JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Perspective

The Open Reaction Database

Steven M. Kearnes,* Michael R. Maser, Michael Wleklinski, Anton Kast, Abigail G. Doyle, Spencer D. Dreher, Joel M. Hawkins, Klavs F. Jensen, and Connor W. Coley*

Cite This: https://doi.org/10.1021/jacs.1c09820
Read Online
ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Chemical reaction data in journal articles, patents, and even electronic laboratory notebooks are currently stored in various formats, often unstructured, which presents a significant barrier to downstream applications, including the training of machine-learning models. We present the Open Reaction Database (ORD), an open-access schema and infrastructure for structuring and sharing organic reaction data, including a centralized data repository. The ORD schema supports conventional and emerging technologies, from benchtop reactions to automated high-throughput experiments and flow chemistry. The data, schema, supporting code, and web-based user interfaces are all publicly available on GitHub. Our vision is that a consistent data representation and infrastructure to support data sharing will enable downstream applications that will greatly improve the state of the art with respect to computer-aided synthesis planning, reaction prediction, and other predictive chemistry tasks.

INTRODUCTION

The opportunity to learn complex patterns of chemical reactivity from organic reaction data is increasingly clear. Data-driven machine-learning models have been designed and applied to planning synthetic pathways, recommending reaction conditions for each putative transformation, and even predicting what the major products of a yet-untested reaction might be, e.g., for impurity prediction. These and other tasks in "predictive synthesis" promise to streamline chemical development at stages where the synthetic route development occurs and eventually enable automated multi-step synthesis.

Chemical reaction data support these modeling efforts by capturing the details of how an experiment was performed and its outcome. There are few curated databases that record organic reactions and even fewer that are generally available to researchers, even commercially. Only one prominent data set—a set of reactions extracted from the USPTO—is open access.[1]

Beyond their role in supporting data-driven models, tabulated reaction data have become an indispensable tool in the workflow of nearly every chemist, if only for information retrieval. Few practicing chemists approach the synthesis of a new compound without performing a database or literature search (e.g., through SciFinder or Reaxys) and referring to procedural details in a precedent article or patent describing a relevant synthesis. While these databases tabulate and make searchable many important aspects of a chemical reaction such as the structures of reactants, agents, and products, full procedural details are left as unstructured text in the original document.

The Open Reaction Database[29] is an initiative to support machine learning and related efforts in reaction prediction, chemical synthesis planning, and experimental design (Figure 1). Our goals are to provide a structured data format for chemical reaction data, to make data freely and publicly available, to encourage sharing of precompetitive proprietary data, and to provide an interface for browsing/downloading that data. One key use case is the standardization and sharing of high-throughput reaction screening data. Underlying this effort are the FAIR principles for scientific data management.[2]

We have defined a reaction schema that provides a thorough coverage of experimental details that we know are important for reproducibility. Importantly, we capture the most essential information in a structured format, rather than an unstructured text format as is currently used in publications (e.g., as Supporting Information). Analytical data, both raw and processed, can be directly coupled to experimental outcomes. Reaction data can be recorded programmatically (e.g., using Python) or using an interactive web editor that is more approachable for those without coding experience. A templating mechanism for data set enumeration allows one reaction entry to be rapidly expanded into hundreds or thousands, as in the case of a high-throughput screen.

SCHEMA

The ORD schema contains structured and unstructured (free text) fields for documenting chemical reactions. It is intended to be descriptive, rather than prescriptive; each `Reaction` record in the database should describe what was actually done in the lab, and not an idealized protocol or instruction set for, e.g., automated synthesis hardware (although this may also be

Received: September 15, 2021

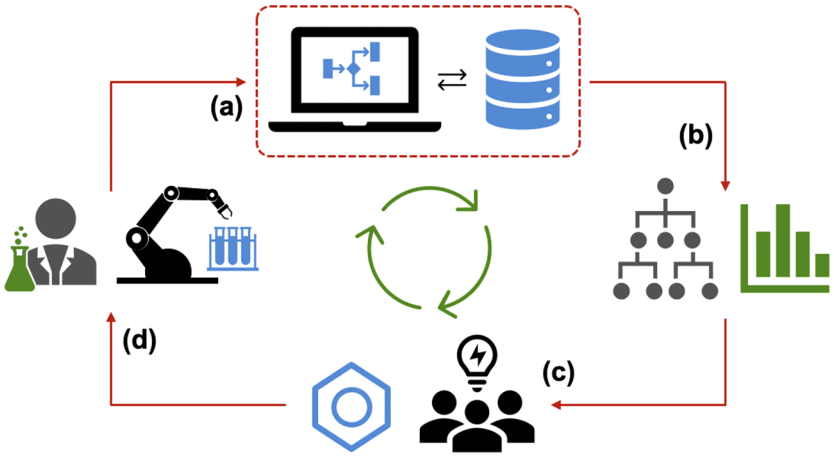

Figure 1. Computer-aided chemical discovery cycle: (a) the Open Reaction Database; (b) machine learning and cheminformatics; (c) human or automated interpretation and material design; (d) manual or robotic chemical synthesis.

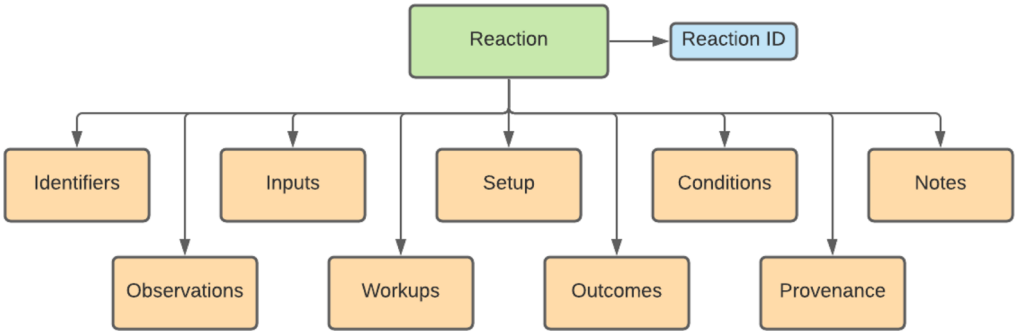

Figure 2. Overview of the Open Reaction Database schema. Each `Reaction` record contains sections for reaction identifiers, inputs, outcomes, etc.

included). The schema is implemented using Protocol Buffers,[30] which have distinct usability advantages over other common technologies such as XML.

At a high level, the schema is divided into nine sections: reaction identifiers, inputs, setup, conditions, notes, observations, workups, outcomes (products and analytics), and provenance (Figure 2). Conceptually, each section can be thought of as a first-class object in the schema. Each of these sections contains data fields and child objects (with their own data fields) for describing the reaction. For example, each `ReactionInput` contains one or more components (usually `Compound` objects), each with their own identifiers(s), amount, and reaction role (Figure 3a). The fields of each schema object are structured to constrain their types or values, such as only allowing positive numeric values for amounts or limiting the units to a set of predefined constants. Many schema objects also include an unstructured `details` field for providing additional information that is not captured by other fields. The set of structured fields is largely driven by (a) a relatively small set of common types, such as for reaction roles, and (b) downstream use cases for machine learning, where structured types can be used as categorical features.

The schema is designed to support arbitrary levels of detail depending on the available information. For reactions taken from the patent literature, it may only be possible to describe the inputs and outputs at the level of their identifiers and amounts. For reactions submitted by the original experimenter, it is possible to use structured and unstructured fields to include every detail required for reproducibility (including and beyond those in Figure 3b). To enforce a base level of consistency between records in the database, we use validation functions written in Python to require the presence of certain fields and check for reasonable values. For instance, each reaction must have at least one input, and every input compound requires an amount. Warnings are issued for ambiguous values, such as percentages entered as fractions. These validations are performed automatically in the interactive web editor (see below) and during the data set submission process.

INTERFACES

Although submissions can be generated programmatically with Python, we are sensitive to the reality that experimentalists are not always comfortable with programming or using the command line. Accordingly, we have built web interfaces for creating submissions and searching the database. We also include instructions for web-only submissions in the online documentation.

The ORD Interactive Editor[31] is a tool for generating submissions to the database. Users can create data sets and use a responsive web form to fill in structured and unstructured data fields. Additionally, the editor can be used to enumerate a factorial data set based on a reaction template and a CSV or Excel spreadsheet with a row for each reaction. We have also recorded tutorial videos demonstrating the use of the editor, available on the ORD YouTube channel. Submissions to the ORD are made as pull requests to the `ord-data` GitHub repository[32] or directly by email.[33]

https://doi.org/10.1021/jacs.1c09820
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

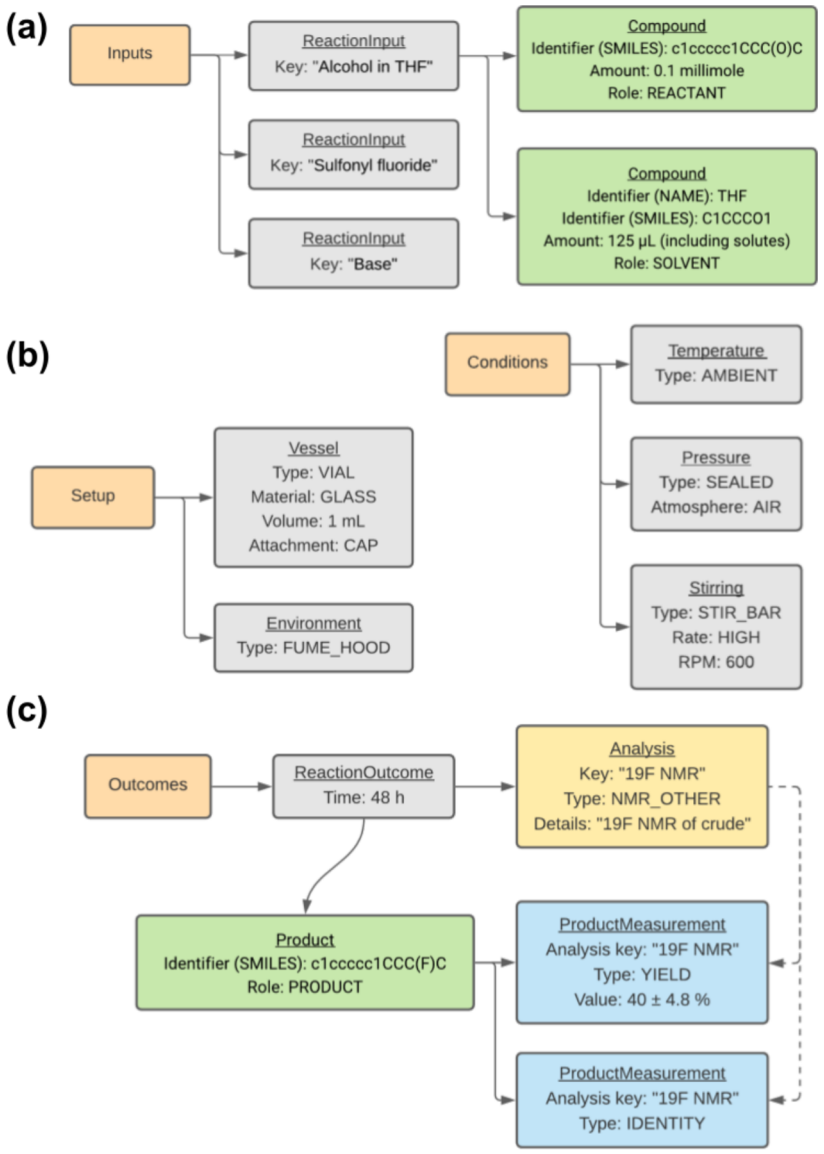

Figure 3. Example schema representations for a deoxyfluorination reaction.[3] (a) Inputs. The "Alcohol in THF" input is expanded to show its solute and solvent components. (b) Setup and conditions. The setup includes a description of the reaction vessel and there are separate objects for temperature, pressure, and stirring conditions. (c) Outcomes. Each outcome has an associated reaction time and one or more analyses that are linked to product-specific measurements of product identity, yield, selectivity, etc. There are several dozen additional fields not shown in this figure.

The ORD Search Interface[34] is used to search for reactions in the database. Reactions can be queried by SMILES/SMARTS substructure patterns for input or output molecules, by reaction ID, by publication DOI, etc. Each reaction has a summary page that displays the full record along with a graphical schematic. We have also written a simple Python API to enable downstream users to easily access data programmatically. The client library supports all of the search functionality in the web-based search interface, as well as the ability to download full data sets or lists of reactions.

■ DATA SET EXAMPLES

The ORD has been designed to accommodate information about organic reactions spanning many distinct types, each requiring the definition of different metadata to ensure reproducibility. Here, we briefly mention several common categories we anticipate and discuss the data that might be associated with each; examples in the ORD are given in Table 1.

The simplest but perhaps broadest category of reactions to capture are those performed through "traditional" benchtop operations. A data set of this type typically contains a small number of reactions defined by their inputs (Figure 3a), conditions/setup (Figure 3b), and product outcomes at a single time point (Figure 3c). Reaction inputs should always contain information about the quantity of each component and any additional metadata that might be relevant, such as the source of the compound. Conditions should at least include temperature and stirring, and the setup should specify what kind of vessel (e.g., flask, vial) was used. The level of detail associated with each reaction is a function of what information is available; for retrospective data entries, such as importing the open-source reaction data set parsed from the USPTO literature, many procedural details will be unspecified.

https://doi.org/10.1021/jacs.1c09820
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

Table 1. Example Data Sets Currently Available in the ORD

category	description	ref	size	example
single-step batch	deoxyfluorination reaction screening as a function of substrate, base, and fluoride source (entries in Figure 1)	3	80	link
single-step batch	microwave synthesis of a small library using the Biginelli multicomponent condensation reaction	4	48	link
kinetic profiling	online monitoring of a Suzuki coupling reaction by HPLC	5	7	link
high throughput	subset of "chemistry informer" screen of copper-catalyzed Buchwald−Hartwig aminations (entries 11−15 in Figure 4)	6	90	link
high throughput	C−N cross-coupling reaction yields varying aryl halide, additive, Pd catalyst, and base identities	7	4312	link
high throughput	Suzuki coupling reaction performance as a function of aryl halide, boronic acid, ligand, base, and solvent performed under pseudoflow conditions	8	5760	link
high-throughput	C−N cross-coupling reaction performance of 3-bromopyridine with various nucleophiles, varying precatalysts and bases (entries in Experiment 2)	9	1536	link
high throughput	combinatorial nanochemistry screen of a complex aryl halide library using dual-metal photoredox C−N coupling (entries in Figure 6)	10	1728	link
photochemistry	substrate scope tables regarding coupling of α-carboxyl sp^3 carbons with aryl halides	11	24	link
photochemistry	Ir-catalyzed debromination conversions as a function of photocatalyst ligands	12	1152	link
electrochemistry	electroreductive coupling of alkenyl and benzyl halides via nickel catalysis (entries in Figures 2 and 3)	13	27	link
flow chemistry	sulfonamide library synthesis in flow	14	39	link
enzymatic	multistep biocatalytic cascade for the manufacture of islatravir	15	3	link
multistep	copper-catalyzed enantioselective hydroamination of alkenes	16	3	link
literature extracted	reactions extracted by text-mining United States published patents; imported from CML documents	1	1771017	link

In most cases, analytical data will at least include the yield of the desired product and how it was calculated (isolated yield, NMR yield, enantiomeric excess by chiral SFC, etc.). Ideally, more detailed analytical data will be included, such as the usual slate of analyses when a new compound is reported (i.e., ^{1}H NMR, ^{13}C NMR, HRMS, IR). Reactions can accommodate arbitrary analytical data as uploaded files or URLs to externally hosted data. Kinetic profiling experiments where multiple samples or aliquots are taken from a single vessel over time are described by one reaction with several distinct outcomes, each with their own analytical and product information; if different reactors are used for each time point, these are described as distinct reactions.

Data sets corresponding to high-throughput experiments will typically follow the case where reactions are identical other than a small number of variable fields such as substrate identity, catalyst identity, reagent identity, temperature, and product yield. Depending on the analytical workflow, the recorded yields from HTE may be uncalibrated: i.e., correspond only to relative UV peak areas in an LC or integrated TIC/EIC traces from LCMS. Specifying how yields were calculated or estimated is essential for understanding the fidelity of the data and how it should be treated in downstream tasks. It is easiest to define data sets of this type through Python code or through the data set enumeration feature.

Additional types of reactions make use of the many optional fields in the schema. Photochemical reactions require additional specification of the illumination conditions, such as the light source and its distance to the vessel. Electrochemical reactions require details about the cell configuration: whether it is divided or undivided and the materials of the cathode and anode. Reactions employing flow chemistry include information about the type of pump and delivery rate used for all stock solutions, in addition to details about the type of flow reactor used. Reactions employing enzymatic catalysts take advantage of the ability to define compounds by their primary amino acid sequence or UnitProt/PDB identifiers.

Multistep reactions in the ORD are split into separate entries within one data set; their treatment depends on whether intermediate products are isolated. Reaction sequences with incomplete intermediate isolations (e.g., one-pot reactions or telescoped reaction sequences) are linked by considering the product of one step as a "crude" input of the subsequent one. Otherwise, an isolated product from an in-house reaction is treated no differently from a purchased starting material, except for a cross-reference to the reaction ID defining how the compound was prepared.

DOWNSTREAM USE CASES

We anticipate that one of the major applications of the ORD will be for the generation of structured data sets for machine learning (ML). Reaction modeling with ML has become increasingly common, with reports frequently citing the importance of high-quality data for success. Examples of the types of data-driven reactivity studies that could benefit from the ORD include synthesis planning,[17,18] reaction product prediction,[19] reaction yield prediction,[3,7,20] reaction condition prediction,[21,22] selective catalyst design,[23,24] and reaction optimization,[25] among others. These predictive chemistry tools illustrate the value of curating reaction data, yet we believe they are just scratching the surface in terms of both their utility and the complexity of tasks that they can help address.

The schema was designed in part around these use cases and provides descriptive, easily accessible fields for reaction featurization (see above). As a result, data sets require minimal processing before model training and can be quickly integrated into Python ML workflows. We have made an example data processing and ML pipeline available as a Jupyter Notebook in the ORD schema repository.[35] This example constructs a yield predictor from a Suzuki−Miyaura coupling data set[8] and reproduces regression results from the literature.[20]

It is our hope that this type of ML workflow will aid in advancing the understanding of reaction performance (e.g., yield, selectivity, etc.) and thus accelerate discovery. We expect that another valuable use case will be the construction of retrosynthesis models. These tools may enable more quantitative evaluations of new reactions given these additional details (e.g., concentrations, orders of addition, vendor information) and eventually allow training directly on raw

https://doi.org/10.1021/jacs.1c09820
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

analytical data instead of processed analytical data. Data captured by the ORD schema is richer than that in previous database efforts and should facilitate training of the next generation of predictive chemistry tools.

DISCUSSION

Commitment to Open Access. As the name implies, the Open Reaction Database is designed for open access and community contributions. All reaction data in the database is available under a CC BY-SA license.[36] The various software tools and code, such as the schema definitions and the interactive web editor, are available under an Apache license.[37] Both the data and code are hosted on GitHub under the Open Reaction Database organization (https://github.com/open-reaction-database).

Data Quality and Validation. There are several mechanisms for ensuring, or at least encouraging, that data are of high quality. At a basic level, Python scripts are used to validate each reaction within a data set, as mentioned above. This requires that all reactions contain a minimal level of information and, for example, that quantities are associated with units. Beyond these checks, the schema itself was designed to capture important metadata such as *how* yields were quantified, so that subsequent analyses can distinguish between true isolated yields and estimates from LCMS peak areas (where the latter is more common in high-throughput workflows).

Additionally, all submissions are reviewed manually for completeness and correctness by a reviewer not involved in the preparation of the submission. During the review process, the reviewer will require the details listed above unless the submitter does not have access to more information about the experiment (e.g., more than what was published); when a previously published data set is reviewed, the publication and submission should be compared directly. If a data set is not associated with a journal or patent article, then the reviewer may only check for internal consistency and completeness and must trust the submitter to provide the correct structures and technical details. As the rate of ORD submissions grows, we anticipate recruiting additional volunteer reviewers and adjusting the review process as needed.

Information about the submitter and experimenter (who might not be the same person) is captured as part of the provenance metadata for each entry, optionally including email addresses. This will allow data consumers to contact either party if more information is needed for their downstream use case. If data quality issues or discrepancies are identified, we will request that the original submitter help revise or review any proposed changes.

Evolution of the Schema. The ORD schema is a "living document" that will change and adapt to the needs of its users and of the community. For example, future contributions from electrochemists or advances in downstream machine learning may require the addition of new structured fields to better represent the data. In other cases, structured fields that are hardly ever used may be deprecated and removed in favor of unstructured descriptions. We will use versioning and migration processes to ensure that backward-incompatible changes (if any) are deployed without disruption to existing workflows pinned to earlier versions of the data and code.

OUTLOOK

By providing a structured schema, submission mechanism, and search/retrieval tools, the ORD reduces several technical barriers to data sharing and model building. For instance, adoption of the ORD schema will allow the Supporting Information in publications to be standardized and readable by anyone without manual parsing or back and forth with the authors. Importantly, the use of a well-defined schema will make the omission of procedural details needed for reproducibility more apparent, as well as enable easier comparisons between reactions (both within and across publications). We plan to support translation software to convert between the ORD format and other emerging data standards[26−28] as well as electronic lab notebook formats. To be clear: we believe that PDFs without accompanying structured data should no longer clear the bar for publication.

However, we are still sensitive to the many social and cultural barriers surrounding publishing that persist in the field. There is not a culture of data sharing beyond what is required for peer-reviewed publication. In particular, the unwillingness to share "unsuccessful" or "failed" reactions with low yields or selectivities—whether due to intrinsic reactivity, procedural details, or human errors—presents an overly optimistic view of chemistry where most reactions succeed, impeding the development of models to explain when they may not. Successful and unsuccessful reactions are not differentiated by the ORD schema beyond product measurements; this categorization is somewhat arbitrary and depends on the downstream use case.

Though the ORD is still young, initial feedback and adoption have been promising. The success of this effort will depend on buy-in from data generators, contributors, and consumers and a shared recognition of its value. We encourage data generators to explore the schema and infrastructure that we have built for capturing their experimental data and invite discussion from the broader community on how to incorporate these structured data formats throughout the life cycle of reaction data—from benchtop to laptop.

AUTHOR INFORMATION

Corresponding Authors

Steven M. Kearnes − *Relay Therapeutics, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0003-4579-4388; Email: skearnes@relaytx.com

Connor W. Coley − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-8271-8723; Email: ccoley@mit.edu

Authors

Michael R. Maser − *Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States;* orcid.org/0000-0001-7895-7804

Michael Wleklinski − *Chemistry Capabilities Accelerating Therapeutics, Merck & Co., Inc., Kenilworth, New Jersey 07033, United States;* orcid.org/0000-0001-6735-4696

Anton Kast − *Google LLC, Mountain View, California 94043, United States*

https://doi.org/10.1021/jacs.1c09820
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

Abigail G. Doyle − *Department of Chemistry & Biochemistry, University of California at Los Angeles, Los Angeles, California 90095, United States;* orcid.org/0000-0002-6641-0833

Spencer D. Dreher − *Chemistry Capabilities Accelerating Therapeutics, Merck & Co., Inc., Kenilworth, New Jersey 07033, United States;* orcid.org/0000-0002-5094-1218

Joel M. Hawkins − *Chemical Research and Development, Pfizer Worldwide Research and Development, Groton, Connecticut 06340, United States*

Klavs F. Jensen − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-7192-580X

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.1c09820

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

There are many people to thank for help in getting this initiative started: all of the respondents to our original use case survey in 2019, the full advisory board, and all of the direct contributors to `ord-data`. We also thank Nathan Kim for helping with infrastructure development, Brian Lee for helping with feedback and beta testing, and Devin Sandberg and Zan Armstrong for design advice. Cloud computing resources and legal support were provided by Google. S.M.K. acknowledges support from Google and Relay Therapeutics. Additional GitHub data storage and bandwidth were provided through the GitHub Education program.

■ REFERENCES

(1) Lowe, D. *Chemical reactions from US patents* (1976−September 2016). https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873. DOI: 10.6084/m9.figshare.5104873.v1. (accessed 5/19/2021).

(2) Wilkinson, M. D. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.

(3) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 5004−5008.

(4) Stadler, A.; Kappe, C. O. Automated Library Generation Using Sequential Microwave-Assisted Chemistry. Application toward the Biginelli Multicomponent Condensation. *J. Comb. Chem.* **2001**, *3*, 624−630.

(5) Christensen, M.; Adedeji, F.; Grosser, S.; Zawatzky, K.; Ji, Y.; Liu, J.; Jurica, J. A.; Naber, J. R.; Hein, J. E. Development of an automated kinetic profiling system with online HPLC for reaction optimization. *Reaction Chemistry & Engineering* **2019**, *4*, 1555−1558.

(6) Kutchukian, P. S.; Dropinski, J. F.; Dykstra, K. D.; Li, B.; DiRocco, D. A.; Streckfuss, E. C.; Campeau, L.-C.; Cernak, T.; Vachal, P.; Davies, I. W.; Krska, S. W.; Dreher, S. D. Chemistry informer libraries: a chemoinformatics enabled approach to evaluate and advance synthetic methods. *Chemical Science* **2016**, *7*, 2604−2613.

(7) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C−N cross-coupling using machine learning. *Science* **2018**, *360*, 186−190.

(8) Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **2018**, *359*, 429−434.

(9) Buitrago Santanilla, A. Nanomole-scale High-throughput Chemistry for the Synthesis of Complex Molecules. *Science* **2015**, *347*, 49−53.

(10) Dreher, S. D.; Krska, S. W. Chemistry Informer Libraries: Conception, Early Experience, and Role in the Future of Cheminformatics. *Acc. Chem. Res.* **2021**, *54*, 1586−1596.

(11) Zuo, Z.; Ahneman, D. T.; Chu, L.; Terrett, J. A.; Doyle, A. G.; MacMillan, D. W. C. Merging photoredox with nickel catalysis: Coupling of α-carboxyl sp3-carbons with aryl halides. *Science* **2014**, *345*, 437−440.

(12) Mdluli, V.; Diluzio, S.; Lewis, J.; Kowalewski, J. F.; Connell, T. U.; Yaron, D.; Kowalewski, T.; Bernhard, S. High-throughput Synthesis and Screening of Iridium(III) Photocatalysts for the Fast and Chemoselective Dehalogenation of Aryl Bromides. *ACS Catal.* **2020**, *10*, 6977−6987.

(13) DeLano, T. J.; Reisman, S. E. Enantioselective Electroreductive Coupling of Alkenyl and Benzyl Halides via Nickel Catalysis. *ACS Catal.* **2019**, *9*, 6751−6754.

(14) Gioiello, A.; Rosatelli, E.; Teofrasti, M.; Filipponi, P.; Pellicciari, R. Building a Sulfonamide Library by Eco-Friendly Flow Synthesis. *ACS Comb. Sci.* **2013**, *15*, 235−239.

(15) Huffman, M. A. Design of an in vitro biocatalytic cascade for the manufacture of islatravir. *Science* **2019**, *366*, 1255−1259.

(16) Liu, R. Y.; Buchwald, S. L. Copper-Catalyzed Enantioselective Hydroamination of Alkenes. *Org. Synth.* **2018**, *95*, 80−96.

(17) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604−610.

(18) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.

(19) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434−443.

(20) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559*, 377−381.

(21) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4*, 1465−1476.

(22) Maser, M. R.; Cui, A. Y.; Ryou, S.; DeLano, T. J.; Yue, Y.; Reisman, S. E. Multilabel Classification Models for the Prediction of Cross-Coupling Reaction Conditions. *J. Chem. Inf. Model.* **2021**, *61*, 156−166.

(23) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363*, No. eaau5631.

(24) Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R. Enantiodivergent Pd-catalyzed C−C bond formation enabled through ligand parameterization. *Science* **2018**, *362*, 670−674.

(25) 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.

(26) Pistoia Alliance, D. Unified Data Model. https://www.pistoiaalliance.org/projects/current-projects/unified-data-model/, Accessed June 1, 2021.

(27) Mehr, S. H. M.; Craven, M.; Leonov, A. I.; Keenan, G.; Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **2020**, *370*, 101−108.

(28) Tremouilhac, P.; Lin, C.-L.; Huang, P.-C.; Huang, Y.-C.; Nguyen, A.; Jung, N.; Bach, F.; Ulrich, R.; Neumair, B.; Streit, A.; Bräse, S. The Repository Chemotion: Infrastructure for Sustainable Research in Chemistry. *Angew. Chem., Int. Ed.* **2020**, *59*, 22771−22778.

(29) https://open-reaction-database.org/.

(30) https://developers.google.com/protocol-buffers.

(31) https://editor.open-reaction-database.org/.

(32) https://github.com/open-reaction-database/ord-data.
(33) submissions@open-reaction-database.org.
(34) https://client.open-reaction-database.org/.
(35) https://github.com/open-reaction-database/ord-schema/blob/main/examples.
(36) https://creativecommons.org/licenses/by-sa/4.0/.
(37) https://www.apache.org/licenses/LICENSE-2.0.

https://doi.org/10.1021/jacs.1c09820
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX