ACS Synthetic Biology

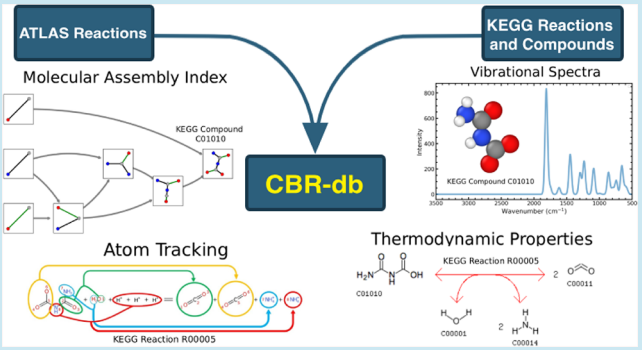
pubs.acs.org/synthbio
Technical Note

CBR-db: A Cheminformatic Database for Biochemical Reaction Analysis

Louie Slocombe, Camerian Millsaps, M. Reza Shahjahan, Kamesh Narasimhan, and Sara Imari Walker*

Cite This: ACS Synth. Biol. 2026, 15, 2684−2692
Read Online

ABSTRACT: We present CBR-db, a database with detailed chemical properties data for biochemical compounds and reactions. We provide the first chemically consistent analyses and detailed chemical property data for a biologically relevant set of 148,673 reactions and 18,716 small molecules derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the ATLAS of Biochemistry. CBR-db provides detailed atom tracking, reaction similarity classification, compound properties, chiral centers, molecular assembly indices, and thermodynamic properties. CBR-db is designed to be continuously updated, open-source, and reproducible. This report highlights the cheminformatic data, explains the curation and data set, and highlights the data's utility for researchers working at the intersection of chemical properties and biology. We anticipate a broad range of potential applications for these data, given their position at the intersection of cheminformatics and bioinformatics, including researchers interested in molecular properties and reaction mechanisms found in metabolism, the chemical underpinnings of synthetic biochemical design and engineering, and how the origins and evolution of biochemistry have selected among properties of molecules and reactions within chemical space.



KEYWORDS: *chemical space, atom tracking, reaction similarity, assembly index, cheminformatics, biochemical data*

INTRODUCTION

Informatics approaches have enabled advances in many areas of science, particularly when they allow for the integration of data across disciplines.[1−3] Cheminformatics has emerged in recent years as a powerful approach to the study of chemical space, defined as the property space of molecules and reactions. This has catalyzed advances in chemistry ranging from drug design to metabolic network reconstruction for synthetic biology, flux balance analysis for bioprocess optimization, retrosynthesis, atom tracking, reaction prediction in cheminformatics, and prebiotic chemistry modeling to explore the origins of metabolism.[4−11] Likewise, bioinformatics has fostered unprecedented progress in the study of biochemistry, including the exploration of compound and reaction diversity and the organization of these into pathways and networks. Yet, few bridges have been built to equip researchers to work at the intersection of cheminformatics and bioinformatics. Critical areas that could be advanced by such connections include some of the most important open questions in science, such as the emergence of biology from chemistry and what alternative chemistries might be possible for biology, which necessitate interdisciplinary integration of insights from researchers studying both chemistry and biology.

In chemical analyses, precise stoichiometry and balanced reactions are essential. Yet, for many biological applications, this level of chemical detail is unnecessary. For example, a bioinformatician might be interested in identifying the enzymes encoded in genes, and what reactions those enzymes catalyze, but not the exact molecular transformations underpinning those reactions, where the atoms go, the temperatures or pressures for the reaction to be spontaneous, or distinctive molecular properties (i.e., in elemental composition or molecular complexity). As such, prominent and highly used bioinformatic databases, like the Kyoto Encyclopedia of Genes and Genomes (KEGG),[1] have not focused on curating reaction chemistry, but instead on detailing only those properties of reactions relevant to understanding their roles in a broader biological context, e.g., in metabolic pathways. For researchers interested in the chemistry underpinning life, limitations include the absence of intermediate reactions, disconnected reaction pathways, the lack of metabolites linked to reaction pathways, and incomplete enzyme characterization.

Received: January 26, 2026
Revised: May 4, 2026
Accepted: May 5, 2026
Published: May 28, 2026

https://doi.org/10.1021/acssynbio.6c00060
ACS Synth. Biol. 2026, 15, 2684−2692

Table 1. Comparison of Key Features in CBR-db, KEGG, ATLAS, Enzyme Mapper, and Rhea

Feature	CBR-db	KEGG	ATLAS	Enzyme Mapper	Rhea
Data set size and scope					
Number of reactions	148,673	12,216	149,052	47,974	18,343
Number of compounds	18,716	19,438	19,438	5,100	15,125
Number of enzymes	-	-	-	12,749	7,600
Data quality improvements				-	-
Compounds with malformed, incomplete, or duplicated data	0	756	756	-	-
Reactions with malformed, incomplete, or duplicated data	0	1711	1249	-	-
Stoichiometrically unbalanced (after pruning)	0	806	11936	-	-
Generic halogen reactions resolved	Full	No	Partial	-	-
Standardized compound representations	Full	Partial	Partial	-	-
Integration of the latest KEGG release	Full	-	Partial	-	-

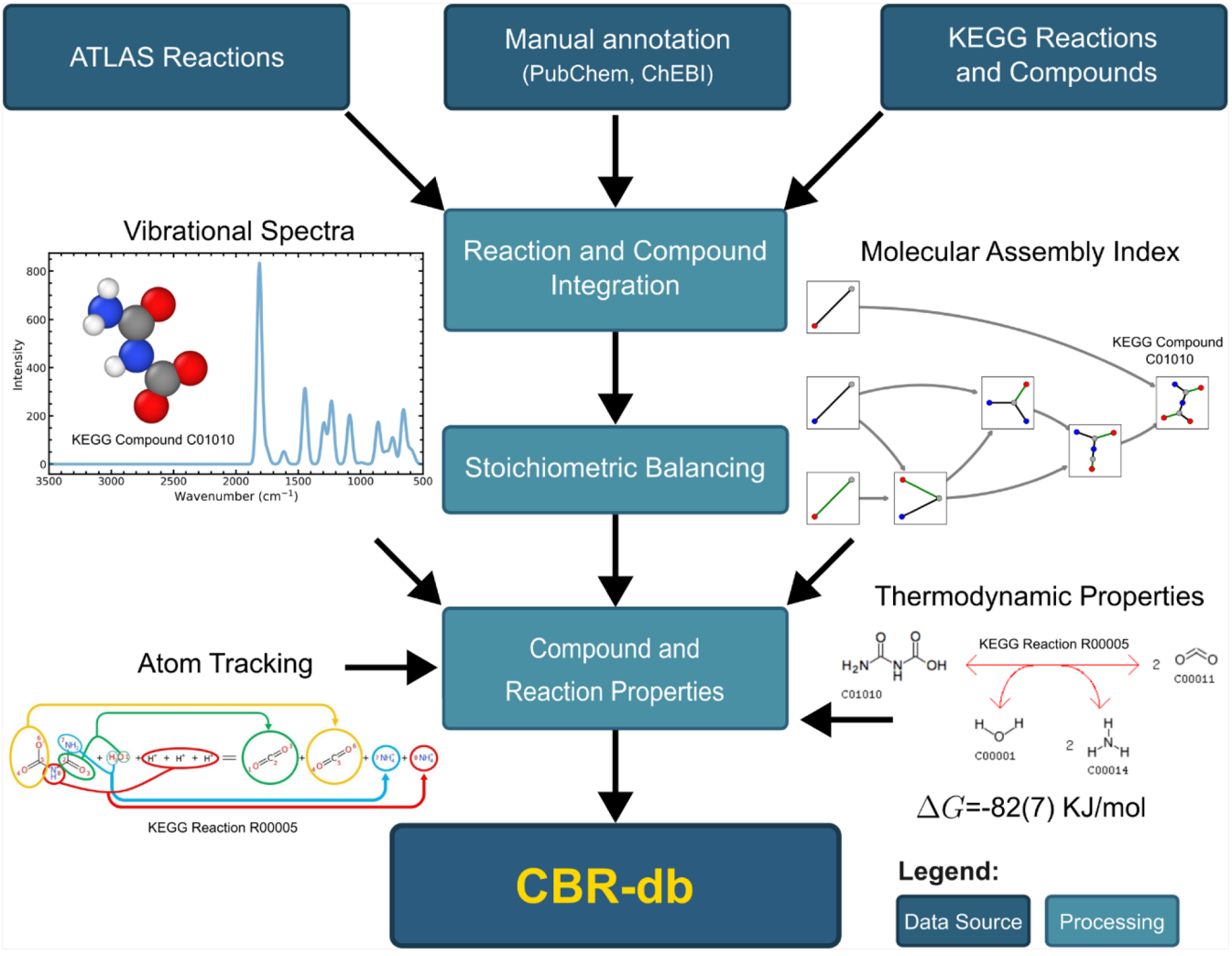


Figure 1. CBR-db includes several molecular properties and reaction properties. The database is constructed by merging the latest versions of KEGG and ATLAS, and it employs a series of quality control measures to fix stoichiometry, add missing mol files, and ensure a unique, high-quality biochemical reaction data set. From this data set, properties such as vibrational spectra, molecular assembly index, atom tracking of reactions, and thermodynamic properties are computed. This figure shows an example of atom tracking and the thermodynamic properties of KEGG reaction R00005. Additionally, the vibrational spectra and molecular assembly index pathway calculation for compound C01010 from the same reaction (R00005) are shown.

The full scope of biologically relevant reactions is not known. Established in 1995, the KEGG database was an important early effort to catalog what was known of biochemistry at the time. It has become an essential resource for research into biochemical reactions, pathways, and networks, with many updates over the years.[1] However, the full set of reactions used by living processes is still believed to be much larger than those identified and cataloged in KEGG. Confirming the biogenicity of a reaction and discovering new biological reaction functions are labor-intensive experimental

https://doi.org/10.1021/acssynbio.6c00060
ACS Synth. Biol. 2026, 15, 2684−2692

efforts. The ATLAS of Biochemistry[2,12] database (ATLAS) was developed to predict a potentially large set of possible biological reactions computationally, by applying reaction rules and prediction algorithms to theoretically build a more expansive reactome of enzyme-catalyzed reactions than what is cataloged in KEGG.[2] ATLAS systematically addresses limitations in KEGG's reaction coverage, including hypothesizing enzymatically catalyzed reactions that are not yet confirmed to occur in biology but are consistent with known enzyme activities. This provides a valuable resource for investigating new reaction pathways and predicting possible biochemical functionality that could be missing from current data.[12] The BNICE.ch rules used to generate ATLAS have been used to computationally design and experimentally validate pathways synthesizing structurally complex plant natural products and their derivatives.[8,9,13]

Biochemical databases such as KEGG and ATLAS provide extensive data that underpin a variety of computational and experimental workflows in synthetic biology, metabolic engineering, drug design, and evolutionary biology. But ATLAS inherited some limitations from KEGG regarding the accuracy of chemical data and introduced several new ones. These include reactions with incomplete stoichiometry (11,936 reactions), missing elements, and inconsistent chemical representations (1,249 reactions). Examples of reactions with missing elements include R0797 (KEGG and ATLAS entries), in which a chlorine atom appears on one side of the reaction but not the other, and A142112 (ATLAS), in which the left- and right-hand sides do not balance. A chemist will easily spot these errors, but an automated workflow can miss them if not specifically designed to detect them. Indeed, in many biological applications, these chemical details do not matter. However, many emerging areas at the interface of biology and chemistry, such as drug design, synthetic biochemistry, and prebiotic chemistry, require both accurate biological functional annotation and accurate chemical property annotations. KEGG and ATLAS emerged from research in bioinformatics and the need of researchers to organize data about enzymes and the reactions they catalyze. Cheminformatics has emerged in recent decades as an important field of research into molecules, reactions, and their properties, but the interface between bioinformatics and cheminformatics remains underdeveloped. More data at this intersection would enrich many fields of research that require knowledge of both biology and chemistry.

To this end, we introduce CBR-db, a Cheminformatic Biochemical Reaction database that provides rich cheminformatic data on molecules and reactions derived from the KEGG and ATLAS databases, see Table 1 for a comparison of CBR-db, KEGG and ATLAS data and the widely used biochemical databases Enzyme Mapper and Rhea. CBR-db contains no malformed compounds, includes balanced stoichiometric equations, provides atom-tracking for all reactions, free energy and vibrational spectra, molecular assembly index values, and other chemical properties. It thus represents a significant advance in the curation of chemical data associated with biochemical reactions in KEGG and ATLAS and includes cheminformatic data for 148,673 reactions and 18,716 compounds.

METHODS

We integrated the latest KEGG (v.116, dated October 1, 2025) release with ATLAS 2020 data.[12] KEGG reaction entries superseded their corresponding ATLAS reaction to ensure the inclusion of the most up-to-date and accurate biochemical information. Duplicated structural or equation line entries were merged, provided all quality control checks (as noted below) were passed. The curation efforts were structured around two interlinked components: compounds and reactions, see Figure 1. For compounds, KEGG provided structural information, which we used as a basis for compound entries. See the section "compound curation" below. For reactions, we used the reaction line from KEGG and ATLAS as our starting point; see the section "reaction curation". All metadata from both databases was retained.

Compound Curation

The first stage of the curation process focused on resolving inconsistencies in the KEGG compound data set, which is referenced by both KEGG and ATLAS reactions (Figure 1) and thus needed to be addressed first. Inconsistencies include malformed elemental entries and inconsistent representations of functional groups, such as R-group notation, which was standardized to reduce ambiguity. Compounds with generic placeholder atoms representing a specific elemental group, most notably halogens, were instantiated as specific compound entries reflecting individual members of that group when such compounds were missing. For most compounds in KEGG, KEGG provides structural information in mol file format, which must be converted. All provided structures were converted to consistent, reproducible SMILES and InChI formats to ensure compatibility with cheminformatics tools and computational workflows.[14,15] Compounds were standardized using the RDKit standardization and sanitize workflow functions to ensure molecular structures were chemically valid, consistent/canonical, and comparable prior to analysis or modeling. Thus, resolving improper stereochemical notations that impacted 3D embedding and chirality calculations.[16] Compounds lacking structural information were separated and flagged for manual curation. Where possible, we obtained structures from PubChem and ChEBI for a subset of these compounds and incorporated them into the database. Receipts of all changes and of items that were not fixed were logged and, if required, can be provided to the user in a separate file.

Compounds with malformed elemental entries (e.g., some KEGG-provided mol files that have an element entry labeled "OH", implying OH as an element instead of a group) were reprocessed to conform to chemical conventions and were appropriately protonated using RDKit.

Compounds lacking essential information, such as structural information or empirical formulas, were separated, logged in a separate file, and hand-curated. Structures were added for 60 compounds from the relevant PubChem entry.[17] This step ensured that only well-defined chemical compounds were retained for subsequent analyses.

Structures from KEGG may represent neutral or ionic forms. To ensure consistency, we standardized all molecules to the major microspecies at pH default 7.4 (standard physiological conditions) using a rules-based pipeline: (i) removal of inorganic counterions/mixtures and retention of the largest organic component; (ii) normalization of functional groups (e.g., nitro, sulfoxide, amide); (iii) reionization of acids/bases to their dominant charge state near neutral pH while preserving permanently charged moieties (e.g., quaternary ammonium); (iv) tautomer canonicalization; and (v) InChIKey generation for deduplication. The pipeline yields protonation-standardized parent structures for fingerprints, similarity, and enumeration.

We enumerate all protonated forms of each compound using dimorphite_dl.[18] With this, CBR-db adds or removes hydrogen atoms from molecular representations to achieve the appropriate protonation state for a user-specified pH range. The released compound entries were run over a pH range of 6.8−7.9 to capture the diverse range of states. Thus, all ionization states with a pH within this range are available in the compound database.

https://doi.org/10.1021/acssynbio.6c00060
ACS Synth. Biol. 2026, 15, 2684−2692

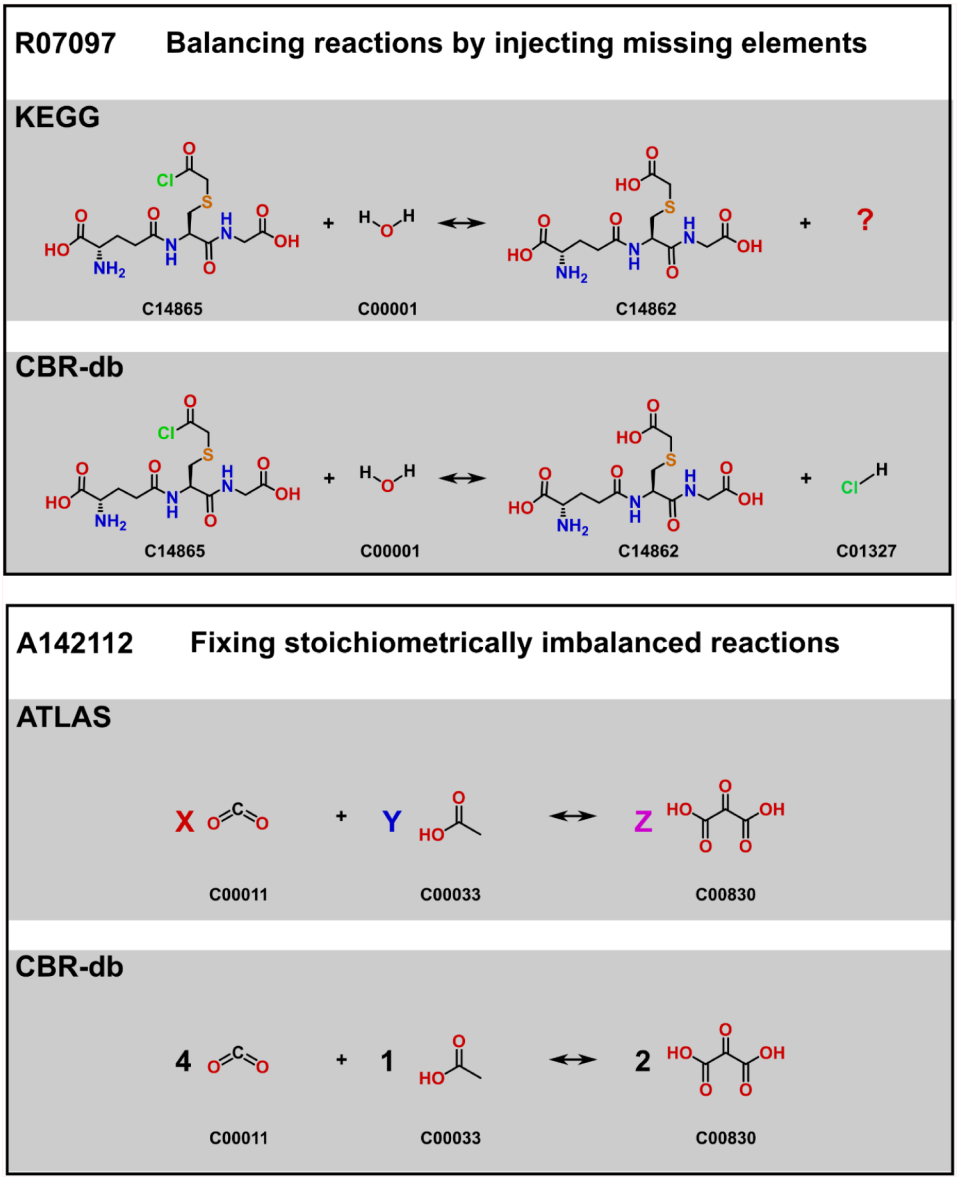


Figure 2. Examples showing accurate chemical reactions in CBR-db for certain error classes that effect both KEGG and ATLAS. The first box shows an example where there are missing compounds; the second box shows an example of an unbalanced reaction.

Compound Properties

Using the Atomic Simulation Environment (ASE) framework,[19] a popular, open-source Python library for building, manipulating, simulating, and analyzing atomic-scale systems, we calculate several properties of compounds, including free energies and vibrational spectra for each molecule. ASE provides a straightforward tool for converting between file inputs and managing calculations within Python. The underlying quantum chemistry package is ORCA,[20] which performs electronic structure calculations on molecules. We use ORCA version 6.1.0 because it efficiently and accurately implements analytic Hessians and hybrid functionals[20,21] and is well-suited for accurate calculations on small to medium-sized molecular systems.

We use the r2SCAN-3c exchange-correlation functional combined with the def2-mTZVPP basis set.[22] The r2SCAN-3c exchange-correlation functional is a suitable choice for vibrational calculations, as it balances predictive accuracy and computational efficiency. This makes it reliable, especially when dealing with complex molecular systems or large data sets. The r2SCAN-3c functional is a variant of the SCAN (strongly constrained and appropriately normed) functional, designed to enhance the description of exchange-correlation interactions. Compared with traditional generalized gradient approximation functionals, it yields more accurate results across a wide range of molecular systems, including organic and inorganic compounds. Additionally, r2SCAN-3c is based on a self-consistent treatment of the exchange-correlation functional, resulting in a more universally applicable and less parameter-dependent approach. The def2-mTZVPP basis set is a suitable choice for vibrational calculations due to its accuracy, computational efficiency, and comprehensive coverage of all elements in the data set, including heavy species. We use ORCA's preset tight-geometry optimization settings to ensure that the configuration is at a local minimum and contains no negative frequencies. We then extract the vibrational states and free energy.

In addition to *ab initio* calculations, we determine the experimental free energy of formation at physiological conditions using the eQuilibrator 3.0 package.[4]

Molecular assembly index is a recently developed and experimentally accessible measure of molecular complexity,[23] with applications to biosignature and origins of life research.[24] We calculate the molecular assembly index[23] of each processed and cleaned molecule by exporting each compound as a.mol file and using a new branch and bound rapid assembly index calculator.[25] This

https://doi.org/10.1021/acssynbio.6c00060
ACS Synth. Biol. 2026, 15, 2684−2692

provides the first comprehensive, publicly available data set of molecular assembly index values for biochemical data.

Reaction Curation

Our refined compound data set guided the selection of reactions, focusing on eliminating incomplete, ambiguous, and chemically invalid entries from both KEGG and ATLAS (see Figure 1).

In KEGG, "overall" or "multi-step" reactions summarize complex processes that encompass chains of other reactions but do not provide additional details. We removed all "overall" reactions and all multistep reactions where the individual steps were not well-defined. Furthermore, some reactions were identified as redundant following compound deduplication, stoichiometric rebalancing, or compound-injection rebalancing. Since these redundancies indicate duplications rather than comprehensive processes, such reactions were retained and deduplicated only during the final processing stage (i.e., database merger).

We rectified left and right-side imbalances to ensure chemically valid transformations, see Figure 2. Here, we define stoichiometric rebalancing as solving for the stoichiometric balance between the left- and right-hand side of the equation. We check for consistency across all elements, including hydrogen. We include hydrogen as the compounds have been consistently populated at pH 7.4. To do this efficiently, we used the ChemPy[26] package and balanced all equation entries in parallel.

For compound-injection rebalancing, we developed an algorithm that identifies differences between the two sides of the equation and checks whether the exact difference corresponds to a compound. An example is the KEGG reaction R07097, in which a Cl−H is missing from the right-hand side, and the algorithm identifies that C01327 can be added to the right-hand side. If the first pass fails, we use an iterative injection cycle. We begin with a list of compounds that match the elemental difference, and working down the list, add the compound and adjust the coefficients to balance the equation. We constrain the algorithm by starting with matches that have one non-hydrogen atom and increasing the number of heavy atoms; we terminate at depth two because the search space grows rapidly due to the large number of small molecules.

Reactions classified and labeled within KEGG as "general," "incomplete," or "unclear" were systematically isolated, flagged, and often affected subsequent processing stages. "General" reactions frequently referenced "General" compounds for which structural data were unavailable, increasing the likelihood of being flagged for correction. Additionally, some "General" compounds were found to be structural duplicates of specific compounds, resulting in many "General" reactions being identified as duplicates of specific ones. "Incomplete" or "unclear" reactions were also more likely to be removed due to imbalance, which often could not be resolved by our rebalancing methods, such as compound-ratio and compound-injection rebalancing.

Additionally, we fixed a subset of 150 KEGG reactions tagged "incomplete" or "unclear" by identifying critical reactants or products via compound-injection rebalancing. Reactions that previously involved generic halogen compounds were refined by creating specific and elementally self-consistent reaction entries. In total, we added 20 unique reactions, thereby significantly enhancing the data set's comprehensiveness in the space of halogen-associated reactions.

Finally, we calculate the reaction similarity using the differential reaction fingerprint DRFP technique, which uses a deep learning-based learned fingerprint for reaction classification.[27] We also determine the overall change in reaction energy and provide a reversibility index,[4] a metric that uses the reaction energy to determine how likely a reaction is to be thermodynamically skewed in a given direction.

In summary, all reactions were assessed for stoichiometric consistency. All reactions, whether fixed or not, are retained, and the user can filter them (using a simple flag) depending on the use case. It is at the user's discretion to determine whether they are critical to their work and to include them accordingly. This enhances the utility of a reaction space that might otherwise be unsuitable for accurate computational analyses.

Atom Mapping

Atom mapping, also known as atom-to-atom mapping, is a concept in cheminformatics and organic chemistry, often used interchangeably with atom tracking, with nuanced differences depending on the context. Atom mapping is a specific computational or formal process that assigns unique identifiers (usually numerical indices) to atoms in reactants and products to demonstrate their explicit correspondence in a chemical reaction. It establishes a one-to-one correspondence between atoms in the reactants and those in the products.[28] Conducting this task is labor-intensive and requires ongoing annotation of chemical reactions and the development of logically consistent directives when employing computational methods.[29] It is used to provide a standardized way to represent how atoms are conserved or rearranged in a reaction, often in SMILES or other chemical notation formats, and for informatics goals such as reaction prediction, synthesis planning, identifying reaction centers, determining bond cleavages and formations, extracting reaction templates, and database curation.[30] It also enables the automation of reaction analysis, for example, using tools such as RXNMapper that perform robust atom mapping without human intervention.[29] Here, we used the RXNMapper and Localmapper Python libraries, given their speed in handling large data sets and minimal human intervention, for reaction mapping. RXNMapper employs transformer architectures with SMILES representations and attention mechanisms to map atoms, achieving high performance and accuracy.[30] It converts these two-way connections into one-way links from product atoms to reactant atoms: starting with product atoms that exhibit the strongest links to matching-type reactant atoms, ignoring links to different types, and assigning matches iteratively until all product atoms are covered. A "neighbor boost" (set to 90) strengthens links between atoms that are close together, which helps when local regions are similar. RXNMapper converts the SMILES representations of compounds into smaller components, treating each atom and its bonds as individual "tokens." The model has a 512-token limit per reaction, which can lead to errors for longer and more complex reactions.[29] Our database contains approximately 450 reactions that exceed this token limit; therefore, we have designed a fallback procedure that uses the Localmapper library for those cases.

RESULTS AND DISCUSSION

The cheminformatics data generated from KEGG and ATLAS entries using CBR-db lead to significant improvements in quality and richness across multiple dimensions. It addresses limitations such as malformed compound representation, incomplete reactions, and imbalanced reaction stoichiometry. The results are grouped into two broad categories: data set size and scope, and data quality improvements. After removing duplicate or superseded entries, the integration yielded 18,716 unique compound entries and 148,673 unique reactions. In total, 2,871 (∼23.5%) of KEGG and 13,185 (∼8.8%) of ATLAS reaction entries required significant updates and chemical corrections or were filtered out for failing to meet chemical accuracy quality-control standards. To address gaps and ambiguities in the reaction space annotation, new inferred reactions were generated by resolving generic halogen species, thereby expanding the inferred biochemical reaction space relative to the ATLAS and KEGG databases. This effort ensures that the data set provides high-quality reaction data, facilitating its application to research questions in which precise chemical reaction models are crucial.

A critical focus of this work was ensuring the chemical validity of reactions. Approximately 806 (∼6.6%) KEGG and 11,936 (∼8%) ATLAS reactions were corrected for stoichiometric imbalances by adding missing elements or molecules

https://doi.org/10.1021/acssynbio.6c00060
ACS Synth. Biol. 2026, 15, 2684−2692

and rebalancing the reactant and product sides (Figure 2). Including the first pass and the injection-depth one, we fixed 3,533 reactions; in the final rounds, the algorithm fixed an additional 481 reactions. Example compounds that were commonly added to the reactions include (but are not limited to) water (C00001), methane (C01438), and calcium oxide (C13140). All compounds were sanitized and standardized using RDKit compounds to enhance compatibility with cheminformatics tools and thermodynamic modeling. Issues with malformed compounds, including improper stereochemical notations and inconsistent R-group (for example, mixed use of R, r, *) and N-group definitions (for example, mixed use of m, n, i, and j for repeated compounds), were resolved. The resulting data set standardizes compound annotations across all chiral compounds, thereby enhancing its utility for tasks that require accurate 3D embeddings or atomic chirality calculations. The compound database has entries that have chiral centers annotated and have SMILES/InChI entries with standardized repeated (n) and wildcard (*) groups.

KEGG reactions labeled as "overall" or multistep and annotated with their constituent parts were removed. Our choice to omit these reactions reduces redundancy. It ensures a more consistent representation of chemical data, with all data at the same level of granularity, suitable for chemical computational workflows and network representations.

We developed a script for CBRdb reactions that automates atom tracking in chemical reactions using advanced algorithms. It identifies specific atoms in substrate molecules and traces their fates through transformations to determine which product molecules contain them. The process starts with analyzing reaction equations, removing coefficients and cofactors, and selecting substrate molecules. Atoms are identified for tracking, focusing on those with the highest connectivity. RXNMapper is the primary tool for atom mapping, with LocalMapper as a backup when needed. As the reaction proceeds, atom trajectories are monitored with advanced detection techniques, and results are archived to ensure errors are managed effectively. The cofactor filtering process boosts tracking accuracy using dynamic and static strategies. The dynamic method filters out compounds on both sides of the reaction, while the static method removes known cofactors from the reactions. CBR-db has 148,651 atom-tracked reactions (complete coverage), compared to 43,000 in MetAMDB.[31]

A continuously updated, comprehensive list of all reactions and compounds identified or modified is available on GitHub (https://github.com/ELIFE-ASU/CBRdb/tree/main/data) and Zenodo (https://zenodo.org/records/18380137). See the Supplementary file for a breakdown of all data fields in both the compound and reaction database files, facilitating easy reference to all data types.

Pruning Hypothetical Reaction Spaces

ATLAS has demonstrated impressive predictive power by successfully suggesting many new reactions. A potential issue is that the BNICE.ch rules used to generate ATLAS may be overly permissive, leading to reactions unlikely to occur in biochemistry. In CBR-db, we provide a heuristic to compare reaction similarity and enable users to set a threshold, thereby pruning reactions that are too dissimilar to a reference reaction space, e.g., reactions unlike any found in biochemistry. Figure 3a shows the reaction similarity score for each reaction in the

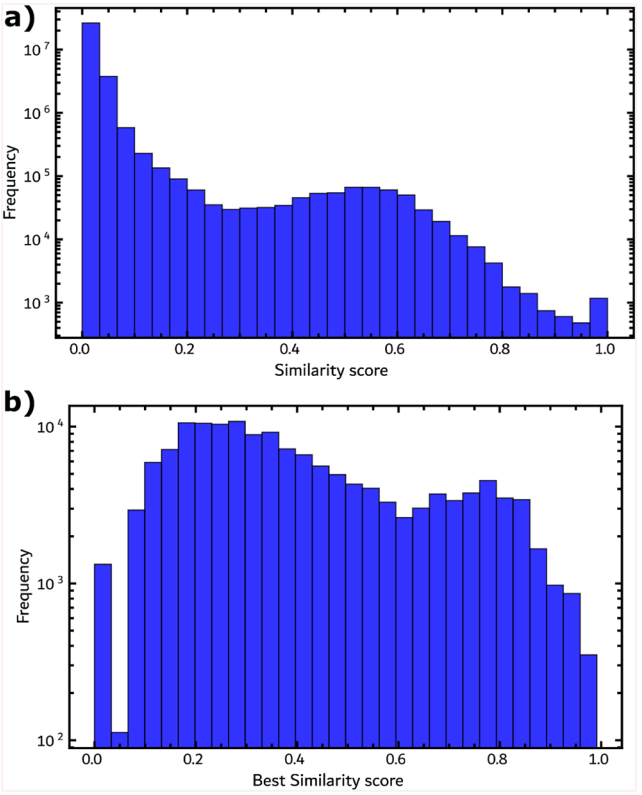


Figure 3. Reaction similarity of components CBR-db can be tested using the differential reaction fingerprint DRFP.[27] a) Comparing every entry of the KEGG subsection to every other reaction in KEGG. b) Comparing every entry of the ATLAS subsection to every reaction in KEGG, the highest score, or closest KEGG entry is shown as the "best score".

KEGG subsection of CBR-db compared to every other reaction. We plot a histogram and remove self-comparisons. The goal is to quantify the similarity among the different subsections of CBR-db and to provide a metric to prune reactions that are dissimilar to those in the KEGG database. Some reactions are remarkably similar and thus fall into the upper similarity-score bin; for example, reactions with inputs or outputs that are highly similar or differ by only one minor point, such as the addition of an atom. A much wider distribution is observed at a similarity score of 0.55, whereas a narrower distribution is centered at zero. Similarity scores below 0.50 between any two reactions are typically considered dissimilar. At the bin near zero, this captures KEGG reactions that are not like any other reactions in KEGG. Examples include disconnected reactions that share no compounds with any other reaction, a point initially highlighted in the first ATLAS paper. ATLAS was introduced to address cases in which disconnected reaction nodes should retain some connectivity.

On the other hand, Figure 3b shows a pairwise comparison of the ATLAS subsection of the CBR-db to the KEGG subsection. Each ATLAS reaction entry is iteratively compared to all KEGG entries, and the highest score is shown. Again, the bin nearest unity corresponds to reactions that ATLAS has added, which are highly similar to each other. Meanwhile, the general distribution is remarkably flat. Consequently, this indicates that ATLAS successfully added a homogeneous mixture of reactions in the close neighborhood of each KEGG reaction. However, the bins closest to zero indicate that some

https://doi.org/10.1021/acssynbio.6c00060
ACS Synth. Biol. 2026, 15, 2684−2692

reactions added by ATLAS lack clear KEGG analogues and may therefore be speculative.

CBR-db database users can prune these reactions using a simple filter, as each has been assigned a similarity score. Thus, a more conservative yet well-connected database can be created to meet the user's needs.

CONCLUSION

While several other cheminformatic and bioinformatic databases already exist, such as BRENDA,[32] MetaCyc,[33] SABIO-RK,[34] and Rhea,[35] each database has a specific scope. BRENDA focuses on enzymes; MetaCyc emphasizes pathway context; SABIO-RK provides reaction kinetics; and Rhea features a vast database of standardized reactions with cross-referencing. Similarly, other databases provide a cleaned, merged, and interlinked resource specifically based on KEGG or BRENDA, such as EnzymeMap,[36] MetAMDB,[31] and Reactzyme.[37] CBR-db is the first data set to integrate chemical space analyses of reactions and chemical properties, with high coverage across KEGG and ATLAS, and is specifically designed for studying the structure and properties of biochemistry within chemical space. The data we provide includes detailed atom tracking, reaction similarity classification, compound properties, chiral centers, molecular assembly values, and thermodynamic properties.

CBR-db provides a human- and machine-readable compound and reaction database that includes comprehensive atom-tracking, reaction classification, compound properties, and thermodynamic properties. The improvements implemented in CBR-db significantly enhance the utility of biochemical compound and reaction data for applications that require chemically accurate reaction models, such as large-scale physicochemical reaction analysis, thermodynamic modeling, machine learning, drug design, retrosynthesis, metabolic networks, origins-of-life, and evolution studies.[8,9,38−40] The curated database is maintained and accessed through its dedicated repository.[41]

Large-scale data cleaning inherently introduces certain limitations. For example, large glycans and macromolecules (e.g., proteins, peptides, and nucleotide oligos) are commonly unaccompanied by structural information in KEGG, meaning that reactions involving them would not be found in CBR-db's curated database. This currently limits the applicability of CBR-db to research questions concerning these molecular species. We invite the community to contribute structural information for compounds that currently lack it.

While stoichiometric corrections resolved many issues, additional experimental validation of specific reactions (e.g., by resolving generic functional groups such as halogens) is necessary to confirm their biochemical relevance. This is also true for hypothesized reactions from the ATLAS database, a majority of which have not yet been experimentally confirmed in living organisms. CBR-db is designed to be extensible, continuously updated, open source, and reproducible. Future updates to CBR-db will expand coverage to include additional biochemical reaction databases (e.g., ATLASx) and ensure seamless updates with newer releases of these repositories.

ASSOCIATED CONTENT

Data Availability Statement

The source code for all packages mentioned or used is open source and freely available. CBR-db is available at https://github.com/ELIFE-ASU/CBRdb. We expect potential users and contributors to CBR-db to bring a variety of computational infrastructures and backgrounds. As such, each major version of CBR-db will be available in multiple formats. First, CBR-db may be installed from source as a Python package. This option is best suited for users who wish to reproduce the full pipeline, download raw data, or replicate thermodynamic calculations. Second, CBR-db may be accessed within a prebuilt Binder container. This option is best suited for those who want to examine or demonstrate most postdownload processes. Lastly, the primary data products for each major version are available for download as CSV files, which can be queried using standard data analysis tools (such as Pandas). This option is best suited to most applications (e.g., chemical network construction, statistical analysis, and compound filtering based on conditions). The exact container digest and the Git commit corresponding to this study (CBR-db v1.9.0) are archived in the Zenodo record at https://zenodo.org/records/18380137. The Binder container is available at https://mybinder.org/v2/gh/ELIFE-ASU/CBRdb/HEAD.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.6c00060.

Table S1 of general information for compound data, including entry titles and description; Table S2 of compound property entry titles and their description; Table S3 of compound metadata, including entry titles for compound IDs; Table S4 of compound metadata, including entry titles related to KEGG-specific information; Table S5 of cross-reference entry titles and descriptions of what these identifiers point to in databases; Table S6 of general information for reaction data, including entry titles and descriptions; Table S7 of reaction metadata, including entry titles for reaction IDs; Table S8 of reaction metadata, including entry titles related to KEGG-specific information; Table S9 of entry titles for ATLAS-specific information; and Table S10 of entry titles for data processing and curation flags (PDF)

AUTHOR INFORMATION

Corresponding Author

Sara Imari Walker − *Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, Arizona 85287−0506, United States; School of Earth and Space Exploration and School of Complex Adaptive Systems, Arizona State University, Tempe, Arizona 85287−0506, United States; Santa Fe Institute, Santa Fe, New Mexico 87501, United States;* orcid.org/0000-0001-5779-2772; Email: sara.i.walker@asu.edu

Authors

Louie Slocombe − *Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, Arizona 85287−0506, United States*

Camerian Millsaps − *School of Earth and Space Exploration, Arizona State University, Tempe, Arizona 85287−0506, United States*

M. Reza Shahjahan − *School of Earth and Space Exploration, Arizona State University, Tempe, Arizona 85287−0506, United States*

https://doi.org/10.1021/acssynbio.6c00060
ACS Synth. Biol. 2026, 15, 2684−2692

Kamesh Narasimhan − *Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, Arizona 85287−0506, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.6c00060

Author Contributions

L.S.: project administration, methodology, validation, original draft, and reviewand editing. C.M.: conceptualization, methodology, and validation. M.R.S.: original draft and methodology. K.N.: original draft and data curation. S.I.W.: conceptualization, project administration, supervision, funding acquisition, and review and editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Pilar Vergeli and Hikaru Furukawa for their helpful feedback on the development and curation of the database. We also thank the ATLAS database team for providing access to the database on which CBR-db was built, enhanced, and used to improve the quality of the compound and reaction datasets. We acknowledge Schmidt Sciences for support and NASA grant number 80NSSC25M7051.

■ REFERENCES

(1) Kanehisa, M. K. Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **2000**, *28* (1), 27−30.
(2) Hadadi, N.; Hafner, J.; Shajkofci, A.; Zisaki, A.; Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol* **2016**, *5* (10), 1155−1166.
(3) López-López, E.; Bajorath, J.; Medina-Franco, J. L. Informatics for Chemistry, Biology, and Biomedical Sciences. *J. Chem. Inf. Model* **2021**, *61* (1), 26−35.
(4) Beber, M. E.; Gollub, M. G.; Mozaffari, D.; Shebek, K. M.; Flamholz, A. I.; Milo, R.; Noor, E. eQuilibrator 3.0: A Database Solution for Thermodynamic Constant Estimation. *Nucleic Acids Res* **2022**, *50* (D1), D603−D609.
(5) Jankowski, M. D.; Henry, C. S.; Broadbelt, L. J.; Hatzimanikatis, V. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. *Biophys. J* **2008**, *95* (3), 1487−1499.
(6) Lin, G.-M.; Warden-Rothman, R.; Voigt, C. A. Retrosynthetic Design of Metabolic Pathways to Chemicals Not Found in Nature. *Curr. Opin. Syst. Biol* **2019**, *14*, 82−107.
(7) Hadadi, N.; Hatzimanikatis, V. Design of Computational Retrobiosynthesis Tools for the Design of de Novo Synthetic Pathways. *Curr. Opin. Chem. Biol* **2015**, *28*, 99−104.
(8) MohammadiPeyhani, H.; Chiappino-Pepe, A.; Haddadi, K.; Hafner, J.; Hadadi, N.; Hatzimanikatis, V. NICEdrug.Ch, a Workflow for Rational Drug Design and Systems-Level Analysis of Drug Metabolism. *eLife* **2021**, *10*, No. e65543.
(9) Hafner, J.; Hatzimanikatis, V. NICEpath: Finding Metabolic Pathways in Large Networks through Atom-Conserving Substrate−Product Pairs. *Bioinformatics* **2021**, *37* (20), 3560−3568.
(10) Lin, A.; Dyubankova, N.; Madzhidov, T. I.; Nugmanov, R. I.; Verhoeven, J.; Gimadiev, T. R.; Afonina, V. A.; Ibragimova, Z.; Rakhimbekova, A.; Sidorov, P.; Gedich, A.; Suleymanov, R.; Mukhametgaleev, R.; Wegner, J.; Ceulemans, H.; Varnek, A. Atom-to-atom Mapping: A Benchmarking Study of Popular Mapping Algorithms and Consensus Strategies. *Mol. Inf* **2022**, *41* (4), 2100138.
(11) Chetry, A. B.; Ohto, K. From Molecules to Data: The Emerging Impact of Chemoinformatics in Chemistry. *J. Cheminf.* **2025**, *17* (1), 121.
(12) Hafner, J.; MohammadiPeyhani, H.; Sveshnikova, A.; Scheidegger, A.; Hatzimanikatis, V. Updated ATLAS of Biochemistry with New Metabolites and Improved Enzyme Prediction Power. *ACS Synth. Biol* **2020**, *9* (6), 1479−1482.
(13) Hafner, J.; Payne, J.; MohammadiPeyhani, H.; Hatzimanikatis, V.; Smolke, C. A Computational Workflow for the Expansion of Heterologous Biosynthetic Pathways to Natural Product Derivatives. *Nat. Commun* **2021**, *12* (1), 1760.
(14) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7* (1), 23.
(15) Weininger, D.; Smiles, A. Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci* **1988**, *28* (1), 31−36.
(16) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; Sriniker; Gedeck, P.; Jones, G.; Kawashima, E., et al. *Rdkit/Rdkit: 2025_09_2 (Q3 2025) Release*, Zenodo 2025.
(17) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Res* **2023**, *51* (D1), D1373−D1380.
(18) Ropp, P. J.; Kaminsky, J. C.; Yablonski, S.; Durrant, J. D. Dimorphite-DL: An Open-Source Program for Enumerating the Ionization States of Drug-like Small Molecules. *J. Cheminf.* **2019**, *11* (1), 14.
(19) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Bjerre Jensen, P.; Kermode, J.; Kitchin, J. R.; Leonhard Kolsbjerg, E.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Bergmann Maronsson, J.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The Atomic Simulation Environment—a Python Library for Working with Atoms. *J. Phys.: Condens. Matter* **2017**, *29* (27), 273002.
(20) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA Quantum Chemistry Program Package. *J. Chem. Phys* **2020**, *152* (22), 224108.
(21) Neese, F. The SHARK Integral Generation and Digestion System. *J. Comput. Chem* **2023**, *44* (3), 381−396.
(22) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. r2SCAN-3c: A "Swiss Army Knife" Composite Electronic-Structure Method. *J. Chem. Phys* **2021**, *154* (6), 064103.
(23) Jirasek, M.; Sharma, A.; Bame, J. R.; Mehr, S. H. M.; Bell, N.; Marshall, S. M.; Mathis, C.; MacLeod, A.; Cooper, G. J. T.; Swart, M.; Mollfulleda, R.; Cronin, L. Investigating and Quantifying Molecular Complexity Using Assembly Theory and Spectroscopy. *ACS Cent. Sci* **2024**, *10* (5), 1054−1064.
(24) Marshall, S. M.; Mathis, C.; Carrick, E.; Keenan, G.; Cooper, G. J. T.; Graham, H.; Craven, M.; Gromski, P. S.; Moore, D. G.; Walker, S. I.; Cronin, L. Identifying Molecules as Biosignatures with Assembly Theory and Mass Spectrometry. *Nat. Commun* **2021**, *12* (1), 3033.
(25) Seet, I.; Patarroyo, K. Y.; Siebert, G.; Walker, S. I.; Cronin, L. Rapid Exploration of the Assembly Chemical Space of Molecular Graphs. *J. Chem. Inf. Model* **2025**, *65* (24), 13203−13214.
(26) Dahlgren, B. ChemPy: A Package Useful for Chemistry Written in Python. *J. Open Source Softw.* **2018**, *3* (24), 565.
(27) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction Classification and Yield Prediction Using the Differential Reaction Fingerprint DRFP. *Digital Discovery* **2022**, *1* (2), 91−97.
(28) Chen, S.; An, S.; Babazade, R.; Jung, Y. Precise Atom-to-Atom Mapping for Organic Reactions via Human-in-the-Loop Machine Learning. *Nat. Commun* **2024**, *15* (1), 2250.
(29) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of Organic Chemistry Grammar from Unsupervised Learning of Chemical Reactions. *Sci. Adv* **2021**, *7* (15), No. eabe4166.
(30) Astero, M.; Rousu, J. Learning Symmetry-Aware Atom Mapping in Chemical Reactions through Deep Graph Matching. *J. Cheminf.* **2024**, *16* (1), 46.

(31) Starke, C.; Wegner, A. MetAMDB: Metabolic Atom Mapping Database. *Metabolites* **2022**, *12* (2), 122.
(32) Placzek, S.; Schomburg, I.; Chang, A.; Jeske, L.; Ulbrich, M.; Tillack, J.; Schomburg, D. BRENDA in 2017: New Perspectives and New Tools in BRENDA. *Nucleic Acids Res* **2017**, *45* (D1), D380−D388.
(33) Caspi, R.; Billington, R.; Keseler, I. M.; Kothari, A.; Krummenacker, M.; Midford, P. E.; Ong, W. K.; Paley, S.; Subhraveti, P.; Karp, P. D. The MetaCyc Database of Metabolic Pathways and Enzymes - a 2019 Update. *Nucleic Acids Res* **2020**, *48* (D1), D445−D453.
(34) Wittig, U.; Rey, M.; Weidemann, A.; Kania, R.; Müller, W. SABIO-RK: An Updated Resource for Manually Curated Biochemical Reaction Kinetics. *Nucleic Acids Res* **2018**, *46* (D1), D656−D660.
(35) Bansal, P.; Morgat, A.; Axelsen, K. B.; Muthukrishnan, V.; Coudert, E.; Aimo, L.; Hyka-Nouspikel, N.; Gasteiger, E.; Kerhornou, A.; Neto, T. B.; Pozzato, M.; Blatter, M.-C.; Ignatchenko, A.; Redaschi, N.; Bridge, A. Rhea the Reaction Knowledgebase in 2022. *Nucleic Acids Res* **2022**, *50* (D1), D693−D700.
(36) Heid, E.; Probst, D.; Green, W. H.; Madsen, G. K. H. EnzymeMap: Curation, Validation and Data-Driven Prediction of Enzymatic Reactions. *Chem. Sci* **2023**, *14* (48), 14229−14242.
(37) Hong, L.; Hua, C.; Luan, S.; Precup, D.; Wolf, G.; Zheng, S.; Zhong, B. ReactZyme: A Benchmark for Enzyme-Reaction Prediction. In *Advances in Neural Information Processing Systems 37*. Neural Information Processing Systems Foundation, Inc. BC, Canada, 2024, pp. 26415−26442. .
(38) Buehler, Y.; Reymond, J.-L. A View on Molecular Complexity from the GDB Chemical Space. *J. Chem. Inf. Model* **2025**, *65* (16), 8405−8410.
(39) Goldford, J. E.; Segrè, D. Modern Views of Ancient Metabolic Networks. *Curr. Opin. Syst. Biol* **2018**, *8*, 117−124.
(40) Li, F. Filling Gaps in Metabolism Using Hypothetical Reactions. *Proc. Natl. Acad. Sci. U. S. A* **2022**, *119* (49), No. e2217400119.
(41) Slocombe, L.; Millsaps, C.; Shahjahan, M. R.; Narasimhan, K. *Sara Imari Walker. ELIFE-ASU/CBRdb: V1.9.0*. Zenodo 2026.

https://doi.org/10.1021/acssynbio.6c00060
ACS Synth. Biol. 2026, 15, 2684−2692