

MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models

Sébastien Moretti¹, Van Du T. Tran¹, Florence Mehl¹, Mark Ibberson and Marco Pagni^{1*}

Vital-IT group, SIB Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland

Received September 15, 2020; Revised October 09, 2020; Editorial Decision October 12, 2020; Accepted October 27, 2020

ABSTRACT

MetaNetX/MNXref is a reconciliation of metabolites and biochemical reactions providing cross-links between major public biochemistry and Genome-Scale Metabolic Network (GSMN) databases. The new release brings several improvements with respect to the quality of the reconciliation, with particular attention dedicated to preserving the intrinsic properties of GSMN models. The MetaNetX website (<https://www.metanetx.org/>) provides access to the full database and online services. A major improvement is for mapping of user-provided GSMNs to MNXref, which now provides diagnostic messages about model content. In addition to the website and flat files, the resource can now be accessed through a SPARQL endpoint (<https://rdf.metanetx.org/>).

INTRODUCTION

MetaNetX/MNXref provides computed cross-references between metabolites as well as biochemical reactions, to reconcile major public biochemical databases and a selection of public genome-scale metabolic networks (GSMN) from a few dedicated resources (1–12). The previous releases of the MetaNetX/MNXref resource and website have already been presented (13,14) as well as the main challenges of the reconciliation (15–17). This paper focuses on the current status of the resource and the recent improvements accomplished through the complete redesign and rewriting of its production pipeline.

GSMN reconstruction is one of the foundations of Systems Biology. The development of such reconstruction involves the integration of knowledge about reactions and metabolites from the scientific literature, public databases and previously published GSMNs. Historically, many GSMNs were formulated using metabolites represented as ‘symbols’, i.e. without explicit reference to a molecular structure. The initial motivation for creating this resource ten years ago was to add molecular structures to

existing GSMNs. More generally, the goal was to establish cross-links between symbols in GSMNs published by different groups and the molecules found in the major biochemical databases. This problem is trivial as long as one-to-one mappings can be established between the different resources, but this is not always the case. Merging metabolites in a metabolic network may lead to the merging of reactions, possibly altering the model properties, and hence the predictions that could be made using it.

MATERIALS AND METHODS

The computation of the reconciliation was performed using various different software components glued together by Bash, Perl 5.24.0 (<https://www.perl.org/>) and R 3.6.1 (<https://www.R-project.org>) scripts. These components include InChI 1.0.5 (18), Chemaxon MarvinBeans 19.26 (<https://chemaxon.com/>), Gurobi 9.0.2 (<https://www.gurobi.com/>), Blast 2.7.1+ (19), MATLAB (<https://ch.mathworks.com/products/matlab.html>), COBRA Toolbox 2.0.0 (20) and LibSBML 5.1.8.0 (21). The MetaNetX website backend is written in Perl with the help of Apache (<https://httpd.apache.org/>) and Virtuoso (<https://virtuoso.openlinksw.com/>). The frontend uses jQuery (<https://jquery.com/>) and DataTables (<https://datatables.net/>) with some of their plugins to ease browsing.

RESULTS

Principles

For MetaNetX/MNXref, a metabolite identifier (e.g. MNXM123) first refers to a *set* of metabolites in external databases, which are merged together because they are assumed to be the same biochemical entity, i.e. distinctions in protonation states, tautomeric forms or isotopes are typically ignored. Secondly, for every set of metabolites, a single external database identifier is selected as the (best) reference, from which a molecular structure is possibly retrieved. ChEBI identifiers that appear in Rhea reactions are preferentially selected as reference.

*To whom correspondence should be addressed. Tel: +41 21 692 40 73; Email: Marco.Pagni@sib.swiss

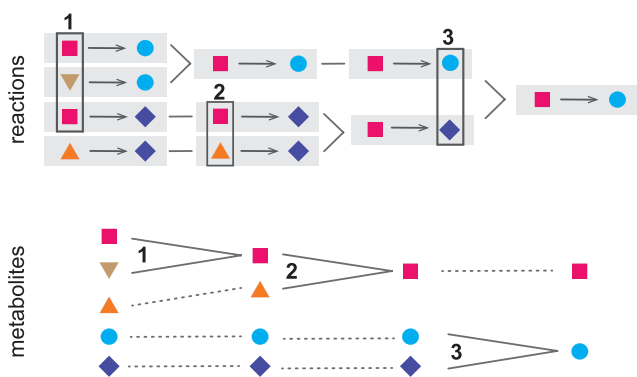


Figure 1. Example of the progressive reconciliation of five metabolites in four reactions into two metabolites in a single reaction. Three lines of evidence are used successively: (1) identical structures; (2) similar structures in reaction context; (3) structure and symbol pair in reaction context.

A consequence of the merging of metabolites is the merging of reactions involving these metabolites (Figure 1). A reaction identifier (e.g. **MNXR456**) designates a *set* of reactions in external databases that are assumed to be the same biochemical entity, because they were merged together. A single external identifier is selected as the reference for this set, preferably a reaction from Rhea when feasible. However, a reaction rewritten with MNXref identifiers may differ from the original equation, for example, with respect to proton balancing. In addition, since release 2.0, a distinction was introduced between transported protons that account for the proton motive force (**MNXM01**) and those introduced to balance chemical equations (**MNXM1**). Such a distinction is also made in some recently published models (22).

Reconciling metabolites and reactions

Molecular structures are systematically exploited to reconcile metabolites. The proportion of chemicals in external databases endowed with a molecular structure has steadily increased over the last 10 years. However, many difficulties remain, such as incomplete information on *cis/trans*- and stereo-isomerisms. The interpretation of partially defined structures, especially those involving R-groups and polymers, also remains a challenge.

The availability of cross-references between reactions often enables the reconciliation of pairs of metabolites if supported by additional evidence, for example the compound name. This reconciliation of metabolites from the *reaction context* evidence was the key idea that allowed the first computation of MetaNetX/MNXref (15). It is to be noted that imported cross-references between metabolites are ignored during the reconciliation, to avoid propagating existing errors present in external databases.

Manual curation

Although we aimed to make the reconciliation procedure fully automated, errors are occasionally imported from external resources and these can affect the computed reconciliation. Automated detection of conflicting evidence and

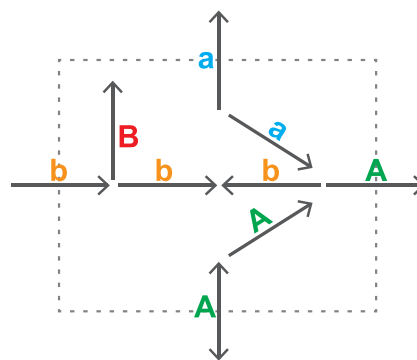


Figure 2. A status (A, a, b, B) was attributed to every reaction of this simple network, as explained in the text. A reaction status ultimately depends on the whole network topology and the directionality of every reaction.

ambiguities has been central to our effort to rewrite the pipeline. Thereafter, manual case-by-case study enables fixing of flagged issues, for example by correcting the reference of a metabolite to a molecular structure. An archetypal example of such difficulties are the two stereo-isomers of a compound—according to the textual description—where the structure of the L-form is fully specified and the structure of the R-form misses the stereochemistry.

Two metabolites to be merged that appear in the same reaction, in the left and right terms for example, also deserve scrutiny. Pairs of tautomers or acid-bases can be validly merged, possibly yielding an *empty* reaction that will disappear from the mapped GSMN. Changes in reaction status (see below) have also been used as a trigger for the manual inspection of a particular reconciliation.

The errors we have detected have been systematically reported to the resource from which they originated. Most of them have since been corrected, thanks to the diligence of their curators.

Impact of reconciliation on model properties

One of the main challenges of the MNXref reconciliation is to ease comparison of GSMNs while preserving their *intrinsic properties*. We have attempted to summarize the latter by attributing a *status* to every reaction in a GSMN (Figure 2): **A**, the reaction can carry a non-zero flux in the unaltered GSMN; **a**, the reaction can carry a flux after all boundary (exchange) reactions were set to bi-directional; **b**, the reaction can carry a flux after all reactions in the GSMN were set to bi-directional; **B**, the reaction cannot carry any flux because of the network topology, for example because of a dead-end metabolite. Computation is performed using a variant of the flux variability algorithm (23).

For example, the iAF1260 model for *E. coli* (24) contains a total of 2376 reactions. After mapping to MNXref, 2368 reactions are retrieved with a one-to-one mapping to an original reaction. The status of these reactions is distributed as follows: **A**, 1518; **a**, 627, **b**, 87; **B**, 136. The reaction status is perfectly conserved in this example, despite the merging of two metabolites namely **1agpg140** (1-acyl-sn-glycero-3-phosphoglycerol (n-C14:0)) and **1tdecg3p** (1-tetradecanoyl-sn-glycerol 3-phosphate). In other GSMNs, the merging of metabolites may cause the status of a few

Table 1. Sources and numbers of metabolites found in the MNXref 4.0 reconciliation. The number of metabolites found in the different resources and how non-redundant they are made through MNXref reconciliation. Secondary and deprecated identifiers are not counted here.

Resources	Prefixes ^a	Number of entries in		
		Total	Reactions	GSMNs
BiGG (1)	biggM	9130	9053	6532
ChEBI (2)	chebi	134607	35025	14274
enviPath (3)	envipath	12306	1587	654
HMDB (4)	hmdb	195008	9596	5159
KEGG (5)	keggC	18673	9901	3140
	keggD	11147	649	308
	keggE	864	0	0
	keggG	11042	405	109
LIPID MAPS (6)	lipidmaps	43085	2890	1030
MetaCYC (7)	metacycM	20296	16192	4064
REACTOME (8)	reactome	5526	2039	1719
RHEA (9)	rheaG^b	1494	1490	21
	rheaP^b	228	195	0
SABIO-RK (10)	sabiorkM	8944	8899	1713
TheSEED (11)	seedM	33995	21630	4964
SwissLipids (12)	slm	777657	1833	624
MetaNetX/ MNXref 4.0		1045319	41687	10533
	<i>ratio^c</i>	<i>1.23</i>	<i>2.91</i>	<i>4.3</i>
	<i>others^d</i>			<i>941</i>

^aThese prefixes are used to refer external resources in the TAB-delimited and RDF distributions of MNXref.

^bThese are the few metabolites in RHEA reactions that are not directly encoded by a ChEBI identifier.

^cAverage number of identifiers in external databases per MNXref entry.

^dMetabolites that are only found in models.

Table 2. Sources and numbers of reactions found in the MNXref 4.0 reconciliation. Secondary and deprecated identifiers are not counted here.

Resources	Prefixes ^a	Number of entries in	
		Total	GSMs
BiGG (1)	biggR	28167*	41863*
KEGG (5)	keggR	11161	3508
MetaCYC (7)	metacycR	17208	4898
RHEA (9)	rheaR^b	12510	3989
SABIO-RK (10)	sabiorkR	8118	2319
TheSEED (11)	seedR	43855*	19230*
MetaNetX/ MNXref 4.0		37103	11968
	<i>ratio^c</i>	<i>3.26</i>	<i>6.87</i>
	<i>others^d</i>		<i>6463*</i>

* Reactions with specific compartments (*vs* generic compartments).

^aThese prefixes are used to refer external resources in the TAB-delimited and RDF distributions of MNXref.

^bOnly master reactions of RHEA are counted here.

^cAverage number of identifiers in external databases per MNXref entry.

^dReactions only found in models, but most of them are formulated with metabolites in MNXref.

reactions to change, the latter not necessarily involving the merged metabolites and possibly augmenting the number of reactions that can carry a non-zero flux. Taking control of such indirect effects has been crucial for testing and improving the reconciliation. More public GSMNs will be added in the future to further challenge and improve the quality of the reconciliation.

Database content

Tables 1 and 2 list the different resources that were included in the latest release. The ratio of the numbers of entries before and after reconciliation is a measure of its efficiency. Indeed, this ratio is most optimal for the metabolites and reactions found in a hundred GSMNs hosted at MetaNetX.

Service

The MetaNetX website (<https://www.metanetx.org>) provides access to the full database and online services. Most notably, the service that permits a user to upload his or her own GSMN and map it to MNXref has been greatly improved to produce diagnostic messages about the model content. The database can be queried through the website user interface, as well as through a SPARQL endpoint (<https://rdf.metanetx.org>).

DISCUSSION

By design a GSMN is a model of the metabolism of low molecular weight compounds, where mass conservation and thermodynamics apply (25,26). GSMN models play a crucial role in the toolbox of synthetic biology (27,28). Beyond popular applications, such as metabolic flux prediction with flux balance analysis and prediction of the gene essentiality, GSMNs have been used in numerous applications (29), e.g. chemicals and materials production (30–33), drug targeting (34,35), human metabolism, disease understanding (36,37) and, multi-organism interaction modeling (38,39). MNXref is now used within tools for testing GSMNs such as MEMOTE (40), and has been proposed as a reference for minimal standard content for metabolic network reconstruction (41). Finally, the ability to accurately reconcile metabolites and reactions within GSMNs paves the way for applications such as multi-omics data interpretation (42,43), and white-box AI models (44). Moving forward, MetaNetX/MNXref efforts to not only unify metabolites, reactions and subcellular compartments but also genes and proteins (45) into a single comprehensive namespace will continue to provide a strong basis for such key systems biology endeavors in the future.

DATA AVAILABILITY

Data is distributed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). The MNXref reconciliation is available in TAB-delimited and RDF/Turtle formats. The mapped GSMNs are available in SBML level 2 and 3 formats from the web site and as part of the RDF graph.

ACKNOWLEDGEMENTS

Computation and maintenance of the MetaNetX.org server are provided by the Vital-IT (<https://www.vital-it.ch>) and Core-IT centers of the SIB Swiss Institute of Bioinformatics (<https://sib.swiss>). We thank Alan Bridge, Anne Morgat and the whole Rhea team for stimulating exchanges and discussions. We thank Jerven Bolleman for convincing us that RDF was an effective technology. We thank Manuela Pruess for help with the manuscript.

FUNDING

Swiss Federal Government through the Federal Office of Education and Science; Swiss Initiative for Systems Biology evaluated by the Swiss National Science Foundation. Funding for open access charge: Swiss Federal Government through the Federal Office of Education and Science. *Conflict of interest statement.* None declared.

REFERENCES

- Norsigian, C.J., Pusarla, N., McConn, J.L., Yurkovich, J.T., Dräger, A., Pálsson, B.O. and King, Z. (2020) BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res.*, **48**, D402–D406.
- Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P. and Steinbeck, C. (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–D1219.
- Wicker, J., Lorschbach, T., Gütlein, M., Schmid, E., Latino, D., Kramer, S. and Fenner, K. (2016) enviPath—the environmental contaminant biotransformation pathway resource. *Nucleic Acids Res.*, **44**, D502–D508.
- Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N. *et al.* (2018) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.*, **46**, D608–D617.
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
- Fahy, E., Subramaniam, S., Murphy, R.C., Nishijima, M., Raetz, C.R.H., Shimizu, T., Spener, F., van Meer, G., Wakelam, M.J.O. and Dennis, E.A. (2009) Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.*, **50**, S9–S14.
- Caspi, R., Billington, R., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P.E., Ong, Q., Ong, W.K. *et al.* (2018) The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **46**, D633–D639.
- Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
- Morgat, A., Lombardot, T., Axelsen, K.B., Aimo, L., Niknejad, A., Hyka-Nouspikel, N., Coudert, E., Pozzato, M., Pagni, M., Moretti, S. *et al.* (2017) Updates in Rhea – an expert curated resource of biochemical reactions. *Nucleic Acids Res.*, **45**, D415–D418.
- Wittig, U., Rey, M., Weidemann, A., Kania, R. and Müller, W. (2018) SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Res.*, **46**, D656–D660.
- Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Lindsay, B. and Stevens, R.L. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.*, **28**, 977–982.
- Aimo, L., Liechti, R., Hyka-Nouspikel, N., Niknejad, A., Gleizes, A., Götz, L., Kuznetsov, D., David, F.P.A., van der Goot, F.G., Riezman, H. *et al.* (2015) The SwissLipids knowledgebase for lipid biology. *Bioinformatics*, **31**, 2860–2866.
- Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A. and Pagni, M. (2016) MetaNetX/MNXref—reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.*, **44**, D523–D526.
- Ganter, M., Bernard, T., Moretti, S., Stelling, J. and Pagni, M. (2013) MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics*, **29**, 815–816.
- Bernard, T., Bridge, A., Morgat, A., Moretti, S., Xenarios, I. and Pagni, M. (2014) Reconciliation of metabolites and biochemical reactions for metabolic networks. *Brief. Bioinform.*, **15**, 123–135.
- Kumar, A., Suthers, P.F. and Maranas, C.D. (2012) MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics*, **13**, 6.
- Lang, M., Stelzer, M. and Schomburg, D. (2011) BKM-react, an integrated biochemical reaction database. *BMC Biochem.*, **12**, 42.
- Boucher, R., Heller, S. and McNaught, A. (2017) The status of the IUPAC InChI chemical structure standard. *Chem. Int.*, **39**, 47.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
- Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V. *et al.* (2019) Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.*, **14**, 639–702.
- Keating, S.M., Waltemath, D., König, M., Zhang, F., Dräger, A., Chaouiya, C., Bergmann, F.T., Finney, A., Gillespie, C.S., Helikar, T. *et al.* (2020) SBML Level 3: an extensible format for the exchange and reuse of biological models. *Mol. Syst. Biol.*, **16**, e9110.
- Smith, A.C., Eyassu, F., Mazat, J.-P. and Robinson, A.J. (2017) MitoCore: a curated constraint-based model for simulating human central metabolism. *BMC Syst. Biol.*, **11**, 114.
- Mahadevan, R. and Schilling, C.H. (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.*, **5**, 264–276.
- Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V. and Pálsson, B.Ø. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 121.
- Henry, C.S., Broadbelt, L.J. and Hatzimanikatis, V. (2007) Thermodynamics-based metabolic flux analysis. *Biophys. J.*, **92**, 1792–1805.
- Kiparissides, A. and Hatzimanikatis, V. (2017) Thermodynamics-based metabolic sensitivity analysis in metabolic networks. *Metab. Eng.*, **39**, 117–127.
- Cameron, D.E., Bashor, C.J. and Collins, J.J. (2014) A brief history of synthetic biology. *Nat. Rev. Microbiol.*, **12**, 381–390.
- Duigou, T., du Lac, M., Carbonell, P. and Faulon, J.-L. (2019) RetroRules: a database of reaction rules for engineering biology. *Nucleic Acids Res.*, **47**, D1229–D1235.
- Gu, C., Kim, G.B., Kim, W.J., Kim, H.U. and Lee, S.Y. (2019) Current status and applications of genome-scale metabolic models. *Genome Biol.*, **20**, 121.
- Tibocha-Bonilla, J.D., Zuñiga, C., Godoy-Silva, R.D. and Zengler, K. (2018) Advances in metabolic modeling of oleaginous microalgae. *Biochemol. Biofuels*, **11**, 241.
- Yang, J.E., Park, S.J., Kim, W.J., Kim, H.J., Kim, B.J., Lee, H., Shin, J. and Lee, S.Y. (2018) One-step fermentative production of aromatic polyesters from glucose by metabolically engineered *Escherichia coli* strains. *Nat. Commun.*, **9**, 79.
- Mishra, P., Lee, N.-R., Lakshmanan, M., Kim, M., Kim, B.-G. and Lee, D.-Y. (2018) Genome-scale model-driven strain design for dicarboxylic acid production in *Yarrowia lipolytica*. *BMC Syst. Biol.*, **12**, 12.
- Prigent, S., Nielsen, J.C., Frisvad, J.C. and Nielsen, J. (2018) Reconstruction of 24 *Penicillium* genome-scale metabolic models shows diversity based on their secondary metabolism. *Biotechnol. Bioeng.*, **115**, 2604–2612.
- Abdel-Haleem, A.M., Hefzi, H., Mineta, K., Gao, X., Gojbori, T., Pálsson, B.O., Lewis, N.E. and Jamshidi, N. (2018) Functional interrogation of *Plasmodium* genus metabolism identifies species- and stage-specific differences in nutrient essentiality and drug targeting. *PLoS Comput. Biol.*, **14**, e1005895.
- Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M. and Nielsen, J. (2014) Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol. Syst. Biol.*, **10**, 721.
- Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K. *et al.* (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.*, **36**, 272–281.
- Dhillon, B.K., Smith, M., Baghela, A., Lee, A.H.Y. and Hancock, R.E.W. (2020) Systems biology approaches to understanding the human immune system. *Front. Immunol.*, **11**, 1683.
- Pacheco, A.R., Moel, M. and Segrè, D. (2019) Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nat. Commun.*, **10**, 103.

39. Kumar,M., Ji,B., Babaei,P., Das,P., Lappa,D., Ramakrishnan,G., Fox,T.E., Haque,R., Petri,W.A., Bäckhed,F. *et al.* (2018) Gut microbiota dysbiosis is associated with malnutrition and reduced plasma amino acid levels: lessons from genome-scale metabolic modeling. *Metab. Eng.*, **49**, 128–142.
40. Lieven,C., Beber,M.E., Olivier,B.G., Bergmann,F.T., Ataman,M., Babaei,P., Bartell,J.A., Blank,L.M., Chauhan,S., Correia,K. *et al.* (2020) MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.*, **38**, 272–276.
41. Carey,M.A., Dräger,A., Beber,M.E., Papin,J.A. and Yurkovich,J.T. (2020) Community standards to facilitate development and address challenges in metabolic modeling. *Mol. Syst. Biol.*, **16**, e9235.
42. Tran,V.D.T., Moretti,S., Coste,A.T., Amorim-Vaz,S., Sanglard,D. and Pagni,M. (2019) Condition-specific series of metabolic sub-networks and its application for gene set enrichment analysis. *Bioinformatics*, **35**, 2258–2266.
43. Sánchez,B.J., Li,F., Kerkhoven,E.J. and Nielsen,J. (2019) SLIMER: probing flexibility of lipid metabolism in yeast with an improved constraint-based modeling framework. *BMC Syst. Biol.*, **13**, 4.
44. Kavas,E.S., Yang,L., Monk,J.M., Heckmann,D. and Palsson,B.O. (2020) A biochemically-interpretable machine learning classifier for microbial GWAS. *Nat. Commun.*, **11**, 2580.
45. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.