ATLASx: a computational map for the exploration of biochemical space

Homa Mohammadi-Peyhani[1,2] & Jasmin Hafner[1,2], Anastasia Sveshnikova[1], Victor Viterbo[1], Vassily Hatzimanikatis[1*]

[1] Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne, EPFL, Lausanne, Switzerland

[2] These authors contributed equally

*Corresponding author

Abstract

Metabolic "dark matter" describes currently unknown metabolic processes, which form a blind spot in our general understanding of metabolism and slow down the development of biosynthetic cell factories and naturally derived pharmaceuticals. Mapping the dark matter of metabolism remains an open challenge that can be addressed globally and systematically by existing computational solutions. In this work, we use 490 generalized enzymatic reaction rules to map both known and unknown metabolic processes around a biochemical database of 1.5 million biological compounds. We predict over 5 million reactions and integrate nearly 2 million naturally and synthetically-derived compounds into the global network of biochemical knowledge, named ATLASx. ATLASx is available to researchers as a powerful online platform that supports the prediction and analysis of novel biochemical pathways and evaluates the biochemical vicinity of molecule classes (https://lcsb-databases.epfl.ch/Atlas2).

Keywords: biochemical space, reaction prediction, biological reactivity, pathway search, metabolic engineering, database

Introduction

Metabolic "dark matter" designates biochemical processes where knowledge is still sparse, limiting our general understanding of metabolism, the discovery of key disease mechanisms [1,2], and the development of medicines derived from plant natural products [3]. Metabolic knowledge gaps also hamper the advancement of bioengineering applications like the creation of sustainable cell factories for the green production of commodity chemicals and pharmaceuticals. Key examples of metabolic dark matter include underground metabolism resulting from promiscuous enzymatic activity [4,5], undetected plant natural products and their uncharacterized biosynthesis pathways, and chemical damage of metabolites [6].

Given the vastness of these and other unknown metabolic elements, it is essential to generate hypotheses on potential biochemical functions that guide the experimental identification of new enzymatic functions and detect novel natural products. While genomic, transcriptomic, proteomic, and metabolomic data have the potential to generate important hypotheses on metabolic dark matter, linking these data to metabolic functions remains difficult [7,8], and major gaps in biochemical and metabolic knowledge remain. As an example, 25% percent of proteins in *E. coli*, one of the best studied model organisms, do not have a function assigned [9]. In addition, almost 10,000 metabolites in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [10] are considered to be orphan, meaning that they are not integrated in any biochemical reaction within KEGG. An unbiased and global approach is needed to systematically explore the metabolic dark matter arising from the elasticity of enzymatic catalysis, which can be achieved by current computational approaches.

Computational approaches to biological questions have attracted increasing interest over the past few decades. Diverse computational tools have emerged that can bridge knowledge gaps in metabolism through cheminformatic predictions of potential metabolic reactions, uncharacterized metabolites, and novel enzyme functions. Most of these tools have been developed for metabolic engineering applications, where the objective is to find biosynthetic routes that produce a desired target compound in a host organism [11–15]. Identification of these biosynthetic routes is accomplished by biochemically "walking back" from the target to precursor metabolites that are produced by, or fed to, the host organism. This procedure is called *retrobiosynthesis* and is implemented in a range of tools such as BNICE.ch [16,17], GEM-

Path [18], NovoPathFinder [19], NovoStoic [20], ReactPRED[21], RetroPath [22,23], and Transform-MinER [24]. Retrobiosynthetic methods rely on the concept of *generalized enzymatic reaction rules*. A reaction rule encodes the biochemistry of a substrate-promiscuous enzyme by describing the pattern of the reactive site recognized by the enzyme, as well as the bond rearrangement performed by the enzyme on the substrate. By applying the rule on a substrate that is non-native to the represented enzyme, the rule can predict if (i) the substrate can be recognized by the enzyme, (ii) if the biotransformation can occur, and (iii) the identity of the product molecule(s). While tools featuring reaction rules have the power to predict novel biochemical reactions, their application is usually limited to a given research or engineering question.

One exception is the ATLAS of Biochemistry database [10], which attempts to map dark matter in biochemistry by predicting novel reactions between metabolites from the KEGG database[25]. The ATLAS contains ~150,000 novel enzymatic reactions predicted by the retrobiosynthesis tool BNICE.ch and annotated with putative enzymes suggested by the enzyme prediction tool BridgIT [26]. In contrast to other tools that use automatic rule generation, the reaction rules in BNICE.ch are designed by experts based on biochemical knowledge and assigned the corresponding three-level Enzyme Commission (EC) number, which is a numerical coding used to classify enzyme-catalyzed reactions. The complementary tool BridgIT uses the knowledge of the reactive site encoded in the BNICE.ch rules to predict enzymes that can potentially catalyze novel and orphan reactions. More than 100 reactions predicted by BNICE.ch and stored in ATLAS were validated in 2019 following their addition to KEGG, supporting the predictive utility of the ATLAS tool [27]. In addition, Yang *et al.* were able to experimentally validate predicted ATLAS reactions while designing novel one-carbon assimilation pathways, further demonstrating the value of predictive ATLAS reactions in metabolic engineering. One major drawback of ATLAS is its limitation to KEGG compounds, which excludes many drugs and plant natural products with undefined or putative biological functions. Predicting enzymatic reactions from biochemical compounds retrieved from databases other than KEGG will help expand the scope of our predictions and enhance the application range and the predictive power of the database.

In the following, we present ATLASx, an online biochemical resource providing reliable predictions of biochemical reactions and pathways for synthetic biologists and metabolic

engineers. The ATLASx workflow (Figure 1) unifies biochemical reactions and compounds from 14 different database sources into one curated dataset called bioDB. bioDB holds 1.5 million unique biological or bioactive compounds and 56,000 unique biochemical reactions, which enable the prediction of a hypothetical biochemical space. By applying 490 bidirectional, generalized reaction rules from BNICE.ch onto biological and bioactive compounds within the database, we predicted around 1.6 million potential biotransformations between bioDB compounds. Another 3.6 million reactions were found to connect bioDB compounds with molecules only found in chemical databases, producing a total of 5.2 million predicted reactions. From these predictions, we characterized the connectivity and reactivity of biologically important molecules, and we showed that ATLASx pathway predictions could recover 97.5% of known biological pathways from MetaCyc. Finally, we provide access to ATLASx through an online web interface that features tools for pathway design and network exploration, which can be accessed at https://lcsb-databases.epfl.ch/Atlas2. The ATLASx platform can be readily used for the design of novel metabolic pathways, and for the exploration and expansion of biosynthesis pathways. ATLASx provides an estimation on the staggering number of unknowns in biochemistry, and can thus foster future research explorations into metabolic dark matter.

Results

bioDB unifies over 1.5 million unique compounds

Current biochemical databases are heterogenous in their organization, biological scope, and level of detail, which complicates the comparison of data across databases. This in turn makes it difficult to reliably detect biochemical knowledge gaps, as information that is missing in one database may be present in another resource. To reconcile the heterogeneity of existing biochemical data sources, we unified compound data from a variety of biological and bioactive databases (KEGG [25], SEED [28], HMDB [29], MetaCyc [30], MetaNetX [31], DrugBank [32], ChEBI [33], ChEMBL [34]) (Supplementary Table S1). The unification resulted in a reference database, named bioDB, containing 1,500,222 unique 2D structural compound entries.

The contributions of each database to the total of bioDB compounds varied significantly. Three of the biological databases (KEGG, SEED, and MetaCyc) exclusively contain biological

compounds, and contributed 22,447 (1.5%) of the total compounds within bioDB. The remaining 1,477,775 (98.5%) compounds were contributed by databases covering all compounds produced by, or known to interact with, biological systems (i.e., bioactive compounds). Remarkably, the lion's share of these bioactive compounds (1,447,079, or 97%) came from ChEMBL, suggesting that this database has the most comprehensive definition of bioactivity. Based on our unification criteria, we found that the ratio of unique compounds varied across databases. The KEGG COMPOUND database had the highest number of unique compounds (80%, 15,064 unique compounds), while KEGG DRUGS had the lowest ratio (40%, 4,514 unique compounds), illustrating the heterogeneity of curation standards between different resources.

To pave the way for integrating chemicals into biochemical pathways, we imported all compounds from the chemical database PubChem [35] (77,934,143 unique molecules). PubChem entries that could not be matched to any existing compound in bioDB were assigned to the chemical compound space (chemDB) in our database, regardless of their true origin (i.e., chemical synthesis, natural biosynthesis, or semisynthetic procedure). Though some of the 77,934,143 unique compounds from the chemical space may still be of biological origin, they may not be labeled as such, or may have the potential to be derived from biological compounds in a bioengineering setting. This artificial classification of biological and chemical compounds presents an opportunity to re-assign compounds from the chemical space to the biological space by integrating them into hypothetical biochemical reaction.

56,000 unique biochemical reactions collected from 9 different sources

To create a unified reaction database as a reference for known metabolic processes, 235,698 reactions entries were collected from KEGG, BRENDA [36], Rhea [37], BiGG models [38], SEED [39], MetaNetX [31], MetaCyc [30], Reactome [40], and BKMS-react [41], and merged into 56,087 unique bioDB entries (Supplementary Table S2). Surprisingly, many databases contained a high number of duplicate reactions. We observed the highest ratio of unique reactions for KEGG (97%), followed by BKMS (95%). Reactome and BiGG had the lowest percentage of unique reactions (49% and 50%, respectively), indicating that many of the reactions are structural duplicates. This quantitative assessment of reaction uniqueness further exposes the

heterogenous nature of biochemical databases, suggesting that the number of entries provided by the database hosts should be handled with care when comparing databases, and highlighting the importance of quality assessment. We therefore checked whether or not the collected reactions were elementally balanced and associated with an EC number.

We first searched for reactions containing undefined or un-processable molecular structures (e.g., polymers, proteins, compounds describing two or more disconnected structures such as salts) and other reactions that were not elementally balanced (mostly missing reaction participants, or their reaction mechanism is not known). We found that 56% (31,711 out of 56,087) of total reactions were well-balanced (Supplementary Table S3), and that 46% (25,651) of the reactions had an assigned EC number. The highest ratio of balanced reactions having an EC number assigned was found in the Brenda and KEGG database (80% and 67%), while BiGG had the lowest ratio (21%). The comparatively high number of unbalanced reactions in collections of genome-scale models is partially explained by gap-filling efforts, where hypothetical, unbalanced reactions are added to the metabolic network to ensure model feasibility. This example illustrates how different applications require different levels, scopes and curation standards, resulting in heterogeneous data collections. Our unification and quality assessment of biochemical databases provides an overview on accumulated biochemical data, and it compares curation standards, biochemical coverage, overlaps and consistency across different resources. The resulting unified biochemical space (bioDB) forms the basis for the subsequent expansion of biochemical knowledge through reaction prediction.

Reactive sites detected in all biological and almost all bioactive database compounds

Functional groups, or reactive sites, designate which parts of a molecule are recognized and transformed by enzymes, and are important features of biochemically active compounds. To determine the biochemical reactivity of our collected biological and bioactive compounds, we applied 490 reaction rules from BNICE.ch to search for reactive sites within the 1,500,222 biological and bioactive compounds in bioDB, excluding those with more than one disjoint molecular structure (e.g., salts). We found that 1,498,307 out of 1,500,222 (99.8%) of collected biological and biochemical compounds had at least one reactive site (Supplementary Figure S1a). The remaining 1,915 compounds do not seem to have the

biochemical capacity to participate in any enzyme-catalyzed reaction, but their presence in biological databases can still be justified through their interaction with living organisms (e.g., chemically synthesized molecules with medical or research applications, Supplementary Figure S1b).

The number and types of reaction rules assigned to a compound is as an indicator for the diversity of functional groups, or the biochemical versatility, of the molecule. By screening our database of biological and bioactive molecules for reactive sites, each compound was assigned a list of reaction rules that can recognize one or more reactive sites on the molecule, thus characterizing its biochemical reactivity (Supplementary Figure S2). We showed that almost all molecules in bioDB have the potential to undergo biochemical transformations.

ATLASx predicts 5.2 million novel, hypothetical reactions

To explore the hypothetical biochemical space and map metabolic dark matter, we predicted novel reactions from biological and bioactive compounds within bioDB. Hypothetical reactions were predicted by applying 490 bidirectional reaction rules from BNICE.ch to the 1,498,307 compounds with at least one reactive site in bioDB. The reaction rules generated known and novel reactions that produced both known and novel product molecules. Reactions whose products were part of the assigned biological and bioactive compounds space were stored in the bioATLAS data collection, and reactions for which at least one product was only found in the chemical compound space were stored in chemATLAS. In total, we reconstructed 11,172 of the metabolic reactions in bioDB, and we predicted 5,236,833 novel reactions from biological and bioactive compounds (Table 2). Out of these reactions, 1,590,057 (30%) occurred exclusively between biological and bioactive compounds (bioATLAS), and the remaining 3,646,776 reactions involved at least one compound from the chemical space (chemATLAS). The Gibbs free energy of reaction was estimated for 81% of predicted reactions, and all predicted reactions were assigned a third-level EC number as defined by the BNICE.ch reaction rule that generated the reaction, and therefore have defined mechanism. Additional analyses showed no significant difference in the distribution of the estimated Gibbs free energy of reaction between known (bioDB) and predicted (chemATLAS) reactions (Supplementary Figure S4). Predicted reactions within bioATLAS integrated 56% (844,316 out of 1,500,222) of bioDB compounds. From the remaining

655,906 bioDB compounds, an additional 163,460 were integrated in at least one chemATLAS reaction, showing the importance of integrating chemical compounds into biochemical reaction network prediction. These results are even more important considering that 1,485,324 compounds in bioDB are not involved in any known reaction, and thus orphan. 67% (992,874) of orphan compounds could be integrated into at least one novel, predicted reaction in ATLASx. We further found that 863,000 compounds, originally only present in PubChem, could be integrated into the predicted chemATLAS network, meaning that these molecules are situated only one reaction step away from a known biological or bioactive compound. This ability to link orphan and chemical compounds to known biochemistry highlights the utility of our tool in drug discovery and metabolic research. These compounds are potential candidates for secondary metabolites (e.g., plant natural products), unwanted products of side reactions (e.g., damaged metabolites [42]), or bioactive compounds (e.g., drugs, pesticides) with the capacity to be transformed by enzyme catalysis.

Network analysis of the biotransformation reveals disjoint components

The connectivity of a biochemical reaction network can help identify missing biochemical links and serve as an indicator for the comprehensiveness of a knowledge base. According to the chemical law of mass conservation, a network that represents perfect biochemical knowledge would be fully connected—every compound would be connected to every other compound through a suite of biotransformations. To assess the connectivity of our reaction networks, we compared network properties of bioDB, bioATLAS, and chemATLAS using graph theory. Graph theory has been previously used to analyze the properties of biochemical networks, but these analyses were either performed on specific organisms or restricted to single databases [43,44]. Here, we estimated the network properties of known and expanded biochemistry through state-of-the-art graph-theoretical metrics.

We created graph-representations of each network, where nodes represent compounds, and edges represent biotransformations between two compounds (detailed in the Methods section). For each of the three networks, we counted the number of connected components (i.e., disjoint graphs or islands) (Figure 2a), and we found that the total number of components increased with the network expansion from bioDB to bioATLAS to chemATLAS

(Supplementary Table S4). However, the number of components relative to the size of the network, represented by the average number compounds per component, decreased from 22.6 in bioDB to 9.0 in bioATLAS, and increased again to 12.2 in chemATLAS. This result suggests that disconnected islands created by integrating bioactive compounds become more interconnected after including chemical compounds. To further characterize the networks, we looked at the size distribution of the different components. We found that all three networks were dominated by one single giant component followed by a large number of secondary components that involved at most 515 compounds (Figure 2b). While the biggest component in bioDB connected 88% of compounds in the network, this number decreased to 58% in bioATLAS and increased again to 68% in chemATLAS. This result supports our first observation, as it indicates that the integration of chemical compounds into the network bridges compound islands in bioATLAS into a denser network in chemATLAS. Our hypothesis is further confirmed by the diameter metrics. To calculate the diameter of a network, one needs to find all the shortest paths between all the possible combination of nodes in the network. The longest shortest path is called *diameter* of the network, and the average length of shortest paths between any two nodes is called the *average path length.* Here, we found that the diameter decreased with the expansion from bioDB (32 steps) to bioATLAS (27 steps), and increased again in chemATLAS (34-step diameter), suggesting expansion of the network towards novel chemistry and integration of previously disconnected components. The average path length increased monotonically from 7 in bioDB to 9 in bioATLAS, and to 12 in chemATLAS, reflecting the continuous growth of the main component (i.e., the core of biochemical reaction knowledge) during the network expansion process.

Searching for biological pathways within ATLASx

The quest for novel biochemical pathways is a key challenge in the bioproduction of natural and chemical compounds in chassis organisms, the elucidation of complex natural product biosynthesis, and the study of biodegradation routes. Here, we employed the algorithm implemented in NICEpath to search for pathways within biochemical reaction [45].

To benchmark our pathway search, we examined whether ATLASx could recover known pathways from MetaCyc, a collection of metabolic pathways. Out of the 3,149 collected

MetaCyc pathways, 1,518 matched our curation standards (see Methods). For each MetaCyc pathway, we determined the precursor and the target compound. We then used the pathway search to extract the top 100 pathways connecting the precursor and the target in ATLASx, and we compared the extracted pathways to the original MetaCyc pathway. We were able to find pathways for 99% (1,508 out of 1,518) of precursor-target pairs (Figure 3). The MetaCyc pathway reconstructions fall into three distinct categories: (i) exact reconstruction within top 100 (i.e., the exact MetaCyc pathway was found within the top 100 pathways extracted from ATLASx), (ii) exact reconstruction beyond top 100 (i.e, all the required biotransformations were present in the ATLASx network, but the original MetaCyc pathway was not found within the top 100 pathways) and (iii) alternative reconstruction (i.e., pathways connecting the precursor to the target pass via alternative intermediates). Out of the reconstructed MetaCyc pathways, 46% were exactly reconstructed within the top 100 pathways, 52% beyond the top 100, and 2% via alternative intermediates, demonstrating that despite the remarkable addition of millions of predicted reactions, the pathway search reliably recovered known pathways.

About 80% of reactions in bioDB do not have their reaction mechanism described in BNICE.ch, for which there are two main reasons: (i) the reaction mechanism is unknown or (ii) the reaction mechanism has not yet been added to the BNICE.ch reaction rule collection. Reactions missing a mechanism are usually also missing further annotations such as protein sequence and cofactor usage. By excluding reactions *without* known BNICE.ch mechanisms from the pathway search, we only find pathways that guarantee a known biochemical reaction mechanism in each step. By repeating the MetaCyc pathway reconstruction using only ATLASx reactions with known BNICE.ch mechanisms, we still found connections between precursor and target compounds for 1318 out of 1518 pathways (87%) (**Figure 3**). Out of the reconstructed pathways, 40% were exactly reconstructed within the top 100 pathways, 18% were found beyond the top 100 pathways, and for the remaining 42% pathways only alternative intermediate sequences were found. This analysis shows that known biochemical pathways can be diligently reproduced using ATLASx, and if reactions without known mechanism are excluded, more routes passing through alternative sequences of intermediates are found. The pathway search tool is available online at

https://lcsb-databases.epfl.ch/Search2. Users can adjust the network scope as discussed above, as well as perform database-specific search scopes for all of the imported reaction databases.

ATLASx fills metabolic gaps and proposes new biosynthesis pathways

ATLASx is a resource with wide-ranging practical applications that include compound classification, metabolic pathway searches, and gap-filling for metabolic models. To illustrate how the unification and expansion of knowledge in ATLASx can inform hypotheses around a given biochemical pathway, we used the ATLASx web tools to explore and expand a biosynthetic pathway of interest. As a case study, we chose the biosynthesis pathway of the anti-fungal and anti-hypertensive compound staurosporine, a secondary metabolite with a complex molecular structure (Figure 4) [46]. According to KEGG, the biosynthesis of staurosporine from tryptophan involves 7 reaction steps, most of them being poorly characterized. To explore the biochemical vicinity of this pathway, we retrieved all compounds that were one step away from the original pathway intermediates. Out of 861 potential pathway derivatives, 60 were found exclusively in bioDB, 407 were contributed by bioATLAS, and the remaining 394 compounds were only integrated when chemATLAS reactions were considered. According to our analysis, most derivatives (93%, or 799) were detected around tryptophan. Secondary hubs were found around K-252c and staurosporine, with each hub contributing 3% (24 compounds) and 3% (22 compounds) to the total number of pathway derivatives, respectively. Intrigued by the high number of potential staurosporine derivatives, we explored the network around this molecule and found 58 derivatives within a distance of four reaction steps (Figure 4). We found 6 staurosporine derivatives within bioDB (4 of them part of the original pathway), 18 derivatives only within bioATLAS, and an additional 34 compounds within chemATLAS. Interestingly, the network exploration converged, and the only derivatives found four steps away from staurosporine were located upstream of the original pathway.

To characterize the potential staurosporine derivatives we identified, we retrieved the number of patents and citations associated with the compounds in the extracted network, and we ranked the compounds by the sum of patents plus citations. These metrics have been previously established to extract compounds with high industrial, pharmaceutical and

academic interest [47]. Among the four top-ranked compounds in the network, we found that staurosporine itself garnered the most attention (29,819 patents plus 15,439 citations), followed by 7-hydroxystaurosporine (1,521 patents, 571 citations) and then K-252c (39 patents, 736 citations), which is part of the staurosporine synthesis pathway. Midostaurin, a cancer therapeutic and protein kinase inhibitor commercially known as Rydapt [48] (158 patents, 570 citations) and was one step away from staurosporine. While "popularity" estimation of chemicals is one of many possible ways to organize and rank compounds, other metrics such as molecular weight, toxicity, or structural similarity to known pharmaceuticals may be applied depending on the objective of the study. This example illustrates how to explore the biochemical vicinity of a compound or pathway within ATLASx, and to retrieve information (e.g., citations and patents) from external sources to rank derivatives based on their academic and industrial relevance.

We further investigated the capability of ATLASx to detect and bridge knowledge gaps in the biosynthesis pathway of staurosporine. Out of the 7 reaction steps in the pathway obtained from KEGG, only one reaction is linked to an enzyme. The other 6 reactions are orphan (i.e., no enzyme assigned), and half of these orphan reactions have unknown reaction mechanisms (Table 3). To show how one can find plausible enzymes(s) for orphan reactions, we examined each orphan reaction within the pathway. First, for reactions with assigned BNICE.ch reaction rules, we applied the computational tool BridgIT to predict potential catalyzing enzymes. For reaction steps without an assigned BNICE.ch rule, we searched for pathways that connect the substrate to the product via alternative sequences of well-annotated bioDB reactions, or novel BNICE.ch reactions that provide the basis for robust enzyme predictions with BridgIT.

The first step of the pathway, the conversion of L-tryptophan to IPA imine, is identified with the partial EC number 1.4.3.- by KEGG. The computational tool BridgIT, which uses enzyme annotations of non-orphan reaction to suggest enzymes catalyzing similar orphan reactions, proposed the enzyme 7-chloro-L-tryptophan oxidase (EC 1.4.3.23) as the best candidate to catalyze this first step. This predication was bolstered by a high BridgIT score of 0.95, which indicates that both substrates have a similar reactive site and surrounding structure. While the native function of this proposed enzyme is to convert 7-chloro-L-tryptophan to 2-imino-

3-(7-chloroindol-3-yl) propanoate, the activity of this candidate enzyme on L-tryptophan has been proven in a study by Nishizawa *et al.* [49]. Another orphan reaction in the staurosporine pathway is the conversion of 3'-demethylstaurosporine to O-demethyl-N-demethyl-staurosporine (step 6). For this reaction, BridgIT suggested that an N-formiminotransferase serves as a catalyzing enzyme (EC 2.1.2.5), although this prediction is accompanied by a relatively low BridgIT score of 0.34. Finally, the last step of the pathway is known to be catalyzed by an O-methyltransferase with EC number 2.1.1.139. In this case, BridgIT successfully mapped this reaction to itself and found the corresponding enzyme. This showcase exemplifies how BridgIT can be used on top of the BNICE.ch reaction prediction to find enzymes for novel or orphan reactions and to fill gaps in metabolic pathways and networks.

Finally, all of the presented analyses can be performed using the computational tools available online, in combination with open-source visualization software as detailed in the Methods section. We provide public access to our database through an online search interface, which includes a powerful pathway search algorithm that can be used for the design of novel metabolic pathways. The web access to ATLASx (https://lcsb-databases.epfl.ch/Atlas2) provides further query tools, such as the ability to identify all reactions associated to a query compound.

Discussion

This work attempts to map the hypothetical vicinity of known biochemistry and to address the vast amount of metabolic "dark matter" by using biochemical reaction principles implemented in 490 generalized reaction rules. Based on 1.5 million known biological and bioactive compounds unified in bioDB, we predicted 1.6 million biochemically possible biotransformations between biological and bioactive compounds (bioATLAS). We further predicted more than 3.6 million reactions that involved compounds from the chemical compound space, resulting in a total of almost 5.2 million in chemATLAS. From this new wealth of information, we extracted insightful numbers on the reactivity and connectivity of biologically relevant molecules.

Assessing the composition of metabolic "dark matter" is by definition difficult, since we lack a way to quantify the unknowns a priori. Fortunately, biochemical data collected and generated from our database allows us to answer a broad range of questions regarding the biochemical reactivity of compounds, the expansion of biochemical space from a graph-theoretical perspective, and the characteristics of our hypothetical reaction network. Potential applications of ATLASx include the prediction of bioproduction or biodegradation pathways involved in the transformation of commodity and specialty chemicals, pharmaceuticals, and plastics. ATLASx can also be used to discover the biosynthesis routes of poorly characterized secondary metabolites, and systematically fill in knowledge gaps surrounding metabolic models. For example, ATLASx can be used to expand the network around all compounds within a given metabolic model, remove dead-end metabolites, and then examine the new, expanded model for potential shortcuts, enhanced predictions, and enzymatic promiscuity.

Since we successfully integrated tens of thousands of chemical compounds into a biochemical network, we hypothesize that many compounds are not yet part of any database, even though they potentially exist in nature or could be created by metabolic engineering. While the integration and accurate prediction of hypothetical compound structures remains an open challenge, ATLASx provides the necessary tools and conceptual framework to predict hypothetical compounds reliably in the future. In order to properly meet that future, ATLASx is designed as a dynamic database, and can be continuously expanded around biochemical pathways or compound classes of interest. We believe that predictive biochemistry is crucial for the advancement of synthetic biology and metabolic engineering, and hope that ATLASx can provide reliable reaction and pathway predictions for the scientific community.

Methods

We developed a unification algorithm based on canonical SMILES to generate consistent and duplicate-free data. Data was collected from KEGG [25], SEED [28], HMDB [29], MetaCyc [30], MetaNetX [31], DrugBank [32], ChEBI [33], ChEMBL [34], BRENDA [36], Rhea [37], BiGG models [38], Reactome [40], and BKMS-react [41] by January 2019. To efficiently store, retrieve, and analyze the increased amount of data, we created an SQL-based database where we imported the collected data (bioDB). The compounds were filtered to keep only those that had a defined, single molecular structure (i.e., no salts, no polymers, no generic molecules with undefined branches). Compounds were merged using their canonical SMILES as a unique identifier, which was calculated using openbabel [50] version 2.4.0. As a result of the unification procedure, a unique compound entry in bioDB can contain different resonance forms, stereoisomers, as well as dissociated and charged states of a same compound.

Reactions were unified based on the structure of their reactants and products. Transport and stereoisomeric reactions were filtered to only keep reactions that modify the connectivity of atoms within the molecule. Unified compounds and reactions were annotated with all available identifiers from different databases

Curation of unified compounds and reactions

Compounds were annotated with the following structural descriptors: chemical formula, molecular weight, InChI, InChIKey. For the annotation we used openbabel [50] version 2.4.0. Reactions were annotated with enzymatic reaction mechanisms obtained from BNICE.ch. We further estimated the standard Gibbs free energy of the reactions in cellular conditions using the Group Contribution Method (GCM) [51].

Reactive site analysis

We used BNICE.ch to determine the reactivity of compounds. For this, we applied all of the 490 bidirectional generalized reaction rules available in BNICE.ch to the unified set of compounds. BNICE.ch reaction rules rely on a bond-electron matrix (BEM) representation of the reactive site that will be recognized by the enzyme. A second matrix (difference BEM) describes the bonds that need to be rearranged in the molecule to form the product. To

determine the potential reactivity of the compounds, we therefore used the first BEM describing the reactive site of each reaction rule to screen all of the molecular structures of the compounds for potential reactive sites. Each compound was assigned all the rules that harbor a matching reactive site for further usage.

For compounds with aromatic rings, all possible kekulé representations were generated to ensure that no potential reactive site is missed. Since the double bonds in aromatic compounds can be drawn in different positions, giving rise to different mesomers, it is important to screen all the different kekulé structures for reactive sites. It has been shown that different kekulé structures lead to different results in cheminformatics approaches like docking, molecular-fingerprint construction, clustering, and any 2D- or 3D-QSAR analysis [52].

Prediction of novel reactions

Each BNICE.ch reaction rule was applied to all compounds recognized by the rule in the previous step, and all the generated products were analyzed. Reactions only producing compounds that belong to the biological or biochemical compound space were imported to the database as bioATLAS reactions, and reactions involving products from the chemical space were imported as chemATLAS reactions. Importing a generated reaction into the database involved checking its equation against all the reactions present in the database. If a reaction was already present, only the reaction rule was added to the reaction description, otherwise the reaction was imported. For each generated reaction, the Group Contribution Method (GCM) [53], as integrated in BNICE.ch, was employed to provide an estimation of Gibbs free energy of reaction and its error.

Network analysis

Most available methods for network analysis are based on manually derived reactant-product pairs (e.g., KEGG RPAIR network [54]), or they define a set of cofactors to be excluded from the analysis to avoid the generation of hubs by currency metabolites. Here, we rely on NICEpath, a pathway search tool proposing a graph representation of metabolic reactions that weighs each substrate-product pair according the number of atoms conserved between the substrate and the product [45]. NICEpath has been previously shown to be able to retrieve

relevant pathways from biochemical networks, and to help in understanding the inherent properties of biochemical networks (e.g., connectivity, diameter).

To calculate the weights on each pair, NICEpath requires that each reaction is annotated with a reaction mechanism that allows the calculation of the atom conservation between the substrate and the product. This condition is met for all bioDB reactions with assigned reaction rules, and for all predicted reactions in bioATLAS and chemATLAS. For reactions in bioDB without reaction mechanism assignment, the atom conservation was estimated by assuming a minimal rearrangement of atoms in the course of the reaction.

For each pair, the conserved atom ratio (CAR) is calculated with respect to the reactant (CAR_r) and with respect to the product (CAR_p), where n_c is the number of conserved atoms between the reactant and the product, n_r is the number of atoms in the reactant, and n_p is the number of atoms in the product. Hydrogen atoms are excluded from the calculation.

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

To calculate a bidirectional CAR, the mean $CAR_{r,p}$ is multiplied with a correction factor that increases with the difference between the number of common atoms and the total number of atoms in the molecule.

$$CAR = \frac{CAR_r + CAR_p}{2}$$

For bioDB reactions that could not be annotated with a BNICE.ch reaction mechanism, n_c was estimated as follows: First, the number of conserved atoms was assumed to be defined for standard cofactors (e.g., NAD^+ and NADH conserve 100% of atoms, hydrogen excluded). If no pre-defined cofactor pair could be identified, we estimated the number of common atoms using the maximum common substructure algorithm (implemented as FMCS library in RDKit). For the remaining pairs, the number of conserved atoms was calculated based on the assumption that the maximum possible number of atoms is conserved between a given

substrate-product pair (e.g., C_6O is conserved between C_7O and C_6O_2, hydrogens excluded). For the remaining reactant pairs, the atom conservation was assumed to be equal to 0.
To create a weighted, searchable biochemical network, the CAR of substrate-product pair is then transformed into a distance between the substrate node and the product node, and the two are connected by an edge representing the biotransformation. Since a same biotransformation, or reactant-product pair, can occur in more than one reaction, each edge in the network represents all the reactions that transform the substrate to the product. The proposed method allows the construction of atom-weighted biochemical networks to find metabolic pathways using common graph-search algorithms. Moreover, the atom-weighted biochemical networks can be analyzed to derive global properties of known biochemistry.

We constructed three networks with different biochemical scopes: The 56,087 reactions in bioDB translated into 62,299 weighted edges connecting 14,914 bioDB compounds. The bioATLAS network connects 844,337 compounds in 1,590,057 reactions, represented by 2,503,627 edges in the network, and chemATLAS connects 1,876,992 compounds in 5,248,711 reactions, represented by 5,717,409 edges (Supplementary Table S1). For many types of network analysis, however, an unweighted graph is required. Since we know that substrate-product pairs with a very low degree of atom conservation may not be biologically relevant, we used a cutoff CAR value to decide whether or not to draw an edge between any given compounds. It has been shown previously that a cutoff of 0.34 in the conserved atom ratio best predicts the manually curated reactant pairs of type "main" in KEGG [45,54] . Hence, we removed edges with a CAR below 0.34, and we removed the weights from the remaining edges. The result is a new set of unweighted graphs, which can be analyzed using standard graph analysis algorithms (i.e., extraction of disjoint components, diameter calculations on disjoint components).

Reconstruction of known reactions with BNICE.ch reaction rules

To estimate the biochemical coverage of BNICE.ch reaction rules, we assessed how many bioDB reactions could be reconstructed by BNICE.ch rules. For this, we first removed reactions missing structural information on their reactants, as well as isomerase and transport reactions. For the remaining reactions, we checked if the reaction could be

reconstructed by one or several BNICE.ch reaction rules. We distinguished between exact reconstruction, where reactants and products of the reaction match exactly, reconstruction of the main biotransformation using alternative cofactors, and reconstruction of the main biotransformation using 2 to 4 reaction steps with BNICE.ch rule within ATLASx. For multi-step reconstruction of reactions, we considered all main reactant pairs (i.e., with a CAR >= 0.34) of a given reaction, and we searched to connect the substrate of the pair to the product. To do this, we extracted the shortest path within the weighted reaction network of chemATLAS, considering only BNICE.ch curated reactions. If the substrate and the product could be connected in k reaction steps, and the product of the CARs along the path did not drop below the threshold of 0.34, we called the reactant pair *reconstructed in k steps*. If the reaction could be split into several "main" reactant pairs, the pair reconstruction with the highest number of steps was considered for the reaction reconstruction. Using this workflow, we found that 72 % of filtered reactions in bioDB could be reconstructed by BNICE.ch (Supplementary Table S5).

Pathway reconstruction of linear pathways from MetaCyc

MetaCyc pathways and reactions were downloaded from MetaCyc (https://metacyc.org/smarttables) on May 25, 2020. The pathways used for benchmarking had to pass the following criteria: The pathway (i) consists of a minimum of two reactions, (ii) it does not contain transport reactions and electron-transfer reactions, (iii) it does not contain reactions that are not listed in the MetaCyc reactions table, (iv) it does not contain compounds with undefined structure, non-carbon compounds, proteins and peptide polymers, RNA molecules, unknown compounds, (v) it is not circular (as we are comparing linear pathways), (vi) is not a polymerization pathway (e.g. bacterial peptidoglycan polymerization), (vii) is not a light-dependent pathway, (viii) is not a superpathway (i.e., pathway consisting of other pathways with no individual unique reaction sequence).

The pathways in MetaCyc are reported as a set of reactions and not as a sequence of compounds. We therefore translated the pathway of reactions into a pathway of compounds to directly compare the output. For these, we generated a graph based on the reaction equations: all the compounds of the reactions were considered as nodes, and

biotransformations were introduced into the graph as edges for every substrate-product pair with the same distance. This is because we do not have information about the atom conservation from MetaCyc's sources. At the same time, we excluded the common cofactors (NAD, ATP, etc.) from the pathway graph to minimize the possible number of shortcuts. We found the shortest loop-less linear path between every pair of compounds in this graph, and we considered the longest of these pathways (therefore the 2 more distant compounds) as the linear pathway of reference (e.g. longest branch of the branching pathway). The resulting linear set of pathways was considered as a reference pathway to benchmark the pathway search algorithm. For the pathway search we applied the following parameters: as the pathway search scope we used two different networks, one involving all ATLASx reactions, and one only including reactions that have a BNICE.ch reaction mechanism assigned. The analysis of respective differences for network with and without BNICE.ch mechanism annotation for all reactions provided in the main text. We restricted the maximum number of pathways to 100 and maximum number of steps in a pathway to 30. Edges with a CAR value below 0.34 were removed from the network, and the distance between nodes was calculated as follows, including a bias of a factor 100 favoring known biotransformations in the pathway search

$$\text{if (known reaction) } \{ \, distance = \frac{e^{1/CAR}}{e} \, \}$$

$$\text{else } \{ \, distance = 100 * \frac{e^{1/CAR}}{e} \, \}$$

We discovered that by excluding reactions without known mechanisms from the network when all required biotransformations were present in the network, we increased the chance of finding the correct pathway within the top 100 pathways. Here, the pathway search exactly reconstructed 524 out of 760 pathways (69%) for which all the necessary biotransformations were present. In the original network that included all bioDB reactions (with incomplete reaction mechanisms), we only found 700 out of 1480 pathways (47%) that had all biotransformations present. We also investigated how the length of the reference pathway from MetaCyc affects the proportion of reconstructed pathways. We found that even pathways as long as 16 reaction steps could be exactly reconstructed from the original MetaCyc pathway, and that alternative pathways for MetaCyc pathways could be up to 26 reaction steps in length (Supplementary Figure S3). These results show that the performance of the pathway search is not significantly compromised when searching for longer pathways.

Pathway expansion

The biosynthesis pathway of staurosporine was obtained from KEGG (https://www.kegg.jp/kegg-bin/show_pathway?map=map00404&show_description=show, 12 May 2020). For each intermediate, the LCSB ID was obtained using the online compound search interface, and the IDs were compiled into a compound list. The ID list was then used as input to extract all compounds one reaction step away from the native pathway for bioDB, bioATLAS, and chemATLAS scopes (https://lcsb-databases.epfl.ch/Atlas2/Analysis, Analysis 1). The default values were applied (CAR threshold of 0.34, reactions without known BNICE.ch reaction mechanism included). The resulting network file for chemATLAS was imported to the open-source graph visualization software Gephi (.csv file extension required for import). To assign the origin to each compound in the network, the list of nodes was obtained from the bioDB and bioATLAS networks, and each node in the chemATLAS network was assigned the corresponding scope. The node table was imported into the Gephi project, and nodes were colored in accordance to their scope. Finally, Gephi's forceAtlas2 algorithm was applied iteratively to the network and the visualization was manually improved.

To assess the "popularity" of the staurosporine derivatives, the number of publications was derived from PubChem and PubMed, while the number of patent annotations was extracted from

PubChem. The PUG-REST service was used to retrieve information from the PubChem website (https://pubchem.ncbi.nlm.nih.gov/) on the number of patents and citations associated to each compound. We further used the Entrez Programming Utilities (E-utilities) API service to search the PubMed database for citations by compound name. To find enzymes for the novel, predicted reaction, the enzyme prediction tool BridgIT was used [26]. BridgIT is available as a resource on our website (https://lcsb-databases.epfl.ch/Bridgit).

Website

The ATLASx search and analysis tools can be found at https://lcsb-databases.epfl.ch/Atlas2. The visualization of compounds is achieved by SMILES viewer, a light-weight JavaScript library developed by Probst and Reymond [55] that visualizes molecular structures in runtime. The pathway search within known and predicted biochemical network uses the concepts and code developed in NICEpath [45]. The Python library NetworkX is used for the pathway search implementation, and the network statistics were obtained using the Python libraries NetworkX and SNAP [56].

References

1. Stanway, R. R. *et al.* Genome-Scale Identification of Essential Metabolic Processes for Targeting the Plasmodium Liver Stage. *Cell* **179**, 1112-1128.e26 (2019).
2. Lewis, N. E. & Abdel-Haleem, A. M. The evolution of genome-scale models of cancer metabolism. *Front. Physiol.* **4**, (2013).
3. Atanasov, A. G. *et al.* Discovery and resupply of pharmacologically active plant-derived natural products: A review. *Biotechnol. Adv.* **33**, 1582–1614 (2015).
4. Notebaart, R. A., Kintses, B., Feist, A. M. & Papp, B. Underground metabolism: network-level perspective and biotechnological potential. *Curr. Opin. Biotechnol.* **49**, 108–114 (2018).
5. Rosenberg, J. & Commichau, F. M. Harnessing Underground Metabolism for Pathway Development. *Trends Biotechnol.* **37**, 29–37 (2019).
6. Lerma-Ortiz, C. *et al.* 'Nothing of chemistry disappears in biology': the Top 30 damage-prone endogenous metabolites. *Biochem. Soc. Trans.* **44**, 961–971 (2016).
7. Oliver, S. G. From DNA sequence to biological function. *Nature* **379**, 597–600 (1996).
8. Galperin, M. Y. & Koonin, E. V. From complete genome sequence to 'complete' understanding? *Trends Biotechnol.* **28**, 398–406 (2010).
9. Keseler, I. M. *et al.* The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.* **45**, D543–D550 (2017).
10. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A. & Hatzimanikatis, V. ATLAS of Biochemistry: A repository of all possible biochemical reactions for synthetic biology and etmabolic engineering studies. *ACS Synth. Biol.* **5**, 1155–1166 (2016).
11. Bachmann, B. O. Biosynthesis: Is it time to go retro? *Nat. Chem. Biol.* **6**, 390–393 (2010).
12. Hadadi, N. & Hatzimanikatis, V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Curr. Opin. Chem. Biol.* **28**, 99–104 (2015).
13. Wang, L., Ng, C. Y., Dash, S. & Maranas, C. D. Exploring the combinatorial space of complete pathways to chemicals. *Biochem. Soc. Trans.* **46**, 513–522 (2018).
14. Lin, G.-M. M., Warden-Rothman, R. & Voigt, C. A. Retrosynthetic design of metabolic pathways to chemicals not found in nature. *Curr. Opin. Syst. Biol.* **14**, 82–107 (2019).
15. Jeffryes, J. G., Seaver, S. M. D., Faria, J. P. & Henry, C. S. A pathway for every product? Tools to discover and design plant metabolism. *Plant Sci.* **273**, 61–70 (2018).

16. Hatzimanikatis, V. *et al.* Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609 (2005).
17. Tokic, M. *et al.* Discovery and Evaluation of Biosynthetic Pathways for the Production of Five Methyl Ethyl Ketone Precursors. *ACS Synth. Biol.* **7**, 1858–1873 (2018).
18. Campodonico, M. A., Andrews, B. A., Asenjo, J. A., Palsson, B. O. & Feist, A. M. Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path. *Metab. Eng.* **25**, 140–158 (2014).
19. Ding, S. *et al.* novoPathFinder: a webserver of designing novel-pathway with integrating GEM-model. *Nucleic Acids Res.* **48**, W477–W487 (2020).
20. Kumar, A., Wang, L., Ng, C. Y. & Maranas, C. D. Pathway design using de novo steps through uncharted biochemical spaces. *Nat. Commun.* **9**, 184 (2018).
21. Sivakumar, T. V., Giri, V., Park, J. H., Kim, T. Y. & Bhaduri, A. ReactPRED: a tool to predict and analyze biochemical reactions. *Bioinformatics* **32**, 3522–3524 (2016).
22. Delépine, B., Duigou, T., Carbonell, P. & Faulon, J.-L. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metab. Eng.* **45**, 158–170 (2018).
23. Koch, M., Duigou, T. & Faulon, J.-L. Reinforcement Learning for Bioretrosynthesis. *ACS Synth. Biol.* **9**, 157–168 (2020).
24. Tyzack, J. D., Ribeiro, A. J. M., Borkakoti, N. & Thornton, J. M. Exploring Chemical Biosynthetic Design Space with Transform-MinER. *ACS Synth. Biol.* **8**, 2494–2506 (2019).
25. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
26. Hadadi, N., MohammadiPeyhani, H., Miskovic, L., Seijo, M. & Hatzimanikatis, V. Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 201818877 (2019).
27. Hafner, J., MohammadiPeyhani, H., Sveshnikova, A., Scheidegger, A. & Hatzimanikatis, V. Updated ATLAS of Biochemistry with New Metabolites and Improved Enzyme Prediction Power. *ACS Synth. Biol.* **9**, 1479–1482 (2020).
28. Overbeek, R. *et al.* The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
29. Wishart, D. S. *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **35**, D521-6 (2007).

30. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **46**, D633–D639 (2018).
31. Moretti, S. *et al.* MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* **44**, D523–D526 (2016).
32. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
33. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214 (2016).
34. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
35. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
36. Schomburg, I. *et al.* BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem. Sci.* **27**, 54–56 (2002).
37. Morgat, A. *et al.* Updates in Rhea--a manually curated resource of biochemical reactions. *Nucleic Acids Res.* **43**, D459-64 (2015).
38. Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11**, 213 (2010).
39. Aziz, R. K. *et al.* SEED Servers: High-Performance Access to the SEED Genomes, Annotations, and Metabolic Models. *PLoS ONE* **7**, e48053 (2012).
40. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
41. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.* **47**, D542–D549 (2019).
42. Linster, C. L., Van Schaftingen, E. & Hanson, A. D. Metabolite damage and its repair or pre-emption. *Nature Chemical Biology* vol. 9 72–80 (2013).
43. Wang, L., Dash, S., Ng, C. Y. & Maranas, C. D. A review of computational tools for design and reconstruction of metabolic pathways. *Synth. Syst. Biotechnol.* **2**, 243–252 (2017).

44. Arita, M. The metabolic world of Escherichia coli is not small. *Proc. Natl. Acad. Sci.* **101**, 1543–1547 (2004).
45. Hafner, J. & Hatzimanikatis, V. Finding metabolic pathways in large networks through atom-conserving substrate-product pairs. *bioRxiv* 2020.11.25.398453 (2020) doi:10.1101/2020.11.25.398453.
46. Omura, S., Sasaki, Y., Iwai, Y. & Takeshima, H. Staurosporine, a Potentially Important Gift from a Microorganism. *J. Antibiot. (Tokyo)* **48**, 535–548 (1995).
47. Hafner, J., Payne, J., MohammadiPeyhani, H., Hatzimanikatis, V. & Smolke, C. A computational workflow for the expansion of heterologous biosynthetic pathways to natural product derivatives. *Under revision*.
48. Kim, E. S. Midostaurin: First Global Approval. *Drugs* **77**, 1251–1259 (2017).
49. Nishizawa, T., Aldrich, C. C. & Sherman, D. H. Molecular Analysis of the Rebeccamycin l-Amino Acid Oxidase from Lechevalieria aerocolonigenes ATCC 39243. *J. Bacteriol.* **187**, 2084–2092 (2005).
50. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminformatics* **3**, 33 (2011).
51. Jankowski, M. D., Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. *Biophys. J.* **95**, 1487–1499 (2008).
52. Hähnke, V. D., Kim, S. & Bolton, E. E. PubChem chemical structure standardization. *J. Cheminformatics* **10**, 36 (2018).
53. Jankowski, M. D., Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* **95**, 1487–1499 (2008).
54. Shimizu, Y., Hattori, M., Goto, S. & Kanehisa, M. Generalized Reaction Patterns for Prediction of Unknown Enzymatic Reactions. *Genome Inform.* **20**, 149–158 (2008).
55. Probst, D. & Reymond, J.-L. SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript. *J. Chem. Inf. Model.* **58**, 1–7 (2018).
56. Leskovec, J. & Sosič, R. SNAP. *ACM Trans. Intell. Syst. Technol.* **8**, 1–20 (2016).
57. WEIDNER, S., KITTELMANN, M., GOEKE, K., GHISALBA, O. & ZÄHNER, H. 3'-Demethoxy-3'-hydroxystaurosporine-O-methyltransferase from Streptomyces longisporoflavus

Catalyzing the Last Step in the Biosynthesis of Staurosporine. *J. Antibiot. (Tokyo)* **51**, 679–682 (1998).

Tables

Table 2: Compound and reaction statistics for bioDB, bioATLAS, and chemATLAS.

	Property	bioDB	bioATLAS	chemATLAS
Compounds	Integrated in reaction	14,902	1,007,776	1,870,776
	Total number of compounds	1,500,222	1,500,222	77,934,143
Reactions	Known reactions	56,087	56,087	56,087
	Out of which BNICE.ch is curated	*11,172*	*11,172*	*11,172*
	Novel reactions	0	1,578,885	5,225,661
	Total number of reactions	56,087	1,634,972	5,281,748

Table 3: Pathway reconstruction and gap-filling of the staurosporine biosynthesis pathway with ATLASx

Step	KEGG ID	EC number	BNICE.ch rule	Top BridgIT hit EC (KEGG ID, score)	Reconstruction within ATLASx
1	R11119	1.4.3.-	1.4.3.-	1.4.3.23 (R09560, 0.95)	Biotransformation with LCSB ID 2600177067
2	R11120				2-step reaction (spontaneous + 1.21.98.2) in bioDB[a]
3	R11121	1.13.12.-			Not reconstructed within ATLASx
4	R11122	2.4.-.-			3-step reaction in chemATLAS[b]
5	R11123				2-step reaction in bioATLAS[c]
6	R11129		2.1.1.-	2.1.2.5 (R03189, 0.34)	Biotransformation with LCSB ID 2600423725
7	R05757	2.1.1.139[57]	2.1.1.-	2.1.1.139 (R05757, 1.00)	Biotransformation with LCSB ID 2600261843

[a] https://lcsb-databases.epfl.ch/Graph2/loadPathway/1/1468050408,1469435049,1468050416/2806125367,2806150968/0
[b] https://lcsb-databases.epfl.ch/Graph2/loadPathway/1/1468050425,1469288899,277921848,1468050433/2603459454,2603467379,2682146339/0
[c] https://lcsb-databases.epfl.ch/Graph2/loadPathway/1/1468050433,1469288674,1468050440/2603455158,2682148818/0

Figures

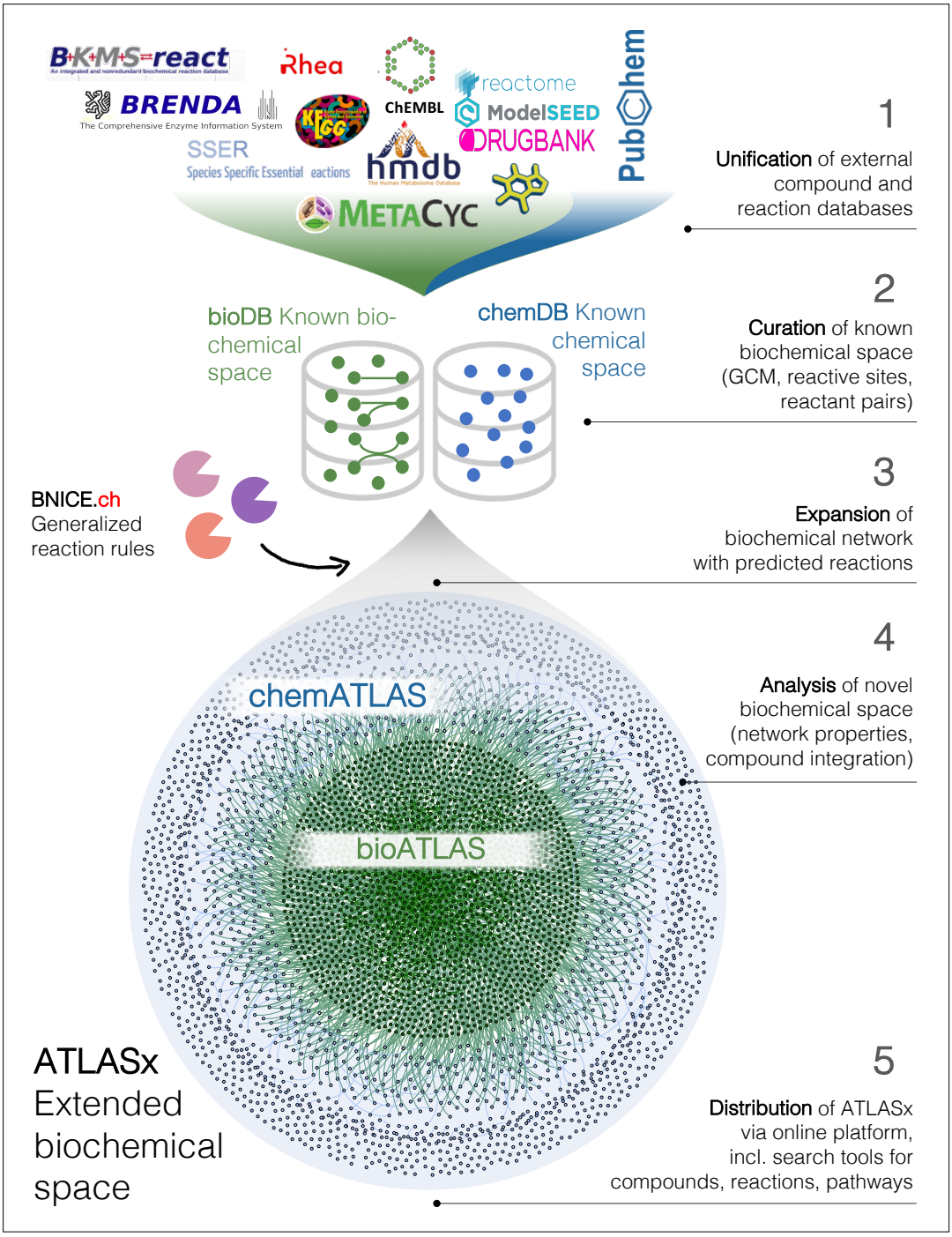

Figure *1*: ATLAS workflow applied to known biological and bioactive compounds. 1. Unification- collection of metabolic reactions and biochemical compounds from different publicly available databases, which were merged into a consistent and duplicate-free database, called bioDB. 2. Curation- compounds were annotated with molecular identifiers and reactions were annotated with reaction mechanisms. 3. Expansion- generalized reaction rules from BNICE.ch were applied to bioDB compounds to generate all possible reactions producing known biological or chemical products. 4. Analysis- the connectivity of the biochemical reaction networks was analyzed before and after reaction prediction, as well as the integration of compounds not previously connected in known biochemical networks. 5. Distribution- the results were made available on our website (https://lcsb-databases.epfl.ch/Atlas2).

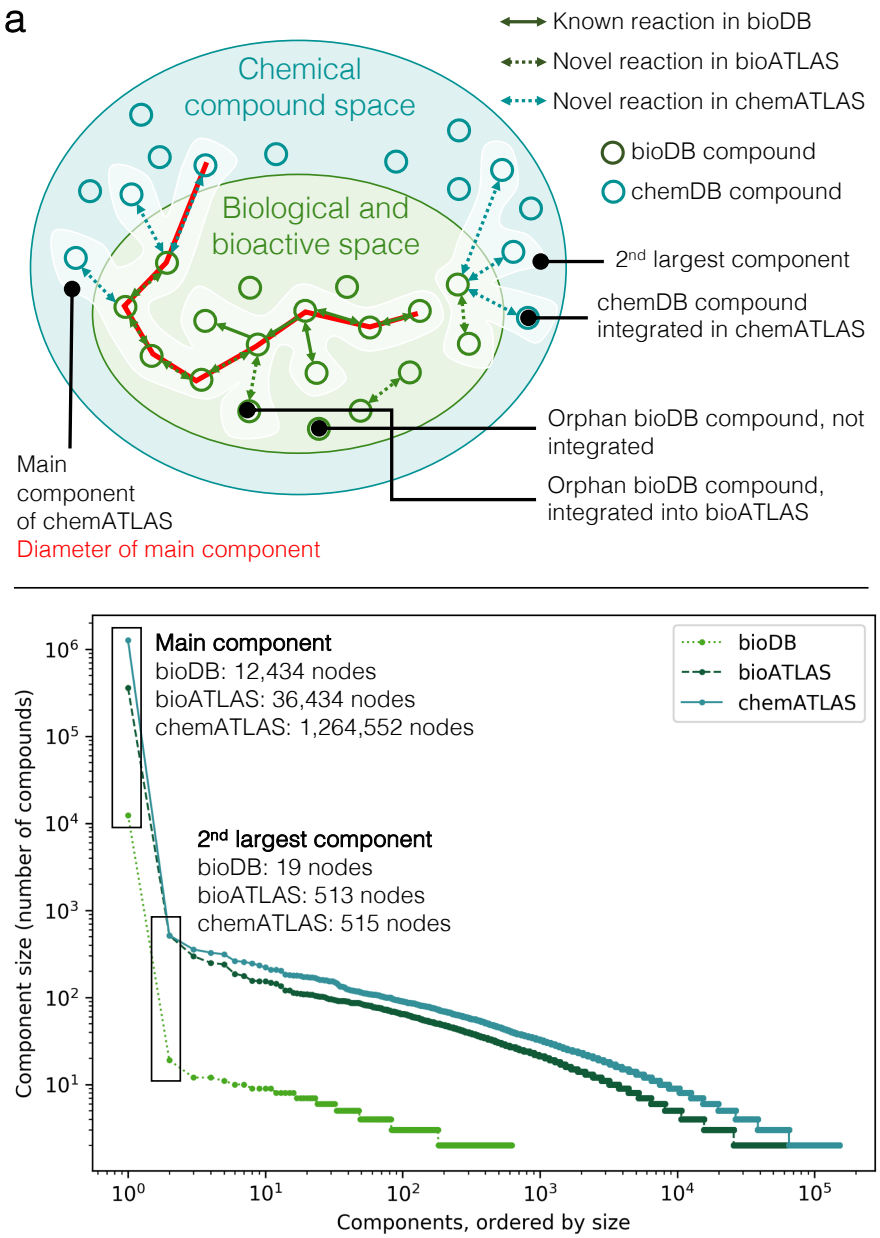

Figure 2: **Graph-theoretical analysis of biotransformation networks. a**, Schematic overview of statistics and network properties calculated for bioDB, bioATLAS, and chemATLAS. Reactions exclusively involving biological and bioactive compounds (green nodes) are assigned to the bioATLAS reaction space, and reactions involving one or more chemical compound (light blue nodes) are assigned to the chemATLAS reaction space. The main component and the second largest component of the network are schematically shown (white highlight). The diameter, or longest shortest path between any two nodes of the main component, has a length of 8 and is highlighted in red. **b**, The components of each reaction scope in ATLASx have been extracted and ordered by size. Here, the number of compounds (nodes) of each component is plotted on a log-log scale to show the size distribution of disconnected components for bioDB, bioATLAS, and chemATLAS. For the main component (highest number of nodes) as well as for the second largest component, exact numbers of compounds are indicated.

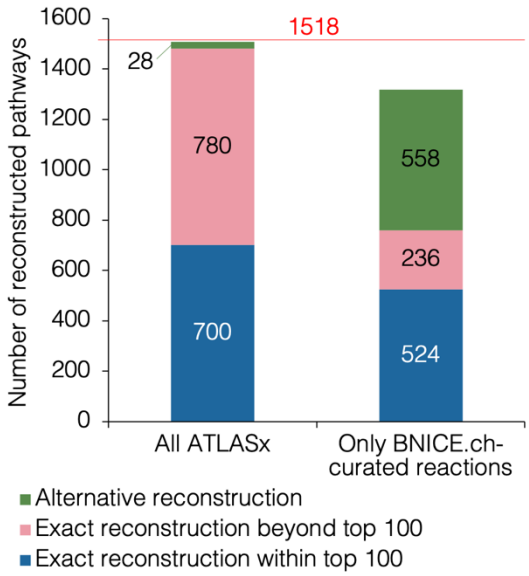

Figure 3: Pathway search comparison to dataset of pathways extracted from MetaCyc. Coverage of the collected MetaCyc pathways dataset (1518 pathways) with ATLAS pathway search tool.

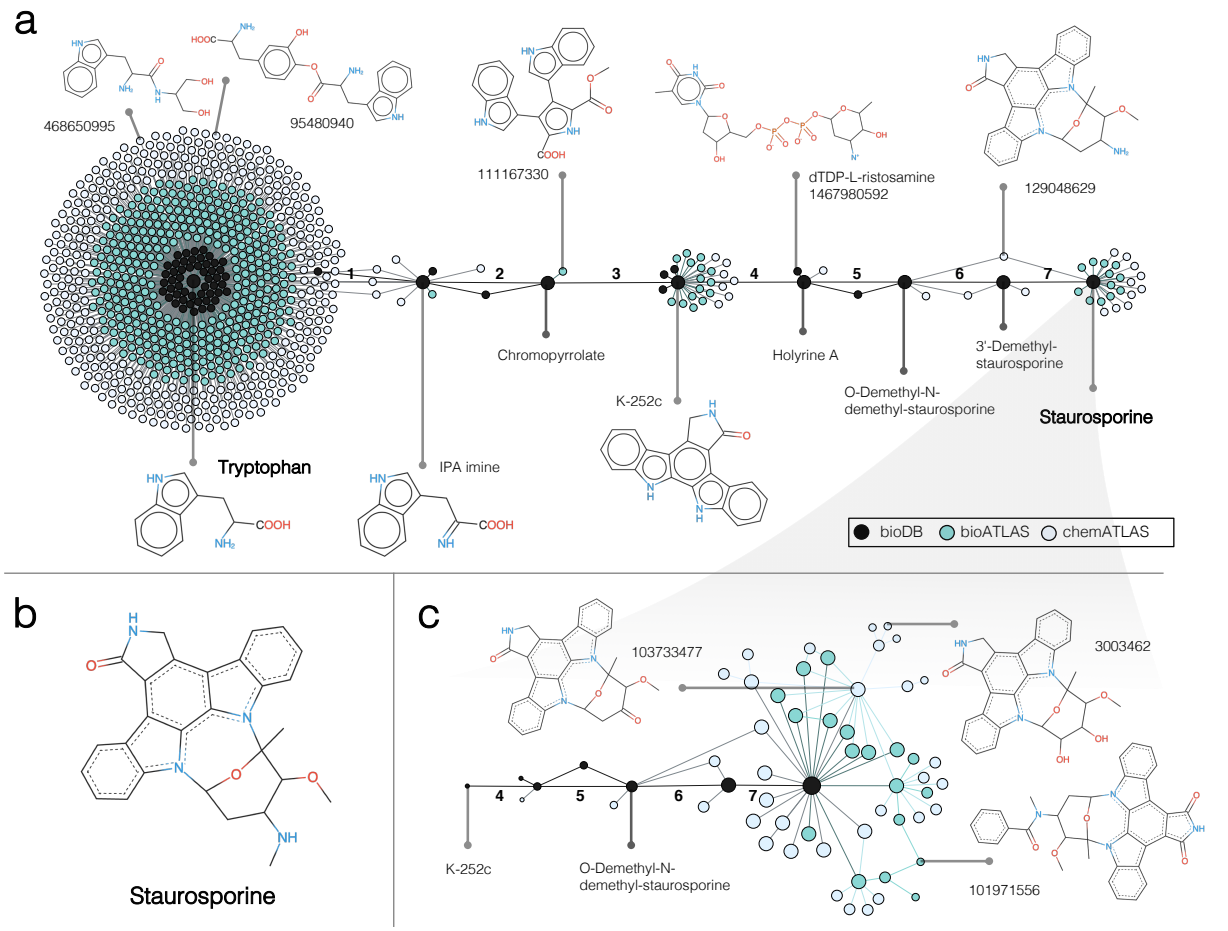

Figure 4: Example pathway expansion showcase for the biosynthesis of the natural product staurosporine. **a,** Visual representation of the biosynthesis pathway from tryptophan to staurosporine (obtained from KEGG, steps numbered in bold black) has been expanded for one generation around the native intermediates. **b**, The molecular structure of staurosporine. **c**, Network of potential staurosporine biochemistry expanded four generations around the target compound. The size of each compound nodes decreases with each generation.

Acknowledgements

We thank Dr. N. Hadadi for her creative inputs in an early stage of this work. We further thank Dr. K. Butler for valuable feedback in the preparation of the manuscript. Funding for this work was provided by the Swiss National Science Foundation (SNSF) (NCCR Microbiomes grant agreement 51NF40_180575 and MicroScapeX grant number 2013/158 project number 200021_188623), the European Union's Horizon 2020 research and innovation programme (Marie Skłodowska-Curie grant agreement No 72228 and ShikiFactory grant agreement No 814408), and the Ecole Polytechnique Fédérale de Lausanne (EPFL).

Author contributions

HM, JH and VH designed the study. HM, JH, AS and VV developed algorithms, performed the analyses, and wrote the manuscript. HM and AS curated the data. HM developed the database architecture. JH developed the web interface. VH was responsible for project administration and funding acquisition, and he conceptualized and supervised the project.

Competing Interests

The authors declare no competing interests.