CBR-db: A Curated Biochemical Reaction Database for Precise Biochemical Reaction Analysis

Louie Slocombe,[†] Camerian Millsaps,[‡] Kamesh Narasimhan,[†] and Sara Imari Walker*[,†,‡,¶,§]

[†]*Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe AZ USA*

[‡]*School of Earth and Space Exploration, Arizona State University, Tempe AZ USA*

[¶]*School of Complex Adaptive Systems, Arizona State University, Tempe AZ USA*

[§]*Santa Fe Institute, Santa Fe, NM USA*

E-mail: sara.i.walker@asu.edu

Abstract

We present CBR-db, a curated biochemical reaction database that integrates and refines data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the ATLAS of Biochemistry databases to enable chemically consistent analyses of biochemical reaction data. This curation effort addresses key limitations of both KEGG and ATLAS, such as malformed chemical representations, inaccurate stoichiometry, and ambiguous or incomplete reaction entries. These limitations are addressed in CBR-db, enabling more chemically realistic analyses of biochemical reactions and hypothesized reactions and reaction networks constructed from these, which are essential for applications in research areas ranging from prebiotic chemistry to metabolism and its evolution, and synthetic biology and metabolic engineering. Altogether, CBR-db features 148,673 high-quality reactions and 18,716 compounds, with details of the refinement and curation procedures highlighted in this report. CBR-db is designed to be continuously updated, incorporating the latest releases from the KEGG and ATLAS databases. Furthermore, it provides a rigorous framework so that the reaction list can be extended, and further issues can be improved.

CBR-db is provided as an open-access resource, enabling researchers to leverage its curated biochemical reaction data[1].

Introduction

Biochemical databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the ATLAS of Biochemistry offer extensive data that underpins a variety of computational and experimental workflows within synthetic biology, metabolic engineering, drug design, and evolutionary biology. Established in 1995, the KEGG database has become essential for research into biochemical reactions, pathways, and networks[2]. However, its limitations include the absence of intermediate reactions, metabolites not linked to reaction pathways, and incomplete enzyme characterization. To attempt to address this deficit, the ATLAS of Biochemistry database was developed by applying a large set of reaction rules and prediction algorithms to theoretically predict the entire possible reactome of enzyme-catalyzed reactions from what is cataloged of known metabolism in KEGG[3]. Consequently, ATLAS systematically addresses the gaps left by KEGG by hypothesizing enzymatically catalyzed reactions that have never been reported to occur in living organisms yet align with known enzyme activities, thus providing a valuable resource for investigating new reaction pathways and predicting possible biochemical functionality missing from current data[4].

ATLAS inherited many data quality issues present in KEGG and introduced several new ones, such as incomplete stoichiometry, missing elements, and inconsistent chemical representations. This limits effectiveness for downstream applications where precise and high-quality reaction datasets are necessary. Building upon these foundational efforts, we introduce CBR-db, a Curated Biochemical Reaction database that integrates KEGG and ATLAS[1]. Our work addresses these challenges and enhances compound and reaction data quality through systematic curation, which

includes correcting malformed compounds, re-balancing stoichiometric equations, and removing ambiguous or incomplete entries. With 148,673 high-quality reactions and 18,716 compounds, CBR-db enables applications that demand high-quality biochemical reaction datasets. These include metabolic network reconstruction for synthetic biology, flux balance analysis for bioprocess optimization, retrosynthesis, atom-tracking, and reaction prediction in cheminformatics, and even prebiotic chemistry modeling to explore the origins of metabolism[5–11]. CBR-db thus represents a significant value to the research community as a complement to KEGG and ATLAS, tailored for applications requiring accurate chemical reaction modeling.

Methods

The latest KEGG (v.113, dated January 1, 2025) release was integrated with ATLAS 2018 data[4]. KEGG reaction entries superseded their corresponding ATLAS reaction, provided all quality control checks (as noted below) were passed. Duplicated entries were removed during the integration to ensure the inclusion of the most up-to-date and accurate biochemical information. The curation efforts were structured around two primary components: compounds and reactions (Figure 1).

Compound Curation

The first stage of the curation process focused on addressing inconsistencies and inaccuracies in the KEGG compound data (Figure 1).

Standardization of Chemical Representations

All compounds were converted to consistent and reproducible SMILES and InChI formats to ensure compatibility with cheminformatics tools and computational workflows[12,13]. Compounds

were standardized and sanitized using the RDKit standardization and sanitize functions to resolve improper stereochemical notations that impacted 3D embedding and chirality calculations[14].

Correction of Elemental Entries and Ambiguous Notations

Compounds with malformed elemental entries (e.g., treating OH as an element instead of a compound) were reprocessed to conform to chemical conventions. Inconsistent representations of functional groups, including R-group and nitrogen notations, were also standardized to remove ambiguities.

Removal of Undefined or Incomplete Entries

Compounds lacking essential information, such as structural information or empirical formulas, were excluded from the curated dataset. This step ensured that only chemically well-defined compounds were kept for subsequent analyses.

Reaction Curation

The curation of reactions was based on the refined compound dataset and aimed at eliminating incomplete, ambiguous, and chemically invalid entries from both KEGG and ATLAS (Figure 1).

Treatment of Incomplete and Redundant Reactions

Reactions labeled as “incomplete” or “unclear” in KEGG were systematically targeted for correction, as a lack of critical reactants or products would otherwise make them unsuitable for accurate computational analyses. Furthermore, “multi-step” and “overall” reactions defined by KEGG, which summarize complex processes without offering unique information, were removed.

Filtering of Reactions with Large Macromolecules and Missing Co-Factors

Entities outside the scope of metabolite-level studies, such as large glycans and proteins, were excluded. This includes reaction entries that implicitly involve large macromolecules, but information on their transformations is insufficiently captured in compound entries. We also removed entries where an implied or missing coenzyme or cofactor has not been fully represented in the reaction to ensure that only high-confidence reactions are included in the dataset.

Stoichiometric Balancing and Reaction Completion

All reactions were reassessed for stoichiometric consistency. When feasible, missing elements and common small molecules were incorporated to balance the equations. Left and right-side imbalances were systematically rectified, ensuring chemically valid transformations (Figure 2).

Handling Halogens

Reactions involving generic halogens were refined by instantiating specific entities. To address these issues, 27 new reactions were added, enhancing the dataset's comprehensiveness.

Results and Discussion

The curation and refinement of KEGG and ATLAS reaction data in CBR-db led to significant improvements 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: dataset size and scope and data quality improvements. These enhancements provide a well-curated dataset that improves upon existing resources while offering a more curated and reliable dataset tailored for precise biochemical modeling required for those downstream applications that require accurate reaction chemistry.

Dataset Size and Scope

CBR-db integrates data from KEGG and ATLAS, resulting in a meticulously curated biochemical reaction dataset that prioritizes accuracy and reliability by eliminating spurious and low-quality reaction entries. After removing duplicate or superseded entries, the integration yielded 18,716 unique compound entries and 148,673 unique reactions. In total, 2517 (~20.62%) KEGG and 13,185 (~8.8%) ATLAS reaction entries either needed significant updates and corrections due to issues such as malformed mol representations, improper protonation states, and incorrect elemental notations, or were filtered out from the initial pool for failing to meet our quality control standards. To address gaps and ambiguities in reactions, new inferred reactions were created by resolving generic halogen species, thus expanding the inferred reaction space compared to the ATLAS and KEGG databases. This effort ensures that the data set provides high-quality reactions, facilitating its application to research questions where precise chemical reaction models are crucial.

Data Quality Improvements

Stoichiometric Balancing, Elemental balancing

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 injecting missing elements or molecules, such as Fe or H_2O, and rebalancing the reactant and product sides (Figure 2). 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 definitions, were resolved. The resulting dataset fully standardizes compound annotations, enhancing its utility for tasks requiring accurate 3D embeddings or chirality calculations.

Removal of Ambiguous and Redundant Reactions

Of unbalanced reactions, those tagged as "general," "unclear," or "incomplete" in KEGG were systematically flagged, attempted to be fixed, and excluded if the reaction could not be salvaged. This step eliminated entries unsuitable for quantitative modeling and ensured the clarity of the curated data. Similarly, KEGG reactions labeled as "overall" or multi-step and annotated with their constituent parts are removed. Our choice to omit these reduces redundancy and ensures a more consistent representation of chemical data, with all data at the same level of granularity, suitable for computational workflows.

Conclusion

The improvements in CBR-db greatly enhance its utility for specific downstream applications. Nonetheless, the deliberate large-scale pruning of low-quality data 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 for research questions associated to these molecular species. However, CBR-db is extensible, and we invite the community to contribute structural information for compounds currently lacking such data. Furthermore, while stoichiometric corrections resolved many issues, additional experimental validation of specific reactions (where algorithmic balancing of reactions was implemented or where generic functional groups such as halogens were resolved) 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. Despite these limitations, the improvements implemented in CBR-db significantly enhance the utility of biochemical compound and reaction data for those downstream applications where chemically accurate models of reactions are required, such as for large-scale

physicochemical analysis of reactions, thermodynamic modeling, machine learning, retrosynthesis, and metabolic network and evolution studies[15,16]. The curated database is maintained and can be accessed through its dedicated repository[1]. Future updates to CBR-db will include expanding coverage to incorporate additional biochemical reaction databases (e.g., BRENDA) and ensuring seamless updates with newer releases of these repositories.

Acknowledgments

The authors thank Pilar Vergeli and Hikaru Furukawa for their helpful feedback on the database's development and curation. We also thank the ATLAS database team for providing access to their database, upon which CBR-db was able to build, enhance, and curate the quality of the compound and reaction dataset. We acknowledge Schmidt Sciences for support and NASA grant number 80NSSC21K1402.

Table 1: Comparison of Key Features in KEGG, ATLAS, and CBR-db.

Feature	KEGG	ATLAS	CBR-db
Dataset Size and Scope			
Number of Reactions	12,203	149,052	148,673
Number of Compounds	19,438	19,438	18,716
Data Quality Improvements			
Compounds with malformed, incomplete, or duplicated data	722	722	0
Reactions with malformed, incomplete, or duplicated data	1711	1249	0
Stoichiometrically unbalanced (after pruning)	806	11936	0
Generic Halogen Reactions Resolved	No	Partial	Full
Standardized Compound Representations	Partial	Partial	Full
Integration of Latest KEGG Release	Full	Partial	Full

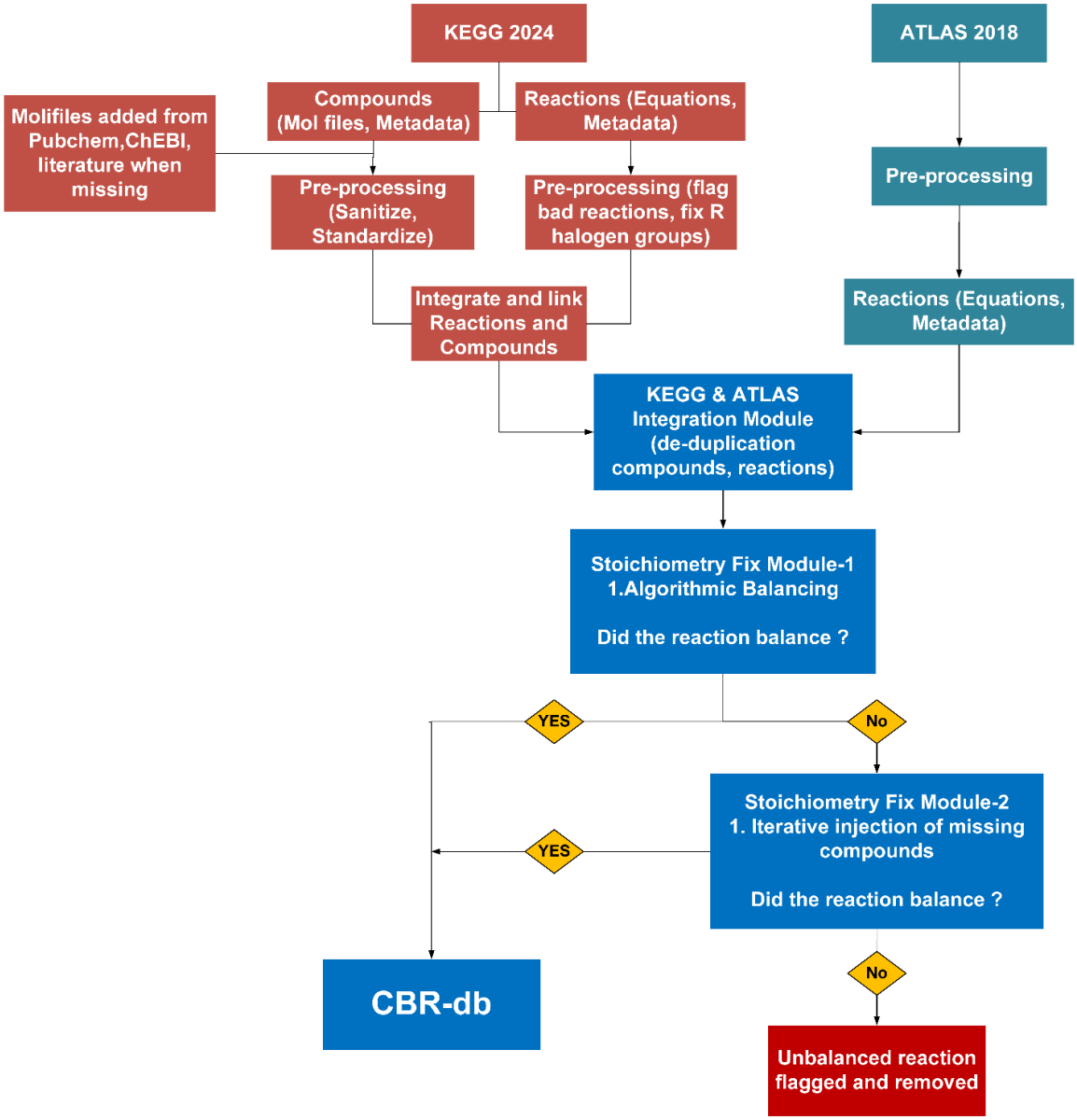

Figure 1: Construction of CBR-db by merging KEGG 2024 and ATLAS 2018 and employing a series of quality control measures by fixing stoichiometry, adding missing mol files, and ensuring a unique, high-quality biochemical reaction dataset.

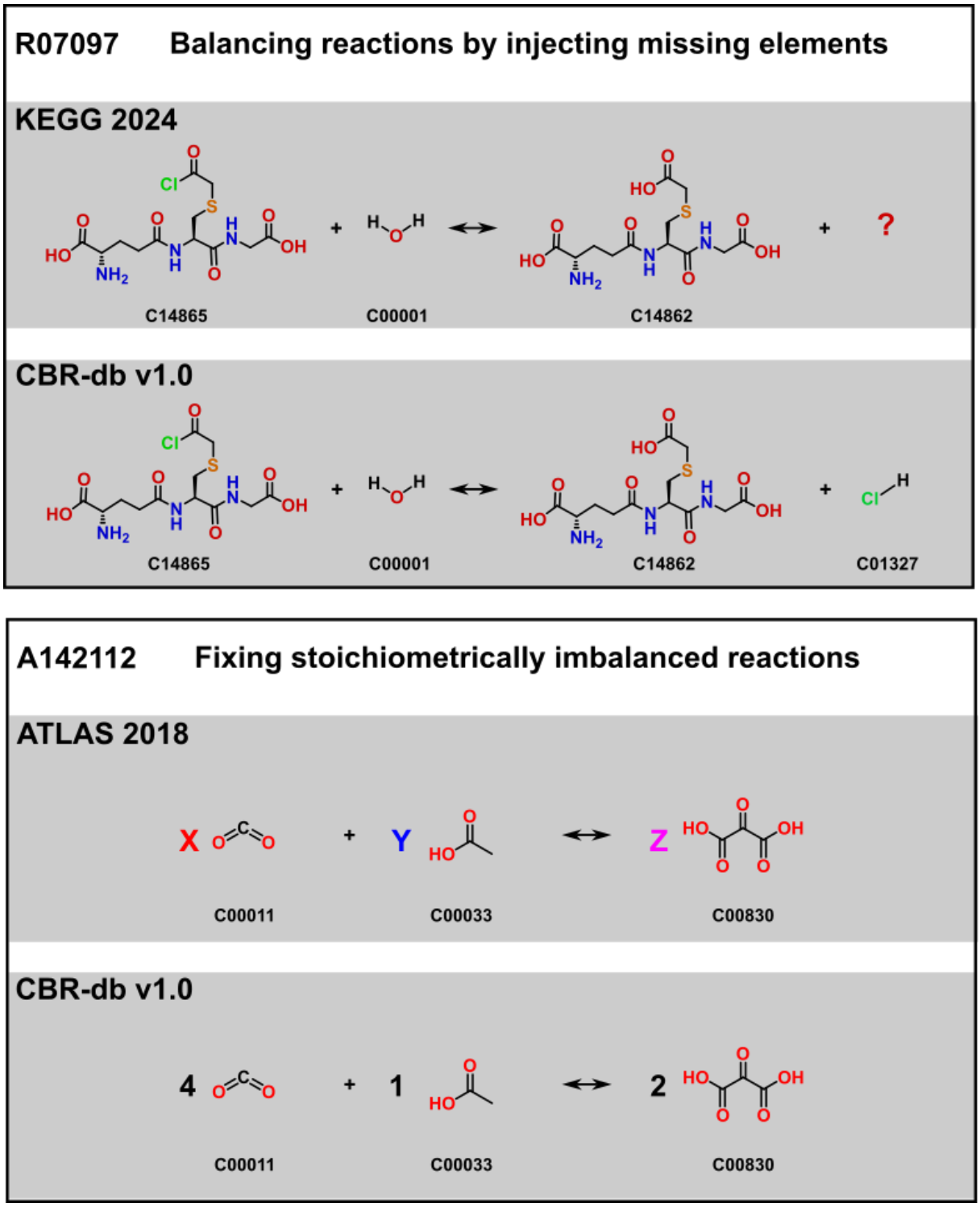

Figure 2: Examples showing improvements to reactions in CBR-db for certain error classes. The first panel shows an example where there are missing compounds; the second panel shows an example of an imbalanced reaction.

References

(1) Slocombe, L.; Millsaps, C.; Narasimhan, K.; Walker, S. ELIFE-ASU/CBRdb: V1.3.0. https://doi.org/10.5281/ZENODO.14948473.

(2) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **2000**, *28* (1), 27–30. https://doi.org/10.1093/NAR/28.1.27.

(3) 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. https://doi.org/10.1021/ACSSYNBIO.6B00054.

(4) 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. https://doi.org/10.1021/ACSSYNBIO.0C00052.

(5) 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. https://doi.org/10.1093/NAR/GKAB1106.

(6) 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. https://doi.org/10.1529/BIOPHYSJ.107.124784.

(7) 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. https://doi.org/10.1016/J.COISB.2019.04.004.

(8) 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. https://doi.org/10.1016/J.CBPA.2015.06.025.

(9) 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*. https://doi.org/10.7554/ELIFE.65543.

(10) Hafner, J.; Hatzimanikatis, V. NICEpath: Finding Metabolic Pathways in Large Networks through Atom-Conserving Substrate–Product Pairs. *Bioinformatics* **2021**, *37* (20), 3560–3568. https://doi.org/10.1093/BIOINFORMATICS/BTAB368.

(11) 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 Inform* **2022**, *41* (4). https://doi.org/10.1002/minf.202100138.

(12) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J Cheminform* **2015**, *7* (1). https://doi.org/10.1186/s13321-015-0068-4.

(13) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J Chem Inf Comput Sci* **1988**, *28* (1). https://doi.org/10.1021/ci00057a005.

(14) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; sriniker; Gedeck, P.; Jones, G.; NadineSchneider; Kawashima, E.; Nealschneider, D.; Dalke, A.; Swain, M.; Cole, B.; Turk, S.; Savelev, A.; Vaucher, A.; Wójcikowski, M.; Take, I.; Scalfani, V. F.; Walker, R.; Probst, D.; Ujihara, K.; tadhurst-cdd; Pahl, A.; godin, guillaume; Lehtivarjo, J.; Bérenger, F.; Bisson, J. Rdkit/Rdkit: 2024_09_2 (Q3 2024) Release. https://doi.org/10.5281/ZENODO.13990314.

(15) Goldford, J. E.; Segrè, D. Modern Views of Ancient Metabolic Networks. *Curr Opin Syst Biol* **2018**, *8*, 117–124. https://doi.org/10.1016/J.COISB.2018.01.004.

(16) Li, F. Filling Gaps in Metabolism Using Hypothetical Reactions. *Proc Natl Acad Sci U S A* **2022**, *119* (49). https://doi.org/10.1073/PNAS.2217400119.