Bioinformatics, 37(20), 2021, 3560–3568
doi: 10.1093/bioinformatics/btab368
Advance Access Publication Date: 18 May 2021
Original Paper

Systems biology

NICEpath: Finding metabolic pathways in large networks through atom-conserving substrate–product pairs

Jasmin Hafner and Vassily Hatzimanikatis *

Laboratory of Computational Systems Biotechnology (LCSB), Institute of Chemical Sciences and Engineering (ISIC), School of Basic Sciences (SB), Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 30, 2020; revised on March 22, 2021; editorial decision on May 7, 2021; accepted on May 17, 2021

Abstract

Motivation: Finding biosynthetic pathways is essential for metabolic engineering of organisms to produce chemicals, biodegradation prediction of pollutants and drugs, and for the elucidation of bioproduction pathways of secondary metabolites. A key step in biosynthetic pathway design is the extraction of novel metabolic pathways from big networks that integrate known biological, as well as novel, predicted biotransformations. However, the efficient analysis and the navigation of big biochemical networks remain a challenge.

Results: Here, we propose the construction of searchable graph representations of metabolic networks. Each reaction is decomposed into pairs of reactants and products, and each pair is assigned a weight, which is calculated from the number of conserved atoms between the reactant and the product molecule. We test our method on a biochemical network that spans 6546 known enzymatic reactions to show how our approach elegantly extracts biologically relevant metabolic pathways from biochemical networks, and how the proposed network structure enables the application of efficient graph search algorithms that improve navigation and pathway identification in big metabolic networks. The weighted reactant–product pairs of an example network and the corresponding graph search algorithm are available online. The proposed method extracts metabolic pathways fast and reliably from big biochemical networks, which is inherently important for all applications involving the engineering of metabolic networks.

Availability and implementation: https://github.com/EPFL-LCSB/nicepath.

Contact: vassily.hatzimanikatis@epfl.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Extracting meaningful metabolic pathways from large metabolic networks is essential for the computational design of bioproduction pathways, for the elucidation of biosynthesis of natural products, and for the fundamental understanding of metabolism. Metabolic pathways describe the transformation from one or several source molecules over consecutive reaction steps into one or several target molecules and provide the roadmap guiding these various applications (Cravens *et al.*, 2019; Lin *et al.*, 2019; Nielsen and Keasling, 2016). Traditionally, metabolic pathways were drawn by hand after directly inferring biochemical transformations from experimental evidence. However, the advent of the omics era and the dramatic increase of computational data resources has drastically changed the way we study biochemistry. Biochemical knowledge is now collected in continuously growing databases, providing new opportunities for fundamental research and metabolic engineering. These resources can be harnessed to design non-canonical pathways that do not exist in nature. While many non-natural pathways have been historically designed by intuition using paper and pencil, it is likely that alternative and more efficient solutions will be missed given the wealth of available biochemical data, thus making systematic and automated pathway extraction indispensable. To address this challenge, computational pathway search tools have been developed to extract metabolic pathways from biochemical databases (Hadadi and Hatzimanikatis, 2015; Wang *et al.*, 2017).

Pathway search tools aim to provide meaningful, easily interpretable metabolic pathways as they are shown in textbooks, but through an automated approach. A typical, linear pathway starts with a precursor molecule that is chemically modified by subsequent enzymatic steps until a target molecule is obtained. The atoms of the precursor compounds that are conserved throughout the pathway can be defined as *core atoms*. Cofactors and co-metabolites are considered *boundary compounds* and their metabolic provenance and fate are not further detailed in the linear pathway. In reality however, pathways are not always linear. Anabolic pathways may involve the concatenation of two or more metabolites to form a bigger compound [e.g. (*S*)-norcoclaurine synthase], and catabolic pathways may break down a compound into two or more smaller molecules. However, every branched pathway can also be described as a combination of linear pathways, and we therefore only consider linear pathways in this work. Here, we define biologically relevant, linear pathways as biochemical routes that fulfill the following criteria: (i) Core atoms are conserved throughout the pathway, (ii) loops are not allowed, meaning that no metabolite appears twice and (iii) other metabolites that contribute to the main biotransformation route in a lesser degree are considered as cofactors or co-metabolites.

Pathway search methods are applied to biochemical networks that define the search space. There are two main approaches to mathematically describe a biochemical network: A stoichiometric matrix or a mathematical graph (Wang *et al.*, 2017). Stoichiometric reaction matrices can be searched for pathways by optimizing the production of a target molecule within a metabolic network (Kumar *et al.*, 2018). However, this technique is not applicable to biochemical networks the size of biochemical databases with tens to hundreds of thousands of reactions and is therefore not further discussed here. Graph-based methods on the other hand are suitable for large-scale applications due to their computational efficiency. They represent metabolic networks as mathematical graph structures, and then find paths within the graph from a given source to a target metabolite. Several approaches have been explored to bias a biochemically blind graph search algorithm toward biologically meaningful pathways, such as the exclusion of cofactors from the network to avoid shortcuts through hub metabolites (e.g. CO_2, ATP, H_2O) (Ma and Zeng, 2003), defining substrate-product pairs through the chemical similarity of reactants (Pertusi *et al.*, 2015), precomputing recurring subpaths (Kim *et al.*, 2020) and atom or substructure conservation throughout the pathway (Sankar *et al.*, 2017). Atom conservation in general, and carbon conservation in particular, is a valuable criterion for finding biologically meaningful pathways (Arita, 2004; Blum and Kohlbacher, 2008; Heath *et al.*, 2010; Huang *et al.*, 2017; Kumar *et al.*, 2018; Tervo and Reed, 2016).

Several solutions to pathway discovery employ the concept of atom conservation. Initially, the tracking of single atoms was used by Arita *et al.* to calculate network properties of the metabolism of *Escherichia coli* (Arita, 2004). Later, atom tracking was used to improve the quality of pathway search tools by ensuring that one or several atoms were conserved throughout the pathway (Fooshee *et al.*, 2013; Heath *et al.*, 2010; Huang *et al.*, 2017; Latendresse *et al.*, 2012; 2014; Pey *et al.*, 2013). Atom-tracking methods have been shown to find biologically relevant pathways, although the high quality came with an increased computational cost. An alternative strategy has been pursued by the Kyoto Encyclopedia of Genes and Genomes (KEGG). Their reactions are annotated with chemical structure alignments, also called substrate-product pairs or reactant pairs (short RPAIRs) (Shimizu *et al.*, 2008). The KEGG RPAIR database consists of manually curated, atom-mapped, substructure-conserving substrate-product pairs. KEGG differentiates between five types of RPAIRS: ‘main’, ‘cofac’, ‘trans’, ‘ligase’ and ‘leave’. The four latter ones describe cofactor pairs, small groups transferred by transferases, nucleotide triphosphate consumption by ligases, and the addition or removal of small inorganic compounds by lyases and hydrolases, respectively. The first type, ‘main’, describes the main biotransformation in a given reaction.

KEGG’s pathway prediction server, named PathPred, uses the ‘main’ reactant pairs to create a searchable graph of biologically meaningful biotransformations (Moriya *et al.*, 2010). Instead of tracking atoms individually, PathPred approximates the atom conservation by defining moiety-conserving reactant pairs, which decreases the complexity of the path search problem. However, their classification system is based on a combination of manual curation and automatic annotation, a strategy that is not easily applicable to large biochemical networks, such as the ATLAS of Biochemistry with its more than 140 000 predicted reactions (Hadadi *et al.*, 2016; Hafner *et al.*, 2020), or its successor database ATLASx with its more than 5 million predicted reactions (Mohammadi-Peyhani *et al.*, 2021). Large biochemical databases, especially those including hypothetical reactions, require reliable and computationally efficient algorithms to extract biologically relevant biochemical pathways.

Here, we address the challenge of efficiently searching and analyzing big biochemical networks. We propose a new method, named NICEpath, that biases the graph search toward atom-conserving pathways. To achieve this, we calculate weighted reactant–product pairs that reflect the atom conservation in each reaction, and we use the atom-conserving pairs to represent biochemical reaction networks as weighted graphs that are compatible with efficient search algorithms. The pathways found by NICEpath therefore fit our definition of ‘biologically meaningful’ in the sense that they fulfill the three criteria mentioned earlier. The algorithm finds atom-conserving pathways first and returns a pathway list ranked by overall atom conservation. NICEpath can be readily employed to extract and compare metabolic pathways from biochemical database (e.g. KEGG) or from metabolic networks specific to an organism (e.g. genome-scale models). The method can be further applied to efficiently search large biochemical networks, as they are generated by reaction prediction tool such as BNICE.ch (Hatzimanikatis *et al.*, 2005).

2 Materials and methods

Our approach can be divided into four steps (Fig. 1): (i) The first step consists of acquiring an atom-level representation of each reaction. The atom maps can come from databases, atom-mapping algorithms, or, in our case, enzymatic reaction rules as implemented in BNICE.ch (Hatzimanikatis *et al.*, 2005). (ii) In a second step, each atom-mapped reaction is decomposed into all the possible reactant–product pairs. For each pair, we calculate the Conserved Atom Ratio (CAR) from the number of conserved atoms between reactant and product and the size of the molecules in terms of number of atoms. (iii) The atom-weighted substrate-product pairs are used to construct a weighted undirected graph, where the distance between reactants and products are inversely proportional to the CAR. (iv) Once the graph of weighted substrate-product pairs is constructed, we can apply well-established graph search methods to find the shortest paths, which will inherently find the pathways that conserve the highest number of atoms. NICEpath uses the Yen’s k-shortest loop-less path (Yen, 1971) algorithm, a standard method to find a predefined number (k) of shortest paths in weighted graph, avoiding the repetition of nodes.

2.1 Biochemically correct atom mapping with BNICE.ch

Atom-mapped reactions are the prerequisite for calculating weighted reactant–product pairs. Here, we use the computational tool BNICE.ch, developed to predict hypothetical biochemical networks, to calculate biochemically correct atom mappings of enzymatic reactions. The core of BNICE.ch consists of 442 bidirectional, generalized biochemical reaction rules that describe the biochemical reaction mechanisms of enzymatic reactions. The reaction rules are applied to a molecular structure to (i) reconstruct atom-mapped, known biochemical reactions; and (ii) to predict all possible biochemical transformations that a given compound can undergo along with the product compounds generated in the process. Here, BNICE.ch calculates atom maps for metabolic reactions using the mechanistic knowledge stored in the reaction rules, as described by Hadadi et al. (2017). In this step, other tools for the automatic atom mapping of reactions may also be applied to generate atom maps (Chen *et al.*, 2013; Fooshee *et al.*, 2013; Latendresse *et al.*, 2012).

2.2 Calculation of weighted reactant–product pairs

The following steps are applied to each reaction in the network to generate atom-weighted reactant–product pairs: (i) Each reaction is split into all possible reactant–product pairs. (ii) For each pair of reactant and product, the number of common atoms (n_c) between reactant and product is calculated along with the total number of atoms in the reactant (n_r) and the total number of atoms in the product (n_p). Hydrogen atoms are omitted from the calculation. (iii) For each pair, the ratio of conserved atoms (in the following, called Conserved Atom Ratio, or CAR) is calculated with respect to the reactant (CAR_r) and with respect to the product (CAR_p).

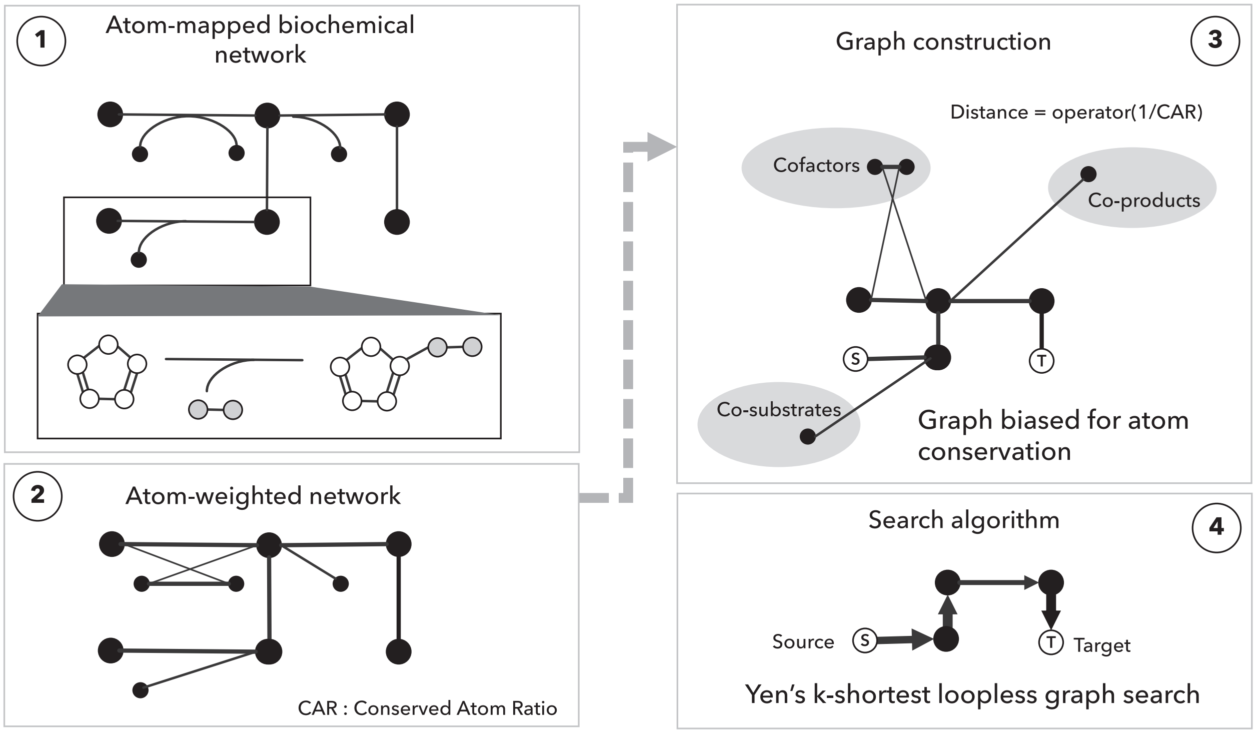
1
Atom-mapped biochemical network
2
Atom-weighted network
CAR : Conserved Atom Ratio
Graph construction
3
Distance = operator(1/CAR)
Cofactors
Co-products
S
T
Co-substrates
Graph biased for atom conservation
Search algorithm
4
Source
S
T
Target
Yen's k-shortest loopless graph search

Fig. 1. The workflow of the pathway search is divided into two parts. The first two steps (left) describe the atom weighting of the network from atom-mapped reactions. In this study, steps 1 and 2 are performed by BNICE.ch. Steps 3 and 4 (right), implemented in NICEpath, take the atom-weighted network as an input to create a searchable graph structure and finally apply a Yen's k-shortest pathway search

$$CAR_r = \frac{n_c}{n_r}, CAR_p = \frac{n_c}{n_p} \tag{1,2}$$

(iv) To calculate a bidirectional CAR, the mean CAR is multiplied with a correction factor that decreases with the difference between the number of common atoms and the total number of atoms in the molecule.

$$CAR = \frac{CAR_r + CAR_p}{2} \cdot \left(1 - |CAR_r - CAR_p|\right) \tag{3}$$

The only exception to this approach is made for reactions involving the cofactor Coenzyme A (CoA). In a molecule, CoA is treated as a single atom when it occurs in both the reactant and in the product, mainly because the high number of conserved atoms between the comparably big CoA leads to high CARs, thus masking the biochemically more interesting connections between the smaller metabolites that are attached to and detached from CoA during metabolic transformations. The final CAR value is used to weight reactant–product pairs in the network.

2.3 Assigning mechanisms to biochemical reactions from the KEGG reference network

We used KEGG as a reference database for enzymatic reactions, from which we extracted all reactions that have an associated mechanism in BNICE.ch. If a given reaction from KEGG could be reconstructed with BNICE.ch, it was assigned a reaction mechanism that allowed us to retrieve the number of conserved atoms between each reactant–product pair. The set of KEGG reactions with assigned reaction mechanisms and pre-calculated CAR values was used for further validation and as an example network for network analysis and pathway search. The set of BNICE.ch curated KEGG reactions is available from the GitHub repository at https://github.com/EPFL-LCSB/nicepath.

2.4 Graph representation of biochemical networks

For a given reaction network, NICEpath loads all the reactant–product pairs to generate a weighted, undirected graph, where metabolites are nodes connected by edges, representing the reactant–product relationship. Edges are assigned a weight that defines the relation between two connected nodes. To use state-of-the-art shortest-path graph search algorithms, highly atom-conserving reactants should be close to each other, and pairs that only share a few atoms should be further away. Hence, we convert the CAR into a distance:

$$\text{Default transformation } \mathit{distance} = \frac{1}{CAR} \tag{4}$$

NICEpath accepts two alternative ways to calculate the distance, which can be used to modulate the influence of the atom conservation on the weight of the reactant–product pair.

$$\text{Square root transformation } \mathit{distance} = \sqrt[2]{\frac{1}{CAR}} \tag{5}$$

$$\text{Exponential transformation } \mathit{distance} = \frac{e^{1/_{CAR}}}{e} \tag{6}$$

The type of transformation can be changed to square root or exponential depending on the nature of the pathway search problem, i.e. the structures of source and target molecules as well as the estimated number of biotransformation used to convert one into the other. The distance measure is used to reconstruct a directed graph whose edge weights represent the atomic distance between reactants and products. The different transformations as a function of the CAR are visualized in Supplementary Figure S1. For longer pathways, we recommend using the exponential transformation because it increases the penalty for pairs with low CARs, which makes the search more conservative in terms of atoms.

For this study, we grouped duplicate KEGG compounds into one node. Duplicates were identified based on the first fourteen letters of the InChIKey that describe the atom connectivity of a compound, but that do not contain additional information (i.e. charge, stereochemistry, isotopes) In practice, this means that different stereoisomers of the same molecular structure were merged into one node.

2.5 Finding metabolic pathways with graph search

NICEpath applies a Yen's k-shortest loop-less path search (Yen, 1971) to extract the shortest pathways from the weighted network of reactant–product pairs using the python package NetworkX. As inputs, the pathway search algorithm takes a weighted graph, a source compound, a target compound and the maximum number of shortest paths (k) to be found. As soon as this number k is reached, the algorithm stops and returns all the k-shortest paths in terms of summed edge weights.

The run time of NICEpath depends on the structure of the network, the distance between the source and target compound in the graph, the number of pathways to be found and the maximum pathway length allowed. As an example, to find 10 000 pathways of maximum length 100, the algorithm runs for about 15 min on a standard desktop computer using a single core. If there are several source compounds given as input, NICEpath can run path searches in parallel for different source compounds using all available cores.

2.6 Network analysis in NICEpath

NICEpath first calculates standard network statistics, such as the number of nodes and edges, and then extracts an undirected, unweighted network from the original network by only considering edges with a CAR higher than a given threshold. For this new network, the number of components, or disjoint graphs, is extracted, and the biggest component is further analyzed regarding its size relative to the previous network as well as its diameter. Since searching for pathways between two compounds belonging to different disconnected graphs will not yield any good pathways, NICEpath will warn the user in this case.

2.7 Software

The NICEpath code can be executed with any python version up to 3.7. The NetworkX python library (https://networkx.github.io/) was used to implement and search the reaction graph. An extensive list of libraries used can be found in the specification file on GitHub.

3 Results

3.1 Weighted substrate-product pairs capture the main biotransformations

To validate the biochemical relevance of weighted substrate-product pairs, we compared them to the KEGG RPAIR database. KEGG RPAIR distinguishes 'main' substrate-product pair of the reaction from secondary types of pairs (e.g. 'cofac', 'leave'). To take the alcohol dehydrogenase reaction as an example, the main pair would be the transformation of the primary alcohol to the aldehyde, and the conversion of the cofactor NAD+ to NADH would be of type 'cofac' (Fig. 2). 'Main' pairs are used to draw the KEGG metabolic pathway maps. Therefore, a method that accurately predicts KEGG RPAIRS of type 'main' can be used to reconstruct biologically relevant metabolic pathways. It should be noted that KEGG discontinued the manual definition and curation of RPAIRS in 2016, and replaced the concept of RPAIRS with an automatically calculated alternative, RCLASS.

We validated the NICEpath method by predicting KEGG RPAIRS of type 'main' using the concept of the Conserved Atom Ratio. We used BNICE.ch to calculate CAR values for a test set of 6546 KEGG reactions for which the exact reaction mechanism is known, and which are, therefore, reconstructed by BNICE.ch (Supplementary Table S1). From these 6546 reactions, we determined 10 747 substrate-product pairs with a non-zero CAR, meaning that at least one non-hydrogen atom is conserved between the substrate and the product (Supplementary Table S2). Out of these 10 747 pairs, 5148 were found to be KEGG RPAIRS of type 'main'. Since RPAIRs are defined based on the conservation of structural moieties within a reaction, we hypothesized that the higher the CAR value, the more atoms conserved between a substrate and a product, and hence the higher the probability that the pair would be a KEGG RPAIR of type 'main'. We should therefore be able to predict the membership of a pair to the set of 'main' KEGG RPAIRS by using a given CAR threshold as a classifier. To test our hypothesis that the CAR is a good predictor for a reactant–product pair to be of KEGG RPAIR type 'main', we performed a Receiver-Operator Characteristic (ROC) analysis (Fig. 3). The reference for true pairs were the 5148 'main' RPAIRs (true positives), and the remaining 5599 pairs were true negatives.

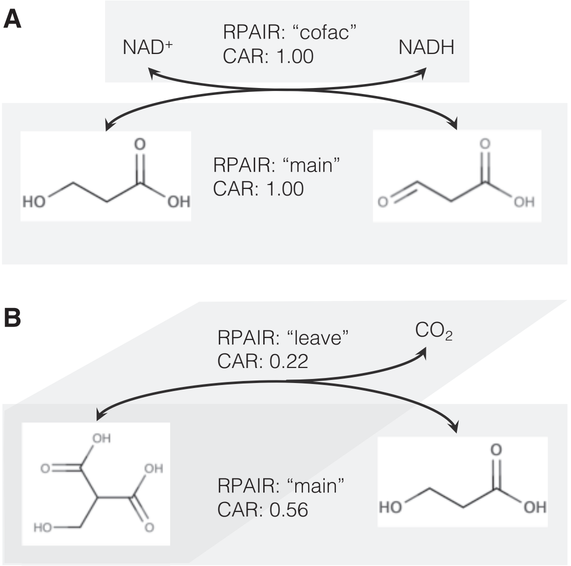

Fig. 2. Example of relation between KEGG RPAIRs and the CAR value in a biochemical reaction. (**A**) Alcohol dehydrogenase: in an oxidoreduction reaction only electrons and protons are exchanged between the reaction participants, resulting in two distinct substrate-product pairs with a maximum CAR value. (**B**) Decarboxylation reaction: the atoms of the reactant are distributed between a leaving CO_2 molecule with a low CAR value and a product molecule with a higher CAR value corresponding to the 'main' RPAIR

For 100 CAR cutoff values between zero and one we calculated the number of good predictions (i.e. number of pairs with a CAR above the cutoff and of type 'main', or true positives) and bad predictions (i.e. number of pairs with a CAR above the cutoff and not of type 'main', or false positives). By drawing true positives versus false positives, we found an Area Under Curve (AUC) of 0.88. An AUC above 0.8 is considered an 'excellent discrimination' (Hosmer and Lemeshow, 2000). We further show the tradeoff between sensitivity and specificity, as well as the Youden's index (i.e. sensitivity + specificity - 1) to characterize this tradeoff (Youden, 1950) and to determine an optimal CAR cutoff. We found that the Youden's index is maximal at a CAR equal to 0.34, which suggests that this is the optimal CAR cutoff to tell whether a given substrate-product pair conserves enough atoms to be considered a 'main' pair. This analysis shows that we can reliably use the CAR to predict KEGG RPAIRS of type 'main'. The network of weighted KEGG reactant pairs for 6546 KEGG reactions is included in the NICEpath software and used as a reaction database in the default search.

3.2 Graph-theoretical analysis of metabolic networks

Characterizing biochemical networks from a graph-theoretical point of view can be used to evaluate the quality and connectivity of the represented network, and also bring new insights into the overall organization of metabolism. Furthermore, knowing the graph-theoretical properties of a biochemical network can be crucial for anticipating potential problems in the pathway search. NICEpath provides basic network statistics that allow us to assess the quality of the data. Here, the weighted graph of the KEGG network used for validation initially contained 5578 compounds, or nodes, and 20 911 directed edges representing reactant–product pairs. Certain graph properties are not defined for weighted directed graphs, **such** as the number of components or the network diameter. For

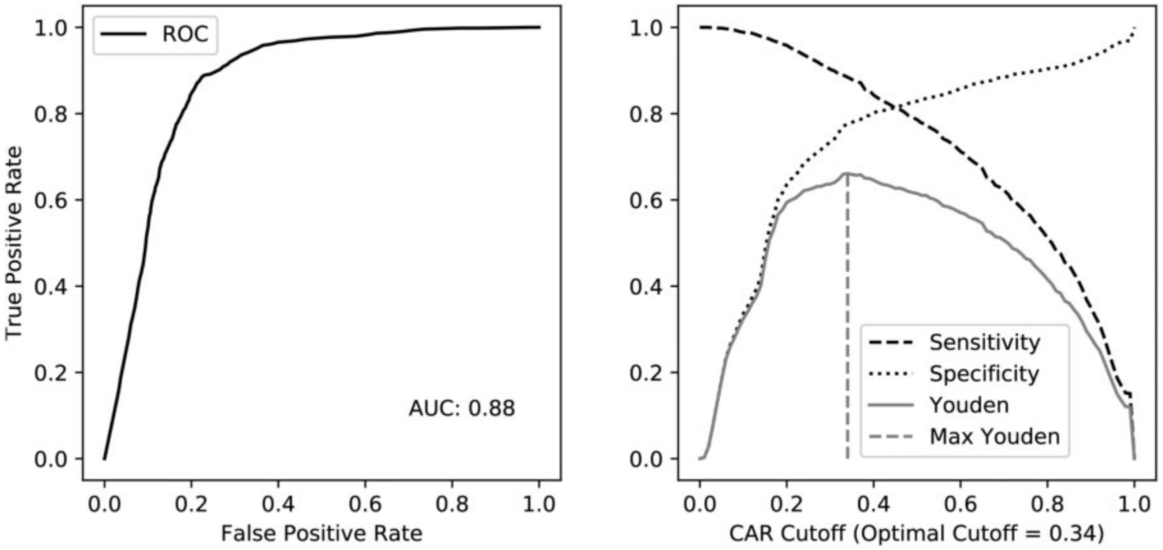

Fig. 3. Sensitivity and specificity analyses. The ROC (left panel) curve shows the prediction of KEGG RPAIRS of type 'main' by CAR score from BNICE.ch. The right panel shows the tradeoff between specificity (black dashed line) and sensitivity (black dotted line). The Youden's index (gray continuous line) reaches its maximum (0.66) at a CAR value of 0.34 (gray dashed line). ROC: Receiver operator characteristic, AUC: Area under the curve

calculating these properties, a simple, non-directed graph was generated by removing reactant–product pairs with a CAR lower than the previously calculated optimal threshold of 0.34 and by removing the weights on the remaining reactant–product pairs. The unweighted graph contains 5518 nodes and 5541 edges, which are distributed over 813 smaller disjoint graphs, or so-called components. The biggest component contains 2663 nodes (48%) and 3422 edges (62%), and it has a network diameter of 40. In other words, the longest shortest pathway connecting two compounds counts 40 biotransformation steps in the main component of the KEGG network. This means that our KEGG network is dominated by one big component, or 'island', that includes half of the metabolites in KEGG and represents the core metabolism plus connected secondary metabolism. The remaining metabolites are organized in small, disconnected subnetworks, which we hypothesize to be mostly secondary metabolites without defined biosynthesis pathways.

3.3 Finding biologically relevant pathways with NICEpath

To illustrate the output of NICEpath, we discuss two example pathway searches. In the first example, we tried to biochemically connect tyrosine to caffeate, and we allowed a maximum number of ten pathways to be found. The pathway search resulted in ten pathways with lengths ranging from two to six consecutive reaction steps (Table 1). The quality of the pathway can be estimated from the pathway score and the average CAR. The pathway score sums the distances for each reactant–product pair in the pathway. The score reflects both the length of the pathway as well as the quality of atom conservation within the pathway, and it is eventually used by NICEpath to rank the paths. The average CAR estimates the quality of the pathway by averaging the atom conservation over each reaction step. Out of these ten best pathways, the pathways ranked first, second and fifth were chosen for visual inspection (Fig. 4). The first pathway had a very low score of 2.24 combined with a high average CAR (0.89) and a length of two, which indicatesthat the pathway is of good quality because it only requires a small number of steps, all of them showing high atom conservation. Indeed, KEGG proposes this pathway in its phenylpropanoid biosynthesis map, meaning that it is biologically relevant. The second pathway, although longer, has a similarly high average CAR of 0.93, a length of four steps, and it can also be found in KEGG. To contrast these two good pathway examples with a poor example, the pathway ranked fifth shows a slightly lower average CAR of 0.81, which is due to the attachment and subsequent detachment of a one-carbon unit. In this pathway, out of five reaction steps, the last step is redundant with the first pathway, while the first four steps describe a detour from tyrosine to coumarate (C00811). This last, suboptimal pathway cannot be found in the KEGG map for phenylpropanoid biosynthesis.

In a second example, we searched for pathways connecting the compounds tyrosine and syringin. The number of pathways to be found was restricted to five, and we used three different transformations to calculate the distance between reactant–product pairs: The default transformation 1/CAR, the square root transformation, and the exponential transformation. Using the default option, NICEpath first listed three short pathways with a low average CAR (~0.5), followed by two longer pathways with high average CAR (~0.8) (Table 2). The square root option yielded only short pathways with a low average CAR, while the exponential option only resulted in longer pathways of high average CAR. Interestingly, all the long pathways with high CAR were identified as known metabolic pathways in KEGG, indicating that increasing the influence of the CAR on the distance by choosing an exponential transformation operator is helpful to reliably extract longer pathways.

Two pathways were chosen to understand in detail the influence of the type of transformation used for calculating the distances between reactants and products: one was short with a low CAR (A) and one was long with a high CAR (D*) (Fig. 5). Pathway A connected tyrosine to syringin in four reaction steps, with a relatively low average CAR of 0.51. As already indicated by the low CAR, the pathway turned out to be a shortcut through pathway was ranked first in the default and the square root transformation types, but, interestingly, ranked 1114th in the exponential case when re-running the search with an upper limit of 2000 pathways. The exponential transformation increases the penalty of atom loss in biotransformation, which leads to a higher pathway score assigned to the shortcut pathway. Pathway D* connected tyrosine to syringin in eight reaction steps, with a high average CAR of 0.86. It was ranked first using an exponential transformation, ranked fourth using the default distance calculation, and ranked 43rd for the square root case. This second pathway kept the molecular core structure of tyrosine and modified it to produce syringin, conserving a maximum number of atoms. Remarkably, this pathway is part of the KEGG pathway map for phenylpropanoid biosynthesis, and it

Table 1. Output of example pathway search from tyrosine (C00082) to caffeate (C01197)

Index	Pathway length	Intermediates	Reaction IDs	Pathway score	Average CAR
1	2	C00082->C00811->C01197	R00737->R07826	2.24	0.89
2	4	C00082->C00811->C00223->C00323->C01197	R00737->R01616->R07436->R01943	4.32	0.93
3	4	C00082->C01179->C03672->C00811->C01197	R00729->R03336->R08766->R07826	4.33	0.93
4	4	C00082->C00079->C00423->C00811->C01197	R07211->R00697->R02253->R07826	4.52	0.89
5	5	C00082->C00826->C00079->C00423->C00811->C01197	R00732->R00691->R00697->R02253->R07826	6.27	0.81
6	6	C00082->C01179->C03672->C00811->C00223->C00323->C01197	R00729->R03336->R08766->R01616->R07436->R01943	6.41	0.94
7	6	C00082->C00811->C00223->C00323->C00406->C01494->C01197	R00737->R01616->R07436->R01942->R02194->R03366	6.44	0.93
8	6	C00082->C00079->C00423->C00540->C00223->C00323->C01197	R07211->R00697->R02255->R08815->R07436->R01943	6.50	0.93
9	6	C00082->C00811->C00423->C00540->C00223->C00323->C01197	R00737->R02253->R02255->R08815->R07436->R01943	6.50	0.93
10	6	C00082->C00079->C00423->C00811->C00223->C00323->C01197	R07211->R00697->R02253->R01616->R07436->R01943	6.60	0.91

Note: KEGG identifiers are used to specify compounds and reactions. The maximum number of pathways (k) was set to 10, and only one reaction alternative was printed when several reactions could do the same biotransformation.

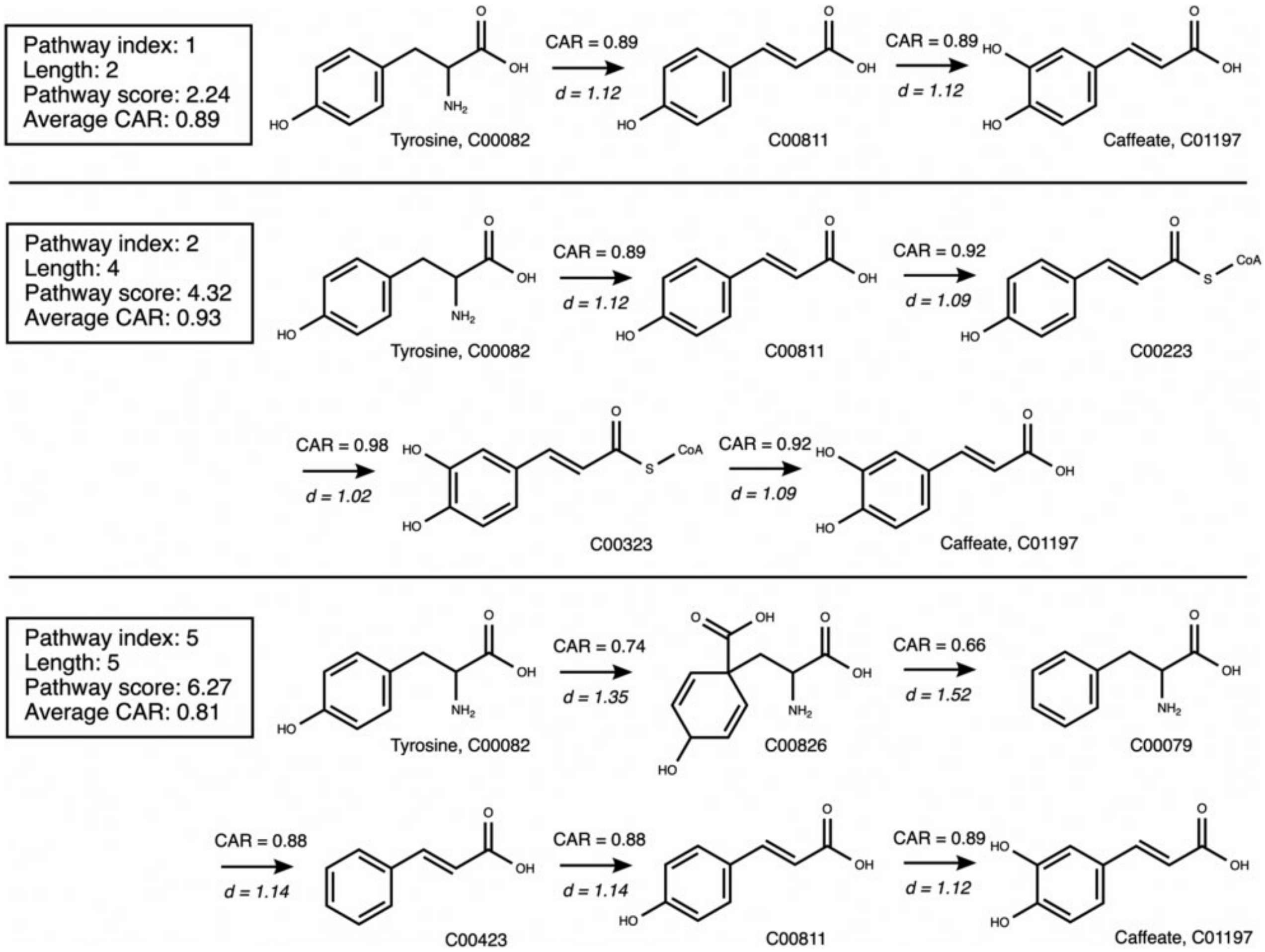

Fig. 4. The pathways from Table 1 connecting tyrosine and caffeate with index numbers 1, 2 and 5 are visualized in detail for comparison. For each biotransformation, the CAR value as well as the default distance (*d*) are indicated

Table 2. Output of pathway search from tyrosine (C00082) to syringin (C01197)

Distance	Index	Pathway length	Intermediates	Mapping	Pathway score	Average CAR
$\frac{1}{CAR}$	1	4	C00082->C00811->C16827->C00031->C01533	A	9.06	0.51
	2	4	C00082->C00811->C04415->C00029->C01533	B	9.86	0.48
	3	4	C00082->C00811->C16827->C00029->C01533	C	9.86	0.48
	4	8	C00082->C00811->C01197->C01494->C05619->C00482->C05610->C02325->C01533	D*	9.88	0.86
	5	8	C00082->C00811->C01197->C01494->C02666->C12204->C05610->C02325->C01533	E*	9.90	0.85
$\sqrt[2]{\frac{1}{CAR}}$	1	4	C00082->C00811->C16827->C00031->C01533	A	5.93	0.51
	2	4	C00082->C00811->C04415->C00029->C01533	B	6.18	0.48
	3	4	C00082->C00811->C16827->C00029->C01533	C	6.18	0.48
	4	5	C00082->C00079->C00423->C04164->C00031->C01533	F	7.00	0.58
	5	5	C00082->C00811->C00423->C04164->C00031->C01533	G	7.00	0.58
$e^{1/CAR}/e$	1	8	C00082->C00811->C01197->C01494->C05619->C00482->C05610->C02325->C01533	D*	11.09	0.86
	2	8	C00082->C00811->C01197->C01494->C02666->C12204->C12205->C02325->C01533	H*	11.13	0.85
	3	8	C00082->C00811->C01197->C01494->C02666->C00590->C12205->C02325->C01533	I*	11.13	0.85
	4	8	C00082->C00811->C01197->C01494->C02666->C12204->C05610->C02325->C01533	E*	11.13	0.85
	5	9	C00082->C00811->C00223->C00323->C00406->C12203->C00411->C05610->C02325->C01533	J*	11.79	0.90

Note: Three different CAR transformations were used to calculate the distance between pairs of reactants and products: The first one (1/CAR) is the default option in NICEpath, the second is a square root transformation that decreases the effect of atom conservation, and the third one is an exponential transformation that increases the effect of conserved atoms in reactant–product pairs. Pathways are mapped across distance transformations with letters indicated in the column labeled 'Mapping'.

*Pathways marked with an asterisk correspond to known metabolic pathways that fulfill the criteria of biologically relevant pathways.

can therefore be called a confirmed, biologically meaningful pathway. These two examples of pathway search problems illustrate the capacity of NICEpath to efficiently extract biologically relevant pathways from large biochemical networks. The algorithm robustly handled searches for long pathways of eight and more biotransformation steps, as they are usually present in secondary metabolism.

To demonstrate the universality of our approach, we performed a systematic validation on 50 pathways collected from KEGG consisting of 40 biosynthesis and 10 biodegradation routes and involving between 3 and 15 reactions steps (Supplementary Table S3, Supplementary Fig. S2). We evaluated the NICEpath performance using the three proposed transformation operators and compared it to a two-sided breadth-first search within (i) an unweighted network of substrate-product pairs with edges where the CAR exceeds the identified threshold value of 0.34, (ii) an unweighted network of 'main' KEGG RPAIRs (as used in the KEGG PathPred server) and (iii) an unweighted network of all possible KEGG RPAIRs without cofactors. The first and second networks are expected to yield similar results, because a CAR threshold of 0.34 has been shown to well predict KEGG RPAIRs of type 'main'. The third network represents the common approach of removing cofactors from a network of all possible substrate-product edges to avoid extracted pathways to shortcut through cofactors acting as hub metabolites.

For each of the reference pathways, we performed a k-shortest path search within each network between the source and the end compound of the pathway. To evaluate the performance, we retrieved the rank of the reference pathway within the top 100 pathways produced by the algorithm, and we measured the runtime of the search algorithm (Supplementary Table S4). We could show that within the weighted networks, the exponential transformation resulted in longer search time, but it had the highest probability of finding the reference pathway first, while the square root transformation yielded the opposite result (Supplementary Figs S3 and S4.). We further found that the runtimes within the unweighted networks were significantly lower than within the weighted networks, which is due to the fact that the unweighted network allows for a two-sided search starting simultaneously from the source and the target compound. Within the unweighted networks, searches within the network without cofactors were found to be the slowest and the less accurate. The RPAIR 'main' network and the CAR > 0.34 network showed a similar performance with respect to runtime, but the ranking performance was better in the CAR > 0.34 network, which is due to the fact that not all reactions within the reference pathway had RPAIRs assigned in KEGG.

Interestingly, the ranking performance of the exponentially transformed network and the unweighted CAR > 0.34 were very similar. This indicates that when the algorithm runtime becomes a limiting factor, the search can be performed within an unweighted graph (CAR > 0.34) to reduce the runtime, allowing a two-sided search starting simultaneously from the source and the target compound and thus speeding up the search.

3.4 Limitations and future challenges

There are cases in which NICEpath will not find satisfactory solutions. Possible reasons for suboptimal results are (i) the network does not contain the necessary reactions to connect the starting compound to the target compound, and (ii) the source and the target compound initially have only a few atoms in common. The first issue can be solved by adding the missing reactant–product pairs to the network. Missing steps can be hypothesized manually or predicted using reaction prediction tools such as BNICE.ch. The second issue is more complex, since it depends on the molecular structure of the source and target compound, as well as on the real number of biochemical transformations needed to transform one into the other. Possible solutions to improve the output include breaking down the search into several sub-searches by identifying intermediates and increasing the penalty on atom loss by using an exponential transformation of the CAR into the distance between

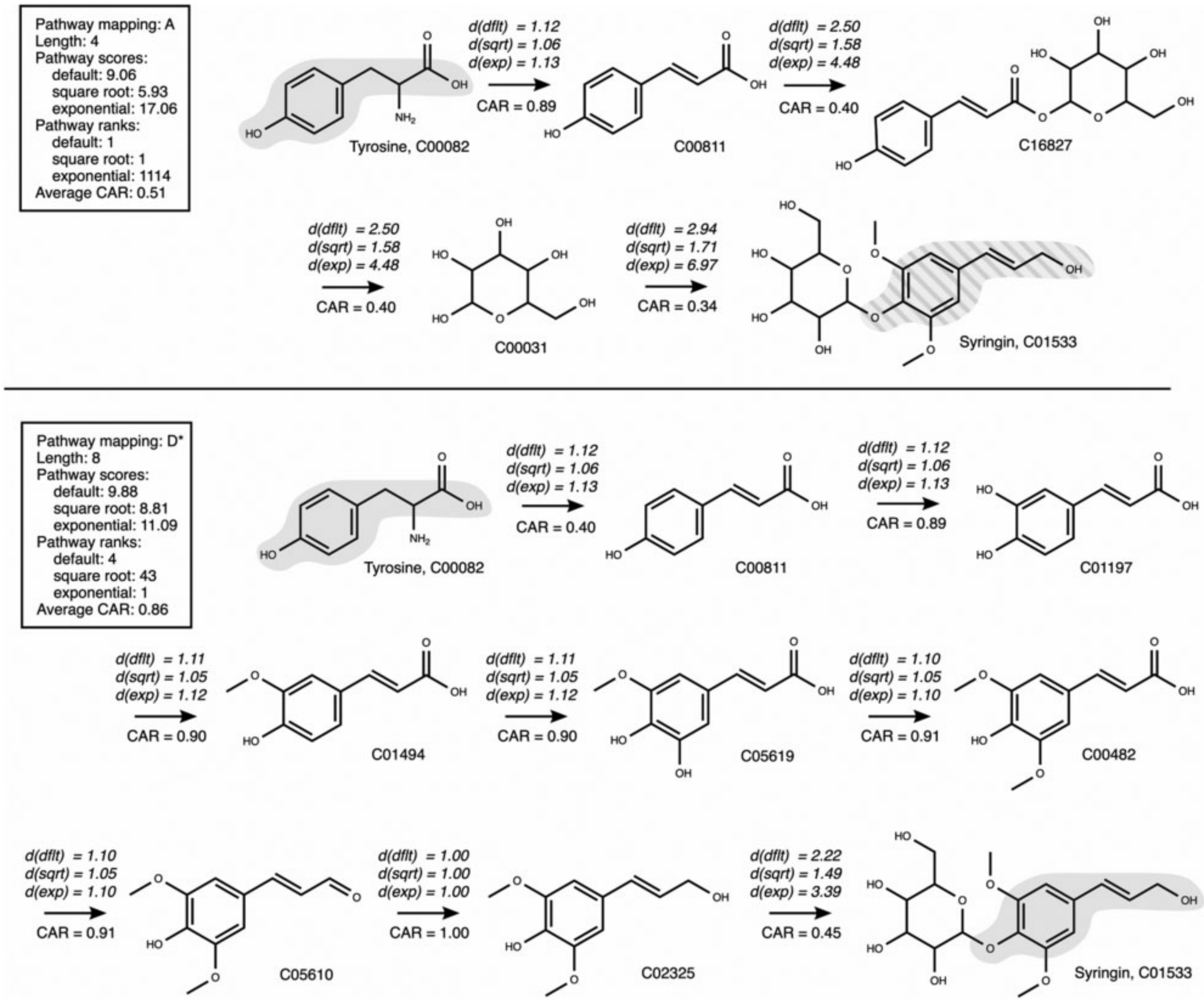

Fig. 5. Comparison of two pathways (A and D*) from the pathway search connecting tyrosine to syringin. For each biotransformation, the CAR value along with the default distances for each transformation are indicated. (dflt): default distance, d(sqrt): square root transformation, d(exp): exponential transformation. The tyrosine moiety is marked in gray if conserved from tyrosine to syringin, and gray striped if entering from co-substrates

reactants and products. While our algorithm successfully circumvents the recurrent problem of shortcuts through small hub metabolites, it does not satisfactorily avoid shortcuts through big hub metabolites such as Coenzyme A (CoA). In fact, reactant pairs involving CoA structures on both sides have a lot of atoms in common, and therefore a high CAR value. For this reason, NICEpath excludes CoA by default from the reactant pair network.

4 Conclusion

We introduce a new pathway search method based on weighted reactant–product pairs. To our best knowledge, this is the first to use automatically generated atom-weighted reactant–product pairs in combination with a k-shortest graph search approach. We benchmark our method for reactant-pair weighting against the KEGG RPAIR database, and we evaluated the performance of the proposed pathway search on 50 pathways obtained from KEGG. The strong point of NICEpath is that it is suitable for big biochemical networks, spanning more than hundreds of thousands biochemical reactions, such as hypothetical reaction networks generated by retrobiosynthesis tools and predictive biochemistry (Hadadi *et al.*, 2016; Hafner *et al.*, 2020). Pathway search constitutes the first step, and hence the foundation, of the overall pathway design pipeline. Downstream pathway analyses include the evaluation of the stoichiometric and thermodynamic feasibility of a pathway within a host organism via Flux Balance Analysis and Thermodynamic Flux Analysis, respectively, the estimation of the kinetic properties of the pathway, and the assessment of enzyme availability (e.g. from databases or from enzyme prediction tools) (Hadadi and Hatzimanikatis, 2015). As all these tools require substantial effort and computational resources, it is key that the pathways proposed initially by a pathway search tool only deliver biologically feasible biotransformation routes.

We estimate that the future development of reaction prediction tools, based on biochemical reaction rules or machine learning methods, will yield big hypothetical reaction networks that require optimized search tools to efficiently extract biochemical pathways. Furthermore, the presented method to translate metabolic networks into a graph structure can be used in the future to analyze the global characteristics of biochemical networks, such as the diameter of a network or its connectivity, and finally to detect and map knowledge gaps in metabolic databases.

Finally, the herein proposed framework will lay the foundation for further developments. Other types of weights, such as kinetic and thermodynamic considerations, can be integrated into the weighting of substrate-product pairs to steer the pathway search toward biochemically feasible pathways, and a set of user-defined parameters will make it easy to fine-tune the pathway search.

Additionaly, metrics other than the here proposed CAR could be utilized to quantify atom conservation within substrate-product pairs, such as the Jaccard index (Jaccard, 1908). The NICEpath code is available on GitHub (https://github.com/EPFL-LCSB/nicepath), and it comes with a collection of 5434 known metabolic reactions with pre-calculated atom-weighted reactant pairs.

Acknowledgements

The authors thank Kaycie Deyle for valuable feedback in the preparation of the manuscript.

Funding

Funding for this work was provided by the Swiss National Science Foundation (SNSF) [2013/158] (MicroScapeX) and the Ecole Polytechnique Fédérale de Lausanne (EPFL).

Conflict of Interest: none declared.

Data availability

The data underlying this article are available in the article and in its online supplementary material and from the GitHub repository at https://github.com/EPFL-LCSB/nicepath.

References

Arita,M. (2004) The metabolic world of *Escherichia coli* is not small. *Proc. Natl. Acad. Sci. USA*, **101**, 1543–1547.

Blum,T. and Kohlbacher,O. (2008) Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J. Comput. Biol.*, **15**, 565–576.

Chen,W.L. *et al.* (2013) Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**, 560–593.

Cravens,A. *et al.* (2019) Synthetic biology strategies for microbial biosynthesis of plant natural products. *Nat. Commun.*, **10**, 2142.

Fooshee,D. *et al.* (2013) ReactionMap: an Efficient Atom-Mapping Algorithm for Chemical Reactions. *J. Chem. Inf. Model.*, **53**, 2812–2819.

Hadadi,N. *et al.* (2016) ATLAS of Biochemistry: a repository of all possible biochemical reactions for synthetic biology and etmabolic engineering studies. *ACS Synth. Biol.*, **5**, 1155–1166.

Hadadi,N. *et al.* (2017) Reconstruction of biological pathways and metabolic networks from in silico labeled metabolites. *Biotechnol. J.*, **12**, 1600464.

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.

Hafner,J. *et al.* (2020) Updated ATLAS of biochemistry with new metabolites and improved enzyme prediction power. *ACS Synth. Biol.*, **9**, 1479–1482.

Hatzimanikatis,V. *et al.* (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics*, **21**, 1603–1609.

Heath,A.P. *et al.* (2010) Finding metabolic pathways using atom tracking. *Bioinformatics*, **26**, 1548–1555.

Hosmer,D.W. and Lemeshow,S.L. (2000) *Applied Logistic Regression*, 2nd edn. Chapter 5, John Wiley and Sons, New York, NY.

Huang,Y. *et al.* (2017) A method for finding metabolic pathways using atomic group tracking. *PLoS One*, **12**, e0168725.

Jaccard,P. (1908) Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, **44**, 223–270.

Kim,S.M. *et al.* (2020) Improving the organization and interactivity of metabolic pathfinding with precomputed pathways. *BMC Bioinformatics*, **21**, 13.

Kumar,A. *et al.* (2018) Pathway design using de novo steps through uncharted biochemical spaces. *Nat. Commun.*, **9**, 184.

Latendresse,M. *et al.* (2012) Accurate atom-mapping computation for biochemical reactions. *J. Chem. Inf. Model.*, **52**, 2970–2982.

Latendresse,M. *et al.* (2014) Optimal metabolic route search based on atom mappings. *Bioinformatics*, **30**, 2043–2050.

Lin,G.-M.M. *et al.* (2019) Retrosynthetic design of metabolic pathways to chemicals not found in nature. *Curr. Opin. Syst. Biol.*, **14**, 82–107.

Ma,H. and Zeng,A.-P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, **19**, 270–277.

Mohammadi-Peyhani,H. *et al.* (2021) ATLASx: a computational map for the exploration of biochemical space. *bioRxiv*, 10.1101/2021.02.17.431583.

Moriya,Y. *et al.* (2010) PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.*, **38**, W138–43.

Nielsen,J. and Keasling,J.D. (2016) Engineering cellular metabolism. *Cell*, **164**, 1185–1197.

Pertusi,D.A. *et al.* (2015) Efficient searching and annotation of metabolic networks using chemical similarity. *Bioinformatics*, **31**, 1016–1024.

Pey,J. *et al.* (2013) Refining Carbon Flux Paths using atomic trace data. *Bioinformatics*, **30**, btt653.

Sankar,A. *et al.* (2017) Predicting novel metabolic pathways through subgraph mining. *Bioinformatics*, **33**, 3955–3963.

Shimizu,Y. *et al.* (2008) Generalized reaction patterns for prediction of unknown enzymatic reactions. *Genome Inf.*, **20**, 149–158.

Tervo,C.J. and Reed,J.L. (2016) MapMaker and PathTracer for tracking carbon in genome-scale metabolic models. *Biotechnol. J.*, **11**, 648–661.

Wang,L. *et al.* (2017) A review of computational tools for design and reconstruction of metabolic pathways. *Synth. Syst. Biotechnol.*, **2**, 243–252.

Yen,J.Y. (1971) Finding the K shortest loopless paths in a network. *Manage. Sci.*, **17**, 712–716.

Youden,W.J. (1950) Index for rating diagnostic tests. *Cancer*, **3**, 32–35.