doi:10.2533/chimia.2023.484 Chimia 77 (2023) 484–488 © T. Laino *et al.*

Fueling the Digital Chemistry Revolution with Language Models

Antonio Cardinale[a], Alessandro Castrogiovanni[a], Theophile Gaudin[a], Joppe Geluykens[a], Teodoro Laino*[ab], Matteo Manica[a], Daniel Probst[a], Philippe Schwaller[ab], Aleksandros Sobczyk[a], Alessandra Toniato[ab], Alain C. Vaucher[ab], Heiko Wolf[a], and Federico Zipoli[ab]

Sandmeyer Award 2022

Abstract: The RXN for Chemistry project, initiated by IBM Research Europe – Zurich in 2017, aimed to develop a series of digital assets using machine learning techniques to promote the use of data-driven methodologies in synthetic organic chemistry. This research adopts an innovative concept by treating chemical reaction data as language records, treating the prediction of a synthetic organic chemistry reaction as a translation task between precursor and product languages. Over the years, the IBM Research team has successfully developed language models for various applications including forward reaction prediction, retrosynthesis, reaction classification, atom-mapping, procedure extraction from text, inference of experimental protocols and its use in programming commercial automation hardware to implement an autonomous chemical laboratory. Furthermore, the project has recently incorporated biochemical data in training models for greener and more sustainable chemical reactions. The remarkable ease of constructing prediction models and continually enhancing them through data augmentation with minimal human intervention has led to the widespread adoption of language model technologies, facilitating the digitalization of chemistry in diverse industrial sectors such as pharmaceuticals and chemical manufacturing. This manuscript provides a concise overview of the scientific components that contributed to the prestigious Sandmeyer Award in 2022.

Keywords: Digital Chemistry · Language Models · Machine Learning · Sandmeyer Award 2022 · Synthetic Organic Chemistry

Group photo taken at the Swiss Chemistry Science Night 2022, on the occasion of the celebration of the Sandmeyer Award 2022. From left to right: Heiko Wolf, Matteo Manica, Alessandra Toniato, Federico Zipoli, Alain Claude Vaucher, Philippe Schwaller, Theophile Gaudin, Aleksandros Sobczyk, Teodoro Laino. Team members not present in the photo: Antonio Cardinale, Alessandro Castrogiovanni, Daniel Probst, Joppe Geluykens. Photo: © SCS

1. Introduction

For over two centuries, the synthesis of organic molecules has played a vital role in organic chemistry, with significant scientific and commercial implications that affect the lives of billions of individuals. However, traditional laboratory practices in this field have seen little disruption. The vast number of reaction classes, continually expanding with new discoveries, poses a challenge for organic chemists working in domains such as materials science, oil and gas, and life sciences. While experts may recall a limited number of reactions within their specific field, becoming an expert generalist is difficult. It typically requires decades of experience to master reactivity patterns and reaction rules through a series of experiments, testing these human-designed rules in diverse reactivity contexts.

Artificial intelligence (AI) has rapidly advanced in various domains, from voice assistants on smartphones to self-driving cars. AI has the potential to significantly increase productivity. By automating mundane or hazardous tasks, AI liberates human labor to focus on endeavors that demand creativity, empathy, and other uniquely human skills.

In 2017, the IBM Research team (or RXN team in the following for short) embarked on a mission to promote the adoption of data-driven technologies in the chemical community, fostering a digital approach to synthetic chemistry. A pioneering concept employed by the team involved treating organic chemistry as a language.

But what defines a ‘language’? Experts characterize language as a system of spoken, written, and signed symbols accompanied by grammar rules, facilitating human communication. With approximately 6,500 languages existing worldwide, each reflecting a specific culture, chemistry, like any other natural science, represents a language of its own. Modeling chemical language holds the potential to capture the behavior of the physical laws governing chemical phenomena without resorting to solving com-

**Correspondence*: Dr. T. Laino[ab], E-mail: teo@zurich.ibm.com
Authors are listed in alphabetical order. [a]IBM Research Europe - Zurich, Säumerstrasse 4, Rüschlikon, CH-8803, Switzerland; [b]National Center for Competence in Research-Catalysis (NCCR-Catalysis), Zurich, Switzerland

plex mathematical equations, analogous to how linguists explore language to gain insights into humanity.

The connection between language models and chemistry elicits both surprise and intrigue. On one hand, the vast diversity among languages demonstrates that, despite the universality of language among humans, there exists a substantial degree of variation. Thus, the parallelism between a domain-specific language like chemistry and a natural language should not be unexpected. On the other hand, the structure, grammar, and inflection observed across different languages exemplify the evolution and contextual adaptation of language. Similarly, chemistry, along with all other natural sciences, has evolved over time through the continual accumulation and interpretation of new experimental evidence. While physical laws provide a detailed description of natural phenomena, language captures their coarse representation. This inherent relationship between chemistry and languages places language modeling in a favorable position compared to mathematical laws governing physical phenomena. In fact, within the realm of chemistry, language models prove to be a more successful approach in predicting the behavior of physical systems, such as chemical reactions, compared to solving complex equations involving many-body interactions. Nonetheless, physical laws remain highly effective in explaining observable phenomena. The application of language modeling extends well beyond chemistry and finds relevance in various branches of natural sciences.

Although there is no mathematical proof that the set of reaction rules, generalized over centuries of organic chemistry, constitutes a consistent grammar of a domain language, the team chose to put the hypothesis to test with the use of state-of-the-art neural machine translation methods. They treated the forward reaction prediction problem as a translation task, employing sequence-to-sequence (seq2seq) models.[1] At that time, the results outperformed existing data-driven solutions by achieving a top-1 accuracy of 80.3% on their own training and test sets.[1] In 2018, the team introduced a new AI architecture for reaction prediction called the Molecular Transformer,[2] pioneering the use of today widely adopted Transformer architecture developed for neural machine translation tasks. The RXN team made the software, trained models, and patent data used for training freely available. Even after more than five years, Molecular Transformers remain the best-performing data-driven models in the field of forward reaction prediction. Over 650 additional papers, authored by numerous research groups worldwide, have utilized this architecture for diverse scientific contributions in the realm of digital synthetic organic chemistry demonstrating the importance of language models in chemoinformatic tasks (Fig. 1).

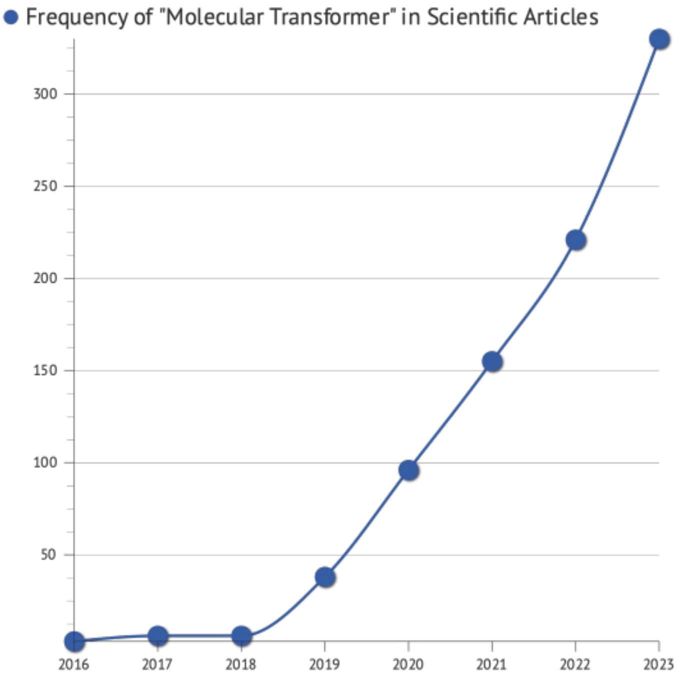

Fig. 1. Frequency of appearance of the term ‘Molecular Transformer’ in scientific publications. Data extracted from Google Scholar: # Years on the x-axis ‘years = [2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023]’; # Frequency of ‘Molecular Transformer’ on the y-axis ‘frequency = [3, 6, 6, 38, 96, 155, 221, 330]’. Data for 2023 is extrapolated based on the trend of the first 6 months (165).

2. Language Models for Digital Chemistry Tasks

The use of language models in the chemical domain requires the establishment of a machine-processable domain language. While humans traditionally employ diagrammatic representations for inputting and representing molecules and reactions, the majority of machine-learning applications in molecular studies rely on line notation, known as SMILES (Simplified Molecular Input Line Entry System). The SMILES notation encapsulates the same information as diagrammatic or graphical representations. However, the SMILES notation proves more advantageous in the context of language modeling as it embodies a linguistic concept rather than a computer data structure. SMILES can be regarded as a bona fide language, albeit with a limited vocabulary of atom and bond symbols and a small number of grammar rules. The succinctness of the SMILES notation, compared to other line notations, facilitates the learning of language models especially in low data regimes.

In the realm of chemical language modeling, the SMILES notation produces ‘words’ to represent molecules (including reactants, reagents, and products) that are combined in ‘sentences’, *i.e.,* chemical reactions. This approach has emerged as one of the most effective and scalable methods for capturing human knowledge and modeling processes and tasks specific to synthetic organic chemistry. These tasks encompass predicting chemical reactions,[1,2] designing retrosynthetic routes,[3] digitizing chemical literature,[4] predicting detailed experimental procedures[5] and yields,[6] designing novel fingerprints,[7] and facilitating the use of biocatalysis in synthetic processes.[8] For instance, predicting the outcome of a reaction involving multiple reactants or reagents is framed as a translation task, converting a sentence representing precursors using the SMILES notation into a sentence representing products using the same notation.

The application of language models in chemistry extends beyond the realm of the digital experience, as language architectures serve as pivotal enablers[5] in the development of autonomous laboratories for chemical compound synthesis. Remote access to these laboratories is made feasible through cloud technologies.[9] Furthermore, language models have recently demonstrated their full potential in various areas of chemical synthesis. These include the ability to recover three-dimensional features from one-dimensional sequences when trained with biocatalytic data,[10] obviating the need for laborious data curation campaigns to eliminate noise in datasets,[11] and emulating human performance in the design of chemical disconnection strategies.[12] These examples underscore the wide-ranging applications where language models exhibit their prowess.

2.1 *Forward Reaction Prediction*

Among the initial models developed, one notable example is the forward prediction model,[2] which does not explicitly incorporate chemical knowledge or employ reaction templates. Instead, it relies solely on the chemical information contained within the reaction data. In our study,[2] we demonstrated the superiority of a multihead attention Molecular Transformer model over existing algorithms, achieving a top-1 accuracy exceeding 90% on a widely used benchmark dataset. The Molecular Transformer model generates predictions by inferring the correlations between the presence or absence of chemical motifs in the reactant, reagent, and product entities present in the dataset. Notably, our model does not rely on handcrafted rules and exhibits remarkable accuracy in predicting subtle chemical transformations. Importantly, our model is capable of accurately estimating its own uncertainty, as demonstrated by an uncertainty score that achieves an 89% accuracy in classifying the correctness of predictions. Furthermore, the forward prediction model based on Transformer can effectively handle inputs without a distinct reactant–reagent split and accommodate stereochemistry, thereby establishing the universal applicability of our approach. In situations where specialization of the model for specific classes with limited coverage is desired, transfer learning of pre-trained models on a comprehensive database of chemical reactions can be employed. By utilizing a small number of chemical reaction examples, the accuracy of our RXN model[13] can be significantly enhanced for reaction classes that were not encountered during the original training phase. Over the years, the forward prediction model has undergone further refinement to improve performance across various metrics, including the diversity of predictions.[14]

2.2 *Retrosynthesis*

For implementing a retrosynthetic analysis, we extended the capabilities of the Molecular Transformer model by incorporating a hyper-graph exploration strategy, enabling automatic retrosynthesis route planning without the need for human intervention. Importantly, the single-step retrosynthetic model utilizes the same Molecular Transformer architecture and training data as the forward reaction prediction model. The single step proved to be superior to other approaches by not only predicting reactants but also providing accurate predictions for reagents, solvents, and catalysts at each retrosynthetic step. To assess the performance of these single-step retrosynthetic models, we introduced four metrics: coverage, class diversity, round-trip accuracy, and Jensen-Shannon divergence. These metrics are evaluated using the forward prediction model and a reaction classification model, both based on the transformer architecture.

Our framework dynamically constructs a hypergraph, where nodes are filtered and expanded based on Bayesian-like probability.[3] Through a thorough evaluation using various retrosynthesis examples from literature and academic exams, we critically assess the end-to-end framework. Overall, the framework demonstrates outstanding performance, although a few limitations related to the training data are observed.

The introduction of the novel metrics presented an opportunity to optimize entire retrosynthetic frameworks, using exclusively the performance of the single-step model.

2.3 *Converting Experimental Procedures to Concise Actions*

The RXN team developed and implemented an AI model designed to extract information on experimental methods for conducting chemical reactions from scientific literature. The AI model ingests experimental procedures (recipes) written in prose, and converts them into a streamlined sequence of concise actions, formatted in a manner that is amenable to automation. Unlike other similar tools that rely on rule-based approaches and often struggle with noisy and variable experimental procedures, our approach is entirely data-driven. It does not depend on explicitly formulated rules for parsing sentences and extracting relevant information; instead, it learns the relevant patterns autonomously. One of the key advantages of this data-driven approach is its reliance solely on data, making it easily improvable by adding more examples.

In practice, our team pre-trained the model using a large volume of automatically generated data and fine-tuned it using a manually annotated set of high-quality sentences, employing a custom-designed framework. We extensively utilized this technology to construct the Smiles2Actions dataset, which comprised one million experimental procedures. The findings and advancements achieved through this work were published in the peer-reviewed journal *Nature Communications* in 2020.[4] This same technology is also employed in RoboRXN, enabling a groundbreaking feature known as ‘copy and paste to synthesis’, where users can directly execute synthesis programs generated by the AI models by copying and pasting experimental procedure paragraphs. This represents a significant advancement in integrating AI with commercial robotic hardware.

2.4 *Inferring Experimental Procedures*

Leveraging the dataset of experimental protocols, we developed an advanced AI model known as Smiles2Actions, which has been trained on a vast dataset of over 1 million experimental procedures. This model possesses the ability to learn the intricacies of chemical reactions and can provide recommendations for the optimal sequence of operations required to synthesize a specific target molecule. This groundbreaking technology serves as the foundation for RoboRXN. What sets Smiles2Actions apart is its capacity to not only determine the necessary steps for a requested chemical reaction but also utilize its embedded knowledge to generate a comprehensive set of instructions for executing a chemical reaction that may not have been part of its training data. This represents a significant advancement, as previous approaches were limited to predicting individual reaction parameters or optimizing synthesis parameters within a specific reaction class.

From an IT perspective, the Smiles2Actions model can be likened to an AI architecture that writes programs for molecular synthesis. When integrated with RoboRXN, this AI model eliminates the laborious task of manually programming commercial automation hardware, allowing chemists to allocate their time towards more creative endeavors. The details of this remarkable work have been published in *Nature Communications* in 2021.[5]

2.5 *Automatic Curation of Datasets*

Ensuring the accuracy and reliability of training datasets is crucial for the successful application of machine-learning models in the field of chemistry. However, the presence of chemically incorrect entries in these datasets can have a negative impact on the user experience. Learning from a significant number of erroneous examples can distort the representation of chemical rules within the models, leading to biased predictions that may involve unreasonable connections and disconnections. Current approaches for removing noise from datasets rely heavily on domain experts.

To address this challenge, the RXN team has developed a highly effective method for reducing noise in chemical reaction datasets, thereby enhancing the performance of predictive models. The core concept behind this method is based on the phenomenon known as ‘catastrophic forgetting’, which refers to the tendency of artificial intelligence (AI) models to forget previously learned information when trained on new tasks. Similar to language models, where difficult data points often indicate instances of incorrect or incoherent grammar, the most challenging examples to learn during the training of reaction prediction models are likely to be cases of incorrect chemistry when compared to the prevailing chemical grammar described by the majority of the dataset.

The team’s innovative approach, featured in a publication in *Nature Machine Intelligence* in 2021,[11] addresses this issue by mitigating catastrophic forgetting and reducing the impact of chemically incorrect examples in training data. By incorporating this method, the RXN team has made significant strides in improving the reliability and accuracy of predictive models in chemistry.

2.6 *Reaction Classification*

The process of classifying chemical reactions into distinct reaction classes is a laborious and time-consuming task. It involves identifying the appropriate reaction class template by annotating various aspects of the reactions, such as the number of molecules involved, the reaction center, and the differentiation between reactants and reagents. Leveraging transformer-based language models, the RXN team has developed a technology capable of inferring reaction classes from non-annotated, simple text-based representations of chemical reactions. Through this innovative approach, the team achieved remarkable results, with the best models achieving a classification accuracy of 98.2%. Furthermore, the learned representations, known as reaction embeddings, have proven to be effective as reaction fingerprints, capturing subtle differences between reaction classes even better than traditional reaction fingerprints. These new reaction fingerprints have received widespread recognition as the top-performing fingerprints in third-party assessments.

The groundbreaking work on reaction classification by the RXN team has been prominently featured in *Nature Machine Intelligence* in 2021.[7] Moreover, the utilization of RXN fingerprints has enabled the development of highly accurate models for predicting reaction yields, as highlighted in a publication in *Machine Learning: Science and Technology* in 2021.[6] This research represents a significant advancement in the field of reaction classification and holds immense potential for various applications within the realm of chemistry.

2.7 *Atom Mapping*

The process of atom mapping, which involves determining how atoms in reactants correspond to atoms in products, is a crucial step in compiling 'reaction rules' in chemistry. Traditionally, atom mapping is a labor-intensive experimental task, and when computational methods are employed, it often requires manual annotation of chemical reactions and the development of consistent guidelines. We made a groundbreaking discovery by demonstrating that Transformer Neural Networks can learn atom-mapping information between reactants and products without any supervision or human labeling. By utilizing the attention weights of the Transformer model, the team developed an attention-guided reaction mapper that can extract coherent chemical grammar from unannotated sets of reactions, regardless of the specific chemical system.

The proposed method exhibits remarkable performance in terms of accuracy and speed, as demonstrated in third-party assessments. Even for challenging scenarios involving imbalanced and chemically complex reactions with nontrivial atom mapping, the approach proves effective. These findings were published in *Science Advances* in 2021, underscoring the significance of this research breakthrough.[15] By eliminating the need for laborious manual annotation and leveraging the power of Transformer Neural Networks, the RXN team's approach streamlines the process of atom mapping, providing a more efficient and automated solution. This advancement has the potential to revolutionize the way reaction rules are compiled and accelerate research and development efforts in the field of chemistry.

2.8 *Extension to Biocatalysis*

The integration of enzyme catalysts plays a crucial role in advancing green chemistry practices, promoting sustainability and efficiency in chemical synthesis. However, incorporating biocatalysis into retrosynthetic planning posed challenges in predicting enzymatic activity on unreported substrates, as well as enzyme-specific stereo- and regioselectivity. We expanded upon the Molecular Transformer architecture to incorporate biocatalysis into both forward reaction prediction and retrosynthetic pathway prediction models.[8] Leveraging an extensive dataset of publicly available biochemical reactions, we employ a new class token scheme based on the enzyme commission classification number to capture catalysis patterns among enzymes belonging to the same hierarchy. This allowed us to learn enzymatic knowledge effectively. We introduced a tokenization system based on enzyme classes, enabling the prediction of substrates and enzyme classes given a target product for retrosynthetic pathway prediction. Moreover, we enhanced the encoding of enzymes by incorporating the enzyme commission (EC) number into the reaction SMILES, providing a more standardized representation.

To compile our enzymatic reactions and EC numbers, we extracted data from various databases, including Rhea, BRENDA, PathBank, and MetaNetX. The processing led to the creation of a comprehensive dataset named ECREACT, encompassing enzyme-catalyzed reactions with their respective EC numbers.

The forward reaction prediction model achieves a top-1 accuracy of 49.6%, while the retrosynthetic pathway prediction model achieves a top-1 single-step round-trip accuracy of 39.6%. This work contributed to expanding the applicability of the Molecular Transformer in the realm of biocatalysis,[8] enabling more accurate predictions and facilitating the integration of enzymatic catalysis in the development of greener chemistry processes.

3. Autonomous Chemical Reaction Laboratory

RoboRXN is an ambitious project that aims to revolutionize autonomous synthesis by combining AI, cloud technology, and commercial automation. The concept of RoboRXN was developed to minimize human intervention and enable machine-learning algorithms to autonomously design and execute the production of molecules in a remotely accessible laboratory. To bring their vision to life, the RXN team integrated the AI models reported in the previous sections, retrosynthesis and prediction of experimental procedures, with existing robotic hardware. Instead of relying on human programming, commercial autonomous chemical hardware is now self-programmed using these AI models.

Following the data-driven approach of the entire RXN framework, RoboRXN operates on purely data-driven models. Users can initiate the process by specifying a target molecule and utilizing the retrosynthetic module. This module enables the construction of a retrosynthetic scheme and subsequently launches the synthesis process on the automation hardware. RoboRXN represents a significant advancement in the field of autonomous synthesis, offering a streamlined and efficient approach to manual chemical synthesis. Through the integration of AI, cloud technology, and robotic hardware, it empowers researchers to conduct experiments remotely, reducing the need for manual intervention and accelerating the pace of scientific discovery. Further details about the RoboRXN project will be published in a separate article.

4. Conclusions

The current advancements in natural language modeling and AI technologies mark the beginning of a larger structural revolution in the natural sciences. The competition among big AI companies to build increasingly large models with vast amounts of domain-specific knowledge has become akin to an arms race. The state-of-the-art in language processing and the impact of these technologies are currently defined by factors such as the number of model parameters, the volume of training data, and computational resources.

This explosion of AI capabilities, both in social and scientific applications, underscores the importance of models being able to capture subtle nuances of languages, whether contextualizing words or molecules with specific functional groups, understanding narrative logic in chemical reactions, or inferring unbiased words or molecules related to specific topics. In social applications, considerations such as gender and ethnicity play a crucial role, while in chemistry, prioritizing novel, greener, and more sustainable alternatives over obsolete chemical processes becomes imperative. However, one constant remains: languages have been instrumental in human evolution in the past, and they will continue to play a pivotal role in driving the advancement and acceleration of scientific research in the future.

Acknowledgements

This publication was created as part of NCCR Catalysis (grant number180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

Received: June 26, 2023

[1] P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, T. Laino, *Chem. Sci*. **2018**, *9*, 6091, https://doi.org/10.1039/C8SC02339E.

[2] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572, https://doi.org/10.1021/acscentsci.9b00576.

[3] P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, T. Laino, *Chem. Sci*. **2020**, *11*, 3316, https://doi.org/10.1039/C9SC05704H.

[4] A. C. Vaucher, F. Zipoli, J. Geluykens, V. H. Nair, P. Schwaller, T. Laino, *Nat. Commun*. **2020**, *11*, 3601, https://doi.org/10.1038/s41467-020-17266-6.

[5] A. C. Vaucher, P. Schwaller, J. Geluykens, V. H. Nair, T. Laino, *Nat. Commun*. **2021**, *12*, 2573, https://doi.org/10.1038/s41467-021-22951-1.

[6] P. Schwaller, A. C. Vaucher, T. Laino and J.-L. Reymond, *Mach. Learn.: Sci. Technol*. **2021**, *2*, 015016, https://doi.org/10.1088/2632-2153/abc81d.

[7] P. Schwaller,, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino, J.-L. Reymond, *Nat. Mach. Intell*. **2021**, *3*, 144, https://doi.org/10.1038/s42256-020-00284-w.

[8] D. Probst, M. Manica, Y. G. Nana Teukam, A. Castrogiovanni, F. Paratore, T. Laino, *Nat. Commun*. **2022**, *13*, 964, https://doi.org/10.1038/s41467-022-28536-w.

[9] RXN for Chemistry: https://rxn.res.ibm.com

[10] Y. G. Nana Teukam, L. Kwate Dassi, M. Manica, D. Probst, P. Schwaller, T. Laino, *ChemRxiv*. Cambridge: Cambridge Open Engage, **2023**, https://doi.org/10.26434/chemrxiv-2021-m20gg-v3.

[11] A. Toniato, P. Schwaller, A. Cardinale, J. Geluykens, T. Laino, *Nat. Mach. Intell.* **2021**, *3*, 485, https://doi.org/10.1038/s42256-021-00319-w.

[12] A. Thakkar, A. C. Vaucher, A. Byekwaso, P. Schwaller, A. Toniato, T. Laino, *ChemRxiv*. Cambridge: Cambridge Open Engage, **2022**, https://doi.org/10.26434/chemrxiv-2022-gx9gb.

[13] G. Pesciullesi, P. Schwaller, T. Laino, J.-L. Reymond, *Nat. Commun*. **2020**, *11*, 4874, https://doi.org/10.1038/s41467-020-18671-7.

[14] A. Toniato, A. C. Vaucher, P. Schwaller, T. Laino, *Digital Discov*. **2023**, *2*, 489, https://doi.org/10.1039/D2DD00110A.

[15] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, T. Laino, *Sci. Adv*. **2021**, *7*, 15, abe4166, https://doi.org/10.1126/sciadv.abe4166.

License and Terms

This is an Open Access article under the terms of the Creative Commons Attribution License CC BY 4.0. The material may not be used for commercial purposes.

The license is subject to the CHIMIA terms and conditions: (https://chimia.ch/chimia/about).

The definitive version of this article is the electronic one that can be found at https://doi.org/10.2533/chimia.2023.484