Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Lien, Martin

Metabolic network layout using biochemical coordinates

Master's thesis in Biotechnology, MBIOT5
Supervisor: Machado, Daniel
Co-supervisor: Márquez-Zavala, Elisa
August 2022

NTNU
Norwegian University of
Science and Technology

Lien, Martin

Metabolic network layout using biochemical coordinates

Master’s thesis in Biotechnology, MBIOT5
Supervisor: Machado, Daniel
Co-supervisor: Márquez-Zavala, Elisa
August 2022

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Metabolic network layout using biochemical coordinates

Martin Eide Lien

August 2022

Master thesis
Department of Biotechnology and Food Science
Norwegian University of Science and Technology

Supervisor : Associate Professor Daniel Machado
Co-supervisor: Ph.D candidate Elisa Márquez-Zavala

Preface

The work presented in this master's thesis was carried out at the Norwegian University of Science and Technology (NTNU) in Daniel Machado's Computational Biology lab at the Department of Biotechnology and Food Science. This thesis is the culmination of the Systems Biology specialization in the five-year M.Sc program in Biotechnology. The work presented is an individual undertaking and was started in June of 2021 and completed in August of 2022.

Marten Eide lien

Acknowledgements

When I got a new graphics card to replace my old one as a teenager, I read an article about how the processing power in newer graphical processing units present on new graphic cards were well suited to do parallel computations. The article mentioned the “Folding@home” project, a global distributed computing endeavour where volunteers offered the processing power in their personal computers to fill the need for computational power that was required to predict a protein’s tertiary structure based on their amino-acid sequence. The possibility of examining biological phenomena computationally was at the time nothing I could have imagined. When I wasn’t playing computer games, the card crunched through allocated work units given to my user profile. I was happy to think that there could be scientific advancements made with my small contributions but I came to realise that I wanted to pursue this field myself. With a fascination for biology, computers, and technology it isn’t strange that I choose to study biotechnology.

My experiences during my time at NTNU have spanned a wide spectrum, with many highs and many lows. I’ve made friends that I find it hard to see myself without in the future, achieved feats I could never have dreamed of but also failed and made more mistakes than I could ever imagine.

I’d like to acknowledge Post.doc Andre Voigt of the Systems Biology group for his suggestion to rank data points as way to solve the issue of planarity for highly clustered graphs. I’d like to extend a special thanks to my advisor, Associate Prof. Daniel Machado, for his help and guidance. I especially appreciate the conversations we’ve had during my time as your student when we didn’t talk about academics. I’d like to thank my co-supervisor Elisa Márquez-Zavala for her assistance with the written work as the deadline approached. I would like to extend my appreciation to my friends from NTNU’s 2021 iGEM team, Jonas Aouay Grønbakken from UiO’s 2020 iGEM team and the Principal Investigator for the teams, Associate Prof. Rahmi Lale and Professor Dirk Linke respectively. They ensured that I leave my studies with the confidence I need to overcome the challenges I will face in the future.

Writing this thesis would be far less enjoyable without the companionship of my dearest friends: Toan Phuc Vo, Haakon Zhang Moene, Victor Jankowski, and Henrik Dybedahl.

Abstract

The chemical properties of molecules are characterized by numerical values known as molecular descriptors and are calculated from molecular representations in cheminformatic tools. Different molecular descriptors are used to describe different chemical characteristics to develop new drugs or to predict a molecule's biological activity. There are a number of descriptors that can be derived from molecular topology, while more complex descriptors can be derived from the geometry of a molecule and others reflect different physicochemical properties.

Genome-scale metabolic models (GEMs) represent the known metabolism of an organism and are used to simulate *in silico* different aspects of its metabolism to understand how biological systems interact with each other to form emergent metabolic properties and capabilities. GEMs allow the study of how perturbations on the metabolic network affect the organism as a whole. The effects of gene deletions and changes in metabolic fluxes in an organism can be simulated and have been used by systems biologists to study metabolism *in silico*.

It is possible to gain new insights into how biological systems interact by interpreting visualizations of metabolic networks. The interpretation of large metabolic networks is difficult and the volume of data visualized often obfuscates biological insight that can be derived from their visualization. This project examined whether biochemical properties, quantified by molecular descriptors, could be used as two-dimensional coordinates, i.e. their biochemical coordinates, for generating meaningful metabolic network layouts.

This project used quantitative structure-activity relationship (QSAR) methods to calculate molecular descriptors for metabolites in two *Escherichia coli* GEMs and used these descriptors to generate metabolic network layouts. These layouts were then evaluated to determine which pair of descriptors produced the most intuitive and meaningful layouts for 7 case studies; 2 metabolic networks and 5 metabolic pathways. As a result of the chemical similarity prevalent in most metabolites, the generated metabolic network layouts exhibit significant node clustering and edge crossings, which negatively affects the readability of the network and its ability to convey biological information. The results of the metabolic pathway case studies suggest that a less complex network comprised of the primary intermediaries in metabolic pathways would be a better candidate for layouts generated using molecular descriptors.

Sammendrag

De kjemiske egenskapene til molekyler er preget av numeriske verdier kjent som molekylære deskriptorer og beregnes fra molekylære representasjoner i kjeminformatiske verktøy. Ulike molekylære deskriptorer brukes for å beskrive ulike kjemiske egenskaper for å utvikle nye medikamenter eller for å forutsi et molekyls biologiske aktivitet. Det er en rekke deskriptorer som kan avledes fra molekylær topologi, mens mer komplekse deskriptorer kan avledes fra geometrien til et molekyl og andre gjenspeiler forskjellige fysisk-kjemiske egenskaper.

Genom-skala metabolske modeller (GEM-er) representerer den kjente metabolismen til en organisme og brukes til å simulere *in silico* forskjellige aspekter av metabolismen for å forstå hvordan biologiske systemer interagerer med hverandre for å danne nye metabolske egenskaper og evner. GEM-er tillater studiet av hvordan forstyrrelser på det metabolske nettverket påvirker organismen som helhet. Effektene av gendelesjoner og endringer i metabolske flukser i en organisme kan simuleres og har blitt brukt av systembiologer for å studere metabolisme *in silico*.

Det er mulig å få ny innsikt i hvordan biologiske systemer samhandler ved å tolke visualiseringer av metabolske nettverk. Tolkningen av store metabolske nettverk er vanskelig, og volumet av data som visualiseres tilslører ofte biologisk informasjon som kan utledes fra visualiseringen. Dette prosjektet undersøkte om biokjemiske egenskaper, kvantifisert av molekylære deskriptorer, kunne brukes som todimensjonale koordinater, dvs. deres biokjemiske koordinater, for å generere meningsfulle metabolske nettverksoppsett.

Dette prosjektet brukte kvantitative struktur-aktivitetsforhold (QSAR) metoder for å beregne molekylære deskriptorer for metabolitter i to *Escherichia coli* GEM-er og brukte disse deskriptorene for å generere metabolske nettverksoppsett. Disse oppsettene ble deretter evaluert for å bestemme hvilket par av deskriptorer som produserte de mest intuitive og meningsfulle oppsettene for 7 casestudier; 2 metabolske nettverk og 5 stoffskifteveier. Som et resultat av den kjemiske likheten som er utbredt i de fleste metabolitter, viser de genererte metabolske nettverksoppsettene betydelige node-klynger og kantkrysninger, noe som negativt påvirker lesbarheten til nettverket og dets evne til å formidle biologisk informasjon. Resultatene fra casestudiene av stoffskifteveier tyder på at et mindre komplekst nettverk bestående av de primære mellomleddene i stoffskifteveier ville være en bedre kandidat for oppsett generert ved bruk av molekylære deskriptorer.

Contents

Figures

Tables

Acronyms

GEM Genome-scale metabolic model

GSMN Genome-scale metabolic network

ORF Open Reading Frame

GPR Gene-Protein reaction

FBA Flux Balance Analysis

SBML Systems Biology Markup Language

PPI Protein-Protein Interactions

PIN Protein-Interaction network

SBGN Systems Biology Graphical Notation

SBGN-ML Systems Biology Graphical Notation Markup Language

SBO Systems Biology Ontology

QED Quantitative Estimation of drug-likeness

QSAR Quantitative Structure-Activity Relationship

QSPR Quantitative Structure-Property Relationship

SMILES Simplified Molecular-Input Line-Entry System

InChI IUPAC Chemical Identifier

Da Dalton

Chapter 1

Introduction

1.1 Introduction

In 1965, a German biochemist, Dr. Gerhard Michal, compiled a list of known chemical reactions that take place in living cells. As a hobby, he created the first version of the Biochemical Pathways diagram, demonstrating the interconnectedness of various pathways. During the launch of the second edition of his biochemical pathway textbook in 2012, he stated [1]:

> “If you ask yourself what’s the most important characteristic of an organism, the usual answer is metabolism and reproduction (...) one of the main problems when presenting this activity is choosing what to show [2]”

In the fifty years since its initial rendition, the Biochemical Pathways have grown larger and more complex as the knowledge of biochemistry has increased. Gerhard’s lifelong effort resulted in two biochemical maps. In the first map, metabolic pathways are depicted, and in the second map, molecular processes are represented. It is thanks to the considerable amount of time Gerhard has spent editing the layout that these maps contain a great deal of detail while remaining meaningful to educated observers, and have been displayed in laboratories around the world [2].

The post-genomic era saw computerized algorithms meet the demand for accurate microbial gene identification [3] as the number of available microbial genome sequences grew. The mapping of microbial genomes led to the functional annotation of genes. Most genes in microbial cells code for gene products that are linked to metabolic function [4], and the majority of their biochemical function are known. It follows that an annotated genome allowed for the reconstruction of the metabolic map and its associated reactions, i.e. its metabolic network [5].

The availability of databases that hosts detailed metabolic pathway diagrams [6, 7] and annotated genomes [8, 9] has made the reconstruction of an organism’s metabolic network possible. Metabolic network reconstruction is first drafted from its annotated genome and then manually curated to ensure the reactions in the

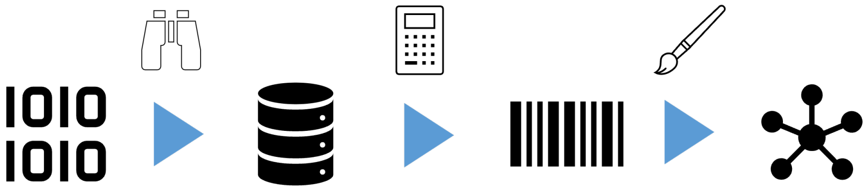


Figure 1.1: Simplified schematic representation of the thesis workflow: From GEMs the information required to build metabolite datasets will be collected. Molecular representations allow for the calculation of chemical features and their values used to position the metabolites when drawn.

draft are correct. The curated draft is then translated into a mathematical representation and converted into a genome-scale metabolic model (GEM). GEMs are used in metabolic modeling, an approach that allows for *in silico* simulation of microbial responses and can be used to verify experimental results [10], predict gene knock-out effects on metabolic fluxes [11], and the identification of essential genes under certain conditions [12].

In addition to its aforementioned usage in systems biology, metabolic networks can allow for the identification of emergent properties that would otherwise not be apparent when analyzing select parts of it [13]. These facets of a network can aid in contextualizing systems behavior, e.g. network robustness and modularity [13].

Analysis of genome-scale simulations lacks a suitable method to visualize large metabolic networks. Due to the interconnectivity of metabolites, reactions, and enzymes, automated layout algorithms tend to generate "hairball" networks [14].

Commonly used visualization tools [15, 16] utilize manually drawn biochemical maps which are intuitive to use but their coverage is limited to central metabolic pathways in well-studied model organisms.

The aim of this project is to create a metabolic network layout that can replace arbitrary layouts by integrating biochemical properties. This will be explored by examining if quantifiable biochemical properties can be utilized to assign metabolites a well-defined location in a two-dimensional coordinate system, i.e. their biochemical coordinates. By utilizing the known connection between molecular features and how these are expressed numerically [17], the modules and network structures present in the layouts generated by biochemical coordinates might reflect recognizable biochemical changes across metabolic pathways or the grouping of metabolites with similar molecular characteristics in metabolic networks. The pathway layouts will also be exported in one of the standard community formats.

The simplified workflow presented in Figure 1.1, outlines the steps that will be performed to achieve the aim of this thesis. By applying tools and methods from the fields of cheminformatics and systems biology to calculate molecular descriptors from molecular representations and collect metabolite identifiers from GEMs, the most promising descriptors can be found by employing PCA to identify the calculated features which best capture variance.

Several well-known pathways and *Escherichia coli* networks will be drawn to evaluate the different combinations of molecular descriptors that generates the most intuitive pathway layouts.

Chapter 2

Background

The field of systems biology incorporates computational and theoretical methods to be able to study how biological systems interact together to form emergent biological properties [18]. In contrast with classical biology's reductionism, systems biology employs a holistic approach to understanding how biological systems interact. The field [19] grew out of the post-genomic era following the volume of available biological data [20]. The advancements and wider adoption of systems biology saw the creation of a format to store metabolic networks [21]. An extension to this format was proposed so it could encode graphical representations of biochemical maps [22] to facilitate the standardization of metabolic network visualizations [23]. Cheminformatics is a field of chemistry that is focused on solving chemical issues by the application of tools and methods related to informatics. It uses representations of molecules [24, 25] to study the relationship between biological activity and chemical structure [17], drug discovery [26], predict structures of protein-ligand complexes [27] and quantify molecular properties by calculating their molecular descriptors [28]. This chapter begins with a section that describes GEMs and associated databases in more detail. Next, several well-known metabolic pathways used in this work are presented to provide a reference for the evaluation of their layouts with biochemical coordinates. Subsequently, an overview of tools and methods used to visualize metabolic networks and pathways are presented. The final section will present an overview of molecular descriptors, their categories, their application in quantitative structure-activity relationship studies, the codification of chemical information in molecular representations, and lastly the field of cheminformatics and cheminformatic tools.

2.1 Genome-scale metabolic models

To reconstruct a representation of the biochemical processes that compose cellular function, systems biology uses a bottom-up approach to simulate and predict the result of the interplay between biological systems [29]. The prediction of how gene-deletions affect metabolic fluxes across reactions in a metabolic network and the detection of essential genes are possible because GEMs represent the sum of reactions, metabolites, and genes of the organism of interest.

2.1.1 GEM reconstruction

In 1997, the genome sequence of Escherichia coli K-12 MG1655 was released [30] and three years later the first GEM [31] based on the annotation of that sequence was reported. The process of reconstructing GEMs encompasses steps that link the available metabolic knowledge to its genome sequence [5]. This process can be broken down into four steps [5]:

Step 1: Draft from genome annotations Genome annotations are used to collect a set of metabolic functions. Open reading frames (ORF) are identified and their function can be ascertained experimentally, through reference databases [5, 6, 32] or through sequence-homology searches with other organisms to approximate their function [5, 33]. For large GEMs, this step is often automated [5].

Step 2: Manual curation of the first draft In the first draft, metabolic functions may not have been correctly included or may not represent the entire metabolic network of an organism. Manually curating the draft to remove incorrect or include missed metabolic functions is necessary. This step is highly iterative due to the process of mapping missing or incorrect metabolites, reactions proteins, or genes and incorporating changes to the draft. Gene-protein reactions (GPRs) are also included to ensure that the proteins present in the model are linked to one or more genes [34]. The reconstruction is a self-contained Biochemical, Genetic, and Genomic database [15] (BiGG) of the metabolism of the organism of interest. The interconnectivity of an organism's metabolism is represented by a metabolic network that is encoded in a stoichiometric matrix. In this matrix, reactions and participating metabolites are stoichiometrically balanced which allows for constraint-based analysis such as flux balance analysis (FBA) [31].

Step 3: Formalizing the reconstruction into a computational mode As soon as the quality of the reconstruction has been verified by manual curation, the model is converted to a genome-scale constraint-based metabolic model. This is possible with tools such as the COBRA toolbox [35] and Cobrapy [36].

Step 4: Evaluation and iterative development of the model By comparing previous GEMs and experimental data, the accuracy of the model is evaluated, and missing metabolic functions can be identified with the utilization of reactome data and the gaps filled with information from reference databases [6, 32, 37]. Tools exist to automate the gap filling process [38, 39] but manual identification and filling of gaps remain the preferred method to ensure that the model is accurate [40].

2.1.2 GEM-based applications

GEMs are mathematical representations of the metabolic make-up of organisms. In GEMs, the molecular interactions are annotated and biochemical reactions are balanced with respect to mass and energy to be in line with the needed stoichiometric balance that allows the model to be used to simulate organism behavior at the metabolic level. An especially well-used approach using GEMs is constraint-based reconstruction and analysis [36] (COBRA). Constraints at a metabolic level are represented by mathematical equations that impose upper and lower bounds of a certain metabolite. The stoichiometric matrix ensures that the reactions are balanced with respect to the constraints. One of the first [5] constraint-based analysis methods is flux balance analysis (FBA). A typical parameter that a GEM is constrained with is substrate or nutrient availability. This can be achieved via modifying its uptake rate via exchange reactions. Measuring the impact of the constraint is done by defining an objective function. Cellular growth is an example of an objective function. Growth is often measured through the production of “biomass”, a colloquial term for biomolecules associated with cellular division and growth. FBA simulates the metabolic network’s response to situations such as gene deletions and growth media optimization by monitoring the flux of metabolites through metabolic networks with respect to its designated objective function [41].

To be able to accurately predict the metabolic properties of metabolically engineered cells can have profound industrial ramifications [42]. The engineering of microbes to overexpress desired products for industrial and medical applications [43] is hampered by reduced growth, and mutation rates which can be caused by the metabolic burden this overexpression incur on the cell, especially with the use of high-copy number plasmids. FBA can be used as a tool for measuring and predicting the metabolic burden associated with metabolic engineering, thereby guiding design decisions and saving time [44].

Using gene-expression data of SARS-CoV-2 infection and GEMs, it was possible to demonstrate extensive metabolic reprogramming of infected cells both *in silico* and experimentally [41]. A GEM of human metabolism [45], Recon 3D, was used to integrate metabolic fluxes generated from gene expression data [46]. Viruses direct the metabolism of infected cells for viral proliferation and this study characterized changes in metabolic flux patterns in pathways related to metabolite transport, glycine, serine, and threonine metabolism, pyrimidine synthesis, and fatty acid synthesis [41]. This study proposed data-supported metabolical drug targets for SARS-CoV-2 that could counter-act the metabolic state set by the virus [41].

Two well-known GEMs that have been used in this work are presented in subsequent subsections.

Escherichia coli central metabolism

The *E. coli* central metabolism GEM was released in 2010 and accompanied a paper that outlined how to reconstruct and use microbial metabolic networks [5]. It is smaller than a fully-fledged GEM and primarily encompasses central metabolic pathways which provide energy to *E. coli*. It comprises 72 metabolites, 96 reactions, and 137 genes. The reduced number of metabolites, reactions, and genes allows the calculation and simulation results to confer insight for those in-training while also yielding feedback for new constraint-based analysis methods that are easier to integrate during the troubleshooting phase due to the small size [5].

*Escherichia coli i*ML1515

*i*ML1515 is one of the most comprehensive GEMs for *E. coli* K-12 MG1655 [47] since the release of *i*JO1366 in 2011 [48] and has been validated with different growth conditions. Its name comes from 1,515 open reading frames which code for 1,192 unique metabolites and contain 2712 different reactions. The *i*ML1515 model is highly accurate for detecting essential genes under a wide array of carbon sources [49]. *i*ML1515 shows that GEMs can be utilized to predict how constraints manifest through changes in cellular and physiological behavior with accuracy and have been used to predict gene essentiality through machine learning [50] and provide insight into the improvement of lysine production in *E. coli* [51].

2.1.3 Systems biology standard format and GEM databases

As the number of GEMs and analysis methods increased, there was a clear need for a standardized format that could contain biological models for computational purposes [13]. The Systems Biology Markup Language [21, 52] (SBML) is the most used format for systems biology purposes and is the standard format in which GEMs are stored in. SBML is written in a dialect of the eXtensible Markup Language (XML, ".xml"). SBML stores cross-references for metabolites, reactions, proteins, and genes in annotations that link to external databases [53, 54].

Since its release in March 2001 [21], SBML has evolved in response to feedback and requests from the systems biology community. At its core, SBML stores a mathematical formalized model of the organism of interest, and subsequent iterations of SBML have developed layers that represent different model characteristics [52]. Of special interest is the layout [55] and rendering [56] layer which extends SBML to encode network and pathway representations. The positions and size of elements to be rendered are stored in the layout package and the parameters that dictate an element's color, font, and other graphical options are stored in the render package.

Due to the varying quality of reported GEMs [57], several databases have been established to meet the demand for centralized repositories of standardized models. Two of the databases used in this work are presented below. Biochemical Genetic and Genomic Models, or BiGG Models [15], is a database that hosts curated, high-quality models. BiGG models follow a systematic naming convention for metabolites and reactions that follows community standards. BiGG Models hosts 108 models that are validated with the model validation tool, MEtabolic MOdel TEsts (Memote) [58], and is available to be downloaded in SBML, JavaScript Object Notation (JSON), and "matfile" (MAT), a MATLAB proprietary data container. It provides visualizations of the desired pathway with the integration of Escher [59] and an application programming interface (API) to access the database [15].

MetaNetX [60] is a database that serves as a: "reconciliation of metabolites and biochemical reactions providing cross-links between major public biochemistry and Genome-Scale Metabolic Network (GSMN) databases". The terms "GEM" and "GSMN" are in the context and scope of this thesis interchangeable but GEM will be used throughout. It provides an automated genome-scale metabolic network reconstruction tool and cross-reference and chemical properties databases. It also provides tools for analyzing user-uploaded, SBML-compliant, GEMs as well as the ones hosted in their repository. While BiGG Models also hosts GEMs, MetaNetX focuses on mapping metabolites in GEMS with their structure. Molecular representations of metabolites within GEMs can be collected through the use of MetaNetX's chemical library mapping of metabolites within user-submitted GEMs [60].

2.2 Metabolic pathways

In the context of metabolic networks, pathways represent the modularity within an organism's metabolism. The following are brief descriptions of a few well-known metabolic pathways as well as their intermediary metabolites' chemical profiles.

2.2.1 The tricarboxylic acid cycle

The tricarboxylic acid (TCA) cycle is a cyclic catabolic pathway present in many organisms [61]. It produces energy via the reduction of common electron carriers such as NADH and $FADH_2$ by oxidizing two-carbon fragments. These carriers in turn can be used to provide energy to the respiratory chain, resulting in the formation of CO_2 and the synthesis of ATP.

Pyruvate, a three-carbon fragment, the end product from carbohydrate, fat, and protein degradation is formed through glycolysis and is oxidized to form acetyl-CoA, a two-carbon fragment attached to coenzyme A, by the pyruvate dehydrogenase complex (PDH). The acetyl moiety from acetyl-CoA enters the citric acid cycle when it is donated to oxaloacetate, forming citrate via a Claisen condensation. The sequence of reactions in the citric acid cycle extracts the energy within the acetyl group via oxidation to form two molecules of CO_2 and reduce

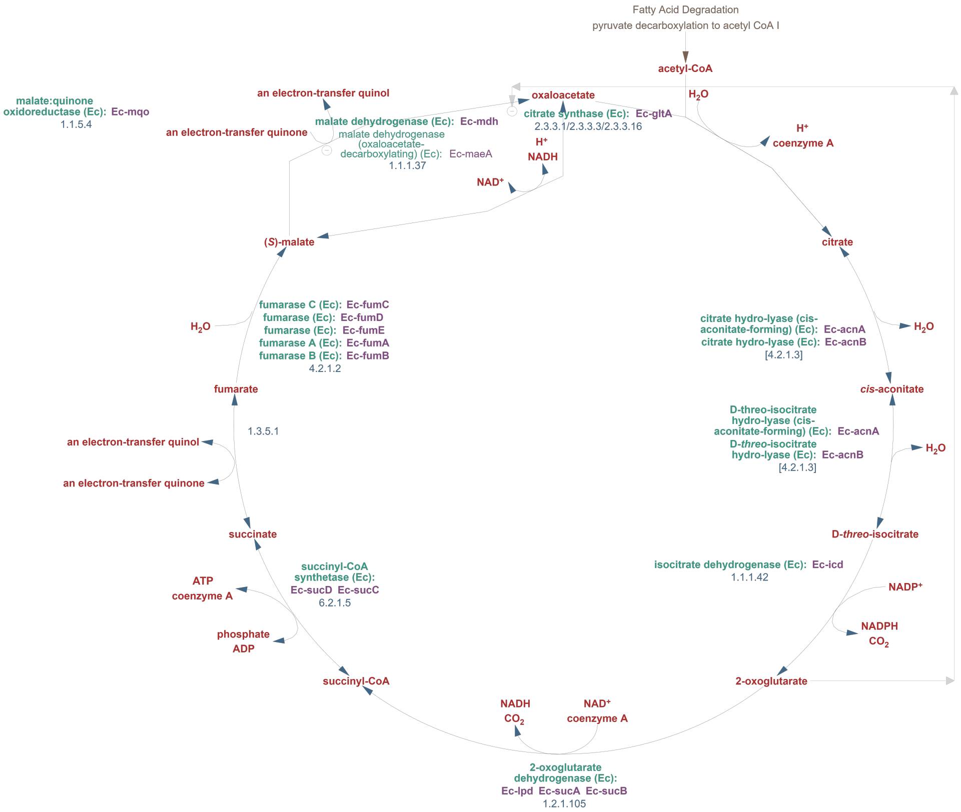

Fatty Acid Degradation
pyruvate decarboxylation to acetyl CoA I
acetyl-CoA
malate:quinone oxidoreductase (Ec): Ec-mqo
1.1.5.4
an electron-transfer quinol
an electron-transfer quinone
oxaloacetate
H_2O
malate dehydrogenase (Ec): Ec-mdh
malate dehydrogenase (oxaloacetate-decarboxylating) (Ec): Ec-maeA
1.1.1.37
citrate synthase (Ec): Ec-gltA
2.3.3.1/2.3.3.3/2.3.3.16
H^+
coenzyme A
H^+
NADH
NAD^+
(*S*)-malate
citrate
fumarase C (Ec): Ec-fumC
fumarase (Ec): Ec-fumD
fumarase (Ec): Ec-fumE
fumarase A (Ec): Ec-fumA
fumarase B (Ec): Ec-fumB
4.2.1.2
H_2O
citrate hydro-lyase (cis-aconitate-forming) (Ec): Ec-acnA
citrate hydro-lyase (Ec): Ec-acnB
[4.2.1.3]
H_2O
fumarate
cis-aconitate
1.3.5.1
D-threo-isocitrate hydro-lyase (cis-aconitate-forming) (Ec): Ec-acnA
D-*threo*-isocitrate hydro-lyase (Ec): Ec-acnB
[4.2.1.3]
an electron-transfer quinol
H_2O
an electron-transfer quinone
succinate
D-*threo*-isocitrate
succinyl-CoA synthetase (Ec): Ec-sucD Ec-sucC
6.2.1.5
isocitrate dehydrogenase (Ec): Ec-icd
1.1.1.42
ATP
coenzyme A
$NADP^+$
phosphate
ADP
NADPH
CO_2
succinyl-CoA
2-oxoglutarate
NADH
CO_2
NAD^+
coenzyme A
2-oxoglutarate dehydrogenase (Ec): Ec-lpd Ec-sucA Ec-sucB
1.2.1.105

Figure 2.1: Pathway representation of the tricarboxylic acid cycle in *E. coli*. Retrieved from MetaCyC [62]. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Figure C.1 shows this pathway with compound structures.

three NAD^+ molecules and one FAD molecule in the course of one cycle. It is in the reduction of these electron carriers that provide the energy to form ATP in the electron transport chain. Figure C.1 illustrates the structural similarities of the TCA cycle intermediaries. With the exception of succinyl-CoA and acetyl-CoA, the majority of intermediaries also have similar molecular weights (see Table B.1).

2.2.2 *De novo* purine nucleotide biosynthesis pathway

The purine nucleotides, adenosine triphosphate (ATP) and guanosine triphosphate (GTP), shown in Figure 2.2 are repeating monomeric units that when linked together form nucleic acids. The structural make-up of nucleotides can be broken down into phosphate groups, a pentose sugar, and a nitrogenous base. In the case of purine nucleotides, their name comes from their nitrogenous base, a purine ring. Purines are heterocyclic amines that have a pyrimidine ring merged with a hept-member ring with two nitrogen atoms, illustrated in Figure 2.2

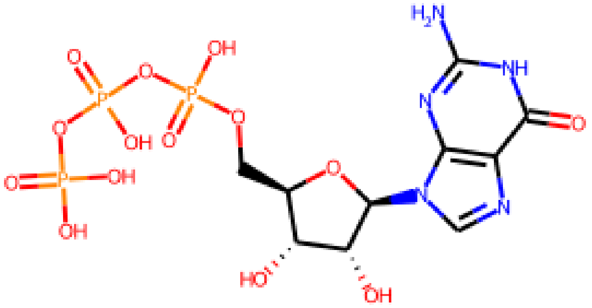


Figure 2.2: Two-dimensional representation of GTP. From left to right: Phosphate groups colored red, a pentose sugar, and a purine ring. Rendered in RDkit.

The *de novo* biosynthesis of the purine nucleotides shares a common intermediate metabolite, inosine-5'-phosphate (IMP), which is the end product of the inosine-5'-phosphate biosynthesis pathway, illustrated in Figure 2.3. The intermediaries in the IMP pathway are structurally similar (see Figure C.2). The ribosyl group in PRPP is chemically transformed to a purine in the IMP pathway by the addition of carbon and nitrogen atoms from glycine, glutamine, aspartic acid, formyl groups, and bicarbonate which is facilitated by enzymatic reactions. Due to the structural similarity between the main intermediaries in this pathway, they also exhibit overlapping chemical similarities (see Tables B.2 and B.4).

De novo biosynthesis of purines begins with the formation of the purine ring backbone in the form of inosinate monophosphate (IMP) which again is synthesized from a series of reactions from phosphoribosyl pyrophosphate (PRPP) in the inosine-5'-phosphate biosynthesis pathway, illustrated in 2.3. From IMP, adenosine and guanosine end products can synthesize, but L-aspartate is required for dATP synthesis as it forms N^6-(1,2-dicarboxyethyl)AMP with IMP in the adenylosuccinate synthetase enzyme reaction, as illustrated in Figure 2.4.

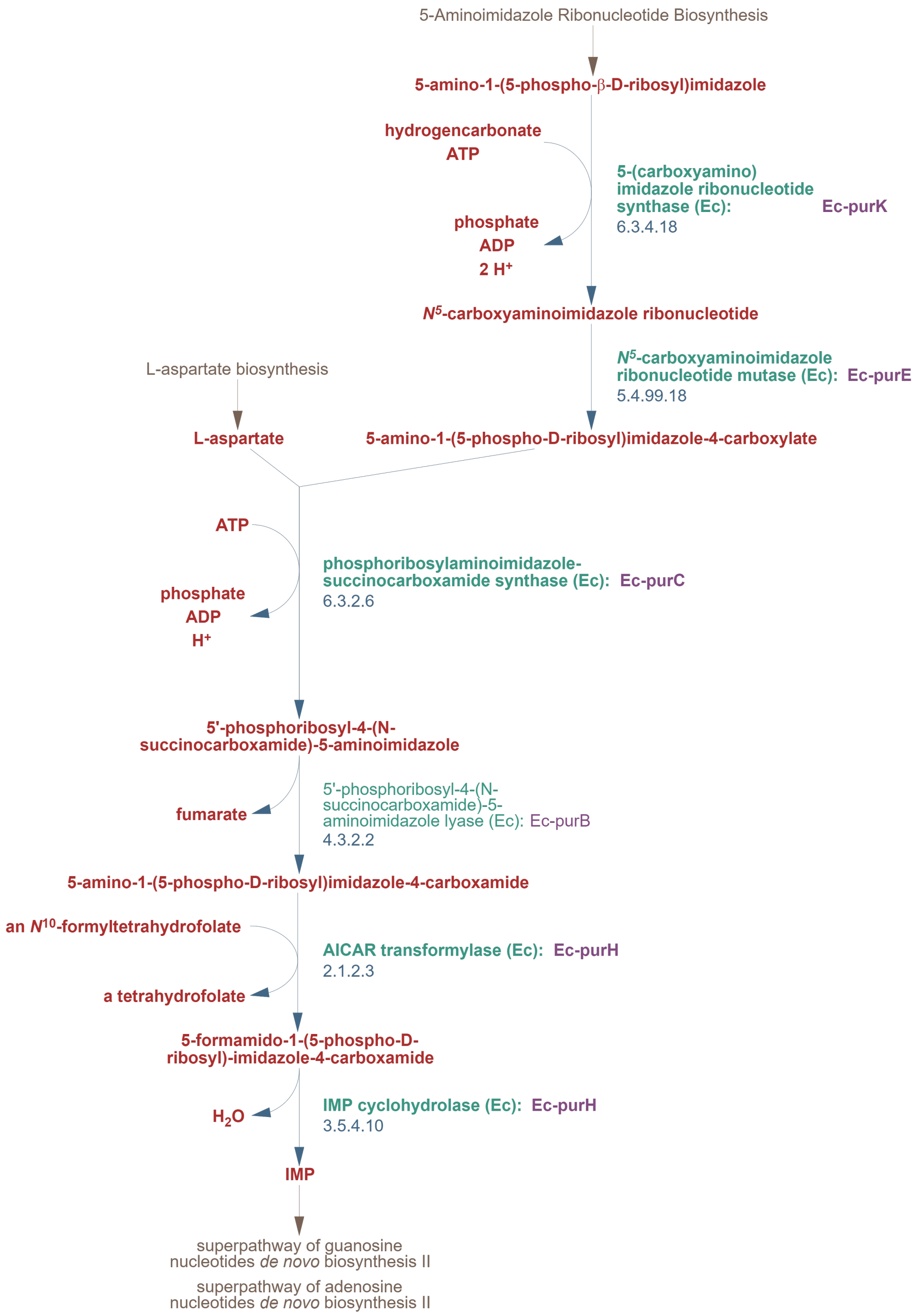

5-Aminoimidazole Ribonucleotide Biosynthesis
5-amino-1-(5-phospho-β-D-ribosyl)imidazole
hydrogencarbonate
ATP
phosphate
ADP
2 H+
5-(carboxyamino) imidazole ribonucleotide synthase (Ec): Ec-purK
6.3.4.18
N5-carboxyaminoimidazole ribonucleotide
N5-carboxyaminoimidazole ribonucleotide mutase (Ec): Ec-purE
5.4.99.18
L-aspartate biosynthesis
L-aspartate
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxylate
ATP
phosphate
ADP
H+
phosphoribosylaminoimidazole-succinocarboxamide synthase (Ec): Ec-purC
6.3.2.6
5'-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole
fumarate
5'-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole lyase (Ec): Ec-purB
4.3.2.2
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide
an N10-formyltetrahydrofolate
a tetrahydrofolate
AICAR transformylase (Ec): Ec-purH
2.1.2.3
5-formamido-1-(5-phospho-D-ribosyl)-imidazole-4-carboxamide
H2O
IMP cyclohydrolase (Ec): Ec-purH
3.5.4.10
IMP
superpathway of guanosine nucleotides de novo biosynthesis II
superpathway of adenosine nucleotides de novo biosynthesis II

Figure 2.3: Pathway representation of inosine-5’-phosphate biosynthesis in *E. coli*. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [63]. Figure C.2 shows this pathway with compound structures.

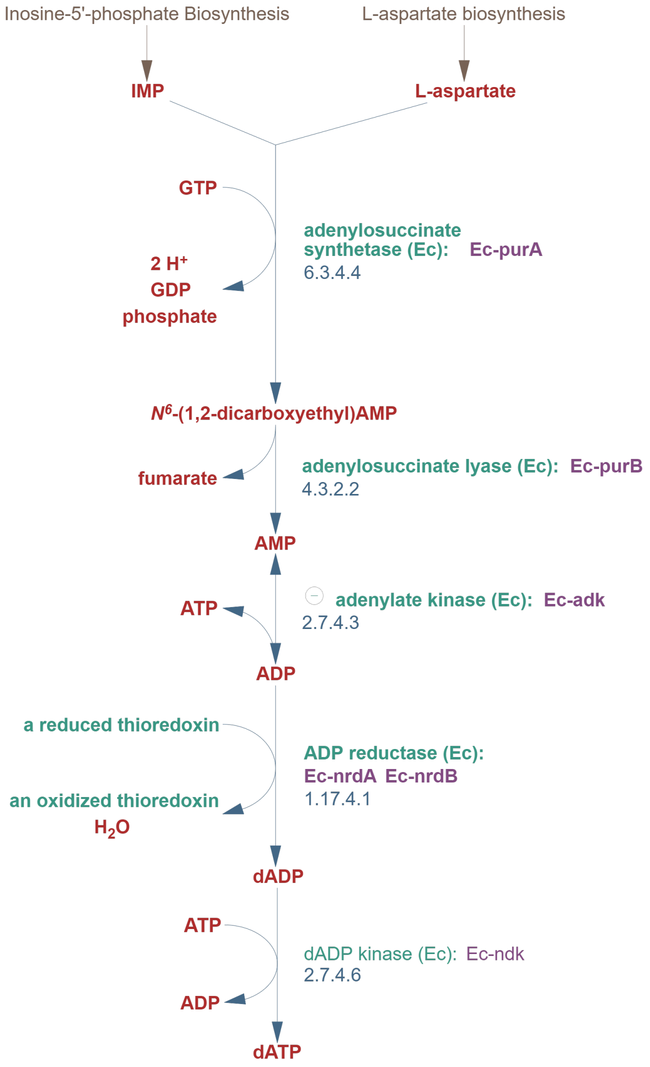


Figure 2.4: Pathway representation of *de novo* adenosine nucleotide biosynthesis in *E. coli*. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [63]. Figure C.3 shows this pathway with compound structures.

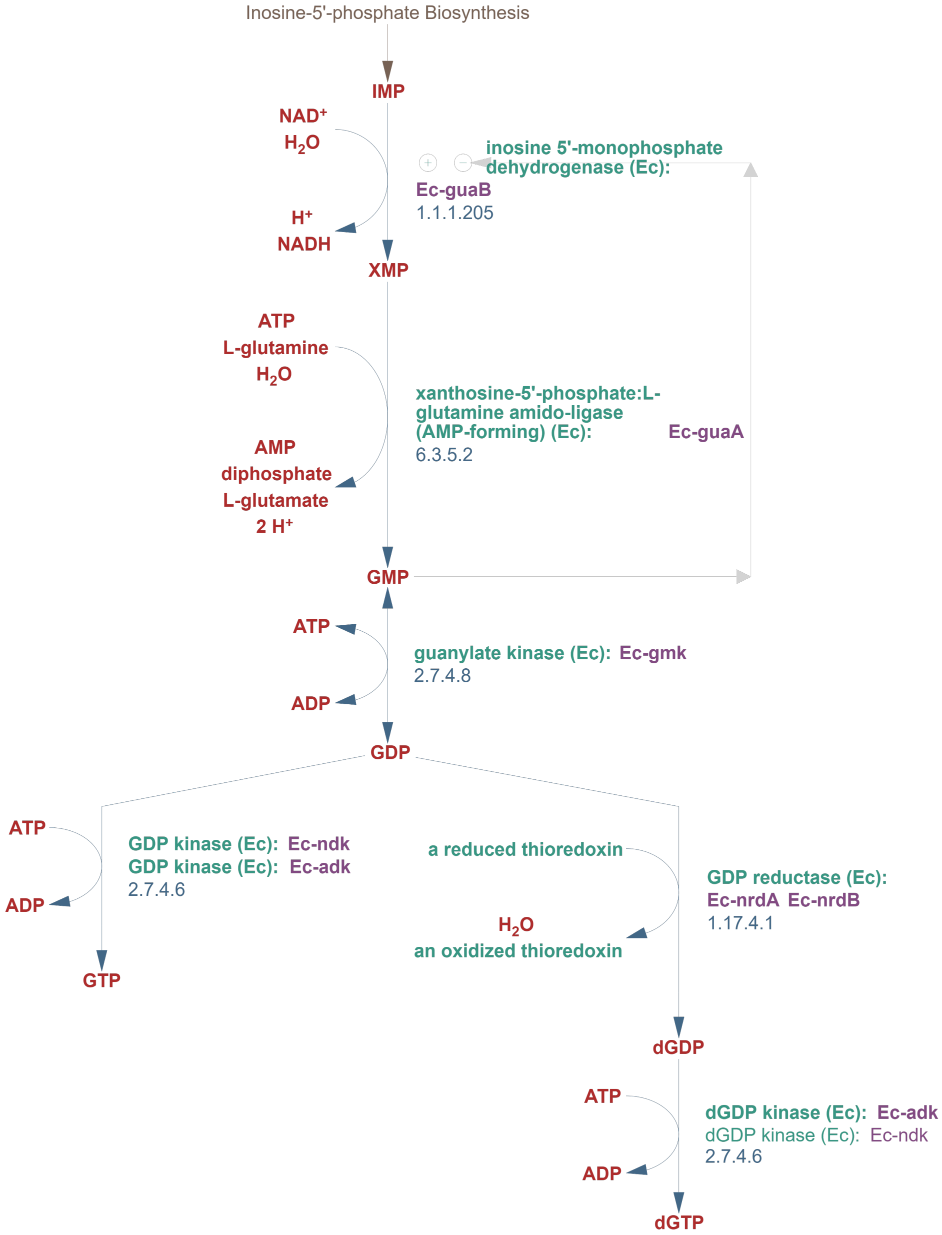


Figure 2.5: Pathway representation of *de novo* guanosine nucleotide biosynthesis in *E. coli*. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [64]. Figure C.4 shows this pathway with compound structures.

2.2.3 *De novo* pyrimidine nucleotide biosynthesis pathway

A major function of any organism is the biosynthesis of pyrimidine nucleotides. The de novo pyrimidine biosynthesis pathway is one of the oldest pathways conserved across eukaryotes and prokaryotes.The pyrimidine nucleotides, cytidine triphosphate (CTP), uridine monophosphate (UMP), and deoxythymidine monophosphate (dTMP) are repeating monomeric units that when linked together form nucleic acids. The structural make-up of pyrimidine nucleotides can be broken down into phosphate groups, a pentose sugar, and a pyrimidine ring, illustrated in Figure 2.6.

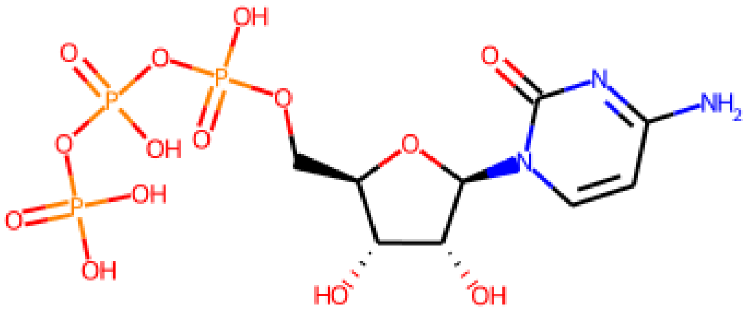


Figure 2.6: Two-dimensional representation of CTP. From left to right: Phosphate groups colored red, a pentose sugar, and a purine ring. Rendered in RDkit.

In the case of pyrimidine nucleotides, their name comes from their nitrogenous base, a pyrimidine ring. Pyrimidines are heterocyclic amines that have a pyrimidine which two nitrogen atoms in its 6-member ring. *De novo* biosynthesis of pyrimidine nucleotides, shown in Figure 2.6, requires aspartate, PRPP, and carbamoyl phosphate. Contrasting the way in which purine nucleotides are synthesized by the use of IMP as the basis for their nitrogenous base, pyrimidines create their nitrogenous base from aspartate and carbamoyl phosphate. Figure C.5 illustrates how aspartate and carbamoyl phosphate form carbamoyl phosphate and in subsequent reactions, the steric size and weight of intermediaries increase as they are becoming gradually similar to that of the end product, CTP.

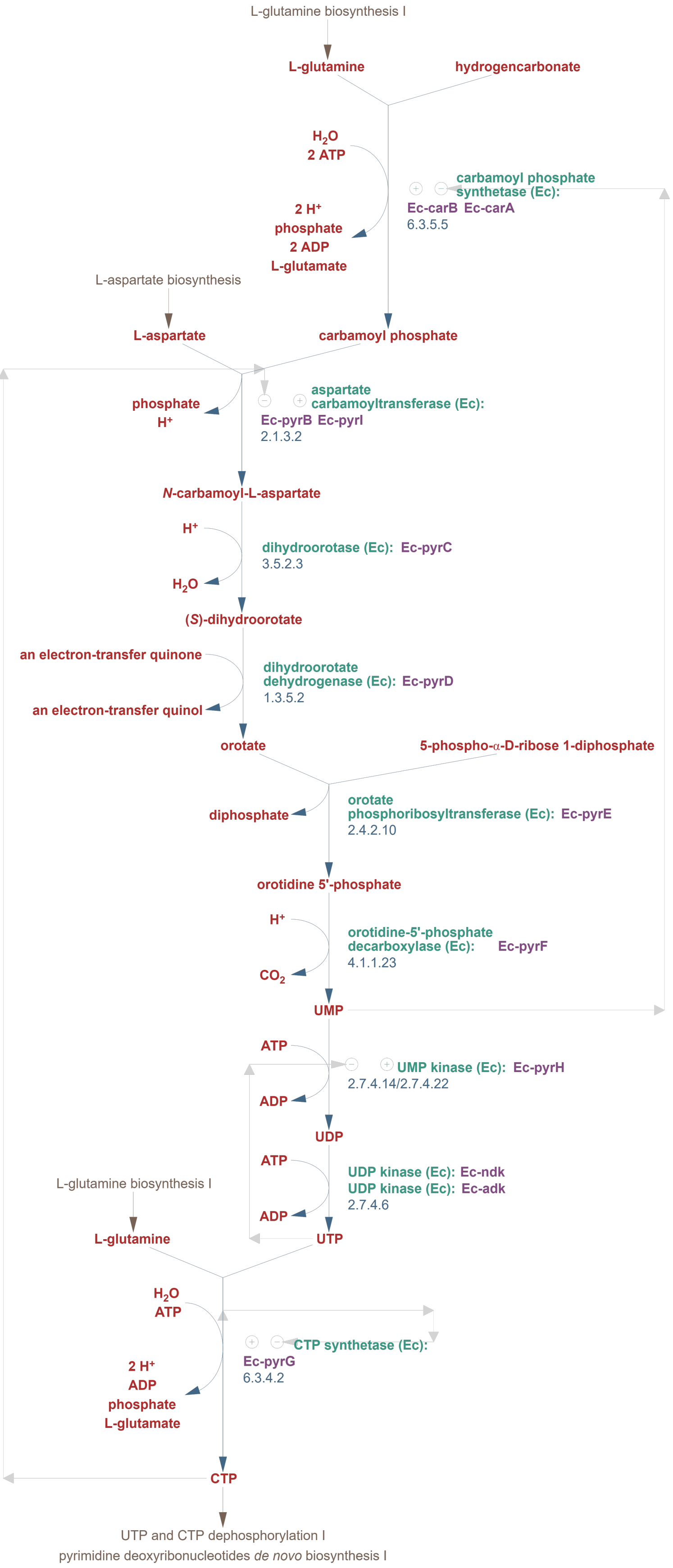


Figure 2.7: Pathway representation of *de novo* CTP biosynthesis in *E. coli*. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [65]. Figure C.5 shows this pathway with compound structures.

2.2.4 Histidine biosynthesis pathway

Histidine biosynthesis is a linear metabolic pathway that results in the formation of the amino acid histidine. Histidine contains an α-amino group, a carboxylic acid, and an imidazole side chain, illustrated in Figure 2.8. Histidine is a precursor for histamine [66], and the imidazole side chain is fundamental for many enzyme reactions [67].

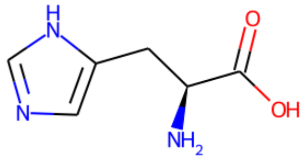


Figure 2.8: Two-dimensional representation of Histidine. Visualized in RDkit.

Histidine is one of the amino acids that cannot be synthesized de novo in humans and is conserved across many organisms [67]. Histidine is synthesized from PRPP in 10 steps, illustrated in 2.9. From PRPP it condenses with ATP to form phosphoribosyl-ATP (PRBATP) and through subsequent reactions, it forms L-histidine. Figure C.6 illustrates how the size of steric bulk and associated molecular weight (see Table B.5 decrease as the intermediaries are becoming more similar to that of the end product, L-histidine.

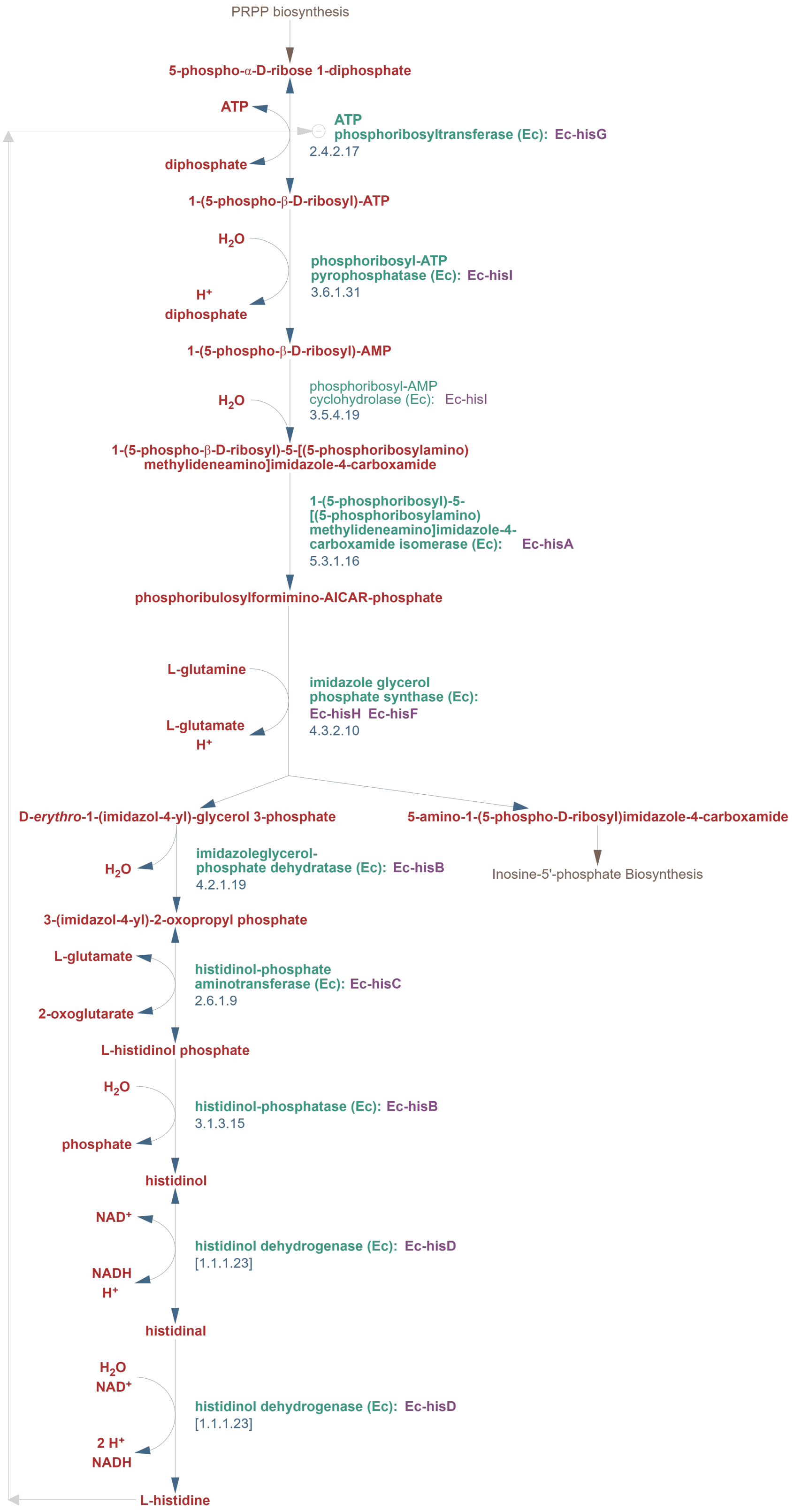

PRPP biosynthesis
5-phospho-α-D-ribose 1-diphosphate
ATP
diphosphate
ATP
phosphoribosyltransferase (Ec): Ec-hisG
2.4.2.17
1-(5-phospho-β-D-ribosyl)-ATP
H_2O
H^+
diphosphate
phosphoribosyl-ATP
pyrophosphatase (Ec): Ec-hisI
3.6.1.31
1-(5-phospho-β-D-ribosyl)-AMP
H_2O
phosphoribosyl-AMP
cyclohydrolase (Ec): Ec-hisI
3.5.4.19
1-(5-phospho-β-D-ribosyl)-5-[(5-phosphoribosylamino)
methylideneamino]imidazole-4-carboxamide
1-(5-phosphoribosyl)-5-
[(5-phosphoribosylamino)
methylideneamino]imidazole-4-
carboxamide isomerase (Ec): Ec-hisA
5.3.1.16
phosphoribulosylformimino-AICAR-phosphate
L-glutamine
L-glutamate
H^+
imidazole glycerol
phosphate synthase (Ec):
Ec-hisH Ec-hisF
4.3.2.10
D-*erythro*-1-(imidazol-4-yl)-glycerol 3-phosphate
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide
Inosine-5'-phosphate Biosynthesis
H_2O
imidazoleglycerol-
phosphate dehydratase (Ec): Ec-hisB
4.2.1.19
3-(imidazol-4-yl)-2-oxopropyl phosphate
L-glutamate
2-oxoglutarate
histidinol-phosphate
aminotransferase (Ec): Ec-hisC
2.6.1.9
L-histidinol phosphate
H_2O
phosphate
histidinol-phosphatase (Ec): Ec-hisB
3.1.3.15
histidinol
NAD^+
NADH
H^+
histidinol dehydrogenase (Ec): Ec-hisD
[1.1.1.23]
histidinal
H_2O
NAD^+
2 H^+
NADH
histidinol dehydrogenase (Ec): Ec-hisD
[1.1.1.23]
L-histidine

Figure 2.9: Pathway representation of *de novo* histidine biosynthesis in *E. coli*. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [68]. Figure C.6 shows this pathway with compound structures.

2.3 Metabolic networks and network visualization

The advancements made in the acquisition of biological data have created a volume of available data that requires automated data management and analysis tools. A variety of scientific areas and fields have adopted data-dependent methodologies in addition to their established methods to save time and costs as well as discover new insights from new and old data [69]. Visualization is a powerful tool that helps contextualize different aspects of complex information and can provide insights that would not be apparent without it. Visualizations of metabolic networks and pathways elucidate how different aspects of metabolism are connected and by extension how one change in metabolism affects associated processes [70].

2.3.1 Metabolic networks

A metabolic network is a complex structure that is formed by the interconnection between metabolic pathways and shared metabolites [71, 72]. Metabolic networks are often visualized as graphs to demonstrate the relationship between metabolites and biochemical reactions [73]. Graphs are mathematical objects that convey the relationship between nodes that are connected by edges. A graph G is defined as

$$G = (N, E) \tag{2.1}$$

Where N is the set of nodes $\{\mathrm{v}_1,\mathrm{v}_2,\mathrm{v}_3,\mathrm{v}_4...,\mathrm{v}_\mathrm{n}\}$ & E is the corresponding set of edges. Graphs can be undirected or directed. In undirected graphs, the edges between nodes have equal endpoints and convey no information besides the fact that the nodes are linked. In directed graphs, edges can be directed and bidirectional. In directed graphs, edges convey relational information from its node of origin to its target. In bidirectional graphs, edges can carry relational information to and from their node of origin and the target node.

The attributes of a network that can be analyzed and used to characterize it are called network properties. The degree of a network is the sum of all edges in a network and a network's average degree is the sum of edges divided by the sum of all nodes in the network. The degree distribution is a statistical description of the probability that a node has a degree k. A degree distribution, $P(k)$ is defined as

$$P(k) = \begin{pmatrix} n-1 \\ k \end{pmatrix} p^k (1-p)^{n-1-k} \tag{2.2}$$

and reflects the probability P that a node has a degree equal to k. When the degree distribution of a network follows a power law distribution it means that a small number of nodes in the network are linked with many others and most others are linked to a fewer number of nodes. Networks that exhibit this power-law distribution are called scale-free networks, where a few highly connected nodes, typically named hubs, have a degree that is much larger than the network average.

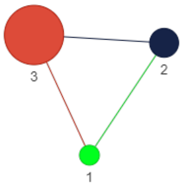


Figure 2.10: A simple directional graph comprised of three nodes. Edge color is inherited from the node it originates from.

Biological networks have been studied to gain insight into their organizational structure, which reflects their function. Metabolic networks are characterized as scale-free networks. This scale-free organization is possibly derived from the chemical basis of the constituents in the metabolic network [74]. Metabolites are typically small molecules and in a metabolic network, they are linked through metabolic reactions which are essentially chemical reactions. The metabolites which drive metabolic reactions are called hubs due to their large degrees compared to the network average. Most metabolic reactions take place in a hydrophilic environment and the metabolites which drive many of these reactions are usually required to be maintained at a higher concentration. In order to reach a high-enough concentration, hubs are often strongly polar. This assumption is supported by the characterization of 100 metabolite concentrations in *E. coli* [75] which revealed that the most abundant compounds exhibited larger average degrees than those with lower concentrations. This possible explanation for the organization of metabolic networks suggests that the chemical property of a metabolite is partly reflected in its nominal degree and highly connected metabolites; hubs, in metabolic networks, exhibit lower variability in their concentration [76, 77].

When metabolic networks and pathways are visualized as graphs, metabolites are typically represented as nodes while edges reflect a reaction with neighboring metabolites and the direction of the edge indicates which way the reaction occurs, illustrated in Figure 2.18.

2.3.2 Visually interpreting biological data using networks

Data from biological systems can be difficult to interpret due to their complexity [78], and visualizations of biological systems often generate more easily intelligible representations than lists of texts and mathematical formulas could [71, 79]. Visualizing protein-protein interactions (PPIs), metabolism, and cell-cell interactions in biological systems are commonly done through network visualization to enable visual interpretation of relationships between elements in these systems and reduce complexity.

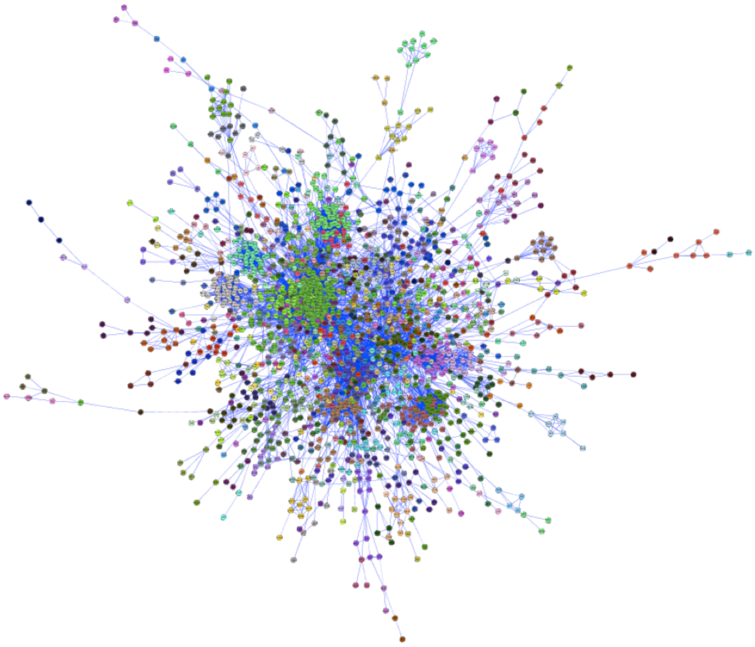


Figure 2.11: Example of a *saccharomyces cerevisiae* protein interaction network (PIN). Edges between proteins indicate protein-protein interactions (PPIs). Taken from Ref [80]

As presented in the preceding section, networks convey relational information about a set of nodes that are linked by edges. In the context of a protein interaction network (PIN), exemplified in Figure 2.11, nodes represent proteins and the edges represent the PPIs of the proteins in that network. Visualizing and analyzing biological systems as networks allows the study of their structure and how they interact together to produce emergent cellular behaviors [79].

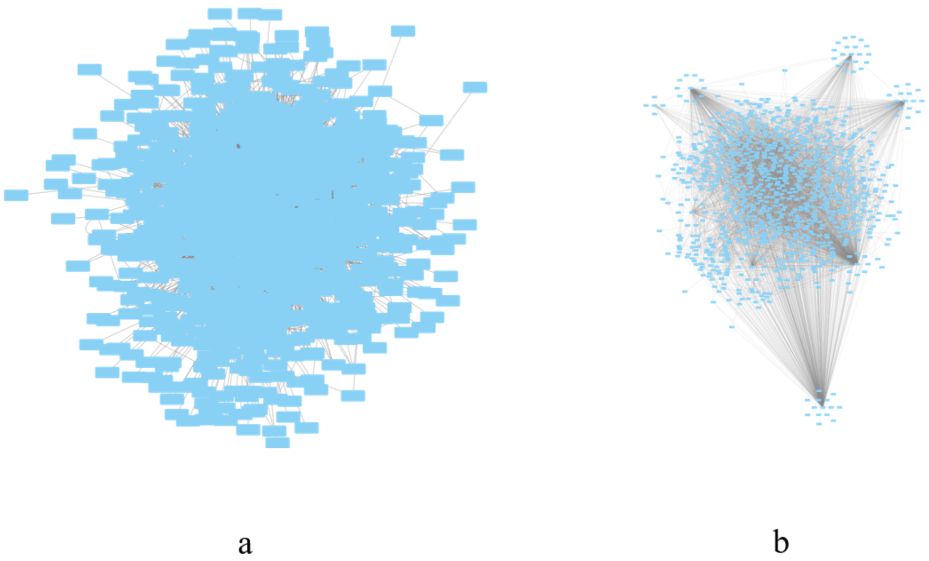


Figure 2.12: Network representations of metabolites in the *E. coli* *i*ML1515 model with (a) default loaded layout and (b) force-directed layout. Edges represent metabolites participating in chemical reactions. Multiple edges indicate that multiple metabolites participate in the reaction. (a) The default loaded layout suffers from some issues of planarity with many nodes overlapping. (b) The force-directed layout is difficult to interpret with the majority of metabolites clustered in the middle. Some metabolites share unique edges with other metabolites, at the periphery of the giant component. Rendered in Cytoscape [81].

Identification and characterization of functional elements in a complex biological system are essential to achieving a systems-level understanding of biology. The function of biological systems is often comprised of modules or motifs of molecular components that work together to perform a specific function. In a network setting, modules comprise a group of interconnected nodes and motifs reflect a recurring pattern of interactions between nodes [82]. Their visual identification is often hampered by overlapping nodes or edge crossings which obfuscates relational information and is also subject to visual bias. Organizing nodes in a layout is often the first step to creating an understandable representation of a network and automatic layout algorithms have been created for this purpose. Popular algorithms that generate "force-based" or circular layouts are often used. A simple explanation of "force-based" layouts is that edges "pull" nodes that "repel" each other with rendered edge lengths as short as possible. An example of a force-based layout is shown in Figure 2.12 Circular layouts arrange nodes sequentially so that the number of edge crossings is reduced. An example of a circular layout is shown in Figure 2.13 In these two examples, the rendered length of edges conveys no biological information [79]. The automatic layout algorithms can organize nodes according to the algorithm's objective in large and highly connec-

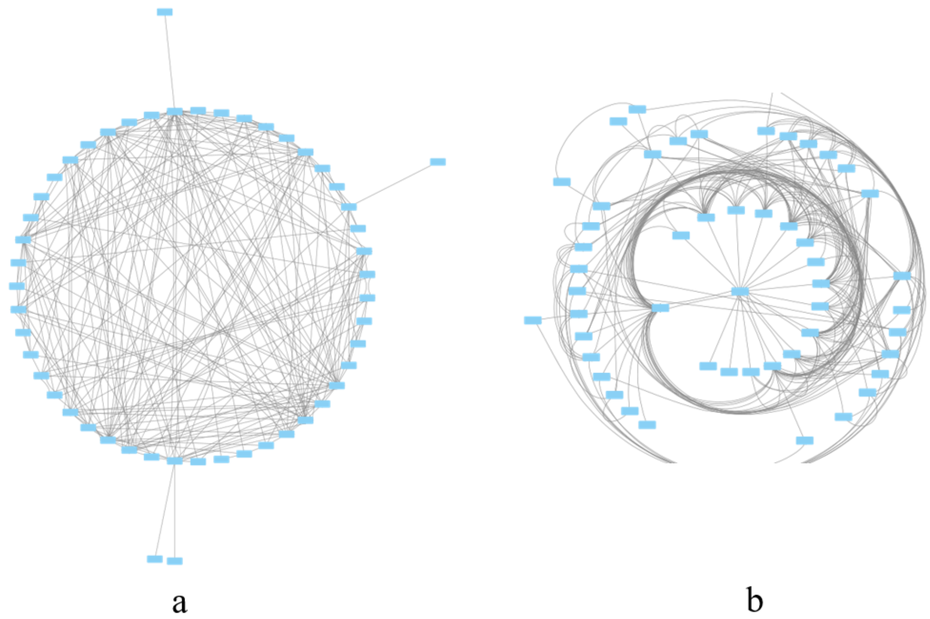


Figure 2.13: Network representations of metabolites in the *E. coli* central metabolism model with (a) circular and (b) radial layout (zoomed in), a variant of the circular layout. Edges represent metabolites participating in chemical reactions. Multiple edges indicate that multiple metabolites participate in the reaction. (a) The circular layout visualization suffers from several interconnected nodes that cause edge-crossings and deciphering information from the visualization is difficult. (b) The radial layout positions a metabolite in the center as the root node. Identifying hubs is easier in the radial layout, when compared to the circular layout in (a). Rendered in Cytoscape [81].

ted networks, but the resulting arrangement is not always optimal from a visual standpoint and may benefit from being divided into smaller, separate components. Manually manipulating node positions after applying layout algorithms can produce understandable representations that facilitate the visual identification of network modules [79]. Network elements can be further distinguished by altering their visual attributes, such as their size, color, shape, or edge thickness, by incorporating different types of data. Examples of how different network visualization software are used to accomplish these are given in subsequent sections.

An intuitive network representation can be used to infer metabolite's function in biochemical reactions of a metabolic network or identify possible protein complexes if several proteins exhibit interconnectivity and form clusters [78]. Identification and characterization of different network structures can then lead to insights into how the interplay of these forms known functions and possibly the dynamics of different biological systems. Several approaches exist for the computational identification of network modules which are based on characterizing modules as sub-graphs within a network using calculated network attributes such as the clustering coefficient. However, interpreting these results still requires expert confirmation of their biological relevance, even though they are less susceptible to bias.

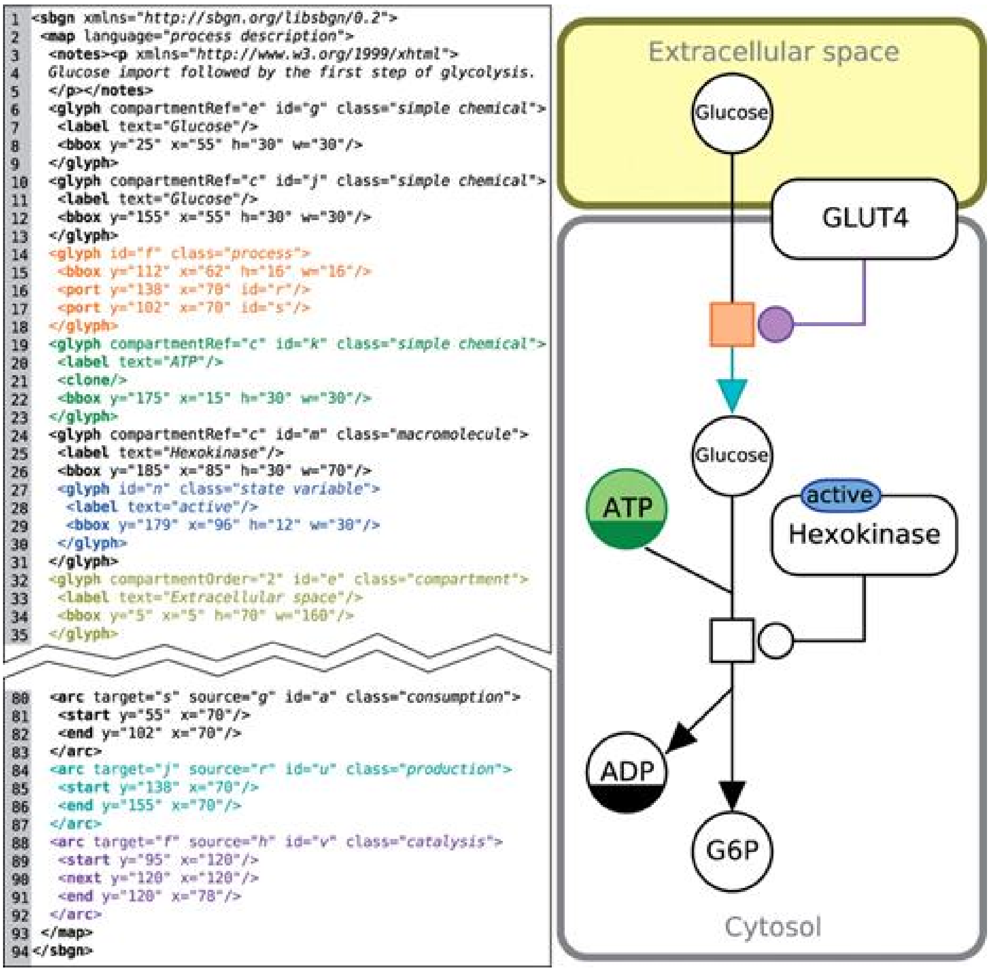
Some types of biological data are not easily represented as networks and some can not be broken down without losing information. especially if the layout or mapping of data interferes with the generation of an intuitive visualization. A reaction in a metabolic pathway usually requires substrate(s), enzyme(s), and product(s). Visualizing reaction elements in a larger network is not easy to achieve without significant edge crossings as many substrates are shared across reactions in different pathways [79]. While several metabolic network visualization tools have implemented methods to visually remove highly connected molecules [83], such as H_2O and CO_2. Alternative methods that more accurately convey biological systems have been suggested but no community or standard solution has been proposed [79].

Analyzing biological data using network representations is crucial for understanding how biological systems work, but the complexity and volume of data often make it difficult to create visualizations that are meaningful [84]. The spatial arrangement of nodes can make the network difficult to interpret, obfuscating the information that can be gleaned from the visualization as node and edge positions influences how the viewer perceives the visualization and lead to erroneous interpretations. The concepts of node centrality and proximity are important when using a spatially meaningful layout [84]. Nodes in proximity to each other are intuitively interpreted as exhibiting similarity while nodes further away, have less similarity. Centrality is a concept that refers to how objects that are placed at the center of a visualization are interpreted as being more essential or important than those at the periphery [84]. A good representation manages to provide meaningful information about the network, demonstrate relationships and not lose significant detail at the same time. There must be sufficient consideration given to how networks and different types of biological data are visualized in order not to overwhelm the viewer while at the same time providing biological insight through an informative visualization of a network.

2.3.3 Encoding visual information in systems biology formats

Several efforts to standardize the visualization of genome-scale metabolic networks and metabolic pathways have been pursued after the publication of SBML [21]. The encoding of graphical information was already proposed as a possible extension to the then unreleased format in 2000 [22]. Various community-supported projects filled the gap to standardize graphical layouts and graphical notations used [23, 85, 86]. The Systems Biology Graphical Notation (SBGN) presented a standardized set of icons to symbolize molecular entities and diagram styles. The SBGN format contains graphical details of the layout.

In an effort to further standardize the layout and graphical details and reduce the use of many fragmented visualization standards, the Systems Biology Graphical Notation Markup Language [87] (SBGN-ML) was developed. SBGN-ML encodes both the layout and the graphical notations assigned. SBGN-ML only contains information pertaining to the positioning and which icon to be used as SBGN-ML compliant software includes a standardized library of graphical symbols.



```
<sbgn xmlns="http://sbgn.org/libsbgn/0.2">
 <map language="process description">
  <notes><p xmlns="http://www.w3.org/1999/xhtml">
  Glucose import followed by the first step of glycolysis.
  </p></notes>
  <glyph compartmentRef="e" id="g" class="simple chemical">
   <label text="Glucose"/>
   <bbox y="25" x="55" h="30" w="30"/>
  </glyph>
  <glyph compartmentRef="c" id="j" class="simple chemical">
   <label text="Glucose"/>
   <bbox y="155" x="55" h="30" w="30"/>
  </glyph>
  <glyph id="f" class="process">
   <bbox y="112" x="62" h="16" w="16"/>
   <port y="138" x="70" id="r"/>
   <port y="102" x="70" id="s"/>
  </glyph>
  <glyph compartmentRef="c" id="k" class="simple chemical">
   <label text="ATP"/>
   <clone/>
   <bbox y="175" x="15" h="30" w="30"/>
  </glyph>
  <glyph compartmentRef="c" id="m" class="macromolecule">
   <label text="Hexokinase"/>
   <bbox y="185" x="85" h="30" w="70"/>
   <glyph id="n" class="state variable">
    <label text="active"/>
    <bbox y="179" x="96" h="12" w="30"/>
   </glyph>
  </glyph>
  <glyph compartmentOrder="2" id="e" class="compartment">
   <label text="Extracellular space"/>
   <bbox y="5" x="5" h="70" w="160"/>
  </glyph>

  <arc target="s" source="g" id="a" class="consumption">
   <start y="55" x="70"/>
   <end y="102" x="70"/>
  </arc>
  <arc target="j" source="r" id="u" class="production">
   <start y="138" x="70"/>
   <end y="155" x="70"/>
  </arc>
  <arc target="f" source="h" id="v" class="catalysis">
   <start y="95" x="120"/>
   <next y="120" x="120"/>
   <end y="120" x="70"/>
  </arc>
 </map>
</sbgn>
```

Figure 2.14: A SBGN-ML process description map of glucose import from extracellular space with its SBGN-ML code to the left. The color-highlighted code on the left corresponds to the same-colored elements on the right side. Taken from Ref [88].

As mentioned in 2.1, the SBML format officially supports the storage of network layouts [88] and how the network is rendered [55]. These packages do not contain information about what an object in the representation represents unless the associated systems biology ontology [53, 89] (SBO) value is referenced. The immediate benefit of using SBML to encode visual information is that one file contains both the computational model and graphical representations of its network and pathways [52].

2.3.4 Visualization approaches, standard formats, and software tools

GEM visualizations should allow for the identification of metabolites, reactions, and pathways [90]. There are two approaches used to visualize metabolic networks and pathways: Automated layouts generated by the use of algorithms and manually drawn maps [90]. Most GEMs contain large amounts of metabolites and reactions and their visualization often requires automation but this tends to result in densely clustered nodes, making identification of metabolites and reactions hard without subsequent manual editing of the layout. For very large networks, this process can take a long time [91]. Manually drawn maps allow for the identification of metabolites and reactions as their layouts prioritize legibility [92] at the cost of time. In the following subsections, a selection of tools used to generate automatic layouts and manually drawn maps are presented.

Automatic layout tools

Automatic layouts used to visualize graphs are generated by layout algorithms. These algorithms prioritize the organization of the network in question and aim to position nodes and edges in accordance with a desired visual style. Some prioritize organizing the network to present hierarchical information at the cost of low-level detail [73] while others employ a physics-based simulation [93] to the network to achieve a force-directed layout that achieves spatial separation of network structures and reduction of edge-crossings that would otherwise interfere with interpreting the relational information they convey [94] in highly connected networks [93].

BioFabric

BioFabric [95] is a software tool that visualizes networks horizontally. This network alignment facilitates the identification of nodes with similar connectivity while simultaneously displaying the network in its entirety [96], as shown in 2.15. The network is given a tabular layout that positions nodes horizontally and in a sequential manner with edges represented as vertical lines that connect to source and target nodes.

BioFabric layouts avoid the “hairball effect” that characterizes traditional automatic layouts applied to large networks with clarity as there are no edge crossings. Each edge is given a dedicated “column” and the width of the visualization scales with the number of edges [95]. BioFabric’s layout is conducive to recognizing similarly connected nodes but does not lend itself to the identification of network structures such as clusters nor does it support the integration of attributes to enrich the visualization [96].

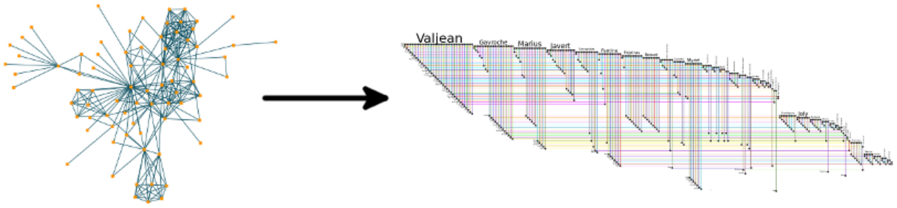


Figure 2.15: BioFabric visualizes complex networks horizontally with each edge as a vertical line between source and target nodes. Taken from Ref [95].

Cytoscape

Cytoscape [81] is the most widely used network visualization and analysis tool used in biology [91, 97, 98] thanks to its ability to integrate biological data in its network properties [99] and display these parameters by changing visual elements of nodes or edges to reflect a desired property in the integrated data via VizMapper [20]. Cytoscape supports a wide range of network formats and an extensive library of automatic layout algorithms which can be expanded upon in the form of downloadable plugins [100]. It supports manual modification and placement of nodes and provides a robust network analysis tool that integrates analysis results as network attributes.

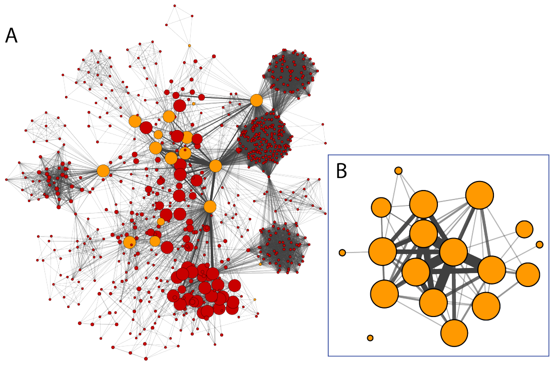


Figure 2.16: Example networks rendered in Cytoscape with a prefuse force-directed layout. Taken from Ref [81]

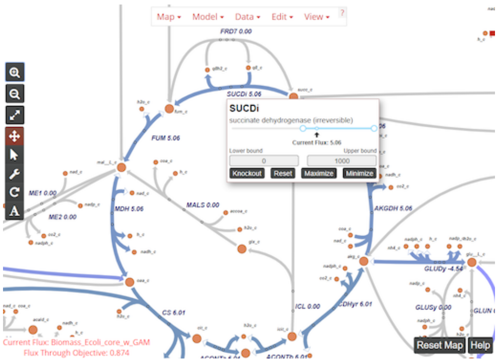
Cytoscape has built-in support for SBML and several systems biology plugins have been developed to enrich SBML-derived networks with external databases [101] and convert Cytoscape network visualizations to standard visualization formats such as SBGN-ML [102].

Manual layout tools

Manually drawn metabolic networks and pathway maps yield interpretable layouts where metabolites and reactions are identifiable. Creating these maps by hand is the result of matching reactions with participating metabolites in a GEM and positioning these elements in a manner that is both readable and does not occlude other elements [90]. This is a tedious and time-consuming process but these visualizations are often able to convey insights more clearly than visualizations with automated layout algorithms [90].

Escher

Escher [59] is a web application that creates visualizations of metabolic pathway maps in a semi-autonomous manner by utilizing information in external databases [15] and GEMs to propose pathways to be visualized with associated reaction, metabolic and genetic data.



SUCDi
succinate dehydrogenase (irreversible)
Current Flux: 5.06
Lower bound: 0
Upper bound: 1000

Current Flux: Biomass_Ecoli_core_w_GAM
Flux Through Objective: 0.874

Figure 2.17: Escher visualization of the TCA cycle from the *E. coli* central metabolism model. Taken from Ref [59].

Building pathways and sequences of reactions are made simple by proposing known reactions connected to a metabolite prompt. Subsequent reactions that are known to follow the previously specified reaction are then auto-suggested.

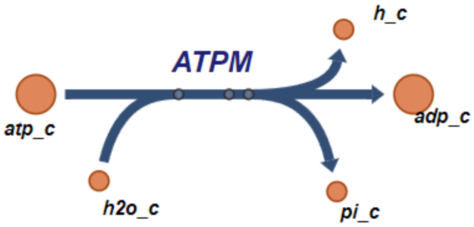


Figure 2.18: The ATP maintenance reaction (ATPM) with substrates; cytoplasmic ATP (atp_c) and H2O (h2o_c) and products; cytoplasmic hydrogen (h_c), cytoplasmic ADP (adp_c) and cytoplasmic inorganic phosphate (pi_c). Metabolite labels are abbreviated and follow a BIGG Models systematic naming convention. Taken from Ref [59].

This ease-of-use enables rapid generation of clear and concise visualizations of metabolic pathways that are enriched with information from external databases

[6, 15, 103–105]. Escher visualizations can be exported and their layouts encoded in JSON format can be converted to SBGN-ML and SBML format by the use of a dedicated client, EscherConverter [59].

CellDesigner

CellDesigner [16] is an editor for diagrams representing systems biology associated information e.g. metabolic pathways and gene regulatory networks. It provides a standardized set of symbols [87] to convey molecular entities, boolean logic gates, and regulatory signals unambiguously [16, 106]. It supports SBML level 2 [107] as of version 4.4 (SBML level 3 support due in version 4.5).

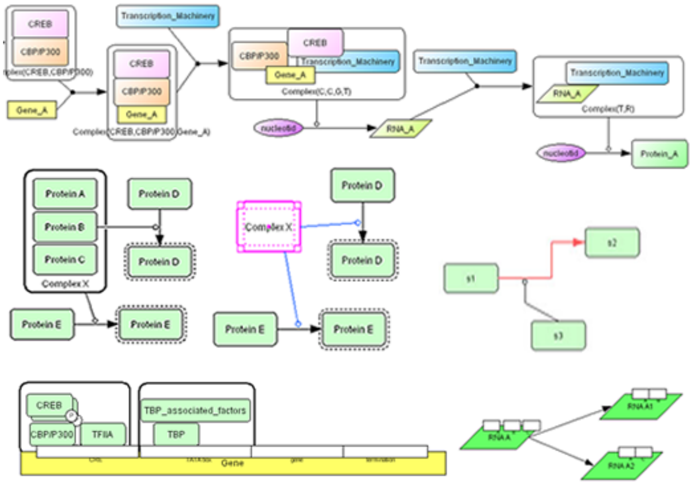


Figure 2.19: An example of a protein-protein-interaction network rendered in CellDesigner. Taken from Ref [16]

CellDesigner has a unique feature over other visualization software in that it can link in-client simulation data to the visualized network. This allows users to utilize CellDesigner as visualization and analytical platform [20, 106, 108].

2.4 Molecular descriptors

2.4.1 Molecular descriptors

Molecular descriptors [28] are quantitative representations [109] of various chemical properties of a molecule that are generated by algorithms from a representation of a molecule and contains information about the physical and chemical information of the molecule [110]. Molecular descriptors play a fundamental role in drug design [111, 112], medical research [113], and evaluation of the environmental toxicity [114] for a given compound. They are also the basis for the virtual screening of molecular datasets and are used in the prediction of biological properties and behavior [115] in organisms [116, 117]. Computational chemistry and cheminformatics facilitates such applications by calculating these molecular properties. A typical molecular graph represents the topology of a molecule and some descriptors are calculated by algorithms processing the topology of the molecular graphs or the correlation of atoms pairs. These are called topological descriptors and autocorrelation descriptors respectively and are categorized as 2D descriptors. Other, more complex descriptors are derived from a geometrical representation of the molecule. Geometrical descriptors require that every atom and chemical bond in the molecule(s) has three-dimensional coordinates [118, 119] in their molecular representation. The following subsections will present a selection of descriptor categories; constitutional, autocorrelation, and geometrical descriptors as well as molecular fingerprints. This selection does not cover every type of descriptor and is merely a simplified overview of the most commonly used descriptor categories. Subsequent chapters will include descriptors and descriptor categories which are not included in this section but a brief explanation will accompany them when they are mentioned. A more thorough list of molecular descriptors and their categories are available in Ref [120].

Constitutional descriptors (0 + 1D)

Constitutional descriptors comprise features that are aggregates of the individual atoms that a molecule is made up of [109]. Examples of constitutional descriptors are the sum of atoms, the sum of different chemical bonds, and molecular weight. Constitutional descriptors are fast to compute as they usually quantify properties which are the sum of constituent atoms and bonds. This infers that two or more uniquely different molecules can have the exact same constitutional descriptor values. Due to this fact, constitutional descriptors are usually used in conjunction with different, more complex descriptors, such as autocorrelation descriptors or geometrical descriptors [121].

Some constitutional descriptors are calculated by summing the number of atom types or structural motifs that exist in the molecule and is often stored as integer values. Examples are the number of hetero atoms, the number of non-carbon atoms that exist in the molecular backbone [122], or the number of ring structures [122].

Autocorrelation and topological descriptors (2D)

Where constitutional descriptors can be calculated from the molecular composition without geometrical or topological information, the calculation of autocorrelation descriptors requires a molecule to be represented as a graph $\boldsymbol{G}$ and the physical properties of atoms be values given to vertices $\boldsymbol{N}$ of $\boldsymbol{G}$ and chemical bonds represented as edges $\boldsymbol{V}$ of $\boldsymbol{G}$. While more complex than constitutional descriptors, they encompass characteristics of a molecule that can be extracted from its size, electron distribution, and two-dimensional shape. Autocorrelation descriptors [123] are the most frequently used descriptor type and carry information about a molecule's topological profile. Given the light computational load required for autocorrelation descriptors, they are often used to compare millions of different molecules in large datasets [124].

An example of a commonly used autocorrelation descriptor is the ratio of a given molecule's surface that is hydrophobic. The hydrophobicity of a molecule can be quantitatively described by the polar partition coefficient, the descriptor MolLogP [125]. It represents the ratio of the molecule that will exist in the uncharged part of an water and n-octanol solution by an atom-based Crippen's approach calculation [123, 126, 127]. MolLogP can infer how well a compound binds to a given protein and its ability to traverse the membrane of a cell [109]. Molecular refractivity or MolMR is a quantitative measure of the bulkiness of a molecule; it is calculated by dividing molecular weight by the density of the molecule which is derived from the topological profile of the molecule [109].

Geometrical descriptors (3 + 4D)

Geometrical descriptors mainly encompass emergent physio-chemical properties that arise from their three-dimensional configuration [128]. This category of molecular descriptors requires molecular representations that contain spatial and coordinate information regarding its constituent atoms. Due to some molecular structures that can exist in multiple stable conformations, this category of descriptors will have numerical differences between these conformations which convey their unique characteristics. Three-dimensional descriptors are widely used to evaluate if a drug's pharmacokinetic profile satisfies desired drug property requirements when screening large numbers of possible drug candidates [112].

Molecular fingerprints and quantification of multiple molecular descriptors

Molecular fingerprints are a method of representing molecules as a mathematical objects where they convey the presence or absence of a chemical structure. This method allows for rapid screening and mapping of the chemical structures that the molecule possesses. The sheer number of different molecules that are of interest as drug candidates increase as they aim to target ever more complex biological processes. Molecular fingerprints describe the structural motifs that a molecule possesses in a vector, where the presence or absence of a structure is denoted by a 1 or 0, respectively [129].

The most commonly used molecular fingerprint is the Morgan fingerprint [129], also known as the extended-connectivity fingerprint (ECFP4). Morgan fingerprints are well suited for describing the characteristics of small molecules but lack the fidelity to convey global features such as the size and shape of larger molecules. Computation of some molecular fingerprints such as MIniHashed fingerprint, MHFP6 [130] requires a molecular representation format that encodes stereochemistry such as InChI or isomeric SMILES to calculate an atom-pair based molecular fingerprint [130].

While not categorized as a molecular fingerprint, the descriptor Quantitative estimation of drug-likeness [131] (QED) incorporates several molecular descriptors in its calculation; molecular weight, MolLogP, topological/total polar surface area (TPSA), the number of hydrogen bonds, donors and acceptors, the number of aromatic rings and rotatable bonds [128, 131]. QED is a viable approximation for the comparison or estimation of a drug's oral bioavailability and is commonly used in evaluating drug candidates at a pre-clinical stage [128].

2.4.2 Molecular descriptor applications

“It is obvious that there must exist a relation between the chemical constitution and the physiological action of a substance [. . .], but as yet scarcely any attempts have been made to discover what this relation is. [. . .] it might be supposed that a careful examination and comparison of known facts would lead to the discovery of some empirical law by means of which we could deduce the action from the chemical constitution.” [132].

The quote is a citation from Nathan and Brown’s article in the Journal of anatomy and physiology in 1868. Molecular descriptors are the product of the purported empirical laws proposed in their article and quantitatively express information about a molecule’s chemical properties and enable the prediction of chemical properties from its structure and composition. Applications such as quantitative structure-activity relationships (QSAR) studies of molecules require molecular descriptors. By using QSAR, it is possible to evaluate the toxicity of chemical compounds in the environment or to organisms as well as to screen molecules for their potential as drug candidates. Two applications of molecular descriptors are described below and contain concepts and principles that will be expanded upon in subsequent sections.

Quantitative Structure-Activity/Property Relationships

Each day, new chemicals are created and theorized. This constant influx of new chemicals makes their experimental characterization a gargantuan task. Cheminformatic tools are necessary to meet the demand for evaluating and analysis of their molecular descriptors. Quantitative Structure-Activity/Property Relationships (QSAR/QSPR) are two, well-used, approaches in the field of computational, medical, and environmental chemistry [133]. QSAR and the related quantitative structure-property relationships (QSPRs) are enabled by the cheminformatic techniques that allow for the linking of molecular structure to molecular behavior. The basis for these approaches is that the activity and behavior of chemicals are derived from their structure. By extension, chemicals with similar descriptor values will have similar activity. QSAR assumes that a molecule’s biological, chemical and physical properties are invariably linked to its structure and composition and that the structure contains the information necessary to predict the behavior of a molecule. The molecular structure must be defined clearly in order to facilitate these approaches. A molecular representation, in cheminformatics terms, is the information required for the generation of a theoretical descriptor based on the structure using cheminformatics algorithms.

Being able to identify compounds that exhibit complementarity with regards to steric configuration, electrostatic, and hydrophobicity is of great interest in the *in silico* screening of drug candidates. Finding candidates that complement or possess bioactive conformations is dependent on the three-dimensional quantitative structure-activity relationship (3D-QSAR) methodologies [112]. These 3D-QSAR methodologies require a molecular representation that possesses spatial and coordinate information of the constituent atoms. This information is either determined experimentally or computationally. Several computational approaches to ascertaining possible and stable structural conformations exist [17, 134, 135] and require extensive use of databases with previously confirmed molecular structures to approximate a given molecule's three-dimensional structure from its 2D molecular representation.

Descriptor selection

Cheminformatics allows for the computation of many molecular descriptors for a given molecular representation but not all descriptors are relevant to each usecase [17]. Descriptor selection is an important of many QSAR workflows [133] and in most cases where molecular descriptors are used. As a rule of thumb, this selection removes descriptors whose values are missing as a result of insufficient data required for its calculation, has constant values for the dataset, and descriptors that are highly correlated with each other. Due to the high dimensionality that comes with the number of available molecular descriptors, multivariate analysis, and dimensionality reduction methods such as principal component analysis (PCA) are often applied [136]. Due to the number of descriptors available, it is prudent to select descriptors based on the molecules they are set to describe. A good descriptor should, in addition, describe physiochemical properties, have a low correlation with other descriptors, and exhibit small degeneracy, i.e. their values should be numerically distinct from each other from molecules with different structures [25], and not contain overlapping information.

2.4.3 Codification of chemical information

Information about chemical structures is represented as molecular graphs on computers. These molecular graphs are reconstructed from molecular representations that contain the necessary information for the generation of a molecular graph. Molecular representations are used as input for cheminformatic software and tools. The two most widely used formats are the Simplified Molecular-Input Line-Entry System [25] (SMILES) and IUPAC chemical Identifier [24] (InChI), these two formats will be introduced in the two following subsections. Each subsection ends with an evaluation of their possible use as the molecular representation chosen for this project.

SMILES

SMILES stands for Simplified Molecular Input Line Entry System. The SMILES notation system was initially developed in 1988 and represents chemical structures in a way that can be read by a computer for the purpose of cheminformatics. Figure 2.20 illustrates the basic notation system for SMILES is a set of five rules. An extended set of rules exists for stereochemistry and isomers but these are omitted for brevity.

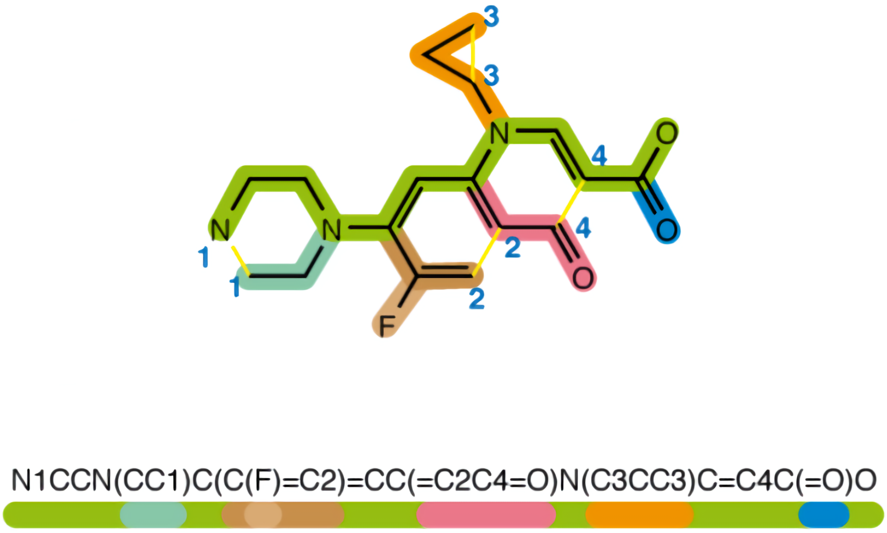


Figure 2.20: Schematic representation of a SMILES string generated for ciprofloxacin. Each element of the molecular structure has a color code and corresponds to a part of the SMILES string which encodes that element, i.e. the light green color represents the molecular backbone of ciprofloxacin. Adapted from [137].

Rule one denotes the atoms and bonds inside a chemical structure by its atomic symbol and bonds. For atoms in an aromatic structure, their atomic symbol is decapitalized. Differentiation of bonds is achieved by the different symbols shown in Table 2.1.

Table 2.1: Examples of how different types of chemical bonds are represented in the SMILES format.

Bonds	**Symbol**
Single bond	-
Double bond	=
Triple bond	#
Aromatic bond	*
Disconnected structures	.

The second rule explains how chains are represented by atomic symbols and bonds, which is shown in Table 2.2. The SMILES notation eschews the inclusion of hydrogens (H) as any unspecified bonds or valences will be filled with hydrogens when read by cheminformatic tools that follow the SMILES specifications.

Table 2.2: Denotation of simple chains using SMILES symbolism from rule 1 and 2.

SMILES	Chemical formula	Chemical name
C-C-C	$CH_3CH_2CH_3$	Propane
CCCC=O	C_4H_8O	Butyraldehyde

Rule three explains, demonstrated in Table 2.3 the use of parentheses to denote branches. Branches or single atoms in parenthesis are placed right after the un-parenthesized atom it comes from.

Table 2.3: The use of parenthesis denotes branching of molecules in SMILES. the SMILES string for 2,3-Butanediol indicates the branching of 2 oxygen atoms and 2,2,4,4-tetramethylpentane indicates how there are 4 molecular branches, denoted by parenthesis encapsulating the atoms that branch.

SMILES	Chemical formula	Chemical name
CC(O)C(O)C	$C_4H_{10}O_2$	2,3-Butanediol
CC(C)(C)CC(C)(C)C	C_9H_{20}	2,2,4,4-tetramethylpentane

Rule four, demonstrated in Table 2.4 explains how ring structures are denoted by marking the opening and closing ring atoms. Note that ring structures made up of aromatic bonds are denoted with “ * ”, as described in rule one.

Table 2.4: The SMILES strings for benzene and glucose demonstrate how aromatic and different ring structures are encoded by SMILES rule 4.

SMILES	Chemical formula	Chemical name
c1ccccc1	C_6H_6	Benzene
C(C1C(C(C(C(O1)O)O)O)O)O	$C_6H_{12}O_6$	Glucose

Rule five demonstrated in Table 2.5 explains how charged atoms are identified within the chemical structure. It is represented via curly brackets. The integer charge value can be included or not as; ({-1}), ({-}), but for charges that are larger than 1, its value must be included after the charge sign.

Table 2.5: The SMILES string for trimethylammonium demonstrates how charged atoms are encoded in rule 5 and that the charge of nitrogen is positive.

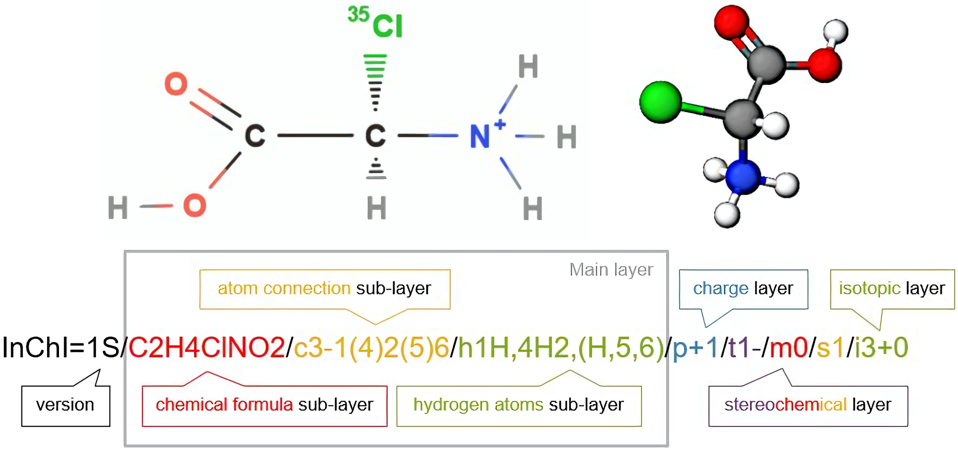
SMILES	Chemical formula	Chemical name
C[NH+](C)C	$C_3H_{10}N$	Trimethylammonium

The SMILES project is not an open project and has led to different software tools creating their own algorithms to generate SMILES [138]. This can lead to one SMILES string representing two different molecules. The SMILES format supports a limited number of stereochemistry categories when compared to that of InChI, which will be presented in the next section. It is for these reasons that the SMILES format was not chosen as a molecular representation for this project.

InChI

The International Union of Pure and Applied Chemistry [139] (IUPAC) saw the need for a standardized and un-ubiquitous chemical identifier to be used in the digital era. It was developed in conjunction with the U.S. National Institute of Standards and Technology [140] (NIST). Borne of this cooperation between IUPAC and NIST came the International Chemical Identifier (InChI), a set of nomenclature rules that can be used to generate unique molecular representations for any chemical substance. The InChI Trust [141] and the IUPAC InChI subcommittee was formed in 2010 with the task of maintaining the standard and developing it further.

InChIs are layered hierarchically, with each layer containing more details about the molecule it is generated from. The layers, as shown in Figure 2.21: the chemical formula, atom connection, hydrogen atoms, charge, stereochemical and isotopic layers. Within the InChI string, each layer is separated by a forward slash.



35Cl
O
C
C
N+
H
H
H
H
H
H – O
Main layer
atom connection sub-layer
charge layer
isotopic layer
InChI=1S/C2H4ClNO2/c3-1(4)2(5)6/h1H,4H2,(H,5,6)/p+1/t1-/m0/s1/i3+0
version
chemical formula sub-layer
hydrogen atoms sub-layer
stereochemical layer

Figure 2.21: Schematic representation of the main layers for a standard InChI of (R)-carboxy(chloro)methyl]azanium. Adapted from online material for the course “Chem1102” taught at Fordham University [142]

Some InChI’s can become excessively long [143]. InChI’s can be hashed via a dedicated hashing algorithm to produce a fixed-length string called InChIKeys. InChIKeys are 27 uppercase letters. InChIKeys contain layers but the 27-character limit confers less resolution and information about the molecule it is generated from when compared with a full InChI string. A potential upside of this format is how its format lends itself to websites as it does not contain special characters.

A drawback of the InChI notation is that it is primarily used as input for cheminformatic tools and is not designed to be readable by humans, as demonstrated in 2.21, contrasting the SMILES notation in 2.20. There exists only one algorithm for generating InChI's and supports more stereochemistry types than SMILES [138]. It is for these reasons that InChI's was chosen as the molecular representation format for metabolites to calculate molecular descriptors from.

2.4.4 Cheminformatics

Cheminformatics refers to the application of computer science to the field of chemistry through the use of *in silico* methods [144]. Cheminformatics represents molecules as molecular graphs, a concept derived from graph theory. Atoms are nodes and edges are the chemical bonds that connect the nodes to form molecules. SMILES and InChI are the most common formats from which cheminformatics tools can create molecular graphs [145]. Drug discovery, drug design, and QSAR are typical tasks where cheminformatics is applied. Cheminformatic tools process molecular graphs of molecules and can visualize molecular structures in three dimensions, highlight chemical groups or visualize electron density or other molecular attributes across a molecular structure and calculate a wide array of molecular descriptors and fingerprints, shown in Figure 2.22.

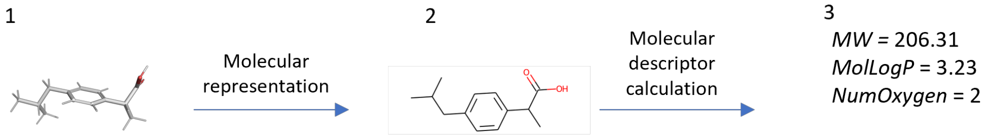


Figure 2.22: A diagram exemplifying how cheminformatic tools are used to calculate molecular descriptors (3) from a molecular representation (2) that is generated from a molecule (1).

Through the use of these descriptors, machine learning methods can be used to model QSAR properties and develop AI-aided models to predict how changes in chemical properties can lead to changes in biological behavior *in silico* [112].

Listed below are several cheminformatic tools that were considered for use in calculating molecular descriptors for this research project, along with evaluations of their suitability based on the following criteria: open-source, easy to use, and capable of calculating a large number of descriptors.

RDkit

RDkit is a widely used open-source cheminformatics tool [122]. Its core methods and data structures are written in C++ but are accessed in Python via a wrapper generated via Boost.Python [146]. Some examples of RDkit functionality are the reading of molecules from different molecular representation formats, substructure searches, substructure transformations, fingerprinting, molecular similarity searches (see 2.23), molecular descriptor calculation, and drawing of molecules and chemical reactions with substructure highlighting [122].

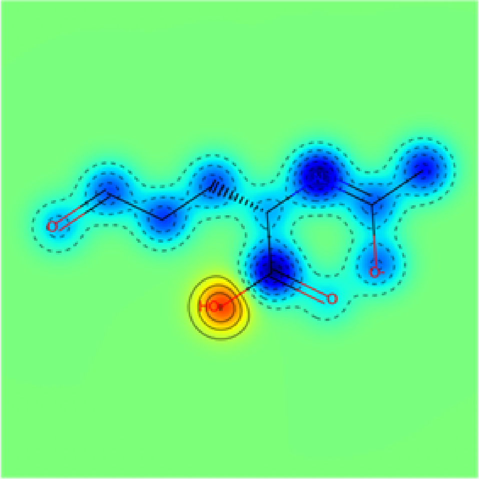


Figure 2.23: N-acetyl-L-glutamate 5-semialdehyde with Gasteiger charge similarity maps drawn with RDkits similarityMaps function.

RDkit is in active development at Novartis Institutes for BioMedical Research [147] and is a component of many widely used software packages such as DeepChem [148] and features full integration in the KNIME Analytics Platform [149].

RDkit calculates 215 constitutive and autocorrelation descriptors from two-dimensional molecular representations and 5 geometrical descriptors if the molecular representation format supports and contains atom coordinates or conformers in a supported file format like Spatial Data File (.sdf). It also supports a suite of solutions for rendering molecules 3D visualizations within Jupyter Notebooks of metabolites using a local PyMol instance [122, 150], shown in 2.24D.

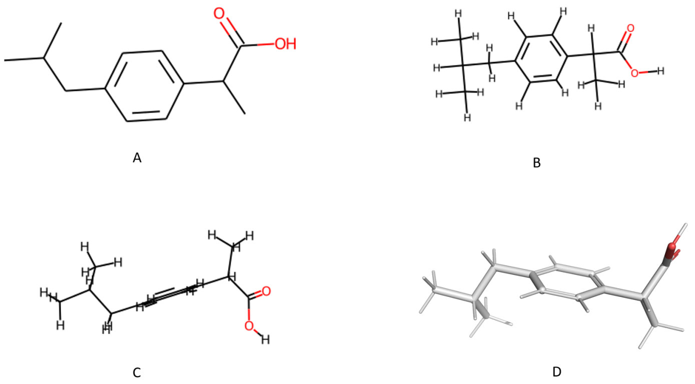


Figure 2.24: (A) planar drawing of ibuprofen without added hydrogen atoms. (B) Ibuprofen drawn linearly with hydrogen atoms. (C) Two-dimensional render of an ibuprofen conformer. (D) A three dimensional visualization of ibuprofen.

Due to its ability to calculate a wide range of molecular descriptors and the fact that it is an open-source tool, RDkit was chosen as the cheminformatic tool to calculate molecular descriptors.

Open Babel

Open Babel is an open-source cheminformatic tool originally developed to interconvert between chemical structure formats [134]. It supports the interconversion of 110 different chemical formats, accommodating interoperability of different programs and workflows. Open Babel supports the calculation of 16 numerical molecular descriptors, several molecular fingerprints, and molecular similarity searching of large molecule datasets. Open Babel can approximate a three-dimensional structure from a two-dimensional molecular representation. Open Babel was not chosen to calculate the molecular descriptor used in this project as the number of descriptors offered by RDkit both supplanted and included the descriptors supported by Open Babel.

alvaDesc

alvaDesc is a cheminformatic software tool that is engineered for the calculation of molecular descriptors [151]. It is the successor to the widely used but now discontinued DRAGON software [110]. alvaDesc calculates more than 5471 descriptors and 3 fingerprints. In addition to calculating descriptors, it also provides a suite of tools for exploratory analysis of molecular datasets by visualization, multivariate statistical analysis, and molecular structure verification by referencing PubChem [152]. Unlike RDkit and Open Babel, alvaDesc requires a paid license.

Chapter 3

Methodology

This chapter describes the methods used in this project. The first section details the collection of molecular representations of metabolites from GEMs and the calculation of their molecular descriptors. The second section describes the descriptor selection process and associated methods are presented in brief. The third section details the process of visualizing metabolic networks and pathways with biochemical coordinate layouts. Finally, the process of exporting the graph layouts to community standard formats is presented.

3.1 Data collection and descriptor calculation

By default, SBML formatted GEMs do not contain non-hashed molecular representations (InChI Key) in the annotations of metabolites [58]. The subsequent subsections include the methods in which the metabolite identifiers are collected, how the molecular representations are extracted from a reference database, and finally the process of calculating molecular descriptors from molecular representations.

Data collection

This step involved the acquisition and use of multiple datasets from two databases; BiGG Models and MetaNetX. Discrepancies in column titles across datasets were removed by their standardization, presented in Table 3.1. This standard was kept across datasets generated to ensure compatibility.

Table 3.1: Standardization of column titles in downloaded and generated datasets. The following column title modifications across all downloaded data.

Column type	Column title	New column title
MetaNetX metabolite identifiers	#ID, id	MNXref
BiGG identifiers	source, BiGG_id	BiGG

The GEMs utilized in this project, *E. coli* central metabolism [5] & *i*ML1515

[47] models were downloaded from BiGG Models [15] in their SBML (.xml) format. Table 3.2 presents an overview of their metrics.

Table 3.2: Overview of model metrics. Metrics retrieved from their respective entries on BiGG Models [15]

Model name	Organism name	Reactions	Metabolites	Unique metabolites	ORF
iML1515	*Escherichia coli*	2705	1877	1169	1515
e_coli_core	*Escherichia coli*	95	72	55	95

3.1.1 Chemical library mapping of GEMs

GEMs does not typically include representations of a metabolites molecular structure. Chemical library mapping of GEMs is the process of mapping chemical library information, such as the MNXref namespace, to metabolites in GEMs [60]. MetaNetX [60] was utilized to perform this function as it supports the upload of SBML formatted GEMs. Chemical compounds and metabolites are given a prefix (NADH is MNXM10) and chemicals or reactions whose products do not participate in other reactions are removed [103]. The results of the chemical library mapping were downloaded as comma-separated value (CSV) files and column titles were standardized as presented in Table 3.1.

3.1.2 Collection of molecular representations

The molecular representations, SMILES & InChI, were collected by matching MNXref to the chemical property database provided by MetaNetX [153]. This was done in python with a local copy of the chemical property database. Separate CSV files for InChI and SMILES were saved for each model to be used for molecular descriptor calculation from both representations.

3.1.3 Alternative approach to the collection of molecular representations

Performing the step described in subsection 3.1.2 can be performed using the MetaNetX website, which eliminates the need for programming knowledge and the need to download the chemical property database. MetaNetX hosts a selection of curated GEMs, which can be selected with the “Pick from repository” option from the toolbox. Both the core and *i*ML1515 models are hosted and can be selected after single queries and ticking the “Select” box for both models. From this point, the process of mapping the chemicals is functionally identical as laid out in 3.1.1. The collection of molecular representations is achieved via using the “MNXref ID mapper” tool and copying the MNXref column from 3.1.1 in the text field. The result of this mapping contains molecular representations (SMILES, InChI) and external database identifiers (BiGG [15], CheBI [104], KEGG [6], MetaCyc [154], and HMDB [155]) where available. This data can be downloaded either as a tab spaced variable file (TSV), JSON, or hypertext markup language (HTML) file. A significant drawback to this alternative approach is that the MNXref ID mapper tool is limited to queries containing 100 identifiers at a time, necessitating additional repetitive steps and the possibility of human errors when processing larger GEMs.

3.1.4 Molecular descriptor selection

To calculate molecular descriptors, the cheminformatic tool RDkit [122] was used. For each model, 215 constitutive and topographical molecular descriptors were calculated from SMILES and InChI representations to facilitate the comparison of numerical differences in molecular descriptor calculations between representation formats. Calculation results were stored in separate CSV files for each format, resulting in two files for each model, and were reviewed to investigate if any molecular descriptors failed to compute.

3.1.5 Case studies

In order to evaluate different layouts generated using different biochemical coordinates across different pathways and networks, 7 case study datasets were created. These case study datasets were:

- Metabolic networks
  - *E. coli* central metabolism
  - *i*ML1515
- Metabolic pathways
  - The TCA cycle
  - ATP biosynthesis
  - CTP biosynthesis
  - GTP biosynthesis
  - Histidine biosynthesis

The IMP biosynthesis pathway (see Figure 2.3) is integrated in the purine biosynthesis datasets; ATP and GTP biosynthesis (see Tables B.2 and B.4).

For the purpose of network visualization, edgelists were compiled using the metabolic modeling package ReFramed. The edgelist consists of pairs of nodes that indicate that there is a link between them. In order to create metabolic pathway datasets, metabolite entries were copied from the *i*ML1515 InChI-based molecular descriptor file. In a dataset, excluding column titles, each row represents one metabolite and contained 225 columns. These columns were:

- Molecular descriptors (215)
- Identifers and descriptions (5)
  - BiGG universal identifier
  - MNXref identifier
  - Full name
  - Description of molecule
  - ChEBI identifier
- Molecular representations (3)
  - InChI
  - InChIKey
  - SMILES
- Edgelist (2)
  - Source node
  - Target node

All case study datasets are available in Table A.1, Supplementary data 3.

3.2 Descriptor selection

A key step in a descriptor selection process is the identification of features that capture as much information while using as few of them as possible [109]. Due to the number of descriptors available (n = 215), it was important to identify those which were orthogonal to one another and to remove those which were correlated, empty, or did not show any variation across datasets. To remove empty features, descriptors that had 90% of their values equal to zero were removed. Spearman rank correlation coefficients were calculated before and after the removal of empty descriptors. Identification of orthogonal descriptors was performed by analyzing descriptor correlations and applying principal component analysis (PCA) and then determining the descriptors that best capture variance in the chemical space by analyzing component loadings. The seaborn [156] statistical visualization package was used to visualize correlation matrices and principal component loadings. In contrast to typical QSAR workflows, the selection of descriptors reflecting chemical features with a clear biological interpretation was favored. An emphasis was placed on biologically relevant descriptors so that the position of each node could convey intuitive biological information about the metabolite and its neighbors.

A step-by-step process was used to select the molecular descriptors. As a result of analyzing principal component loadings, a final selection of biologically relevant descriptors was made based on the identification of descriptors that best captured variance in the biochemical space.

3.3 Layout visualization

3.3.1 Layout visualization

The network analysis package NetworkX [157] was used to create graphs of case study datasets. Different pairwise combinations of selected descriptors were used as x- and y- cartesian coordinates to generate graph layouts for all case studies. The plotting package Matplotlib [158] was used to visualize the graph layouts.

3.3.2 Layout evaluation

It is necessary to assess the layouts themselves to determine whether they are biologically meaningful and understandable in order to evaluate the combinations of descriptors that generate the most intuitive layouts. An intuitive and comprehensible layout is not difficult to interpret. Bad layouts can frustrate or confuse viewers if they place objects near one another and result in node stacking and edge crossings. There may be a number of factors that can negate the benefits of a visualization, including distractions, overlapping information, or simply a large amount of information, as demonstrated in Figure 2.13 and Figure 2.11

An intuitive visualization would have a network layout that allows exploration and identification of known biological patterns so insight can be acquired, whereas a non-intuitive visualization would have a layout that arranges information in a way that is extremely hard to understand. In this project, a biologically meaningful layout can be characterized as being able to identify metabolic pathways in a network or biological patterns based on the position of metabolites when visualized. The opposite is a non-intuitive layout, which is characterized by the aggregation of metabolites in such a manner that their placement prevents the interpretation of network relationships between metabolites on a local or global level.

Layout evaluation was performed by characterizing visualizations of network and pathway as either biologically meaningful or non-intuitive. Metabolic pathways were also highlighted in the *i*ML1515 case study visualization network to allow for their evaluation on a global scale.

3.4 Layout export

To export generated layouts to formats that can be read by commonly used network visualization software, NetworkX's export to Cytoscape functionality was used. This functionality did not produce a file that was readable by Cytoscape and required several steps to ensure full compatibility. A step-by-step guide on how to do this as well as an example file to validate this process is provided in Table A.1 Supplementary data 4, Appendix A. Upon completion, the file can be read by Cytoscape and the graph layout is visualized by mapping the node x- and y- positions to the “x” and “y” node attributes. In addition, the complete set of molecular descriptors was imported to Cytoscape and exported in the Graph Markup Language (GraphML). This format stores node attributes such as the imported molecular descriptor data and can be converted to the systems biology community standard format SBGN-ML by using the ySBGN [159] tool.

Chapter 4

Results

4.1 Descriptor calculation

4.1.1 Collecting molecular representations for descriptor calculation

Not every metabolite in the *E. coli* central metabolism and *i*ML1515 models were chemically mapped. 55 metabolites were mapped to the MNXref namespace from the core metabolism model and 1169 from the *i*ML1515 model, an overview of the results of the chemical library mapping is listed in Table 4.1. As presented in section 3.1.1, chemicals or reactions whose products do not participate in other reactions in the model were not included in the mapping process, and metabolites were listed in multiple compartments; cytoplasm, extracellular space, and the periplasm were merged and represented with one universal entry. These results are available in Table A.1, Supplementary data 1.

Table 4.1: Comparison of results from the chemical library mapping and collection of molecular representations for both GEMs. Rounded percentage of molecular representation coverage in parenthesis.

GEM	Metabolites	Mapped metabolites	InChI	SMILES
E. coli central metabolism	72	55	55 (100%)	55(100%)
*i*ML1515	1192	1169	950 (81%)	1012(86%)

All molecular representations, SMILES, and InChI were collected for the mapped metabolites in the *E. coli* central metabolism model. 157 (13%) of the 1169 mapped metabolites from the *i*ML1515 model were unable to be matched with a molecular representation, 950 (81%) metabolites could be matched with an InChI and 1012 (86%) metabolites had associated SMILES.

Further inspection of the 157 mapped metabolites from the *i*ML1515 model which was unable to match with an associated molecular representation revealed two trends. Large and complex metabolites with one or several of the following cofactors bound; coenzyme A, acyl substructure, and acyl carrier protein (ACP) cofactors have no molecular representation formats available. An excerpt of metabolites that followed this trend is listed in Table 4.2, and full results are available in Supplementary data 1, Table A.1.

Table 4.2: A sample of metabolites without any chemical representation following chemical library mapping. Excerpt from "Metabolites_lacking_chemical_representation.tsv" in Supplementary data 1, Table A.1. It is unclear how the question mark symbols were incorporated into the metabolite names, as a compound similar to MNXM1094051 already exists in PubChem under the name "alpha-D-ribose 1-methylphosphonate 5-phosphate" [160].

MNXref	**Metabolite**
MNXM1094051	?-D-ribose-1-methylphosphonate 5-phosphate
MNXM1094075	Poly-?-1,6-N-acetyl-D-glucosamine
MNXM1094075	Poly-?-1,6-N-acetyl-D-glucosamine
MNXM1094141	1-(?-D-sulfoquinovosyl)glycerol
MNXM1101270	Bis-molybdenum cofactor
MNXM725917	3-Hydroxypimeloyl-[acyl-carrier protein] methyl ester
MNXM727020	Enoylglutaryl-[acyl-carrier protein] methyl ester
MNXM728594	3-Oxodecanoyl-[acyl-carrier protein]
MNXM728614	3-Oxotetradecanoyl-[acyl-carrier protein]
MNXM728730	(R)-3-Hydroxydecanoyl-[acyl-carrier protein]
MNXM729358	2-Acyl-sn-glycero-3-phosphoglycerol (n-C18:0)
MNXM739590	Methane

Additionally, many of the MNXref namespaces attributed to metabolites during chemical library mapping were outdated or deprecated. This made them incompatible with the chemical property database from MetaNetX, which contained updated namespaces. To ensure that the number of molecular representations was as accurate as possible, the *i*Ml1515 model hosted on MetaNetX, was used. This model contained a list of metabolites with updated and MetaNetX database-compatible namespaces. Using the previously described methodology, 14 additional metabolites were matched with InChIs, totaling 964, but no new SMILES. To investigate if the metabolites without a molecular representation existed under different names, a list of synonyms for these metabolites was collected. Following a review of literature [161] and the possibility of erroneously including metabolites that are not present in the *i*ML1515 model, this investigation was not undertaken.

Further inspection of the remaining 142 metabolites without molecular representations revealed that some of their MNXref namespace values were placeholders or had been deprecated. Querying the MetaNetX database with these MNXref values returned empty metabolite. The MetaNetX entries for, “1-hexadec-9-enoyl-sn-glycerol 3-phosphate” (MNXM91047) and “3-Hydroxyglutaryl-[acyl-carrier protein] methyl ester” (MNXM741743), did not contain any chemical information but did have links to their respective BiGG database entries. These contained only information on their molecular charge and chemical formula.

Comparing the *i*ML1515 SMILES and InChI files revealed that every metabolite that only had a SMILES representation lacked an InChI associated with them. This comparison also identified 8 invalid SMILES. These were proteins in a reduced or oxidized state and are listed in Table 4.3.

Table 4.3: Metabolites with invalid SMILES strings.

MNXref	Metabolite	SMILES
MNXM12615	Periplasmic disulfide isomerase/thiol-disulphide oxidase (reduced)	“ * “
MNXM12618	Periplasmic protein disulfide isomerase I (reduced)	“ * “
MNXM12772	Protein disulfide isomerase II (reduced)	“ * “
MNXM8620	Fused thiol:disulfide interchange protein (reduced)	“ * “
MNXM97001	Fused thiol:disulfide interchange protein (oxidized)	“ * “
MNXM97003	Periplasmic disulfide isomerase/thiol-disulphide oxidase (oxidized)	“ * “
MNXM97004	Periplasmic protein disulfide isomerase I (oxidized)	“ * “
MNXM97006	Protein disulfide isomerase II (oxidized)	“ * “

In the case of metabolites listed in Table 4.3, the “ * “ symbol does not represent an aromatic bond (see Table 2.2) but the SMILES notation for “unspecified atomic number”, explained in Ref [162]. Consequently, it was prudent to remove metabolites that did not have a valid molecular representation since they could not be used for the calculation of molecular descriptors. The final number of metabolites available in each format for both models are listed in Table 4.4.

Table 4.4: The final number of collected molecular representations following manual curation and corroborating previous results (see Table 4.1) with the *i*ML1515 GEM hosted on MetaNetX.

Model	Metabolites in InChI file	Metabolites in SMILES file
E. coli central metabolism	55	55
*i*Ml1515	964	1004

4.1.2 Molecular descriptor calculation

All 215 molecular descriptors were calculated from available molecular representations for both models and complete *i*ML1515 results from InChI and SMILES calculations are available in Supplementary data 2, Table A.1. Upon inspecting the resultant calculations, several descriptors quantifying intramolecular forces and partial molecular charges had “not-a-number” (NaN) values across both formats in the *i*ML1515 model. The BCUT [163] and EstateIndex [164] families of descriptors had failed to be successfully calculated for 45 metabolites and in lieu of calculated descriptor values, had NaN’s. A complete list of the metabolites with NaN values for these descriptors is made available in Supplementary data 2, Table A.1 and an excerpt is presented in Table 4.5. These metabolites were primarily ionized atoms, such as Co^{2+}, Na^{+} and Zn^{2+} or larger molecules with very complex structures such as adenosylcob(III)inamide-GDP. Molecular descriptor calculation was repeated multiple times and compared with the first calculation results to be able to ascertain if these results were random or indicative of an underlying flaw in their molecular representation. Upon comparison, these descriptors had failed to be calculated for the same group of 45 metabolites that failed initially. A solution to this issue was adapted from an answer from RDkit's chief developer, Greg Landrum, to a similar problem [165]. Upon implementing this solution, all molecular descriptors were successfully calculated for all metabolites in the *i*ML1515 model.

Table 4.5: Excerpt of metabolites that failed to compute BCUT and EstateIndex family of descriptors. Excerpt from supplementary data 2 "*i*Ml1515_Metabolites_failed_Estate_BCUT_family_calculation.tsv".

MNXref	**Metabolite**
MNXM1094276	tellurite
MNXM1101279	adenosylcob(III)inamide-GDP
MNXM1101937	Aerobactin
MNXM1103700	Mo-molybdopterin cytosine dinucleotide
MNXM1103715	adenosylcob(III)inamide
MNXM1104368	Mo-molybdopterin
MNXM1104551	adenosylcob(III)alamin
MNXM1104756	cob(I)alamin
MNXM1104857	siroheme

After discovering instances in which RDkit was unable to calculate molecular descriptors, an effort was made to verify the accuracy of those descriptors between SMILES and InChI molecular representation formats. To explore a potential scenario in which either of the molecular representation formats led to different descriptor values or failed to be calculated, the python package DataComPy was used. Comparison of the InChI and SMILES molecular descriptors for both models was done with an absolute deviation tolerance of 0.001 and a relative tolerance of 0.0. Due to the different number of metabolites in the InChI and SMILES datasets, the comparison was limited to the metabolites which had both SMILES and InChI representations. The result was that 144 descriptors had un-equal values while 65 had equal values. A total of 19,967 values were not equal. The complete comparison analysis of the molecular descriptor calculations between InchI and SMILES in the *i*Ml1515 model is available in Supplementary data 2, Table A.1 in "iML1515_SMILES_InChI_comparison.txt". Further inspection of the comparison results revealed that descriptors related to the surface charge of a molecule, its excitation state, and Crippens atom-based descriptors MolLogP and MolMR had the largest number of differences. An excerpt of the descriptors with the most and fewest numbers of un-equal values from the *i*Ml1515 model is presented in Table 4.6 and a sampling of metabolites with their InChI and SMILES descriptor values are presented in Table 4.7.

Table 4.6: A sample of molecular descriptors with the most and fewest number of unequal values between the SMILES and InChI based calculations of metabolites in the *i*Ml1515 model. Excerpt from supplementary data 2, Table A.1 “iML1515_SMILES_InChI_comparison.txt”.

Descriptor name	Descriptor type	Num. of unequal values
MolMR	Topological	460
MolLogP	Topological	455
estate_vsa1	Topological	427
FpDensityMorgan3	Fingerprint	87
FpDensityMorgan1	Fingerprint	70
qed	Fingerprint	49
ExactMolWt	Constitutional	2
NumValenceElectrons	Constitutional	1

The comparison report revealed a large numerical difference in descriptor values for some metabolites. MNXM87122, listed in Table 4.7 and named: “three disacharide linked murein units (tetrapeptide crosslinked tetrapeptide (A2pm->D-ala), one uncrosslinked tetrapaptide) (middle of chain)" is a large molecule with a molecular weight of 2740 Da. This metabolite had the largest difference in its MolLogP value. The veracity of both the SMILES and InChI based calculations was corroborated by an external database, PubChem [152]. According to PubChem, MNXM87122 has a MolLogP value of -16.5.

Table 4.7: A sample of metabolites with their calculated descriptor values from their InChI and SMILES representations. Excerpt from "iML1515_SMILES_InChI_comparison.txt " in Supplementary data 2, Table A. The molecular descriptor and its values are written in the form of "descriptor name"(descriptor value").

MNXref	Metabolite	InChI value	SMILES value
MNXM1107906	Superoxide	bcut2d_mwlow (1.008)	bcut2d_mwlow (14.999)
MNXM87122	Large multi-unit metabolite	mollogp (-9.447)	mollogp (-30.3377)
MNXM732007	[4Fe-4S]$^{2+}$ cluster	exactmolwt (351.6)	exactmolwt (355.6)
MNXM584	Allantoate	qed (0.154931)	qed (0.327322)

4.1.3 Case study datasets

All 7 case study datasets were successfully created as described in section 3.1.5 and an overview of their metrics are listed in Table 4.8 and excerpts of the pathway case study datasets are available in Appendix B.

Table 4.8: Comparison of case study dataset metrics. The number of nodes represents the number of metabolites in that dataset. The number of edges represents the number of reaction links between metabolites in the dataset.

Case study Num.	Case study	Note	Nodes	Edges
1	*E. coli* central metabolism	Metabolic network	964	4463
2	*i*ML1515	Metabolic network	55	240
3	The TCA cycle	Metabolic pathway	11	11
4	ATP biosynthesis	Metabolic pathway	17	16
5	CTP biosynthesis	Metabolic pathway	9	8
6	GTP biosynthesis	Metabolic pathway	16	15
7	Histidine biosynthesis	Metabolic pathway	10	9

4.2 Descriptor selection

4.2.1 Molecular descriptor analysis

Spearman rank correlation coefficients were calculated for the molecular descriptors calculated in the *E. coli* central metabolism and *i*ML1515 models and their correlation matrices are presented in Figures 4.1 and 4.2 respectively.

The correlation matrix between descriptors in the *E. coli* central metabolism illustrates a strong positive correlation between features that quantify chemical structures or molecular properties that are positively linked with molecular weight and a corresponding negative correlation when that relationship is reversed. Several descriptors demonstrate low or no correlation with each other, demonstrated by PEOE_VSA and fr_benzene (benzene substructure fragments). PEOE_VSA quantifies molecular surface charges using partial charges and van der Waal surface area contributions, a physiochemical property, while fr_benzene quantifies the number of benzene substructure fragments. Substructure fragment descriptors tend to calculate zero for most biomolecules [166] as they are closely related to biological function [167].

The correlation matrix for the *i*ML1515 model shows in Figure 4.2 a larger number of less correlated descriptors when compared to the core metabolism model, shown in Figure 4.1. The majority of the uncorrelated features listed in the lower-right quadrant are substructure fragment descriptors. These types of descriptors are often used in conjunction with physiochemical descriptors to estimate the activity and potency of compounds in a biological context [167–169].

The removal of empty features revealed that the majority of uncorrelated features for both the *E. coli* central metabolism and *i*ML1515 models were empty descriptors. The removed descriptors were substructure fragment counts ($n = 85$), molecular shape descriptors, intramolecular descriptors, and chemical structure counts. The number of features available after the removal of empty descriptors was 45 and 55 for the *E. coli* central metabolism and *i*ML1515 model respectively.

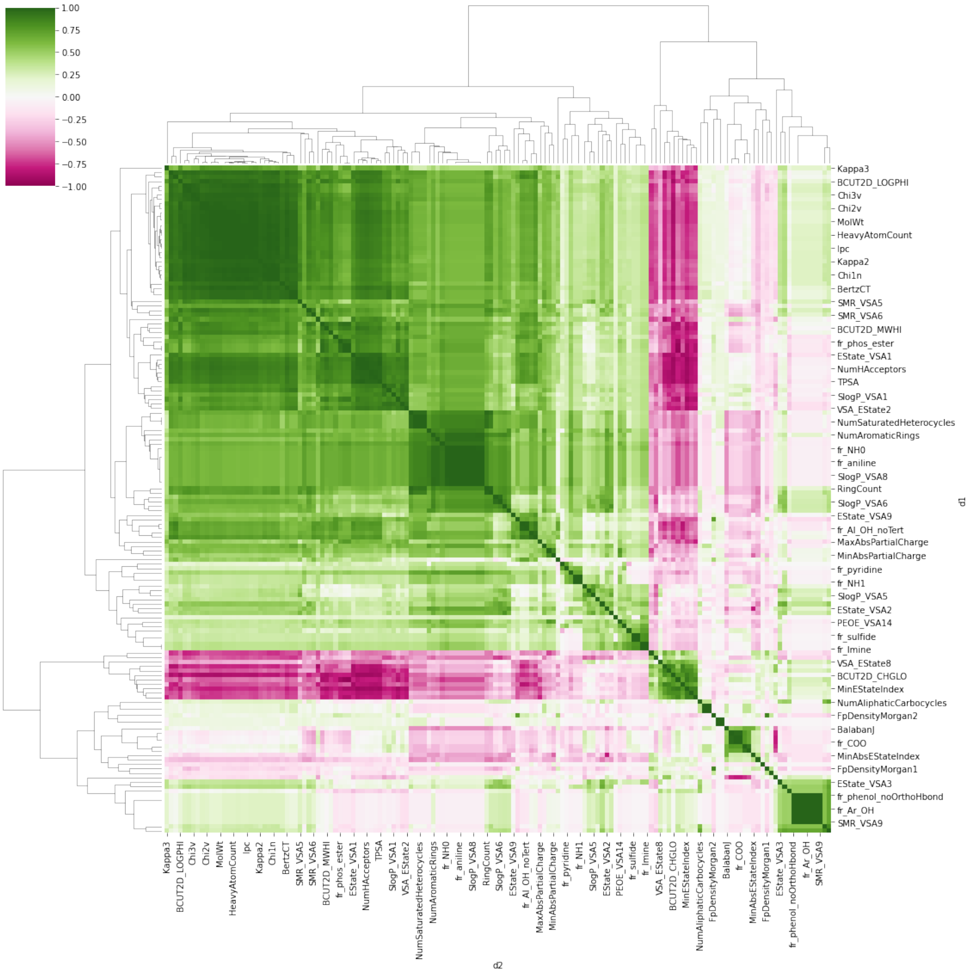


Figure 4.1: Correlation matrix of the calculated molecular descriptors for the *E. coli* central metabolism model. Green indicates a positive correlation. Purple indicates a negative correlation. White indicates low or no correlation.

A simplified overview of the number of removed descriptors and their categories for the *i*ML1515 model is illustrated in Figure 4.9.

The remaining descriptors in the pruned set of molecular descriptors that exhibited orthogonality were molecular fingerprints such as FpDensityMorgan1, FpDensityMorgan2 in the central metabolism model, and MolLogP in the *i*ML1515 model. It is possible to explain FpDensityMorgan3's higher correlation as compared with FpDensityMorgan1 and FpDensityMorgan2 by examining how morgan fingerprints are calculated. Morgan fingerprints are calculated by assigning a numeric identifier to an initial atom and grouping nearby atoms within a certain diameter into fragments [170]. Morgan fingerprints differ based on the size of the diameter indicated by the number at the end of the name. FpDensityMorgan1, for

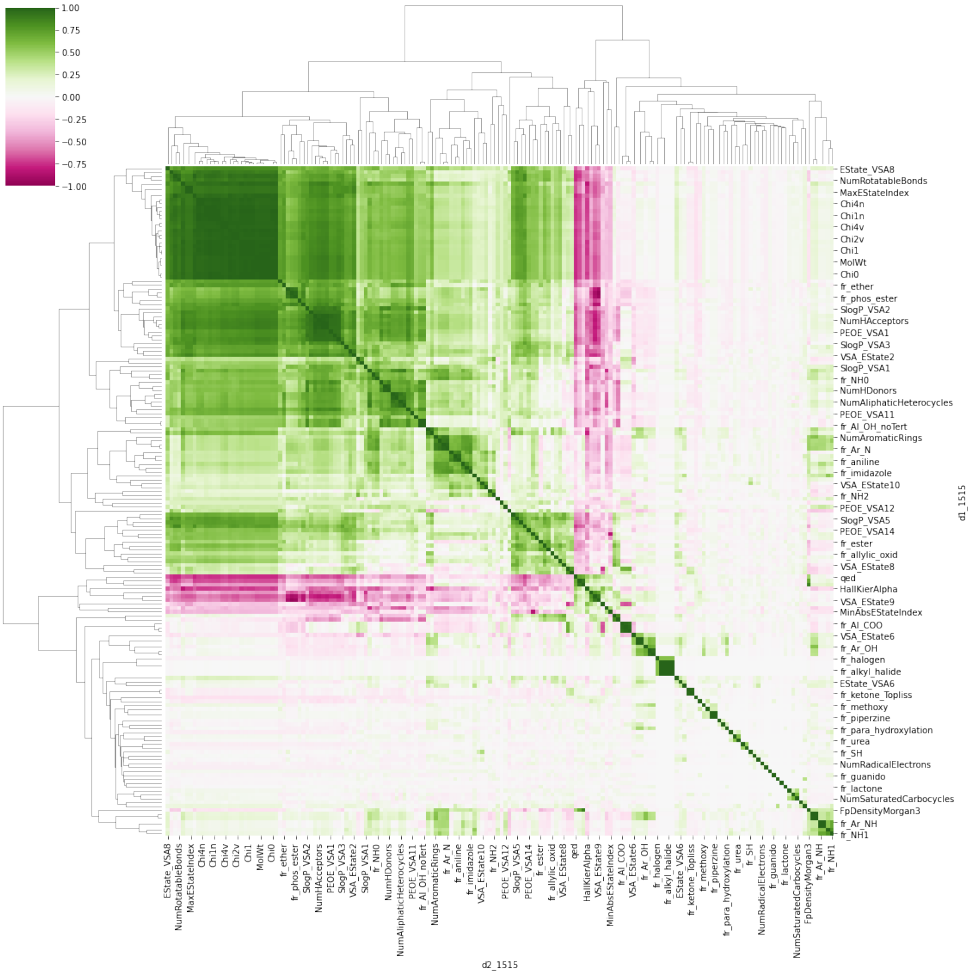


Figure 4.2: Correlation matrix of the calculated molecular descriptors for the *E. coli i*ML1515 model. Green indicates a positive correlation. Purple indicates a negative correlation. White indicates low or no correlation.

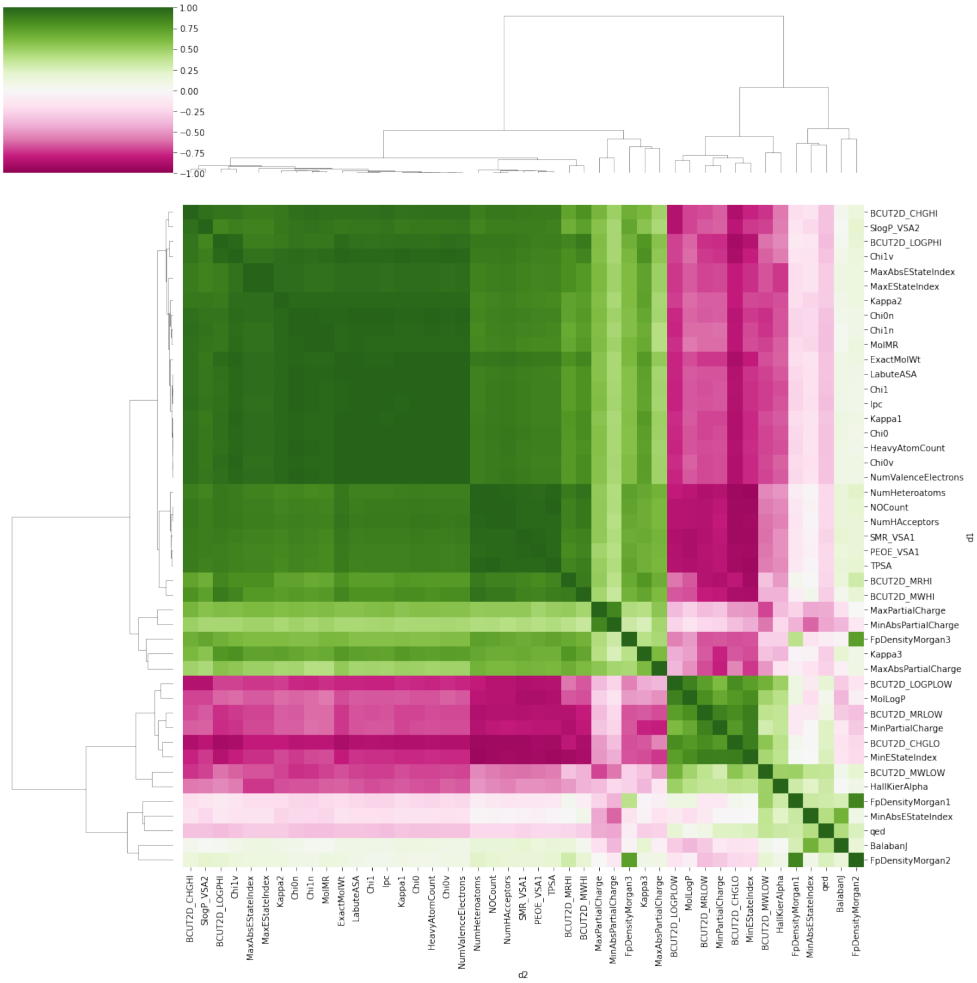


Figure 4.3: A correlation matrix of the calculated molecular descriptors for the *E. coli* core metabolism model after removal of empty descriptors. Green indicates a positive correlation. Purple indicates a negative correlation. White indicates low or no correlation.

example, has a diameter of one. An increasing negative correlation is observed in the central metabolism model as diameter increases. FpDensityMorgan3 incorporates more structural elements and chemical bonds in its fragments, resulting in greater correlation than fingerprints with a smaller diameter.

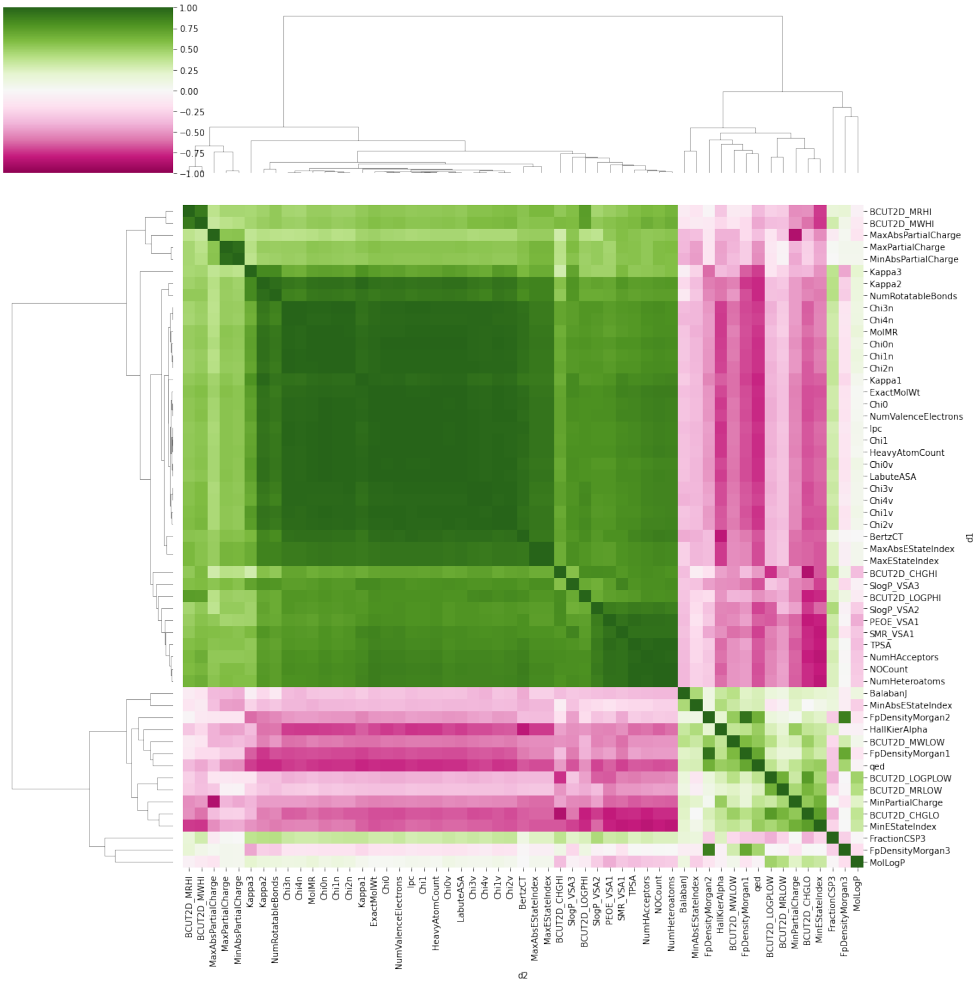


Figure 4.4: Correlation matrix of the calculated molecular descriptors for the *E. coli i*ML1515 GEM after removal of empty descriptors.

This is reversed in the *i*Ml1515 model where FpDensityMorgan3 is the least correlated molecular fingerprint, shown in Figure 4.4. This could be attributed to the diversity of molecular structures of metabolites in the *i*ML1515 model. MolLogP also differs in its correlation between models. In the central metabolism model, MolLogP is negatively correlated with the Chi and Kappa family of topographical descriptors while in the *i*ML1515 model MolLogP is less correlated with these topographical descriptors. This lack of uniformity in correlations between models could be attributed that the central metabolism model having significantly fewer metabolites, which are more chemically similar to each other when compared to the chemical diversity amongst the metabolites in the *i*Ml1515 model.

In data analysis, the dimensionality of an object is the number of variables that are used to describe each it [109]. Based on significant correlations between many of the molecular descriptors, principal component analysis (PCA) was used to reduce the dimensionality of the pruned molecular descriptor set for both models. Therefore, most of the variation in data can be explained by fewer principal components. The first principal component (PC1) is comprised of the descriptors that capture the most variance in the data while the second principal component (PC2) includes those that were not included in PC1. As principal components are orthogonal to each other, their loadings infer possible descriptor combinations that are orthogonal to each other [109].

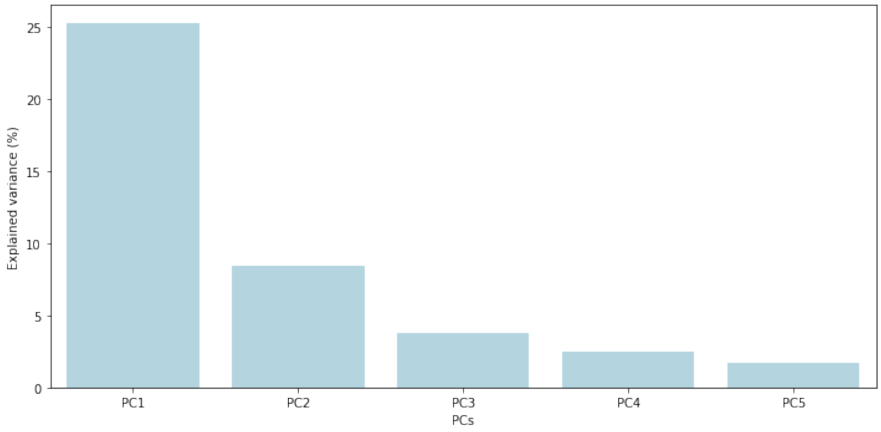


Figure 4.5: Scree plot of the explained variance of 5 calculated principal components for the *E. coli* central metabolism GEM after removal of empty descriptors.

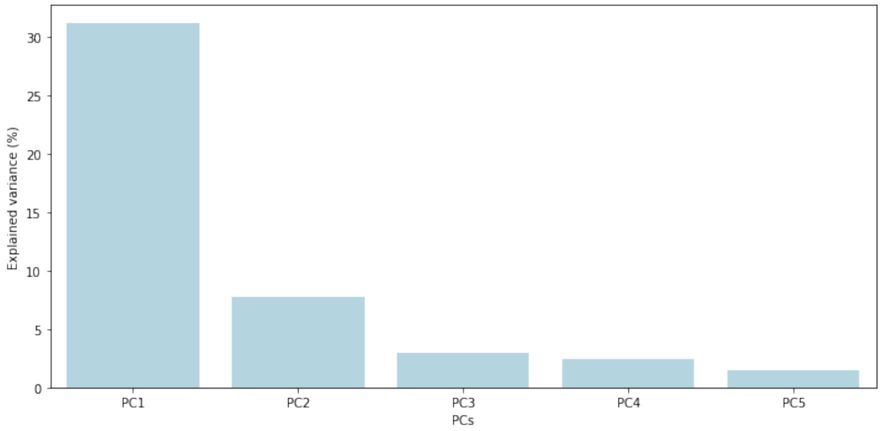


Figure 4.6: Scree plot of the explained variance of 5 calculated principal components for the *E. coli* *i*ML1515 GEM after removal of empty descriptors.

As can be seen from the scree plots in Figures 4.5 and 4.6, the cumulative explained variance (CEV) of PC1 and PC2 represents 34.7% in the central metabolism model and 39.2% in the *i*ML1515 model. To determine the amount of variance explained by removed molecular descriptors, PCA was performed on the un-pruned molecular descriptor datasets for both models. Based on the CEVs for PC1 and PC2, listed in Table 4.9, it is apparent that empty descriptors contributed to explaining variance in biochemical space for both models.

Table 4.9: Cumulative explained variance for PC1 and PC2 for molecular descriptor datasets before and after removal of empty descriptors

Model	**CEV before removal**	**CEV after removal**
*i*Ml1515	74%	39.2%
E. coli central metabolism	96%	34.7%

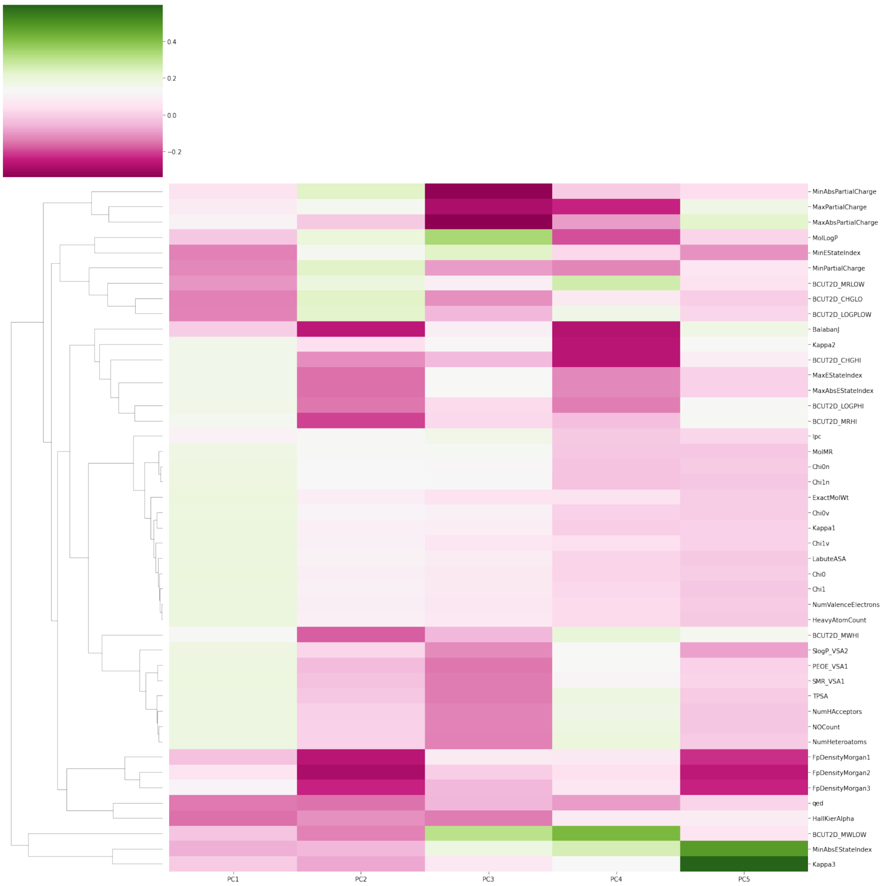


Figure 4.7: Correlation matrix plot of the for 5 calculated principal components loadings for the ***E. coli*** core metabolism GEM after removal of empty descriptors.

We see from Figures 4.7 and 4.8 that the loadings of PC1 for both models are comprised of topographical descriptors such as Chi and kappa descriptors, descriptors associated with molecular mass (ExactMolWt, HeavyAtomCount, NumValenceElectrons), the number of hydrogen acceptors and heteroatoms (NumHAcceptors & NumHeteroatoms). The PC2 loadings differ between models but are primarily comprised of descriptors that quantify intramolecular force descriptors (BCUT) and molecular partial charges (Min/MaxAbsPartialCharge, Min/MaxPartialCharge).

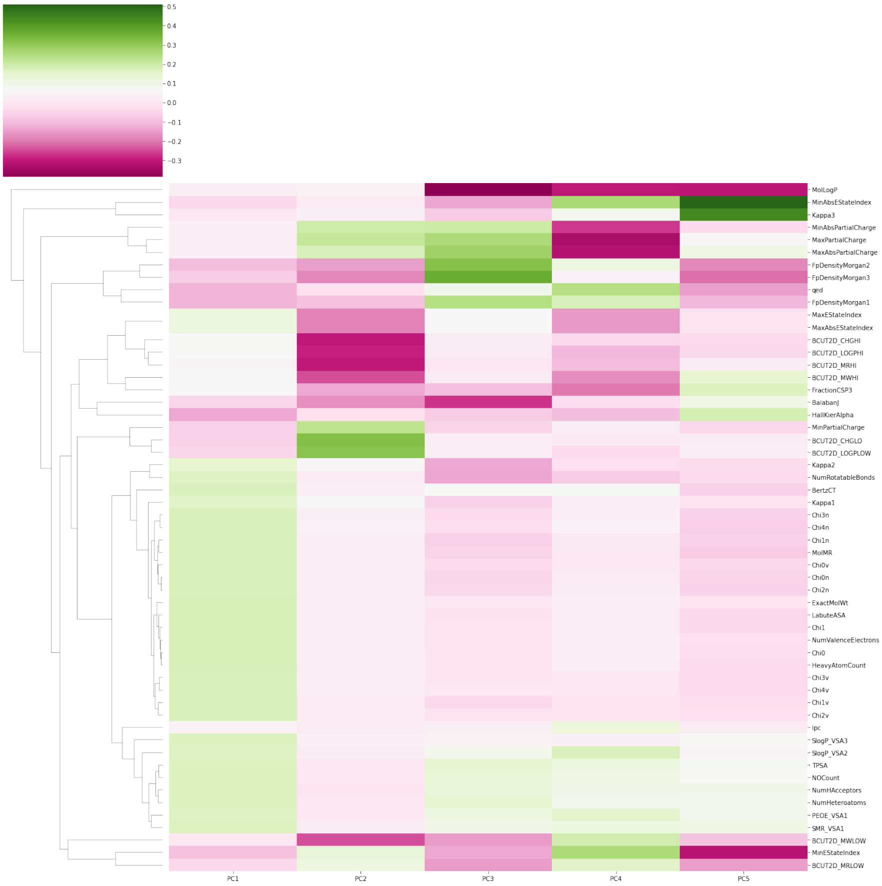


Figure 4.8: Principal component loadings of the principals calculated from the pruned set of molecular descriptors from the *E. coli i*Ml1515 model.

The varying loadings of the components can be attributed to the difference in the number of metabolites between the models, as suggested previously. The metabolites (n = 55) in the core model exhibit more similarity in molecular structure and properties when compared to the diversity of metabolites (n = 964) in the significantly larger *i*Ml1515 model.

4.2.2 Selection of biologically relevant descriptors

In this project, the process for selecting descriptors differs from QSAR norms as described in section 3.2. For the layouts to be able to convey biological information, features reflecting molecular properties with known biological functions were prioritized. This led to the selection of 9 molecular descriptors from the pruned set of descriptors; molecular charge, MolMR, MolLogP, ExactMolWt, NumHDonors, NumHAcceptors, NumHeteroatoms, NumValenceElectrons, ExactMolWt and quantitative estimate of drug-likeness (QED). The descriptors are listed in Table 4.10 and a graphical overview of the descriptor selection process is illustrated in Figure 4.9. The rationale applied for the selection process is listed below:

- Whenever two or more descriptors contained overlapping information, the one with the most intuitive and clear interpretation was selected; for example, between ExactMolWt and HeavyAtomCount, ExactMolWt was selected.
- Descriptors that reflect physiochemical properties of molecules that are more easily understood by a non-expert, for example; molecular weight, number of hydrogen donors & acceptors, the number of non-carbon atoms, number of valence electrons, molecular charge, MolLogP, and MolMR were selected
- It was decided not to include descriptors that reflect intramolecular forces, topographical properties (Chi, Kappa, and HallKierAlpha), partial molecular charges, and molecular fingerprints because they are less intuitive and not easily understood.
- The descriptor QED was selected due to its ability to capture how simple physiochemical properties affect molecular behavior and function. While it is typically used to estimate a drug's oral bioavailability, it is based on eight descriptors (see section 2.4.1) which have all been shown to influence the ability of metabolites to associate with proteins that catalyze biochemical reactions [131].

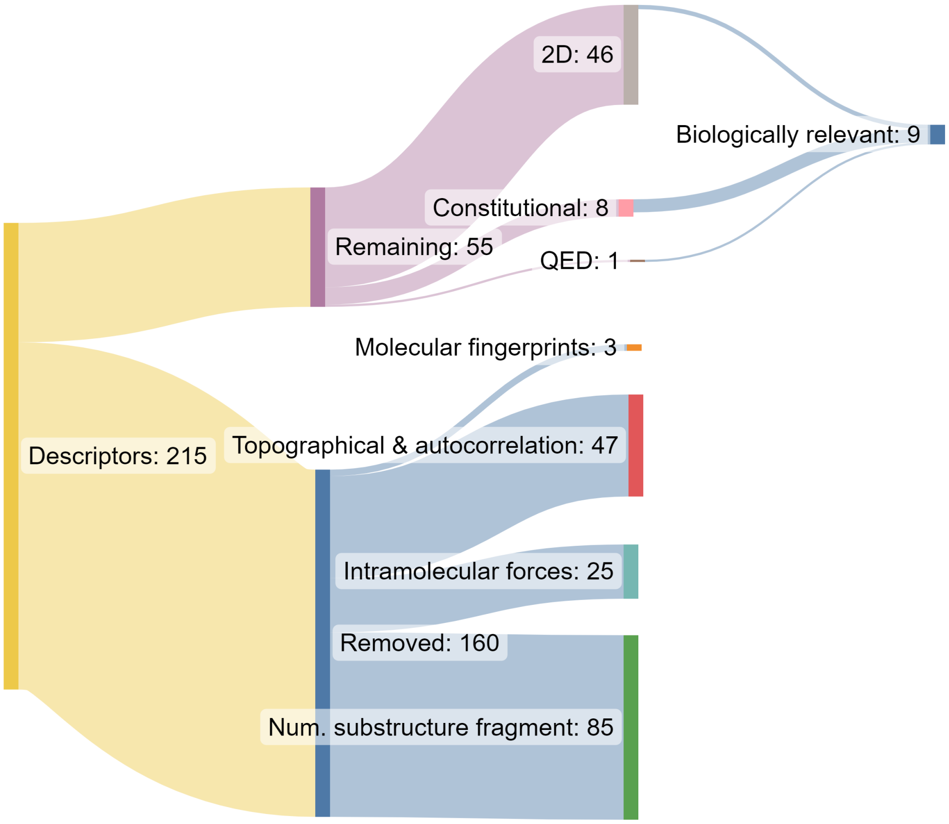

2D: 46
Biologically relevant: 9
Constitutional: 8
Remaining: 55
QED: 1
Molecular fingerprints: 3
Descriptors: 215
Topographical & autocorrelation: 47
Intramolecular forces: 25
Removed: 160
Num. substructure fragment: 85

Figure 4.9: Sankey diagram showing a simplified overview of the categories of molecular descriptors that were removed and selected. Adopted from the *i*ML1515 model.

Table 4.10: Final selection of molecular descriptors: Selected descriptors comprise six constitutional descriptors, two 2D descriptors, and one drug-likeness descriptor.

Descriptor	Descriptor category	Definition
charge	Constitutional	Molecular charge
ExactMolWt	Constitutional	Exact molecular weight
NumValenceElectrons	Constitutional	Num. valence electrons
NumHDonors	Constitutional	Num. hydrogen atoms
NumHAcceptors	Constitutional	Num. hydrogen acceptors
NumHeteroatoms	Constitutional	Num. non-carbon atoms
MolLogP	2D (Crippen atom-based)	Octanol-water partition coefficient
MolMR	2D (Crippen atom-based)	Molecular refraction
QED	Drug-likeness descriptor	Quantification of drug-likeness

4.2.3 Ranking a subset of selected descriptors

Since some of the selected features had overlapping values for many metabolites, ranked versions of three features were created. Upon reviewing descriptor correlations in the pruned descriptor set, the most likely set of descriptors that would yield interesting results when ranked would be the following: MolLogP, molecular weight, and molecular charge. These ranked features, listed in Table 4.11, augmented the selected molecular descriptors listed in Table 4.10.

Table 4.11: Ranking of molecular descriptor values. These descriptors augment the selected molecular descriptors listed in Table 4.10

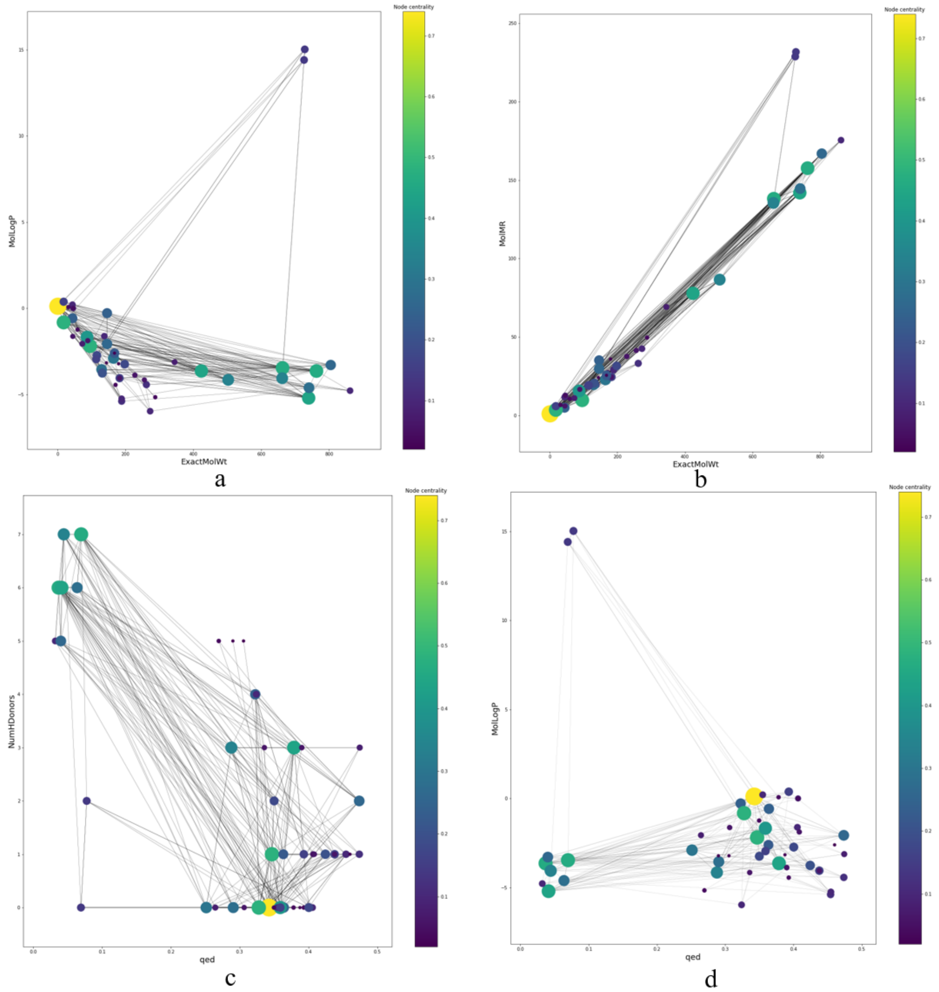
Descriptor	Descriptor category	Description
mass_rank	Ranked descriptor value	Ranking of ExactMolWt values
MolLogP_rank	Ranked descriptor value	Ranking of MolLogP values
charge_rank	Ranked descriptor value	Ranking of molecular charge

4.3 Layout visualization

Layouts of the 7 case studies listed in Table 4.8 are visualized with pairs of selected and ranked descriptors, listed in Tables 4.10 and 4.11 respectively. To facilitate the evaluation of each pathway and their layouts in a global context, the intermediate metabolites in the respective case study dataset are highlighted and labeled in two *i*ML1515 network visualizations where relevant. A summary of the layout visualization results is provided at the end of this section.

4.3.1 Metabolic networks

Case study 1: *Escherichia coli* central metabolism



a b c d

Figure 4.10: Plots of the *Escherichia coli* core metabolism network graphs using: (a): Exact molecular weight and MolLogP, (b): exact molecular weight and MolMR, (c): quantitative estimate of drug-likeness and number of hydrogen donors and (d): quantitative estimate of drug-likeness and MolLogP. Node hue is set by its centrality measure. Node size is set by its connectivity degree.

In Figure 4.10 we see that nodes are very clustered and the number of edge-crossings makes identifying a network motif or module that could represent a metabolic pathway difficult. Two nodes are readily identifiable, ubiquinone-8 (Coenzyme Q8) and ubiquinol-8 (reduced Coenzyme Q8), which because of their large and similar descriptor values, are positioned at a greater distance away from the majority of metabolites.

From Figures 4.10b and 4.11c, we see that the steric bulk (indicated by MolMR

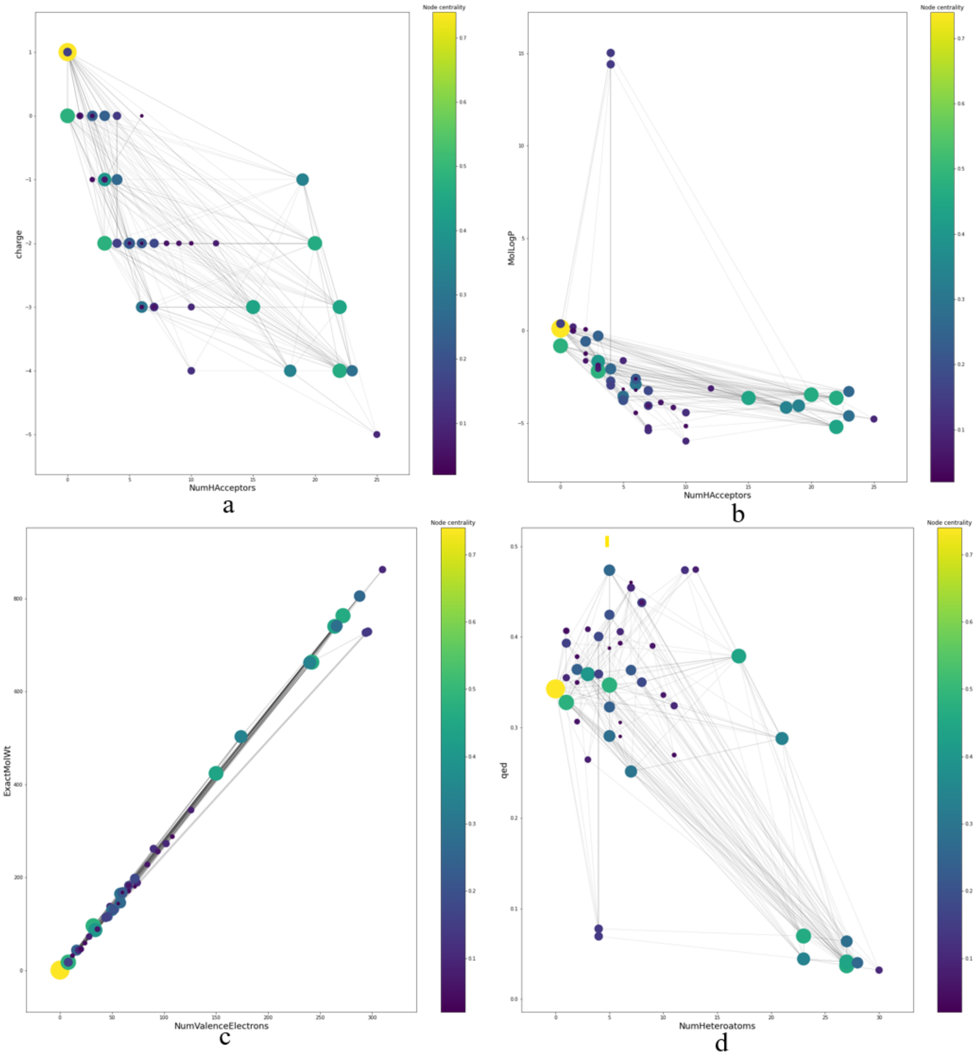
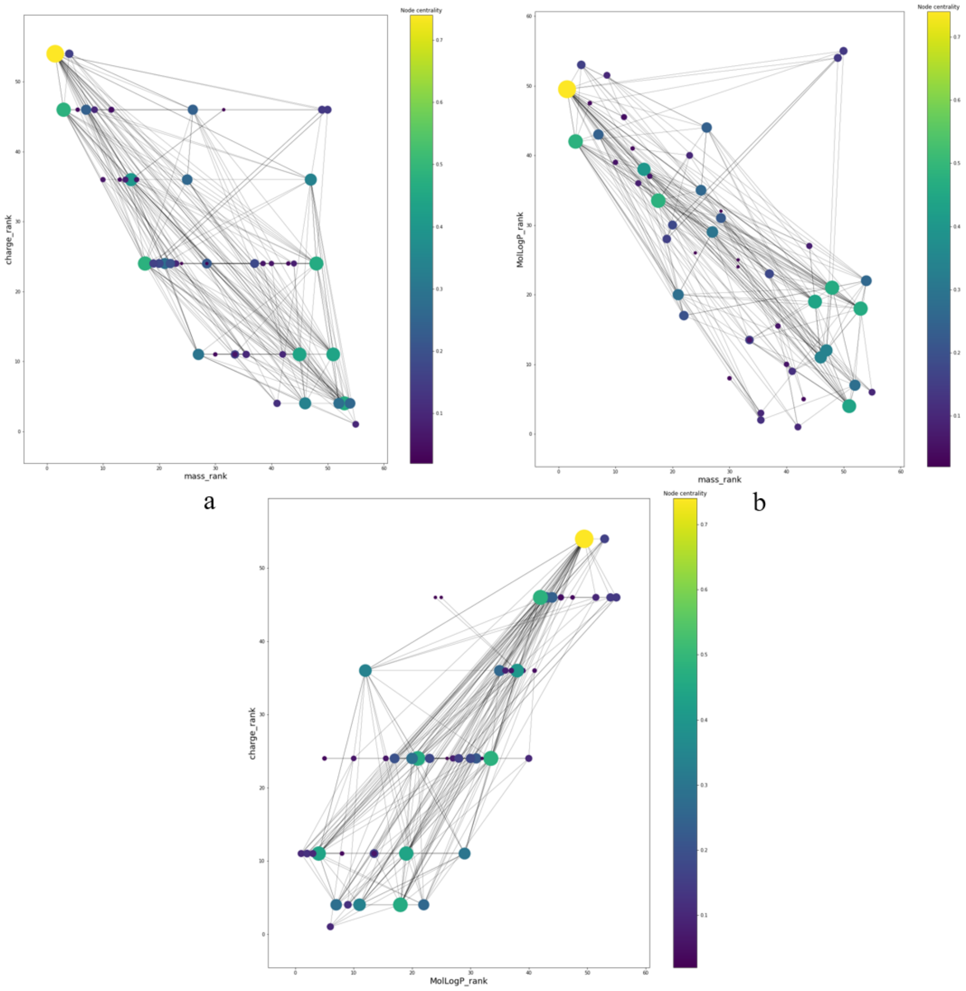


Figure 4.11: Graphs of the *Escherichia coli* central metabolism network using: (a): NumHAcceptors and molecular charge, (b): NumHAcceptors and MolLogP, (c): NumValenceElectrons and ExactMolWt (d): NumValenceElectrons and quantitative estimate of drug-likeness. Node hue is set by its centrality measure. Node size is set by its connectivity degree.

values) and the number of valence electrons increase with molecular weight. In Figures 4.11 and 4.12 we observe similar patterns as in Figure 4.10, except for Figure 4.11d where a slight improvement in the layout is observed. In Figure 4.11d, NumHeteroatoms and QED have positioned central nodes in such a way that the relationship between central metabolites (hydrogen, water, inorganic phosphate, and NADH) is easier to understand as they are not being obfuscated by edge crossings. Figure 4.12b shows that ranked features separate nodes better than their non-ranked equivalent, as shown in Figure 4.10a. As shown in Figure 4.10d, the majority of metabolites in the core metabolism model display chemical similarity, quantified by the QED descriptor and the spatial proximity of the metabolites. Furthermore, Figure 4.10a,d illustrates that the nodes with the highest degree of connectivity exhibit similar polarity, as indicated by their MolLogP values.



a

b

c

Figure 4.12: Metabolic network visualization of the Escherichia coli core metabolism model using (a): ranked molecular mass and ranked molecular charge, (b): ranked ExactMolWt and ranked MolLogP, and (c): ranked MolLogP and ranked molecular charge. Node hue is set by its centrality measure. Node size is set by its connectivity degree.

Case study 2: *i*ML1515

As shown in Figures 4.13, 4.14 and 4.15, the larger number of metabolites and reactions in the *i*ML1515 model results in visualizations that are very hard to read. Due to the chemical similarity between metabolites, the node clustering and edge crossings that result from their spatial arrangement make identifying a pathway or network motif difficult.

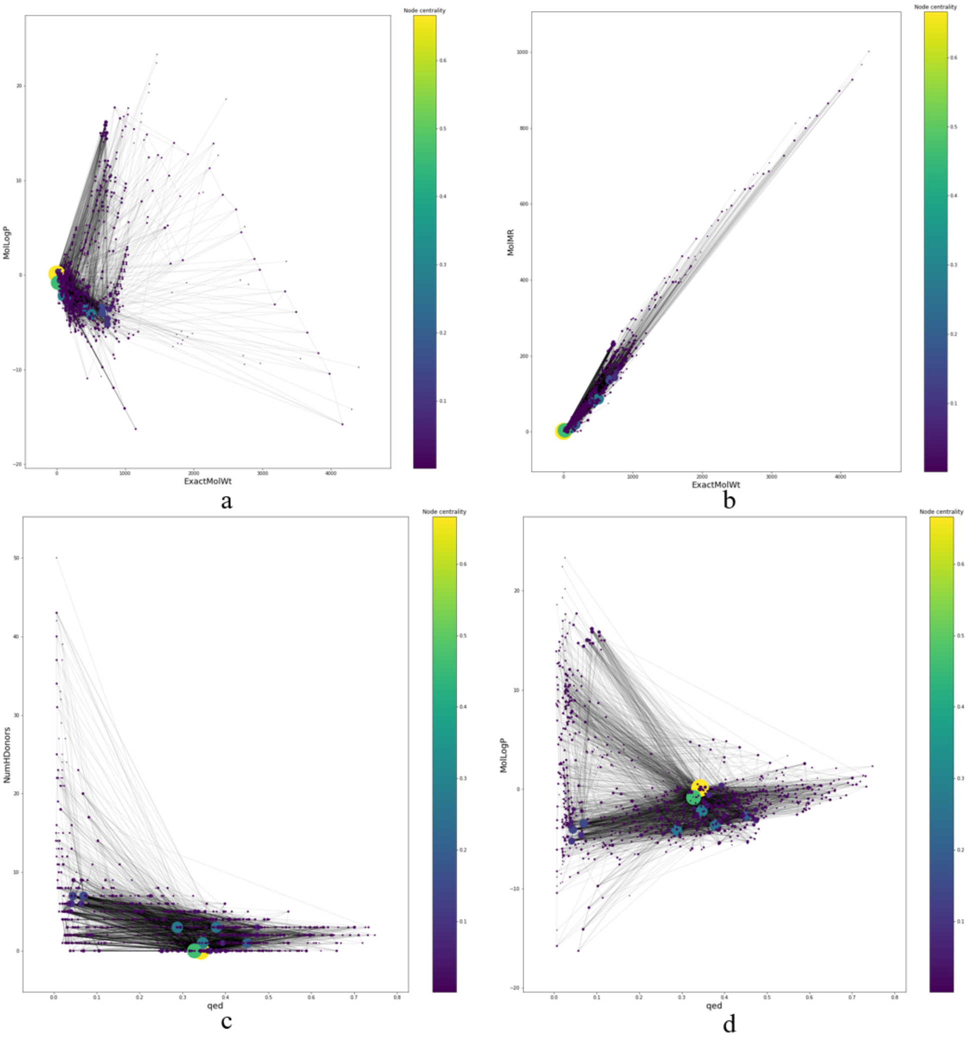


Figure 4.13: Plots of the ***E. coli i*ML1515** network using: (a): Exact molecular weight and MolLogP, (b): exact molecular weight and MolMR, (c): quantitative estimate of drug-likeness and number of hydrogen donors and (d): quantitative estimate of drug-likeness and MolLogP. Node hue is set by its centrality measure. Node size is set by its connectivity degree.

Fig 4.13d illustrates that QED and MolLogP position the majority of the highly connected metabolites (hydrogen, water, inorganic phosphate, ATP, ADP, and NADH) close to one another, but to a greater extent than shown in Figure 4.10d.

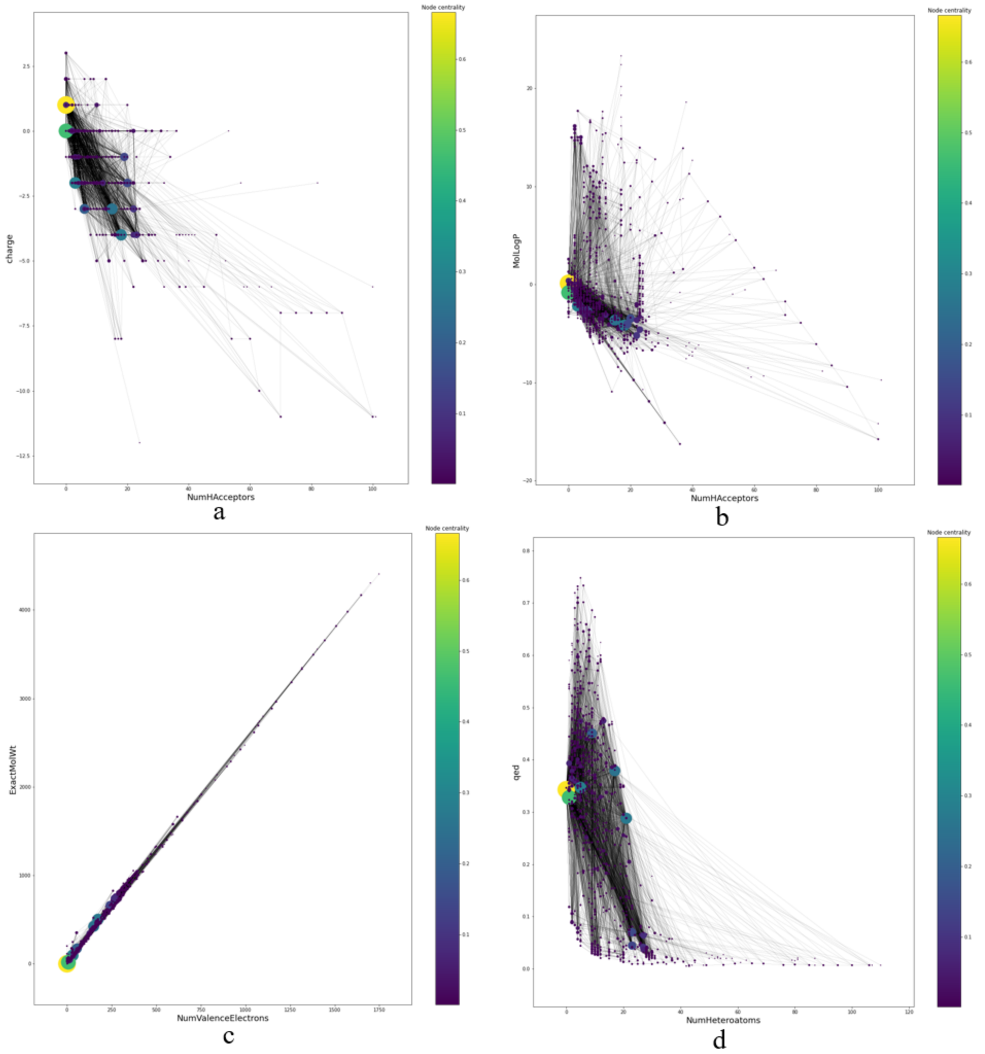


Figure 4.14: Graphs of the Escherichia coli iML1515 network using: (a): NumHAcceptors and molecular charge, (b): NumHAcceptors and MolLogP, (c): NumValenceElectrons and ExactMolWt (d): NumValenceElectrons and quantitative estimate of drug-likeness. Node hue is set by its centrality measure. Node size is set by its connectivity degree.

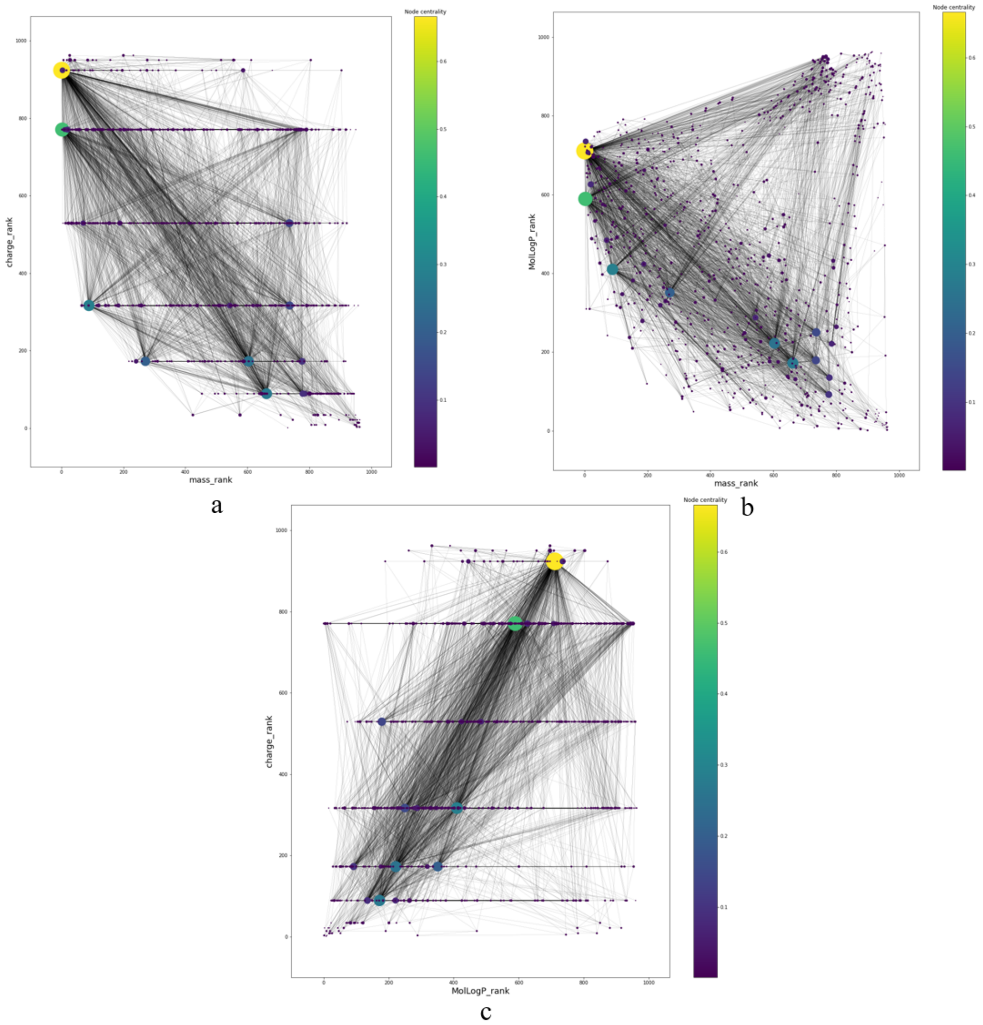


Figure 4.15: Plots of the ***E. coli i*ML1515** network using: (a): Exact molecular weight and MolLogP, (b): exact molecular weight and MolMR, (c): quantitative estimate of drug-likeness and number of hydrogen donors and (d): quantitative estimate of drug-likeness and MolLogP. Node hue is set by its centrality measure. Node size is set by its connectivity degree.

Comparing the layout visualizations in Figures 4.13 and 4.14 with Figure 4.15b, demonstrates that the ranked descriptor variants of ExactMolWt and MolLogP improve the overall clarity of the visualization, but the number of edges and resultant edge crossings makes visual identification of network motifs and modules difficult. According to Figures 4.15a and c, the charge_rank descriptor allows vertical stratification of metabolites with different molecular charges and those that have a high degree of connectivity, but it is very difficult to identify network motifs and pathways visually.

4.3.2 Metabolic pathway case studies

Case study 3: The citric acid cycle

We see in Figures 4.16, 4.17 and 4.18, a recurring network motif that is comprised of a group of nodes representing chemically similar intermediates; 2-oxoglutatare (akg), oxaloacetate (oaa), succinate (succ), fumarate (fum), and D-malate (mal__D), are consistently positioned in proximity to each other and makes it difficult to recognize the direction of reactions between these. When corroborating these results with their molecular structures shown in Figure C.1, we see that the structural similarities in these intermediaries are the likely cause.

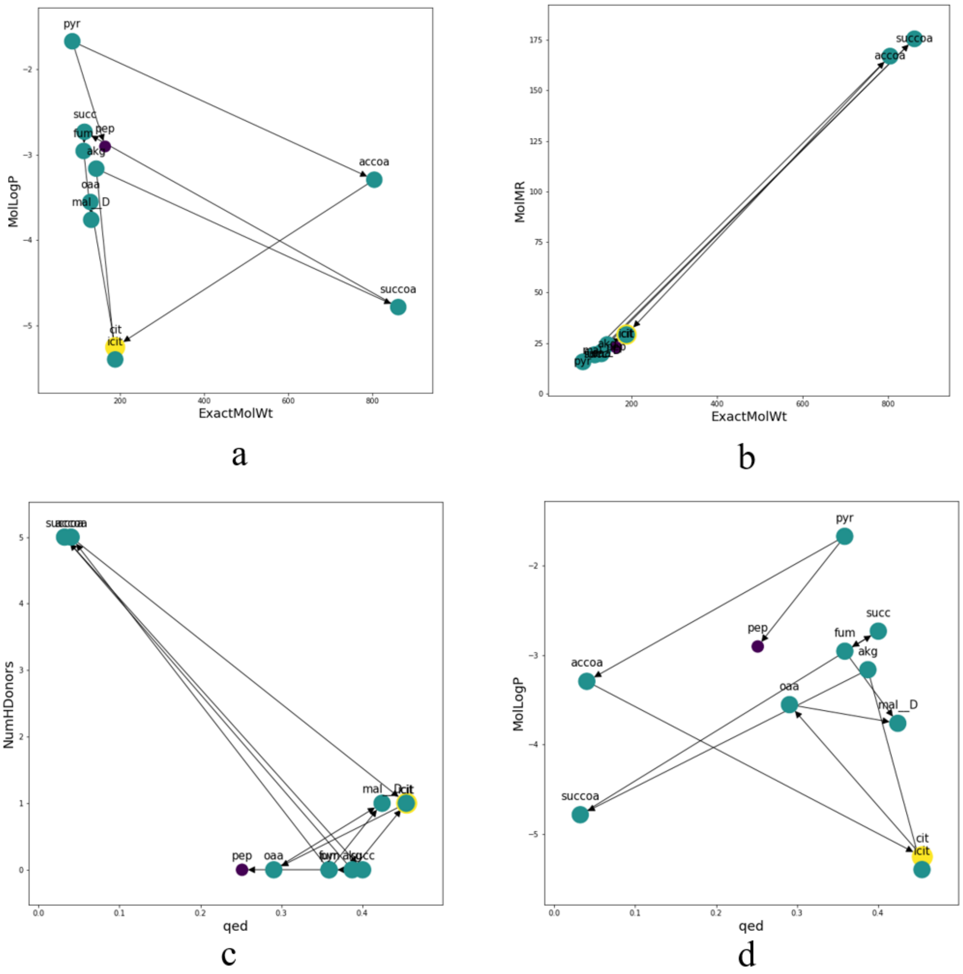


Figure 4.16: Plots of the citric acid cycle using: (a): Exact molecular weight and MolLogP, (b): exact molecular weight and MolMR, (c): quantitative estimate of drug-likeness and number of hydrogen donors and (d): quantitative estimate of drug-likeness and MolLogP. Unabbreviated metabolite names are available in Table B.1

With the layout generated with QED and MolLogP in Figure 4.16d, it is possible to visualize chemical similarities in the motif while remaining somewhat more legible than most layouts for this case study. One possible exception is 4.18b, which shows a similar spatial separation amongst the nodes. The improvement in legibility for layout in Figure 4.16d is likely caused by the way in which the descriptor QED is able to capture multiple biochemical properties of the metabolites. A sharp decrease in chemical similarity is observed with metabolites that are bound with coenzyme-A; succinyl-CoA (succoa) and acetyl-CoA (accoa), which results in layouts that are hard to read.

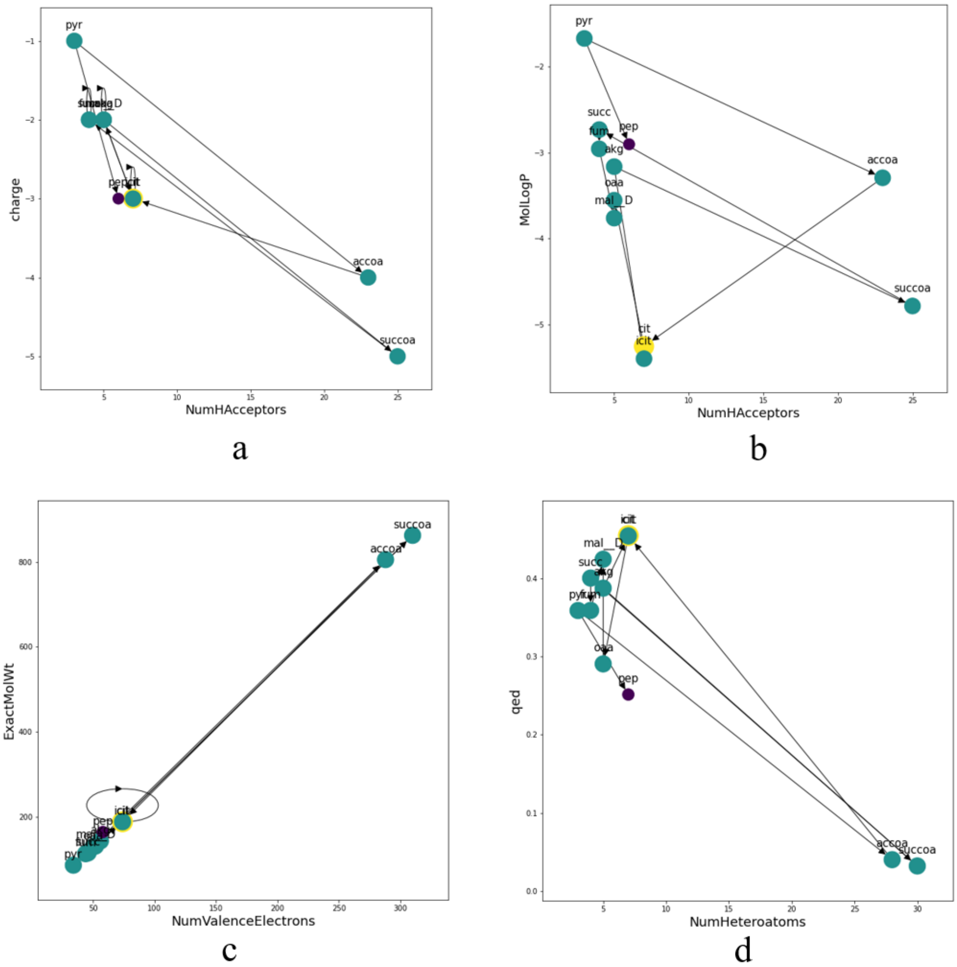


Figure 4.17: Graphs of the TCA cycle using: (a): NumHAcceptors and molecular charge, (b): NumHAcceptors and MolLogP, (c): NumValenceElectrons and ExactMolWt (d): NumValenceElectrons and quantitative estimate of drug-likeness. Unabbreviated metabolite names are available in Table B.1. The circular arrow in panel (c) is caused by the nodes representing citrate and iso-citrate stacking on top of each other due to similar descriptor values.

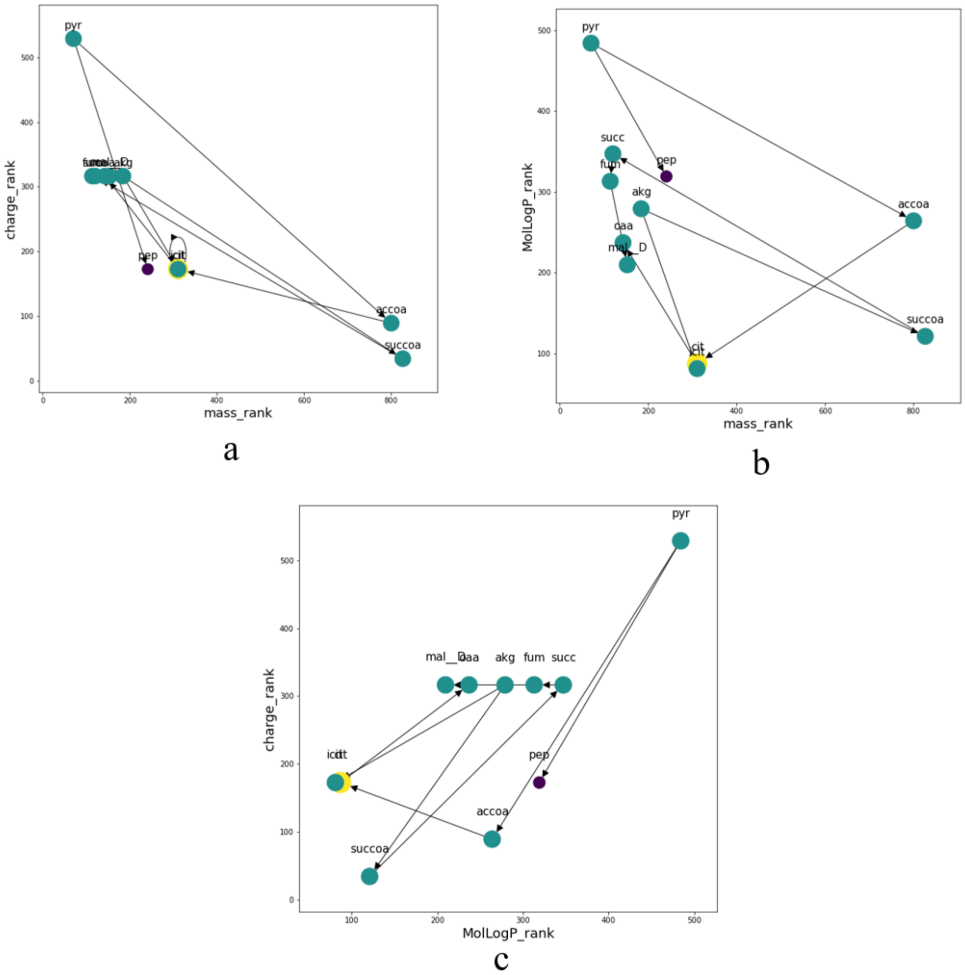


Figure 4.18: Plots of the TCA cycle using (a): ranked molecular mass, and ranked molecular charge, (b): ranked molecular mass, and ranked MolLogP, and (c): ranked MolLogP and ranked molecular charge. Unabbreviated metabolite names are available in Table B.1

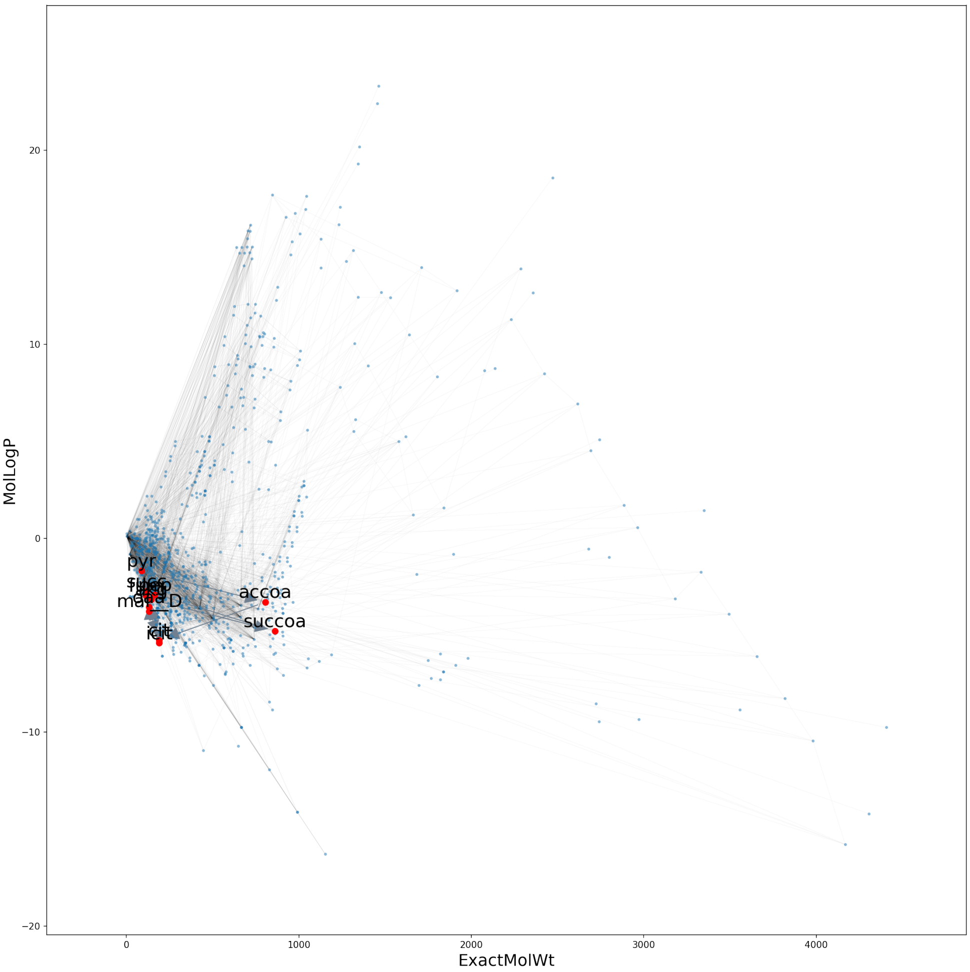


Figure 4.19: The TCA cycle is highlighted in red in the *i*ML1515 metabolic network visualized with ExactMolWt and MolLogP. Unabbreviated metabolite names are available in Table B.1

A comparison between Figure 4.13a and Figure 4.19 illustrates that it is very difficult to identify the TCA cycle pattern observed in Figure 4.16a, even when the intermediaries are highlighted. It is difficult to discern the pattern due to the number of nodes and edge crossings.

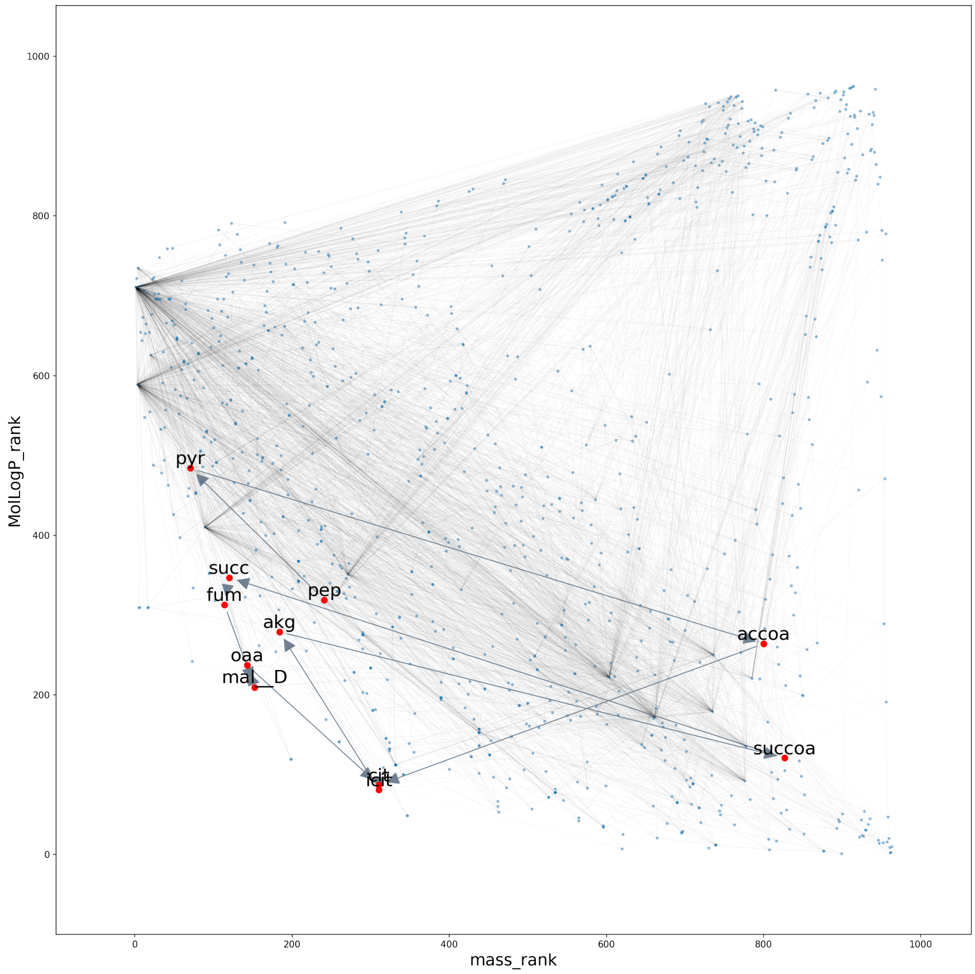


Figure 4.20: The TCA cycle is highlighted in red in the *i*ML1515 metabolic network and visualized with mass_rank and MolLogP_rank. Unabbreviated metabolite names are available in Table B.1

An improvement in the legibility of the pathway, when highlighted, is observed in Figure 4.20 but identification of the pattern found in Figure 4.18b in Figure 4.15b is very difficult. As shown in Figure 4.20, intermediates in the TCA cycle are more legible than in Figure 4.19 due to the use of ranked descriptor variants rather than actual molecular descriptor values. As a result of the positioning of the nodes, it is possible to identify how citrate synthase binds acetyl from acetyl-CoA (accoa) to oxaloacetate (oaa) to form citrate (cit). As shown in Figure 4.15b, this cannot be achieved when the intermediaries are not highlighted as there is too much information visualized.

Case study 4: ATP biosynthesis

In Figures 4.21, 4.22 and 4.23 we see a recurring network motif, a group of nodes representing chemically similar metabolites that cluster together, making identification of the direction of reactions between them hard. With the exception of (2S)-2-[5-amino-1-(5-phospho-beta-D ribosyl)imidazole-4-carboxamido]succinate (25aics) and N^6-(1,2-Dicarboxyethyl)-AMP (dcamp), most of the intermediate metabolites in the ATP biosynthesis pathway dataset has overlapping molecular weights and 7 metabolites (see Table B.2) have molecular weights between 300-360 Da which comprises this loosely clustered group of nodes.

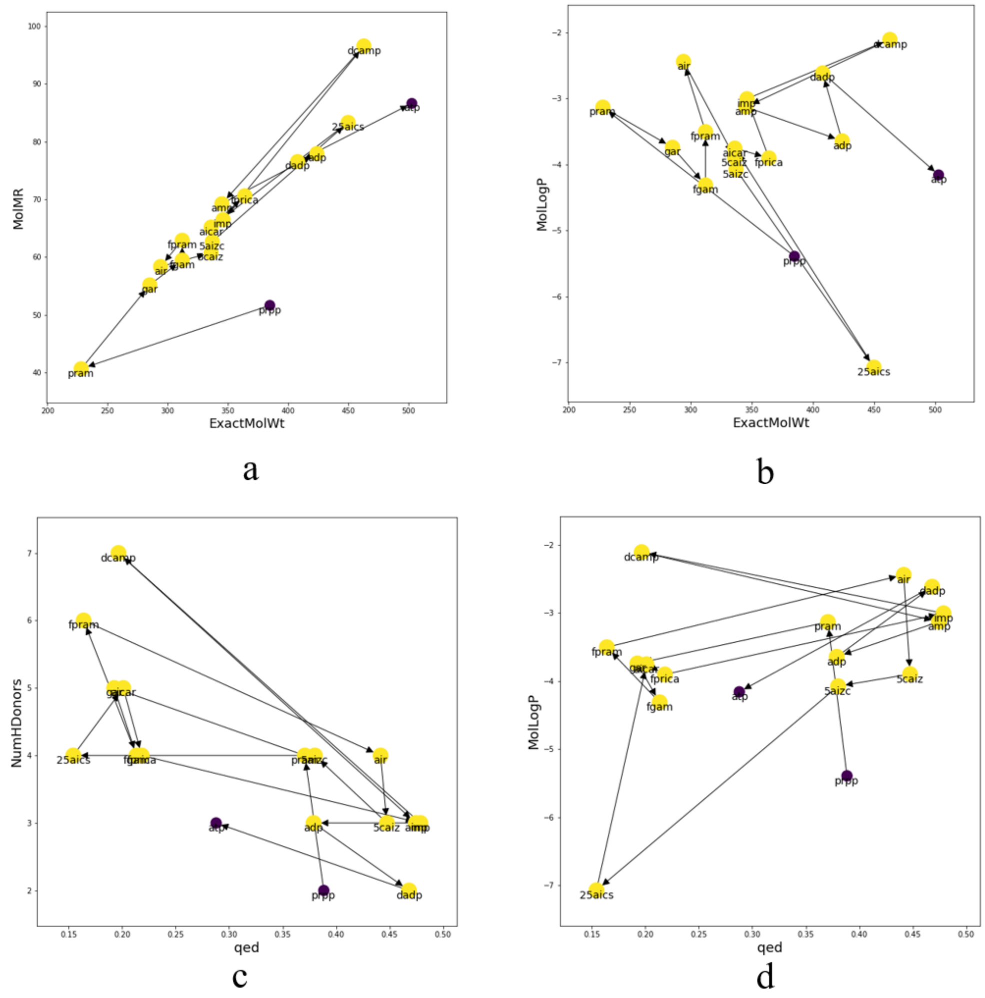
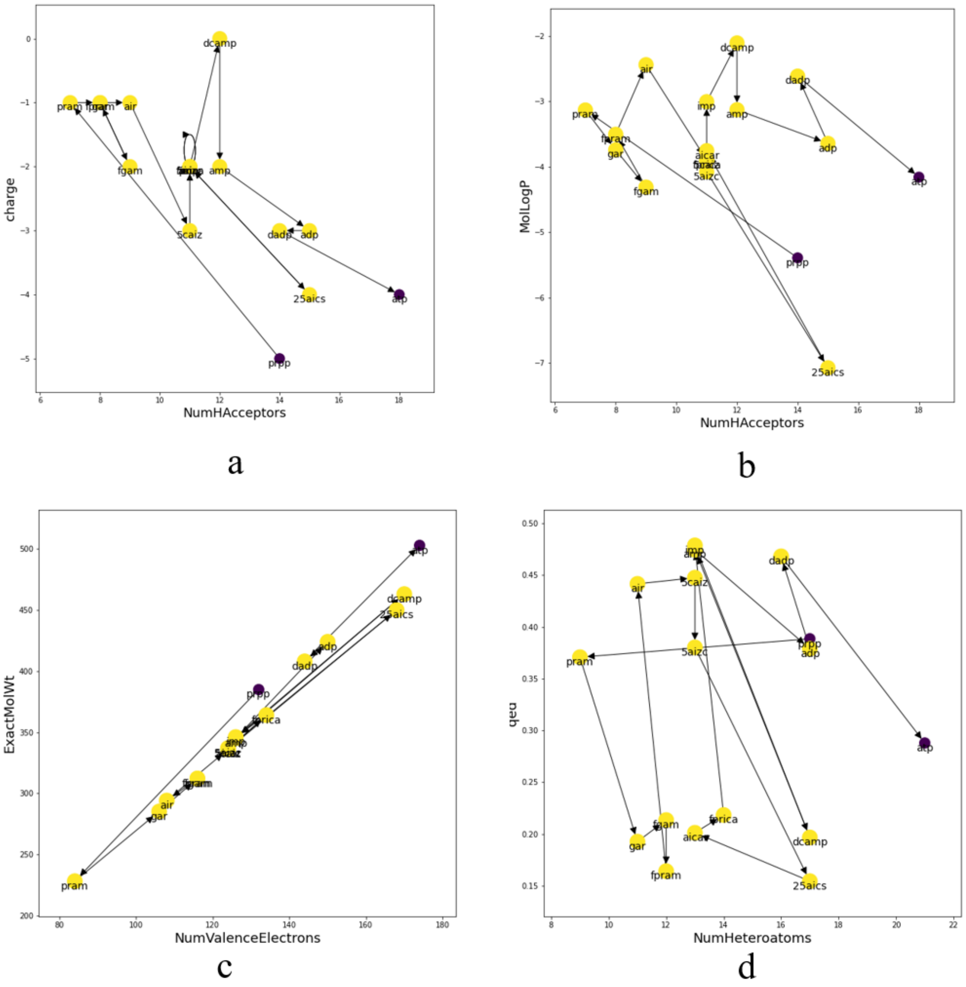


Figure 4.21: Plots of adenosine triphosphate biosynthesis using: (a): Exact molecular weight and MolMR, (b): exact molecular weight and MolLogP, (c): quantitative estimate of drug-likeness and number of hydrogen donors and (d): quantitative estimate of drug-likeness and MolLogP. Unabbreviated metabolite names are available in Table B.2

The nodes that comprise the recurring motif represent the intermediaries of the IMP biosynthesis pathway (see Figure 2.3). The short distance between these nodes is reflected in the structural similarity in the intermediaries (see Figure C.2). Some intermediaries in the *de novo* adenosine nucleotide biosynthesis pathway are positioned in proximity to this motif and the associated edge crossings are visually distracting and detract from its legibility. This pattern is best represented in Figures 4.21a where two positively correlated descriptors; ExactMolWt and MolMR, have positioned nodes from the IMP biosynthesis pathway so close to each other that following the pathway is very difficult.



a

b

c

d

Figure 4.22: Graphs of adenosine triphosphate biosynthesis using: (a): NumHAcceptors and molecular charge, (b): NumHAcceptors and MolLogP, (c): NumValenceElectrons and ExactMolWt (d): NumValenceElectrons and quantitative estimate of drug-likeness. Unabbreviated metabolite names are available in Table B.2.

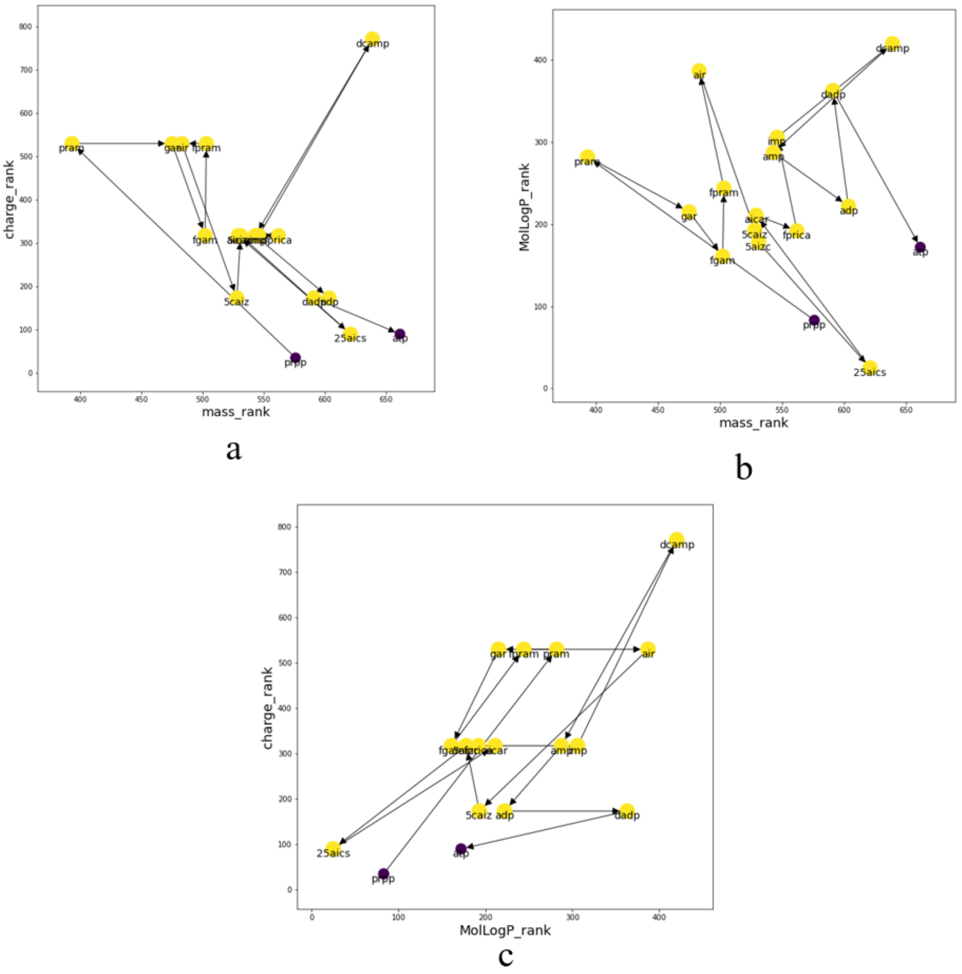


Figure 4.23: Plots of adenosine triphosphate biosynthesis using (a): ranked molecular mass, and ranked molecular charge, (b): ranked molecular mass, and ranked MolLogP, and (c): ranked MolLogP and ranked molecular charge. Unabbreviated metabolite names are available in Table B.2

When highlighted in the *i*ML1515 network, shown in Figure 4.24, distinguishing between the intermediate metabolites and identifying the direction of reactions between them is very difficult. As we have seen previously for the TCA cycle, using ranked descriptors shown in Figure 4.25 leads to improvement in overall clarity due to spatial separation of nodes, and the pathway is more easily read when compared to its representation in Figure 4.24. With the exception of ADP and ATP, identifying a metabolite in the ATP biosynthesis case study dataset is nearly impossible, as demonstrated when comparing the *i*ML1515 network in Figure 4.15b and with the ATP biosynthesis dataset highlighted in Figure 4.25.

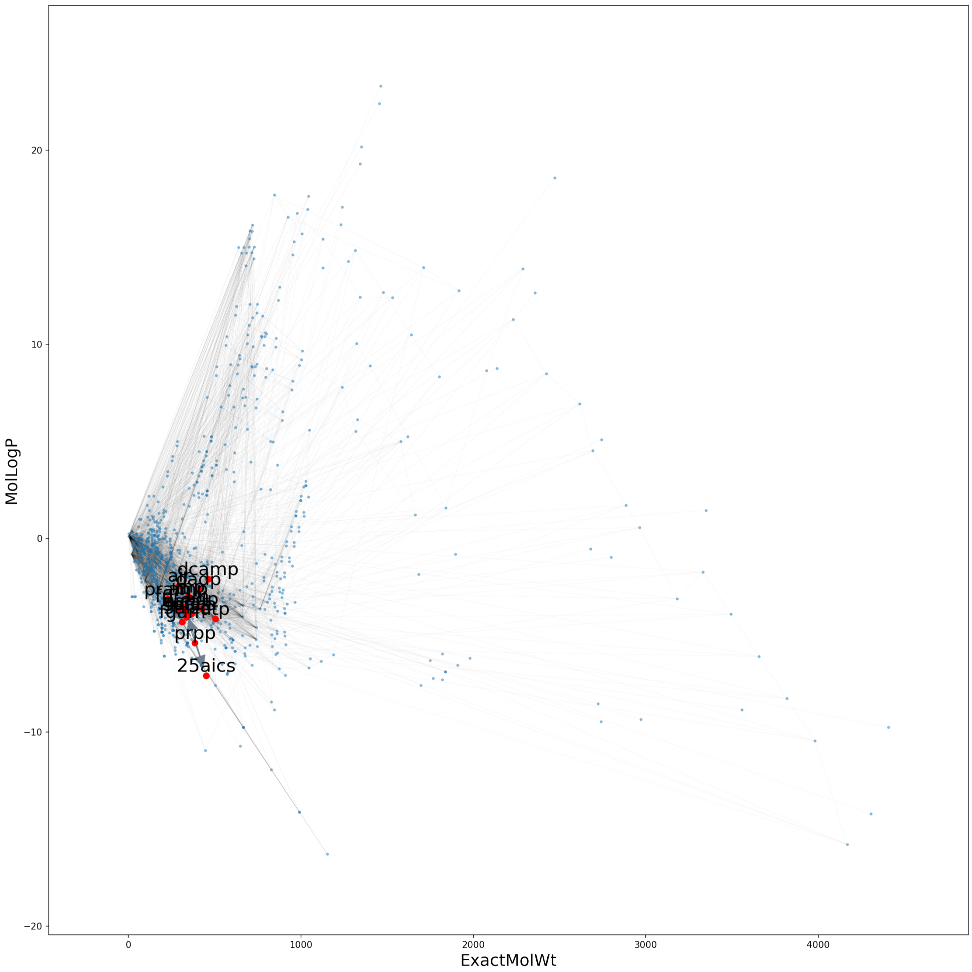


Figure 4.24: The ATP biosynthesis pathway is highlighted in red in the *i*ML1515 metabolic network visualized with ExactMolWt and MolLogP. Unabbreviated metabolite names are available in Table B.2

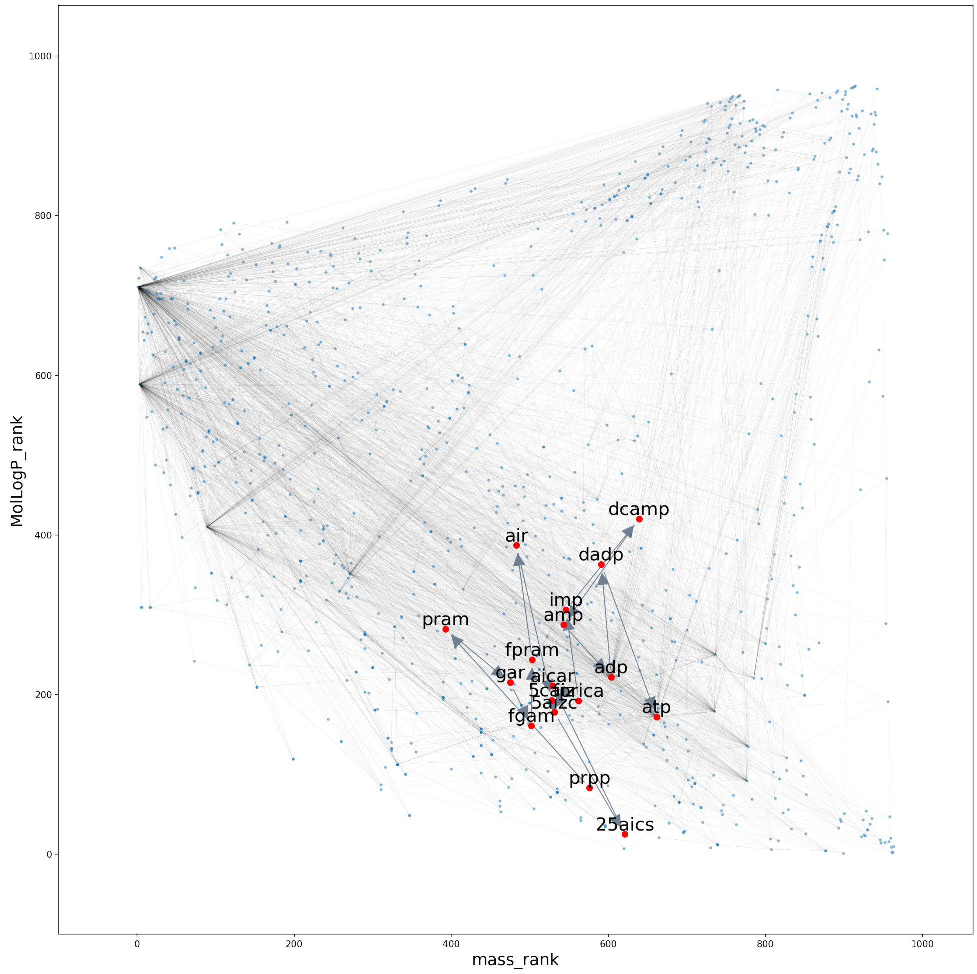


Figure 4.25: The ATP biosynthesis pathway highlighted in red in the *i*ML1515 metabolic network is visualized with mass_rank and MolLogP_rank. Unabbreviated metabolite names are available in Table B.2

Case study 5: CTP biosynthesis

The increase in molecular weight between the intermediate metabolites in subsequent reactions in the CTP biosynthesis pathway (see Figure C.5) generates layouts that when visualized without the context of a metabolic network, are illustrated in Figures 4.26, 4.27 and 4.28, allow for the identification of metabolites, and reactions easier. Exceptions to this are shown in Figures 4.26c-d and 4.28c where following the direction of the pathway is made difficult with nodes stacking on top of each other.

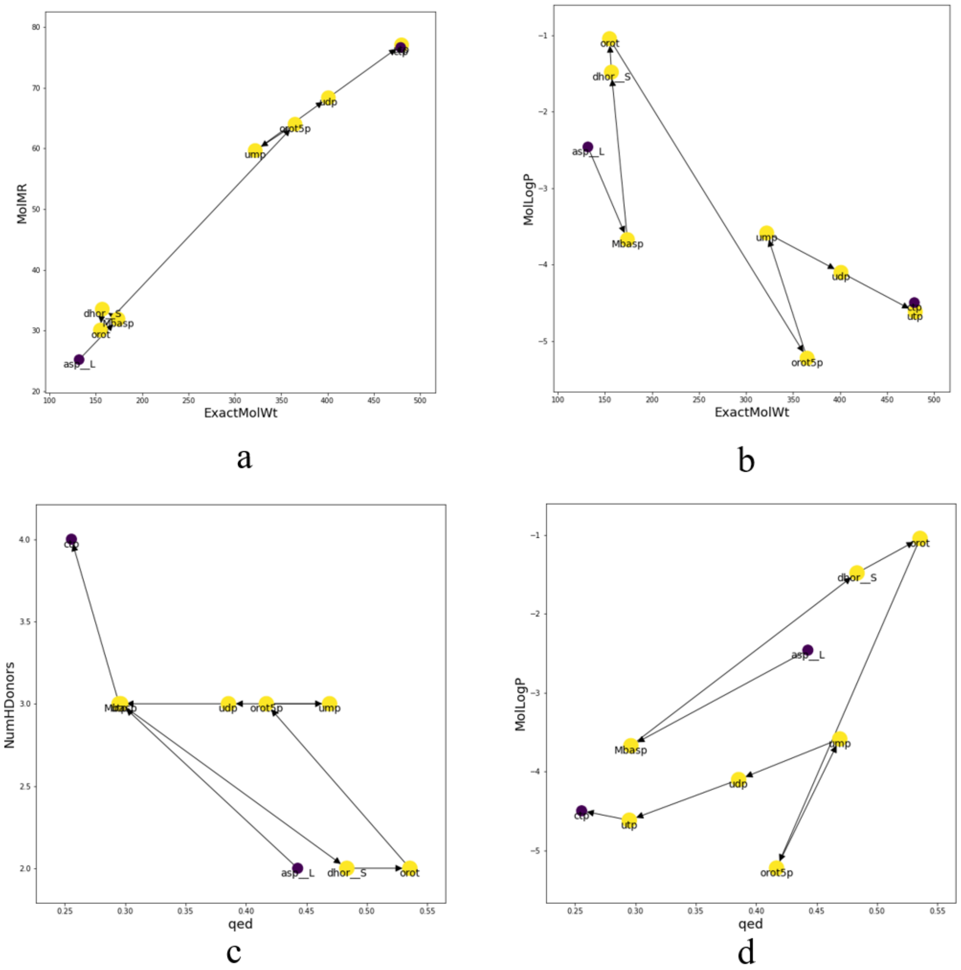


Figure 4.26: Plots of cytidine triphosphate biosynthesis using: (a): Exact molecular weight and MolLogP, (b): exact molecular weight and MolMR, (c): quantitative estimate of drug-likeness and number of hydrogen donors and (d): quantitative estimate of drug-likeness and MolLogP. Unabbreviated metabolite names are available in Table B.3

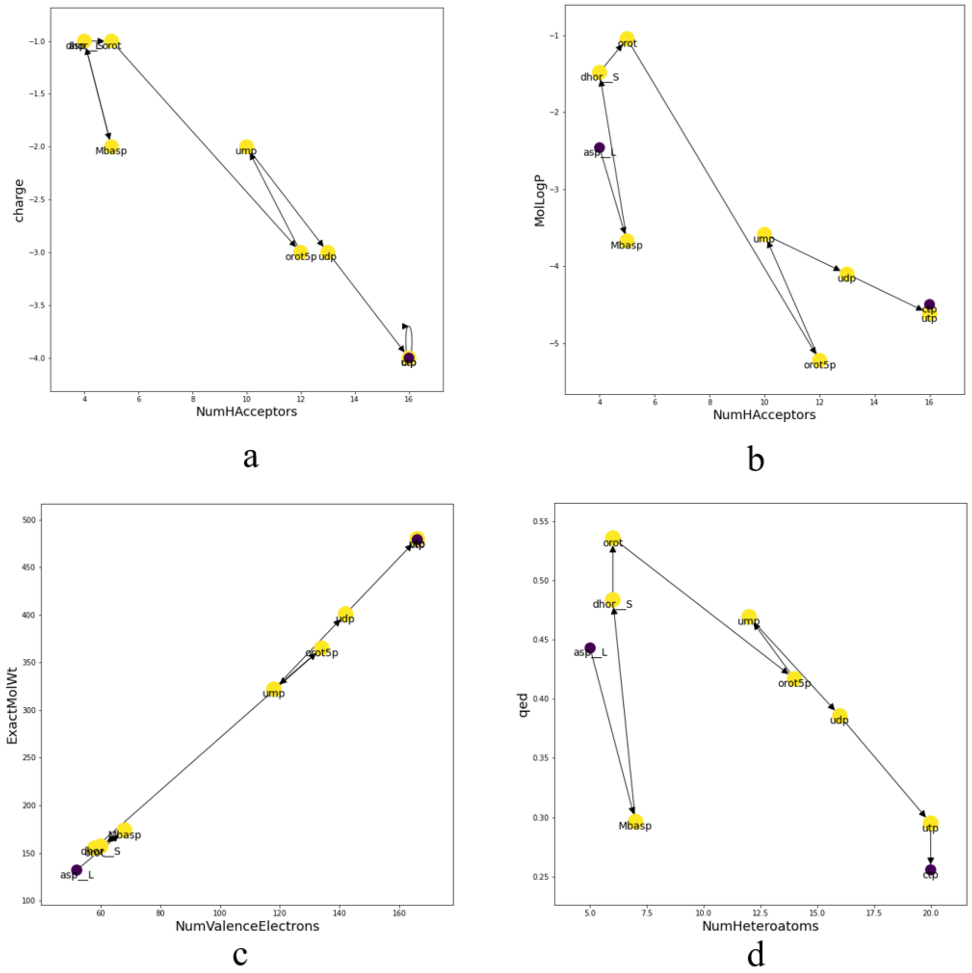


Figure 4.27: Graphs of cytidine triphosphate biosynthesis using: (a): NumHAcceptors and molecular charge, (b): NumHAcceptors and MolLogP, (c): NumValenceElectrons and ExactMolWt (d): NumValenceElectrons and quantitative estimate of drug-likeness. Unabbreviated metabolite names are available in Table B.3

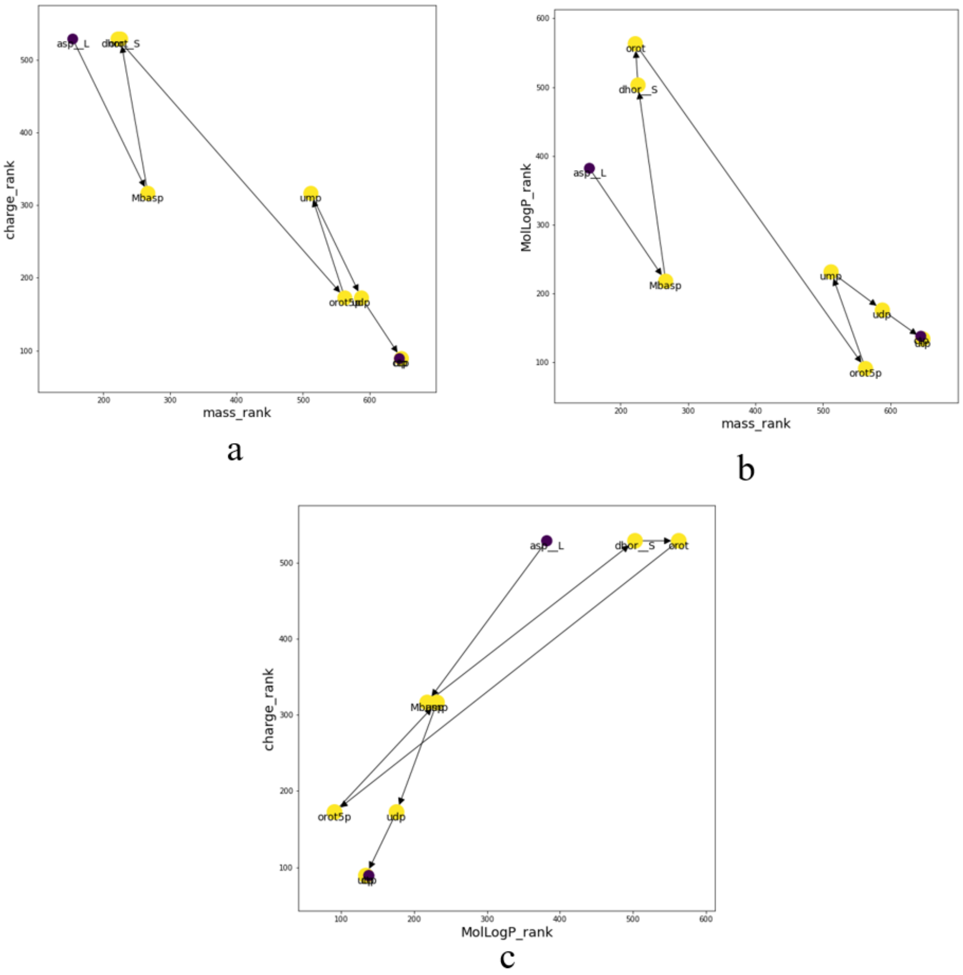


Figure 4.28: Graphs of cytidine triphosphate biosynthesis using (a): ranked molecular mass, and ranked molecular charge, (b): ranked molecular mass, and ranked MolLogP, and (c): ranked MolLogP and ranked molecular charge. Unabbreviated metabolite names are available in Table B.3.

When viewed in the context of a metabolic network, the use of un-ranked descriptors leads intermediate metabolites to be positioned in close proximity with each other. This detracts from the readability we observed when the pathway was visualized locally, demonstrated when comparing 4.26b and 4.29. An improvement in readability when visualized with ranked descriptors, is shown in Figure 4.30.

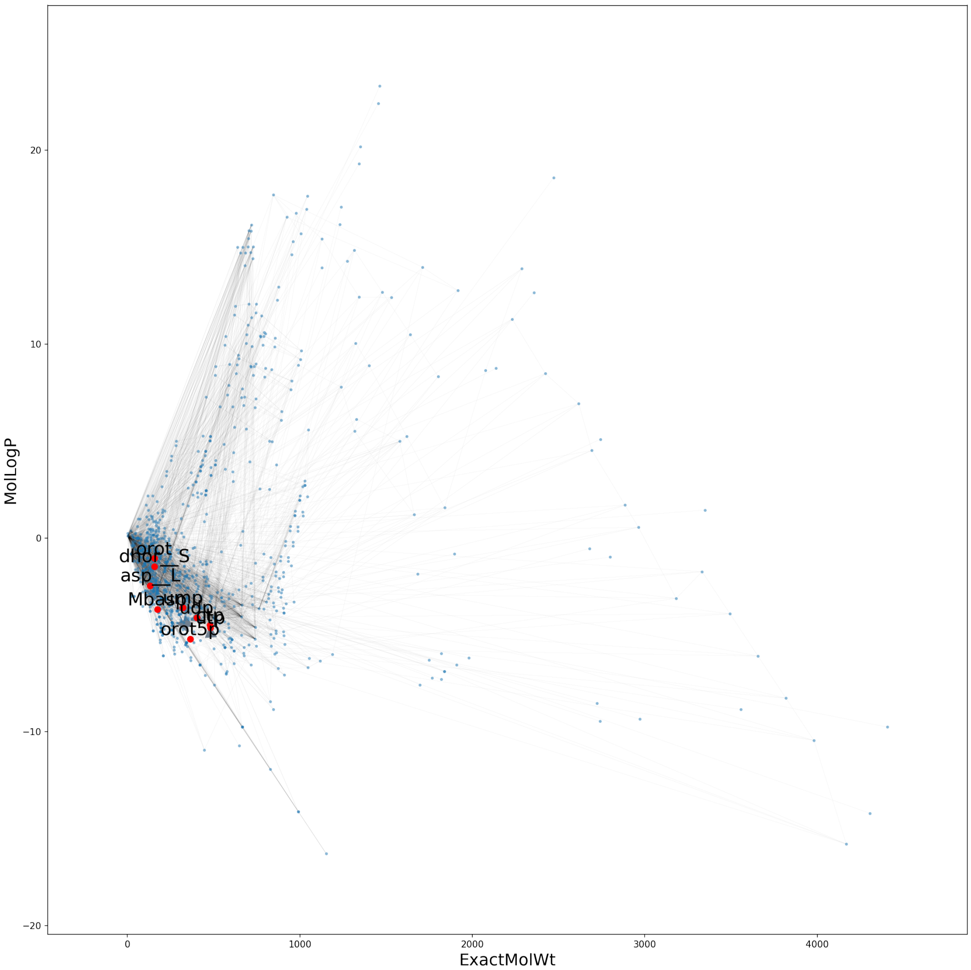


Figure 4.29: The CTP biosynthesis pathway is highlighted in red in the *i*ML1515 metabolic network visualized with ExactMolWt and MolLogP. Unabbreviated metabolite names are available in Table B.3.

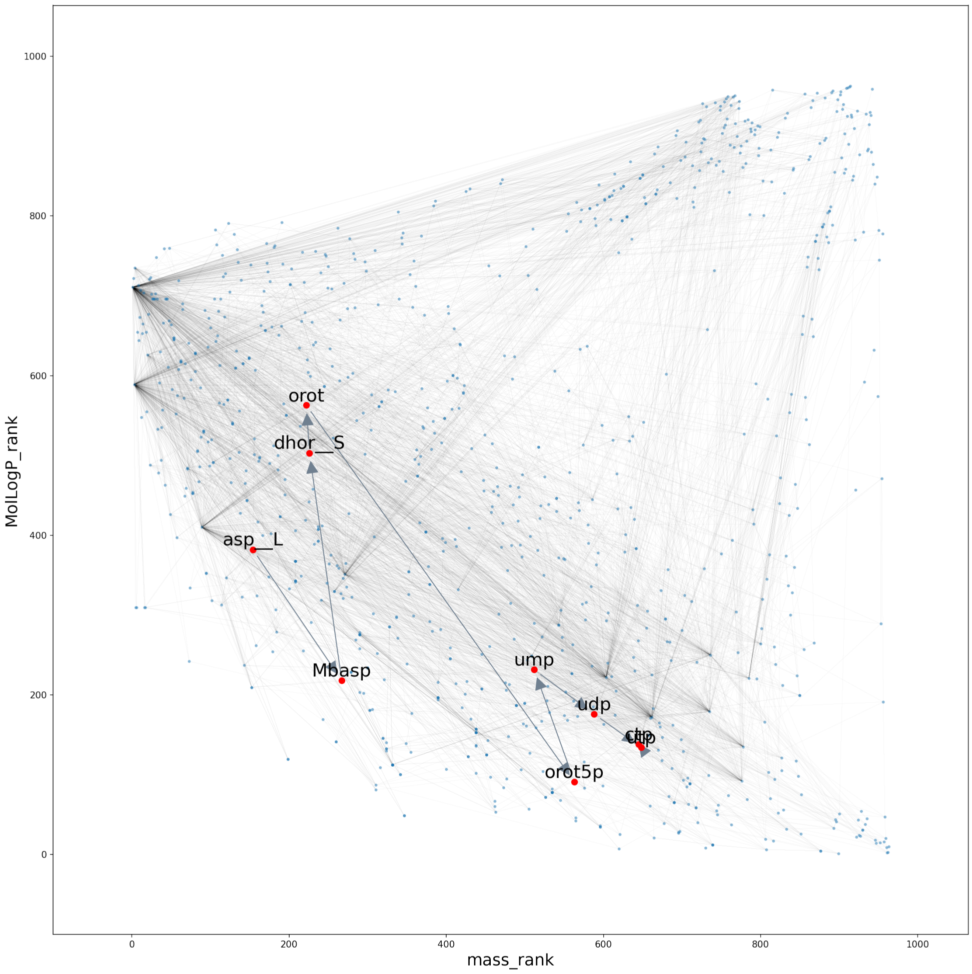


Figure 4.30: The CTP biosynthesis pathway is highlighted in red in the *i*ML1515 metabolic network visualized with mass_rank and MolLogP_rank. Unabbreviated metabolite names are available in Table B.3.

Case study 6: GTP biosynthesis

Most layouts generated for the GTP biosynthesis pathway have similar patterns as those observed for the ATP biosynthesis dataset. This is expected due to the number of intermediate metabolites that participate in both pathways. This creates a recurring motif of nodes with similar chemical properties that cluster together, as shown in Figures 4.31, 4.32 and 4.33.

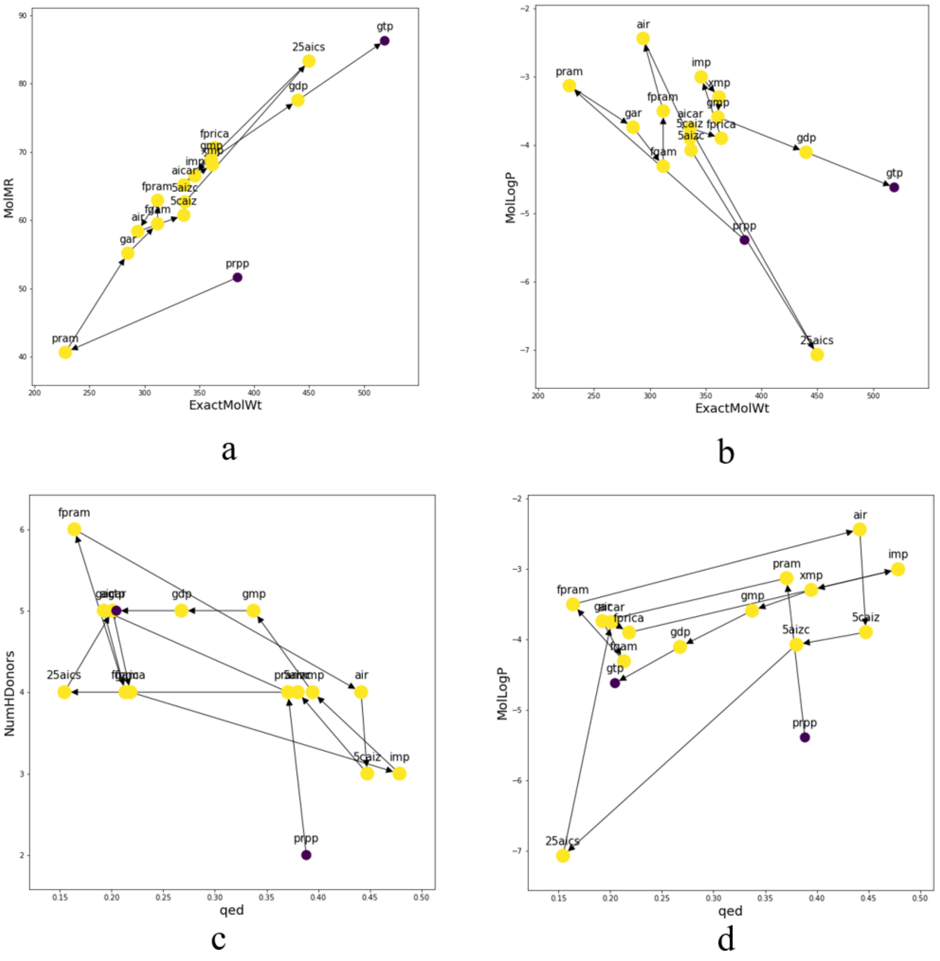


Figure 4.31: Plots of guanosine triphosphate biosynthesis using: (a): Exact molecular weight and MolLogP, (b): exact molecular weight and MolMR, (c): quantitative estimate of drug-likeness and number of hydrogen donors and (d): quantitative estimate of drug-likeness and MolLogP. Unabbreviated metabolite names are available in Table B.4.

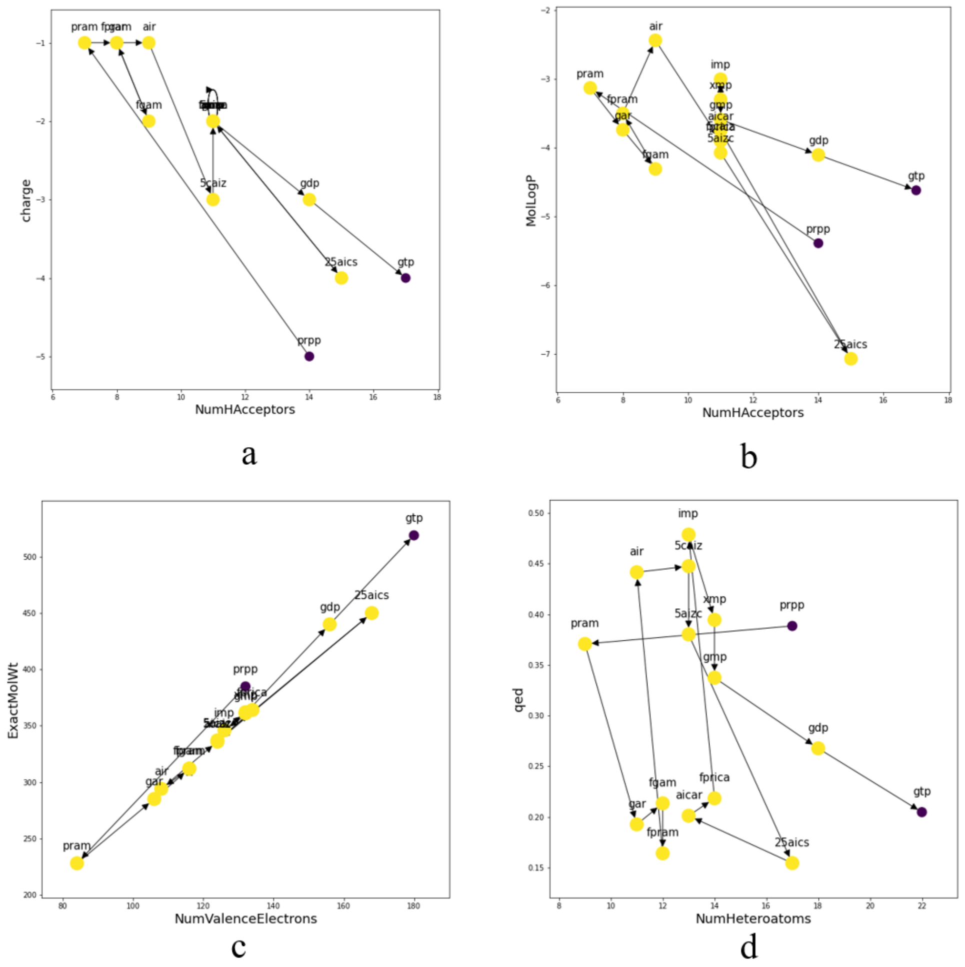


Figure 4.32: Graphs of guanosine triphosphate biosynthesis using: (a): NumHAcceptors and molecular charge, (b): NumHAcceptors and MolLogP, (c): NumValenceElectrons and ExactMolWt (d): NumValenceElectrons and quantitative estimate of drug-likeness. Unabbreviated metabolite names are available in Table B.4.

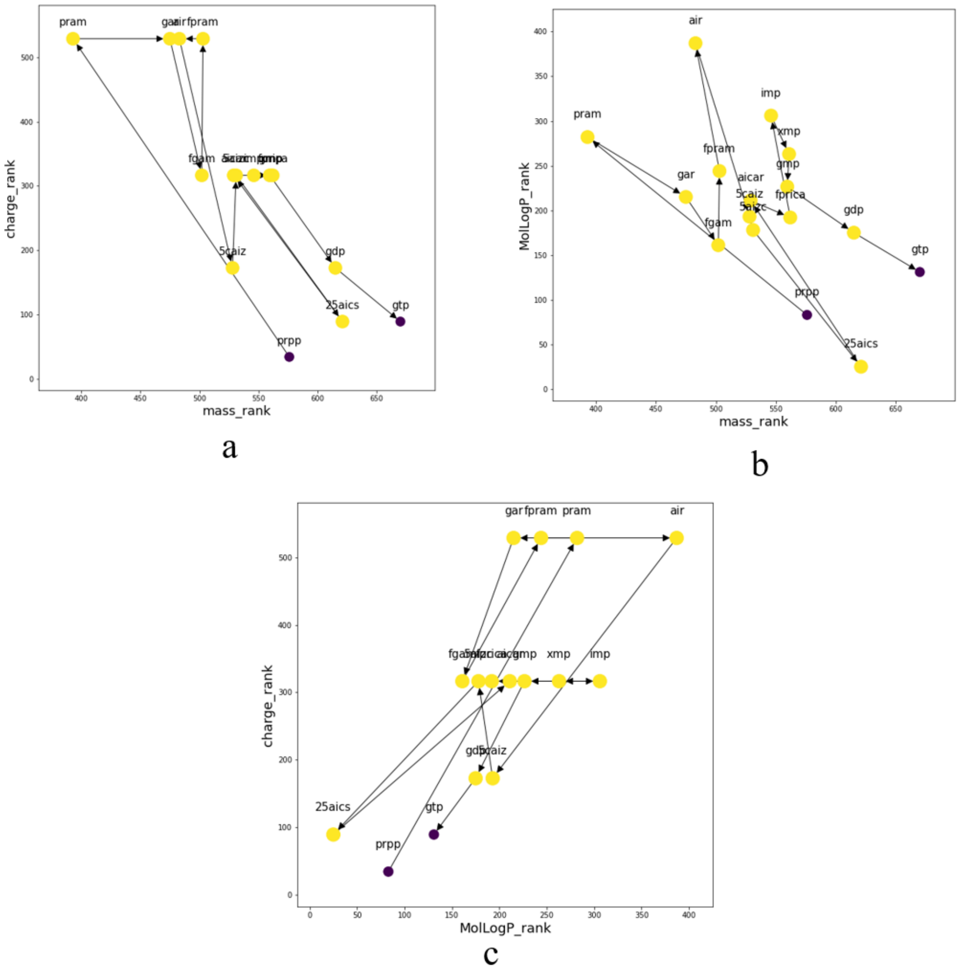


Figure 4.33: Plots of guanosine triphosphate biosynthesis using (a): ranked molecular mass, and ranked molecular charge, (b): ranked molecular mass, and ranked MolLogP, and (c): ranked MolLogP and ranked molecular charge. Unabbreviated metabolite names are available in Table B.4.

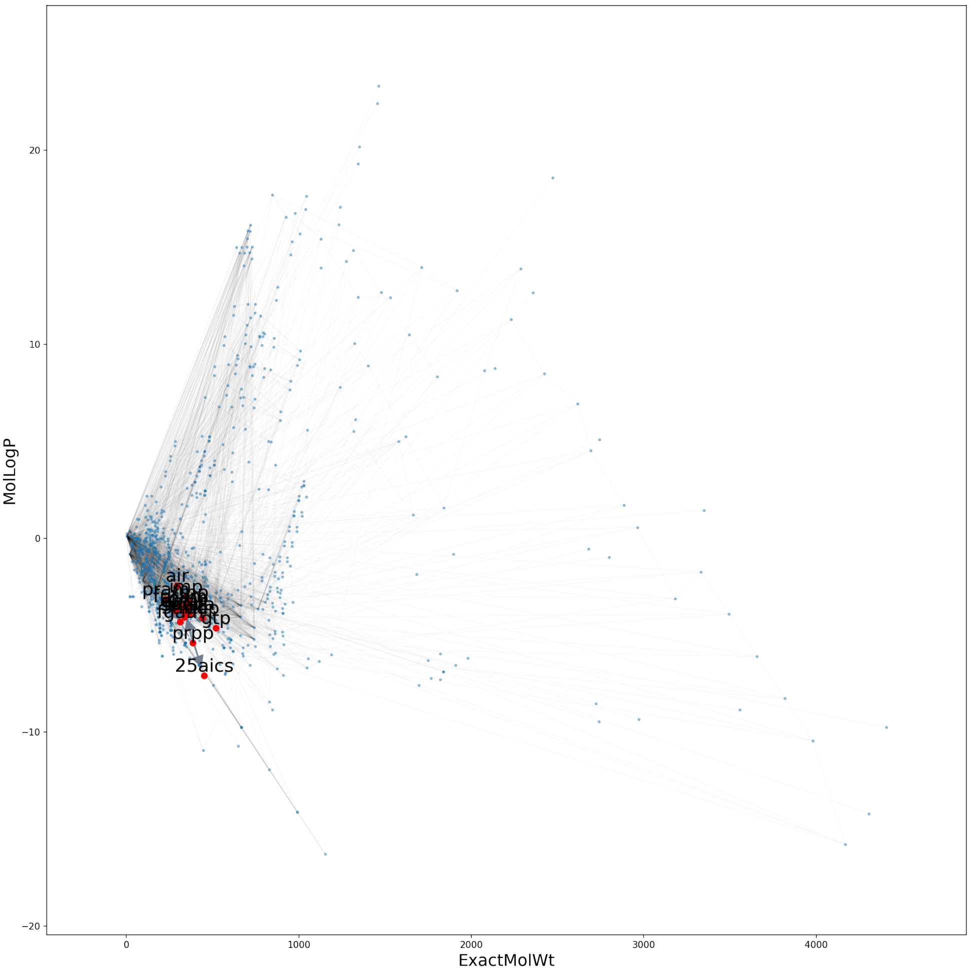


Figure 4.34: The GTP biosynthesis pathway is highlighted in red in the ML1515 metabolic network visualized with ExactMolWt and MolLogP. Unabbreviated metabolite names are available in Table B.4.

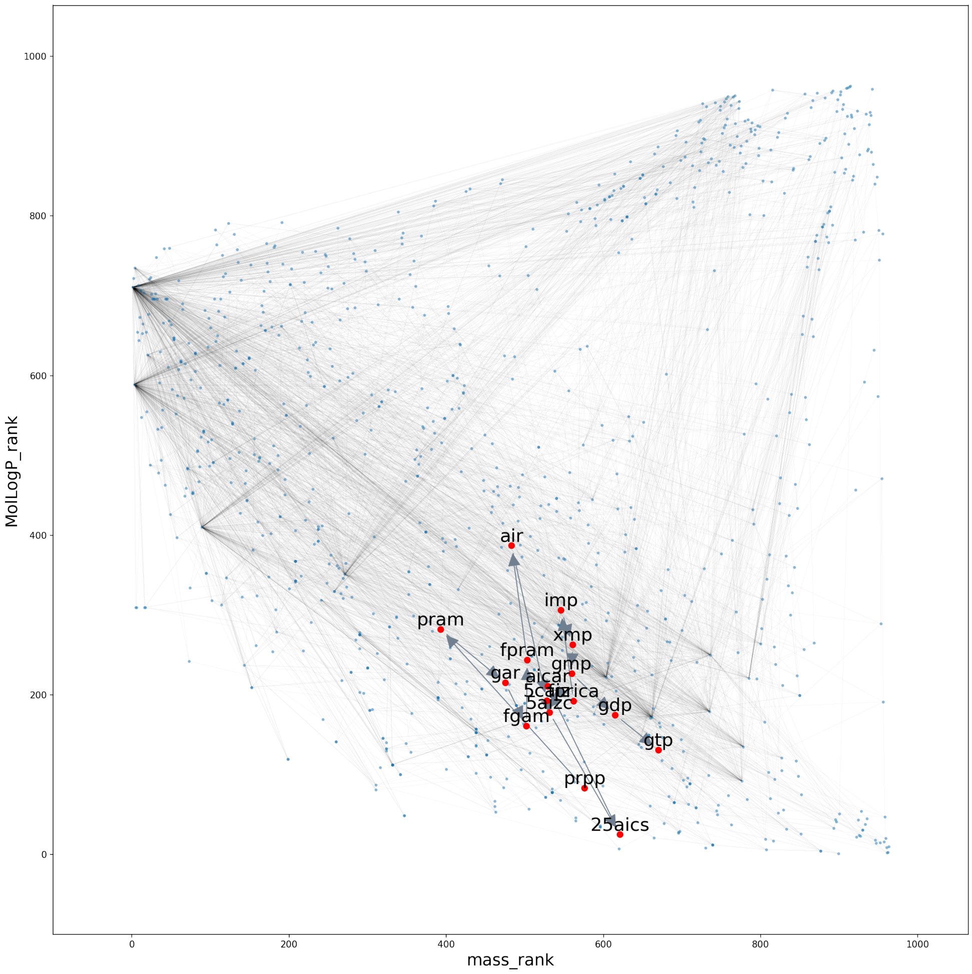
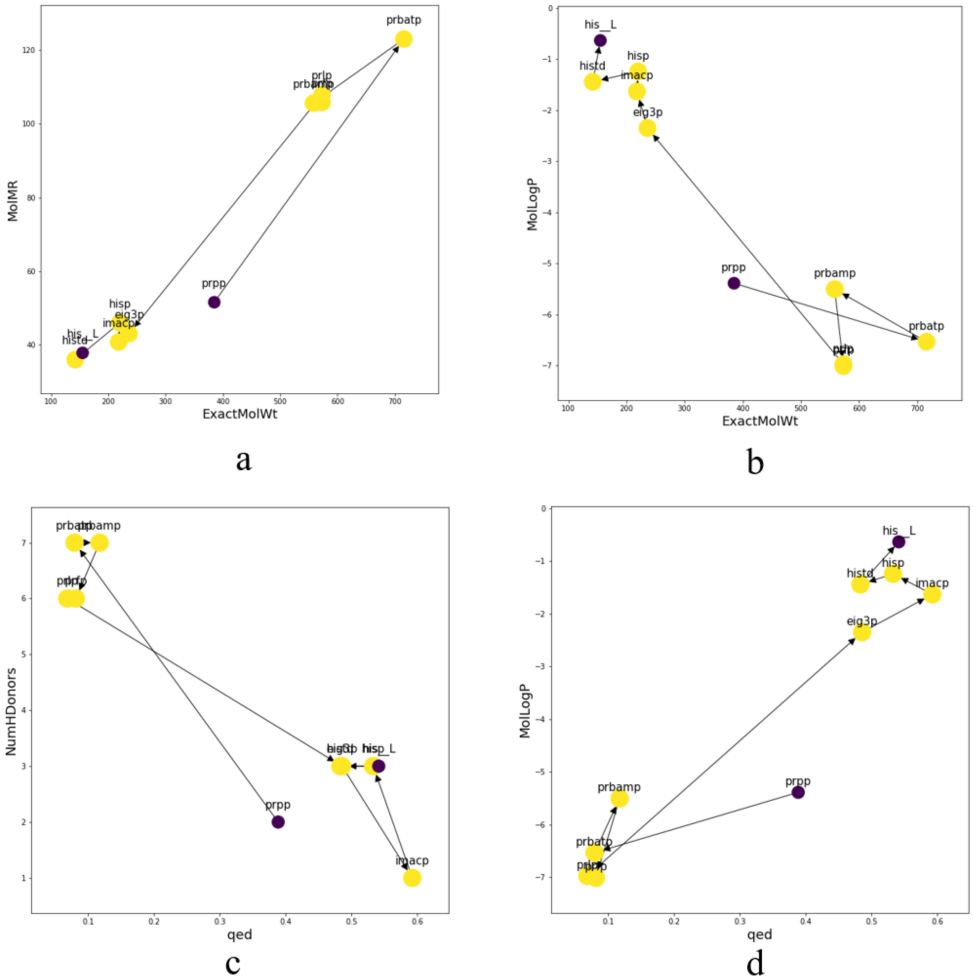


Figure 4.35: The GTP biosynthesis pathway is highlighted in red in the ML1515 metabolic network visualized with mass_rank and MolLogP_rank. Unabbreviated metabolite names are available in Table B.4.

Case study 7: Histidine biosynthesis

In general, due to the gradual gain in chemical similarity in intermediates toward the formation of histidine (his_L), most layouts for the histidine biosynthesis dataset are legible, however, a recurring motif of nodes at the end detracts from the overall clarity of the pathway. Most layouts for this case study manage to capture the reduction in molecular weight between the intermediate metabolites 5-[(5-phospho-1-deoxy-D-ribulos-1-ylimino)methylamino]-1-(5-phospho-beta-D-ribosyl)imidazole-4-carboxamide (prlp, 573 Da) and D-erythro-1-(imidazol-4-yl)glycerol 3-phosphate (eig3p, 236 Da) that occurs in the glutamidotransferase reaction which is catalyzed by the enzyme imidazole glycerol phosphate synthase (EC 4.3.2.10) [67, 171]. This departure in the chemical similarity between intermediate metabolites is visually quantified by the rendered length of the edge associated with this reaction. The edge representing the aforementioned reaction varies in length, depending on the molecular descriptors that the respective layout is generated from. Figures 4.36, 4.37 and 4.38 show that the length of the edge representing this reaction remains approximately the same except for Figures 4.37a and 4.38a, where the difference in molecular charge positions eig3p and prlp closer and the edge is shorter.



a

b

c

d

Figure 4.36: Plots of histidine biosynthesis using: (a): Exact molecular weight and MolLogP, (b): exact molecular weight and MolMR, (c): quantitative estimate of drug-likeness and number of hydrogen donors, and (d): quantitative estimate of drug-likeness and MolLogP. Unabbreviated metabolite names are available in Table B.5.

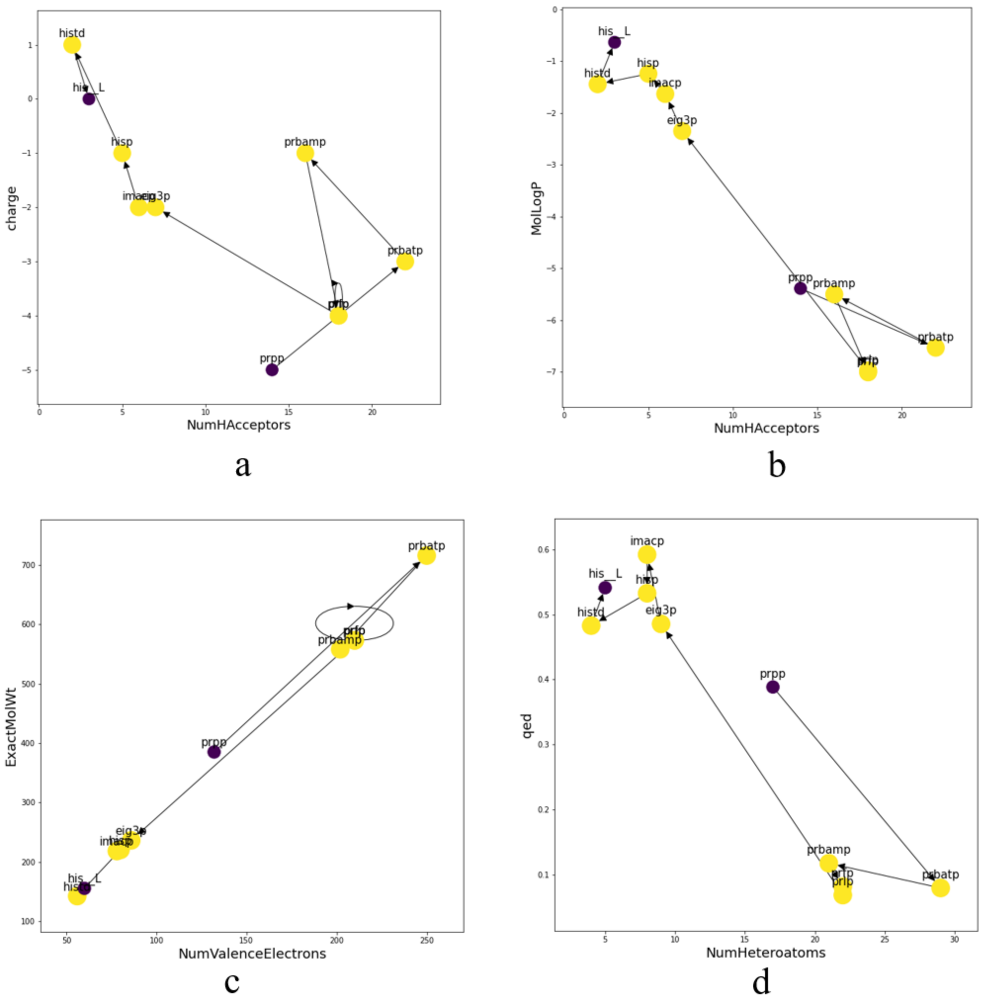


Figure 4.37: Graphs of histidine biosynthesis using: (a): NumHAcceptors and molecular charge, (b): NumHAcceptors and MolLogP, (c): NumValenceElectrons and ExactMolWt (d): NumValenceElectrons and quantitative estimate of drug-likeness.Unabbreviated metabolite names are available in Table B.5.

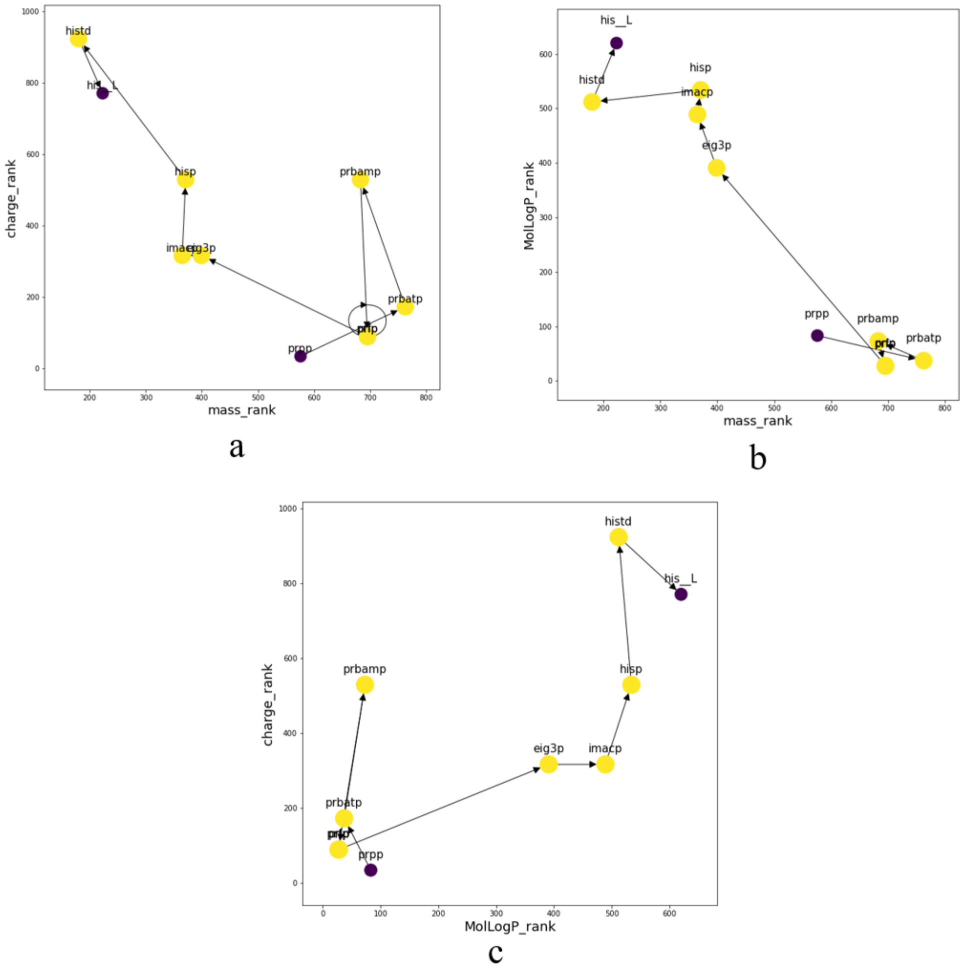


Figure 4.38: Plots of histidine biosynthesis using (a): ranked molecular mass, and ranked molecular charge, (b): ranked molecular mass, and ranked MolLogP, and (c): ranked MolLogP and ranked molecular charge. Unabbreviated metabolite names are available in Table B.5.

We have seen in the previous case studies that, when highlighted in a metabolic network, the descriptor pair mass_rank and MolLogP_rank improve the legibility of the metabolic pathway layout when compared with their unranked counterparts, shown in Figures 4.40 and 4.39 respectively.

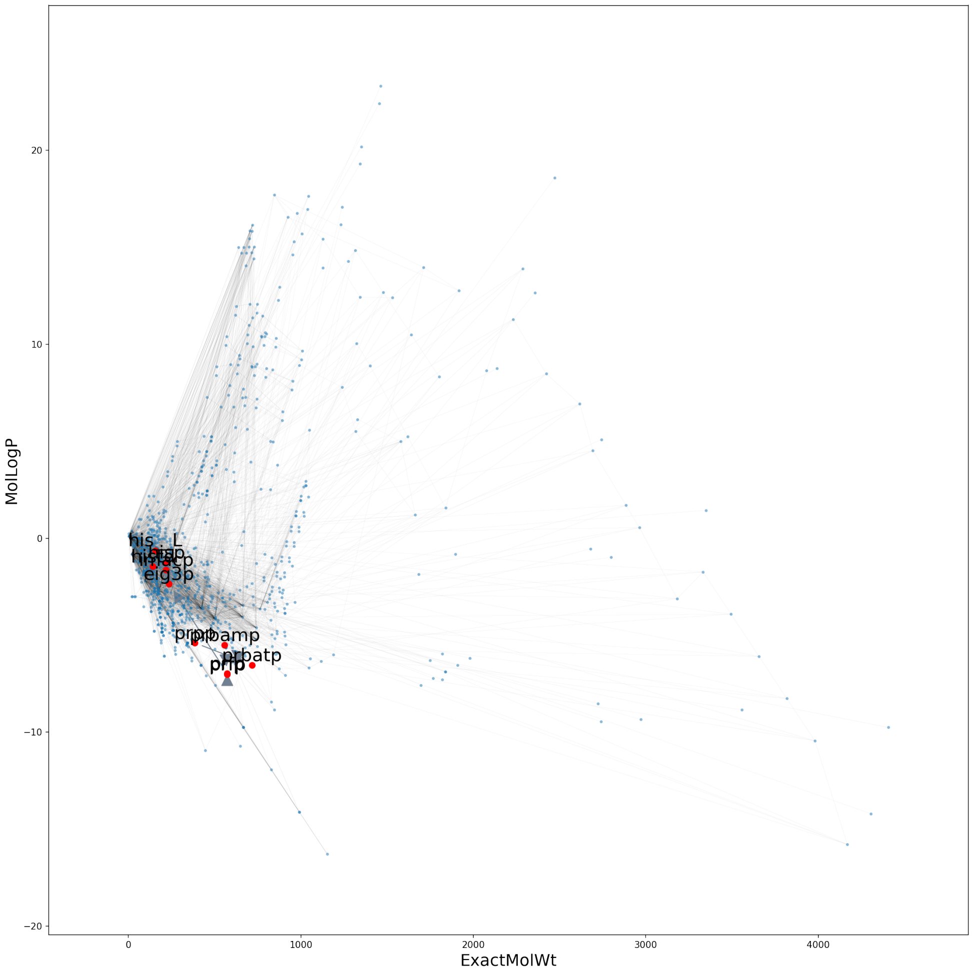


Figure 4.39: The histidine biosynthesis pathway is highlighted in red in the *i*ML1515 metabolic network visualized with ExactMolWt and MolLogP. Unabbreviated metabolite names are available in Table B.5.

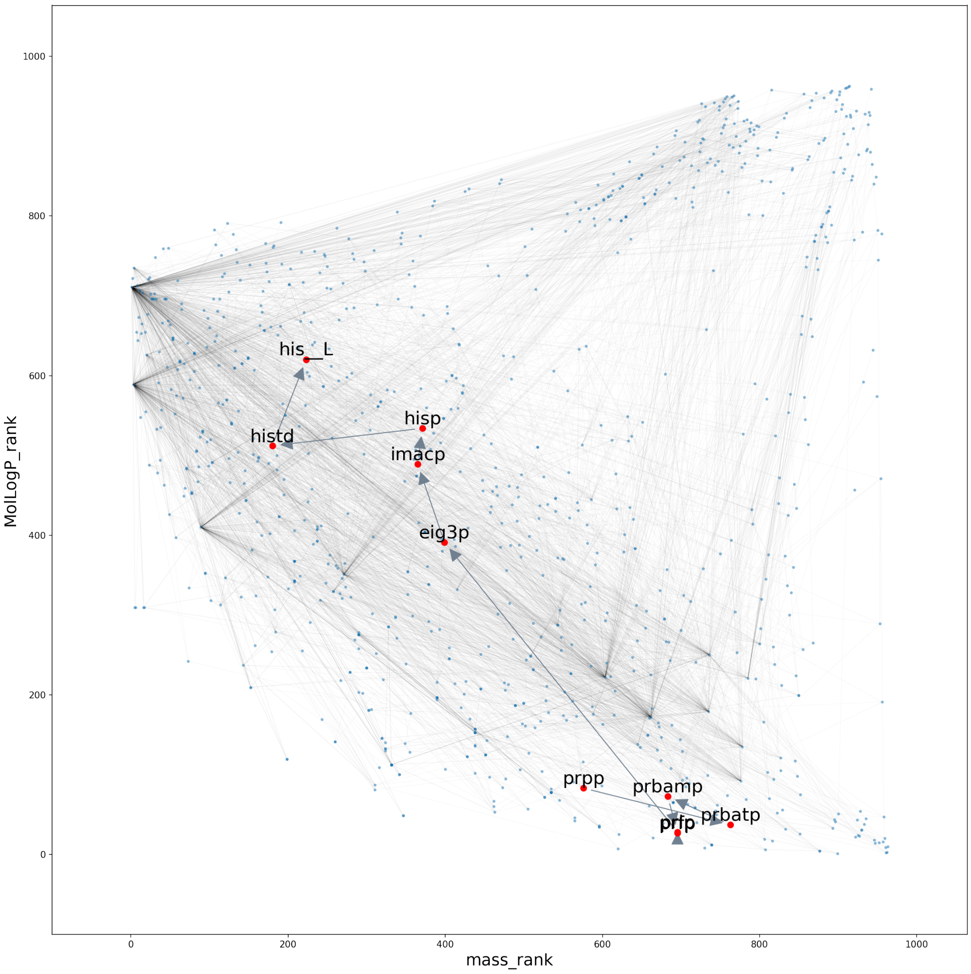


Figure 4.40: Histidine biosynthesis is highlighted in red in the ML1515 metabolic network visualized with mass_rank and MolLogP_rank. Unabbreviated metabolite names are available in Table B.5.

4.3.3 Summary of case study results

- As a result of the clustering of nodes and the consequent edge crossings, the layouts of the metabolic networks are difficult to comprehend, resulting in no distinct network motif or structure that can be interpreted as a metabolic pathway.
- Shorter metabolic pathways with steady incremental changes in the chemical profile of intermediate metabolites towards that of the end product provided the most intuitive layouts when visualized in a local context.
- It was found that the layouts generated using the pair of ranked descriptors, mass_rank, and MolLogP_rank, was easier to read as they reflect the ranks of the chemical similarity of metabolites.
- When highlighted in a metabolic network, metabolic pathways were consistently easier to read when visualized using mass_rank and MolLogP_rank, rather than their unranked counterparts, ExactMolWt and MolLogP.
- There is a correlation between the rendered length of an edge between two metabolites and their chemical similarity, as defined by the pair of molecular descriptors employed to generate the layout. Longer edges indicate less chemical similarity, while shorter edges indicate greater similarity.

4.4 Export of biochemical coordinate layouts

A number of formats compatible with commonly used network visualization software were exported for layouts generated in each of the 7 case studies using the method described in section 3.4. In Figure 4.41 we see the metabolic network case studies rendered in Cytoscape. Table A.1 Supplementary data 4, Appendix A, contains the exported layouts of *E. coli*'s central metabolism, *i*ML1515, and the TCA cycle in the following formats: JSON, Cytoscape JSON (CyJS), and Graph Markup Language (GraphML).

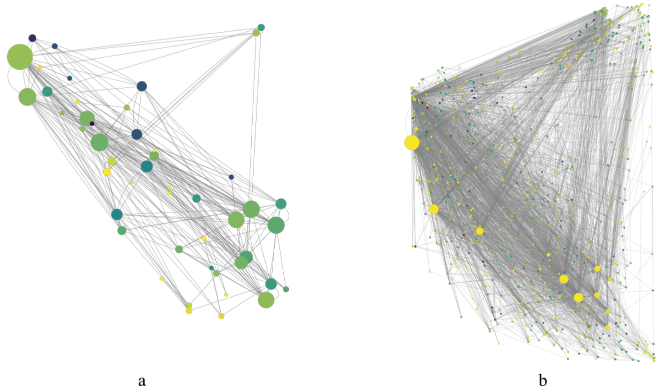


Figure 4.41: Metabolic network layouts generated with mass_rank and MolLogP_rank of (a) the *E. coli* central metabolism and (b) *i*ML1515. Node hue set by centrality measure. Node size is set by its connectivity degree. Visualized in Cytoscape [81].

CyJS and GraphML formats store all available node attributes when generated from Cytoscape. Using these formats, it was possible to import and save molecular descriptor data. The graphs stored in these formats can then be used to visualize layouts based on the molecular descriptors available. The Cytoscape JSON (CyJS) format demonstrated the capacity to visualize the layout when opened in Cytoscape without mapping node x-and y-coordinates to descriptor values. The GraphML and JSON formats lacked this ability and required re-mapping node positions when read by Cytoscape. The GraphML format is compatible with ySBGN [159], a tool that interconverts between GraphML and SBGN-ML. This tool was not compatible with Cytoscape-derived GraphML files, as the necessary data for converting to SBGN-ML was lost when using NetworkX's export to Cytoscape function.

Chapter 5

Discussion

This chapter encompasses the discussion of the results of this project. First, the results from the data collection and molecular descriptor calculation are presented. A brief discussion of the implications that several metabolites in the *i*ML1515 model lacked molecular representations is presented, along with a possible explanation for why there were differences in descriptor values between molecular representation formats. Next, the results from the descriptor analysis and selection process are discussed. Last, an assessment of the visualized layouts' ability to permit biological interpretation and an evaluation of which descriptor combinations produced the most intuitive visualizations are provided.

5.1 Molecular descriptors

5.1.1 Some MetaNetX identifiers were ambiguous or deprecated

Reviewing the chemical library mapping results from the *i*ML1515 model revealed that metabolites that were not matched with a molecular representation after their collection had MNXref identifiers that were either deprecated or represented large molecules that lacked an associated molecular representation (see Table 4.2). Several databases used in genome-scale metabolic modeling, including MetaNetX, allow for non-systematic, ambiguous naming conventions [161]. Recognizing that these metabolites could exist under different names, a list of possible synonyms was made to be parsed through the external databases, ChEBI [104] and ChemSpider [172]. Because of the inherent uncertainty, described in Ref [161], and the possibility of adding a similarly named metabolite that did not exist in the model, metabolites without molecular representations was removed. Consequently, the number of metabolites without a molecular representation (see Table 4.4) for the *i*ML1515 model eliminated the possibility of generating layouts for some pathways, e.g. fatty acid synthesis. This was because the majority of the metabolites lacking chemical representations had large cofactors associated with them, for example, acyl carrier proteins (ACP). It has been suggested to use unique, database-independent identifiers such as InChI to represent metabolites in GEMs, but adoption of such a standard has not been widespread [58]. A molecular representation format, BiGSMILES [173], has been proposed as a means of encoding larger molecules such as macromolecules and polymers. However, as of 2022, it is unknown if this format will be compatible with RDkit.

5.1.2 Molecular descriptor calculation

A broad range of molecular descriptors can be calculated from representations of biomolecules, but some of these descriptors have null values or remain constant for many [116]. Following the removal of these descriptors during the descriptor analysis process which will be discussed in the next section, it was found that there was a lack of uniform correlation between the models for the same descriptor. In the *i*ML1515 model, MolLogP showed a low correlation to molecular weight while in the core metabolism model it is negatively correlated (see Figure 4.3). MolLogP, which quantifies lipophilicity, increases with a decrease in the polar fraction of molecules. A likely explanation is that the metabolites in the core metabolism model are more chemically similar than those in *i*ML1515 with regard to the chemical basis of their MolLogP calculation. This can be attributed to the number of functional groups and sub-structures which contribute to hydrophobicity in the calculation of MolLogP are numerically more prominent among metabolites with lower weight and less prominent with those that are heavier (see fig 4.10a) in the central metabolism model. In *i*ML1515 there is a significantly larger number of metabolites with similar molecular weight and MolLogP values and relatively fewer, heavy metabolites with low MolLogP values (see Figure 4.13a). This might explain this lack of uniformity in the correlation of MolLogP. Similar dynamics might explain differences in the correlation of other descriptors between these models.

A less likely explanation is rooted in the observed differences in molecular descriptor values for the same metabolite calculated from different molecular representations. When calculated from their SMILES representations, some molecules with complex stereochemistry had significantly higher MolLogP and BCUT values compared to their InChI based calculations [174]. The large multi-subunit metabolite, listed in Table 4.7, named "three disacharide linked murein units (tetrapeptide crosslinked tetrapeptide (A2pm->D-ala), one uncrosslinked tetrapaptide) (middle of chain)" had the largest difference in MolLogP between the representations used in this project. Evaluating the veracity of the calculated values from both representations was done by comparing them to the listed MolLogP value for its entry in PubChem [152]. It was revealed that the InChI calculated value of -9.45 was the closest to the PubChem listed value of -16.5 since the SMILES value was -30.3. This indicates the possibility that the positioning of the polar groups is based on the different interpretations of molecular structures and topology between these formats in RDkit. As InChI is superior at encoding multiple types of stereochemistry than SMILES (see section 2.4.3), the numerical differences that occur in topological descriptors and their frequency (see Table 4.6) might be caused by different molecular structures reconstructed from their molecular representations. This is compounded by the fact that the PubChem listed MolLogP value for this molecule is calculated by XlogP3 3.0 [175], a tool dedicated to calculating LogP for molecules. This tool only accepts MolFiles [176] (.mol), a molecular representation format that contains the three dimensional coordinates of atoms to reconstruct a three-dimensional representation of the molecular structure. In the context of molecular descriptors, this format allows for the calculation of descriptors based on their three-dimensional structure, which is conducive to accurate calculations of MolLogP and similar descriptors derived from a molecule's topology or geometrical shape.

Corroborating the calculated ExactMolWt values of [4Fe-4S](2+) (MNXM732007), presented in Table 4.7, with its entries in MetaNetX and ChEBI revealed that the difference in calculated molecular weight was a result of how the InChI representation format stores information about its protonation state. In this case, the InChI string for [4Fe-4S](2+) was its unprotonated state, i.e. no hydrogen atoms because the hydrogen sub-layer is empty (see Figure 2.21) and was calculated based on that information. By design, SMILES do not encode hydrogen atoms but are added after loading into cheminformatic software. RDkit's implementation of ExactMolWt calculation from SMILES differs from its InChI in that it adds hydrogens to the atoms, in line with the SMILES notation rule 2 (see Section 2.4.3). This is the probable cause for this specific molecule, as the differences in molecular weights amount to 4 hydrogen atoms. This difference in molecular structure interpretation is unlikely the sole reason for the majority of numerical discrepancies between the molecular representation formatsformats.

InChITrust [141], the organization which oversees the maintenance and development of the InChI format allows a single InChI to represent the zwitterionic and neutral states of a molecule. This means that one InChI can represent two different entries on PubChem, Chemspider, or MetaNetX while requiring two InChIs for its anionic and cationic states [177]. As a result of the specific nature of structural information in the InChI format, numerical discrepancies between the two formats are most likely to occur. A sufficiently accurate and unique chemical representation is difficult to establish, and although both formats use a line notation system that confers advantages in practicality, it also limits tautomeric and geometrical information about molecules.

In the context of this project, the choice of molecular representation format needs to be addressed. There were more SMILES available for the *i*ML1515 dataset than InChIs and the consequences of excluding 40 metabolites when choosing InChIs are hard to disparage. It is difficult to determine the precise reason for the numerical differences in the topographical descriptor values between the representation formats in the absence of a thorough analysis of how stereochemistry is encoded differently across formats for all metabolites. To determine which format best represents the chemical properties of molecules, it would be prudent to verify the calculated molecular descriptor values with those hosted in external databases. In this regard, it is important to note that the molecular representation format reported for a metabolite is not necessarily canonical. As described in section 2.4.3, there are multiple approaches to generating SMILES for molecules, whereas InChI has a single algorithm. In some cases, SMILES may be calculated using different algorithms depending on the metabolite. In spite of the fact that SMILES is the most widely used molecular representation format, it suffers from ambiguity and uncertainty associated with its representation of stereochemistry [138]. InChI strings have been used as the basis for a universal SMILES format, but this format has not yet been widely adopted [138].

5.1.3 RDkit failed to compute two descriptors for a small number of metabolites

The majority of metabolites had molecular descriptors successfully calculated. In the case of 45 metabolites, two families of descriptors representing electropological and intramolecular interactions, EstateIndex and BCUT, could not be calculated for both molecular representation formats. Because most of these compounds are ionized atoms, they were initially thought to not have the necessary structure to calculate these descriptors. An answer to this question came from the analysis of the metabolite adenosylcob(III)inamide-GDP, a conjugate base of adenosylcobinamide guanosyl diphosphate. In this case, RDkit was only unable to calculate BCUT descriptors for this molecule. This was found to be caused by the large size of the molecule, resulting in a memory crash mid-calculation caused by the Boost.Python implementation for this particular descriptor calculation function. Peculiarly, this event does not throw an error after calculation [178], notifying the user. Following the suggested solution to a similar issue, [165] by RDkit's chief developer, Greg Landrum, BCUT descriptors were successfully computed for adenosylcob(III)inamide-GDP, and the 44 remaining metabolites had their missing descriptors calculated as well.

The results from molecular descriptor calculation demonstrate the importance of rigorous and comprehensive evaluation of their calculation across a variety of cheminformatic tools and molecular representation formats.

5.2 Descriptor selection

5.2.1 Descriptor analysis

When correlated, empty, and constant descriptors are removed, we ensure that certain types of molecular information are not overrepresented. If one or more descriptors show little or no variation in their values, their inclusion and use in subsequent steps do not provide any benefit to the project.

Analysis of spearman rank correlations coefficients between descriptors revealed that the majority of values in the uncorrelated features (see Figures 4.1 and 4.2) were empty. This was observed when descriptors with 90% of their values equal to zero were removed (see Figure 4.3 and 4.4). Removed descriptors were primarily counts of substructure fragments and functional groups that are not common in biomolecules [179]. These descriptors reflect physiochemical attributes that can be used to infer possible biological activity or function [17, 179].

In *i*ML1515's pruned descriptor set, MolLogP demonstrated the least correlation with any other descriptor which indicated that this feature described information that the other descriptors did not. In the pruned descriptor set for the core metabolism model, MolLogP exhibited more correlation with Chi, kappa, ExactMolWt, NumValenceElectrons, BCUT, and van der Waals surface area descriptors which in *i*ML1515 was less correlated with MolLogP. In the core metabolism, the molecular fingerprint FpDensityMorgan2 was the least correlated.

We also observed that several descriptors were correlated with molecular weight. As biomolecules become larger, their carbon-backbone tends to increase, and thus traits are associated with the number of carbon atoms and length of a carbon-backbone scale in tandem with molecular weight [180]. MolMR which reflects the steric bulk of a molecule and NumValenceElectrons which quantify the number of electrons that can take part in chemical bonds are both positively correlated with ExactMolWt. Overall, the descriptors in the pruned sets were more correlated with each other in the core model than in *i*ML1515. This difference can be attributed to the chemical similarity among the 55 mapped metabolites in the core metabolism model and the diversity of the 964 molecular structures of the mapped metabolites in *i*ML1515, as discussed in the previous section.

5.2.2 Selection of biologically relevant descriptors

Of the 9 descriptors selected from the pruned set of descriptors, listed in Table 4.10, 6 are constitutive and 3 incorporate more information about molecular structure in their calculation. Restricting the number of descriptors that would be used to generate layouts was important. This was to ensure that the time spent evaluating the generated layouts would be within the scope of this project. A number of the descriptors that were not selected either contained overlapping information or quantified computational characteristics. Molecular fingerprints and descriptors such as LabuteASA [181], BCUT, Chi, Kappa, HallKierAlpha, and PEOE were not selected as they represent a molecule's per-atom contribution to the molecular polar surface, intramolecular interactions, types of chemical bonds, and van der Waals surface area(VSA) [181] respectively. Despite carrying important molecular information [163, 164, 182], these descriptors are difficult for a non-expert to interpret in a biological context. The selected descriptors were complemented by the addition of three ranked versions of the descriptors ExactMolWt, MolLogP, and molecular charge. In a biological context, these 3 descriptors provide information about a molecule that is easier to understand.

5.3 Layout visualization

A good visualization of a biological network will enable the viewer to gain an understanding of the degree of interactions between network objects, as well as infer why those interactions are occurring as they do. It is often the spatial layout of data that determines whether the desired information is communicated and interpreted or obscured and difficult to comprehend [84].

Graph layout visualizations of the 7 case studies were made with layouts generated by using the selected descriptors and three ranked descriptor variants (see Tables 4.10 and 4.11) as coordinates. The metabolic network layouts were generally difficult to read and the relationship and information that could be divined from the interaction between nodes are lost in the clustering of nodes and edge-crossings. The chemical similarities among the highly connected nodes [74] lead to a significant amount of edges that cross over each other from nodes that are clustered together. This obfuscates any potential network motif that could be identified as a possible pathway. Using molecular descriptors as coordinates does however arrange nodes in a way so that their position reflects their biochemical properties and consequently groups chemically similar molecules. This is demonstrated perhaps best in Figure 4.10d; where one can see how ubiquinol-8 and ubiquinone-8 are spatially separate from the rest of the network in the upper left corner and two distinct groups of metabolites are located in the lower left and right corners. The group in the lower left is comprised of cofactors; NADPH, coenzyme A and activated metabolites such as succinyl-CoA and acetyl-CoA, and the group in the lower right is comprised of smaller metabolites; hydrogen, ATP, H_20, malate, pyruvate, and inorganic phosphate.

Unsurprisingly, the number of metabolites and reactions in the *i*ML1515 model makes extracting similar insights about the biochemical properties of the metabolites more difficult. Identifying distinct groups of metabolites similar to those shown in Figure 4.10d is no longer possible and the large number of edges that originate from highly connected nodes also makes it difficult to determine whether an edge originates from a nearby node. The scale-free organization of metabolic networks [74] is made more evident in the *i*ML1515 metabolic network visualizations and the positioning of hubs reflects how highly connected metabolites exhibit relatively strong polarity [74], as described in Section 2.3.2. Figure 4.15d is the most legible representation of the *i*ML1515 metabolic network because the use of ranked descriptors reduces some of the issues of planarity caused by chemical similarities between metabolites. The spatial arrangement of the highly connected nodes in Figure 4.15d improves the readability of edges, but identifying a motif or pattern that describes a pathway remains problematic. It would be interesting to explore whether a network half the size of *i*ML1515 and not as densely connected would be easier to read given the differences in overall legibility between the central metabolism and *i*ML1515 networks. It may be possible to reduce the amount of clustering and edge crossing by using a metabolic network with a lower level of detail, such as one containing only the main intermediate metabolites of pathways, to communicate the biological information that can be derived from their two-dimensional positions more effectively [84]. A particular property of the network layouts generated using molecular descriptors is not present in many other network layout algorithms. In typical layout algorithms, nodes are assigned positions in accordance with the algorithm, and the rendered length of the edge between two nodes does not communicate distinct information, it is a product of the spatial arrangement of nodes. An edge's rendered length in a generated layout indicates the distance between two metabolites in terms of chemical similarity. It is by this property that a metabolic network primarily comprised of the main compounds of pathways could be more intuitively understood when visualized with molecular descriptors as coordinates. It has been demonstrated in literature that as a linear metabolic pathway progresses toward a specific product, the intermediates become increasingly similar to the final product [183]. Exploiting this trend could allow for the recognition of network motifs that describe pathways but at the loss of information provided by the metabolites that drive biochemical reactions.

The metabolic pathway case studies indicate a potential gain in visual clarity when applying the concept of biochemical coordinates to a metabolic network with reduced complexity. As shown in case studies 4 and 7, CTP and histidine biosynthesis results demonstrated that visualizing the main intermediates of shorter pathways with steady incremental increases in chemical similarity to that of the end product was more intuitive than cyclical and longer pathways, as shown in cases 3, 5 and 6.

There is a significant difference in biochemical properties between metabolites that have bound coenzyme A (succinyl-CoA and acetyl-CoA) and those that do not. This makes it difficult to identify a cyclical network motif in the visualizations of case study 3. In order to avoid the sudden shift in chemical similarity associated with coenzyme A bound metabolites, it has been proposed to represent these cofactors as single atoms [162]. This could be implemented using the SMILES notation for "unspecified atomic number", " * ", for cofactors such as coenzyme A and acyl carrier protein (ACP). It is possible that a similar mindset is responsible for why the metabolites in Table 4.3 were represented by the symbol " * " in their SMILES. Although the rationale for including non-computable SMILES in a chemical property database is a bit unclear, it is most likely because these 8 metabolites have undefined structures, charges, and molecular weights in both MetaNetX and BiGG Models databases.

The inclusion of the IMP biosynthesis pathway in the purine biosynthesis datasets led to an observable and recurring network motif in most visualizations. Figure C.2 illustrates the structural similarity among intermediaries in the IMP biosynthesis pathway. This structural similarity is reflected in the short distance between the nodes that make up this recurring motif in most results for case studies 4 and 6. While the positioning of the nodes in this motif conveys that the metabolites they represent are chemically similar, which in and of itself is a motif, they are so tightly clustered that the relational information represented by edges is obscured. The layouts do however manage to capture and to a lesser degree, communicate, that the intermediaries between PRPP and the end products, GTP and ATP, exhibit different chemical profiles that are reflected in their positioning.

When viewed in a global context, the only pair of molecular descriptors that managed to arrange nodes far enough apart to allow for the identification of pathways and extract insight of the biochemical properties of their metabolites, when highlighted, were the ranked equivalents of ExactMolWt and MolLogP: mass_rank and MolLogP_rank. The Figures 4.18b, 4.23b, 4.33b, 4.28b, and 4.38b, illustrate that the change in molecular weight and hydrophobicity of intermediaries in de novo CTP biosynthesis increases and decreases respectively. A similar chemical pattern, but reversed, is observable for histidine biosynthesis, intermediaries decrease in molecular weight and become increasingly more hydrophobic as it approaches the terminal reaction in the pathway. The chemical similarity of the intermediaries in the *de novo* purine biosynthesis pathways still limits the readability of the pathway and the coenzyme A associated metabolites in the TCA cycle might does not create an aesthetically pleasing layout.

There is unfortunately no way of gaining such insight about the pathways from these figures unless they are highlighted. When observing the visualizations in the metabolic network case studies, it might be possible to make an educated guess as to whether or not a highly connected node is ATP or ADP based on the knowledge that they drive or is the result of biochemical reactions, their molecular composition, structure, and weight. As a result of their biochemical properties reflecting their position, that level of estimation is possible. An expert may be able to read the positioning of nodes intuitively and carefully trace the edges and map out pathways as if they were puzzles. This is not possible for a non-expert, and the layouts using molecular descriptors as coordinates do not sufficiently communicate or allow the identification of pathways or other biological patterns without substantial effort. Most of the information in the metabolic network layouts generated in this project is obscured behind a multitude of edges and closely clustered nodes. The layouts do, however, manage to position the metabolites according to their chemical properties, but they do not provide a good visual representation of a metabolic network. The spatial layout of the data, using the selected descriptors as coordinates, is not able to enable the viewer to gain an understanding of the degree of interactions between metabolites or gain an understanding of why the interactions are occurring as they do.

5.4 Export of biochemical coordinate layouts

Following a post-processing step, each layout generated for each of the 7 case studies was successfully exported to a Cytoscape compatible format. Interestingly enough, NetworkX's write to Cytoscape function writes a JSON file containing the information required to reconstruct the network in Cytoscape, but the data contained within the file is not compatible with JSON and will not read unless formatted correctly. The solution to the formatting issue, presented in Supplementary data 4, is tedious and can be replaced with an automated script invoked by a command line interface (CLI).

Using the write to Cytoscape function only stores the node positions of the exact layout rendered in NetworkX. An additional step was involved in adding molecular descriptor data and storing it in file formats that allowed for the storage of node attributes. The file format that required the least amount of handling to generate the molecular descriptor layout in Cytoscape was Cytoscape JS but the GraphML format showed the most promise as a vehicle to convert to systems biology standard formats such as SBGN-ML.

Chapter 6

Conclusion

The goal of this project was to replace arbitrary metabolic network layouts generated with layout algorithms with a layout where each metabolite had a well-defined location based on its biochemical properties. To achieve this, the project leveraged QSAR methods and developed a workflow that collected molecular representations for metabolites in two GEMs; the *E. coli* central metabolism and *i*ML1515, and calculated their molecular descriptors using the cheminformatic tool RDkit. From the calculated molecular descriptors, 9 descriptors were selected because they represent known and easily comprehensible biochemical properties that can be understood by a non-expert. In addition to the selected descriptors, 3 of the selected descriptors had their values ranked they exhibited a large number of overlapping descriptor values. These descriptors were used to generate network graph layouts for 7 case studies. These layouts were visualized and evaluated to determine which pair of molecular descriptors produced the most intuitive and meaningful network visualizations. The network layouts were successfully exported to a file format readable by the commonly used biological network visualization tool, Cytoscape.

The results demonstrated that metabolic network layouts derived from using molecular descriptors as 2D coordinates exhibit significant node clustering and edge crossings. The node clustering was a result of the chemical similarity among the metabolites in the GEMs used and the edge crossings were caused by the scale-free organization that is characteristic of metabolic networks. The number of edge crossings can be attributed to the fact that hubs in metabolic networks exhibited stronger polarity. As a direct consequence of using molecular descriptors as coordinates, hubs were positioned closer to each other which resulted in a large number of edge crossings from the position of hubs. This was especially prevalent in the visualization of the *i*ML1515 case study which was the largest and most complex case study in this project. These factors negatively affected the readability of the network visualizations and its ability to convey biological information. In spite of this, the generated layouts exhibit a unique characteristic not present in arbitrary network layouts. There is a direct correlation between the rendered length of the edge between nodes representing intermediate metabolites in a pathway and the change in the biochemical properties described by the descriptors used to create the layout. Unfortunately, extracting this information in a global context is difficult as it is obscured by node clustering and edge crossings. The pair of descriptors which generated the most meaningful visualizations were the descriptor pair mass_rank and MolLogP_rank. This pair of descriptors generated layouts that represented an improvement in legibility. Compared to layouts generated using the descriptors, ExactMolWt and MolLogP, the improved legibility can be attributed to the fact that each metabolite has a unique position within the layout due to the ranking of the descriptor values, leading to less node clustering.

The results from the metabolic pathway case studies suggest that a metabolic network of reduced complexity, such as a network comprised of the main intermediaries in metabolic pathways, is a better candidate for visualization using molecular descriptors as 2D coordinates.

This project examined if the biochemical properties of metabolites, quantified by molecular descriptors, could be used as two-dimensional coordinates for generating meaningful metabolic network layouts. As a result of the chemical similarity prevalent in most metabolites, and the scale-free organization of metabolic networks, the layout visualizations were incapable of communicating meaningful biological information.

Future work

Going to the third dimension: pathways in three-dimensions

When visualizing data in two dimensions, one must choose to compromise between presenting as much data in as large space as needed and losing varying levels of low-level detail or restricting the amount of information to present in order to produce a visualization that is easier to understand. There is potential in visualizing metabolic networks and metabolic pathways [184] in three dimensions as it can provide a compromise between an overview of the network and its modules (pathways) and the detail that is otherwise lost due to planarity while still providing a higher information density compared to two-dimensional visualizations. Rojdestvenski proposed in 2003 a method to visualize metabolic pathways in three dimensions using virtual reality technology, pointing out its potential as a tool for exploratory data analysis. A similar approach has been applied to large biological networks [185] for the purpose of exploratory analysis. Using the game development tool Unity [186], a VR model of a gene regulatory network was created that allowed users to manipulate the network visualization, and view network structures and node information in three dimensions [187]. In spite of this study's positive results among evaluators, the small sample size ($n = 7$) warrants further inquiry into its wider applicability.

A 3D animation and visualization tool such as Blender [188] is well suited for realizing this vision. Protein Database (.pdb) and XYZ (.xyz) file formats are natively supported in Blender, and Python can be used to automate tasks via scripting. For the purpose of validating a machine-learning algorithm trained to detect amoeboid gill disease in Atlantic salmon, the author developed a software tool in 2020 that generated 3D visualizations of fish school behavior over time in Blender [189]. A Python script was developed to automate this process by assigning each representation of fish a set of X, Y, and Z coordinates at each time interval. Modifications to this script could facilitate the import of metabolites in .pdb or .xyz format and assign each metabolite a set of X, Y, and Z coordinates, such as molecular descriptors. Through the use of Open Babel [134], which can convert InChI and SMILES to .pdf and .xyz formats, it could be possible to visualize the metabolic networks and pathways in 3D using the data generated in this project using a dedicated network visualization package for Blender, BlendNet [190]. Additionally, Blender can be used to generate VR models that can be viewed using VR headsets directly in Blender or exported to tools such as Unity [186] for the purpose of adding interactivity, as demonstrated in Ref [187].

The prospect of visualizing metabolic networks in three dimensions is an intriguing one, whether as a three dimensional, interactive VR model or presented as a plot in 2.5 dimensions [191] and using molecular descriptors to determine node positions. By doing so, some of the problems with planarity seen in the results of this project may be rectified.

Community standards

CytoScape2Escher [192], a tool made by Jeremy Zucker at the Pacific Northwest National Laboratory could be used to extend the generated layouts to another commonly used community standard, Escher [59]. This tool uses graphs of pathway visualizations generated from Pathway-tools [193] which are then exported and rendered in Cytoscape. Unfortunately, this tool does not work with the Cytoscape formatted layouts produced in this project. It is likely that modifying this tool to work with the layout format presented in this work could lead to the biochemical coordinate layout being encoded in additional community standard formats. The EscherConverter [59] tool could then be used to convert Escher compliant layouts to SBGN-ML and SBML which could build on the standardized graphical notation and styles encoded in these formats.

Force-directed algorithm applied to biochemical coordinates

The use of a force-directed algorithm [194] to untangle a network generated with biochemical coordinates could result in a graph that retains biochemical information to a certain extent but provides aesthetic layouts. In a force-directed algorithm, nodes and edges are given a repulsion force that distances them so their positions are in mechanical equilibrium. Adapting this algorithm to retain biochemical coordinates could lead to improved spatial separation between metabolites and remedy some issues of planarity as node stacking would be reduced by running this algorithm.

Bibliography

[1] Michal, *Roche Biochemical Pathways Wall Charts by Gerhard Michal*, Jan. 2014. DOI: 10.5281/zenodo.4446229. [Online]. Available: https://zenodo.org/record/4446229 (visited on 30/05/2022).

[2] *Roche | Biochemical Pathways*, en. [Online]. Available: https://www.roche.com/about/philanthropy/science-education/biochemical-pathways/ (visited on 01/06/2022).

[3] A. L. Delcher, D. Harmon, S. Kasif, O. White and S. L. Salzberg, 'Improved microbial gene identification with GLIMMER,' *Nucleic Acids Research*, vol. 27, no. 23, pp. 4636–4641, Dec. 1999, ISSN: 0305-1048. DOI: 10.1093/nar/27.23.4636. [Online]. Available: https://doi.org/10.1093/nar/27.23.4636 (visited on 09/07/2022).

[4] C. Ouzounis, G. Casari, C. Sander, J. Tamames and A. Valencia, 'Computational comparisons of model genomes,' eng, *Trends in Biotechnology*, vol. 14, no. 8, pp. 280–285, Aug. 1996, ISSN: 0167-7799. DOI: 10.1016/0167-7799(96)10043-3.

[5] J. D. Orth, R. M. T. Fleming and B. Ø. Palsson, 'Reconstruction and Use of Microbial Metabolic Networks: The Core Escherichia coli Metabolic Model as an Educational Guide,' *EcoSal Plus*, vol. 4, no. 1, Jan. 2010, Publisher: American Society for Microbiology, ISSN: 23246200. DOI: 10.1128/ECOSALPLUS.10.2.1/ASSET/AF2F042E-12AF-48A1-967D-9627A8D81014/ASSETS/GRAPHIC/10.2.1_FIG_022.GIF. [Online]. Available: https://journals.asm.org/doi/full/10.1128/ecosalplus.10.2.1 (visited on 20/04/2022).

[6] M. Kanehisa and S. Goto, 'KEGG: Kyoto Encyclopedia of Genes and Genomes,' *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000, ISSN: 0305-1048. DOI: 10.1093/nar/28.1.27. [Online]. Available: https://doi.org/10.1093/nar/28.1.27 (visited on 09/07/2022).

[7] P. D. Karp, M. Riley, S. M. Paley, A. Pellegrini-Toole and M. Krummenacker, 'EcoCyc: Encyclopedia of Escherichia coli genes and metabolism,' eng, *Nucleic Acids Research*, vol. 26, no. 1, pp. 50–53, Jan. 1998, ISSN: 0305-1048. DOI: 10.1093/nar/26.1.50.

[8] P. Hingamp, A. E. van den Broek, G. Stoesser and W. Baker, 'The EMBL Nucleotide Sequence Database. Contributing and accessing data,' eng, *Molecular Biotechnology*, vol. 12, no. 3, pp. 255–267, Oct. 1999, ISSN: 1073-6085. DOI: `10.1385/MB:12:3:255`.

[9] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers, 'GenBank,' eng, *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D36–42, Jan. 2013, ISSN: 1362-4962. DOI: `10.1093/nar/gks1195`.

[10] *Systems approach to refining genome annotation | PNAS*. [Online]. Available: `https://www.pnas.org/doi/abs/10.1073/pnas.0603364103` (visited on 09/07/2022).

[11] *Identification of Genome-Scale Metabolic Network Models Using Experimentally Measured Flux Profiles | PLOS Computational Biology*. [Online]. Available: `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0020072` (visited on 09/07/2022).

[12] *GrowMatch: An Automated Method for Reconciling In Silico/In Vivo Growth Predictions | PLOS Computational Biology*. [Online]. Available: `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000308` (visited on 09/07/2022).

[13] D. Hoksza, P. Gawron, M. Ostaszewski, J. Hasenauer and R. Schneider, 'Closing the gap between formats for storing layout information in systems biology,' *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1249–1260, Jul. 2020, Publisher: Oxford Academic, ISSN: 14774054. DOI: `10.1093/BIB/BBZ067`. [Online]. Available: `https://academic.oup.com/bib/article/21/4/1249/5527141` (visited on 11/03/2022).

[14] L. Röttjers and K. Faust, 'From hairballs to hypotheses–biological insights from microbial networks,' *FEMS Microbiology Reviews*, vol. 42, no. 6, pp. 761–780, Nov. 2018, ISSN: 0168-6445. DOI: `10.1093/femsre/fuy030`. [Online]. Available: `https://doi.org/10.1093/femsre/fuy030` (visited on 14/07/2022).

[15] Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson and N. E. Lewis, 'BiGG Models: A platform for integrating, standardizing and sharing genome-scale models,' *Nucleic Acids Research*, vol. 44, no. D1, pp. D515–D522, Jan. 2016, Publisher: Oxford Academic, ISSN: 0305-1048. DOI: `10.1093/NAR/GKV1049`. [Online]. Available: `https://academic.oup.com/nar/article/44/D1/D515/2502593` (visited on 20/04/2022).

[16] A. Funahashi, M. Morohashi, H. Kitano and N. Tanimura, 'CellDesigner: A process diagram editor for gene-regulatory and biochemical networks,' *Biosilico*, vol. 1, no. 5, pp. 159–162, 2003, Publisher: Elsevier.

[17] M. Shahlaei, 'Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study,' *Chemical Reviews*, vol. 113, no. 10, pp. 8093–8103, Oct. 2013, Publisher: American Chemical Society, ISSN: 0009-2665. DOI: `10.1021/cr3004339`. [Online]. Available: `https://doi.org/10.1021/cr3004339` (visited on 23/06/2022).

[18] B. Palsson, *Systems Biology*, en. Cambridge University Press, Jan. 2015, Google-Books-ID: XvxDBgAAQBAJ, ISBN: 978-1-107-03885-1.

[19] T. Ideker, T. Galitski and L. Hood, 'A new approach to decoding life: Systems biology,' eng, *Annual Review of Genomics and Human Genetics*, vol. 2, pp. 343–372, 2001, ISSN: 1527-8204. DOI: `10.1146/annurev.genom.2.1.343`.

[20] H.-Y. Chuang, M. Hofree and T. Ideker, 'A Decade of Systems Biology,' *Annual Review of Cell and Developmental Biology*, vol. 26, pp. 721–744, Nov. 2010, ISSN: 1081-0706. DOI: `10.1146/annurev-cellbio-100109-104122`. [Online]. Available: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3371392/` (visited on 09/07/2022).

[21] M. Hucka, a. t. r. o. t. S. Forum: A. Finney, a. t. r. o. t. S. Forum: H. M. Sauro, a. t. r. o. t. S. Forum: H. Bolouri, a. t. r. o. t. S. Forum: J. C. Doyle, a. t. r. o. t. S. Forum: H. Kitano, a. t. r. o. t. S. Forum: A. P. Arkin, a. t. r. o. t. S. Forum: B. J. Bornstein, a. t. r. o. t. S. Forum: D. Bray, a. t. r. o. t. S. Forum: A. Cornish-Bowden, a. t. r. o. t. S. Forum: A. A. Cuellar, a. t. r. o. t. S. Forum: S. Dronov, a. t. r. o. t. S. Forum: E. D. Gilles, a. t. r. o. t. S. Forum: M. Ginkel, a. t. r. o. t. S. Forum: V. Gor, a. t. r. o. t. S. Forum: I. I. Goryanin, a. t. r. o. t. S. Forum: W. J. Hedley, a. t. r. o. t. S. Forum: T. C. Hodgman, a. t. r. o. t. S. Forum: J.-H. Hofmeyr, a. t. r. o. t. S. Forum: P. J. Hunter, a. t. r. o. t. S. Forum: N. S. Juty, a. t. r. o. t. S. Forum: J. L. Kasberger, a. t. r. o. t. S. Forum: A. Kremling, a. t. r. o. t. S. Forum: U. Kummer, a. t. r. o. t. S. Forum: N. Le Novère, a. t. r. o. t. S. Forum: L. M. Loew, a. t. r. o. t. S. Forum: D. Lucio, a. t. r. o. t. S. Forum: P. Mendes, a. t. r. o. t. S. Forum: E. Minch, a. t. r. o. t. S. Forum: E. D. Mjolsness, a. t. r. o. t. S. Forum: Y. Nakayama, a. t. r. o. t. S. Forum: M. R. Nelson, a. t. r. o. t. S. Forum: P. F. Nielsen, a. t. r. o. t. S. Forum: T. Sakurada, a. t. r. o. t. S. Forum: J. C. Schaff, a. t. r. o. t. S. Forum: B. E. Shapiro, a. t. r. o. t. S. Forum: T. S. Shimizu, a. t. r. o. t. S. Forum: H. D. Spence, a. t. r. o. t. S. Forum: J. Stelling, a. t. r. o. t. S. Forum: K. Takahashi, a. t. r. o. t. S. Forum: M. Tomita, a. t. r. o. t. S. Forum: J. Wagner, a. t. r. o. t. S. Forum: J. Wang and a. t. r. o. t. S. Forum: 'The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models,' *Bioinformatics*, vol. 19, no. 4, pp. 524–531, Mar. 2003, Publisher: Oxford Academic, ISSN: 1367-4803. DOI: `10.1093/BIOINFORMATICS/BTG015`. [Online]. Available: `https://academic.oup.com/bioinformatics/article/19/4/524/218599` (visited on 06/04/2022).

[22] A. Finney, *Internal Discussion Document Possible extensions to the Systems Biology Markup Language*, English, Nov. 2000. [Online]. Available: `https://drive.google.com/open?id=120NQsFjhAoGnp-tNJwFOq5aXIG58_a21` (visited on 05/06/2022).

[23] R. Gauges, U. Rost, S. Sahle and K. Wegner, 'A model diagram layout extension for SBML,' *Bioinformatics*, vol. 22, no. 15, pp. 1879–1885, Aug. 2006, ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btl195`. [Online]. Available: `https://doi.org/10.1093/bioinformatics/btl195` (visited on 12/06/2022).

[24] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi and I. Pletnev, 'InChI - the worldwide chemical structure identifier standard,' *J Cheminformatics*, vol. 5, no. 1, p. 7, Jan. 2013, ISSN: 17582946. DOI: `10.1186/1758-2946-5-7`. (visited on 14/04/2022).

[25] D. Weininger, 'SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules,' *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, Feb. 1988, ISSN: 00952338. DOI: `10.1021/CI00057A005`. (visited on 14/04/2022).

[26] N. B. A Bender, 'Special issue: Cheminformatics in drug discovery,' *ChemMedChem*, vol. 13, no. 6, pp. 467–469, Mar. 2018, Publisher: John Wiley and Sons Ltd. DOI: `10.1002/cmdc.201800123`. (visited on 03/11/2021).

[27] X. Martinez, M. Chavent and M. Baaden, 'Visualizing protein structures — tools and trends,' *Biochemical Society Transactions*, vol. 48, no. 2, pp. 499–506, Mar. 2020, ISSN: 0300-5127. DOI: `10.1042/BST20190621`. [Online]. Available: `https://doi.org/10.1042/BST20190621` (visited on 30/06/2022).

[28] A. Mauri, V. Consonni and R. Todeschini, 'Molecular Descriptors,' *Handbook of Computational Chemistry*, pp. 2065–2093, Jan. 2017, Publisher: Springer, Cham. DOI: `10.1007/978-3-319-27282-5_51`. [Online]. Available: `https://link.springer.com/referenceworkentry/10.1007/978-3-319-27282-5_51` (visited on 01/10/2021).

[29] E. J. O'Brien, J. M. Monk and B. O. Palsson, 'Using genome-scale models to predict biological capabilities,' *Cell*, vol. 161, no. 5, pp. 971–987, May 2015, Publisher: Cell Press, ISSN: 10974172. DOI: `10.1016/j.cell.2015.05.019`. (visited on 20/04/2022).

[30] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau and Y. Shao, 'The complete genome sequence of Escherichia coli K-12,' eng, *Science (New York, N.Y.)*, vol. 277, no. 5331, pp. 1453–1462, Sep. 1997, ISSN: 0036-8075. DOI: `10.1126/science.277.5331.1453`.

[31] J. S. Edwards and B. O. Palsson, 'The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities,' eng, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 10, pp. 5528–5533, May 2000, ISSN: 0027-8424. DOI: 10.1073/pnas.97.10.5528.

[32] M. Scheer, A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Söhngen, M. Stelzer, J. Thiele and D. Schomburg, 'BRENDA, the enzyme information system in 2011,' *Nucleic Acids Research*, vol. 39, no. suppl_1, pp. D670–D676, Jan. 2011, ISSN: 0305-1048. DOI: 10.1093/nar/gkq1089. [Online]. Available: https://doi.org/10.1093/nar/gkq1089 (visited on 12/07/2022).

[33] D. R. Kelley, B. Liu, A. L. Delcher, M. Pop and S. L. Salzberg, 'Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering,' *Nucleic Acids Research*, vol. 40, no. 1, e9, Jan. 2012, ISSN: 0305-1048. DOI: 10.1093/nar/gkr1067. [Online]. Available: https://doi.org/10.1093/nar/gkr1067 (visited on 11/07/2022).

[34] J. L. Reed, I. Famili, I. Thiele and B. O. Palsson, 'Towards multidimensional genome annotation,' en, *Nature Reviews Genetics*, vol. 7, no. 2, pp. 130–141, Feb. 2006, Number: 2 Publisher: Nature Publishing Group, ISSN: 1471-0064. DOI: 10.1038/nrg1769. [Online]. Available: https://www.nature.com/articles/nrg1769 (visited on 12/07/2022).

[35] L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, H. S. Haraldsdóttir, J. Wachowiak, S. M. Keating, V. Vlasov, S. Magnusdóttir, C. Y. Ng, G. Preciat, A. Žagare, S. H. J. Chan, M. K. Aurich, C. M. Clancy, J. Modamio, J. T. Sauls, A. Noronha, A. Bordbar, B. Cousins, D. C. El Assal, L. V. Valcarcel, I. Apaolaza, S. Ghaderi, M. Ahookhosh, M. Ben Guebila, A. Kostromins, N. Sompairac, H. M. Le, D. Ma, Y. Sun, L. Wang, J. T. Yurkovich, M. A. P. Oliveira, P. T. Vuong, L. P. El Assal, I. Kuperstein, A. Zinovyev, H. S. Hinton, W. A. Bryant, F. J. Aragón Artacho, F. J. Planes, E. Stalidzans, A. Maass, S. Vempala, M. Hucka, M. A. Saunders, C. D. Maranas, N. E. Lewis, T. Sauter, B. Ø. Palsson, I. Thiele and R. M. T. Fleming, 'Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0,' en, *Nature Protocols*, vol. 14, no. 3, pp. 639–702, Mar. 2019, Number: 3 Publisher: Nature Publishing Group, ISSN: 1750-2799. DOI: 10.1038/s41596-018-0098-2. [Online]. Available: https://www.nature.com/articles/s41596-018-0098-2 (visited on 12/07/2022).

[36] A. Ebrahim, J. A. Lerman, B. O. Palsson and D. R. Hyduke, 'COBRApy: COnstraints-Based Reconstruction and Analysis for Python,' en, *BMC Systems Biology*, vol. 7, no. 1, p. 74, Aug. 2013, ISSN: 1752-0509. DOI: 10.1186/1752-0509-7-74. [Online]. Available: https://doi.org/10.1186/1752-0509-7-74 (visited on 10/06/2022).

[37] R. Breitling, D. Vitkup and M. P. Barrett, 'New surveyor tools for charting microbial metabolic maps,' en, *Nature Reviews Microbiology*, vol. 6, no. 2, pp. 156–161, Feb. 2008, Number: 2 Publisher: Nature Publishing Group, ISSN: 1740-1534. DOI: `10.1038/nrmicro1797`. [Online]. Available: `https://www.nature.com/articles/nrmicro1797` (visited on 12/07/2022).

[38] I. Thiele, N. Vlassis and R. M. T. Fleming, 'fastGapFill: Efficient gap filling in metabolic networks,' *Bioinformatics*, vol. 30, no. 17, pp. 2529–2531, Sep. 2014, ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btu321`. [Online]. Available: `https://doi.org/10.1093/bioinformatics/btu321` (visited on 13/07/2022).

[39] *Discovering missing reactions of metabolic networks by using gene co-expression data | Scientific Reports*. [Online]. Available: `https://www.nature.com/articles/srep41774` (visited on 13/07/2022).

[40] S. Pan and J. L. Reed, 'Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries,' en, *Current Opinion in Biotechnology*, Systems biology • Nanobiotechnology, vol. 51, pp. 103–108, Jun. 2018, ISSN: 0958-1669. DOI: `10.1016/j.copbio.2017.12.012`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0958166917302045` (visited on 13/07/2022).

[41] 'Genome-scale metabolic modeling reveals SARS-CoV-2-induced metabolic changes and antiviral targets,' *Molecular Systems Biology*, vol. 17, no. 11, e10260, Nov. 2021, Publisher: John Wiley & Sons, Ltd, ISSN: 1744-4292. DOI: `10.15252/msb.202110260`. [Online]. Available: `https://www.embopress.org/doi/full/10.15252/msb.202110260` (visited on 12/07/2022).

[42] *Dynamic metabolic control: Towards precision engineering of metabolism | Journal of Industrial Microbiology and Biotechnology | Oxford Academic*. [Online]. Available: `https://academic.oup.com/jimb/article/45/7/535/5996684?login=true` (visited on 22/06/2022).

[43] C. Zhang and Q. Hua, 'Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine,' *Frontiers in Physiology*, vol. 6, 2016, ISSN: 1664-042X. [Online]. Available: `https://www.frontiersin.org/article/10.3389/fphys.2015.00413` (visited on 09/06/2022).

[44] *Team:Virginia/model - 2019.igem.org*. [Online]. Available: `https://2019.igem.org/Team:Virginia/model` (visited on 12/07/2022).

[45] E. Brunk, S. Sahoo, D. C. Zielinski, A. Altunkaya, A. Dräger, N. Mih, F. Gatto, A. Nilsson, G. A. Preciat Gonzalez, M. K. Aurich, A. Prlić, A. Sastry, A. D. Danielsdottir, A. Heinken, A. Noronha, P. W. Rose, S. K. Burley, R. M. T. Fleming, J. Nielsen, I. Thiele and B. O. Palsson, 'Recon3D enables a three-dimensional view of gene variation in human metabolism,' eng, *Nature Biotechnology*, vol. 36, no. 3, pp. 272–281, Mar. 2018, ISSN: 1546-1696. DOI: `10.1038/nbt.4072`.

[46] T. Shlomi, M. N. Cabili, M. J. Herrgård, B. Ø. Palsson and E. Ruppin, 'Network-based prediction of human tissue-specific metabolism,' en, *Nature Biotechnology*, vol. 26, no. 9, pp. 1003–1010, Sep. 2008, Number: 9 Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/nbt.1487. [Online]. Available: https://www.nature.com/articles/nbt.1487 (visited on 12/07/2022).

[47] J. M. Monk, C. J. Lloyd, E. Brunk, N. Mih, A. Sastry, Z. King, R. Takeuchi, W. Nomura, Z. Zhang, H. Mori, A. M. Feist and B. O. Palsson, 'iML1515, a knowledgebase that computes Escherichia coli traits,' *Nature biotechnology*, vol. 35, no. 10, p. 904, Oct. 2017, Publisher: NIH Public Access, ISSN: 15461696. DOI: 10.1038/NBT.3956. [Online]. Available: /pmc/articles/PMC6521705/ (visited on 09/02/2022).

[48] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist and B. Palsson, 'A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011,' *Molecular systems biology*, vol. 7, 2011, Publisher: Mol Syst Biol, ISSN: 1744-4292. DOI: 10.1038/MSB.2011.65. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21988831/ (visited on 26/04/2022).

[49] C. Gu, G. B. Kim, W. J. Kim, H. U. Kim and S. Y. Lee, 'Current status and applications of genome-scale metabolic models,' en, *Genome Biology*, vol. 20, no. 1, p. 121, Jun. 2019, ISSN: 1474-760X. DOI: 10.1186/s13059-019-1730-3. [Online]. Available: https://doi.org/10.1186/s13059-019-1730-3 (visited on 07/07/2022).

[50] L. J. Freischem, M. Barahona and D. A. Oyarzún, *Prediction of gene essentiality using machine learning and genome-scale metabolic models*, en, Pages: 2022.03.31.486520 Section: New Results, Mar. 2022. DOI: 10.1101/2022.03.31.486520. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.03.31.486520v1 (visited on 12/07/2022).

[51] C. Ye, Q. Luo, L. Guo, C. Gao, N. Xu, L. Zhang, L. Liu and X. Chen, 'Improving lysine production through construction of an Escherichia coli enzyme-constrained model,' en, *Biotechnology and Bioengineering*, vol. 117, no. 11, pp. 3533–3544, 2020, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bit.27485, ISSN: 1097-0290. DOI: 10.1002/bit.27485. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.27485 (visited on 12/07/2022).

[52] M. Hucka, F. T. Bergmann, C. Chaouiya, A. Dräger, S. Hoops, S. M. Keating, M. König, N. L. Novère, C. J. Myers, B. G. Olivier, S. Sahle, J. C. Schaff, R. Sheriff, L. P. Smith, D. Waltemath, D. J. Wilkinson and F. Zhang, 'The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core Release 2,' *Journal of integrative bioinformatics*, vol. 16, no. 2, Jun. 2019, Publisher: NLM (Medline), ISSN: 16134516. DOI: 10.1515/JIB-2019-0021/MACHINEREADABLECITATION/RIS. [Online].

Available: https://www.degruyter.com/document/doi/10.1515/jib-2019-0021/html (visited on 24/05/2022).

[53] N. Juty, N. Le Novère, H. Hermjakob and C. Laibe, 'Towards the Collaborative Curation of the Registry underlying identifiers.org,' *Database*, vol. 2013, bat017, Jan. 2013, ISSN: 1758-0463. DOI: 10.1093/database/bat017. [Online]. Available: https://doi.org/10.1093/database/bat017 (visited on 09/06/2022).

[54] N. L. Novère, A. Finney, M. Hucka, U. S. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J. L. Snoep, H. D. Spence and B. L. Wanner, 'Minimum information requested in the annotation of biochemical models (MIRIAM),' en, *Nature Biotechnology*, vol. 23, no. 12, pp. 1509–1515, Dec. 2005, Number: 12 Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/nbt1156. [Online]. Available: https://www.nature.com/articles/nbt1156 (visited on 12/07/2022).

[55] R. Gauges, U. Rost, S. Sahle, K. Wengler and F. T. Bergmann, 'The Systems Biology Markup Language (SBML) Level 3 Package: Layout, Version 1 Core,' en, *Journal of Integrative Bioinformatics*, vol. 12, no. 2, pp. 550–602, Jun. 2015, Publisher: De Gruyter, ISSN: 1613-4516. DOI: 10.1515/jib-2015-267. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/jib-2015-267/html (visited on 12/07/2022).

[56] F. T. Bergmann, S. M. Keating, R. Gauges, S. Sahle and K. Wengler, 'SBML Level 3 package: Render, Version 1, Release 1,' en, *Journal of Integrative Bioinformatics*, vol. 15, no. 1, Mar. 2018, Publisher: De Gruyter, ISSN: 1613-4516. DOI: 10.1515/jib-2017-0078. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/jib-2017-0078/html (visited on 12/07/2022).

[57] J. Monk, J. Nogales and B. O. Palsson, 'Optimizing genome-scale network reconstructions,' en, *Nature Biotechnology*, vol. 32, no. 5, pp. 447–452, May 2014, Number: 5 Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/nbt.2870. [Online]. Available: https://www.nature.com/articles/nbt.2870 (visited on 12/07/2022).

[58] C. Lieven, M. E. Beber, B. G. Olivier, F. T. Bergmann, M. Ataman, P. Babaei, J. A. Bartell, L. M. Blank, S. Chauhan, K. Correia, C. Diener, A. Dräger, B. E. Ebert, J. N. Edirisinghe, J. P. Faria, A. Feist, G. Fengos, R. M. T. Fleming, B. García-Jiménez, V. Hatzimanikatis, W. v. Helvoirt, C. S. Henry, H. Hermjakob, M. J. Herrgård, H. U. Kim, Z. King, J. J. Koehorst, S. Klamt, E. Klipp, M. Lakshmanan, N. L. Novère, D.-Y. Lee, S. Y. Lee, S. Lee, N. E. Lewis, H. Ma, D. Machado, R. Mahadevan, P. Maia, A. Mardinoglu, G. L. Medlock, J. M. Monk, J. Nielsen, L. K. Nielsen, J. Nogales, I. Nookaew, O. Resendis-Antonio, B. O. Palsson, J. A. Papin, K. R. Patil, M. Poolman, N. D. Price, A. Richelle, I. Rocha, B. J. Sanchez, P. J. Schaap, R. S. M.

Sheriff, S. Shoaie, N. Sonnenschein, B. Teusink, P. Vilaça, J. O. Vik, J. A. Wodke, J. C. Xavier, Q. Yuan, M. Zakhartsev and C. Zhang, *Memote: A community driven effort towards a standardized genome-scale metabolic model test suite*, en, Pages: 350991 Section: New Results, Jul. 2018. DOI: 10.1101/350991. [Online]. Available: https://www.biorxiv.org/content/10.1101/350991v3 (visited on 13/07/2022).

[59] Z. A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N. E. Lewis and B. O. Palsson, 'Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways,' *PLOS Computational Biology*, vol. 11, no. 8, P. P. Gardner, Ed., e1004321, Aug. 2015, Publisher: Public Library of Science, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004321. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1004321 (visited on 24/08/2021).

[60] S. Moretti, V. D. T. Tran, F. Mehl, M. Ibberson and M. Pagni, 'MetaNetX/MNXref: Unified namespace for metabolites and biochemical reactions in the context of metabolic models,' *Nucleic Acids Research*, vol. 49, no. D1, pp. D570–D574, Jan. 2021, ISSN: 0305-1048. DOI: 10.1093/nar/gkaa992. [Online]. Available: https://doi.org/10.1093/nar/gkaa992 (visited on 15/07/2022).

[61] V. Lacroix, L. Cottret, P. Thébault and M. F. Sagot, 'An introduction to metabolic networks and their structural analysis,' *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 4, pp. 594–617, Oct. 2008, ISSN: 15455963. DOI: 10.1109/TCBB.2008.79. (visited on 13/04/2022).

[62] *MetaCyc inosine-5'-phosphate biosynthesis I*. [Online]. Available: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6123&detail-level=4 (visited on 02/08/2022).

[63] *MetaCyc inosine-5'-phosphate biosynthesis I*. [Online]. Available: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6123&detail-level=2 (visited on 02/08/2022).

[64] *MetaCyc EC 4.3.2.10*. [Online]. Available: https://biocyc.org/META/NEW-IMAGE?type=REACTION&object=GLUTAMIDOTRANS-RXN (visited on 19/07/2022).

[65] *MetaCyc inosine-5'-phosphate biosynthesis I*. [Online]. Available: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6123&detail-level=4 (visited on 02/08/2022).

[66] A. L. Lehninger, D. L. Nelson, M. M. Cox, M. M. Cox *et al.*, *Lehninger principles of biochemistry*. Macmillan, 2005.

[67] R. K. Kulis-Horn, M. Persicke and J. Kalinowski, ‘Histidine biosynthesis, its regulation and biotechnological application in Corynebacterium glutamicum,’ en, *Microbial Biotechnology*, vol. 7, no. 1, pp. 5–25, 2014, _eprint: https://onlinelibrary.wiley.com/d
7915.12055, ISSN: 1751-7915. DOI: `10.1111/1751-7915.12055`. [Online]. Available: `https://onlinelibrary.wiley.com/doi/abs/10.1111/1751-7915.12055` (visited on 14/07/2022).

[68] *MetaCyc L-histidine biosynthesis*. [Online]. Available: `https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=HISTSYN-PWY&detail-level=1` (visited on 05/07/2022).

[69] A. Vellido, ‘The importance of interpretability and visualization in machine learning for applications in medicine and health care,’ en, *Neural Computing and Applications*, vol. 32, no. 24, pp. 18 069–18 083, Dec. 2020, ISSN: 0941-0643, 1433-3058. DOI: `10.1007/s00521-019-04051-w`. [Online]. Available: `http://link.springer.com/10.1007/s00521-019-04051-w` (visited on 10/06/2022).

[70] A. Schultz and R. Akbani, ‘SAMMI: A semi-automated tool for the visualization of metabolic networks,’ *Bioinformatics*, vol. 36, no. 8, pp. 2616–2617, Apr. 2020, ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btz927`. [Online]. Available: `https://doi.org/10.1093/bioinformatics/btz927` (visited on 13/07/2022).

[71] D. Machado, R. S. Costa, M. Rocha, E. C. Ferreira, B. Tidor and I. Rocha, ‘Modeling formalisms in Systems Biology,’ *AMB Express*, vol. 1, no. 1, p. 45, Dec. 2011, ISSN: 2191-0855. DOI: `10.1186/2191-0855-1-45`. [Online]. Available: `https://doi.org/10.1186/2191-0855-1-45` (visited on 13/07/2022).

[72] *Tools for visualization and analysis of molecular networks, pathways, and -omics data - PMC*. [Online]. Available: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4461095/` (visited on 13/07/2022).

[73] S. He, Y. J. Liu, F. Y. Ye, R. P. Li and R. J. Dai, ‘A new grid- And modularity-based layout algorithm for complex biological networks,’ *PLoS ONE*, vol. 14, no. 8, Aug. 2019, Publisher: Public Library of Science, ISSN: 19326203. DOI: `10.1371/journal.pone.0221620`. (visited on 19/05/2022).

[74] Q. Zhu, T. Qin, Y.-Y. Jiang, C. Ji, D.-X. Kong, B.-G. Ma and H.-Y. Zhang, ‘Chemical Basis of Metabolic Network Organization,’ en, *PLOS Computational Biology*, vol. 7, no. 10, e1002214, Oct. 2011, Publisher: Public Library of Science, ISSN: 1553-7358. DOI: `10.1371/journal.pcbi.1002214`. [Online]. Available: `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002214` (visited on 22/07/2022).

[75] B. D. Bennett, E. H. Kimball, M. Gao, R. Osterhout, S. J. Van Dien and J. D. Rabinowitz, ‘Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli,’ en, *Nature Chemical Biology*, vol. 5, no. 8, pp. 593–599, Aug. 2009, Number: 8 Publisher: Nature Publishing

Group, ISSN: 1552-4469. DOI: 10.1038/nchembio.186. [Online]. Available: https://www.nature.com/articles/nchembio.186 (visited on 22/07/2022).

[76] A. Küken, D. Langary and Z. Nikoloski, 'The hidden simplicity of metabolic networks is revealed by multireaction dependencies,' *Science Advances*, vol. 8, no. 13, eabl6962, Mar. 2022, Publisher: American Association for the Advancement of Science. DOI: 10.1126/sciadv.abl6962. [Online]. Available: https://www.science.org/doi/full/10.1126/sciadv.abl6962 (visited on 22/07/2022).

[77] H.-F. Yang, X.-N. Zhang, Y. Li, Y.-H. Zhang, Q. Xu and D.-Q. Wei, 'Theoretical Studies of Intracellular Concentration of Micro-organisms' Metabolites,' en, *Scientific Reports*, vol. 7, no. 1, p. 9048, Aug. 2017, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-017-08793-2. [Online]. Available: https://www.nature.com/articles/s41598-017-08793-2 (visited on 22/07/2022).

[78] E. Almaas, 'Biological impacts and context of network theory,' *Journal of Experimental Biology*, vol. 210, no. 9, pp. 1548–1558, May 2007, ISSN: 0022-0949. DOI: 10.1242/jeb.003731. [Online]. Available: https://doi.org/10.1242/jeb.003731 (visited on 16/07/2022).

[79] D. Merico, D. Gfeller and G. D. Bader, 'How to visually interpret biological data using networks,' en, *Nature Biotechnology*, vol. 27, no. 10, pp. 921–924, Oct. 2009, Number: 10 Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/nbt.1567. [Online]. Available: https://www.nature.com/articles/nbt.1567 (visited on 16/07/2022).

[80] *Protein-Protein Interactions*. [Online]. Available: https://cytoscape.org/cytoscape-tutorials/presentations/ppi-tools1-2017-mpi.html#/9/1 (visited on 17/07/2022).

[81] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, 'Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks,' *Genome Research*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, ISSN: 1088-9051. DOI: 10.1101/gr.1239303. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403769/ (visited on 07/07/2022).

[82] M. Koutrouli, E. Karatzas, D. Paez-Espino and G. A. Pavlopoulos, 'A Guide to Conquer the Biological Network Era Using Graph Theory,' *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020, ISSN: 2296-4185. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00034 (visited on 22/07/2022).

[83] M. Chazalviel, C. Frainay, N. Poupin, F. Vinson, B. Merlet, Y. Gloaguen, L. Cottret and F. Jourdan, 'MetExploreViz: Web component for interactive metabolic network visualization,' *Bioinformatics*, vol. 34, no. 2, pp. 312–313, Jan. 2018, Publisher: Oxford Academic, ISSN: 1367-4803. DOI: 10.

1093/BIOINFORMATICS/BTX588. [Online]. Available: https://academic.oup.com/bioinformatics/article/34/2/312/4158790 (visited on 24/08/2021).

[84] G. E. Marai, B. Pinaud, K. Bühler, A. Lex and J. H. Morris, 'Ten simple rules to create biological network figures for communication,' en, *PLOS Computational Biology*, vol. 15, no. 9, e1007244, Sep. 2019, Publisher: Public Library of Science, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007244. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007244 (visited on 22/07/2022).

[85] A. Deckard, F. T. Bergmann and H. M. Sauro, 'Supporting the SBML layout extension,' *Bioinformatics*, vol. 22, no. 23, pp. 2966–2967, Dec. 2006, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl520. [Online]. Available: https://doi.org/10.1093/bioinformatics/btl520 (visited on 13/07/2022).

[86] H. M. Sauro, M. Hucka, A. Finney, C. Wellock, H. Bolouri, J. Doyle and H. Kitano, 'Next Generation Simulation Tools: The Systems Biology Workbench and BioSPICE Integration,' *OMICS: A Journal of Integrative Biology*, vol. 7, no. 4, pp. 355–372, Dec. 2003, Publisher: Mary Ann Liebert, Inc., publishers. DOI: 10.1089/153623103322637670. [Online]. Available: https://www.liebertpub.com/doi/abs/10.1089/153623103322637670 (visited on 13/07/2022).

[87] F. T. Bergmann, T. Czauderna, U. Dogrusoz, A. Rougny, A. Dräger, V. Touré, A. Mazein, M. L. Blinov and A. Luna, 'Systems biology graphical notation markup language (SBGNML) version 0.3,' en, *Journal of Integrative Bioinformatics*, vol. 17, no. 2-3, Jun. 2020, Publisher: De Gruyter, ISSN: 1613-4516. DOI: 10.1515/jib-2020-0016. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/jib-2020-0016/html (visited on 12/06/2022).

[88] M. P. van Iersel, A. C. Villéger, T. Czauderna, S. E. Boyd, F. T. Bergmann, A. Luna, E. Demir, A. Sorokin, U. Dogrusoz, Y. Matsuoka, A. Funahashi, M. I. Aladjem, H. Mi, S. L. Moodie, H. Kitano, N. Le Novère and F. Schreiber, 'Software support for SBGN maps: SBGN-ML and LibSBGN,' *Bioinformatics*, vol. 28, no. 15, pp. 2016–2021, Aug. 2012, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts270. [Online]. Available: https://doi.org/10.1093/bioinformatics/bts270 (visited on 22/06/2022).

[89] N. Juty, 'Systems Biology Ontology: Update,' en, *Nature Precedings*, pp. 1–1, Oct. 2010, Publisher: Nature Publishing Group, ISSN: 1756-0357. DOI: 10.1038/npre.2010.5121.1. [Online]. Available: https://www.nature.com/articles/npre.2010.5121.1 (visited on 30/07/2022).

[90] J. J. Kelley, S. Maor, M. K. Kim, A. Lane and D. S. Lun, 'MOST-visualization: Software for producing automated textbook-style maps of genome-scale metabolic networks,' *Bioinformatics*, vol. 33, no. 16, pp. 2596–2597, Aug.

2017, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx240. [Online]. Available: https://doi.org/10.1093/bioinformatics/btx240 (visited on 13/07/2022).

[91] *Visualizing genome and systems biology: Technologies, tools, implementation techniques and trends, past, present and future | GigaScience | Oxford Academic*. [Online]. Available: https://academic.oup.com/gigascience/article/4/1/s13742-015-0077-2/2707594?login=true (visited on 13/07/2022).

[92] G. Michal and D. Schomburg, *Biochemical pathways: an atlas of biochemistry and molecular biology*. Wiley, 2013.

[93] T. Wittkop, J. Baumbach, F. P. Lobo and S. Rahmann, 'Large scale clustering of protein sequences with FORCE -A layout based heuristic for weighted cluster editing,' *BMC Bioinformatics*, vol. 8, no. 1, p. 396, Oct. 2007, ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-396. [Online]. Available: https://doi.org/10.1186/1471-2105-8-396 (visited on 13/07/2022).

[94] R. Jianu, A. Rusu, A. J. Fabian and D. H. Laidlaw, 'A Coloring Solution to the Edge Crossing Problem,' in *2009 13th International Conference Information Visualisation*, ISSN: 2375-0138, Jul. 2009, pp. 691–696. DOI: 10.1109/IV.2009.66.

[95] W. J. Longabaugh, 'Combing the hairball with BioFabric: A new approach for visualization of large networks,' en, *BMC Bioinformatics*, vol. 13, no. 1, p. 275, Dec. 2012, ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-275. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-275 (visited on 30/07/2022).

[96] C. Nobre, M. Meyer, M. Streit and A. Lex, 'The State of the Art in Visualizing Multivariate Networks,' en, *Computer Graphics Forum*, vol. 38, no. 3, pp. 807–832, 2019, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13728, ISSN: 1467-8659. DOI: 10.1111/cgf.13728. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13728 (visited on 13/07/2022).

[97] D. Otasek, J. H. Morris, J. Bouças, A. R. Pico and B. Demchak, 'Cytoscape Automation: Empowering workflow-based network analysis,' en, *Genome Biology*, vol. 20, no. 1, p. 185, Sep. 2019, ISSN: 1474-760X. DOI: 10.1186/s13059-019-1758-4. [Online]. Available: https://doi.org/10.1186/s13059-019-1758-4 (visited on 13/07/2022).

[98] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang and T. Ideker, 'Cytoscape 2.8: New features for data integration and network visualization,' *Bioinformatics*, vol. 27, no. 3, pp. 431–432, Feb. 2011, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq675. [Online]. Available: https://doi.org/10.1093/bioinformatics/btq675 (visited on 13/07/2022).

[99] C. Gaiteri, Y. Ding, B. French, G. C. Tseng and E. Sibille, 'Beyond modules and hubs: The potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders,' en, *Genes, Brain and Behavior*, vol. 13, no. 1, pp. 13–24, 2014, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111 ISSN: 1601-183X. DOI: 10.1111/gbb.12106. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/gbb.12106 (visited on 13/07/2022).

[100] R. Saito, M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, S. Lotia, A. R. Pico, G. D. Bader and T. Ideker, 'A travel guide to Cytoscape plugins,' en, *Nature Methods*, vol. 9, no. 11, pp. 1069–1076, Nov. 2012, Number: 11 Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/nmeth.2212. [Online]. Available: https://www.nature.com/articles/nmeth.2212 (visited on 13/07/2022).

[101] M. König, A. Dräger and H.-G. Holzhütter, 'CySBML: A Cytoscape plugin for SBML,' *Bioinformatics*, vol. 28, no. 18, pp. 2402–2403, Sep. 2012, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts432. [Online]. Available: https://doi.org/10.1093/bioinformatics/bts432 (visited on 13/07/2022).

[102] E. Gonçalves, M. v. Iersel and J. Saez-Rodriguez, 'CySBGN: A Cytoscape plug-in to integrate SBGN maps,' *BMC Bioinformatics*, vol. 14, no. 1, p. 17, Jan. 2013, ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-17. [Online]. Available: https://doi.org/10.1186/1471-2105-14-17 (visited on 13/07/2022).

[103] *MetaNetX: Mnxref*, en. [Online]. Available: https://www.metanetx.org/mnxdoc/mnxref.html (visited on 15/07/2022).

[104] *ChEBI: A database and ontology for chemical entities of biological interest | Nucleic Acids Research | Oxford Academic*. [Online]. Available: https://academic.oup.com/nar/article/36/suppl_1/D344/2506390?login=true (visited on 12/07/2022).

[105] R. Caspi, R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver and P. D. Karp, 'The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases,' *Nucleic Acids Research*, vol. 44, no. D1, pp. D471–D480, 2016, Publisher: Oxford University Press. DOI: 10.1093/NAR/GKV1164. (visited on 24/08/2021).

[106] Y. Matsuoka, A. Funahashi, S. Ghosh and H. Kitano, 'Modeling and Simulation Using CellDesigner,' en, in *Transcription Factor Regulatory Networks: Methods and Protocols*, ser. Methods in Molecular Biology, E. Miyamoto-Sato, H. Ohashi, H. Sasaki, J.-i. Nishikawa and H. Yanagawa, Eds., New York, NY: Springer, 2014, pp. 121–145, ISBN: 978-1-4939-0805-9. DOI:

10.1007/978-1-4939-0805-9_11. [Online]. Available: https://doi.org/10.1007/978-1-4939-0805-9_11 (visited on 13/07/2022).

[107] M. Hucka, F. T. Bergmann, S. Hoops, S. M. Keating, S. Sahle, J. C. Schaff, L. P. Smith and D. J. Wilkinson, 'The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core,' *Journal of Integrative Bioinformatics*, vol. 12, no. 2, pp. 382–549, Jun. 2015, Publisher: De Gruyter, ISSN: 1613-4516. DOI: 10.1515/JIB-2015-266. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/jib-2015-266/html (visited on 24/05/2022).

[108] A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi and H. Kitano, 'CellDesigner 3.5: A versatile modeling tool for biochemical networks,' *Proceedings of the IEEE*, vol. 96, no. 8, pp. 1254–1265, 2008, Publisher: Institute of Electrical and Electronics Engineers Inc., ISSN: 00189219. DOI: 10.1109/JPROC.2008.925458. (visited on 26/04/2022).

[109] A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, en. Springer, Sep. 2007, Google-Books-ID: 4z7Q87HgBdwC, ISBN: 978-1-4020-6291-9.

[110] A. Mauri, V. Consonni, M. Pavan and R. Todeschini, 'Dragon software: An easy approach to molecular descriptor calculations,' *Match*, vol. 56, no. 2, pp. 237–248, 2006.

[111] A. Helguera, R. Combes, M. Gonzalez and M. N. Cordeiro, 'Applications of 2D Descriptors in Drug Design: A DRAGON Tale,' en, *Current Topics in Medicinal Chemistry*, vol. 8, no. 18, pp. 1628–1655, Dec. 2008, ISSN: 15680266. DOI: 10.2174/156802608786786598. [Online]. Available: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1568-0266&volume=8&issue=18&spage=1628 (visited on 20/06/2022).

[112] C. H. T. d. P. da Silva and C. A. Taft, '3D descriptors calculation and conformational search to investigate potential bioactive conformations, with application in 3D-QSAR and virtual screening in drug design,' *Journal of Biomolecular Structure and Dynamics*, vol. 35, no. 13, pp. 2966–2974, Oct. 2017, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07391102.2016.1237382, ISSN: 0739-1102. DOI: 10.1080/07391102.2016.1237382. [Online]. Available: https://doi.org/10.1080/07391102.2016.1237382 (visited on 14/06/2022).

[113] *Chemogenomic approaches to rational drug design - Rognan - 2007 - British Journal of Pharmacology - Wiley Online Library*. [Online]. Available: https://bpspubs.onlinelibrary.wiley.com/doi/full/10.1038/sj.bjp.0707307 (visited on 14/07/2022).

[114] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins and W. Tong, 'Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics,' *Journal of Chemical Information and*

Modeling, vol. 48, no. 7, pp. 1337–1344, Jul. 2008, Publisher: American Chemical Society, ISSN: 1549-9596. DOI: 10.1021/ci800038f. [Online]. Available: https://doi.org/10.1021/ci800038f (visited on 14/07/2022).

[115] M. K. Matlock, T. B. Hughes and S. J. Swamidass, 'XenoSite server: A web-available site of metabolism prediction tool,' *Bioinformatics*, vol. 31, no. 7, pp. 1136–1137, Apr. 2015, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu761. [Online]. Available: https://doi.org/10.1093/bioinformatics/btu761 (visited on 14/07/2022).

[116] S. Chaube, S. Goverapet Srinivasan and B. Rai, 'Applied machine learning for predicting the lanthanide-ligand binding affinities,' en, *Scientific Reports*, vol. 10, no. 1, p. 14 322, Aug. 2020, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-020-71255-9. [Online]. Available: https://www.nature.com/articles/s41598-020-71255-9 (visited on 26/06/2022).

[117] B. Chandrasekaran, S. N. Abed, O. Al-Attraqchi, K. Kuche and R. K. Tekade, 'Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties,' en, in *Dosage Form Design Parameters*, Elsevier, 2018, pp. 731–755, ISBN: 978-0-12-814421-3. DOI: 10.1016/B978-0-12-814421-3.00021-X. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B978012814421300021X (visited on 14/06/2022).

[118] P. J. Ropp, J. O. Spiegel, J. L. Walker, H. Green, G. A. Morales, K. A. Milliken, J. J. Ringe and J. D. Durrant, 'Gypsum-DL: An open-source program for preparing small-molecule libraries for structure-based virtual screening,' en, *Journal of Cheminformatics*, vol. 11, no. 1, p. 34, May 2019, ISSN: 1758-2946. DOI: 10.1186/s13321-019-0358-3. [Online]. Available: https://doi.org/10.1186/s13321-019-0358-3 (visited on 14/07/2022).

[119] S. Urbaczek, A. Kolodzik, J. R. Fischer, T. Lippert, S. Heuser, I. Groth, T. Schulz-Gasch and M. Rarey, 'NAOMI: On the Almost Trivial Task of Reading Molecules from Different File formats,' *Journal of Chemical Information and Modeling*, vol. 51, no. 12, pp. 3199–3207, Dec. 2011, Publisher: American Chemical Society, ISSN: 1549-9596. DOI: 10.1021/ci200324e. [Online]. Available: https://doi.org/10.1021/ci200324e (visited on 14/07/2022).

[120] D.-S. Cao, Q. Xu, Q. Hu and Y.-Z. Liang, *Manual for chemopy*, Mar. 2013.

[121] T. Engel, 'Basic overview of chemoinformatics,' *Journal of Chemical Information and Modeling*, vol. 46, no. 6, pp. 2267–2277, 2006, Publisher: American Chemical Society, ISSN: 1549960X. DOI: 10.1021/CI600234Z. (visited on 03/05/2022).

[122] G. Landrum, *RDKit: Open-Source Cheminformatics Software*. [Online]. Available: http://www.rdkit.org/ (visited on 15/06/2022).

[123] V. Bojovic, B. Lucic, D. Bešlo, K. Skala and N. Trinajstic, 'Calculation of topological molecular descriptors based on degrees of vertices,' *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019 - Proceedings*, pp. 266–269, May 2019, Publisher: Institute of Electrical and Electronics Engineers Inc. ISBN: 9789532330984. DOI: 10.23919/MIPRO.2019.8757128. (visited on 21/03/2022).

[124] A. B. T Scior G Tresadern, 'Recognizing pitfalls in virtual screening: A critical review,' *J Chem Inf Model*, vol. 52, no. 4, pp. 867–881, Apr. 2012, Publisher: American Chemical Society. DOI: 10.1021/ci200528d. (visited on 03/11/2021).

[125] G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski and A. Gambin, 'Predicting the outcomes of organic reactions via machine learning: Are current descriptors sufficient?' *Scientific Reports 2017 7:1*, vol. 7, no. 1, pp. 1–9, Jun. 2017, Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-017-02303-0. [Online]. Available: https://www.nature.com/articles/s41598-017-02303-0 (visited on 07/09/2021).

[126] A. K. Ghose and G. M. Crippen, 'Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity,' en, *Journal of Computational Chemistry*, vol. 7, no. 4, pp. 565–577, 1986, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540070419, ISSN: 1096-987X. DOI: 10.1002/jcc.540070419. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540070419 (visited on 05/07/2022).

[127] H. Sun, 'A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption,' *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 2, pp. 748–757, Mar. 2004, Publisher: American Chemical Society, ISSN: 0095-2338. DOI: 10.1021/ci030304f. [Online]. Available: https://doi.org/10.1021/ci030304f (visited on 17/06/2022).

[128] D. C. Kombo, K. Tallapragada, R. Jain, J. Chewning, A. A. Mazurov, J. D. Speake, T. A. Hauser and S. Toler, '3D Molecular Descriptors Important for Clinical Success,' *Journal of Chemical Information and Modeling*, vol. 53, no. 2, pp. 327–342, Feb. 2013, Publisher: American Chemical Society, ISSN: 1549-9596. DOI: 10.1021/ci300445e. [Online]. Available: https://doi.org/10.1021/ci300445e (visited on 14/07/2022).

[129] *Extended-Connectivity Fingerprints | Journal of Chemical Information and Modeling*. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/ci100050t?casa_token=GRV3FTBMAmQAAAAA%3A1b20yfDJ1fKEys-

4waFFgjbRJF8BowP4BzBrlfczvWpxL2zY5TuZxBChG1ugySEL3MN_3U8EhlBxbj1I (visited on 16/06/2022).

[130] D. Probst and J.-L. Reymond, 'A probabilistic molecular fingerprint for big data settings,' en, *Journal of Cheminformatics*, vol. 10, no. 1, p. 66, Dec. 2018, ISSN: 1758-2946. DOI: 10.1186/s13321-018-0321-8. [Online]. Available: https://doi.org/10.1186/s13321-018-0321-8 (visited on 14/07/2022).

[131] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, 'Quantifying the chemical beauty of drugs,' *Nature chemistry*, vol. 4, no. 2, p. 90, Feb. 2012, Publisher: Europe PMC Funders, ISSN: 17554330. DOI: 10.1038/NCHEM.1243. [Online]. Available: /pmc/articles/PMC3524573/ (visited on 18/02/2022).

[132] A. C. Brown and T. R. Fraser, 'On the Connection between Chemical Constitution and Physiological Action; with special reference to the Physiological Action of the Salts of the Ammonium Bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia,' *Journal of Anatomy and Physiology*, vol. 2, no. 2, pp. 224–242, 1868. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1318606/ (visited on 21/06/2022).

[133] A. Rácz, D. Bajusz and K. Héberger, 'Intercorrelation Limits in Molecular Descriptor Preselection for QSAR/QSPR,' en, *Molecular Informatics*, vol. 38, no. 8-9, p. 1 800 154, 2019, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/min ISSN: 1868-1751. DOI: 10.1002/minf.201800154. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201800154 (visited on 08/06/2022).

[134] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, 'Open Babel: An Open chemical toolbox,' *Journal of Cheminformatics*, vol. 3, no. 10, pp. 1–14, Oct. 2011, Publisher: BioMed Central, ISSN: 17582946. DOI: 10.1186/1758-2946-3-33/TABLES/2. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-3-33 (visited on 06/04/2022).

[135] *Conformational Sampling of Druglike Molecules with MOE and Catalyst: Implications for Pharmacophore Modeling and Virtual Screening | Journal of Chemical Information and Modeling*. [Online]. Available: https://pubs.acs.org/doi/10.1021/ci800130k (visited on 18/07/2022).

[136] A. Lauria, M. Ippolito and A. M. Almerico, 'Principal component analysis on molecular descriptors as an alternative point of view in the search of new Hsp90 inhibitors,' en, *Computational Biology and Chemistry*, vol. 33, no. 5, pp. 386–390, Oct. 2009, ISSN: 1476-9271. DOI: 10.1016/j.compbiolchem.2009.07.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1476927109000656 (visited on 08/06/2022).

[137] *Simplified molecular-input line-entry system*, en, Page Version ID: 1099548399, Jul. 2022. [Online]. Available: `https://en.wikipedia.org/w/index.php?title=Simplified_molecular-input_line-entry_system&oldid=1099548399` (visited on 30/07/2022).

[138] N. M. O'Boyle, 'Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI,' *Journal of Cheminformatics*, vol. 4, p. 22, Sep. 2012, ISSN: 1758-2946. DOI: `10.1186/1758-2946-4-22`. [Online]. Available: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3495655/` (visited on 14/07/2022).

[139] *International Union of Pure and Applied Chemistry*, en-US. [Online]. Available: `https://iupac.org/` (visited on 14/07/2022).

[140] *National Institute of Standards and Technology*, en, text, Last Modified: 2022-07-13T10:02-04:00. [Online]. Available: `https://www.nist.gov/` (visited on 14/07/2022).

[141] *InChI Trust – InChI: Open-source chemical structure representation algorithm*. [Online]. Available: `https://www.inchi-trust.org/` (visited on 14/07/2022).

[142] *Chem1102: Drug Discovery, Fordham University."* en, Aug. 2020. [Online]. Available: `https://chem.libretexts.org/Courses/Fordham_University/Chem11023A_Drug_Discovery_-_From_the_Laboratory_to_the_Clinic/053A_Organic_Molecules/5.083A_Line_Notation_(SMILES_and_InChI)` (visited on 06/07/2022).

[143] C. Southan, 'InChI in the wild: An assessment of InChIKey searching in Google,' en, *Journal of Cheminformatics*, vol. 5, no. 1, p. 10, Feb. 2013, ISSN: 1758-2946. DOI: `10.1186/1758-2946-5-10`. [Online]. Available: `https://doi.org/10.1186/1758-2946-5-10` (visited on 14/07/2022).

[144] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick and J. W. Davies, 'How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space,' *Journal of Chemical Information and Modeling*, vol. 49, no. 1, pp. 108–119, Jan. 2009, Publisher: American Chemical Society, ISSN: 1549-9596. DOI: `10.1021/ci800249s`. [Online]. Available: `https://doi.org/10.1021/ci800249s` (visited on 15/07/2022).

[145] J. Gasteiger, 'Chemoinformatics: Achievements and Challenges, a Personal View,' *Molecules 2016, Vol. 21, Page 151*, vol. 21, no. 2, p. 151, Jan. 2016, Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1420-3049. DOI: `10.3390/MOLECULES21020151`. [Online]. Available: `https://www.mdpi.com/1420-3049/21/2/151/htm` (visited on 03/05/2022).

[146] A. David and R. W. Grosse-Kunstleve, *Building hybrid systems with Boost.Python | OSTI.GOV*, 2003. [Online]. Available: `https://www.osti.gov/biblio/815409` (visited on 23/04/2022).

[147] *Open Source Science*, en. [Online]. Available: https://www.novartis.com/research-development/open-source-science (visited on 14/07/2022).

[148] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing and Z. Wu, *Deep Learning for the Life Sciences*. O'Reilly Media, 2019.

[149] A. Fillbrunn, C. Dietz, J. Pfeuffer, R. Rahn, G. A. Landrum and M. R. Berthold, 'KNIME for reproducible cross-domain analysis of life science data,' en, *Journal of Biotechnology*, Bioinformatics Solutions for Big Data Analysis in Life Sciences presented by the German Network for Bioinformatics Infrastructure, vol. 261, pp. 149–156, Nov. 2017, ISSN: 0168-1656. DOI: 10.1016/j.jbiotec.2017.07.028. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168165617315651 (visited on 14/07/2022).

[150] S. Yuan, H. S. Chan and Z. Hu, 'Using PyMOL as a platform for computational drug design,' en, *WIREs Computational Molecular Science*, vol. 7, no. 2, e1298, 2017, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1298, ISSN: 1759-0884. DOI: 10.1002/wcms.1298. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1298 (visited on 14/07/2022).

[151] A. Mauri, 'alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints,' en, in *Ecotoxicological QSARs*, ser. Methods in Pharmacology and Toxicology, K. Roy, Ed., New York, NY: Springer US, 2020, pp. 801–820, ISBN: 978-1-07-160150-1. DOI: 10.1007/978-1-0716-0150-1_32. [Online]. Available: https://doi.org/10.1007/978-1-0716-0150-1_32 (visited on 30/06/2022).

[152] PubChem, *Three disacharide linked murein units (tetrapeptide crosslinked tetrapeptide (A2pm->D-ala), one uncrosslinked tetrapaptide) (middle of chain)*, en. [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/compound/46173100 (visited on 03/07/2022).

[153] *MetaNetX: ID mapper*, en. [Online]. Available: https://www.metanetx.org/cgi-bin/mnxweb/id-mapper (visited on 15/07/2022).

[154] R. Caspi, H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S. Y. Rhee, C. Tissier, P. Zhang and P. D. Karp, 'MetaCyc: A multiorganism database of metabolic pathways and enzymes.,' *Nucleic acids research*, vol. 34, no. Database issue, 2006, ISSN: 13624962. DOI: 10.1093/nar/gkj128.

[155] *HMDB 4.0: The human metabolome database for 2018 | Nucleic Acids Research | Oxford Academic*. [Online]. Available: https://academic.oup.com/nar/article/46/D1/D608/4616873 (visited on 03/08/2022).

[156] M. L. Waskom, 'Seaborn: Statistical data visualization,' en, *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, Apr. 2021, ISSN: 2475-9066. DOI: 10.21105/joss.03021. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.03021 (visited on 17/07/2022).

[157] A. Hagberg, P. Swart and D. S Chult, 'Exploring network structure, dynamics, and function using networkx,' English, Los Alamos National Lab. (LANL), Los Alamos, NM (United States), Tech. Rep. LA-UR-08-05495; LA-UR-08-5495, Jan. 2008. [Online]. Available: `https://www.osti.gov/biblio/960616` (visited on 17/07/2022).

[158] *Matplotlib: A 2D Graphics Environment*. [Online]. Available: `https://www.computer.org/csdl/magazine/cs/2007/03/c3090/13rRUwbJD0A` (visited on 17/07/2022).

[159] *Sbgn/ySBGN: Translation between GraphML (yED) and SBGN-ML*. [Online]. Available: `https://github.com/sbgn/ySBGN` (visited on 17/07/2022).

[160] PubChem, *Alpha-D-ribose 1-methylphosphonate 5-phosphate*, en. [Online]. Available: `https://pubchem.ncbi.nlm.nih.gov/compound/70678985` (visited on 04/08/2022).

[161] N. Pham, R. G. A. van Heck, J. C. J. van Dam, P. J. Schaap, E. Saccenti and M. Suarez-Diez, 'Consistency, Inconsistency, and Ambiguity of Metabolite Names in Biochemical Databases Used for Genome-Scale Metabolic Modelling,' en, *Metabolites*, vol. 9, no. 2, p. 28, Feb. 2019, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2218-1989. DOI: `10.3390/metabo9020028`. [Online]. Available: `https://www.mdpi.com/2218-1989/9/2/28` (visited on 12/07/2022).

[162] *SMILES Tutorial: Atoms*. [Online]. Available: `https://www.daylight.com/meetings/summerschool98/course/dave/smiles-atoms.html` (visited on 25/07/2022).

[163] D. T. Stanton, 'Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies,' en, *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 1, pp. 11–20, Jan. 1999, ISSN: 0095-2338. DOI: `10.1021/ci980102x`. [Online]. Available: `https://pubs.acs.org/doi/10.1021/ci980102x` (visited on 23/07/2022).

[164] *The Electrotopological State: An Atom Index for QSAR - Hall - 1991 - Quantitative Structure-Activity Relationships - Wiley Online Library*. [Online]. Available: `https://onlinelibrary.wiley.com/doi/abs/10.1002/qsar.19910100108` (visited on 25/06/2022).

[165] *Re: [Rdkit-discuss] Using the RDKit with Dask*. [Online]. Available: `https://www.mail-archive.com/rdkit-discuss@lists.sourceforge.net/msg10510.html` (visited on 01/07/2022).

[166] E. Lounkine and J. Bajorath, *Topological Fragment Index for the Analysis of Molecular Substructures and Their Topological Environment in Active Compounds*, EN, research-article, Archive Location: world Publisher: American Chemical Society, Nov. 2008. DOI: `10.1021/ci8002599`. [Online]. Available: `https://pubs.acs.org/doi/pdf/10.1021/ci8002599` (visited on 18/07/2022).

[167] E. Lounkine, J. Auer and J. Bajorath, ‘Formal concept analysis for the identification of molecular fragment combinations specific for active and highly potent compounds,’ eng, *Journal of Medicinal Chemistry*, vol. 51, no. 17, pp. 5342–5348, Sep. 2008, ISSN: 1520-4804. DOI: 10.1021/jm800515r.

[168] F. Krüger, E. Lounkine and J. Bajorath, ‘Fragment formal concept analysis accurately classifies compounds with closely related biological activities,’ eng, *ChemMedChem*, vol. 4, no. 7, pp. 1174–1181, Jul. 2009, ISSN: 1860-7187. DOI: 10.1002/cmdc.200900035.

[169] H. E. A. Ahmed and J. Bajorath, ‘Methods for computer-aided chemical biology. Part 5: Rationalizing the selectivity of cathepsin inhibitors on the basis of molecular fragments and topological feature distributions,’ eng, *Chemical Biology & Drug Design*, vol. 74, no. 2, pp. 129–141, Aug. 2009, ISSN: 1747-0285. DOI: 10.1111/j.1747-0285.2009.00848.x.

[170] Á. Orosz, K. Héberger and A. Rácz, ‘Comparison of Descriptor- and Fingerprint Sets in Machine Learning Models for ADME-Tox Targets,’ *Frontiers in Chemistry*, vol. 10, Jun. 2022. DOI: 10.3389/fchem.2022.852893.

[171] *MetaCyc L-histidine biosynthesis*. [Online]. Available: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=HISTSYN-PWY&detail-level=1 (visited on 05/07/2022).

[172] H. E. Pence and A. Williams, ‘ChemSpider: An Online Chemical Information Resource,’ *Journal of Chemical Education*, vol. 87, no. 11, pp. 1123–1124, Nov. 2010, Publisher: American Chemical Society, ISSN: 0021-9584. DOI: 10.1021/ed100697w. [Online]. Available: https://doi.org/10.1021/ed100697w (visited on 12/07/2022).

[173] T.-S. Lin, C. W. Coley, H. Mochigase, H. K. Beech, W. Wang, Z. Wang, E. Woods, S. L. Craig, J. A. Johnson, J. A. Kalow, K. F. Jensen and B. D. Olsen, ‘BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules,’ *ACS Central Science*, vol. 5, no. 9, pp. 1523–1531, Sep. 2019, Publisher: American Chemical Society, ISSN: 2374-7943. DOI: 10.1021/acscentsci.9b00476. [Online]. Available: https://doi.org/10.1021/acscentsci.9b00476 (visited on 03/08/2022).

[174] E. Schreiner, L. G. Trabuco, P. L. Freddolino and K. Schulten, ‘Stereochemical errors and their implications for molecular dynamics simulations,’ *BMC Bioinformatics*, vol. 12, p. 190, May 2011, ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-190. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3124434/ (visited on 23/07/2022).

[175] T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li, R. Wang and L. Lai, ‘Computation of OctanolWater Partition Coefficients by Guiding an Additive Model with Knowledge,’ *Journal of Chemical Information and Modeling*, vol. 47, no. 6, pp. 2140–2148, Nov. 2007, Publisher: American Chemical Society, ISSN: 1549-9596. DOI: 10.1021/ci700257y. [Online]. Available: https://doi.org/10.1021/ci700257y (visited on 03/07/2022).

[176] D. S. Wigh, J. M. Goodman and A. A. Lapkin, 'A review of molecular representation in the age of machine learning,' en, *WIREs Computational Molecular Science*, vol. n/a, no. n/a, e1603, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ ISSN: 1759-0884. DOI: 10.1002/wcms.1603. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1603 (visited on 23/07/2022).

[177] *Technical FAQ – for 1.05 – InChI Trust*, en-GB. [Online]. Available: https://www.inchi-trust.org/technical-faq/ (visited on 01/07/2022).

[178] *Empty smiles will cause a memory crash issue when calculating the BCUT2D descriptors. · Issue #4064 · rdkit/rdkit*, en. [Online]. Available: https://github.com/rdkit/rdkit/issues/4064 (visited on 04/07/2022).

[179] L. Xue and J. Bajorath, 'Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principal Component Analysis Identified by a Genetic Algorithm,' *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 3, pp. 801–809, May 2000, Publisher: American Chemical Society, ISSN: 0095-2338. DOI: 10.1021/ci000322m. [Online]. Available: https://doi.org/10.1021/ci000322m (visited on 23/07/2022).

[180] A. K. Nigam, A. A. Ojha, J. G. Li, D. Shi, V. Bhatnagar, K. B. Nigam, R. Abagyan and S. K. Nigam, 'Molecular Properties of Drugs Handled by Kidney OATs and Liver OATPs Revealed by Chemoinformatics and Machine Learning: Implications for Kidney and Liver Disease,' en, *Pharmaceutics*, vol. 13, no. 10, p. 1720, Oct. 2021, Number: 10 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1999-4923. DOI: 10.3390/pharmaceutics13101720. [Online]. Available: https://www.mdpi.com/1999-4923/13/10/1720 (visited on 04/07/2022).

[181] P. Labute, 'A widely applicable set of descriptors,' en, *Journal of Molecular Graphics and Modelling*, vol. 18, no. 4-5, pp. 464–477, 2000, ISSN: 10933263. DOI: 10.1016/S1093-3263(00)00068-1. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1093326300000681 (visited on 14/06/2022).

[182] S. Prasanna and R. J. Doerksen, 'Topological Polar Surface Area: A Useful Descriptor in 2D-QSAR,' *Current medicinal chemistry*, vol. 16, no. 1, pp. 21–41, 2009, ISSN: 0929-8673. DOI: 10.2174/092986709787002817. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7549127/ (visited on 14/06/2022).

[183] D. A. Pertusi, A. E. Stine, L. J. Broadbelt and K. E. Tyo, 'Efficient searching and annotation of metabolic networks using chemical similarity,' *Bioinformatics*, vol. 31, no. 7, pp. 1016–1024, Apr. 2015, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu760. [Online]. Available: https://doi.org/10.1093/bioinformatics/btu760 (visited on 22/07/2022).

[184] I. Rojdestvenski, 'Metabolic pathways in three dimensions,' *Bioinformatics*, vol. 19, no. 18, pp. 2436–2441, Dec. 2003, Publisher: Oxford Academic, ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTG342. [Online]. Available: https://academic.oup.com/bioinformatics/article/19/18/2436/194521 (visited on 13/05/2022).

[185] Á. M. Fernández, L. A. Bongo and E. Pedersen, *GeneNet VR: Interactive visualization of large-scale biological networks using a standalone headset*, arXiv:2109.02937 [cs], Sep. 2021. DOI: 10.48550/arXiv.2109.02937. [Online]. Available: http://arxiv.org/abs/2109.02937 (visited on 28/07/2022).

[186] U. Technologies, *Unity Real-Time Development Platform | 3D, 2D VR & AR Engine*, en. [Online]. Available: https://unity.com/ (visited on 04/08/2022).

[187] *GeneNet VR: Large Biological Networks in Virtual Reality - YouTube*. [Online]. Available: https://www.youtube.com/watch?v=N4QDZiZqVNY (visited on 28/07/2022).

[188] B. Foundation, *Blender.org - Home of the Blender project - Free and Open 3D Creation Software*, en. [Online]. Available: https://www.blender.org/ (visited on 04/08/2022).

[189] *iGEM2020_uioslo_norway (sal.coli)*, original-date: 2020-09-24T15:05:37Z, Nov. 2020. [Online]. Available: https://github.com/igemsoftware2020/UiOslo-Norway (visited on 04/08/2022).

[190] N. Curti, *BlendNet*, 2019. [Online]. Available: https://github.com/Nico-Curti/BlendNet.

[191] U. Brandes, T. Dwyer and F. Schreiber, 'Visualizing Related Metabolic Pathways in Two and a Half Dimensions,' en, in *Graph Drawing*, G. Goos, J. Hartmanis, J. van Leeuwen and G. Liotta, Eds., vol. 2912, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 111–122, ISBN: 978-3-540-20831-0 978-3-540-24595-7. DOI: 10.1007/978-3-540-24595-7_10. [Online]. Available: http://link.springer.com/10.1007/978-3-540-24595-7_10 (visited on 17/06/2022).

[192] J. Zucker, *Cytoscape2Escher*, en. [Online]. Available: https://gist.github.com/djinnome/aead21ef178fc0fd691d5fa71ad5ddb1 (visited on 12/07/2022).

[193] *Pathway Tools Software*. [Online]. Available: http://brg.ai.sri.com/ptools/ (visited on 12/07/2022).

[194] A. Disney and 2. February 2021, *Force-directed graph layouts explained*, en, Feb. 2021. [Online]. Available: https://cambridge-intelligence.com/keylines-faq-force-directed-layouts/ (visited on 16/07/2022).

Appendix A

Appendix A: Supplementary data

Supplementary data

The supplementary data listed in table A.1 is available in the GitHub repository in: `https://github.com/meidelien/Biochemical_coordinate_layout`

Table A.1: Supplementary data is located in the associated GitHub repository. Fields in the Name/Description column are interactive and will link to the associated data when clicked

Supplementary data	Name/Description
1	Data collection
2	Molecular descriptor calculation
3	Case study datasets
4	Exported layouts

Software

This section covers the different software and python libraries used in this project.

Python

The code used to compute molecular descriptors, perform data analysis and visualisations were written in the Python programming language and has been verified to work under versions 3.7 through 3.9.

RDkit

The RDKit software package [RDKit reference here] is an open source cheminformatics toolkit. Its core data structures and algorithms arewritten in C++, but are accessed through a Python wrapper. RDkit calculates molecular descriptors for molecules using their SMILES or InChI strings as input. The version used in this project was 2021.03.5.

ReFramed

ReFramed is a Python package that simulates metabolic models using SBML input. This package was used to generate metabolite links to enable network generation and visualisation in NetworkX and Pyvis. The version used for this project was 1.2.1.

NetworkX

NetworkX is a python package for creating and studying complex networks. The version used in this project was 2.6.3.

Scikit-learn

Scikit-learn is a machine learning package for python. Scikit-learn was used to standardise data and perform PCA on the data. The version used in this project was 1.0.1.

Pandas

Pandas is a python package for data analysis and manipulation. The version used in this project was 1.3.2.

NumPy

NumPy is a python package that provides a large suite of mathematical operations that can be applied to arrays and data structures. The version used in this project was 1.20.3.

Seaborn

Seaborn is a python package for statistical data visualisation and was used to visualise the correlation between features of the GEM. The version used in this project was 0.11.2.

SciPy

SciPy is a python package built on top of NumPy for use in data manipulation and visualisation. The version used in this project was 2.7.1.

DataComPy

DataCompy is a python package that is made to compare two pandasvdataframes on an index key that is present across both dataframes. The version used in this project was 0.81.

Cytoscape

The version used in this project was 3.9.1

Appendix B

Appendix B: Excerpts of metabolic pathway case study datasets

Table B.1: Excerpt from the tricarboxylic acid cycle dataset in supplementary data 3 A.1, A.

Metabolite	BiGG abbreviation	Molecular weight	MolLogP
Pyruvate	pyr	87.00877	-1.674
Acetyl-CoA	accoa	805.09667	-3.294
2-Oxoglutarate	akg	144.00697	-3.164
Oxaloacetate	oaa	129.99132	-3.554
Succinate	succ	116.01206	-2.733
Phosphoenolpyruvate	pep	164.96055	-2.904
Succinyl-CoA	succoa	862.09487	-4.783
Fumarate	fum	113.99641	-2.957
D-Malate	mal__d	132.00697	-3.762
Citrate	cit	189.00517	-5.252
Isocitrate	icit	189	-5.396

Table B.2: Excerpt from the ATP biosynthesis pathway dataset in supplementary data 3 A.1, A.

Metabolite	**BiGG abbreviation**	**Molecular weight**	**MolLogP**
5-phospho-alpha-D-ribose 1-diphosphate	prpp	384.91	-5.391
5-phospho-beta-D-ribosylamine	pram	228.02	-3.130
N(1)-(5-phospho-beta-D-ribosyl)glycinamide	gar	285.049	-3.7407
N(2)-formyl-N(1)-(5-phospho-beta-D-ribosyl)glycinamide	fgam	312.03	-4.310
2-formamido-N(1)-(5-O-phospho-beta-D-ribosyl)acetamidine	fpram	312.06	-3.507
5-amino-1-(5-phospho-beta-D-ribosyl)imidazole	air	294.04	-2.438
5-carboxyamino-1-(5-phospho-D-ribosyl)imidazole	5caiz	336.02	-3.897
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxylate	5aizc	337.03	-4.074
(2S)-2-[5-amino-1-(5-phospho-beta-D-ribosyl)imidazole-4-carboxamido]succinate	25aics	450.04	-7.072
5-amino-1-(5-phospho-beta-D-ribosyl)imidazole-4-carboxamide	aicar	336.04	-3.752
5-formamido-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide	fprica	364.04	-3.901
Inosine monophosphate	imp	346.03	-3.003
N6-(1,2-Dicarboxyethyl)-AMP	dcamp	463.07	-2.105
Adenosine monophoshphate	amp	345.04	-3.126
Adenosine diphoshphate	adp	424	-3.641
Deoxyadenosine diphosphate	dadp	408.01	-2.612
Adenosine triphoshphate	atp	502.96	-4.156

Table B.3: Excerpt from the CTP biosynthesis pathway dataset in supplementary data 3 A.1, A.

Metabolite	BiGG abbreviation	Molecular weight	MolLogP
L-aspartate	asp__L	132.03	-2.461
N-carbamoyl-L-aspartate	Mbasp	174.02	-3.672
(S)-dihydroorotate	dhor__S	157.02	-1.483
Orotate	orot	155.00	-1.046
Orotidine 5'-phosphate	orot5p	365.00	-5.223
Uridine monophosphate	ump	322.02	-3.586
Uridine diphosphate	udp	400.98	-4.101
Uridine triphosphate	utp	479.93	-4.616
Cytidine triphosphate	ctp	478.95	-4.497

Table B.4: Excerpt from the GTP biosynthesis pathway dataset in supplementary data 3 A.1, A.

Metabolite	**BiGG abbreviation**	**Molecular weight**	**MolLogP**
5-phospho-alpha-D-ribose 1-diphosphate	prpp	384.91	-5.391
5-phospho-beta-D-ribosylamine	pram	228.02	-3.130
N(1)-(5-phospho-beta-D-ribosyl)glycinamide	gar	285.04	-3.740
N(2)-formyl-N(1)-(5-phospho-beta-D-ribosyl)glycinamide	fgam	312.03	-4.310
2-formamido-N(1)-(5-O-phospho-beta-D-ribosyl)acetamidine	fpram	312.06	-3.502
5-amino-1-(5-phospho-beta-D-ribosyl)imidazole	air	294.04	-2.438
5-carboxyamino-1-(5-phospho-D-ribosyl)imidazole	5caiz	336.02	-3.897
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxylate	5aizc	337.03	-4.074
(2S)-2-[5-amino-1-(5-phospho-beta-D-ribosyl)imidazole-4-carboxamido]succinate	25aics	450.04	-7.072
5-amino-1-(5-phospho-beta-D-ribosyl)imidazole-4-carboxamide	aicar	336.04	-3.752
5-formamido-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide	fprica	364.04	-3.901
Inosine monophosphate	imp	346.03	-3.003
Xanthosine monophosphate	xmp	362.02	-3.297
Guanidyl monophosphate	gmp	361.04	-3.591
Guanidyl diphosphate	gdp	440.	-4.106
Guanidyl triphosphate	gtp	518.96	-4.621

Table B.5: Excerpt from the histidine biosynthesis pathway dataset in supplementary data 3 A.1, A.

Metabolite	**BiGG abbreviation**	**Molecular weight**	**MolLogP**
5-phospho-alpha-D-ribose	prpp	384.91	-5.39
1-(5-Phosphoribosyl)-ATP	prbatp	715.98	-6.536
1-(5-Phosphoribosyl)-AMP	prbampt	558.06	-5.505
Phosphoribosyl-formimino-AICAR-P	prfp	573.05	-7.012
phosphoribulosylformimino-AICAR-P	prlp	573.05	-6.977
D-erythro-1-(imidazol-4-yl)glycerol 3-phosphate	eig3p	236.02	-2.35
3-(imidazol-4-yl)-2-oxopropyl phosphate	imacp	218.01	-1.633
L-histidinol phosphate	hisp	220.04	-1.243
L-histidinol	histd	142.09	-1.444
L-histidine	his__L	155.06	-0.635

Appendix C

Appendix C: MetaCyC metabolic pathway visualizations with compound structures

C.1 The tricarboxylic acid cycle with compound structures

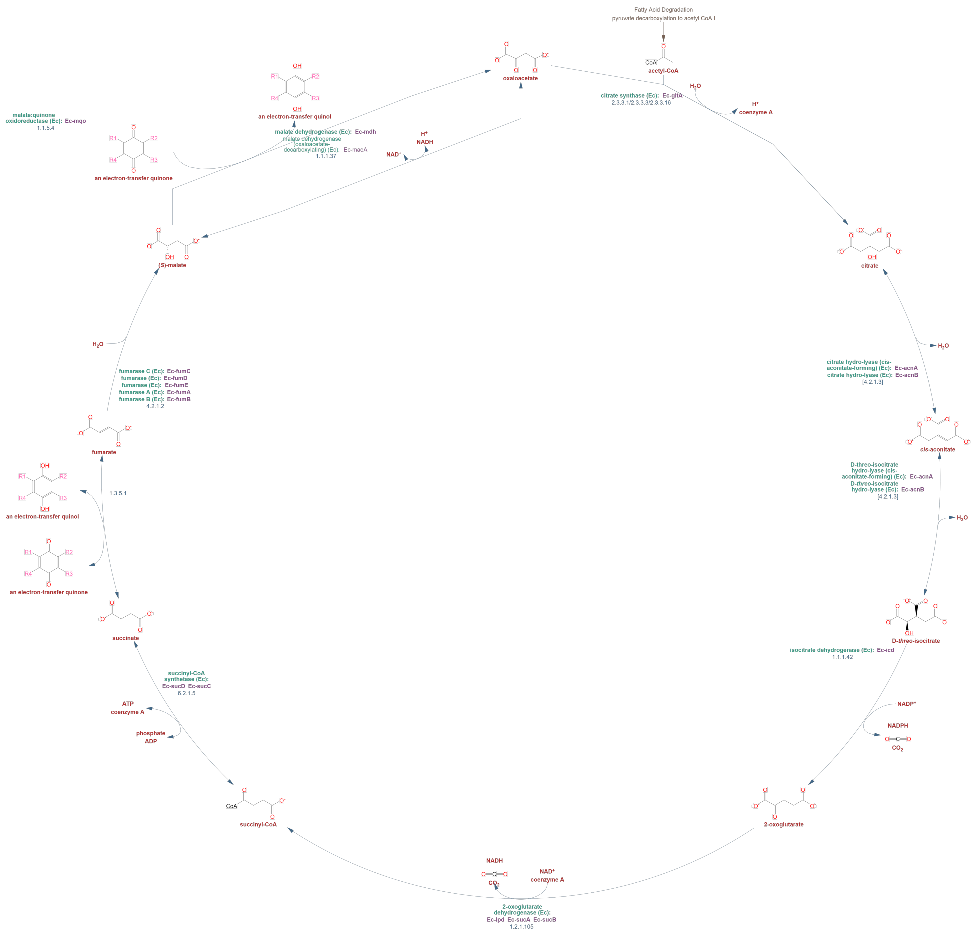

Fatty Acid Degradation
pyruvate decarboxylation to acetyl CoA I
CoA
acetyl-CoA
oxaloacetate
citrate synthase (Ec): Ec-gltA
2.3.3.1/2.3.3.3/2.3.3.16
H_2O
H^+
coenzyme A
citrate
H_2O
citrate hydro-lyase (cis-aconitate-forming) (Ec): Ec-acnA
citrate hydro-lyase (Ec): Ec-acnB
[4.2.1.3]
cis-aconitate
D-threo-isocitrate hydro-lyase (cis-aconitate-forming) (Ec): Ec-acnA
D-*threo*-isocitrate hydro-lyase (Ec): Ec-acnB
[4.2.1.3]
H_2O
D-*threo*-isocitrate
isocitrate dehydrogenase (Ec): Ec-icd
1.1.1.42
$NADP^+$
NADPH
CO_2
2-oxoglutarate
NADH
CO_2
NAD^+
coenzyme A
2-oxoglutarate dehydrogenase (Ec):
Ec-lpd Ec-sucA Ec-sucB
1.2.1.105
succinyl-CoA
CoA
succinyl-CoA synthetase (Ec):
Ec-sucD Ec-sucC
6.2.1.5
ATP
coenzyme A
phosphate
ADP
succinate
an electron-transfer quinol
an electron-transfer quinone
1.3.5.1
fumarate
fumarase C (Ec): Ec-fumC
fumarase (Ec): Ec-fumD
fumarase (Ec): Ec-fumE
fumarase A (Ec): Ec-fumA
fumarase B (Ec): Ec-fumB
4.2.1.2
H_2O
(*S*)-malate
malate:quinone oxidoreductase (Ec): Ec-mqo
1.1.5.4
an electron-transfer quinone
an electron-transfer quinol
malate dehydrogenase (Ec): Ec-mdh
malate dehydrogenase (oxaloacetate-decarboxylating) (Ec): Ec-maeA
1.1.1.37
NAD^+
H^+
NADH

Figure C.1: Pathway representation of the tricarboxylic acid cycle in *E. coli* with compound structures. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [62].

C.2 *De novo* purine biosynthesis pathway with compound structures

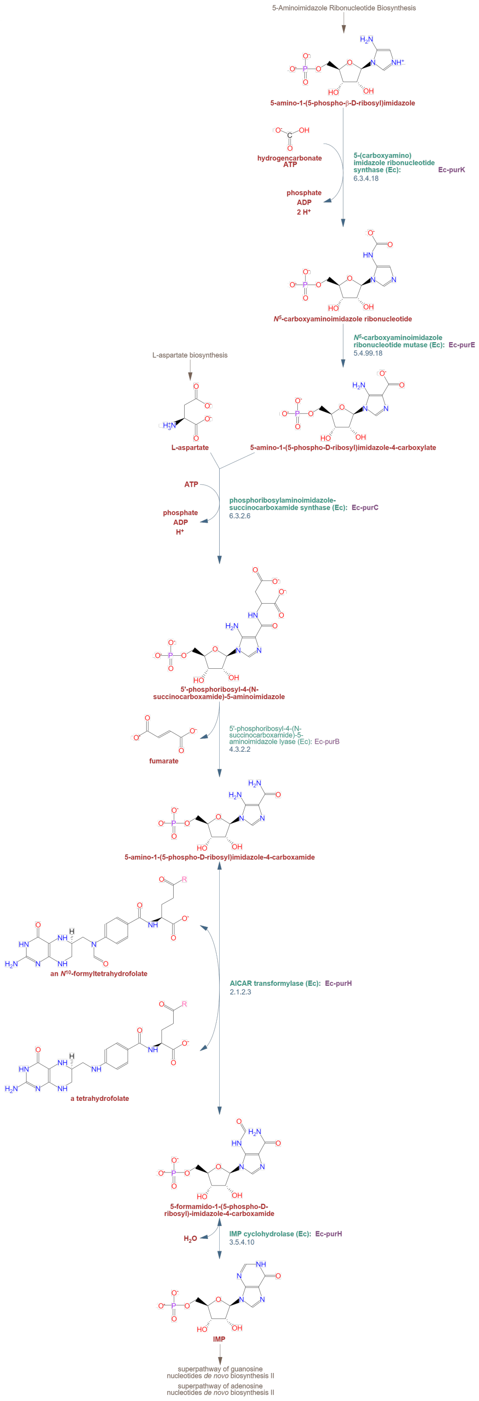

5-Aminoimidazole Ribonucleotide Biosynthesis
5-amino-1-(5-phospho-β-D-ribosyl)imidazole
hydrogencarbonate
ATP
phosphate
ADP
2 H+
5-(carboxyamino) imidazole ribonucleotide synthase (Ec): Ec-purK
6.3.4.18
N5-carboxyaminoimidazole ribonucleotide
N5-carboxyaminoimidazole ribonucleotide mutase (Ec): Ec-purE
5.4.99.18
L-aspartate biosynthesis
L-aspartate
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxylate
ATP
phosphate
ADP
H+
phosphoribosylaminoimidazole-succinocarboxamide synthase (Ec): Ec-purC
6.3.2.6
5'-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole
fumarate
5'-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole lyase (Ec): Ec-purB
4.3.2.2
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide
an N10-formyltetrahydrofolate
AICAR transformylase (Ec): Ec-purH
2.1.2.3
a tetrahydrofolate
5-formamido-1-(5-phospho-D-ribosyl)-imidazole-4-carboxamide
H2O
IMP cyclohydrolase (Ec): Ec-purH
3.5.4.10
IMP
superpathway of guanosine nucleotides de novo biosynthesis II
superpathway of adenosine nucleotides de novo biosynthesis II

Figure C.2: Pathway representation of inosine-5'-phosphate biosynthesis in *E. coli* with compound structures. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [63].

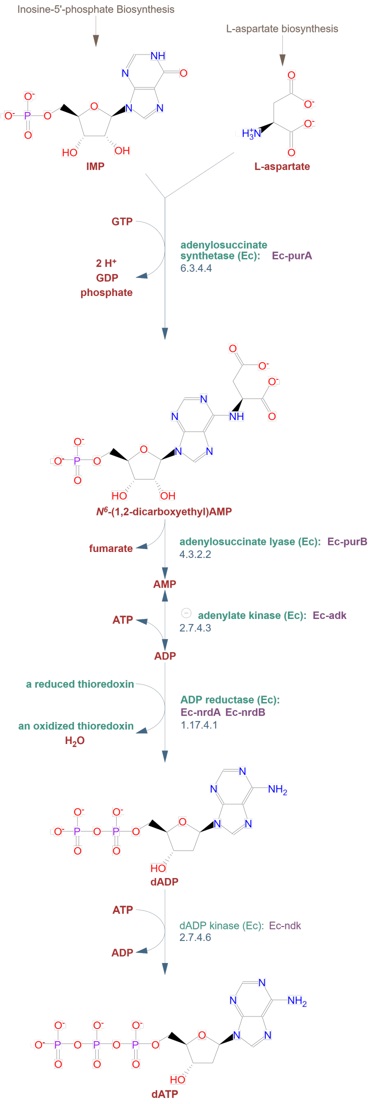


Figure C.3: Pathway representation of *de novo* adenosine nucleotide biosynthesis in *E. coli*. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [63].

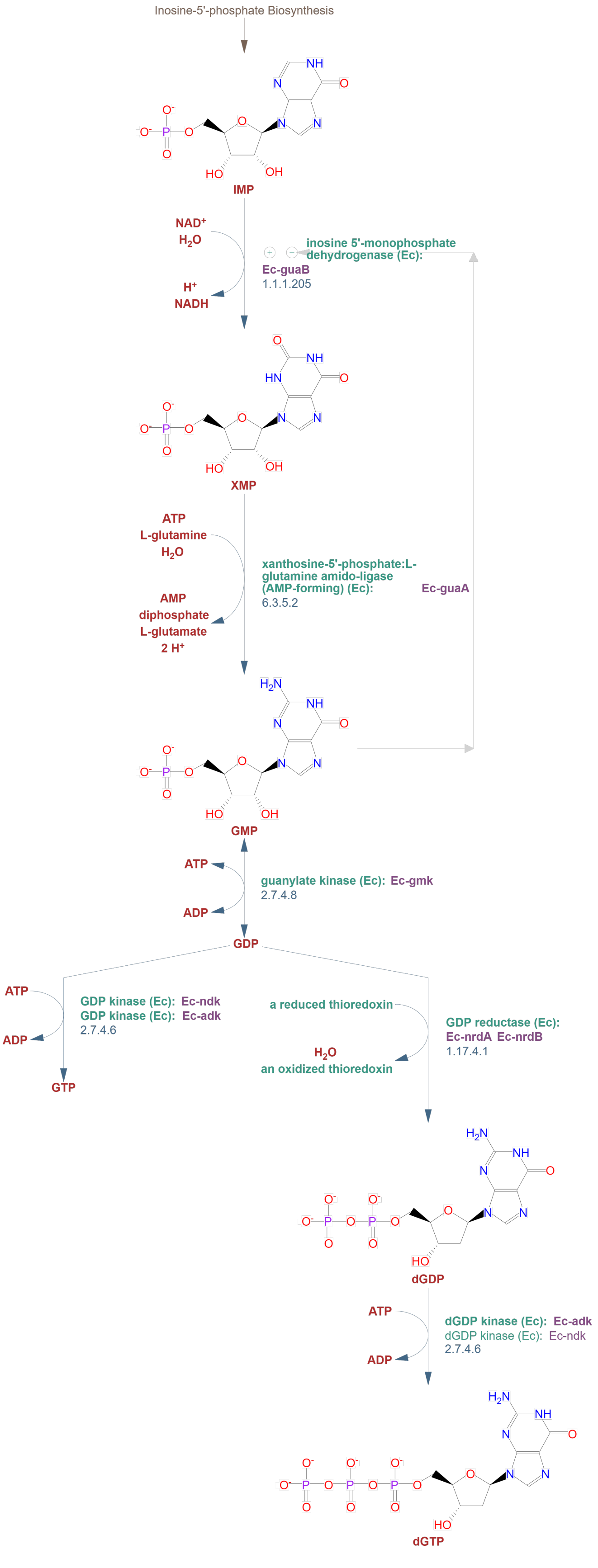


Figure C.4: Pathway representation of *de novo* guanosine nucleotide biosynthesis in *E. coli* with compound structures. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [64].

C.3 *De novo* pyrimidine biosynthesis pathway with compound structures

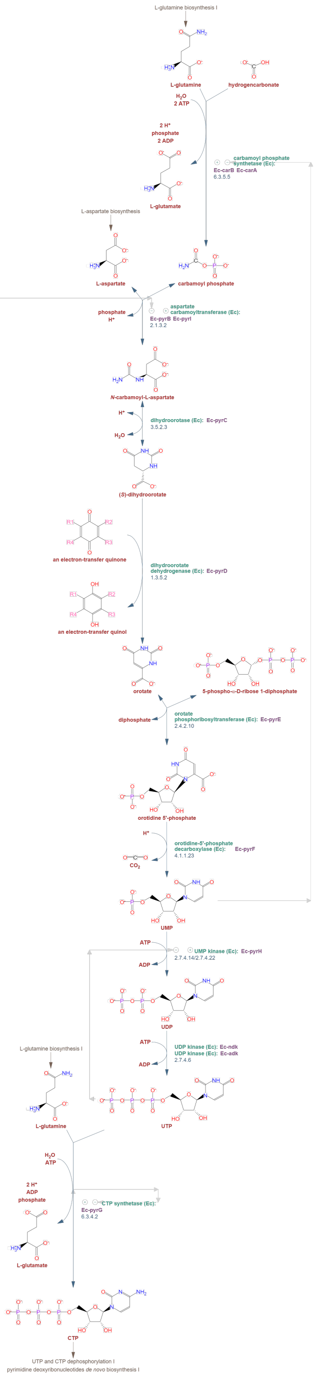


Figure C.5: Pathway representation of *de novo* CTP biosynthesis in *E. coli* with compound structures. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [65].

C.4 Histidine biosynthesis with compound structures

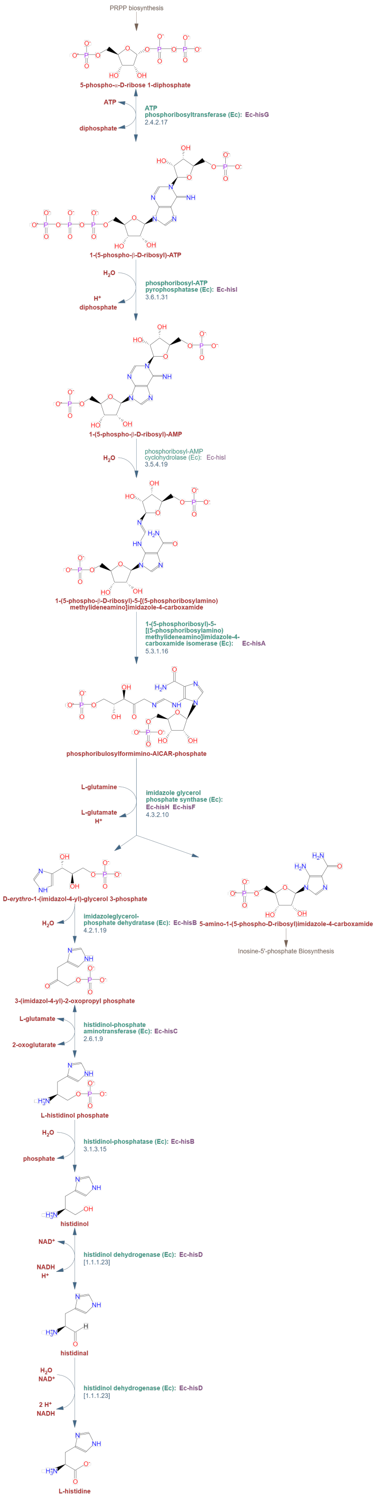

PRPP biosynthesis
5-phospho-α-D-ribose 1-diphosphate
ATP
diphosphate
ATP phosphoribosyltransferase (Ec): Ec-hisG
2.4.2.17
1-(5-phospho-β-D-ribosyl)-ATP
H_2O
H^+
diphosphate
phosphoribosyl-ATP pyrophosphatase (Ec): Ec-hisI
3.6.1.31
1-(5-phospho-β-D-ribosyl)-AMP
H_2O
phosphoribosyl-AMP cyclohydrolase (Ec): Ec-hisI
3.5.4.19
1-(5-phospho-β-D-ribosyl)-5-[(5-phosphoribosylamino) methylideneamino]imidazole-4-carboxamide
1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino) methylideneamino]imidazole-4-carboxamide isomerase (Ec): Ec-hisA
5.3.1.16
phosphoribulosylformimino-AICAR-phosphate
L-glutamine
L-glutamate
H^+
imidazole glycerol phosphate synthase (Ec): Ec-hisH Ec-hisF
4.3.2.10
D-*erythro*-1-(imidazol-4-yl)-glycerol 3-phosphate
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide
Inosine-5'-phosphate Biosynthesis
H_2O
imidazoleglycerol-phosphate dehydratase (Ec): Ec-hisB
4.2.1.19
3-(imidazol-4-yl)-2-oxopropyl phosphate
L-glutamate
2-oxoglutarate
histidinol-phosphate aminotransferase (Ec): Ec-hisC
2.6.1.9
L-histidinol phosphate
H_2O
phosphate
histidinol-phosphatase (Ec): Ec-hisB
3.1.3.15
histidinol
NAD^+
NADH
H^+
histidinol dehydrogenase (Ec): Ec-hisD
[1.1.1.23]
histidinal
H_2O
NAD^+
2 H^+
NADH
histidinol dehydrogenase (Ec): Ec-hisD
[1.1.1.23]
L-histidine

Figure C.6: Pathway representation of *de novo* histidine biosynthesis in *E. coli* with compound structures. Compound names are colored red. Enzyme names are colored green and their corresponding genes are colored purple. Enzyme Commission (EC) number are colored light blue. Retrieved from MetaCyC [68].

NTNU

Norwegian University of
Science and Technology