

Updated ATLAS of Biochemistry with New Metabolites and Improved Enzyme Prediction Power

Jasmin Hafner,[‡] Homa MohammadiPeyhani,[‡] Anastasia Sveshnikova, Alan Scheidegger, and Vassily Hatzimanikatis*



Cite This: *ACS Synth. Biol.* 2020, 9, 1479–1482



Read Online

ACCESS |



Metrics & More



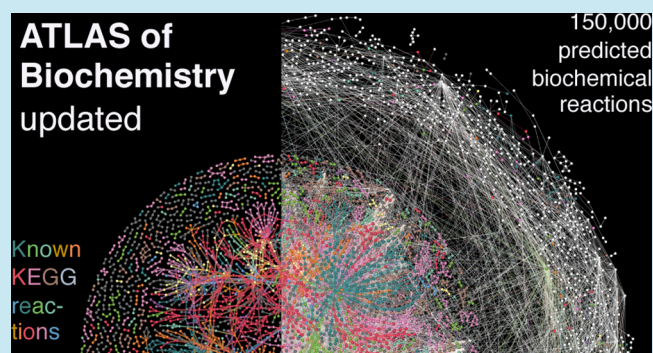
Article Recommendations



Supporting Information

ABSTRACT: The ATLAS of Biochemistry is a repository of both known and novel predicted biochemical reactions between biological compounds listed in the Kyoto Encyclopedia of Genes and Genomes (KEGG). ATLAS was originally compiled based on KEGG 2015, though the number of KEGG reactions has increased by almost 20 percent since then. Here, we present an updated version of ATLAS created from KEGG 2018 using an increased set of generalized reaction rules. Furthermore, we improved the accuracy of the enzymes that are predicted for catalyzing novel reactions. ATLAS now contains ~150 000 reactions, out of which 96% are novel. In this report, we present detailed statistics on the updated ATLAS and highlight the improvements with regard to the previous version. Most importantly, 107 reactions predicted in the original ATLAS are now known to KEGG, which validates the predictive power of our approach. The updated ATLAS is available at <https://lcsb-databases.epfl.ch/atlas>.

KEYWORDS: reaction prediction, enzyme prediction, enzyme promiscuity, metabolic networks, biochemical database



Predicting hypothetical biochemical reactions and catalyzing enzymes is needed to design novel pathways in metabolic engineering and to fill knowledge gaps in our understanding of metabolism. The ATLAS of Biochemistry¹ is a database of known and predicted biochemical reactions that was compiled by taking the biological data available in the Kyoto Encyclopedia of Genes and Genomes (KEGG) and predicting the biochemical reactions that would produce the contained compounds. Published in 2016, the utility of ATLAS has been recognized by several reviews as a source of novel metabolic reactions for enzyme and metabolic engineering.^{2–4} More recently, Yang *et al.* experimentally validated hypothetical ATLAS reactions and used them to construct novel one-carbon assimilation pathways.⁵ However, ATLAS was created based on the biochemical knowledge available in KEGG 2015.⁶ Since then, KEGG has added 802 new metabolites, 918 new reactions, and 633 enzymes to its collection.

The expansion of biochemical reactions within ATLAS relies on the reaction prediction tool BNICE.ch^{7–12} (Biochemical Network Integrated Computational Explorer), which consists of (i) a large set of expert-curated, generalized reaction rules that mimic the promiscuous activity of enzymes, and (ii) a network-generating algorithm that applies the reaction rules to molecular structures to generate possible biochemical reactions and compounds. The BNICE.ch reaction rules can reconstruct

known biochemical reactions, as well as generate novel, hypothetical reactions. Currently, BNICE.ch has 400 bidirectional reaction rules that account for both the forward and reverse reaction directionality. More than 130 000 novel biochemical reactions between known biological compounds have been predicted using this strategy.

Herein, we integrated the new KEGG 2018 data into our database and expanded the biochemical space covered by ATLAS from 137 877 to 149 052 reactions. Interestingly, we found that the newly available data validated 107 novel reactions predicted in ATLAS 2015. In the following, we discuss the updated ATLAS statistics and illustrate the improvements compared to the first version. The latest version of ATLAS is available online (<https://lcsb-databases.epfl.ch/atlas>).

METHODS

The ATLAS Workflow. To generate the new version of ATLAS, we applied the BNICE.ch reaction rules to all of the

Received: January 29, 2020

Published: May 18, 2020



Table 1. Overview of Compound, Reaction, and Enzyme Statistics in KEGG and ATLAS

		ATLAS 2015	ATLAS 2018	percent change
KEGG compounds	Total number of compounds	17 450	18 254	+5%
	Filtered compounds (<i>fc</i>)	16 798	17 255	
	Orphan KEGG compounds (<i>okc</i>)	9371 (56% of <i>fc</i>)	9857 (57% of <i>fc</i>)	
KEGG reactions	Total number of reactions	9135	10 829	+19%
	Filtered reactions	8592	10 753	
BNICE.ch	Number of bidirectional enzymatic reaction rules	360	400	+11%
KEGG reaction reconstruction	Covered reactions total	6651	8118	+22%
	Exact coverage	5270	5779	
	Alternative cofactor usage	916	1705	
	2-step reconstruction	387	408	
	3-step reconstruction	78	145	
ATLAS statistics	Total number of reactions	137 877	149 052	+8%
	Novel reactions	132 607	143 272	
	Total number of compounds	10 362	10 939	
	Number of orphan compounds integrated in ATLAS	3945 (42% of <i>okc</i>)	4587 (47% of <i>okc</i>)	
Consistency of EC numbers ^a	1st level EC match	79 058	138 168	+75%
	2nd level EC match	65 854	126 689	+92%
	3rd level EC match	47 918	94 168	+96%

^aNumber of matches between the EC assignment from the reaction rules and the EC numbers assigned by BridgIT for novel reactions in ATLAS.

metabolites available in KEGG to generate all possible biochemically consistent reactions between any two or more KEGG compounds. Two types of additional annotations were performed on the generated reactions: First, the new ATLAS reactions were curated with Gibbs free energy of reaction estimated with the Group Contribution Method (GCM).¹³ Second, the computational tool BridgIT was used to assign known enzymes to novel, predicted reactions¹⁴ by comparing the molecular structure of the participants in a novel, predicted reaction to a database of known, well-curated reactions with full gene-protein-reaction assignment. It calculates a similarity score between the novel and the known reactions, which makes it possible to find the enzyme with the highest probability of catalyzing the novel reaction.

Updated Tools and Methods. Since 2015, two main aspects of our workflow have been updated, which were applied to generate the updated version of ATLAS. First, the set of bidirectional reaction rules was increased from 360 to 400. Second, we applied the most recent version of BridgIT to predict putative enzymes for novel compounds, and we report the top three enzyme matches for each. The 40 new rules were created to reconstruct the exact reaction mechanism of an additional number of 510 KEGG reactions that were not considered previously (*i.e.*, KEGG reaction R03223) (Table S1).

Marvin was used for drawing, displaying, and characterizing chemical structures, substructures and reactions, Marvin 17.28.0, 2017, ChemAxon (<http://www.chemaxon.com>).

RESULTS AND DISCUSSION

ATLAS 2018, based on KEGG 2018, now has 149 052 reactions, out of which 5779 are known to KEGG. Compared to 2015, we added 385 known and 11 173 novel reactions (Table S2). Thanks to the predicted reactions, ATLAS now integrates 4587 out of 9857 disconnected, or “orphan”, KEGG metabolites, which were not participating in any known biochemical reaction.

Increased Coverage of KEGG Reactions. The KEGG database contained 18 254 compounds as of February 2018

(Table 1). In a first preprocessing step, we removed 999 compounds without clearly defined molecular structures (*e.g.*, polymers, proteins). The filtered data set comprised 17 255 compounds, out of which 9857 were not involved in any KEGG reaction. These orphan compounds did not participate in any known biotransformation in the KEGG metabolic space.

Out of the 10 829 reactions in KEGG, 76 involved compounds with an undefined structure that were removed, resulting in a filtered set of 10 753 reactions. Out of these, 8118 reactions were reconstructed with BNICE.ch reaction rules. We observed three different types of reaction reconstruction: 5779 reactions were exactly reconstructed, meaning that the reactions generated by BNICE.ch use the same cofactors as in KEGG. Another 1705 reactions were reconstructed using alternative cofactors, out of which 123 reactions were poorly characterized in KEGG (*i.e.*, reaction mechanism not known, incomplete reaction). The remaining 634 reactions were reconstructed in two (408 reactions), three (145 reactions), or four (81 reactions) consecutive reaction steps.

A total of 2635 KEGG reactions were not reconstructed with BNICE.ch (Table S3). First, 1546 reactions did not fulfill the BNICE.ch requirements for reconstruction, such as reactions involving polymer structures, generic compounds, or compounds without a defined molecular structure, as well as elementally unbalanced reactions and stereoisomerase reactions. Additionally, the reaction rules are organized according to the Enzyme Classification (EC) system, so each reconstructed or predicted reaction is automatically assigned a third-level EC number corresponding to the nonsubstrate specific EC classification of the reconstructing reaction rule. Another 308 reactions had partial or missing EC number annotations, indicating that the reaction mechanisms are not known and therefore no rule has been created for these reactions. The remaining 862 reactions were not reconstructed because their reaction mechanisms are very specific and hence not readily generalizable.

Predicted ATLAS Reactions Validated in KEGG and Other Databases. To validate the predicted reactions in

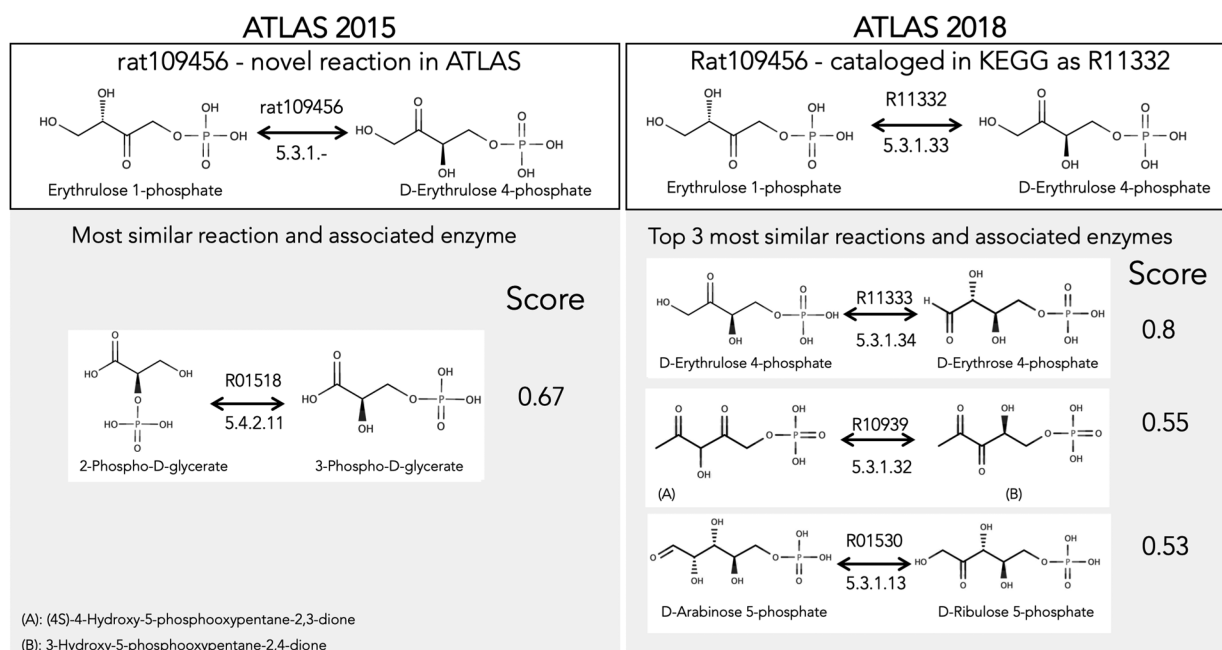


Figure 1. Reaction with ATLAS identifier rat109456 is an example of a reaction that was novel in ATLAS 2015 and that is now cataloged in KEGG. (left) In ATLAS 2015, the earlier version of BridgIT provided the most similar known reaction, and associated enzyme, for the ATLAS reaction with the ID. (right) In ATLAS 2018, the same reaction is now cataloged in KEGG as R11332 with EC 5.3.1.33. Other than the native enzyme with EC 5.3.1.33, BridgIT provides three alternative enzyme candidates that might also catalyze the reaction.

ATLAS, we analyzed the novel reactions predicted in 2015 that became known in KEGG 2018. Out of the 958 reactions newly added to KEGG, only 239 reactions involved compounds that were already present in KEGG 2015, meaning that they could have been predicted in the original ATLAS. Out of these 239 reactions, 107 were already present in ATLAS. In other words, the existence of hypothetical reactions in ATLAS 2015 was confirmed in KEGG 2018, demonstrating the predictive power of BNICE.ch.

Next, we examined the enzymes that BridgIT suggested in ATLAS 2015 for these 107 novel reactions, out of which 75 had an enzyme assigned. Interestingly, we found that the predicted EC numbers for 64 out of 75 reactions match the EC number proposed in KEGG up to the third level. For example, the novel reaction rat104204 was predicted to have an EC number of 2.4.1.-. BridgIT suggested R08946 as the most similar reaction, which was known to be catalyzed by 2.4.1.245. In 2018, KEGG confirmed the promiscuous activity of 2.4.1.245 for this reaction and named it R11306.

In ATLAS 2018, we additionally mapped the novel reactions to reaction databases other than KEGG. Interestingly, we found that 1118 predicted reactions in ATLAS were not actually novel, but known to at least one of the repositories Brenda, Reactome, HMR, MetaCyc, MetaNetX, BIGG, or Rhea, which shows that the predictive power of ATLAS goes beyond KEGG (Table S4). ATLAS reactions that can be found in any of these databases are linked accordingly in the updated version.

Improvements in the Prediction of Enzymes for ATLAS Reactions. To find putative enzymes for the reactions in ATLAS, we applied the enzyme prediction tool BridgIT. With the latest version of the tool, the new predictions were significantly better in the updated ATLAS: BridgIT correctly matched 92% of ATLAS reactions to the same EC class as BNICE.ch rules, whereas the previous version only matched

around 60% (Table 1). For each ATLAS reaction, we provide the top three candidate enzymes, and we also include BridgIT results for known KEGG reactions to provide alternative enzymes for a known reaction.

As a qualitative example of an improved prediction, we analyzed the ATLAS reaction rat109456, whose closest BridgIT candidate had a low matching score of 0.67. In ATLAS 2018, the reaction is now known and BridgIT found three very similar reactions, the first of which having a higher score than in the previous version (Figure 1).

CONCLUSION

We have updated the ATLAS of Biochemistry to integrate new biochemical data from KEGG 2018 using an updated set of generalized reaction rules and by employing an improved version of BridgIT to enhance the enzyme predictions for novel reactions. This study demonstrates the dynamic nature of biochemical knowledge and highlights the need for continuous updates of database-dependent applications. The updated ATLAS database contributes to fill the gaps in our current knowledge of metabolism by expanding the boundaries to novel predicted metabolic reactions. The updated ATLAS database is freely available online for academia upon request.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acssynbio.0c00052>.

Table S1: Comparison of reaction rules used in the ATLAS workflow in 2015 and 2018; Table S2: Reactions added to ATLAS in 2018, and the reason why they were added; Table S3: List of 2635 reactions that were not reconstructed in ATLAS 2018, and why they were not reconstructed; Table S4: Additional

identifiers for the 1118 reactions novel to KEGG, but known to other biochemical databases (XLSX)

AUTHOR INFORMATION

Corresponding Author

Vassily Hatzimanikatis – Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne, EPFL, CH-1015 Lausanne, Switzerland; orcid.org/0000-0001-6432-4694; Email: vassily.hatzimanikatis@epfl.ch

Authors

Jasmin Hafner – Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne, EPFL, CH-1015 Lausanne, Switzerland; orcid.org/0000-0002-1474-5869

Homa MohammadiPeyhani – Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne, EPFL, CH-1015 Lausanne, Switzerland

Anastasia Sveshnikova – Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne, EPFL, CH-1015 Lausanne, Switzerland

Alan Scheidegger – Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne, EPFL, CH-1015 Lausanne, Switzerland

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acssynbio.0c00052>

Author Contributions

[‡]J.H. and H.M. contributed equally to this work.

Funding

The research was funded by Ecole Polytechnique Fédérale de Lausanne (EPFL) (J.H., H.M., and V.H), MicroScapeX (J.H), PAcMEN (H.M), SNSF (A.S), ShikiFactory100 (A.S). MicroScapeX: Grant 2013/158 awarded by the Swiss National Science Foundation (SNSF). PacMen: European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 72228. SNSF: Swiss National Science Foundation, project number 200021_188623. ShikiFactory100: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 814408.

Notes

The authors declare no competing financial interest.

REFERENCES

- (1) Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A., and Hatzimanikatis, V. (2016) ATLAS of Biochemistry: A repository of all possible biochemical reactions for synthetic biology and etmabolic engineering studies. *ACS Synth. Biol.* 5, 1155–1166.
- (2) Lin, G.-M. M., Warden-Rothman, R., and Voigt, C. A. (2019) Retrosynthetic design of metabolic pathways to chemicals not found in nature. *Curr. Opin. Syst. Biol.* 14, 82–107.
- (3) Choi, K. R., et al. (2019) Systems metabolic engineering strategies: integrating systems and synthetic biology with metabolic engineering. *Trends Biotechnol.* 37, 817–837.
- (4) Lee, S. Y., et al. (2019) A comprehensive metabolic map for production of bio-based chemicals. *Nat. Catal.* 2, 18–33.
- (5) Yang, X., et al. (2019) Systematic design and in vitro validation of novel one-carbon assimilation pathways. *Metab. Eng.* 56, 142–153.
- (6) Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.
- (7) Hatzimanikatis, V., et al. (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics* 21, 1603–1609.

(8) Finley, S. D., Broadbelt, L. J., and Hatzimanikatis, V. (2009) Computational framework for predictive biodegradation. *Biotechnol. Bioeng.* 104, 1086–1097.

(9) Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2010) Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnol. Bioeng.* 106, 462–473.

(10) Soh, K. C., and Hatzimanikatis, V. (2010) DREAMS of metabolism. *Trends Biotechnol.* 28, 501–508.

(11) Hadadi, N., and Hatzimanikatis, V. (2015) Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Curr. Opin. Chem. Biol.* 28, 99–104.

(12) Tokić, M., et al. (2018) Discovery and evaluation of biosynthetic pathways for the production of five methyl ethyl ketone precursors. *ACS Synth. Biol.* 7, 1858–1873.

(13) Jankowski, M. D., Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2008) Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. *Biophys. J.* 95, 1487–1499.

(14) Hadadi, N., MohammadiPeyhani, H., Miskovic, L., Seijo, M., and Hatzimanikatis, V. (2019) Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. *Proc. Natl. Acad. Sci. U. S. A.* 116, 7298.