting pattern (even the stereochemistry preservation of the reaction center), but predicts a different isomer where two non-reacting chiral atoms have been inverted. Finally, in terms of regiochemistry, reaction c shows how MT predicts the correct regioisomer of a Friedel-Crafts acylation, whereas G2S generates the incorrect isomer where the acylation happens on the less activated aromatic carbon on the reacting ring. These examples showcase limitations and differences between the two methods, and may help to propose model architecture improvements to correct the selectivity difference.
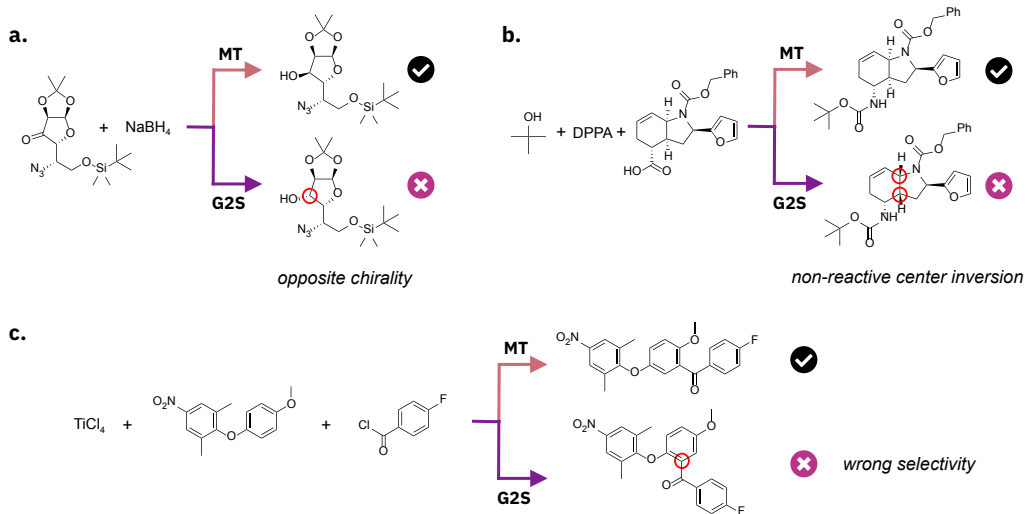

Figure 3: **Model limitations revealed by chemistry-specific metrics.** MT performance is better than G2S in reactions where stereocenters are formed or different regioisomers are possible. **a.** Example where MT predicts the correct product and G2S fails because it predicts the opposite chirality in the reacting center **b.** Example where MT predicts the correct product and G2S fails because it inverts the chirality of non-reacting atoms. **c.** Example where MT predicts the correct product and G2S generates a different regioisomer.

An opposite trend is observed in the OOD splits (low and high PMW), where G2S outperforms MT with differences of up to 7% top-1 accuracy in high PMW. As shown in Figure 4, this difference is further exacerbated as the PMW increases, indicating the strong sensibility of MT to distribution shifts. When analyzing the accuracy by MW split, MT has an initial performance advantage over G2S (higher top-1 accuracy for the reactions where PMW is below or above 100 g/mol to the PMW of the training reactions). MT performance drastically decreases when moving away from training PMW, especially in the high PWM split. G2S is, on the other hand, more robust, as shown by the lower rate of decay in Figure 4. This behavior is hypothesized to be due to the graph-based encoder of G2S, which potentially suffers less from shifts in the molecular size of the input reactants by focusing on local molecular features. The MT instead suffers more from such distribution shifts as PMW directly affects input sequence length. This issue becomes more important for longer sequences, a fact that has been documented previously for language models[67,68] and that can be reflected on the poorer performance of MT on the high PWM. In the out-of-distibution scenarios, MT thus tends to predict larger molecules in the lowPMW split, and smaller molecules in the highPMW split, as shown in Figure 4. In addition, the general performance of both models decreases in the ODD scenario, suggesting that this distribution shift can be used to propose architecture improvements that make models less susceptible to this shift in non-relevant features. Finally, these results highlight the strengths of graph-based models and G2S in particular, which make them a more robust option for scenarios where a shift in property distribution like PMW is expected.

A final perspective based on model efficiency is provided by the sustainability metrics. Detailed analyses of each model's CO_2 production and training and inference time are possible thanks to
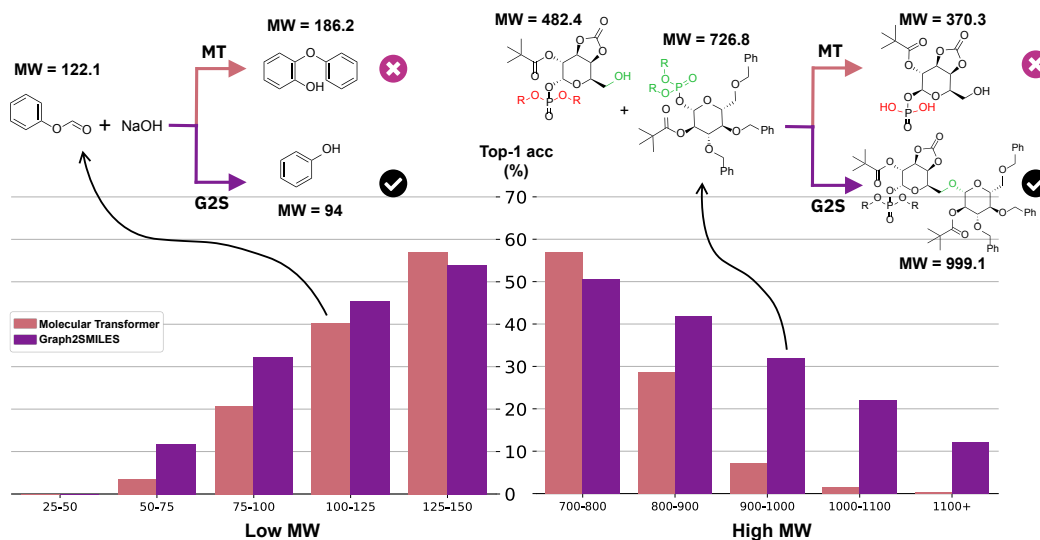
Figure 4: **Model limitations revealed by OOD metrics.** Analysis of model predictions in OOD setting shows that MT's accuracy decreases abruptly as PMW increases, while G2S is more nuanced, showing advantages in OOD generalization. Selected reactions show how MT tends to predict products with PMWs that are within the range of the training data, while G2S is generally better at extrapolation. In green, the reaction center is highlighted, showing how MT predicts the correct reaction, whereas G2S selects an incorrect transformation that lowers the MW of the resulting product and gives an incorrect prediction.

recent tools[69] developed as per recent sustainability concerns of computational research, especially in ML and AI[70]. MT produced 0.32 kg of CO_2 and took 19.8 h to train on the CHORISO data, whereas G2S produced 0.57 kg CO_2 and 40.8 h to train (full training and inference consumption for the benchmarking in Appendix B.2). The higher consumption and training time of G2S renders it less environmentally friendly than the MT, however, the environmental impact is still far less than models in other fields[71]. On the other hand, inference time is similar for both models (3.1 vs 3.7 h). Overall, MT is a more efficient model for this task compared to G2S. This metric may help orient model selection considering a possible sustainability budget for model training and evaluation.

It must be stressed that the goal of holistic evaluation is not to determine the absolute best method among all the available models, as evaluation with top-k accuracy would. Instead, the objective is to provide a detailed map of the strengths and weaknesses of each model, ultimately producing a guide into each model's scope and applicability. Therefore, we have not performed hyperparameter tuning of the models, and instead used previously reported parameters to compare general accuracy. As it has been mentioned, both models were trained for the same number of steps for a fair comparison. As increasing model accuracy was not the main goal of the work, the performance values may be lower than others found in previous reports. However, our results gave a complete comparison in terms of chemistry applicability, robustness, and sustainability of MT and GS2. Following this, MT is a better choice for stereochemically challenging reactions or a limited computational budget, whereas G2S is recommended for scenarios where the novel reaction products fall far from the training property distribution. Furthermore, these efforts will help orient researchers toward addressing specific aspects of models, leading to increasingly better models across multiple chemically relevant directions.

4 Conclusion

We have introduced a new holistic evaluation method for chemical reaction prediction models. This work aims to improve current model evaluation practices in chemistry, allowing a stronger assessment of their real capabilities. Following advances in other fields, we discuss and implement a set of evaluation metrics and scenarios relevant to the task of reaction outcome prediction. In addition, a new academic reactions dataset is released —CHORISO, that is better suited for recreating some of these scenarios, as compared to other existing benchmarks. Leveraging this approach allowed

us to compare two state-of-the-art reaction prediction models, revealing pitfalls and trade-offs in the models, as well as limitations in previous evaluation methods. In particular, holistic evaluation suggests that the Molecular Transformer is better suited for stereochemically challenging reactions, while requiring a fraction of the energetic budget. On the other hand, Graph2SMILES showed much stronger performance in certain out-of-distribution scenarios. These results can easily be rationalized in terms of the models' architecture, with graph-based methods generalizing better to larger graphs, while text-based methods encode spatial features like stereochemistry better. More importantly, the results out-of-the-box reveal key features from models, that are hindered by the commonly used top-k accuracy.

Overall, this holistic evaluation proposes an improved pipeline for thorough reaction prediction model evaluation. Further development is required in the design of richer chemistry-relevant metrics and the identification of additional marginal out-of-distribution splits. In spite of this, this methodology already shows great potential for evaluating models, and opens the way towards the definition of functional guidelines for enhanced model development and selection in chemistry. Selection and improvement of the best-performing models for specific types of reactions and chemical spaces would unlock their routinary application and leverage the current low-data regime. This would lastly enable a future AI-accelerated chemical research.

Data & Code availability

All the data and code used in this work is made freely available. The CHORISO dataset, along with the train and test splits described in this paper can be found at `https://figshare.com/s/5e57a3399c52701cbc15` (DOI: 10.6084/m9.figshare.22598230). The code used for data pre-processing and analysis, metrics, and model evaluation, can be found at `https://github.com/schwallergroup/CHORISO` (data processing, analysis and metrics) and at `https://github.com/schwallergroup/CHORISO-models` (benchmarking).

Acknowledgements

References

[1] Zhang, X. et al. Artificial Intelligence for Science in Quantum, Atomistic, and Continuum Systems. 2023.

[2] 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. *Chemical science* **2020**, *11*, 3316–3325.

[3] Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Is GPT-3 all you need for low-data discovery in chemistry? **2023**,

[4] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

[5] Bran, A. M.; Cox, S.; White, A. D.; Schwaller, P. ChemCrow: Augmenting large-language models with chemistry tools. 2023.

[6] Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gomez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints.

[7] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. 2017; `http://arxiv.org/abs/1706.03762`, arXiv:1706.03762 [cs].

[8] Petersen, F.; Kuehne, H.; Borgelt, C.; Deussen, O. Differentiable Top-k Classification Learning. International Conference on Machine Learning (ICML). 2022.

[9] Lowe, D. M. Extraction of chemical structures and reactions from the literature. **2012**,

[10] Chanussot*, L. et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catalysis* **2021**,

[11] Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. 2019.

[12] Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials* **2020**, *6*, 138.

[13] Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9*, 513–530.

[14] Li, L.; Zeng, L.; Gao, Z.; Yuan, S.; Bian, Y.; Wu, B.; Zhang, H.; Yu, Y.; Lu, C.; Zhou, Z.; Xu, H.; Li, J.; Zhao, P.; Heng, P.-A. ImDrug: A Benchmark for Deep Imbalanced Learning in AI-aided Drug Discovery. 2022; `http://arxiv.org/abs/2209.07921`, arXiv:2209.07921 [cs, q-bio].

[15] Karton, A.; Sylvetsky, N.; Martin, J. M. W4-17: a diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods. *Journal of Computational Chemistry* **2017**, *38*, 2063–2075.

[16] White, A. D.; Hocky, G. M.; Gandhi, H. A.; Ansari, M.; Cox, S.; Wellawatte, G. P.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; Ccoa, W. J. P. Do large language models know chemistry? 2022; `https://chemrxiv.org/engage/chemrxiv/article-details/62c5c622244ce03b8e3c4f21`.

[17] Choudhary, K. et al. Large Scale Benchmark of Materials Design Methods. *arXiv preprint arXiv:2306.11688* **2023**,

[18] Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. - Int. Ed.* **2016**, *55*, 5904–5937.

[19] Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.

[20] Coley, C. W.; Thomas, D. A.; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H., et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365*.

[21] 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. *Journal of cheminformatics* **2020**, *12*, 1–9.

[22] Molga, K.; Szymkuć, S.; Grzybowski, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54*, 1094–1106.

[23] Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T. Machine intelligence for chemical reaction space. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, *12*, e1604.

[24] Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Central Science* **2018**, *4*, 1465–1476, Publisher: American Chemical Society.

[25] Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS central science* **2017**, *3*, 434–443.

[26] Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370–377.

[27] 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 central science* **2019**, *5*, 1572–1583.

[28] 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*, 1–8.

[29] Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* **2022**, *3*, 015022.

[30] 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, PMID: 29532027.

[31] Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: an AI tool for de novo drug design. *Journal of chemical information and modeling* **2020**, *60*, 5918–5922.

[32] Tao, Q.; Xu, P.; Li, M.; Lu, W. Machine learning for perovskite materials design and discovery. *npj Computational Materials* **2021**, *7*, 1–18, Number: 1 Publisher: Nature Publishing Group.

[33] Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* **2016**, *15*, 1120–1127, Number: 10 Publisher: Nature Publishing Group.

[34] Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89–96.

[35] Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *Journal of the American Chemical Society* **2022**, *144*, 19999–20007.

[36] Ramos, M. C.; Michtavy, S. S.; Porosoff, M. D.; White, A. D. Bayesian Optimization of Catalysts With In-context Learning. *arXiv preprint arXiv:2304.05341* **2023**,

[37] Wołos, A.; Koszelewski, D.; Roszak, R.; Szymkuć, S.; Moskal, M.; Ostaszewski, R.; Herrera, B. T.; Maier, J. M.; Brezicki, G.; Samuel, J.; Lummiss, J. A. M.; McQuade, D. T.; Rogers, L.; Grzybowski, B. A. Computer-designed repurposing of chemical wastes into drugs. *Nature* **2022**, *604*, 668–676, Number: 7907 Publisher: Nature Publishing Group.

[38] Weber, J. M.; Lió, P.; Lapkin, A. A. Identification of strategic molecules for future circular supply chains using large reaction networks. *Reaction Chemistry & Engineering* **2019**, *4*, 1969–1981.

[39] 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. *Chemical Science* **2020**, *11*, 3316–3325, Publisher: The Royal Society of Chemistry.

[40] Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical name to structure: OPSIN, an open source solution. 2011.

[41] Lowe, D. Chemical reactions from US patents (1976-Sep2016). **2017**,

[42] Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018**, *9*, 6091–6098.

[43] Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. Advances in Neural Information Processing Systems. 2017.

[44] 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 Central Science* **2019**, *5*, 1572–1583, Publisher: American Chemical Society.

[45] Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chemistry – A European Journal* **2017**, *23*, 5966–5971, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/chem.201605499.

[46] Tu, Z.; Coley, C. W. Permutation Invariant Graph-to-Sequence Model for Template-Free Retrosynthesis and Reaction Prediction. *Journal of Chemical Information and Modeling* **2022**, *62*, 3503–3513, Publisher: American Chemical Society.

[47] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009; pp 248–255.

[48] Mu, N.; Gilmer, J. MNIST-C: A Robustness Benchmark for Computer Vision. 2019.

[49] Chen, M. et al. Evaluating Large Language Models Trained on Code. **2021**,

[50] Srivastava, A. et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. 2022; `http://arxiv.org/abs/2206.04615`, arXiv:2206.04615 [cs, stat].

[51] Teney, D.; Kafle, K.; Shrestha, R.; Abbasnejad, E.; Kanan, C.; Hengel, A. v. d. On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law. 2020; `http://arxiv.org/abs/2005.09241`, arXiv:2005.09241 [cs].

[52] Shen, Z.; Liu, J.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; Cui, P. Towards Out-Of-Distribution Generalization: A Survey. 2021; `http://arxiv.org/abs/2108.13624`, arXiv:2108.13624 [cs].

[53] Hu, R.; Sang, J.; Wang, J.; Hu, R.; Jiang, C. Understanding and Testing Generalization of Deep Networks on Out-of-Distribution Data. 2021; `http://arxiv.org/abs/2111.09190`, arXiv:2111.09190 [cs].

[54] Liang, P. et al. Holistic Evaluation of Language Models. 2022; `http://arxiv.org/abs/2211.09110`, arXiv:2211.09110 [cs].

[55] Kovács, D. P.; McCorkindale, W.; Lee, A. A. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nat. Commun.* **2021**, *12*, 1695.

[56] Toniato, A.; Vaucher, A. C.; Lehmann, M. M.; Luksch, T.; Schwaller, P.; Stenta, M.; Laino, T. Fast customization of chemical language models to out-of-distribution data sets. *Chem. Mater.* **2023**, *35*, 8806–8815.

[57] Wołos, A.; Roszak, R.; Żądło-Dobrowolska, A.; Beker, W.; Mikulak-Klucznik, B.; Spólnik, G.; Dygas, M.; Szymkuć, S.; Grzybowski, B. A. Synthetic Connectivity, Emergence, and Self-Regeneration in the Network of Prebiotic Chemistry. *Science* **2020**, *369*, eaaw1955.

[58] Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. thesis, University of Cambridge, 2012.

[59] Jiang, S.; Zhang, Z.; Zhao, H.; Li, J.; Yang, Y.; Lu, B.-L.; Xia, N. CJHIF original. *IEEE Access* **2021**, *9*, 85071–85083, Conference Name: IEEE Access.

[60] Toniato, A.; Vaucher, A. C.; Schwaller, P.; Laino, T. Enhancing diversity in language based models for single-step retrosynthesis. *Digital Discovery* **2023**, *2*, 489–501.

[61] Gal, J. Molecular chirality in chemistry and biology: Historical milestones. *Helvetica Chimica Acta* **2013**, *96*, 1617–1657.

[62] Ree, N.; Göller, A. H.; Jensen, J. H. RegioML: predicting the regioselectivity of electrophilic aromatic substitution reactions using machine learning. *Digital Discovery* **2022**, *1*, 108–114.

[63] Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels–Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angewandte Chemie International Edition* **2019**, *58*, 4515–4519.

[64] Schwartz, R.; Dodge, J.; Smith, N. A.; Etzioni, O. Green AI. **2019**,

[65] Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2016.

[66] Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature communications* **2020**, *11*, 1–11.

[67] Press, O.; Smith, N. A.; Lewis, M. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. 2022.

[68] Sun, Y.; Dong, L.; Patra, B.; Ma, S.; Huang, S.; Benhaim, A.; Chaudhary, V.; Song, X.; Wei, F. A Length-Extrapolatable Transformer. 2022.

[69] Budennyy, S.; Lazarev, V.; Zakharenko, N.; Korovin, A.; Plosskaya, O.; Dimitrov, D.; Akhripkin, V.; Pavlov, I.; Oseledets, I.; Barsola, I., et al. Eco2ai: carbon emissions tracking of machine learning models as the first step towards sustainable ai. Doklady Mathematics. 2023; pp 1–11.

[70] Budennyy, S. A.; Lazarev, V. D.; Zakharenko, N. N.; Korovin, A. N.; Plosskaya, O. A.; Dimitrov, D. V.; Akhripkin, V. S.; Pavlov, I. V.; Oseledets, I. V.; Barsola, I. S.; Egorov, I. V.; Kosterina, A. A.; Zhukov, L. E. eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI. *Doklady Mathematics* **2022**, *106*, S118–S128.

[71] Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; Lample, G. LLaMA: Open and Efficient Foundation Language Models. 2023.

[72] National Center for Biotechnology Information, PubChem. `https://pubchem.ncbi.nlm. nih.gov/`, 2023.

[73] Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

[74] Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* **2021**, *7*, eabe4166, Publisher: American Association for the Advancement of Science.

[75] Ghiandoni, G. M.; Bodkin, M. J.; Chen, B.; Hristozov, D.; Wallace, J. E. A.; Webster, J.; Gillet, V. J. Development and Application of a Data-Driven Reaction Classification Model: Comparison of an Electronic Lab Notebook and Medicinal Chemistry Literature. *Journal of Chemical Information and Modeling* **2019**, Publisher: American Chemical Society.

[76] Lin, A. et al. Atom-to-atom Mapping: A Benchmarking Study of Popular Mapping Algorithms and Consensus Strategies. *Molecular Informatics* **2022**, *41*, 2100138, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.202100138.

A   Data

Recently, Jiang et al.[59] published a dataset containing 3.2M organic reactions mined from high-impact journals (CJHIF). Each entry contains a reaction SMILES together with the reported yield, as well as the reagents, solvents, and catalysts used, all given in natural text (i.e. no machine readable formats).

A.1   Data curation

For the cleaning and curation of the CJHIF dataset, we followed the pipeline described in Figure 5. In a first step, all the names of reagents, solvents and catalysts are extracted as natural text from the CJHIF dataset and counted by occurrence. This allowed us to translate individually each name to its corresponding SMILES string. To do that, we used Py2OPSIN[40], and PubChem API[72] in case the name could not be translated by the former. Once this translation dictionary is obtained, we analysed more deeply the names that could not be translated to SMILES, and corrected them manually based on the highest occurrence to the lowest. In addition to the manual correction by occurrence and in order to get high quality data on stereochemical reactions, we also looked for specific symbols in the names such as "+", "-", "r", "s", and manually corrected them by occurrence. These characters are normally present in chiral catalysts and are often present in stereoselective transformations. Since we still noticed important compounds that were not translated because of typos in the names or additional characters, we clustered the remaining names based on string similarity using the package fuzzywuzzy in combination with a DBSCAN algorithm provided in the scikit-learn library[73]. Once a quasi-complete translation dictionary is obtained, the compounds names in the CJHIF dataset were translated to SMILES, to form a set of full reaction SMILES. The resulting reactions are filtered to keep only those where no species were lost during translation. This guarantees a one to one correspondance and ensures the fidelity of the resulting reaction SMILES with respect to the original entry. A second filter is also applied to ensure that the products contain at most fewer atoms than the reactants, and that no atom appear in the product that was not in the reactants (to avoid uncomplete entries). A third filter is used to check if the products have any stereocenter that the reactants do not have. These reactions are then modified to remove any stereocenter, in order to keep the chemistry associated. Naturally, a catalyst with axial chirality cannot be encoded as SMILES, so we believe that to be consistent for the transformer model, stereocenters in SMILES have to be induced from the reactants SMILES. Reactions are then canonicalized, and duplicates are removed.
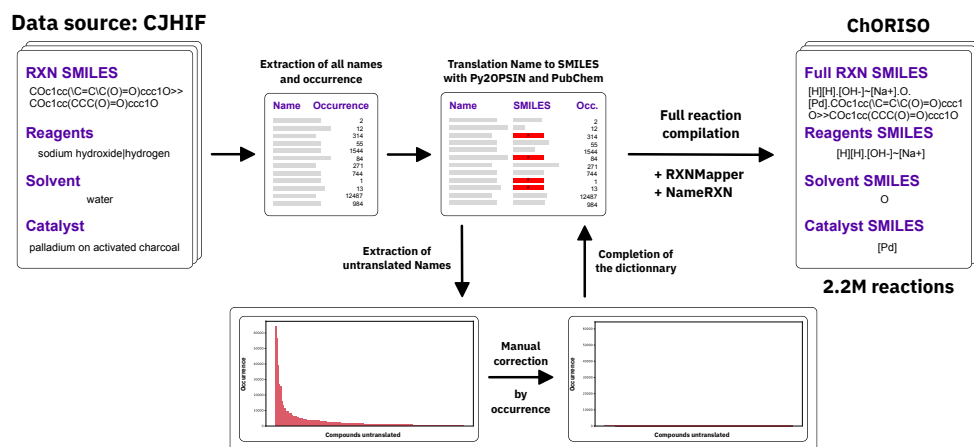

Figure 5: **Pipeline used for data curation.** The original compounds names in the CJHIF are extracted and listed by occurrence. Then, Py2OPSIN[40] and PubChem[72] are used to translate those names to SMILES strings. A manual correction of untranslated names is performed to get a more complete translation dictionary. Next, the original rows in the CJHIF are translated to SMILES to form full reaction SMILES. These reactions SMILES are then mapped using RXNMapper[74] and NameRxn[75].

The next step involves computing the atom mapping[76] for each reaction using both RXNMapper[74] and NameRxn[75].

The resulting dataset, CHORISO, contains 2'224'239 unique chemical reactions, encoded as reaction SMILES.

A.2 Data analysis - Comparison with USPTO

The comparison between CHORISO and USPTO reveals differences in terms of reaction types and product distributions. The same cleaning pipeline was applied in both datasets in order to have a meaningful comparison. Reaction superclasses obtained with NameRxn for each dataset are shown in figure 2a. Dominant superclasses are different in both cases. Acylation and heteroatom alkylation and arylation are the most common classes in USPTO, representing more than 35% of the dataset. In the case of CHORISO, deprotection reactions are the most abundant (around 20%), followed by functional group interconversion (FGI) and heteroatom alkylation and arylation. The differences between classes in both datasets reflect the contrast in data sources. While USPTO samples were extracted from patents (with a higher proportion of pharma-related processes), CHORISO reactions come from academic journals, showing a high proportion of protective chemistry. Additionally, in both datasets more than 30% of the reactions cannot be assigned to a superclass (and therefore not shown in the previous figure plot).

As shown in Figure 2b and c, the distribution of product molecular weight in CHORISO shows longer tails than USPTO, with a better representation of lighter and heavier products which makes it more suitable for the OOD evaluations proposed here. Additionally, the proportion of products containing stereocenters in CHORISO is bigger than in USPTO. This diversity justifies its use as a new benchmarking dataset for reaction prediction models.

A.3 Chemistry-specific metrics

Chemistry-specific metrics are designed as a refinement for top-n accuracy that considers a carefully selected subset of the test set, to test particularly the model's performance on challenging reactions with potential stereo- and regio-selectivity issues.

For the stereo-selective accuracy, reactions in the test set are filtered to include only those where the generation or inversion of a stereocenter occurs (Figure 6a). To do this, a reaction template with radius=0 is extracted from the queried reaction; if the reaction center in the product contains a '@' or '@@' character, it is used for model evaluation. Top-1 accuracy is then computed for this subset of reactions. A similar approach is followed for the regio-selective accuracy, however, the objective is now to identify reactions that could have undergone different regio-selectivity. For this, a reaction template with radius=1 is extracted from the queried reaction and then applied to the main reactants using RDKit RunReactants method. If several products are obtained, the reaction is used for model evaluation.

The resulting evaluation subsets contain 8440 (6.0%) and 13114 (9.4%) of the reactions in the test set of CHORISO's standard split, for stereo and regio, respectively. Despite the comparatively small size of these subsets with respect to the bigger testing set, these evaluations already permit an insightful decomposition of the model's performance across different chemically relevant aspects. Furthermore, stereo and regio accuracies represent a more realistic estimation of the model's ability to predict the results of a subset containing only stereoselective and regioselective reactions respectively.

A.4 Out-of-distribution splits

After doing the splits, the resulting testing sets contain the following number of reaction SMILES:

- Product: 141130
- Random: 121144
- High PMW: 110056
- Low PWM: 90453

The product set contains reactions whose products are not contained in any of the reactions from the training set. The random split contains a random subset of the training reaction. The high PMW and low PWM splits contain reactions whose products are above 700 g/mol and below 150 g/mol
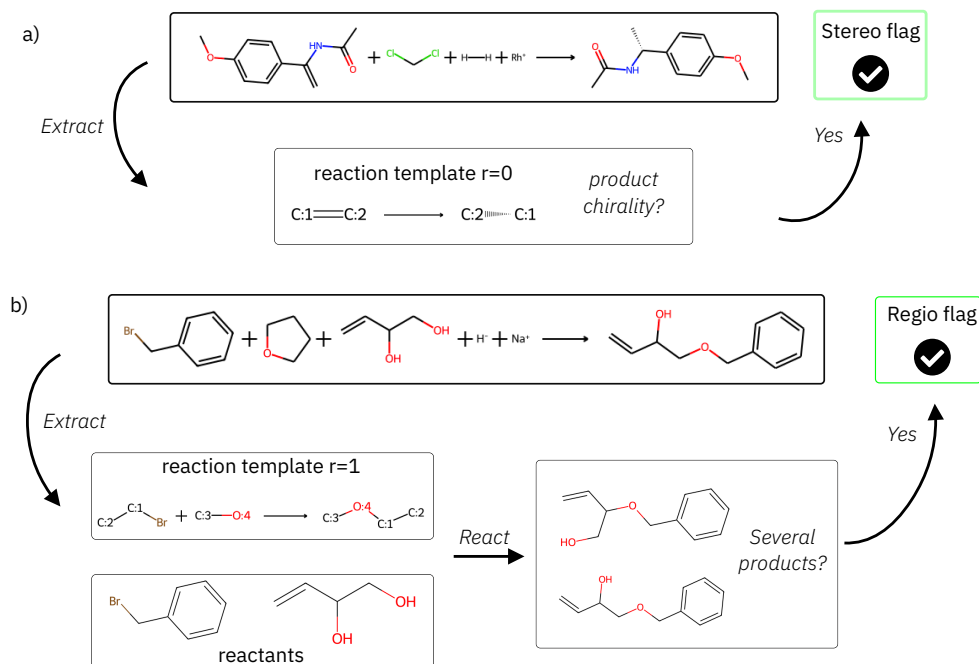
Figure 6: **Chemistry-specific metrics workflow**. a) Stereo-score checks if a chiral center is created or inverted in a reaction and flags it if True to compute the accuracy. b) Regio-score extracts the reaction template with radius=1 and applies it to the reactants. If more than one product is generated, it flags the reaction to compute the accuracy.

respectively, featuring an out-of-distribution shift on a non-relevant feature for reaction prediction as it is product molecular weight (PMW).

B Model benchmarking

In order to demonstrate the benefits of the proposed evaluation scheme with CHORISO, two SOTA models were selected for benchmarking: Molecular Transformer (MT) and Graph2SMILES (G2S). Detailed results are shown, including a comparison with models trained on previous benchmarks.

B.1 Model performance

Table 2 displays the results of the evaluation over 4 metrics, across different data sources and split types (full benchmark).

Data source	Split type	Model	Top-1 acc (%)	Top-2 acc (%)	Stereo-acc (%)	Regio-acc (%)
CHORISOv1	Product	MT	71.9	79.4	35.0	64.4
		G2S	62.8	69.4	20.6	55.3
	Random	MT	72.6	80.2	37.7	64.9
		G2S	62.9	69.6	22.2	54.5
	low PMW	MT	48.2	57.0	23.5	34.8
		G2S	48.9	56.6	20.3	31.7
	high PMW	MT	26.7	29.8	12.8	24.7
		G2S	33.6	37.4	17.0	32.8
USPTO full*	Product	MT	72.4	79.1	35.5	53.1
		G2S	70.3	75.5	32.6	51.4
	Random	MT	73.5	80.1	40.3	54.9
		G2S	70.7	76.0	34.4	51.7
	low PMW	MT	37.4	44.5	11.0	14.1
		G2S	45.1	50.5	14.8	17.0
	high PMW	MT	28.1	30.2	6.4	13.3
		G2S	37.5	40.3	9.6	17.8
USPTO MIT**	Random	MT[42]	88.6	93.7	–	–
		G2S[46]	90.3	–	–	–

Table 2: **Model benchmarking results.** The performance of two models is shown, for each of two data sources, over the different types of data splits discussed here.
* Dataset curated using the same pipeline as for CHORISOv1. **Top-3 accuracies, top-2 accuracies were not reported in these studies.

As can be seen, the additional evaluation metrics and scenarios add new layers of insight to the comparison. In particular, MT is better at handling stereochemistry than G2S, while the latter largely outperforms in OOD scenarios. As highlighted in the main article, this information is instrumental not only for model selection for different use cases, but also serves to guide the research towards tackling specific objectives for different applications of the models. The results in Table 2 also shed light on the influence and importance of proper data splitting. In particular, this has a great influence on performance in the OOD settings, where performances are heavily affected by this choice of splitting type.

B.2 Sustainability assessment

Sustainability assessments are central to our approach. More specifically, we focus on measuring CO_2 emissions and model traning and inference time . Table 3 shows the results for MT and G2S, across different data splits. For fair comparison, every model was trained for 200,000 training steps. Other hyperparameters for each model can be found on the benchmarking repository. All the sustainability metrics were tracked using the eco2AI package[69].

For purposes of visualization (Figure 1), both sustainability metrics were scaled between 0 and 100 to facilitate comparison between models and compatibility with the ranges of the other metrics. The resulting magnitudes therefore follow the same rational than top-k accuracy where a higher value normally means better performance. In particular, a formula of exponential decay was used for each metric. CO_2 production and Duration were scaled using eq. 1, with $k = 0.1$ for CO_2 and $k = 0.001$ for time.

$$S_x = 100e^{-kx} \tag{1}$$

The choice of k was based on the scale of each metric in order to adjust it to the values on the main plot. These values are only used for visualization, and the ones that can be used for the analysis and comparison of models are the absolute values. Table 3 shows the sustainability metrics for the MT and G2S models trained on the CHORISO dataset. Sustainability metrics for the models trained on USPTO were not fully computed due to an error of the Eco2AI on the hardware. This shows .... and suggests that further improvement is needed in order to have a standardized energy consumption report.

Model	Step	CO_2 (kg)	CO_2 scaled	Duration (h)	Duration scaled
MT	Training	0.32	85.01	19.76	90.59
	Inference	0.04	97.86	3.14	98.44
G2S	Training	0.57	75.08	40.77	81.56
	Inference	0.05	97.76	3.71	98.16

Table 3: **Model training sustainability benchmarking.**

C   Extra

All the results obtained for the CHORISO dataset were executed on a single GeForce RTX 3090 (24 GB) GPU. All results obtained for the USPTO dataset were executed on a single Tesla V100-PCIE (32GB) GPU. The model wrappers for training and evaluation can be found at `https://github.com/schwallergroup/CHORISO-models`.