Exploring Organic Syntheses through Natural Language

Andres M. Bran[12*] **Cheng-Hua Huang**[1] **Philippe Schwaller**[12]
[1] Laboratory of Artificial Chemical Intelligence (LIAC), ISIC, EPFL
[2]National Centre of Competence in Research (NCCR) Catalysis, EPFL
philippe.schwaller@epfl.ch

Abstract

Chemists employ a number of levels of abstraction for describing objects and communicating ideas. Most of this knowledge is in the form of natural language, through books, articles and oral explanations, due to its flexibility and capacity to connect the different levels of abstraction. Despite of this, machine-learning chemical models are typically limited to low-level abstractions like graph representations or dynamic point clouds that, although powerful, ignore important aspects like procedural details. In this work, we propose methods for exploring the chemical space at the rich level of natural language. In this setting, synthetic procedure paragraphs are split into segments in four possible classes, and are subsequently mapped into a latent space where they can be conveniently studied. We explore the structure of this space, and find interesting connections with experimental realisation that are beyond the scope of commonly used reaction SMILES. This work aims to draw a path towards LLM-based data processing and chemical space exploration, by analyzing chemical data in previously inaccessible ways that will ultimately allow for better understanding of materials design.

1 Introduction

Chemistry, as a scientific discipline, is characterized by the multiple levels of abstraction that can be used depending on the scientific question[1]. Such representations range from point clouds in physics-based modeling in quantum chemistry[2–4], to graphs and hypergraphs in fields like cheminformatics[1,5,6]. Despite this diversity, the majority of data in chemistry is stored and presented in natural (human) language, as evidenced by the growing volume of scientific articles published annually in the field. Indeed, natural language serves as the principal medium for disseminating chemical knowledge and discoveries. In the realm of synthetic organic chemistry, it conveys rich information not only about the transformation in question but also mechanistic insights, hypotheses, and connections with analytical data. It extends further to link with the physical world through descriptions of conditions, experimental procedures and setups, and beyond. Natural language has thus the potential to represent chemical knowledge in a way no other representation can, thanks to its rich information density and wide availability.

The last years have seen the rise of models for organic reactions based on the molecular structure abstraction[7,8], succeeding at tasks like reaction outcome prediction[9], retrosynthetic planning[10,11], among others[12–18]. Progress in this filed has been fueled by the publication of open reaction datasets mined from patents through traditional NLP techniques[19,20]. Despite the usefulness of this abstraction —which chemists also routinely use for understanding reactions— a natural barrier is established between this abstraction and any form experimental realisation, by cleaning out information like experimental conditions, action sequences, among others. Some works have shown success in

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

attempting to compromise the two worlds[21,22], and also demonstrated how the direct use of this abstraction requires treating other relevant information as separate additions.

An additional issue of this abstraction is the potential for data corruption during conversion. As shown in Figure 1, extraction of reaction smiles (**a**) from the source paragraph (**b**) leads to its contamination with information from different parts of the text, namely the work-up and purification. In this example, the reaction SMILES not only includes the addition of DCM and sodium bicarbonate —actually part of the work-up, but also combines two separated reaction steps into a single one, distorting the original meaning of the reaction and clearly adding noise to the reaction datasets. All this information is, nevertheless, clearly conveyed in the procedure paragraph.

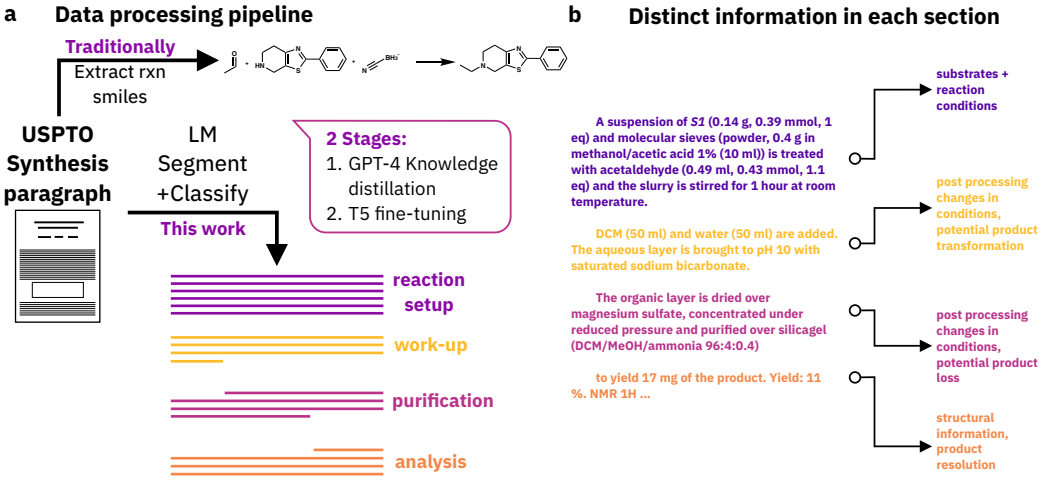

Figure 1: Decomposition of all the levels of information that are extracted from a synthetic procedure paragraph. **a.** Reaction SMILES mined from paragraphs are the most popular choice of representation, however it may come with noise from reagents from potentially irrelevant parts of the paragraph. **b.** Synthetic procedures can be split into steps, with different meanings depending on the context. Machine understanding of such context is key for capturing only the relevant reagents. **c.** Examples of the information present in each semantic segment. This decomposition gives a better idea of what type of reaction is going on, as well as other procedural details.

Despite the drawbacks of directly modeling chemistry through language, recent advances in LLMs have demonstrated the potential in tasks like regression, classification[23] and optimization[18], but also language-modulated reasoning, shown to autonomously solve chemical tasks from organic synthesis to materials design[24,25]. Systematic analysis of chemical data from this perspective has been, however, largely ignored despite success in these applications. This work aims to fill this gap and propose a path towards demonstrating the power of natural language for modeling and exploration of the chemical space.

To assess the predictive power of the additional data extracted as explained in Section 2.2, we consider the task of reaction prediction, where the outcome of a reaction is predicted based on some input information. This task has previously been tackled as an open ended product prediction task from the reaction precursors[26,27], simulating what human experts would solve when planning a reaction sequence, that is by trying to guess the outcome given some hypothetical precursors. More recently, work has been focused on predicting the product given some analytical data as an input[28,29], a task which aims to tackle the product elucidation challenge once a reaction has already been performed and analytical data is available; despite good performance, these models still use only one type of analytical data, like NMR or IR spectroscopy. A more realistic chemical scenario would include both the precursors used for a reaction —narrowing down the space of possible products, as well as the analytical data, which allows to further resolve the obtained product. Here we show how the reaction data processing pipeline described in this work produces a valuable asset to tackle this task.

2 Methods

2.1 Data

The USPTO patent database contains a large collection of chemical synthesis procedures –paragraphs extracted from patent grants (from 1976 to 2016) and applications (from 2001 to 2016). The paragraphs were originally mined and processed by [19,20], from which reaction SMILES and other properties were extracted. More than 3.7M paragraphs are obtained in this way. For the subsequent analysis, a subset of this data was selected that corresponds to paragraphs in the quartiles 2 and 3 of the paragraph length distribution, to limit the data distribution to short enough, but informative enough paragraphs. This process resulted in a total 1.87M samples.

2.2 Semantic procedure segmentation

As shown in Figure 1, organic synthesis procedures can typically be split in four main semantically distinct types of segments —"reaction set-up", "work-up", "purification", and "analysis"— that are clearly differentiated in the practice, but not explicitly separated in texts. Segmenting and classifying these paragraphs is an important part of the work towards analyzing these data form the perspective of natural language, and due to the subtle linguistic distinctions between different segment classes, particularly reaction set-up and work-up, it is only feasible through algorithms capable of capturing and leveraging local context, such as modern language models.

To achieve this goal, we followed a knowledge distillation approach to generate pairs of samples "paragraph":"segmented paragraph" from state-of-the-art LLMs like GPT-3.5 [30] and GPT-4 [31]. Paragraphs were segmented by prompting the LLMs using a combination of prompting techniques such as zero and few-shot [32] in-context learning [23], chain-of-thought [33] and instruction tuning [30]; the complete prompt is given in the Appendix B.1. A training set with 29349 datapoints —23629 from GPT-3.5 + 5720 from GPT-4— was generated in this manner, containing paragraphs X associated with JSON files Y, which specifies the information from each paragraph segment along with its segment class and step order, as shown in Figure 4. To scale this procedure to the complete USPTO dataset (Section 2.1), a Flan-T5 model [32] was then fine-tuned for this task using LoRA adapters [34], improving the efficiency of segmentation and significantly reducing costs. Training details and testing results are provided in the Appendix B.2. With this model, the complete database of 1.8M reaction paragraphs was mapped to segments. After filtering of defective segmentations (see Appendix B.3), a total of 4'881.123 text segments are obtained, each with a segment class and associated step order, corresponding to a total of 1'743.928 paragraphs.

2.3 Sentence embeddings and reaction fingerprints

As proposed by [35], the chemical reaction space can be suitably explored by mapping reaction SMILES into convenient vectorial representations —reaction fingerprints (RXNFP), that encode relevant information about the reactions into a high dimensional vector space. Such representations allow similarity quantification, but also clustering and visualization. This method revealed important insights about reaction datasets, and unlocked a number of applications within the RXN4Chemistry program [36], including reaction property prediction and reaction search.

In a similar fashion, the paragraph segments (Section 2.2) are encoded into a high dimensional vector space using pre-trained open-source Language Embedding Models (LEM) [37]. In particular, the **BAAI/bge-large-en-1.5** (BGE) [37] model was used, as it ranks first in the https://huggingface.co/spaces/mteb/leaderboard, which compares multiple LEMs in tasks such as text classification, clustering, among others [38]. This model encodes input text into a 1024-dimensional representation. Although both embeddings point in principle to the same object (i.e. a chemical reaction), both are fundamentally different as RXNFP encodes the reaction SMILES —a graph representation—, while BGE encodes more global information, such as procedure execution details, reaction loadings and times used, and other conditions, while also offering a linguistic molecular representation, typically *IUPAC* names.

2.4 Reaction prediction

Following the work of[39], we use OpenNMT as a framework for training and running inference of encoder-decoder transformers. We tackle the task of product prediction from reaction precursors smiles and NMR analytical data. For this, the obtained dataset was filtered to obtain those including NMR data and reaction smiles. The NMR data was preprocessed in a similar manner as proposed in[28], and the resulting spectra were tokenized using an in-house trained BPE tokenizer[40] with 5k vocabulary size and WhiteSpace pre-tokenizer. The precursor and product SMILES are tokenized using the regex tokenizer proposed in[39]. The inputs to the model are the concatenated strings of tokenized NMR data and precursor SMILES, while the output is the product SMILES.

For comparison, multiple models were trained on the selected subset of reactions with reported NMR: A baseline model, for the task of precursors to products prediction, along with a number of models trained for the task of NMR+precursors to product, with varying sizes. Results for this are given in Table 1.

To further test the approach, we select a subset of Buchwald-Hartwig reactions from USPTO for which NMR data has been reported, leading to 8,593 test reactions. The results shown in Table 1 correspond to models trained on the full USPTO dataset described above, and tested on this subset.

3 Results & Discussion

In the current setting, each reaction paragraph maps to 5 different vector spaces, each encoding different information about the reaction. Many questions arise regarding the inner structure of this spaces, as well as the interrelations between them. In particular, analogous to the different reaction types that can be found using RXNFPs[35], the different types of workup and purification may as well be explored. Exploring the interrelations between spaces might shed light on the factors driving choices of workup or purification, given a reaction schema or reaction set-up. Furthermore, the effect of work-up and purification on reaction yield may now be explored with this dataset.

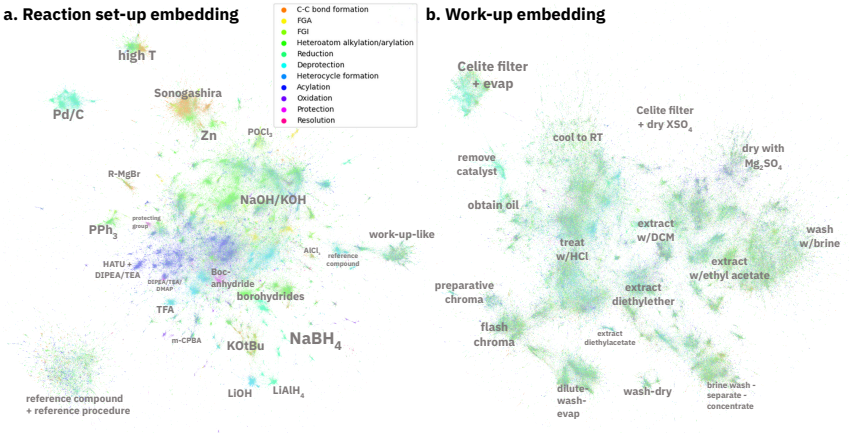

Figure 2: Language Model embeddings of reaction set-ups and work-ups in the USPTO database, as colored by reaction type. The subspaces look highly structured and clusters correlate to a certain extent with the reaction classes. Common topics within the clusters include specific reagents, or types of reagents, as well as specific types of procedures. Using procedural representations such as these is ought to improve models for various tasks, by providing more context regarding experimental realisation.

As shown in Figure 2, LEMs encode reaction set-ups and work-ups into a somewhat structured vector space. In particular, we find that reaction set-ups are mostly clustered by reagents –or type of reagents used, which non-surprisingly correlates with reaction class (extracted from the reaction SMILES). This clustering seems to be largely independent from other irrelevant factors, such as text segment length or writing style, and indeed was seen to capture relevant information regarding experimental conditions. Namely, the "tert-amine" cluster groups reactions executed with DIPEA, TBA, DMAP, etc, or other clusters of reactions executed at high temperatures, among others. The

same model, applied to work-up segments, is seen to roughly cluster them into multiple different types, characterized now by the type of procedure followed, e.g. filtration and drying with sulfates, or washing and removal of catalyst. The clustering no longer correlates with the associated reaction types, which is to be expected given the generality of some of these methods. Other segment classes and colorings are shown in the Appendix C.

To demonstrate the power of the data obtained through the approach described in this work, a number of reaction prediction models were trained, as specified in Section 2.4. The results in Table 1 show that, as expected, including NMR data improves reaction prediction by a small margin in a large, non-specific test set, but also in the more specific Buchwald-Hartwig set, where baseline accuracy is already >95%. Scaling of the models also improve performance, expectedly as NMR data increases the vocabulary size of the transformer encoder.

Input type	Model size	USPTO Test acc (%)	Buchwald-Hartwig Test acc (%)
SMILES	20M	85.22	95.49
SMILES + NMR	24M	85.31	96.03
	69M	**86.69**	**97.04**
	133.4M	**86.70**	**97.02**

Table 1: **Reaction prediciton results.** Performance of reaction prediction models for different data accessibility regimes. The effect of additional NMR data is assessed on two distinct test sets. Results show that including NMR data systematically improves performance of models. Especially on the Buchwald-Hartwig test set, where baseline accuracy is already >95%, including experimental NMR data further improves the results, indicating that these data are necessary to resolve the structure of some products.

To illustrate the advantage of a SMILES+NMR model, consider the reaction in Figure 3. In this example, the SMILES-only model fails to predict the correct product, despite it producing correct chemistry, namely by predicting an amide formation reaction. The correct reaction however, as predicted by the NMR-augmented model, installs a triazole heterocycle in the product. Both possibilities are clearly distinguishable from NMR spectra: the product below would display peaks characteristic to the N-H hydrogens, which is not observed in the experimental data. Our model successfully leverages this information to make its prediction.

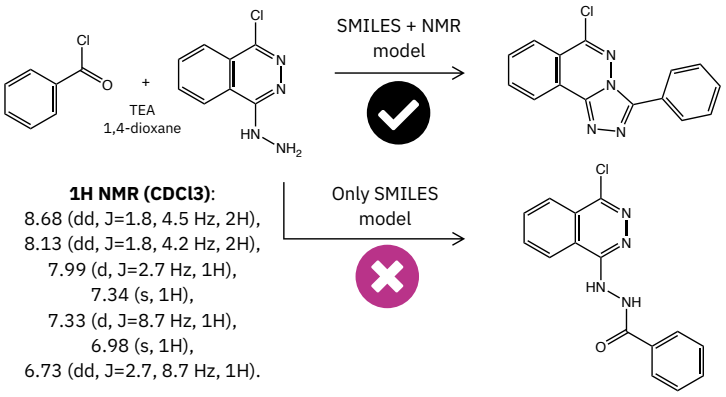

Figure 3: Example reaction that is incorrectly predicted by smiles-only reaction prediction models, but correctly predicted when NMR data is used.

These results not only the capacity of large pre-trained language models at extracting relevant procedural information from reactions, but also the potential of this type of data for applications in cheminformatics, like multi-modal reaction search and chemical space exploration.

4 Conclusions

We have shown an approach for including partially structured procedural information into chemical reaction representations. A model for semantic segmentation of synthesis procedure paragraphs was trained and applied on 1.8M such paragraphs from the USPTO dataset, allowing for their efficient mapping into semantically distinct parts. The model is shown to perform well even in challenging cases, where simpler algorithms could fail by design. This newly-created database allowed us to explore the chemical space in a new, natural language driven manner. Mapping of these semantically distinct segments into high-dimensional vector spaces was shown to encode relevant procedural information, which is entirely ignored in the commonly used reaction SMILES or fingerprints representations, specially in the "reaction set-up" and "work-up" spaces. While the reaction set-up embeddings are seen to cluster reactions by reagent type used or even similarity in conditions, and roughly by reaction type, work-up embeddings encode procedural information like the types of filtering or washing performed. Just as in experimental chemistry, analytical data plays a key role for reaction prediction. We have shown that by including NMR data mined from the USPTO reaction database, reaction prediction models learn leverage this additional information to better resolve products, especially in cases where multiple products are possible.

Analyses of this type allow us to explore the chemical space from a more procedural perspective, an important topic specially with recent advances in one-pot multi-step synthetic chemistry, automated platforms and even self-driving labs. More than that, this work sets the basis for future applications in multi-modal models for reaction search and generation of synthetic procedures. Indeed, understanding the relationships between the different segments, and how one follows from the others, has the potential to unlock enhanced automatized prediction of work-up and purification steps. Future research might involve exploring the role of these extended reaction steps in critical reaction results like reaction yield and selectivities, classification, and procedure generation, among others. Additionally, the proposed semantic segmentation is expected to reduce the noise added in reaction SMILES, thereby directly improving future predictive models on this modality.

Acknowledgments

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

References

[1] Andersen, J. L.; Flamm, C.; Merkle, D.; Stadler, P. F. An intermediate level of abstraction for computational systems chemistry. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2017**, *375*, 20160354, Publisher: Royal Society.

[2] Zuker, M.; Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research* **1981**, *9*, 133–148.

[3] Weinberg, S. *The Quantum Theory of Fields: Volume 1: Foundations*; Cambridge University Press: Cambridge, 1995; Vol. 1.

[4] McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585–590, Number: 5612 Publisher: Nature Publishing Group.

[5] Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *The Journal of Chemical Physics* **2004**, *39*, 1397–1412.

[6] Hückel, E. Quantentheoretische Beiträge zum Benzolproblem. *Zeitschrift für Physik* **1931**, *70*, 204–286.

[7] 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.

[8] 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.

[9] 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.

[10] Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science* **2017**, *3*, 1103–1113.

[11] Dai, H.; Li, C.; Coley, C.; Dai, B.; Song, L. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems* **2019**, *32*.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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**,

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

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

[21] Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T. Inferring experimental procedures from text-based representations of chemical reactions. *Nature Communications* **2021**, *12*, 2573, Number: 1 Publisher: Nature Publishing Group.

[22] Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Cho, K.; Ji, H. Translation between Molecules and Natural Language. 2022.

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

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

[25] Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Is GPT-3 all you need for low-data discovery in chemistry? 2023; `https://chemrxiv.org/engage/chemrxiv/article-details/63eb5a669da0bc6b33e97a35`.

[26] 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.

[27] 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.

[28] Alberts, M.; Zipoli, F.; Vaucher, A. C. Learning the Language of NMR: Structure Elucidation from NMR spectra using Transformer Models. 2023; https://chemrxiv.org/engage/chemrxiv/article-details/64d5e4ccdfabaf06ff1763ef.

[29] Alberts, M.; Laino, T.; Vaucher, A. C. Leveraging Infrared Spectroscopy for Automated Structure Elucidation. 2023; https://chemrxiv.org/engage/chemrxiv/article-details/645df5cbf2112b41e96da616.

[30] Ouyang, L. et al. Training language models to follow instructions with human feedback. 2022.

[31] OpenAI GPT-4 Technical Report. 2023.

[32] Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; Le, Q. V. Finetuned Language Models are Zero-Shot Learners. 2021.

[33] Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* **2022**,

[34] Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. 2021; http://arxiv.org/abs/2106.09685, arXiv:2106.09685 [cs].

[35] Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature machine intelligence* **2021**, *3*, 144–152.

[36] rxn4Chemistry rxn4Chemistry. https://github.com/rxn4chemistry/rxn4chemistry, 2020; Accessed: April 2023.

[37] Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N. C-Pack: Packaged Resources To Advance General Chinese Embedding. 2023.

[38] Muennighoff, N.; Tazi, N.; Magne, L.; Reimers, N. MTEB: Massive Text Embedding Benchmark. 2023.

[39] 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.

[40] Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany, 2016; pp 1715–1725.

[41] 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].

[42] Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; Leahy, C. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. 2020.

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

[44] Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761* **2023**,

[45] Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T. B. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[46] Pfeiffer, J.; Rücklé, A.; Poth, C.; Kamath, A.; Vulić, I.; Ruder, S.; Cho, K.; Gurevych, I. AdapterHub: A Framework for Adapting Transformers. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations. Online, 2020; pp 46–54.

[47] Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine Learning Yield Prediction from NiCOlit, a Small-Size Literature Data Set of Nickel Catalyzed C–O Couplings. *Journal of the American Chemical Society* **2022**, *144*, 14722–14730, Publisher: American Chemical Society.

[48] Jiang, S.; Zhang, Z.; Zhao, H.; Li, J.; Yang, Y.; Lu, B.-L.; Xia, N. When SMILES Smiles, Practicality Judgment and Yield Prediction of Chemical Reaction via Deep Chemical Language Processing. *IEEE Access* **2021**, *9*, 85071–85083.

[49] 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.

[50] 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.

[51] Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Jannik Bjerrum, E. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical Science* **2020**, *11*, 154–168, Publisher: Royal Society of Chemistry.

A Language Models

Language models have revolutionized the field of natural language processing thanks to recent advancements in model design[41], along with a wide availability of text datasets[42] and capacity to scale to large computational budgets. These models are generally trained to predict the likelihood of tokens in text sequences. The most successful models in the field, the Transformers[41], employ an attention mechanism[43] to weigh the importance of each word in a sentence when predicting the next word, thereby learning to extract long-rage dependencies from text sequences.

Two relevant pretrained models are GPT-4[31] and Flan-T5[32]. These state-of-the-art models have been built and trained for different purposes, and thus serve different purposes.

A.1 GPT-4

GPT-4 is a decoder-only model developed by OpenAI[31] trained with an autoregressive objective on large text datasets to generate human text. Capabilities of this, and similar models, include translation, question-answering and general content creation, however additional capabilities have been demonstrated such as chain-of-though reasoning[33], in-context learning[18], and capacity to use tools[44].

In combination, these capabilities make it possible for users to solve generic NLP problems by simply prompting the model with explanations about how to complete the task, along with examples and other relevant information.

A.2 FLAN-T5

FLAN-T5 is a model developed by Google[32] whose training paradigm is that any NLP problem is a text-to-text problem. Under this setting, instead of training individual models for each task, T5 unifies a number of tasks into a single framework as a text generation task.

For our purposes, a key property of the FLAN-T5 model is that it can be fine-tuned to perform any text-to-text task, for which enough data is available, which yields a model with a small parameter count, facilitating local inference and escalation to large datasets. GPT-4, on the other hand, can only be accessed through a costly API, that is additionally restricted to one API call per generation, preventing batches of data to be processed.

B Semantic segmentation model

As discussed, the analysis centers initially in decomposing synthesis procedure paragraphs into semantically distinct segments belonging to different classes, namely "reaction set-up", "work-up", "purification" and "analysis". An example of a solution to this task is given in Figure 4. Solving this task requires that the model learns to copy and paste text from the input, into the output, however separating the different segments based on their meaning in the context of a chemical synthesis, while also assigning a label or class to each.

The task cannot be trivially formulated as a per-sentence classification task as, as shown in Figure 4, some segments can actually extend up to the first words of the next sentence and beyond, as is the case of the piece "Stir for 30 hours,", which belongs together with the "reaction set-up" segment. The semantic segmentation task thus requires certain level of contextual understanding, making (large) language models suitable candidates for solving the task.

B.1 Knowledge Distillation

Knowledge distillation is the process where the knowledge from a more capable model is *distilled* to be transfered to another, potentially cheaper model[45]. In the case of the paragraph segmentation task, we found that LLMs like GPT-3.5 and GPT-4 excelled when adequately prompted, thanks to their demonstrated abilities to follow instructions and formats, and produce step-by-step reasoning sequences. The following text was used as a template prompt to achieve the desired behavior from the models:

Input paragraph:

Suspend anhydrous AlCl3 (156 g, 1.15 mol) in toluene (1500 mL) and cool to 2-4° C. Add, by slow addition, a solution of 4-chlorobutyryl chloride (165.5 g, 1.15 mol) in toluene (300 mL). Stir for 15 minutes and pour into stirring ice-water (2.5 L). Stir for 30 hours, decant the toluene and extract the aqueous phase with toluene (700 mL). Combine the organic layers and wash three times with water (1 L, 1 L, 500 mL). Evaporate the solvent in vacuo to give the title compound as a pale yellow oil (292.3 g, 95%).

Output segmentation:

```
{
    'text segment': 'Suspend anhydrous AlCl3 (156 g, 1.15 mol) in toluene (1500 mL)
and cool to 2-4° C. Add, by slow addition, a solution of 4-chlorobutyryl chloride
(165.5 g, 1.15 mol) in toluene (300 mL). Stir for 15 minutes and pour into stirring
ice-water (2.5 L). Stir for 30 hours,',
    'text class': 'reaction set-up',
    'step order': 1,
}
{
    'text segment': 'decant the toluene and extract the aqueous phase with toluene
(700 mL). Combine the organic layers and wash three times with water (1 L, 1 L,
500 mL).',
    'text class': 'work-up',
    'step order': 2
},
 {
    'txt_sgm': 'Evaporate the solvent in vacuo to give the title compound as a pale
yellow oil (292.3 g, 95%).',
    'segment class': 'purification',
    'step order': 3
}
```

Figure 4: Example of the semantic segmentation task for synthetic procedure paragraphs. The color code shows the origin of each extracted segment from the original paragraph.

You are an adept experimentalist in chemistry. Your role is to teach new researchers how to recognize reaction steps of a chemical reaction and to chunk the procedure into steps based on steps' meanings in the context of a chemical reaction.

Steps in a chemical reaction have an outline to follow as below:

- 'reaction set-up': the preparation of a chemical synthesis procedure, where reactants, solvents, and catalysts are specified. Specific conditions in which the reaction is initiated, such as temperature, pressure, atmosphere, are indicated. Chemical treatments may come along to stop the reaction, such as the portionwise addition of acid, base, water or liquid.
- 'work-up': the process of isolating the desired product from the reaction mixture after the chemical reaction has taken place. It always comes after the completion of reaction-set up in order to separate products from unreacted starting materials, byproducts, and other impurities. Common techniques in work-up includes quenching, extraction, washing, phase separation, evaporation and filtration. Some key words of work-up steps in sentence include 'adding acid (ex. HCL, H2SO4) or base (ex. NaOH) into reaction mixture/residue', 'cooling the mixture to ambient temperature or below 0 degree celsius', 'solvents being removed/filtered/concentrated by rotary evaporation', 'diluting the solution or forming two layers to do extraction'.
- 'purification': Purification is the process of removing impurities and unwanted byproducts from the desired product to obtain a pure compound. It sometimes comes after the work-up step to obtain a high-quality product with the desired properties. Common purification techniques include crystallization, recrystallization, chromatography, and distillation.

- 'analysis': Analysis refers to the characterization and evaluation of the synthesized product to confirm its identity, purity, and properties. This step involves the use of various analytical techniques to determine the product's structure, composition, and physical properties. Common analytical methods include melting point determination, nuclear magnetic resonance (NMR) spectroscopy, infrared spectroscopy (IR), mass spectrometry (MS), Ultraviolet-visible (UV-Vis) spectroscopy, and X-ray crystallography. "Assay", "analysis" are key words usually found in analysis steps.

To do the task, please follow the approach:

1. First, you receive a paragraph of text 'input'. Read the paragraph clause-by-clause (ps. a clause means a group of words separated by a semicolon(;), a comma(,), or a period(.)); when reading a sentence, reason the meaning of this individual reaction step to a chemical reaction by recognizing the keywords; label in mind this reaction step by thinking of their meaning in the context
2. Then, start chunking the paragraph as output
3. Finally, when giving the output, give directly the formatted output; do not output your reasonings on how to chunk the paragraph

To chunk the text, you must follow the format below:

text segment: text segment from step 1 goes here

text class: the category of the segment; it can be 'reaction set-up', 'work-up', 'purification', or 'analysis'

explanation: the explanation of this step; write down why you assign the class to this segment and why you think the next part of text differs from this segment.

step order: the number of steps already done, starting from 'No.1'

Step end #.

You should follow key points below when chunking; the key points are given in order of importance:

1. Copy literally the text; do not paraphrase the text when transcribing texts into a segment.
2. If a sentence contains information that pertains to two different text classes, divide this sentence into 2 steps
3. If the segmented text has the same text class as its preceding segment, this segmented text should be involved into the preceding segment; if the segmented text has the same text class as its following segment, this segmented text should be involved into the following segment.

Here's a ground truth example you could take into consideration:

{example}

Here's the paragraph you need to complete:

{paragraph}

Think step-by-step. Then give the output!

Begin!

The placeholder *example* is replaced by the text below, that gives an idea to the LLM of what the output should look like.

Input:

Methyl (1R)-2-[(2S,4S)-2-(5-2-[(2S,4S)-1-(2S)-2-[(methoxycarbonyl)amino]-3-methylbutanoyl-4-methylpyrrolidin-2-yl]-1,11-dihydroisochromeno[4',3':6,7]naphtho[1,2-d]imidazol-9-yl-1H-imidazol-2-yl)-4-(methoxymethyl)pyrrolidin-1-yl]-2-oxo-1-phenylethylcarbamate: Tert-butyl (2S,4S)-2-[5-(2-(2S,4S)-1-[N-(methoxycarbonyl)-L-valyl]-4-methylpyrrolidin-2-yl-1,11-dihydroisochromeno[4',3':6,7]naphtho[1,2-d]imidazol-9-yl)-1H-imidazol-2-yl]-4-(methoxymethyl)pyrrolidine-1-carboxylate (166 mg, 0.21 mmol) was dissolved in DCM (4 mL), MeOH (1 mL) and HCl (4 M in dioxane, 1 mL) was added. The reaction mixture was stirred for 2 h and then

concentrated under reduced pressure. The crude residue was treated with (R)-2-(methoxycarbonylamino)-2-phenylacetic acid (44 mg, 0.21 mmol), COMU (100 mg, 0.21 mmol) and DMF (5 mL), then DIPEA (0.18 mL, 1.05 mmol) was added dropwise. After 1 h, the mixture was diluted with 10% MeOH/EtOAc and washed successively with saturated aqueous NaHCO3 and brine. The organics were dried over MgSO4, filtered and concentrated under reduced pressure. The crude residue was purified by HPLC to afford title compound (71 mg, 38%). LCMS-ESI+: calculated for C49H54N8O8: 882.41; observed [M+1]+: 884.34. 1H NMR (CD3OD): 8.462 (s, 1H), 8.029-7.471 (m, 7H), 7.394-7.343 (m, 5H), 5.410 (d, 2H, J=6.8 Hz), 5.300 (m, 1H), 5.233 (m, 2H), 4.341 (m, 1H), 4.236 (d, 1H, J=7.2 Hz), 3.603 (s, 3H), 3.551 (s, 3H), 3.522-3.241 (m, 8H), 2.650 (m, 1H), 2.550 (m, 2H), 1.977-1.926 (m, 4H), 1.221 (d, 3H, J=3.2 Hz), 0.897-0.779 (dd, 6H, J=19.2, 6.8 Hz).

Let's think step by step before giving the output:

1. Let's read the first sentence, "Tert-butyl (2S,4S)-2-[5-(2-(2S,4S)-1-[N-(methoxycarbonyl)-L-valyl]-4-methylpyrrolidin-2-yl-1,11-dihydroisochromeno[4',3':6,7]naphtho[1,2-d]imidazol-9-yl)-1H-imidazol-2-yl]-4-(methoxymethyl)pyrrolidine-1-carboxylate (166 mg, 0.21 mmol) was dissolved in DCM (4 mL), MeOH (1 mL) and HCl (4 M in dioxane, 1 mL) was added. " In this sentence, reactants (Tert-butyl (2S,4S)-2-[5-(2-(2S,4S)-1-[N-(methoxycarbonyl)-L-valyl]-4-methylpyrrolidin-2-yl-1,11-dihydroisochromeno[4',3':6,7]naphtho[1,2-d]imidazol-9-yl)-1H-imidazol-2-yl]-4-(methoxymethyl)pyrrolidine-1-carboxylate, MeOH, and HCl) and sovlent (DCM), together with their amounts and concentrations, are given.
2. As reactants, solvents, and catalysts are specified in the reaction set-up step, and as reactants and solvents are given in this sentence, this sentence should be categorized as 'reaction set-up'.
3. Let's read the next sentence, "The reaction mixture was stirred for 2 h and then concentrated under reduced pressure." In this sentence, the duration of the reaction (2 h) and the pressure under which the reaction was undergone (redcued pressure) are given.
4. The step giving the reaction condition is a 'reaction set-up' step; thus, this sentence is categorized as 'reaction set-up'.
5. Let's move on to the next sentence, "The crude residue was treated with (R)-2-(methoxycarbonylamino)-2-phenylacetic acid (44 mg, 0.21 mmol), COMU (100 mg, 0.21 mmol) and DMF (5 mL), then DIPEA (0.18 mL, 1.05 mmol) was added dropwise." In this step, the acid ((R)-2-(methoxycarbonylamino)-2-phenylacetic acid), and other liquids (COMU, DMF, DIPEA) are added to stop the reaction.
6. Given that the step describes how to stop the reaction, it is categorized as a reaction set-up step.
7. In next sentence, "After 1 h, the mixture was diluted with 10% MeOH/EtOAc and washed successively with saturated aqueous NaHCO3 and brine", the clause "after 1 h" tells the duration to wait before the work-up get started. Hence, this is a reaction set-up step. Then, the sentence "the mixture was diluted with 10% MeOH/EtOAc and washed successively with saturated aqueous NaHCO3 and brine" specifies approaches to isolate desired products from the reaction mixture ('diluted' with 10% MeOH/EtOAc, 'washed' successively with saturated aqueous NaHCO3 and brine). Thus, it is a 'work-up' step.
8. The next sentence, 'The organics were dried over MgSO4, filtered and concentrated under reduced pressure', indicates actions to isolate product from mixture ('dried' over MgSO4, 'filtered' and 'concentrated' under reduced pressure). Thus, it is a work-up step.
9. In the next sentence, 'The crude residue was purified by HPLC to afford title compound (71 mg, 38%)', the verb 'purify' is mentioned and a purification method (HPLC) is given; therefore, it is a purification step.

10. Next, a series of characterization data (LCMS-ESI+: calculated for C49H54N8O8: 882.41; observed [M+1]+: 884.34. 1H NMR (CD3OD): 8.462 (s, 1H), 8.029-7.471 (m, 7H), 7.394-7.343 (m, 5H), 5.410 (d, 2H, J=6.8 Hz), 5.300 (m, 1H), 5.233 (m, 2H), 4.341 (m, 1H), 4.236 (d, 1H, J=7.2 Hz), 3.603 (s, 3H), 3.551 (s, 3H), 3.522-3.241 (m, 8H), 2.650 (m, 1H), 2.550 (m, 2H), 1.977-1.926 (m, 4H), 1.221 (d, 3H, J=3.2 Hz), 0.897-0.779 (dd, 6H, J=19.2, 6.8 Hz).) are given. Analytical techinques (LCMS-ESI+, 1H NMR) are specified, which shows that the step is an 'analysis' step.
11. Integrate segmented paragraphs with the same category into a segment and then give the formatted output

Output:

text segment: 'Tert-butyl (2S,4S)-2-[5-(2-(2S,4S)-1-[N-(methoxycarbonyl)-L-valyl]-4-methylpyrrolidin-2-yl-1,11-dihydroisochromeno[4',3':6,7]naphtho[1,2-d]imidazol-9-yl)-1H-imidazol-2-yl]-4-(methoxymethyl)pyrrolidine-1-carboxylate (166 mg, 0.21 mmol) was dissolved in DCM (4 mL), MeOH (1 mL) and HCl (4 M in dioxane, 1 mL) was added. The reaction mixture was stirred for 2 h and then concentrated under reduced pressure. The crude residue was treated with (R)-2-(methoxycarbonylamino)-2-phenylacetic acid (44 mg, 0.21 mmol), COMU (100 mg, 0.21 mmol) and DMF (5 mL), then DIPEA (0.18 mL, 1.05 mmol) was added dropwise. After 1 h, ', text class: reaction set-up, explanation: this is the reaction set-up because the main reactants (Tert-butyl (2S,4S)-2-[5-(2-(2S,4S)-1-[N-(methoxycarbonyl)-L-valyl]-4-methylpyrrolidin-2-yl-1,11-dihydroisochromeno[4',3':6,7]naphtho[1,2-d]imidazol-9-yl)-1H-imidazol-2-yl]-4-(methoxymethyl)pyrrolidine-1-carboxylate), MeOH and HCL were added into the solvent (DCM). Also, the time of the reaction undergoing (stir for 2h for the reaction mixture, 1h for the crude residue), the condition of the reaction (reduced pressure) and the portionwise addition of (R)-2-(methoxycarbonylamino)-2-phenylacetic acid, COMU and DMF are specified in this paragraph. step order: 1

Step end #

text segment: 'the mixture was diluted with 10% MeOH/EtOAc and washed successively with saturated aqueous NaHCO3 and brine. The organics were dried over MgSO4, filtered and concentrated under reduced pressure.'

text class: work-up

explanation: this is the work-up step because the paragraph states that the products is diluted, washed, dried, filtered and concentrated.

step order: 2

Step end #

text segment: 'The crude residue was purified by HPLC to afford title compound (71 mg, 38%).'

text class: purification

explanation: this is the purification step because the purification method (HPLC) is indicated in this segment.

step order: 3

Step end #

text segment: 'LCMS-ESI+: calculated for C49H54N8O8: 882.41; observed [M+1]+: 884.34. 1H NMR (CD3OD): 8.462 (s, 1H), 8.029-7.471 (m, 7H), 7.394-7.343 (m, 5H), 5.410 (d, 2H, J=6.8 Hz), 5.300 (m, 1H), 5.233 (m, 2H), 4.341 (m, 1H), 4.236 (d, 1H, J=7.2 Hz), 3.603 (s, 3H), 3.551 (s, 3H), 3.522-3.241 (m, 8H), 2.650 (m, 1H), 2.550 (m, 2H), 1.977-1.926 (m, 4H), 1.221 (d, 3H, J=3.2 Hz), 0.897-0.779 (dd, 6H, J=19.2, 6.8 Hz).'

text class: analysis

explanation: this is the analysis as the analytical methods (LCMS-ESI+, 1H NMR (CD3OD)) are given in this paragraph.

step order: 4

Step end #

B.2 Model training

Nearly 30k samples were obtained from GPT-4 and GPT-3.5 using the prompt above. To transfer this task to a smaller specialist model, we fine-tuned a **flan-t5-large** model using the adapters[46] library.

To fully profit from the generated dataset, a 2-stage training procedure was followed, where at first the model is fine-tuned on the more abundant –however potentially less accurate– GPT-3.5 dataset in order for it to learn the format and an initial representation of what the task is about. The model is subsequently fine-tuned on the GPT-4 dataset, which is more scarse but assumed to be better quality.

For every stage of training a batch size of 2 was used, over 20 epochs, with a linear learning rate decay starting from 5e-4.

B.3 Output post-processing

Although the resulting model behaves well in multiple situations, in some cases it can generate erroneous outputs by copying the same sentence multiple times, or by missing some text in the output. These cases can easily be detected by calculating the edit distance between the original paragraph and the concatenation of all the output segments which, if correctly done, should equal zero.

With this, we found that the resulting model produces output with satisfactory results in around 66% of cases. This filtering technique is further extended to the inference step to the whole USPTO database, to ensure data quality.

C Segment Embedding Maps

To explore the rich structure of the newly defined semantic subspaces, the sentence embeddings for each segment were calculated and plotted using different labels, in order to facilitate pattern-finding. Yield was chosen as it was readily available as a part of the dataset; the resulting plots are shown in Figure 5. As can be seen, despite the rich structure observed in each space, there is very little correlation with yield. Although some localization of colors can be seen in e.g. work-up and purification, it must be noted that these two types of segments typically contain the yield textually, so the patterns shown may be an artifact. Still, as previously noted by other authors, yield prediction is a very challenging issue[47–50], due to the noisy nature of data[51] and other social factors such as lack of overlap of different research works[47].

Inspection of the purification and analysis plots (Figure 5c,d) shows even more structure than the other two, however these are less interesting as clustering in this case is correlated with clearly defined concepts in each subspace, such as different types of purification, or the multiple analytical techniques. A more in-depth exploration of these spaces would be required to discover new insights, such as for instance clusterings by type of products in the analysis space, which would make sense knowing that results from analytical chemistry typically encode structural information about the analysed susbtances.

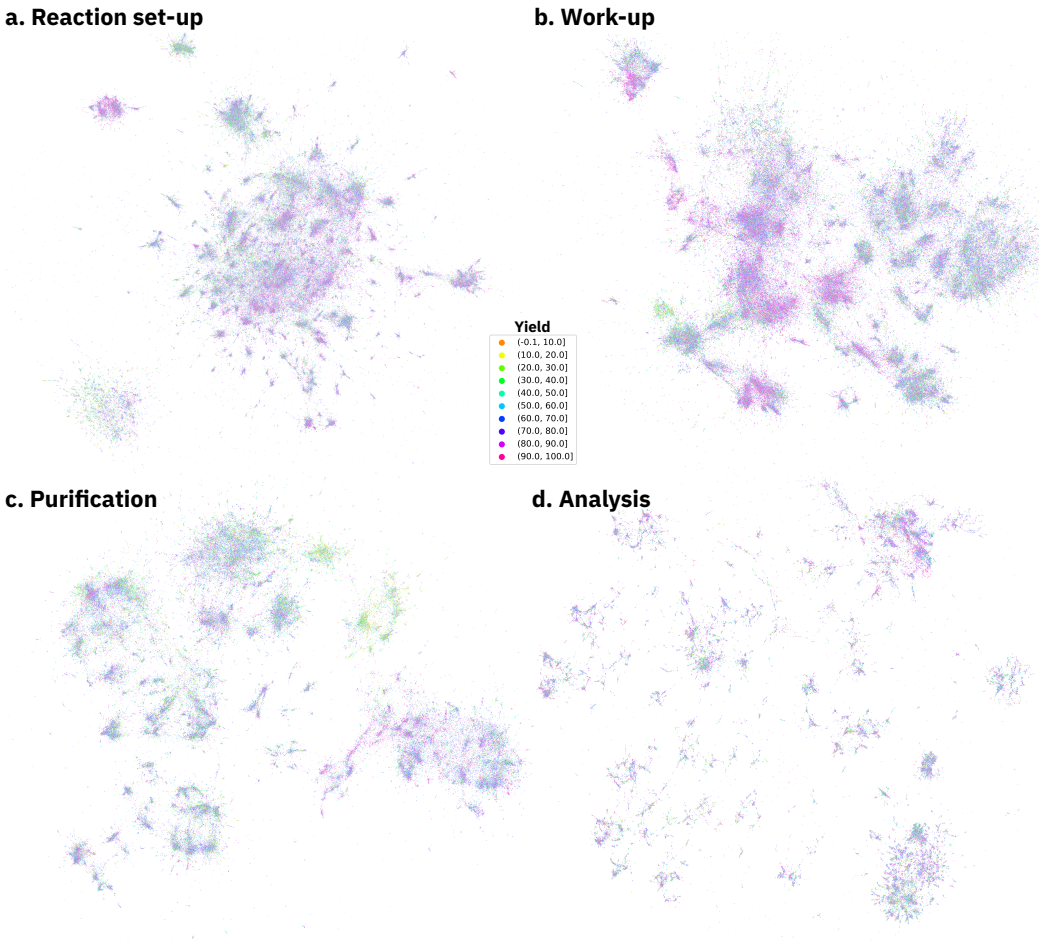

Figure 5: UMAP of each of the defined semantic subspaces, as colored by reaction yield.