Simple User-Friendly Reaction Format

David F. Nippa[1,2,†], Alex T. Müller[1†], Kenneth Atz[1,3,†], David B. Konrad[2,*],
Uwe Grether[1,*], Rainer E. Martin[1,*] & Gisbert Schneider[3,*]

[1]Roche Pharma Research and Early Development (pRED), Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Grenzacherstrasse 124, 4070 Basel, Switzerland.
[2]Department of Pharmacy, Ludwig-Maximilians-Universität München, Butenandtstrasse 5, 81377 Munich, Germany.
[3]ETH Zurich, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.
† These authors contributed equally to this work.
∗ To whom correspondence should be addressed.
E-mail: david.konrad@cup.lmu.de, uwe.grether@roche.com, rainer_e.martin@roche.com, gisbert@ethz.ch

Abstract

Utilizing the growing wealth of chemical reaction data can boost synthesis planning and increase success rates. Yet, the effectiveness of machine learning tools for retrosynthesis planning and forward reaction prediction relies on accessible, well-curated data presented in a structured format. Although some public and licensed reaction databases exist, they often lack essential information about reaction conditions. To address this issue and promote the principles of findable, accessible, interoperable, and reusable (FAIR) data reporting and sharing, we introduce the Simple User-Friendly Reaction Format (SURF). SURF standardizes the documentation of reaction data through a structured tabular format, requiring only a basic understanding of spreadsheets. This format enables chemists to record the synthesis of molecules in a format that is understandable by both humans and machines, which facilitates seamless sharing and integration directly into machine learning pipelines. SURF files are designed to be interoperable, easily imported into relational databases, and convertible into other formats. This complements existing initiatives like the Open Reaction Database (ORD) and Unified Data Model (UDM). At Roche, SURF plays a crucial role in democratizing FAIR reaction data sharing and expediting the chemical synthesis process.

Introduction

The synthesis of chemical matter is often viewed as a rate-limiting step in material sciences, crop protection and drug discovery [1–4]. Crafting complex molecules typically involves multi-step syntheses, encompassing various reaction steps, each presenting multi-parameter optimization challenges [5, 6]. This high complexity makes chemical reactions time- and resource-intensive [7, 8]. Exploiting the growing volume of chemical reaction data could enhance synthesis planning and potentially boost success rates [9–11]. In recent years, machine learning has shown applications to a broad variety of challenges in chemistry [12–19]. Graph neural networks, transformers, and recurrent neural networks have proven effective in reaction prediction and synthesis planning tasks [20–27].

However, these tools can only excel when trained on high-quality data formatted in a structured, machine-readable manner [28]. Usually, laboratory experiment records are documented in varied ways by scientists, leading to complexities in retrieving and applying essential underlying metadata [29, 30]. With the advent of semi-automated reaction screening capable of running hundreds of reactions in parallel, [31, 32] the detailed and digital capturing of chemical reactions and procedures is becoming paramount. Consequently, there is a pressing need to close the gap between the laboratory and the data science worlds (Figure 1).

The challenges in documentation practices also extend to publications, where comprehensive reaction data, including parameters, reagents, quantities, and roles, should ideally be disclosed. However, this information is often buried within the supplementary materials of publications, presented as unstructured text or, occasionally, substrate scope tables. These tables may also appear in the main manuscript of methodology publications but frequently include footnotes highlighting exceptions, further complicating systematic analysis. Furthermore, both types of documents are typically available in the challenging-to-process Portable Document Format (PDF). Consequently, the barrier to accessing complete reaction data sets in a time- and cost-efficient manner remains high [33]. Additionally, data sourced from scientific literature and patents frequently omit details about unsuccessful reaction outcomes. Yet, these negative results are vital for training machine learning models, as they crucially contribute to generating reliable predictions [34–37].
These challenges are evident in the state of currently accessible public and commercial databases that encompass chemical reactions. Public resources in this domain are notably limited, with examples including the dataset covering chemical reactions from US patents spanning from 1976 to 2016 [38]. There are commercial offerings like Reaxys [39] and SciFinder [40], but these, too, face constraints in providing comprehensive and well-structured reaction data. Although these databases do contain a considerable number of reactions from scientific literature and patents, they frequently fall short in terms of providing essential information regarding reaction conditions and outcomes. Moreover, there can be a noticeable bias towards including high-yielding reactions, potentially neglecting the valu-

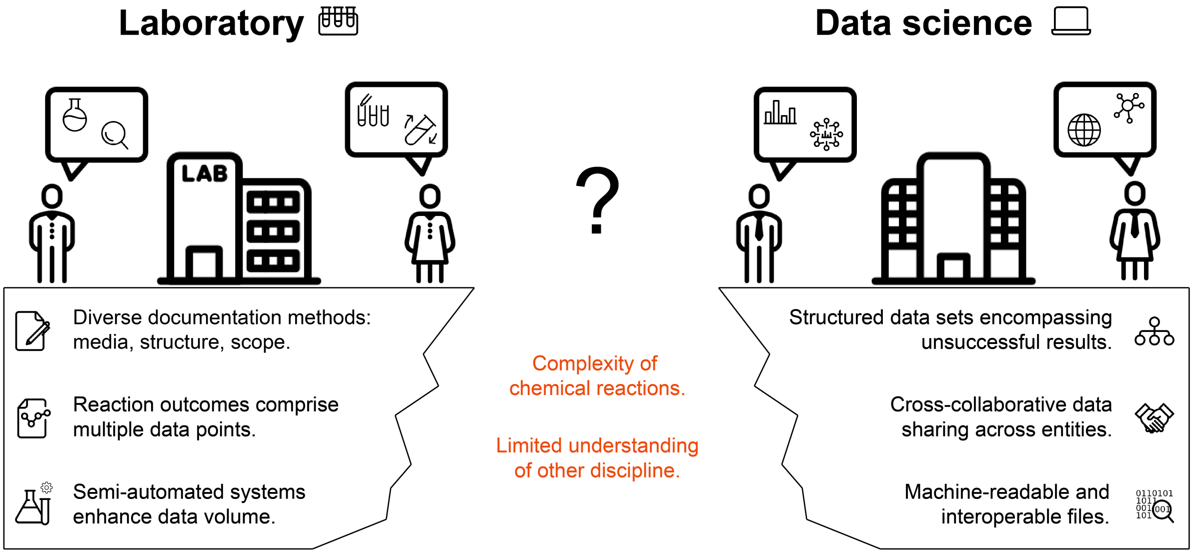

Laboratory
LAB
Data science
?
Diverse documentation methods: media, structure, scope.
Reaction outcomes comprise multiple data points.
Semi-automated systems enhance data volume.
Complexity of chemical reactions.
Limited understanding of other discipline.
Structured data sets encompassing unsuccessful results.
Cross-collaborative data sharing across entities.
Machine-readable and interoperable files.

Figure 1: **The missing connection between the laboratory and the data science worlds.** Information covering chemical reactions is highly complex as it compromises multiple data points leading to a wide range of documentation methods and, often, to unstructured and not machine-readable data (left). As a consequence, it remains difficult to make use of the experimental information for data science applications. This discipline requires structured data including unsuccessful reaction outcomes (right). Importantly, the data must be machine-readable and interoperable with other file types to allow seamless sharing, analysis and utilization of the reaction data. Bridging the gap between the two disciplines therefore remains a challenging task.

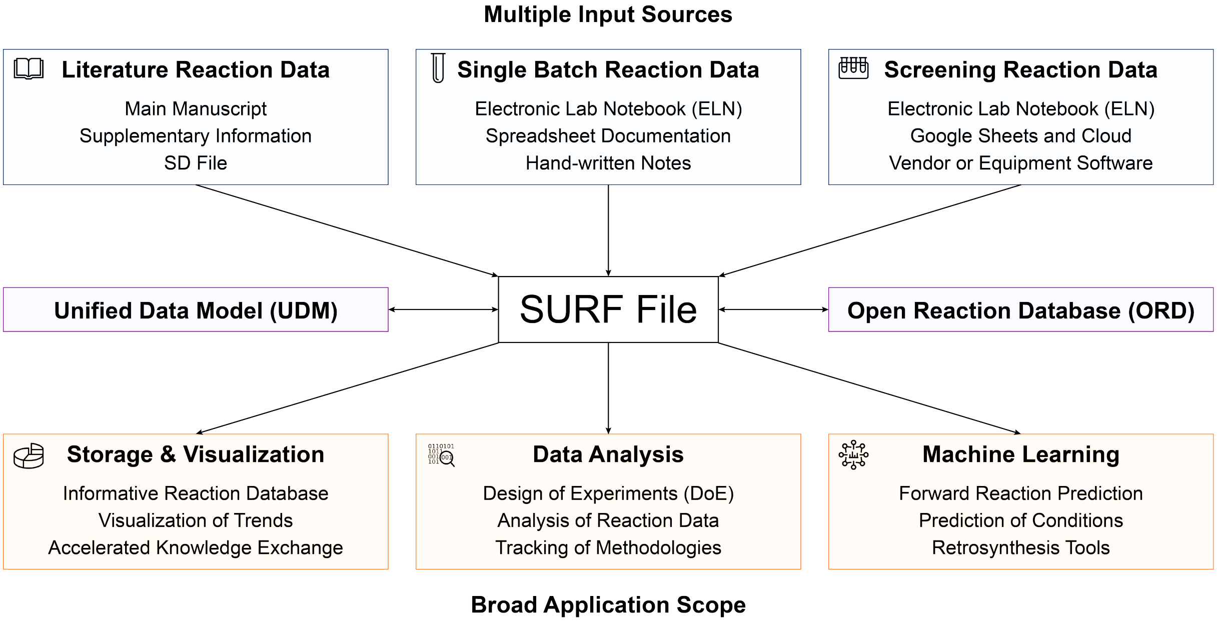
able insights that can arise from reactions with lower yields or unsuccessful outcomes [41, 42]. A multitude of different file formats, in which this data is stored, further complicates access to and harmonization of reaction data. Among the most common formats are Reaction Data File (RDFile), ChemDraw Extensible Markup Language (CDXML), Reaction International Chemical Identifier (RInChI), Reaction File (RXNFile), JavaScript Object Notation (JSON), and Chemical Markup Language Reaction(CMLReact) [43–46]. While these formats can effectively store molecular structures and corresponding chemical reaction diagrams, they tend to lack a controlled vocabulary and detailed reaction conditions, such as equivalents. Their usability is often compromised by the specialized technical knowledge required to work with them, which can hinder accessibility and understanding. Hence, there exists a notable gap in achieving findable, accessible, interoperable, and reusable (FAIR) standards for the reporting, collection, and storage of reaction data. Addressing this gap is imperative to facilitate and advance data-driven research in the field of chemistry. [47].

Recently, two initiatives have been introduced with the aim of capturing reaction data in machine-readable and uniform formats.

1. The Unified Data Model (UDM), initially developed by Roche and Reaxys and now managed by the Pistoia Alliance, is an open, extendable, and freely available data format for exchanging experimental information on compound synthesis and testing [46]. UDM employs a controlled vocabulary, an explicit hierarchical data model, and supports various molecule and reaction representations. UDM, implemented through an Extensible Markup Language (XML) schema, provides the advantage of utilizing widely accessible, generic tools for parsing, validation, and transformation. The format also captures analytical data, literature references, and legal information, with extension points allowing the inclusion of vendor- or process-specific data.

2. The Open Reaction Database (ORD) was introduced as an open-access platform for making chemical reaction data available in a structured format [48]. The ORD schema, implemented using Protocol Buffers [49], offers nine sections to comprehensively cover all experimental details, including the integration of raw and processed analytical data, ensuring reproducibility. ORD's high flexibility accommodates varying levels of detail based on available information. Moreover, the authors of the ORD emphasize usability by enabling data submission via software programs and through a web interface. Leveraging these features, ORD data is compatible with machine learning applications and even provides descriptive fields for reaction featurization.

While UDM and ORD represent important steps towards improving the standardization of reaction data for information sharing and machine learning applications, they pose certain challenges in day-to-day laboratory and data science environments: (i) Complexity: The availability of numerous fields and options



Multiple Input Sources
Literature Reaction Data
Main Manuscript
Supplementary Information
SD File
Single Batch Reaction Data
Electronic Lab Notebook (ELN)
Spreadsheet Documentation
Hand-written Notes
Screening Reaction Data
Electronic Lab Notebook (ELN)
Google Sheets and Cloud
Vendor or Equipment Software
Unified Data Model (UDM)
SURF File
Open Reaction Database (ORD)
Storage & Visualization
Informative Reaction Database
Visualization of Trends
Accelerated Knowledge Exchange
Data Analysis
Design of Experiments (DoE)
Analysis of Reaction Data
Tracking of Methodologies
Machine Learning
Forward Reaction Prediction
Prediction of Conditions
Retrosynthesis Tools
Broad Application Scope

Figure 2: **The simple user-friendly reaction format (SURF) can handle various input sources from the laboratory to aid a broad range of data science applications.** SURF files serve as a connector between multiple input sources from literature and the laboratory environment (blue, top) with a broad range of output applications in data science and machine learning (orange, bottom). The format is interoperable with the Unified Data Model (UDM) and the Open Reaction Database (ORD) (purple, middle). The data flow is demonstrated through arrows, highlighting the central role of SURF in connecting data from the laboratory with data science utilization.

for data entry may lead to fewer entries and missing data, as laboratory scientists have limited time for documentation. Focus and simplification, within the constraints of chemistry, are essential for capturing as many data points as possible, including unsuccessful reactions. (ii) IT barrier: Although ORD offers the option of entering and searching reaction data through a web interface in addition to programmed input, this still necessitates multiple manual steps in an external environment. UDM provides programmed input only, requiring IT skills or dedicated specialists, which precludes most chemists from using UDM for their reaction data. (iii) Data sharing between disciplines: Efficient exchange of reaction data within and across research groups, departments, or companies can accelerate research. With UDM and ORD, direct sharing of data between scientists in the same discipline, such as chemist to chemist, or across disciplines, such as chemist to machine learning scientist, may be hindered depending on available IT skills and infrastructure, as these formats are not easily human-readable for untrained individuals. Finally, the nested data structure complicates streaming reactions from these file formats.

Adopting accessible data practices in chemistry is paramount for advancing machine learning applications in the field [42]. We have developed the "Simple User-Friendly Reaction Format" (SURF) at Roche. SURF addresses certain limitations of UDM and ORD, complementing these existing data formats while maintaining interoperability. It structures reaction data reporting through a straightforward, yet comprehensive tabular format, requiring only a basic understanding of spreadsheets. SURF eliminates the need for coding experience, advanced IT skills, or a web interface, empowering every chemist to document and share their chemical syntheses in a human- and machine-readable format. As a result, the SURF format has the potential to further democratize reaction data. We advocate making the attachment of a SURF file to the supplementary information of manuscripts mandatory, thereby improving reaction data reporting and ultimately allowing a broad scientific community simplified access to valuable data.

Simple-user friendly reaction format

The development of SURF emerged from the need for efficient sharing of reaction data among laboratory chemists, data scientists, and machine learning researchers. Given the involvement of such a diverse group of stakeholders with different backgrounds in computer science and chemistry, creating a structured model interpretable by both humans and machines was of paramount importance for improving the drug discovery process. Based on these considerations, we opted to use simple spreadsheets, as they facilitate data capture in a tabular format, are widely used, and require minimal training. Using spreadsheets addresses the existing information technology barrier of other formats and democratizes FAIR reaction data documentation and sharing. Figure 2 illustrates the current

Table 1: **Overview of the reaction data documentation and storage landscape.** The major options for the documentation of reaction data are assessed based on a range of criteria relevant to bridging the gap between the laboratory and data science worlds. Three ✓denotes best, one ✓indicates worst. Examples for databases are Reaxys or Chemical Abstracts Service (CAS). ELN: Electronic Lab Notebook, ORD: Open Reaction Database, RD / RXN: reaction data file formats, UDM: Unified Data Model.

	Handwritten	**ELN**	**Databases**	**ORD**	**UDM**	**RD / RXN**	**SURF**
Human editable	✓✓✓	✓✓✓	✓	✓	✓✓	✓	✓✓✓
Machine readable	✓	✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓
Vendor reliance	✓✓✓	✓	✓	✓✓✓	✓✓✓	✓✓	✓✓✓
IT requisite	✓✓✓	✓✓	✓✓	✓✓	✓	✓	✓✓✓
Structured data	✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓
Data richness	✓✓	✓✓	✓✓	✓✓✓	✓✓	✓	✓✓✓

role of SURF at Roche, serving as a connector between the laboratory and data world, enabling FAIR reaction data capture, storage, sharing, and application.

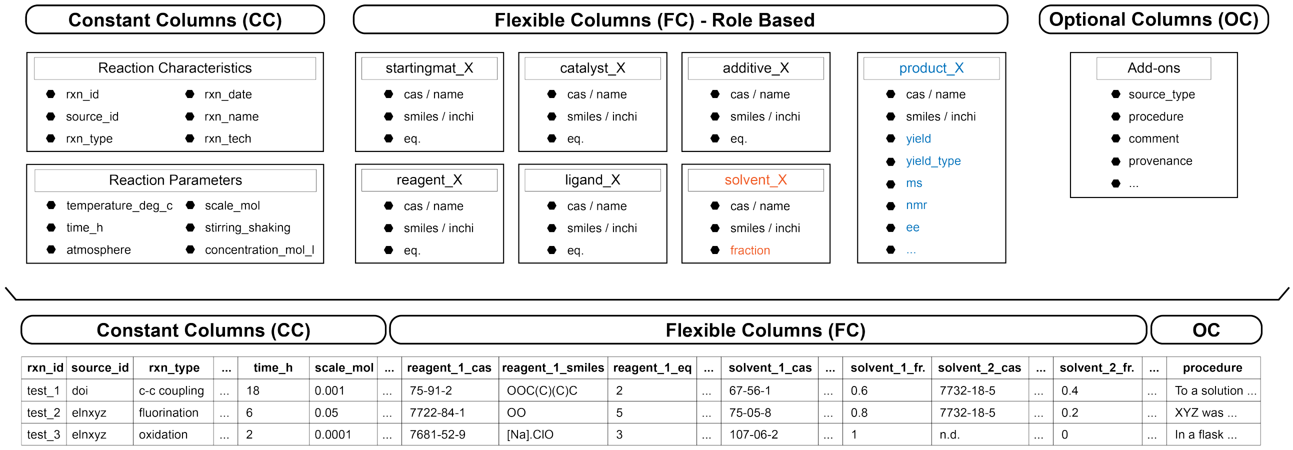
Through SURF, laboratory scientists can independently report their reaction data, eliminating the need for expensive software or specialized training. Other means of single-batch reaction documentation, such as data from electronic laboratory notebooks or spreadsheets, can also be imported or transformed. Furthermore, various types of literature data can be curated into SURF. At Roche, we are funnelling all high-throughput experimentation reaction data from multiple sources into SURF.

SURF enables direct data loading into machine learning models, as structural molecular features are captured through public compound identifiers, *i.e.*, Chemical Abstracts Service (CAS) numbers, simplified molecular input line entry system (SMILES) or international chemical identifier (InChI) strings. This feature enables forward reaction prediction, supports the determination of useful reaction conditions and training of retrosynthesis prediction tools. Due to its structure, reaction databases and corresponding visualization can be easily built and harnessed. Moreover, SURF files enable scientists in the laboratory to efficiently track their reaction data, directly work with their data by conducting analyses, and make data-driven decisions for designing new experiments based on previous outcomes.

Due to these characteristics, SURF has advantages when compared to other means of currently available reaction data documentation and storage possibilities (Table 1). Specifically, the hybrid structure of SURF, *i.e.*, usable by humans and machines, bridges the gap between the laboratory and data science worlds. Previously available reaction formats have focused on being either human-editable or machine-readable, often focusing on the latter. SURF, on the other hand, empowers experimentalists to document reaction data in a flexible yet comprehensive manner, without necessitating any IT expertise. It concurrently generates machine-readable files suitable for data analysis or machine learning applications.

Structure of SURF

In a SURF spreadsheet, each row stores data for one reaction. The column headers structure the data and are divided into constant (CC) and flexible (FC) columns. CCs remain unchanged and should always be present, independent of the number of reaction components. They capture the identifiers and provenance of the reaction, as well as basic characteristics (reaction type, named reaction, reaction technology) and conditions (temperature, time, atmosphere, scale, concentration, stirring/shaking). Add-ons, such as the procedure or comments, also belong to the CCs. The FCs describe the more variable part of a reaction, including different starting materials, solvents, reagents, and products. Each reaction component is represented by an identifier, such as the CAS number or molecule name, a SMILES or an InChI string storing the chemical structure. While the SMILES/InChI string is available for every compound and can serve as structural input for machine learning models, the CAS number can be useful for chemists in the laboratory to order, itemize, and find chemicals. To account for starting materials and reagents, including catalysts, ligands, and additives, a third column is incorporated to specify the stoichiometric amount, that is, equivalents. SURF's flexibility enables the capture of multiple starting materials and reagents, as these can be accommodated by adding three additional columns (CAS/name, SMILES/InChI, and equivalents). If desired, further columns for additional identifiers or lot numbers can be added. As shown in Figure 3, the headers are populated by adding ascending numbers to record all used components. The same applies to multiple solvents or products; however, due to their role, they possess more and partly different column headers. While the CAS number/name and/or the SMILES/InChI string remain as identifiers, the solvent fraction (recorded in decimals from [0,1]) is used instead of equivalents, allowing for the exact determination of the ratio between solvents. The product category contains the largest number of headers, as the basic SURF records the reaction yield (in percent, %), complemented by the reaction yield type (*i.e.*, isolated, LCMS, GCMS, etc.), as well as the detected mass by mass spectrometry and the nuclear magnetic resonance (NMR) spectroscopy sequence(s) in addition to the common identifiers CAS and SMILES/InChI. Additional information, such as detailed product char-



rxn_id	source_id	rxn_type	...	time_h	scale_mol	...	reagent_1_cas	reagent_1_smiles	reagent_1_eq	...	solvent_1_cas	...	solvent_1_fr.	solvent_2_cas	...	solvent_2_fr.	...	procedure
test_1	doi	c-c coupling	...	18	0.001	...	75-91-2	OOC(C)(C)C	2	...	67-56-1	...	0.6	7732-18-5	...	0.4	...	To a solution ...
test_2	elnxyz	fluorination	...	6	0.05	...	7722-84-1	OO	5	...	75-05-8	...	0.8	7732-18-5	...	0.2	...	XYZ was ...
test_3	elnxyz	oxidation	...	2	0.0001	...	7681-52-9	[Na].ClO	3	...	107-06-2	...	1	n.d.	...	0	...	In a flask ...

Figure 3: **The structure of the simple user-friendly reaction format (SURF)**. Top: Detailed structure of a SURF file, which contains constant (CC), flexible (FC), and optional columns (OC) to comprehensively capture reaction data information. Reaction components are described with two identifiers, one of them containing structural information, *e.g.*, SMILES or InChI, and the used equivalents. For solvents, an exception applies, instead of the equivalents, the fraction is recorded (orange). In the product section, depending on the granularity required, multiple columns for product characterization can be added (blue). In the basic SURF structure, yield, yield type, nuclear magnetic resonance and mass spectroscopy information are added. Bottom: Condensed example of a SURF file that demonstrates the simple structure of the format.

acterization (*e.g.*, enantiomeric excess (ee) or purity), can be captured by introducing respective columns with headers following the standard snake case nomenclature.

Utilizing the basic structure of SURF, all relevant data for reproducing the experiment is readily available. Laboratory chemists can order chemicals, draw structures, calculate the masses of molecules, or compare NMR data without the need to consult separate files. Since most electronic laboratory journals already record the aforementioned parameters of the basic SURF structure, enforcing documentation compliance combined with automated data extraction and cleaning pipelines has the potential to make numerous new reaction data accessible in the SURF format and available for machine learning applications.

File formats and interoperability

As SURF captures data in a tabular format, we recommend using universally readable file formats such as TXT, CSV, or TSV files. Since chemical data can contain delimiters such as commas or semicolons, we suggest using only TAB-delimited TXT or TSV files. These file types can be written and read with all popular spreadsheet or text editor software available on multiple operating systems. One point to consider when using SURF is that data is not validated upon capture. We acknowledge that this does not prevent users from entering false or incomplete reaction data. However, we recommend performing validation only upon reading SURF files into a database, transforming them to other formats, or using the data for machine learning purposes. SURF files are interoperable, as they can be introduced into hierarchical databases and converted into other existing reaction formats, such as the ORD Protocol Buffers format or UDM XML format. As part of this manuscript, we open-source the respective Python code enabling the transformation between different data formats (http://reaction-surf.com).

Applications

When preparing for a new series of reactions, such as in a high-throughput setting, chemists have the capability to populate a SURF file with all the necessary conditions and reagents to be tested in advance. They can link these entries to the specific vessels, tubes, or plates used for the reactions through the reaction identifier. Furthermore, having the CAS numbers available for all compounds greatly aids in locating the corresponding materials in the laboratory. Subsequently, as the reactions are executed and data on their outcomes are recorded, any potential gaps or missing data become immediately visible and accessible within the SURF format.

A frequently observed barrier to machine learning application is data pre-processing and cleaning. With SURF, reaction data is presented in a structured, both human and machine-readable format. Hence, SURF has shown to be a key enabler for several reaction prediction case studies at Roche [28, 50]. The use of SURF necessitated minimal data cleaning, mainly focusing on structural information validation and the exclusion of non-relevant columns. This approach allowed for the rapid extraction and analysis of reaction data using standard data science libraries. The SURF header convention as shown in Figure 3 ensures reproducibility and allows for easy identification of relevant columns needed for model training.

The tabular SURF format allows users to browse and filter available reaction data directly in a spreadsheet. Straightforward analyses to visualize yields or find all reactions of a certain type, using a specific technology, substrate, or reagent, can be conducted without the need to load the data into a database. Correlating individual columns like reaction characteristics with reaction outcomes becomes a straightforward task in SURF. Lastly, capturing reactions in a universally readable spreadsheet format facilitates data sharing. Using the snake case naming convention for headers generates tables that are both human and machine-readable. Additionally, by utilizing CAS numbers as identifiers, compounds can be universally recognized even without loading the SMILES/InChI.

Discussion and Conclusion

SURF offers a streamlined and accessible solution for chemists to document and share their chemical syntheses in a format that is both human- and machine-readable. By adopting SURF, researchers can overcome the limitations of existing data formats, promote successful data-driven chemistry research, and foster a culture of open data sharing and collaboration, thereby accelerating the pace of discovery and innovation in the field. The availability of reliable data and accompanying code provided by SURF enables other researchers to rapidly verify research findings, thereby reducing the risk of publishing irreproducible results. Importantly, the adoption of SURF facilitates efficient exchange of reaction data within and across research groups, departments, and companies, which can accelerate research progress.

Funding agencies and journals have an opportunity to play a more prominent role in promoting open access and FAIR publication of reaction data, ensuring that the necessary incentives and support are in place for researchers to embrace these principles. By encouraging the adoption of SURF as a standard for publications and requiring its attachment to the supplementary information of manuscripts, the scientific community can facilitate reaction data sharing and ultimately advance chemistry research.

Acknowledgements

D.B.K. acknowledges funding from the Fonds der Chemischen Industrie (FCI) through a Liebig Fellowship and Roche Basel for funding the PhD position of D.F.N. K.A. and G.S. acknowledge support by the Swiss National Science Foundation (SNSF, grant no. 205321_182176). We thank Nadja Flückiger, Yannick Stenzhorn and Remo Hohler for their valuable contributions to the SURF curated data sets.

Competing interest

G.S. declares a potential financial conflict of interest as co-founder of inSili.com LLC, Zurich, and in his role as a scientific consultant to the pharmaceutical industry. D.F.N., A.T.M., K.A., U.G. and R.E.M. are full employees of F. Hoffmann-La Roche Ltd. The authors have not disclosed any additional potential conflicts of interest.

Data and Code Availability

Three SURF files containing reaction data from literature covering Minisci-type alkylations [50], C-H borylation [28] and post-borylation modification reactions as well as program code for seamless interoperability with other data formats are available at `http://reaction-surf.com`.

References

1. Blakemore, D. C. *et al.* Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **10,** 383–394 (2018).
2. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17,** 97–113 (2018).
3. Tabor, D. P. *et al.* Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mat.* **3,** 5–20 (2018).
4. Schneider, P. *et al.* Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19,** 353–364 (2020).
5. Corey, E. J. The logic of chemical synthesis: multistep synthesis of complex carbogenic molecules (nobel lecture). *Angew. Chem. Int. Ed.* **30,** 455–465 (1991).
6. Coley, C. W. *et al.* A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365,** 1566 (2019).
7. Regalado, E. *et al.* Organic chemistry. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347,** 49–53 (2014).
8. Ley, S. V., Fitzpatrick, D. E., Ingham, R. J. & Myers, R. M. Organic synthesis: march of the machines. *Angew. Chem. Int. Ed.* **54,** 3449–3464 (2015).
9. Schneider, G. Mind and machine in drug design. *Nat. Mach. Intell.* **1,** 128–130 (2019).
10. Lowe, D. M. *Extraction of chemical structures and reactions from the literature* PhD thesis (University of Cambridge, 2012).
11. Schneider, G. & Clark, D. E. Automated de novo drug design: are we nearly there yet? *Angew. Chem. Int. Ed.* **58,** 10792–10803 (2019).
12. Von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4,** 347–358 (2020).
13. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nat. Mach. Intell.* **3,** 1023–1032 (2021).

14. Unke, O. T. *et al.* Machine learning force fields. *Chem. Rev.* **121,** 10142–10186 (2021).

15. Atz, K., Isert, C., Böcker, M. N., Jiménez-Luna, J. & Schneider, G. Δ-Quantum machine-learning for medicinal chemistry. *Physical Chemistry Chemical Physics* **24,** 10775–10783 (2022).

16. Isert, C., Atz, K. & Schneider, G. Structure-based drug design with geometric deep learning. *Curr. Op. Struct. Biol.* **79,** 102548 (2023).

17. Huang, B., von Rudorff, G. F. & von Lilienfeld, O. A. The central role of density functional theory in the AI age. *Science* **381,** 170–175 (2023).

18. Isert, C., Atz, K., Riniker, S. & Schneider, G. Exploring protein-ligand binding affinity prediction with electron density-based geometric deep learning. *ChemRxiv preprint 10.26434/chemrxiv-2023-585vf* (2023).

19. Atz, K. *et al.* Prospective de novo drug design with deep interactome learning. *Nat. Commuun.* **15,** 3408 (2024).

20. Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network. *J. Neural Inf. Process.* **30** (2017).

21. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51,** 1281–1289 (2018).

22. Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555,** 604–610 (2018).

23. Schwaller, P. *et al.* Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5,** 1572–1583 (2019).

24. Shen, Y. *et al.* Automation and computer-assisted planning for chemical synthesis. *Nat. Rev. Methods Primers* **1,** 1–23 (2021).

25. Somnath, V. R., Bunne, C., Coley, C., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. *J. Neural Inf. Process.* **34,** 9405–9415 (2021).

26. Guan, Y. *et al.* Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **12,** 2198–2208 (2021).

27. Thakkar, A., Chadimová, V., Bjerrum, E. J., Engkvist, O. & Reymond, J.-L. Retrosynthetic accessibility score (RAscore)–rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **12,** 3339–3349 (2021).

28. Nippa, D. F. *et al.* Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning. *Nat. Chem.* **16,** 239–248 (2024).

29. Rudolphi, F. & Goossen, L. J. Electronic laboratory notebook: The academic point of view. *J. Chem. Inf. Model.* **52,** 293–301 (2012).

30. Kanza, S. *et al.* Electronic lab notebooks: Can they replace paper? *J. Cheminf.* **9,** 1–15 (2017).

31. Shevlin, M. Practical high-throughput experimentation for chemists. *ACS Med. Chem. Lett.* **8,** 601–607 (2017).

32. Cook, A., Clément, R. & Newman, S. G. Reaction screening in multiwell plates: High-throughput optimization of a Buchwald–Hartwig amination. *Nat. Protoc.* **16,** 1152–1169 (2021).

33. Guo, J. *et al.* Automated chemical reaction extraction from scientific literature. *J. Chem. Inf. Model* **62,** 2035–2045 (2021).

34. Engkvist, O. *et al.* Computational prediction of chemical reactions: current status and outlook. *Drug Discov. Today* **23,** 1203–1218 (2018).

35. Strieth-Kalthoff, F. *et al.* Machine learning for chemical reactivity: The importance of failed experiments. *Angew. Chem. Int. Ed.* **61,** e202204647 (2022).

36. King-Smith, E. *et al.* Predictive Minisci and P450 Late Stage Functionalization with Transfer Learning. *ChemRxiv preprint https://doi.org/10.26434/chemrxiv-2022-7ddw5-v2* (2023).

37. Caldeweyher, E. *et al.* Hybrid Machine Learning Approach to Predict the Site Selectivity of Iridium-Catalyzed Arene Borylation. *J. Am. Chem. Soc.* (2023).

38. Lowe, D. Chemical reactions from US patents (1976-Sep2016). *Figshare https://doi.org/10.6084/m9. figshare* **5104873** (2017).

39. Limited, E. *Reaxys* `https://reaxys.com/` (2023).

40. Society, A. C. *Reaxys* `https://scifinder.cas.org/` (2023).

41. Fitzner, M., Wuitschik, G., Koller, R., Adam, J.-M. & Schindler, T. Machine Learning C–N Couplings: Obstacles for a General-Purpose Reaction Yield Prediction. *ACS Omega* **8,** 3017–3025 (2023).

42. Mercado, R., Kearnes, S. M. & Coley, C. W. Data Sharing in Chemistry: Lessons Learned and a Case for Mandating Structured Reaction Data. *J. Chem. Inf. Model* (2023).

43. Dalby, A. *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput.* **32,** 244–255 (1992).

44. Grethe, G., Blanke, G., Kraut, H. & Goodman, J. M. International chemical identifier for reactions (RInChI). *J. Cheminformatics* **10,** 1–9 (2018).

45. Holliday, G. L., Murray-Rust, P. & Rzepa, H. S. Chemical markup, XML, and the world wide web. 6. CMLReact, an XML vocabulary for chemical reactions. *J. Chem. Inf. Model* **46,** 145–157 (2006).

46. Tomczak, J. *et al.* UDM (Unified Data Model) for chemical reactions - past, present and future. *Pure Appl. Chem.* (2022).

47. Jablonka, K. M., Patiny, L. & Smit, B. Making the collective knowledge of chemistry open and machine actionable. *Nat. Chem.* **14,** 365–376 (2022).

48. Kearnes, S. M. *et al.* The open reaction database. *J. Am. Chem. Soc.* **143,** 18820–18826 (2021).

49. LLC, G. *Protocol Buffers* `https://protobuf.dev/` (2023).

50. Nippa, D. F. *et al.* Identifying opportunities for late-stage CH alkylation with high-throughput experimentation and in silico reaction screening. *Commun. Chem.* **6,** 256 (2023).