Universiteit Leiden
The Netherlands

Omics data integration with genome-scale modelling of dopaminergic neuronal metabolism

Preciat Gonzalez, G.A.

Chapter 5

Atomic resolution of genome-scale metabolic models

Based on:

German Preciat, Emma Schymanski, Ronan M.T. Fleming†, Thomas Hankemeier†

fluxTrAM: Integration of tracer-based metabolomics data into atomically-resolved genome-scale metabolic networks; application to human neuronal metabolism

In preparation

Abstract

Genome-scale modelling is a widely used approach for describing the metabolism of biological systems, but its description at the atomic level allows for a broader range of biological, biomedical and biotechnological applications than stoichiometry alone. For the atomic resolution of genome-scale metabolic models, it is necessary to know all of the metabolite structures in a model, but with the molecular diversity, finding a metabolite structure consistent with a genome-scale model is a difficult task. This work describes a pipeline for creating a chemoinformatic database with standardised metabolite structures that are consistent with a genome-scale model and atom-mapped metabolic reactions. To find a metabolite structure that is consistent with a genome-scale model, an InChI-based comparison was performed using the InChIs from different sources. Each InChI was assigned a score based on the information contained in its layers, which was calculated by comparing features, such as stereochemistry, charge, similarity to other InChIs and the chemical formula. The InChI with the highest score was deemed to have the most consistent metabolite structure in comparison to a genome-scale model, and the metabolite structure was saved in a metabolite database. The atom-mapped reactions generated with the obtained metabolites were used to calculate the number of bonds broken and formed and the enthalpy change, which were then compared to the molecular mass of the substrates to test the consistency of the atom-mapping predictions. This method showed that more than 90% of the atom-mapped reactions of the models used to test the pipeline were atomically balanced, demonstrating a consistent representation of metabolite structures in a genome-scale metabolic model.

The pipeline runs in MATLAB using the COBRA Toolbox and external software, such as CXCALC, Open Babel, and the Reaction Decoder Tool.

Introduction

Systems biology is defined as a comprehensive approach to linking molecular changes with functional regulation in living organisms (Nielsen). The advancement of high-throughput technology accelerates systems biology toward the resolution of more biological questions concerning various aspects of disease and drug research (Zierer et al., Cisek et al., Das et al.). As a result, genome-scale metabolic network reconstructions have emerged as a useful tool in modern biology for studying the metabolic pathways of biological systems *in silico* and predicting the flux of metabolic reactions at the genome-scale level. However, a more detailed representation at the underlying level of atom-mappings opens the possibility for a broader range of biological, biomedical and biotechnological applications than with stoichiometry alone.

A metabolic reaction is a process that leads to the chemical transformation of one set of molecular entities to another. The reaction mechanism of such reactions can be represented by a set of atom-mappings, each of which connects an atom in a substrate metabolite to an atom of the same element in a product metabolite. To atom map metabolic reactions at the genome-scale, one needs the molecular structures of the metabolites in the metabolic network, the reaction stoichiometries and an atom-mapping algorithm. This methodology has been used in a variety of platforms, including Pathway Tools (Karp et al.), EC-BLAST (Rahman et al.) and the COBRA Toolbox (Heirendt et al.). However, obtaining metabolic structures is a difficult task, since metabolites in a metabolic network can have multiple names, and many resources do not include every possible name. Furthermore, a metabolite structure can be represented by different isomers, and a consistent representation of a metabolite is critical for modelling biological systems at the atomic level, i.e., atomically resolving the metabolic network with atomically balanced metabolic reactions. It has been demonstrated that inconsistent structure representations can result in a significant loss of the predictive capacity of computer models, affecting downstream computation (Young, Martin, Venkatapathy and others).

Metabolite structures are obtained in a variety of ways, including drawing them from the literature using chemoinformatic software or obtaining them from metabolic databases manually or with computer software using different identifiers, as suggested in by (Thiele, Swainston, Fleming et al.). The metabolite structures are represented in

a variety of chemoinformatic formats, including 1) Metabolite chemical tables (MDL MOL) that list all of the atoms in a molecule, as well as their coordinates and bonds (Dalby et al.); 2) The simplified molecular-input line-entry system (SMILES), which uses a string of ASCII characters to describe the structure of a molecule (Weininger); or 3) The International Chemical Identifier (InChI) developed by the IUPAC, which provides a standard representation for encoding molecular structures using multiple layers to describe a metabolite structure (Heller et al.; see Figure 1).

Different approaches are being used to differentiate isomers in order to obtain a consistent representation of a metabolite structure from a biological system. The CAS Registry Number is a unique numerical identifier that is independent of any chemical nomenclature system and is widely used to identify chemical substances with a maximum capacity of $1e^9$ identifiers. An InChI represents a chemical structure in a standardised manner, allowing the generation of a unique identifier for a chemical structure (Heller et al.), unlike other chemoinformatic formats (SMILES and chemical tables), which may use different annotations to represent the same molecule. Furthermore, chemical databases, such as the PubChem database (Kim et al.), may create standardisation algorithms, which aim to eliminate invalid/incomplete structures using structure verification and structure normalisation approaches (Hähnke et al.).

This work describes a COBRA Toolbox v3.0 (Heirendt et al.) pipeline called `generateChemicalDatabase`, which generates a chemoinformatic database of standardised metabolite structures and atom-mapped reactions on a genome-scale metabolic reconstruction. In order to identify the metabolite structure that most closely resembles the metabolite in the genome-scale reconstruction, identifiers from different sources are compared based on their InChI. The molecular structures and reaction stoichiometries from the genome-scale reconstruction are used to generate reaction chemical tables containing information about the chemical reactions (MDL RXN). The metabolic reactions are atom-mapped using the Reaction Decoder Tool (RDT) algorithm (Rahman et al.), which was chosen after comparing the performance of published atom-mapping algorithms (Preciat et al.). Finally, the obtained atom-mapped reactions are used to identify the number of broken and formed bonds and the enthalpy change of the reactions in the genome-scale reconstruction.

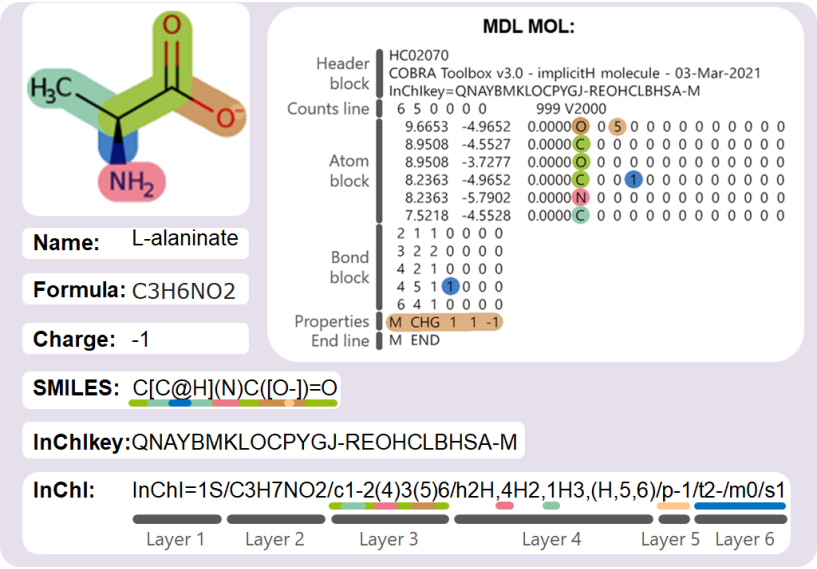


Figure 1: **Comparison of metabolite structure formats.** An L-alaninate molecule represented by a hydrogen-suppressed molecular graph (implicit hydrogens). The main branch of the molecule can be seen in green; the additional branches can be seen in brown, pink and turquoise. The stereochemistry of the molecule is highlighted in blue, the double bond in dark green and the charges in light brown. The same colours are used to indicate where this information is represented in the different metabolite strucure formats. The InChI is divided into layers, each of which begins with a lowercase letter, except for Layers 1 and 2. Layer 1 indicates if the InChI is standardised. Layer 2 shows the chemical formula in a neutral state. Layer 3 indicates the connectivity between the atoms (ignoring hydrogen atoms). Layer 4 demonstrates the connectivity of the hydrogen atom. Layer 5 indicates the charge of the molecule. Layer 6 shows the stereochemistry. Additional layers can be added, but they cannot be represented with a standard InChI.

The metabolite structures and atom-mapped reactions obtained from the `generateChemicalDatabase` pipeline can be found in `https://github.com/opencobra/ctf`.

Materials and Methods

The `generateChemicalDatabase` pipeline is run in MATLAB with the COBRA Toolbox v3.0 (Heirendt et al.), as well as various external software tools, for which installation is optional but recommen-

ded for the best results. The external software tools include Open Babel (O'Boyle et al.) and CXCALC (ChemAxon) for chemoinformatic data processing and the RDT (Rahman et al.) to atom map metabolic reactions (Java SE Development Kit is required). The pipeline generates the standardised database of metabolic structures in different chemoinformatic formats, including MDL MOL (Dalby et al.), SMILES (Weininger), InChI and the InChIkey for metabolic structures (Heller et al.). Atom-mapped metabolic reactions are represented in MDL RXN (Dalby et al.) and reaction SMILES, and unmapped reactions are represented with the International Chemical Identifiers for Reactions (RInChI; Grethe et al.). Additionally, if selected, the pipeline writes a diary file with all the outputs of the function for debugging. An overview of the methodology is given in Figure 2 and further described below.

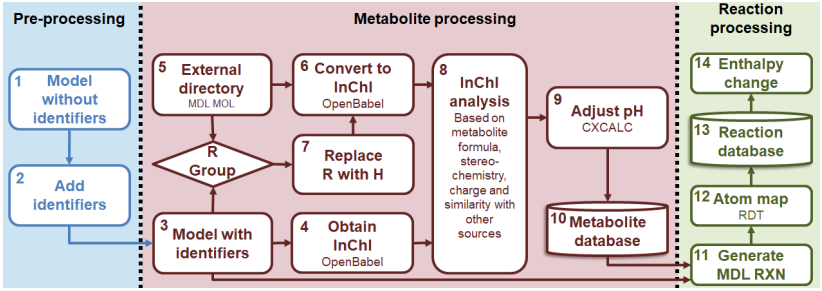


Figure 2: `generateChemicalDatabase` **pipeline.**

If a genome-scale model lacks database identifiers and chemoinformatic structures or more identifiers are required (1), the `addMetInfoInCBmodel` function reads an external file containing new identifiers and incorporates them into the model (2). Each identifier in the model (3) is used to generate an InChI with the tools `obtainMetStructures` and `openBabelConverter` (4). In parallel, metabolite structures stored in local directories can be integrated into the pipeline (5) and converted to InChI using `openBabelConverter` (6). For InChI analysis, any R group in a metabolite structure is replaced by a hydrogen atom to generate an InChI (7 and 6). All the InChI strings obtained are analysed with the tool `compareInchis` (8). The pH of the highest-scoring metabolite structures is adjusted using CXCALC to match the pH at which the metabolite is represented in the model (9), resulting in the standardised metabolite database (10). The model stoichiometries and metabolite database are used to write the reactions in MDL RXN (11), which are then atom-mapped using the function `obtainAtomMappingsRDT` (12) to generate the atom-mapped reaction database (13). Finally, the mapped reactions are used to identify enthalpy changes at the genome scale `findBEandBBF` (14).

Metabolite database

Metabolite structures are represented with database identifiers, e.g., Virtual Metabolic Human (VMH; Noronha et al.), PubChem (Kim et al.), the Kyoto Encyclopaedia of Genes and Genomes database (KEGG; Kanehisa et al.), Chemical Entities of Biological Interest (ChEBI; Hastings et al.) or the Human Metabolome Database (HMDB; Wishart et al.). In addition to database identifiers, the metabolite structures are also represented using several formats including metabolite chemical tables (MDL MOL) that list all of the atoms in a molecule, as well as their coordinates and bonds (Dalby et al.); the Simplified Molecular-Input Line-Entry System (SMILES), which uses a string of ASCII characters to describe the structure of a molecule (Weininger); or the International Chemical Identifier (InChI) developed by the IUPAC, which provides a standard representation for encoding molecular structures using multiple layers to describe a metabolite structure (Heller et al.) as illustrated in Figure 1.

For a given metabolite, comparing different metabolic databases, one may obtain substantial structural diversity due to selection of different isomers in the different sources. The `generateChemicalDatabase` pipeline uses all of the aforementioned databases to obtain and ultimately select a single InChI for each metabolite. InChI was chosen due to its standard structure for encoding molecular information and it is a database independent representation (Heller et al.). `generateChemicalDatabase` analyses all the InChI strings obtained from each source database assigning a score based on the criteria shown in Table 1 to identify the metabolic structure that most closely resembles the metabolite described in the genome-scale reconstruction. The InChI for each metabolite was analysed considering the chemical formula excluding hydrogen atoms, similarity with other databases, stereochemistry and charge in order to avoid loss of the predictive capacity of computer models, due to propagation of an inconsistent structure (Young, Martin, Venkatapathy and others).

Table 1: **InChI based comparison.**
Each InChI layer contains information that can be used to identify the metabolite structure that is most similar to the metabolite in the genome-scale reconstruction.

Concept	Score	Description
Chemical formula	0 or 10	The chemical formula indicated in the genome-scale model is compared with that obtained from the InChI. This feature is given more weight in order to keep the metabolite in the genome-scale model, where all reactions are stoichiometrically consistent and mass balanced. Hydrogen atoms were ignored in this comparison since they can be modified based on the charge (Figure 1).
Charge	0 or 1	The charge indicated in the genome-scale model is compared with the charge obtained from the source.
Stereochemical information	0 or 1	Indicates whether the InChI contains stereochemical information or not.
Standard	0 or 1	Indicates whether the InChI is standardised or not.
Similarity with other databases	0 to 1	The number of sources where the InChI strings are identical, divided by the total number of sources.
Main layer similarity	0 to 1	The number of sources where the main layers are identical, divided by the total number of sources.
InChI with more layers	0 or 1	InChI with more layers.

Letters in formulas that do not correspond to a chemical element, typically described with an R, represent any group of atoms attached to the rest of the molecule by a carbon or hydrogen atom. To compare metabolite structures with R groups, these atoms are modified by replacing each non-chemical atom with hydrogen only for the InChI comparison (Figure 3.1). If the number of hydrogen atoms in the metabolite structure obtained from a database and the metabolite obtained from the genome-scale model was different, software (CXCALC; ChemAxon) was used used to adjust the charge and number of hydrogen atoms of a metabolite based on the pH (Figure 3.2). To ensure database consistency, the molecular structures of the metabolites in the database

were represented as either molecular graphs or hydrogen-suppressed molecular graphs (Figure 3.3). Finally, the results are presented as molecular structures in five different formats, metabolite chemical tables (MDL MOL; Dalby et al.) , the Simplified Molecular-Input Line-Entry System (SMILES; Weininger) , the International Chemical Identifier (InChI; Heller et al.), the InChIkey (a fixed length of 27 characters condensed representation of the InChI; Heller et al.) and an image (JPEG file) containing a graphical representation of the metabolite structure (when CXCALC is installed).

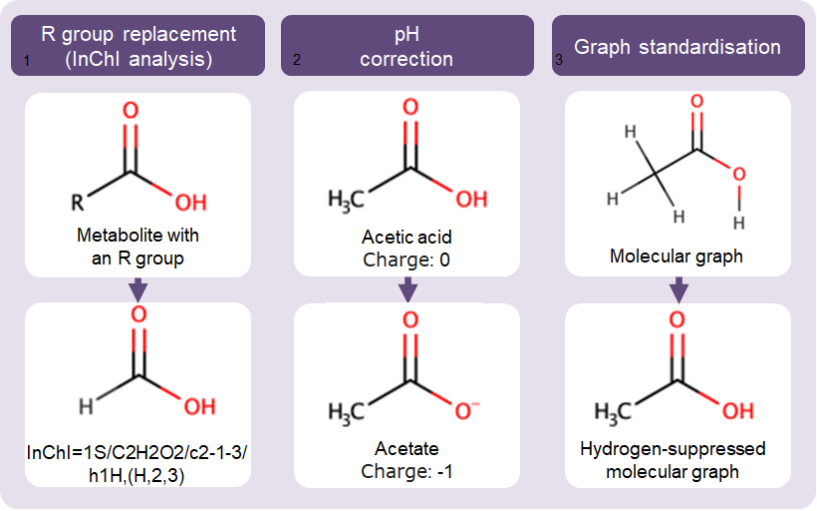


Figure 3: **Example molecular graphs.**
An example of the metabolite structure of acetic acid represented in different forms, as well as how it is processed by the pipeline. If a carbon attached to the carboxyl group is replaced by a non-chemical atom, the non-chemical is replaced by a hydrogen to form a molecule that can be converted into an InChI. (1). The pH of the acetic acid molecule is changed in order to obtain the charge described in the model (2). In order to achieve implicit hydrogen standardisation, a molecular graph representing a chemical compound's structural formula is converted to a hydrogen-suppressed molecular graph, which is a molecular graph with the hydrogen vertices removed (3).

Reaction database

The database of metabolite structures obtained and the reaction stoichiometries from the genome-scale reconstruction are used to represent

the metabolic reactions as MDL RXN files and RInChI. In order to generate a metabolic reaction in the database, all of the metabolites must be present in the metabolite database. If the metabolite database is represented with hydrogen-suppressed molecular graphs, the stoichiometries of the reactions are modified by deleting reacting protons to keep the reaction atomically balanced. Atom-mappings for mass-balanced reactions in the genome-scale reconstruction are performed by
the COBRA Toolbox v3.0 (Heirendt et al.) function `obtainAtomMappingsRDT`, where the inputs are a genome-scale reconstruction and the directory metabolite database with the standardised metabolite structures in MDL MOL format. The function prepares the reactions in MDL RXN format, which are then atom-mapped using the RDT algorithm, an open-source atom-mapping software tool that uses the molecule stereochemistry to predict the atom-mappings and returns the atom-mapping with the minimum number of modified bonds based on four different approaches (Rahman et al.). Post-processing of the predicted atom-mapped reactions is performed to maintain the InChI-based standardised format of the metabolite database. Atom-mapped reactions are represented in MDL RXN files, SMILES and PNG files (see Figure 4).

Bond enthalpies and bonds broken and formed

Chemical bonds are the forces of attraction that bind atoms together. In a chemical reaction, the energy is changed from one form to another; for example, chemical energy may be changed to thermal energy. In such, reactions no energy is lost; it is simply converted from one form to another. For bonds to be broken, energy must be added to the system; conversely, the formation of a bond releases energy. Bond energy values can be used to estimate the heat that passes into or out of the system in a reaction, which is known as enthalpy change (ΔH). The reactions are classified from the overall enthalpy change as 1) Exothermic, where the ΔH is negative, since the enthalpy of the product is smaller than the enthalpy of the reactants; therefore, an exothermic reaction will release heat, meaning there is a heat loss from the system to the surroundings; and 2) Endothermic reactions that absorb heat, since there is a heat gain from the surroundings and the enthalpy of the products is greater than the enthalpy of the reactants, making the ΔH positive.

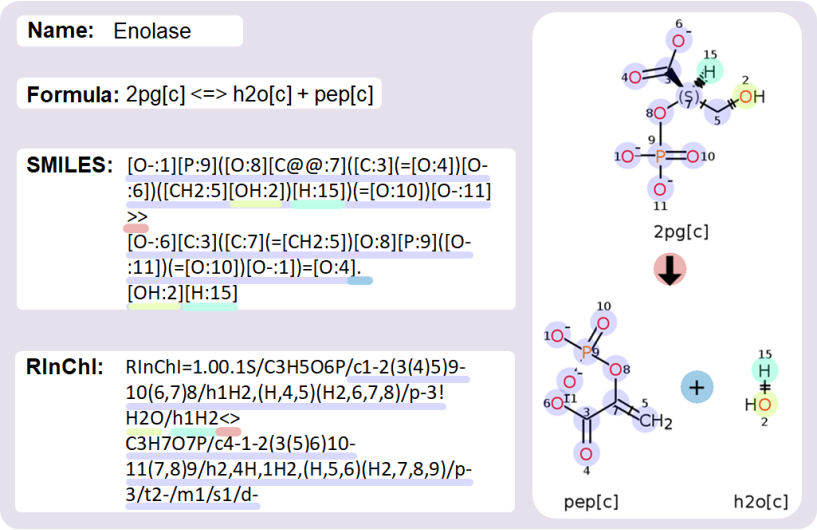


Figure 4: **Chemoinformatic formats for chemical reactions.** Atom-mapped representation of enolase (ENO) reaction in different chemoinformatic formats.

Using bond energy values, the enthalpy of a reaction can be written as Equation 1:

$$\Delta H = \sum n \times bonds\ broken - \sum n \times bonds\ formed \qquad (1)$$

where n denotes the number of moles of a specific type of bond (Zumdahl et al.).

Using atom-mappings to illustrate the reaction mechanism, an additional tool was included in the `generateChemicalDatabase` pipeline to identify the number of bonds that were broken and formed in a metabolic reaction by looking for the reaction centres to estimate the enthalpy change. The information about the bonds was obtained from the atom-mapped MDL RXN file, as the bond block from the chemical tables describes each of the bonds in a very detailed manner (see Figure 1). The information for each of the bonds are annotated in the matrices $R \in \mathbb{R}^{a \times a}$ and $P \in \mathbb{R}^{a \times a}$, in which the rows and columns represent the atoms in the reactants and the products respectively, and a is the total number of atoms in the reactants and products for both matrices, as they represent a one-to-one correspondence of atoms. If an atom in the

ith position is bonded to the atom in the *jth* position, the bond energy is represented in the position $R_{i,j}$ or $P_{i,j}$, an the Equation 1 can be rewritten as:

$$\Delta H = \frac{\sum_{i=1}^{a} \sum_{j=1}^{a} R_{i,j} - P_{i,j}}{2} \tag{2}$$

where the upper and lower triangular $R \in \mathbb{R}^{a \times a}$ represents the energy required to break the bonds of the substrates, and $P \in \mathbb{R}^{a \times a}$ denotes the energy released when forming the bonds of the products. Each of the bond energies used in the pipeline is obtained using the average bond enthalpies presented in Zumdahl's *General Chemistry, 10th edition* (Zumdahl et al.), assuming constant pressure. The elements on matrices R and P can also represent the number of bonds, having the total number of bonds broken and formed with Equation 2.

Results

Metabolite sources comparison.

A chemoinformatic database of metabolite structures and atom mapped-reactions from the genome-scale model `iDopaNeuroC` (Preciat, L. Moreno, Wegrzyn and others) was generated to test the `generateChemicalDatabase` pipeline functionality. The `iDopaNeuroC` model represents a cell culture of dopaminergic neurons derived from neuroepithelial stem cells differentiated using the Reinhardt protocol (Reinhardt et al., Lucumi Moreno et al.) using transcriptomic, metabolomic and bibliomic data. Figure 5 shows a comparison of the sources with identifiers where the metabolite structures in the `iDopaNeuroC` were collected. The Spearman correlation was used to calculate the similarity of the scores obtained in the InChI based on Table 1. The VMH database (Noronha et al.) contains the molecular structures that most closely resemble the metabolites described in the `iDopaNeuroC` model, providing the largest number of atomically balanced reactions. The more specific an InChI is, the more likely it is to obtain the metabolite structure for a genome-scale metabolic network, as evidenced by the InChIs from the VMH database. It contains the greatest number of InChIs with layers representing the stereochemistry and the charge of the metabolites. The metabolite structures in the `iDopaNeuroC` model represented with an InChI or SMILES did not include stereochemical information that showed a low Spearman correlation in comparison with the other sources. Not all of

the metabolites from the HMDB database were collected, resulting in a low collection of metabolites. This was due to connection problems with the database, as shown by the The HyperText Transfer Protocol (HTTP) status 503, which indicates that the server is not ready to handle the request.

Metabolite and reaction database coverage.

The `generateChe- micalDatabase` pipeline obtained 422/438 metabolite structures from the `iDopaNeuroC` model, covering 96% of them. In atomically balanced reactions, 364 metabolites were always present. In atomically unbalanced reactions, three metabolites were always present and 47 metabolites were occasionally present. Eight molecular structures were not used, because at least one molecular structure was required to complete the reaction in which they were involved. For the reaction database, 920/925 internal metabolic reactions were written, covering 99% of the reactions. There were 898 atomically balanced and 22 unbalanced. Five metabolic reactions could not be written, because at least one molecular structure was missing from the metabolite database. In addition to the `iDopaNeuroC` model, the `generateChemical-Database` pipeline was used to generate chemoinformatic databases from different genome-scale models, such as EcoliCore (Orth, Palsson and Fleming), AGORA2 (Heinken et al.) and Recon 3D (Brunk et al.), from which coverage is shown in Table 2. A list of the problematic reactions, along with an explanation of why each reaction is problematic, is found in the supplementary information.

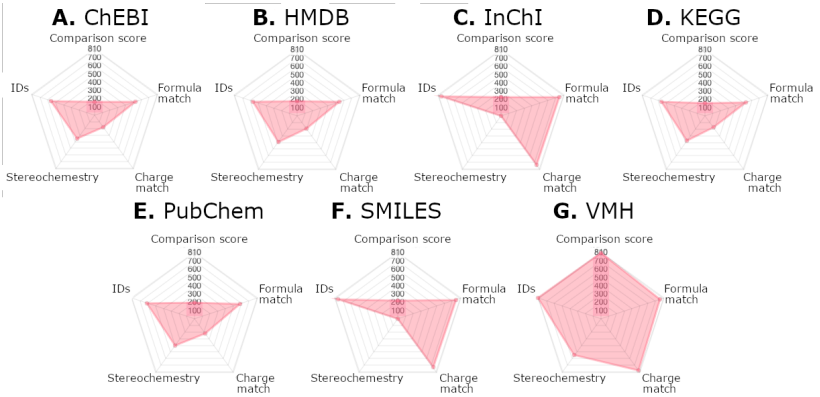


Figure 5: **Comparison of metabolite structure identifiers in the `iDopaNeuroC` model.**
Each of the metabolite identifiers evaluated in this work was compared using various criteria, including 1) IDs: the number of identifiers compared; 2) Comparison score: the number of times they received the best score; 3) Formula match: the number of times the chemical formula matched the metabolite formula in the `iDopaNeuroC` model (excluding the number of hydrogen atoms); 4) Charge match: the charge consistency with the model; and 5) Stereochemistry: the number of times stereochemical information was included.

Table 2: **Chemoinformatic data coverage for different genome-scale models.**

	`iDopaNeuroC`	EcoliCore	AGORA2	Recon3D
Metabolites in the database	810	53	1,894	3,532
Metabolites in balanced reactions	677	53	1,510	2,640
Metabolites occasionally in unbalanced reactions	116	0	260	640
Metabolites always in unbalanced reactions	15	0	66	192
Metabolites not used	2	0	58	60
Missing metabolites	8	1	1,697	668
Reactions in the database	2254	71	4,803	10,276
Atom-mapped reactions	2254	71	4,623	10,097
Balanced reactions	2155	71	4,583	9,512
Unbalanced reactions	99	0	220	764
Missing reactions	18	3	1,697	668

Bond enthalpies

The atom-mapped reactions of the `iDopaNeuroC` model were used to calculate the number of bonds broken and formed and the enthalpy change. However, not all of the atom-mapped reactions in the database were taken into account, as 32 were excluded: 25 were unbalanced, 5 were missing, and 2 were inconsistent because the atom-mapping algorithm changed their molecular structure. The number of bonds broken and formed, as well as the amount of energy consumed or released, correlates to the molecular weight of the substrates, as shown in Figure 6.

Discussion

Due to the different molecular structure representations in different sources, an effective comparison is required for an accurate chemoinformatic representation of a genome-scale reconstruction. The `generateChemicalDatabase` pipeline was presented in this work, which uses the tools in the COBRA Toolbox v3.0 (Heirendt et al.) to generate a standardised database of metabolites and atom-mapped reactions

consistent with a genome-scale reconstruction. Different metabolite structures were compared using InChI analysis to determine which metabolite should be represented in the genome-scale reconstruction.

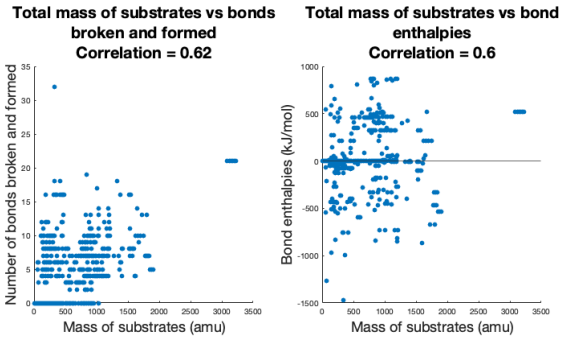


Figure 6: **Bond enthalpies and bonds broken.**
The total mass of substrates vs. bonds broken and formed, as well as the change in enthalpies and their Spearman correlation.

For the comparison of molecular structures, the `generateChemicalDatabase` pipeline assigns a score to each molecular structure to determine the best match for a genome-scale reconstruction. A diverse set of identifiers increases the chances of obtaining the metabolites that best represent those described in the reconstruction. As the specificity of the molecular structure increases, so does the number of atomically balanced reactions in the database. Figure 1 shows a case in which an InChI represents various metabolites depending on the level of detail included in it. The main layer, which, in this case, consists of layers 1, 2, 3 and 4, represents the alanine molecule CHEBI:16449. With the addition of layer 5, a negative charge is added on the molecule representing, the alaninate molecule CHEBI:32439. Finally, layer 6 adds the stereochemistry in the molecule, representing the L-alaninate molecule CHEBI:32431. Assigning the correct identifier for a metabolite in a genome-scale model is a difficult task due to the various isomers that a molecule may have, and more work should be done in this area to avoid mismatches. This is the case of the metabolite nicotinamide adenine dinucleotide, which is represented in the `iDopaNeuroC` model (Preciat, L. Moreno, Wegrzyn and others) with the identifier CHEBI:15846 from the ChEBI database, whereas the consistent iden-

tifier should be CHEBI:57540. Tools for mapping identifiers between different biological databases, such as BridgeDB (van Iersel et al.) or the chemical translation service (Wohlgemuth et al.), can be useful for assigning metabolite identifiers in accordance with the genome-scale model.

The InChI-based comparison and the adjustment of pH used in this pipeline increase the coverage of atomically balanced reactions due to the consistency of metabolite structures with the genome-scale reconstruction. More than 90% of the obtained reactions in Table 2 were atomically balanced. Nevertheless, not all the reactions in the `iDopaNeuroC` model could be atom-mapped for several reasons, including atomically unbalanced reactions, missing metabolites or reactions with more than 1,000 atoms in the substrates and/or the products. Furthermore, despite demonstrating higher accuracy in comparison with other atom-mapping algorithms (Preciat et al.), 112/432 passive transport reactions were atom-mapped incorrectly with the RDT algorithm, as shown in Figure 7, therefore, these reactions were mapped with the `transportRxnAM` function to ensure correct predictions.

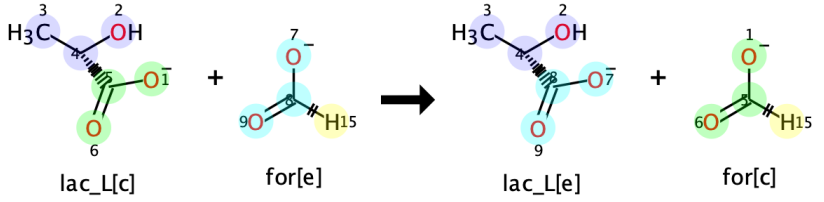


Figure 7: **Inconsistent transport reactions.**
Transport of lactate (VMH ID: r2117) atom-mapped incorrectly by the RDT algorithm.

An atomically resolved genome-scale metabolic model may have multiple applications, one of which was explored in this work. The identification of bonds broken and formed and the enthalpy change describe the mechanism of the metabolic reaction in greater detail. To break a bond, energy is needed, and a biological system tends to optimise energy use. For example, the most accurate atom-mapping algorithm tested in (Preciat et al.), aims to minimise the number of modified bonds. Considering this, the number of bonds broken and formed and the enthalpy change can be used to optimise fluxes in a metabolic network by minimising the modified bonds or the enthalpy

change. The gain or loss of energy depends on the reaction directionality (e.g., the enolase reaction is endothermic in glycolysis, with a ΔH of 54 KJ/mol, whereas the reverse reaction in glycogenesis is exothermic with a ΔH of -54 KJ/mol). Analytical chemistry techniques, such as calorimetry, allow detection of the thermal energy change in a biological system by predicting the flux in a genome-scale model, and with the total enthalpy change, a calorimeter can be used to validate genomic model predictions in different conditions. Another application of an atomically resolved metabolic network is tracing labelled atoms in the network, thus identifying the metabolic pools that form reactions using the structural properties of a biochemical network (Haraldsdóttir and Fleming, Ghaderi et al.) to optimise for tracer-based design or calculate the potential energy in molecules (Wang, Upadhyay and Maranas).

Author contributions

German Preciat: Conceptualisation; Pipeline; Writing.
Emma Schymanski: Supervision
Ines Thiele: Supervision
Ronan M.T. Fleming: Supervision
Thomas Hankemeier: Supervision

German Preciat, Emma Schymanski, Ronan M.T. Fleming[†], Thomas Hankemeier[†]

References

Brunk, E., Sahoo, S., Zielinski, D. C. et al.: 2018, Recon3d enables a three-dimensional view of gene variation in human metabolism, *Nature Biotechnology* **36**, 272.

ChemAxon: 2015, *Standardizer, Was Used for Structure Canonicalization and Transformation, JChem 16.1.11.0*.

Cisek, K., Krochmal, M., Klein, J. and others: 2016, The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease, *Nephrology Dialysis Transplantation* **31**(12), 2003–2011.

Dalby, A., Nourse, J. G., Hounshell, W. D. et al.: 2002, Description of several chemical structure file formats used by computer programs developed at molecular design limited.

Das, T., Andrieux, G., Ahmed, M. and others: 2020, Integration of Online Omics-Data Resources for Cancer Research., *Frontiers in genetics* **11**.

Ghaderi, S., Haraldsdóttir, H. S., Ahookhosh, M. and others: 2019, Structural conserved moiety splitting of a stoichiometric matrix, *Journal of Theoretical Biology* **499**.

Grethe, G., Blanke, G., Kraut, H. and others: 2018, International chemical identifier for reactions (rinchi), *Journal of Cheminformatics* **10**(1), 22.

Hähnke, V. D., Kim, S. and Bolton, E. E.: 2018, Pubchem chemical structure standardization, *Journal of Cheminformatics* **10**, 36.

Haraldsdóttir, H. S. and Fleming, R. M. T.: 2016, Identification of conserved moieties in metabolic networks by graph theoretical analysis of atom transition networks, *PLOS Computational Biology* **12**(11), e1004999.

Hastings, J., de Matos, P., Dekker, A. et al.: 2013, Chebi: A database and ontology for chemical entities of biological interest, *Nucleic Acids Research* **41**(D1), D456–D463.

Heinken, A., Acharya, G., Ravcheev, D. and others: 2021, AGORA2: Large scale reconstruction of the microbiome highlights wide-spread drug-metabolising capacities | bioRxiv, *bioRxiv* .

Heirendt, L., Arreckx, S., Pfau, T. et al.: 2019, Creation and analysis of biochemical constraint-based models using the COBRA toolbox v.3.0, *Nature Protocols* **14**(3), 639.

Heller, S. R., McNaught, A., Pletnev, I. and others: 2015, Inchi, the iupac international chemical identifier, *Journal of Cheminformatics* **7**(1), 23.

Kanehisa, M., Sato, Y., Kawashima, M. et al.: 2016, KEGG as a reference resource for gene and protein annotation, *Nucleic Acids Research* **44**(D1), D457–D462.

Karp, P. D., Latendresse, M., Paley, S. M. et al.: 2016, Pathway tools version 19.0 update: Software for pathway/genome informatics and systems biology, *Briefings in Bioinformatics* **17**(5), 877–890.

Kim, S., Thiessen, P. A., Bolton, E. E. et al.: 2016, Pubchem substance and compound databases, *Nucleic Acids Research* **44**(D1), D1202–1213.

Lucumi Moreno, E., Hachi, S., Hemmer, K. and others: 2013, Differentiation of neuroepithelial stem cells into functional dopaminergic neurons in 3D microfluidic cell culture, *Lab on a Chip* **5**.

Nielsen, J.: 2017, Systems biology of metabolism, *Annual review of biochemistry* **86**(6), 245–275.

Noronha, A., Modamio, J., Jarosz, Y. and others: 2019, The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease, *Nucleic Acids Research* **47**.

O'Boyle, N. M., Banck, M., James, C. A. et al.: 2011, Open babel: An open chemical toolbox, *Journal of Cheminformatics* **3**(1), 33.

Orth, J. D., Palsson, B. Ø. and Fleming, R. M. T.: 2010, Reconstruction and use of microbial metabolic networks: the core escherichia coli metabolic model as an educational guide, *EcoSal Plus* **1**(10).

Preciat, G., El Assal, L. R. P., Noronha, A. and others: 2017, Comparative evaluation of atom mapping algorithms for balanced metabolic reactions: application to recon3d, *Journal of Cheminformatics* **9**.

Preciat, G., L. Moreno, E., Wegrzyn, A. and others: 2021, Mechanistic model-driven exometabolomic characterisation of human dopaminergic neuronal metabolism, *In preparation* .

Rahman, S. A., Cuesta, S. M., Furnham, N. et al.: 2014, Ec-blast: A tool to automatically search and compare enzyme reactions, *Nature Methods* **11**(2), 171–174.

Rahman, S. A., Torrance, G., Baldacci, L. et al.: 2016, Reaction decoder tool (rdt): Extracting features from chemical reactions, *Bioinformatics* **32**(13), 2065–2066.

Reinhardt, P., Glatza, M., Hemmer, K. and others: 2013, Derivation and expansion using only small molecules of human neural progenitors for neurodegenerative disease modeling., *PloS one* **8**.

Thiele, I., Swainston, N., Fleming, R. M. T. et al.: 2013, A community-driven global reconstruction of human metabolism, *Nature Biotechnology* **31**(5), 419–425.

van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J. and others: 2010, The bridgedb framework: standardized access to gene, protein and metabolite identifier mapping services, *BMC Bioinformatics* **11**(1), 5.

Wang, L., Upadhyay, V. and Maranas, C. D.: 2021, dgpredictor: Automated fragmentation method for metabolic reaction free energy prediction and de novo pathway design, *bioRxiv* .

Weininger, D.: 1988, Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules, *Journal of Chemical Information and Computer Sciences* **28**(1).

Wishart, D. S., Feunang, Y. D., Marcu, A. and others: 2018, Hmdb 4.0: the human metabolome database for 2018, *Nucleic Acids Research* **46**.

Wohlgemuth, G., Haldiya, P. K., Willighagen, E., Kind, T. and others: 2010, The chemical translation service –a web-based tool to improve standardization of metabolomic reports, *Bioinformatics (Oxford, England)* **26**(20), 2647–2648.

Young, D., Martin, T., Venkatapathy, R. and others: 2008, Are the chemical structures in your qsar correct?, *QSAR & Combinatorial Science* **27**, 1337–1345.

Zierer, J., Menni, C., Kastenmüller, G. and others: 2015, Integration of 'omics' data in aging research: from biomarkers to systems biology., *Aging cell* **14**(6), 933–944.

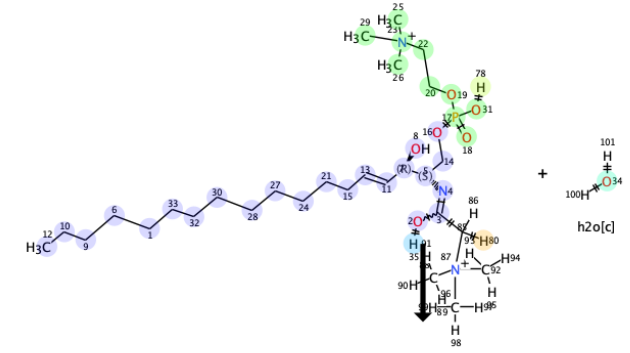
Zumdahl, S. S., Zumdahl, S. A. and DeCoste, D. J.: 2017, *Chemistry*, 10th edition edn, Cengage Learning.

Appendix

Supplementary Information 1 - Problematic metabolites and reactions

Inconsistencies caused by R groups

The R groups can cause inconsistencies because they represent a substructure that, because it is not defined, can result in unbalanced reactions, such as reaction sphingomyelin phosphodiesterase (VMH ID: HMR_0795).



Reaction: Glucosylceramidase
VMH id: GBA
Fromula: h2o[c] + gluside_hs[c] -> glc_D[c] + crm_hs[c]

Reaction: Glucosylceramidase (lysosome)
VMH id: GBAl
Fromula: h2o[l] + gluside_hs[l] -> crm_hs[l] + glc_D[l]

Reaction: S-Adenosyl-L-Methionine:Phosphatidylethanolamine N-Methyltransferase
VMH id: HMR_0653
Formula: amet[c] + pe_hs[c] -> 2 h[c] + ahcys[c] + M02686[c]

Reaction: Ceramide glucosyltransferase
VMH id: HMR_0761
Formula: crm_hs[c] + udpg[c] -> h[c] + udp[c] + gluside_hs[c]

Reaction: Sphingomyelin phosphodiesterase
VMH id: HMR_0795
Formula: h2o[c] + sphmyln_hs[c] -> h[c] + crm_hs[c] + cholp[c]

Reaction: Phosphatidic acid phosphatase
VMH id: PPAP
Formula: h2o[c] + pa_hs[c] -> pi[c] + dag_hs[c]

Reaction: Phosphatidylserine decarboxylase
VMH id: PSDm_hs
Formula: h[m] + ps_hs[m] -> co2[m] + pe_hs[m]

Reaction: Phosphatidylserine synthase
VMH id: PSSA1_hs
Formula: ser_L[c] + pchol_hs[c] <=> chol[c] + ps_hs[c]

Reaction: Phosphatidylserine synthase
VMH id: PSSA2_hs
Formula: ser_L[c] + pe_hs[c] <=> etha[c] + ps_hs[c]

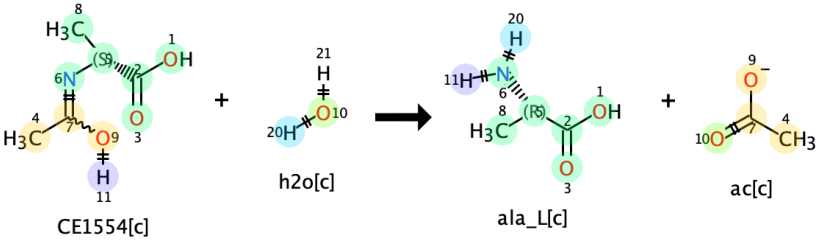
Reaction: -
VMH id: PSDm_hsc

Formula: h[c] + ps_hs[c] -> co2[c] + pe_hs[c]

Reaction: Sphingomyelin synthase
VMH id: SMS
Formula: pchol_hs[c] + crm_hs[c] -> dag_hs[c] + sphmyln_hs[c]

Inconsistencies caused by protons

Inconsistencies with protons are caused by metabolite pH, which changes the number of hydrogens, resulting in unbalanced reactions, such as reaction N-Acyl-Aliphatic-L-Amino Acid Amidohydrolase (VMH ID: RE2642C)



Reaction: Cardiolipin synthase
VMH id: CLS_hs
Fromula: cdpdag_hs[c] + pglyc_hs[c] -> h[c] + cmp[c] + clpn_hs[c]

Reaction: Glucosylceramidase
VMH id: GBA
Fromula: h2o[c] + gluside_hs[c] -> glc_D[c] + crm_hs[c]

Reaction: Glucosylceramidase (lysosome)
VMH id: GBAl
Fromula: h2o[l] + gluside_hs[l] -> crm_hs[l] + glc_D[l]

Reaction: S-adenosyl-L-methionine: phosphatidyl-N-dimethylethanolamine N-methyltransferase
VMH id: HMR_0657
Formula: amet[c] + M02758[c] -> ahcys[c] + pchol_hs[c]

Reaction: Ceramide glucosyltransferase
VMH id: HMR_0761

Formula: crm_hs[c] + udpg[c] -> h[c] + udp[c] + gluside_hs[c]

Reaction: Metabolism of LeuSerTrp (formation/degradation)
VMH id: LEUSERTRPr
Formula: 2 h2o[c] + leusertrp[c] <=> ser_L[c] + leu_L[c] + trp_L[c]

Reaction: Lysophospholipase
VMH id: LPASE
Fromula: h2o[c] + lpchol_hs[c] -> h[c] + Rtotal[c] + g3pc[c]

Reaction: Nicotinate D-ribonucleoside kinase
VMH id: NICRNS
Fromula: atp[c] + nicrns[c] -> h[c] + adp[c] + nicrnt[c]

Reaction: Nucleotide phosphatase
VMH id: NP1
Formula: h[c] + nac[c] + r1p[c] -> pi[c] + nicrns[c]

Reaction: Phospholipase A2
VMH id: PLA2_2
Formula: h2o[c] + pchol_hs[c] -> h[c] + Rtotal2[c] + lpchol_hs[c]

Reaction: Phospholipase A2 (extracellular)
VMH id: PLA2_2e
Formula: h2o[e] + pchol_hs[e] -> h[e] + lpchol_hs[e] + Rtotal2[e]

Reaction: -
VMH id: RE1266C
Formula: o2[c] + 4mop[c] -> h[c] + co2[c] + CE2028[c]

Reaction: Aminoacid N-acetyltransferase
VMH id: RE2031M
Formula: accoa[m] + ala_L[m] <=> h[m] + coa[m] + CE1554[m]

Reaction: N-acyl-aliphatic-L-amino acid amidohydrolase
VMH id: RE2642C
Formula: h2o[c] + CE1554[c] <=> ac[c] + ala_L[c]

Missing reactions

Finally, three of the missing reactions have stoichiometry that cannot be represented with integers, resulting in the failure to generate an MDL RXN reaction. Additionally, not all metabolites were present in the degradation (VMH id: HDL_HSDEG) and formation (VMH id: HDL_HSSYN) of HDL due to the limited information available about metabolites in the `iDopaNeuroC` model.

Reaction: Cytochrome C oxidase; mitochondrial complex IV
VMH id: CYOOm3
Formula: o2[m] + 7.92 h[m] + 4 focytC[m] -> 1.96 h2o[m] + 4 h[c] + 4 ficytC[m] + 0.02 o2s[m]

Reaction: Degradation of HDL
VMH id: HDL_HSDEG
Formula: h2o[e] + hdl_hs[e] -> 2 chsterol[e] + 2 pchol_hs[e] + Rtotal[e] + Rtotal2[e] + Rtotal3[e] + glyc[e] + HC00004[e] + HC00006[e] + HC00007[e] + HC00008[e] + HC00009[e]

Reaction: Formation of HDL
VMH id: HDL_HSSYN
Formula: 2 chsterol[e] + 2 pchol_hs[e] + tag_hs[e] + HC00004[e] + HC00006[e] + HC00007[e] + HC00008[e] + HC00009[e] -> hdl_hs[e]

Reaction: -
VMH id: HMR_0017
Formula: M02909[e] -> 0.125 octa[e] + 0.125 dca[e] + 0.125 but[e] + 0.125 C01601[e] + 0.125 M02108[e] + 0.125 M03117[e] + 0.125 M03134[e] + 0.125 caproic[e]

Reaction: Phosphate transport via Na+ symporter
VMH id: PIt8
Formula: 1.5 na1[e] + pi[e] <=> pi[c] + 1.5 na1[c]

Supplementary Information 2 - Chemoinformatic pipeline tutorial

Atomically resolve a metabolic reconstruction

Author: German Preciat, Analytical BioSciences, Leiden University

INTRODUCTION

Genome-scale metabolic network reconstructions have become a relevant tool in modern biology to study the metabolic pathways of biological systems *in silico*. However, a more detailed representation at the underlying level of atom mappings opens the possibility for a broader range of biological, biomedical and biotechnological applications than with stoichiometry alone.
This tutorial will demonstrate how to use the chimoinformatic tools in the COBRA Toolbox[1]. The chemoinformatic database will be generated using information from the ecoliCore model[2].

MATERIALS

Open Babel[3]

To convert molecular structures, open babel must be installed. To install it, follow the steps below.
On Windows, download the OpenBabel installation and follow the instructions.
On macOS, run the following command in the Terminal:

```
$ brew install open-babel
```

CXCALC[4]

The CXCALC tools are used to adjust the pH and create images of the metabolic structures. To install CXCALC download MarvinSuite and follow the instructions.

```
$ brew install open-babel
```

JAVA

To atom map reactions it is required to have Java version 8. Follow the instruction in https:// www.openlogic.com/openjdk-downloads.
On macOS, please make sure that you run the following commands in the Terminal before continuing with this tutorial:

```
$ /usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/
 install/master/install)"
$ brew install coreutils
```

PATH

On Linux, please make sure that Java, OpenBabel[3] and CXCALC[4] directories are included. To do this, run the following commands

```
$ export PATH=$PATH:/usr/local/bin (default location of OpenBabel)
$ brew install coreutils
```

PROCEDURE

Chemoinformatic database

The `generateChemicalDatabase` pipeline generates a chemoinformatic database of standardised metabolite structures and atom-mapped reactions on a genome-scale metabolic reconstruction (Figure 1).

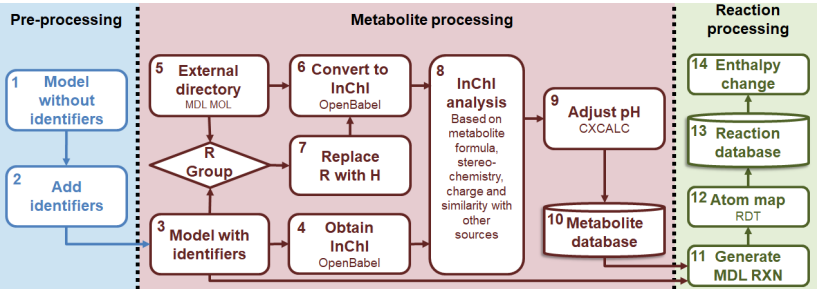


Figure 1. `generateChemicalDatabase` **workflow.**

In order to identify the metabolite structure that most closely resembles the metabolite in the genome-scale reconstruction, identifiers from different sources are compared based on their InChI (Table 1). Finally, the obtained atom mapped reactions are used to identify the number of broken and formed bonds, as well as the enthalpy change of the reactions in the genome-scale reconstruction.

Table1. Each InChI layer contains information that can be used to identify the metabolite structure that is most similar to the metabolite in the genome-scale reconstruction.

Concept	Score	Description
Chemical formula	0 or 10	The chemical formula indicated in the genome-scale model is compared with that obtained from the InChI. This feature is given more weight in order to keep the metabolite in the genome-scale model, where all reactions are stoichiometrically consistent and mass balanced. Hydrogen atoms were ignored in this comparison since they can be modified based on the charge.
Charge	0 or 1	The charge indicated in the genome-scale model is compared with the charge obtained from the source.
Stereochemical information	0 or 1	Indicates whether the InChI contains stereochemical information or not.
Standard	0 or 1	Indicates whether the InChI is standardised or not.
Similarity with other databases	0 to 1	The number of sources where the InChI strings are identical, divided by the total number of sources.
Main layer similarity	0 to 1	The number of sources where the main layers are identical, divided by the total number of sources.
InChI with more layers	0 or 1	InChI with more layers.

The goal of the comparison is to obtain a larger number of atomically balanced metabolic reactions. The Reaction Decoder Tool algorithm[5] (RDT) is used to obtain the atom mappings of each metabolic reaction. The atom mapping data is used to calculate the number of bonds formed or broken in a metabolic reaction, as well as the enthalpy change. The information gathered is incorporated into the COBRA model.

We will obtain chemoinformatic database of the Ecoli core model in this tutorial.

The user-defined parameters in the function `generateChemicalDatabase` will activate various processes. Each parameter is contained in the struct array options and described in detail below:

- **outputDir:** The path to the directory containing the chemoinformatic database (default: current directory).
- **printlevel:** Verbose level.
- **standardisationApproach:** String containing the type of standardisation for the molecules (default: 'explicitH' if openBabel[3] is installed, otherwise default: 'basic'):

  - **explicitH:** Chemical graphs;
  - **implicitH:** Hydrogen suppressed chemical graph;
  - **basic:** Update the header.

- **keepMolComparison:** Logical value, indicate if all metabolite structures per source will be saved or not.
- **onlyUnmapped:** Logic value to select create only unmapped MDL RXN files (default: FALSE, requires Java to run the RDT[5]).
- **adjustToModelpH:** Logic value used to determine whether a molecule's pH must be adjusted in accordance with the COBRA model. (default: TRUE, requires MarvinSuite[4]).
- **addDirsToCompare:** Cell(s) with the path to directory to an existing database (default: empty).
- **dirNames:** Cell(s) with the name of the directory(ies) (default: empty).
- **debug:** Logical value used to determine whether or not the results of different points in the function will be saved for debugging (default: empty).

```
options.outputDir = [resultsDir 'database'];
options.printlevel = 1;
options.debug = true;
options.standardisationApproach = 'explicitH';
options.adjustToModelpH = true;
options.keepMolComparison = false;
options.onlyUnmapped = false;
```

Use the function `generateChemicalDatabase`

```
load  ecoli_core_model.mat
info = generateChemicalDatabase(model, options);
```

Bibliography

1. Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, et al., "Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0", Nature protocols (2019);

2. Orth, J. D., Palsson, B. Ø. and Fleming, R. M. T., "Reconstruction and use of microbial metabolic networks: the core escherichia coli metabolic model as an educational guide", EcoSal Plus (2010).

3. O'Boyle et al,. "Open Babel: An open chemical toolbox." Journal of Cheminformatics (2011).

4. "Marvin was used for drawing, displaying and characterizing chemical structures, substructures and reactions, ChemAxon (http://www.chemaxon.com)".

5. Rahman et al,. "Reaction Decoder Tool (RDT): Extracting Features from Chemical Reactions", Bioinformatics (2016).