Master's thesis

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

Elin Sandertun Røed

Predicting Allosteric Regulation at Genome Scale

Master's thesis in Biotechnology
Supervisor: Daniel Machado
Co-supervisor: Elisa Márquez-Zavala
May 2023

NTNU
Norwegian University of
Science and Technology

Elin Sandertun Røed

Predicting Allosteric Regulation at Genome Scale

Master's thesis in Biotechnology
Supervisor: Daniel Machado
Co-supervisor: Elisa Márquez-Zavala
May 2023

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

NTNU
Norwegian University of
Science and Technology

Preface

This Master thesis is conducted under the Department of Biotechnology and Food Science at the Faculty of Natural Sciences, of the Norwegian University of Science and Technology.

For as long as I can remember, I have been captivated by the subjects of natural science. While this personality trait did not make me the most popular kid in school, it did provide me with a lot of motivation and opportunities for the future. Choosing what path to pursue at the university was probably one of the hardest choices I have ever had to make, but today, I am very happy to say that I have no regrets regarding my decision to study biotechnology. Throughout these last five years, I have often found myself fascinated by the wonders of the biological world. Despite all of the unavoidable challenges and frustrating moments faced during numerous courses and assignments, I have found a lot of joy in the work and knowledge this degree has provided me. My time as a student has also uncovered an unexpected interest for the field of informatics, which greatly contributed to my motivation for pursuing a specialty within systems biology. At the end of this era in my life, I suddenly realize how quickly the time went by, and the experiences and lessons from this time are something I will carry with me forever.

There are several people to whom I owe gratitude for the completion of this thesis. Firstly, a special thanks is owed to my supervisor Daniel Machado. Your support and guidance have been essential during the work with this project, and I owe a great deal of my motivation and knowledge to your involvement in my work. I would also like to thank my co-supervisor, Elisa Márquez-Zavala, for helping with the final part of the project. I greatly appreciate all knowledge, advice, and time you both have shared with me. Lastly, I want to thank my friends and family for all support throughout the years, and all teachers and advisors who have shared their knowledge, and spiked my love and curiosity for natural science.

I am happy to present this completed thesis. I hope you enjoy the read!

Elin Sandertun Røed

Elin Sandertun Røed
Trondheim, May 15, 2023

Abstract

Cells orchestrate an incomprehensible number of pathways and reactions to stay alive. In order to ensure the correct production of reactants for the current conditions, the cell is dependent on protein-activity regulation. This regulation is often conducted transcriptionally, in which the activity of the protein is determined by its abundance as set by the rate of gene expression. However, there also exist several posttranscriptional modes of regulation, including a mechanism named allostery. Allosteric regulation is the reversible interaction between a protein and a metabolite which induces a shift in the protein's activity. This shift is the result of an altered affinity of the protein for another molecule caused by conformational changes following binding of the metabolite, often termed the effector molecule, to the allosteric site.

While the mechanisms of transcriptional regulation are thoroughly researched and documented, the case of allostery is somewhat different. Historically, posttranscriptional regulation has been considered less important in the grand scheme of protein-activity modulation. This is represented in the poor documentation of allostery, and in the lack of an established, systematic approach to study and detect such interactions. Despite the increased development of molecular biology tools, discovering allosteric interactions is typically very laborious and resource-demanding work, requiring substantial amounts of time, equipment, and previous knowledge. The recent increased understanding of protein-metabolite interactions as essential modes of metabolic regulation has, however, motivated the research community to find more appropriate, efficient ways of studying this phenomenon. In this context, the use of computational power to uncover the secrets of the immense interaction space is especially appealing.

The aim of this thesis is to evaluate the potential of predicting allosteric interactions from genome sequences, as represented by the sequence and structure of proteins. To achieve this aim, data on protein-activating and -inhibiting metabolic interactions was retrieved from the BRENDA database and assembled into a standardized dataset. This dataset consists of 32 535 organism-specific interactions among 3 097 proteins and 1 002 metabolites, and displays a trend of biased regulation towards central pathways of carbon metabolism. Annotation of common interactions to a phylogenetic tree revealed both a taxonomy-wise conservation and lacking documentation of allostery. In order to predict interactions from protein structure, data on protein sequence and structural features annotated to the proteins of the assembled database were downloaded. These features included eight different protein classifiers: active sites, binding sites, conserved sites, domains, families, homologous superfamilies, PTMs, and repeats. The protein features were associated with the interactions of the assembled database, and associations were further quantified through Fisher's exact tests using an odds ratio of 10 and an adjusted p-value less than 0.05 as the significant threshold values.

The enrichment analysis identified in total 32 276 statistically significant associations. The feature types family and domain were found to be most important for the prediction of metabolite-interactions, and assessing a subgroup of the highly and exclusively associated features and interactions revealed that the approach identified several biologically justifiable connections. Extending the phylogenetic tree with predicted interactions further confirmed the validity of the approach, but also caused the identification of a few false-positive predictions. Despite this inaccuracy, the aim of the thesis was achieved: association of protein features with metabolite-interactions demonstrates that protein sequence and structure, and thereby genome sequence, has the potential for being used as a predictor of allosteric interactions.

Sammendrag

Celler administrerer et ubegripelig antall reaksjonsveier og reaksjoner for å holde seg i live. For å sikre korrekt produksjon av reaktanter for de aktuelle forholdene er cellen avhengig av å regulere proteinaktivitet. Denne reguleringen utføres ofte transkripsjonelt, der aktiviteten til proteinet bestemmes av dets konsentrasjon som er kontrollert av hastigheten på genuttrykk. Imidlertid eksisterer det også flere posttranskripsjonelle reguleringsmåter, inkludert en mekanisme kalt allosterisk regulering. Allosterisk regulering er den reversible interaksjonen mellom et protein og en metabolitt som induserer et skifte i proteinets aktivitet. Dette skiftet er resultatet av en endret proteinaffinitet for et annet molekyl som forårsakes av konformasjonsendringer etter binding av metabolitten, ofte kalt effektormolekylet, til det allosteriske setet.

Mens mekanismene for transkripsjonell regulering er grundig undersøkt og dokumentert, er tilfellet med allosterisk regulering noe annerledes. Historisk sett har posttranskripsjonell regulering blitt ansett som mindre viktig i proteinaktivitetmodulering. Dette er representert i den mangelfulle dokumentasjonen av allosterisk regulering, og i mangelen på en etablert, systematisk tilnærming til å studere slike interaksjoner. Til tross for økt utvikling av molekylærbiologiske verktøy, er det å oppdage allosteriske interaksjoner typisk svært arbeids- og ressurskrevende ved at det krever betydelige mengder tid, utstyr og tidligere kunnskap. Protein-metabolitt interaksjoner har imidlertid nylig blitt ansett som essensielle metabolske reguleringsmekanismer, noe som har motivert forskningsmiljøet til å finne mer hensiktsmessige og effektive måter å studere dette fenomenet på. I denne sammenhengen er bruken av beregningskraft for å avdekke hemmelighetene til det enorme interaksjonsomfanget spesielt tiltalende.

Målet med denne oppgaven er å evaluere potensialet for å forutse allosteriske interaksjoner fra genomsekvenser, som representert ved sekvensen og strukturen til proteiner. For å oppnå dette målet ble data om proteinaktiverende og -hemmende metabolske interaksjoner hentet fra databasen BRENDA og satt sammen til et standardisert datasett. Dette datsettet består av 32 535 organismespesifikke interaksjoner mellom 3 097 proteiner og 1 002 metabolitter, og viser en trend av partisk regulering mot sentrale reaksjonsveier i karbonmetabolismen. Annotering av populære interaksjoner til et fylogenetisk tre avslørte både en taksonomisk bevaring og manglende dokumentasjon av allosterisk regulering. For å forutse interaksjoner fra proteinstruktur ble data om sekvensensielle og strukturelle trekk annotert til proteinene i databasen lastet ned. Disse trekkene inkluderte åtte forskjellige proteinklassifiserere: aktive seter, bindingsseter, konserverte seter, domener, familier, homologe superfamilier, PTMer, og repetisjoner. Disse proteintrekkene ble assosiert med interaksjonene i den sammensatte databasen, og assosiasjoner ble ytterligere kvantifisert via Fishers eksakte tester der en odds-ratio på 10 og justert p-verdi mindre enn 0,05 ble brukt som signifikante terskelverdier.

Anrikningsanalysen identifiserte totalt 32 276 statistisk signifikante assosiasjoner. Familie og domene var de proteintrekkene som ble funnet til å være viktigst for å forutse metabolitt interaksjoner, og undersøkelse av en undergruppe av de sterkt og ekslusivt assosierte trekkene og interaksjonene viste at analysen identifiserte flere biologisk relevante forbindelser. Utvidelse av det fylogenetiske treet med foreslåtte interaksjoner bekreftet ytterligere gyldigheten av tilnærmingen, men forårsaket også identifiseringen av noen få falske positive konklusjoner. Til tross for denne unøyaktigheten ble målet med prosjektet oppnådd: assosiasjonen av proteintrekk med metabolitt interaksjoner viser at proteinsekvens og struktur, og dermed genomsekvens, har potensial som base for å forutse allosteriske interaksjoner.

Contents

List of Figures

List of Tables

Abbreviations

(c)AMP	(cyclic) adenosine monophosphate
(m/t)RNA	(messenger/transfer)-ribonucleic acid
ADP	adenosine diphosphate
ASD	Allosteric Database
ATP	adenosine triphosphate
BRENDA	Braunschweig Enzyme Database
CETSA	cellular thermal shift assay
ChEBI	Chemical Entities of Biological Interest
DARTS	drug affinity responsive target stability
DNA	deoxyribonucleic acid
DRaCALA	differential radical capillary action of ligand assay
EC	Enzyme Commission
FDR	false discovery rate
GSH	glutathione
GTP	guanosine triphosphate
InChI	International Chemical Identifier
iTOL	Interactive Tree Of Life
MI-DAS	mass spectrometry integrated with equilibrium dialysis
NAD(H)	nicotinamide adenine dinucleotide
NADP	nicotinamide adenine dinucleotide phosphate
NCBI	National Library of Medicine
NMR	nuclear magnetic resonance
OR	odds ratio
PLP	pyridoxal phosphate
PMI	protein-metabolite interaction
PTM	post-translational modification
SMRN	small-molecule regulatory network
SPROX	stability of proteins from rates of oxidation
UniProt(KB)	The Universal Protein Resource (Knowledgebase)

1 Introduction

As a biologist, it is impossible to not be familiar with ”The Central Dogma”, also referred to as ”The Secret of Life”. This concept derives from a lecture given by Francis Crick in 1957 [1], and today the definition of the central dogma is usually given as the following:

$$DNA \rightarrow RNA \rightarrow Protein$$

This concept has usually been considered the most significant part of metabolic control, implying that the level of metabolic flux is decided by the amount of available protein, which again is determined by the level of gene expression. However, when defining the central dogma, this conception is not what Crick had in mind [1]. In fact, his definition of the central dogma states ”Once information has got into a protein it can’t get out again” [1], which refers to the phenomenon where once information has gone from deoxyribonucleic acid (DNA) to protein, it can not be reversed back into the genetic code.

Today we know it to be scientifically true that ribonucleic acid (RNA) is enzymatically encoded from DNA in the process named transcription, and that proteins are synthesized through the work of ribosomal proteins utilizing messenger-RNA (mRNA), transfer-RNA (tRNA) and amino acids in the process called translation. We are also aware that for most proteins to be active they require post-translational modifications such as acetylation or the removal of certain structures, and that the activity of many proteins is regulated by the binding of small metabolites, such as feedback inhibition by end products in a pathway. Despite this knowledge of post-translational protein regulation, regulation at the transcriptional level is still considered to be the main mode of metabolic control [2].

However, if the message expressed by Crick in his lecture in 1957 is so misunderstood by the research community today, what if there is actually more to ”the secret of life” than what has been historically anticipated? Perhaps the main mode of metabolic flux regulation does not reside within the scopes of the central dogma, but is in fact executed after the protein has been formed. One such mode of post-translational metabolic regulation, which is now considered ”the second secret of life” [3], is the type of protein-metabolite interaction called allosteric regulation.

This Chapter aims at providing the information necessary to understand the context of and the motivation behind this current work. Firstly, the definition and role of allosteric regulation as a mode of metabolic control will be given and discussed in light of more recent research. The second Section will then address different methods for discovering protein-metabolite interactions, classified by experimental- and computational-based approaches. The third Section is devoted to describing an approach for studying and investigating the network of small-molecule regulatory interactions, that aids in elucidating the current state of allosteric research. Finally, the last Section of this Chapter will provide the motivational defense for this project based on the information provided in the previous Sections, while also describing this project’s aim in view of these incentives.

1.1 The role of allosteric regulation

1.1.1 The definition of allosteric regulation

Allosteric regulation is a type of protein-metabolite interaction (PMI) that regulates the activity of a protein by affecting its affinity for another molecule [4, 5, 6]. The regulated protein has two different conformational states - active and inactive. Which state the protein acquires is determined by binding of a certain metabolite, termed the effector molecule, to a region in the protein that is distant from the active site, termed the allosteric site [7].

Allosteric interactions are reversible regulatory mechanisms that may either increase or decrease a protein's function. While this function is often the catalytic ability of an enzyme [7], all proteins could, in theory, be allosterically regulated [8]. An effector that increases the protein's function is called an allosteric activator, while an effector that decreases its function is called an allosteric inhibitor [7]. Figure 1.1 illustrates different modes of allosteric behavior, including the change in conformational state induced by the binding of an effector molecule to a protein (a), allosteric inhibition (b), and allosteric activation (c) [7].

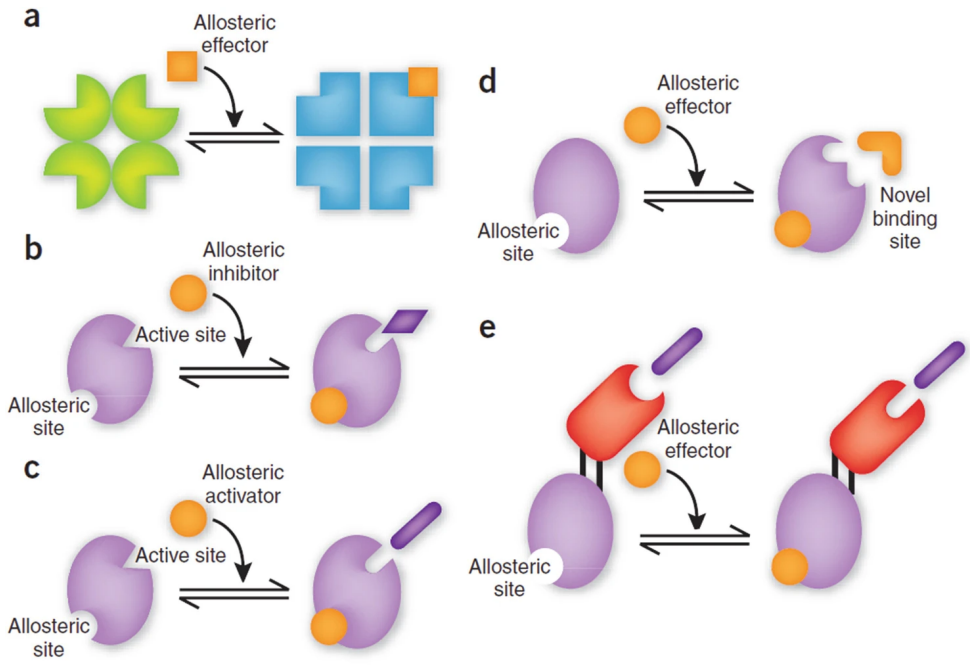


Figure 1.1: Different modes of allosteric behavior: (a) the conformational change induced by ligand-binding, (b) allosteric inhibition, (c) allosteric activation, (d) the introduction of a new binding site induced by ligand-binding, and (e) an allosteric switch consisting of an enzyme fused to an allosteric protein [7].

Figure 1.1 also depicts two other, perhaps more sophisticated, modes of allosteric behavior, namely the introduction of a new binding site (d) and a phenomenon referred to as an allosteric switch (e). As mentioned, the binding of an allosteric effector induces a conformational state in the protein. Even though this usually leads to a change in the active site, causing the protein to become either active or inactive, it might also lead to the formation of a new binding site on the protein. This new binding site might also be bound by a ligand, triggering another conformational shift and change in activity. Furthermore, an enzyme might be fused to an allosterically controlled protein, thereby being under allosteric control via the conformational state of its construct partner. While allosteric control may be divided into the two main categories of ”activation” and ”inhibition”, these two mechanisms of indirect allosteric control are also present in nature [7]. In order to reduce the complexity

of the work conducted in this thesis, however, activation and inhibition are the only two modes of allostery that will be regarded.

1.1.2 The role of allostery in metabolic regulation

As mentioned at the beginning of this introduction, it has long been thought that transcriptional regulation, and thereby enzyme concentration and availability, is responsible for the regulation of metabolism in microbes. The main reason for this belief is the modern view of the central dogma of molecular biology, stating that DNA encodes mRNA that in turn encodes proteins which then execute tasks in the cell [2]. However, several both recent and not-so-recent studies have shown that this type of regulation is insufficient to explain the observed metabolic fluxes [9]. In a study conducted by Chubukov *et.al.*, they found that very few reactions are fully controlled by enzyme concentrations and that there is, in general, an excess level of enzymes available in the cell. They also found that substrate concentration had a negligible effect on most reaction fluxes, leading them to the conclusion that the flux must be controlled by other mechanisms, most likely allosteric regulation by non-substrate metabolites or enzyme modification such as phosphorylation or acetylation [9].

According to Kochanowski *et.al.*, several additional studies have found a mismatch between the relative change in enzyme abundance and the relative change in flux. These findings are further indicators that posttranscriptional modifications and metabolic control such as allosteric interactions play a large role in the regulation of metabolic fluxes compared to transcriptional regulation [10]. The authors further suggest three hypotheses as to why metabolic fluxes are not very sensitive to moderate changes in enzyme levels. Firstly, cells produce an abundance of enzymes as a mode of protection. Insufficient metabolic flux may be more harmful to the cell than spending resources on excess protein production, and the cells therefore maintain higher amounts of protein to protect themselves against unavoidable variations in protein levels. Secondly, a cell may need to change fluxes faster than what can be accomplished by transcriptional regulation. For example, the response to oxidative stress in yeast cells involves allosteric regulation which changes fluxes in seconds, while transcriptional regulation works on the scale of minutes. The utilization of allostery instead of transcriptional control may therefore be a question of life or death for a simple microbe. Finally, the last hypothesis states that a perfect regulatory strategy that always produces an optimal enzyme level is very difficult to design, leading to the generation of an enzyme level that is unnecessarily high [10].

The role of transcriptional regulation in regulating metabolic fluxes, as explained by Kochanowski *et.al.*, is illustrated in Figure 1.2. As can be seen in the graph, the metabolic fluxes (F) are enzyme-limited at low levels of gene expression (enzyme abundance, E), but as the enzyme concentration increases, the change in enzyme abundance has a weaker effect on the measure of metabolic flux. In wild-type microbes, most enzymes are expressed in overabundant levels. This means that the enzyme abundance will in most cases be located within the blue area of the graph in which the enzyme abundance has very little effect on the metabolic flux, indicating the importance of other mechanisms for the regulation of metabolic fluxes [10].

On the other hand, in their article concerning a study of the regulation of fluxes through individual enzymes of the glycolytic pathway in *Saccharomyces cerevisiae*, Daran-Lapujade *et.al.* [2] emphasize that the mismatch observed between different levels of gene-expression might simply be caused by the time delay that transpires between changes at the mRNA level and the changes in protein concentrations and enzyme activities. However, they also

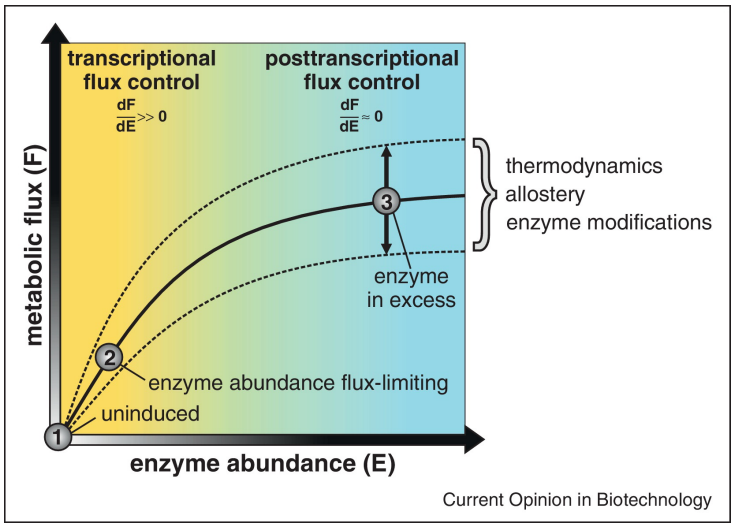


Figure 1.2: Transcriptional regulation in metabolic control, illustrating the role of transcriptional control, represented by enzyme abundance (E), in regulating metabolic flux (F). In wild-type microorganisms, most enzyme abundances are within the blue range [10].

point out that studies utilizing steady-state chemostat cultures, where the cells grow under constant conditions, find poor correlation between mRNA levels, protein concentrations and fluxes, which indicates that time delay cannot be the only responsible factor. The authors further present regulation of gene expression mainly at the posttranscriptional level as a plausible explanation for this observed relationship [2].

In their study of the glycolytic pathway in *S. cerevisiae*, Daran-Lapujade *et.al.* did in fact find that most of the gene-expression regulation is practiced at the protein synthesis-degradation level and posttranslational level, rather than at the mRNA level [2]. Their findings also include the identification of metabolic regulation, namely the regulation of enzymes by interactions with metabolic compounds such as substrates, products, or allosteric effectors, as a ”substantial component of almost all regulation observed”. This means that the transcriptional regulation of yeast glycolysis is less extensive than anticipated and that this metabolic pathway is controlled by several regulatory mechanisms rather than one simple regulation strategy. As glycolysis is a very central metabolic process, they further conclude that this might be the case for other pathways, organisms, and conditions as well, and states that identification of more important regulatory mechanisms deserves to be prioritized [2].

1.2 Methods for discovering allosteric interactions

Historically, allosteric regulators have mainly been discovered by what is referred to as random events, before they were later verified experimentally. While the development of molecular biology tools has resulted in routine methods for gene and protein discovery, this is yet to happen for allosteric effectors. However, in the later years, further attention has been designated to finding a systematic way of discovering allosteric interactions. This has caused an increase in the use of high-throughput chemical screens in the search for specific enzyme activators or inactivators [8], and also in the development of computational methods for this same purpose. This Section is designated to describe some of the experimental and computational approaches for discovering allosteric interactions that have been developed and applied in the not-so-distant past.

1.2.1 Experimental approaches

There are several experimental ways of discovering and studying allosteric interactions. Some frequently applied methods over the last few decades include X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, fluorescence resonance energy transfer (FRET), and hydrogen-deuterium exchange mass spectrometry (HDXMS). While X-ray crystallography provides detailed structural information about the protein, it cannot be used for studying dynamical properties. NMR spectroscopy is, however, better at handling the transient conformations of allostery, and labeling proteins, such as with FRET and HDXMS, can also be applied in order to track protein conformational changes [6]. These technologies are only examples of methods that have been used to detect protein-metabolite interactions, and there exist several additional approaches that utilize either these exact methods or similar concepts for the detection of PMIs.

In the later years, approaches for discovering new allosteric interactions are typically divided into metabolite-centric and protein-centric methods. The metabolite-centric approaches aim at identifying protein targets for specific metabolites, while the protein-centric methods aim at identifying interacting metabolites for specific proteins. The two following Subsections will describe a few metabolite- and protein-centric methods that have been successfully used for discovering allosteric interactions, while also mentioning some of their associated challenges.

The last part of this Subsection is devoted to an innovative approach that may not necessarily be classified as either metabolite- or protein-centric. Instead of focusing on a small subgroup of either proteins or metabolites and a larger group of the opposing variable, this ligand-detected NMR-approach aims at mapping interactions within a small subnetwork of enzymes and metabolic compounds, causing a reduction in the possible interaction space compared to other existing methods.

1.2.1.1 Metabolite-centric approaches

This Subsection will describe a few experimental methods that identify protein targets of small molecules. The approaches in focus are DARTS, SPROX, CETSA, and the utilization of functionalized small molecules.

The workflows of DARTS, SPROX, and CETSA are shown in Figure 1.3 [11]. While these three approaches have different workflows, they all share the same concept; two mixtures of proteins, where one is mixed with a ligand and one is not, are examined in order to distinguish between the enzymes that interact and those that do not interact with the ligand in question. The main distinction between the methods is how the bound proteins are identified from the unbound, which will be further explained below.

DARTS, drug affinity responsive target stability, detects the increased proteolysis-resistance of target proteins that is induced by the binding of a small molecule. The experiment is performed by treating cell lysate with the compound of interest and proteases before mass spectrometry is used to identify the bound and proteolysis-protected proteins present in protein bands on SDS-gels [12]. While DARTS has been used to identify the protein targets of cancer drugs and the role of protein-metabolite interactions in aging, it does have the limitation of poor identification of low-abundance proteins due to them not being clearly visible on the gel [11].

SPROX, stability of proteins from rates of oxidation, is another lysate-based approach in which proteins are exposed to different concentrations of a denaturant and an oxidizing agent. When bound to a ligand, the proteins will be protected against denaturation, and the

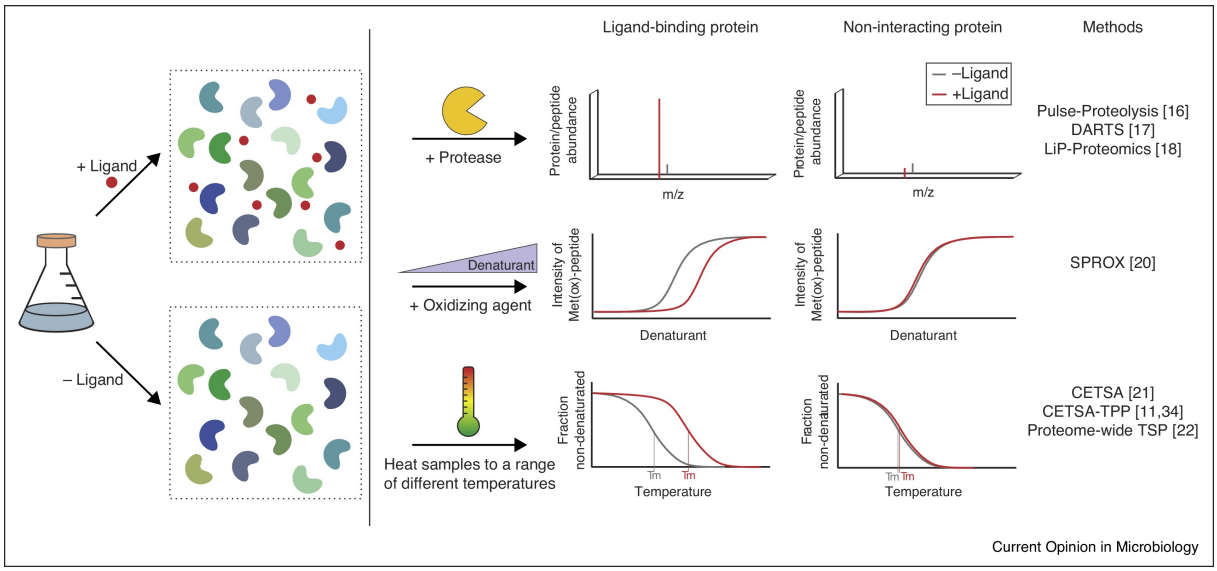


Figure 1.3: Outline of DARTS, SPROX and CETSA, used for identifying protein targets of small molecules. Two mixtures of enzymes, one ligand-treated and one untreated, are treated in different ways that allow for the distinction between ligand-bound and not ligand-bound proteins [11].

ligand-bound proteins can therefore be identified from the non-interacting proteins [11, 13]. The limitations of SPROX include its limited application in detecting low-affinity interactions and therefore the identification of weak regulatory protein-metabolite interactions [11].

CETSA, cellular thermal shift assay, is an approach that can be used for monitoring the stability of proteins in vivo [13]. The method exploits the change in thermal stability that is induced by ligand-binding, and identifies target proteins by comparing the melting curves of proteins in the presence and absence of a metabolite. CETSA has been used to identify protein targets of several compounds, including ATP, but it is known to yield many false negatives as not all ligands affect the proteins' stability [11].

Lastly, functionalized small molecules can also be used to identify protein-metabolite interactions, as illustrated in Figure 1.4. The functionalized ligands are crosslinked to the proteins, tagged with a purification tag, and affinity purification is used to isolate the interacting proteins which are subsequently identified with mass spectrometry [11]. There are several options for altering the chemical functionality of such functionalized compounds that can enable the identification of protein-metabolite interactions in different ways and under different circumstances. However, a limitation of this approach, which is valid for all of the experimental approaches mentioned above, is that they are limited to compounds that are chemically stable throughout the experiment [11].

1.2.1.2 Protein-centric approaches

While there are several available approaches for identifying the protein targets of specific metabolites, there, as mentioned, also exist methods for identifying interacting metabolites for specific proteins. These approaches include DRaCALA and MI-DAS. While these methods have been successfully utilized for the identification of protein-metabolite interactions, there are a few general challenges with the protein-centric methods that cause them to be less frequently applied. These challenges include a lower throughput than the metabolite-centric methods, and that the approaches typically require purified proteins [11].

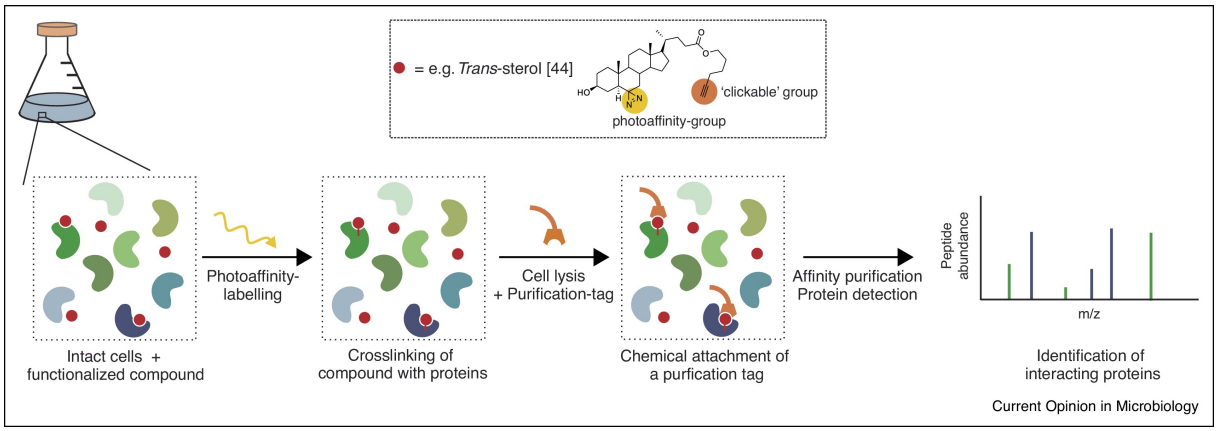


Figure 1.4: Identifying protein-metabolite interactions by using functionalized small molecules, applying crosslinking, affinity purification and mass spectrometry [11].

DRaCALA, differential radical capillary action of ligand assay, shown in Figure 1.5(A), is based on the ability of a nitrocellulose membrane to sequester proteins, and their potentially bound ligand, from unbound ligands. Spotting a mixture of proteins and radiolabeled metabolites onto such a membrane causes the protein-metabolite complexes to be immobilized at the site of interaction, while free ligand is mobilized and distributed together with the liquid phase. The membranes are then quantitated using a Phosphorimager and the fraction bound for the proteins is calculated using equations shown in Figure 1.5(B) [14]. While DRaCALA is a rapid, high-throughput method that overcomes the limitation of having to use purified proteins, there are some expected challenges associated with this approach. These challenges include the limited availability of radiolabeled ligands and the poor detection of transient interactions. Transiency is quite common for regulatory protein-metabolite interactions [11], and the accuracy of this approach is thus reduced as a consequence of this issue.

A protein-centric method that is supposedly suited to deal with the problem of detecting transient interactions is MI-DAS, mass spectrometry integrated with equilibrium dialysis, which identifies target metabolites using mass spectrometry. In the proof-of-concept experiment for this method, where 5 enzymes were dialyzed against 138 metabolites, 13 novel interactions were discovered. This method does however require large amounts of purified protein which does complicate its application [11].

1.2.1.3 Ligand-detected NMR

The Braunschweig Enzyme Database (BRENDA) reports over 4500 unique, regulatory interactions in *Escherichia coli* metabolism and over 1500 in *Saccharomyces cerevisiae* [11], and there have been performed several large-scale studies on protein-metabolite interactions in which there is almost no overlap [15]. These facts indicate a very large interaction space of proteins and metabolites, and Diether *et.al.* therefore suggest that it might be useful to map the protein-metabolite interactions within a defined subnetwork, rather than on a large scale. They report an NMR approach that allows for the detection of interactions between a set of water-soluble proteins and metabolites, that they used for investigating interactions between 29 enzymes and 55 metabolites of the *E. coli* central carbon metabolism [15].

The workflow of this approach is illustrated in Figure 1.6. The enzymes, tagged pre-experiment, are mixed with the metabolites distributed in four different mixtures. Each combination of protein and metabolite mix is then incubated for several hours before their

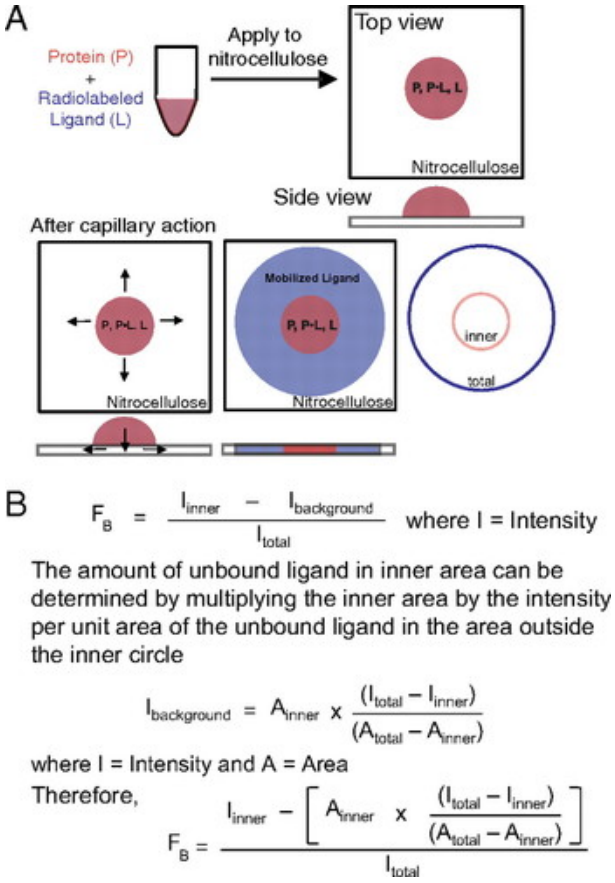


Figure 1.5: The prinicples of the DRaCALA approach, for identifying interacting metabolites for specific proteins. Protein-metabolite complexes are immobilized on a nitrocellulose membrane while unbound ligand is distributed throughout the membrane with the liquid phase (A). Equations are then used to analyze the data for the fraction bound of proteins (B) [14].

NMR spectra are recorded. The result of this experiment was the detection of 98 interactions, of which 40% of the interacting metabolites were predicted to be allosteric effectors due to low chemical similarity to their target's substrate. While 22 of these detected interactions were already known, 76 had not been previously reported. These results demonstrate the potential of using ligand-detected NMR for discovering novel interactions within a defined subnetwork of metabolites and proteins, while also illustrating our lack of knowledge regarding protein-metabolite interactions [15].

Although the ligand-detected NMR approach shows great potential for discovering novel PMIs, there are a few disadvantages that can be associated with this procedure. Firstly, the strength of this method is situated in utilizing a subnetwork of metabolism for studying allosteric regulation. Even though this approach was demonstrated to be successful, it does imply that the method is of a low-throughput and not easily applicable to the entire metabolism of an organism. This circumstance complicates the large-scale application of the current approach, thus weakening its relevance in a systems biology context. Secondly, as with all experimental approaches, this type of method requires quite a lot of resources. Using 29 different enzymes and four mixes of metabolites gives a total of 116 samples to examine, which is demanding both in time, equipment, and funding. If this approach was to be applied on a larger scale of PMIs and for several different organisms, these requirements would increase accordingly, causing the generation of great expenses.

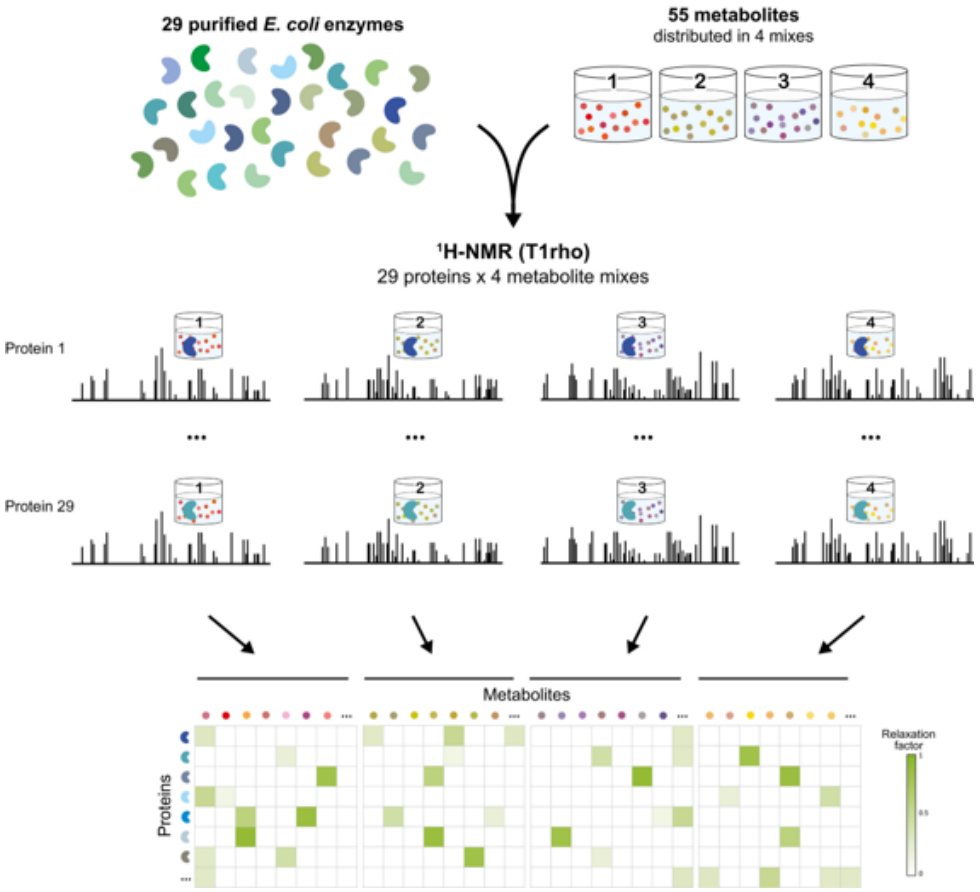


Figure 1.6: The workflow of the ligand-detected NMR approach, for identifying protein-metabolite interactions within a subnetwork of enzymes and metabolites. Pre-tagged enzymes are mixed with four different metabolite mixtures, and the NMR-spectrum of each protein-metabolite combination is recorded post-incubation [15].

1.2.2 Computational approaches

Historically, computational approaches have typically not been used to study allostery alone, but rather in combination with experimental methods. They do however provide powerful tools, for example by allowing for the simulation of protein conformational dynamics and by having a prediction power that can enable the identification of allosteric sites [6]. Additionally, even though experimental approaches have great potential for identifying protein-metabolite interactions, evaluating the in vivo functionality of such interactions can be a tedious process requiring many follow-up experiments. The process of resolving this currently major bottleneck might be assisted by the use of computational power [11], further increasing the research community's initiative in developing new computational approaches for discovering protein-metabolite interactions.

This Subsection will describe both purely computational and combined experimental and computational approaches for identifying and analyzing protein-metabolite interactions. Methods for determining the functionality of PMIs will first be discussed, followed by approaches that facilitate the identification of novel allosteric sites. As computational biology is in focus in this project, special attention will be paid to approaches that do not rely on immediate experimental work, such as the development of computational allosteric services by Lu *et.al.*

1.2.2.1 Determining in vivo functionality computationally

In their review of methods for detecting regulatory protein-metabolite interactions, Diether and Sauer [11] describe two computational approaches for determining the in vivo functionality and mechanisms of protein-metabolite interactions.

The first of these two methods is SIMMER, systematic identification of meaningful metabolic enzyme regulation. This approach assesses whether the experimental measurements from separate reactions can be explained by Michealis-Menten kinetics, or if more complex models that include allosteric interactions are required. Application of the approach resulted in the successful identification of novel allosteric interactions in yeast, but its application to organisms that are not as well characterized is made difficult by its requirement for prior knowledge of certain kinetic parameters [11].

The other method that is described by Diether and Sauer is a combined experimental and computational approach for identifying allosteric protein-metabolite interactions that control enzyme activity, documented by Link *et.al.* in 2013 [4]. The authors highlight the problem of quantifying the in vivo activity of allosteric regulations only being possible by computational modeling, of which all cases rely on a priori knowledge of either one or several of the allosteric interactions. They therefore developed an approach, which is based on fitting data from dynamic metabolomics and ^{13}C isotopic labeling experiments to kinetic models of the same pathway and testing putative allosteric interactions, that doesn't require any prior knowledge of interactions [4].

Applying the approach to investigate how allosteric interactions control the switch between the gluconeogenesis and glycolysis pathways, Link *et.al.* identified the most likely regulatory interactions together with hypotheses of their function. All combinations of allosteric activation and inhibition for the nine irreversible enzymes by the seven metabolites in the model resulted in 126 putative allosteric interactions, of which 17 were already known. One part of their test results suggested that active regulation of the enzyme pair phosphofructokinase and FBPase was necessary for flux reversal in upper glycolysis, and the identified effector metabolites were in fact consistent with data from previous studies. The authors describe their own method as a way of systematically mapping biologically relevant allosteric interactions under certain conditions, but validation of the interactions' functional importance does, however, require experiments with mutant enzymes or approaches that focus on single interactions [4]. Furthermore, even though the approach does not require a priori knowledge of relevant PMIs, it does rely on prior knowledge of kinetic parameters [11].

1.2.2.2 Allosteric site prediction for drug design

As previously mentioned, the binding of an allosteric modulator to a protein may change its functional activity. These changes may alter different phenotypic traits, thus making allosterically regulated proteins potential targets of medical treatment by drugs that are designed to bind their allosteric sites and affect the proteins in the desired way [7]. It has also been established that these types of drugs achieve higher specificity due to the higher selectivity found among allosteric sites, resulting in fewer side effects and lower toxicity compared to ligands binding at the active site [16, 17]. As the improvement of medical treatments is of generally high interest in the research community, there have been many efforts made in the last two decades to identify such allosteric sites.

As experimental approaches for discovering allosteric sites are typically very demanding in terms of time and resources, predicting approaches based on computational methods

are very attractive. One of the first developed predictive methods is COREX, a structure-based algorithm that produces a list of possible protein conformations and their respective probabilities [18]. This probability distribution function can then be examined in terms of the effects of ligands and other chemical or physical properties [18]. COREX has successfully been used to identify already known allosteric sites, to define the communication pathway between regulatory and catalytic sites, and for predicting effector binding [18, 19]. Other, newer methods for studying the dynamical details of allostery include statistical coupling analysis for identifying amino acid residues involved in allosteric signaling within proteins [19, 20], and a structure-based statistical mechanical model that allows for the analysis of allosteric communication energetics [21].

Despite the progress made within the field of allostery, there are still challenges associated with the identification of allosteric sites and mechanisms. These challenges constitute the motivation of Lu and his team in developing several allosteric services that can be used for studying relevant topics [3]. These tools include the Allosteric Database (ASD) (v. 3.0) consisting of data about experimentally confirmed allosteric proteins and modulators [22], the ASBench consisting of datasets of allosteric sites that can be used for the development of computational methods to predict unknown allosteric sites [23], Allosite and AllositePro for the prediction of allosteric sites [24, 16], and Alloscore for predicting binding affinities of allosteric protein-modulator interactions [25].

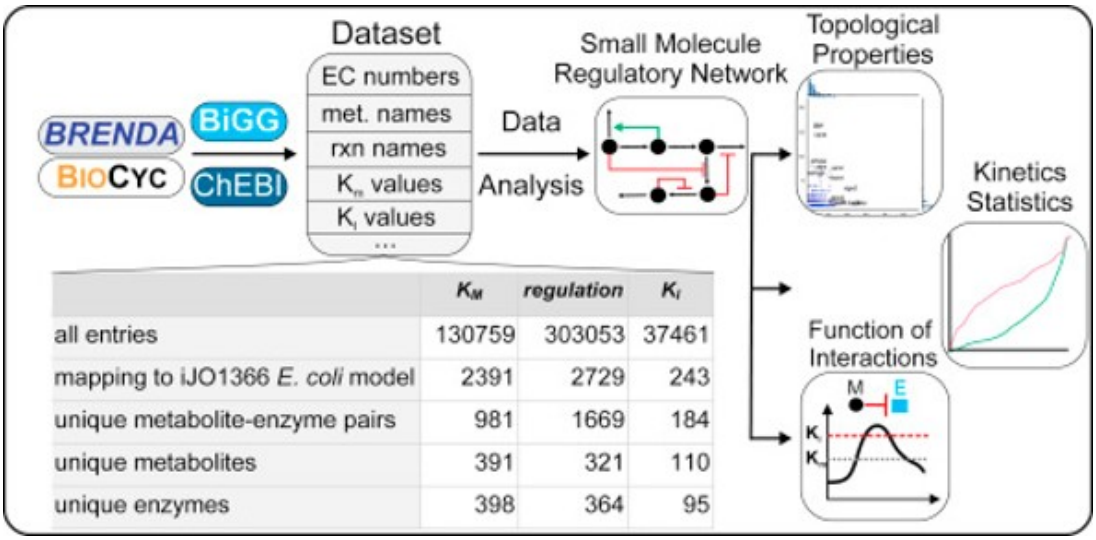
While these tools provided by the Allosteric Database are useful for studying allostery, especially in the case of allosteric drug discovery, there are a few issues that complicate their application for systems biology purposes. For example, while the discovery of novel allosteric sites is useful for drug discovery, many of these sites have no known natural binding effectors [19]. A quick query of the ASD (v. 3.0) [22, 26] revealed that the database in fact contains many inorganic compounds that may not regulate proteins in natural systems, and are thereby not necessarily useful in an in vivo systems biology context. Also, the content of the database is of low diversity organism-wise. Of the reported proteins, 43% belong to humans and 31% to bacterial species, leaving 26% of the proteins to other species [26]. This complicates the application of the Allosteric Database in studies of less popular organisms or larger groups of species. Another complication, that is also relevant for the other approaches described in this Subsection, is that these tools appear to be most suitable for the study of singular interactions or pathways. For example, Allosite and AllositePro seem to be adapted to identifying allosteric sites in only one protein at a time [24, 16], and although Alloscore does allow the user to upload multiple ligands for the assessment of binding affinity between them and the protein in question [25], the assessment of unknown interactions still require the putative effector molecule to be present in the regarded dataset. These matters complicate the application of these tools for discovering novel interactions on a larger scale, possibly making them just as resource-demanding as previously described experimental approaches, especially when considering that the identified interactions may require experimental validation.

1.3 Related work

As a part of the current bottleneck that is validation of the functional in vivo relevance of protein-metabolite interactions, technical limitations have also stood in the way of mapping small-molecule-enzyme regulatory interactions on a genome-scale. These limitations were the motivation behind the efforts of Reznik *et.al.* [27] in developing a framework for reconstructing and analyzing the small-molecule regulatory network (SMRN), which is an alternative strategy to study small-molecule regulation. The group used *Escherichia coli* as their model organism and gathered interaction data from the BRENDA and BioCyc

databases. For every Enzyme Commission (EC) number they obtained a list of possible regulating small molecules, the type of interaction (activation/inhibition), and the interaction constant (K_I), and then mapped this data onto a genome-scale metabolic reconstruction of *E. coli* [27].

The pipeline developed by the group for obtaining small-molecule regulation data and using computational tools for integrating it with a genome-scale metabolic model is shown in Figure 1.7 [27]. This computational framework is freely available on GitHub and can be used to reconstruct and analyze the SMRN of other organisms, given that enough data is available at the necessary databases [27].



	K_M	regulation	K_I
all entries	130759	303053	37461
mapping to iJO1366 *E. coli* model	2391	2729	243
unique metabolite-enzyme pairs	981	1669	184
unique metabolites	391	321	110
unique enzymes	398	364	95

Figure 1.7: The framework for reconstructing and analysing the SMRN, by Reznik *et.al.* [27]. Databases are mined for regulatory interactions that are subsequently mapped onto a genome-scale metabolic reconstruction of the organism in question, *E. coli* in the current study. The resulting network can be further analyzed to elucidate the role of regulatory interactions or other properties.

The data mining resulted in 1669 unique regulatory interactions among 321 unique metabolites and 364 unique enzymes, of which 84% were inhibitions. The most frequent regulators were ATP, AMP, ADP, PI, PPI, NADPH, GTP, cysteine, pyruvate, and phosphoenolpyruvate (PEP), among which ATP was the most frequent regulatory metabolite with participation in 57 different reactions. Metal ions also constituted a significant fraction of the group. Statistics on the *E. coli* metabolites (A) and enzymatic reactions (B) and scatterplots of activating and inhibiting interactions in which each metabolite (C) and reaction (D) participates are shown in Figure 1.8 [27].

When classifying the reactions according to functional metabolic subsystem, Reznik *et.al.* found that most interactions targeted cofactor biosynthesis, nucleotide salvage pathway, arginine/proline metabolism, alternate carbon metabolism, nucleotide biosynthesis, cell envelope biosynthesis, and glycolysis/gluconeogenesis. Some of the other high-flux pathways, such as the citric acid cycle and pentose phosphate pathway, were regulated by relatively few metabolites, which supports evidence that they are mostly regulated transcriptionally [27].

It is very well known that the central carbon metabolism (CCM), which provides energy and biosynthetic precursors to the cell, is highly regulated, both transcriptionally, post-translationally, and allosterically. Our understanding of this regulation is, however, still incomplete. The reconstruction of the CCM SMRN by Reznik *et.al.*, depicted in Figure 1.9, shows that the majority of the CCM enzymes are regulated and that they interact with more

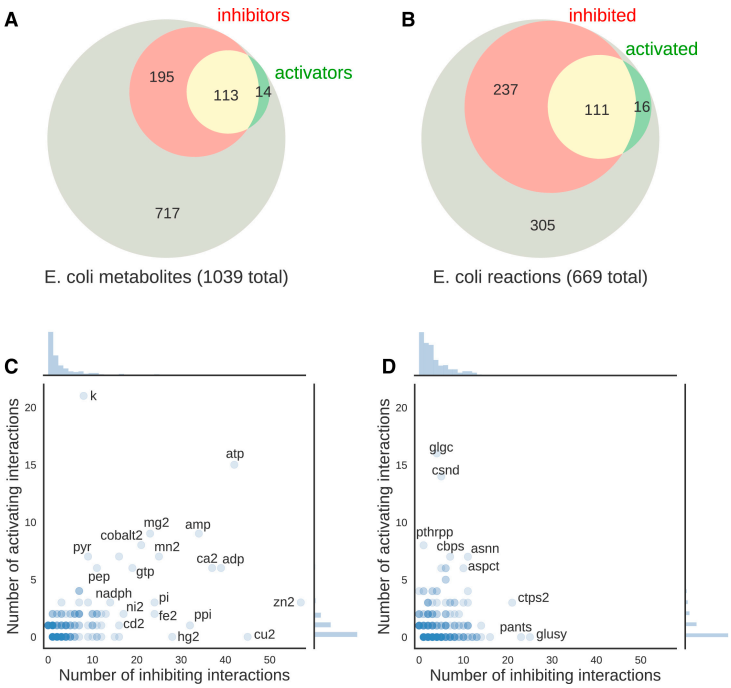


Figure 1.8: Statistics on the components of the SMRN constructed by Reznik *et.al.* for *E. coli*: number of activating and inhibiting interactions among the overall groups of metabolites (A) and enzymatic reactions (B), and scatterplots of activating and inhibiting interactions in which each metabolite (C) and reaction (D) participates [27].

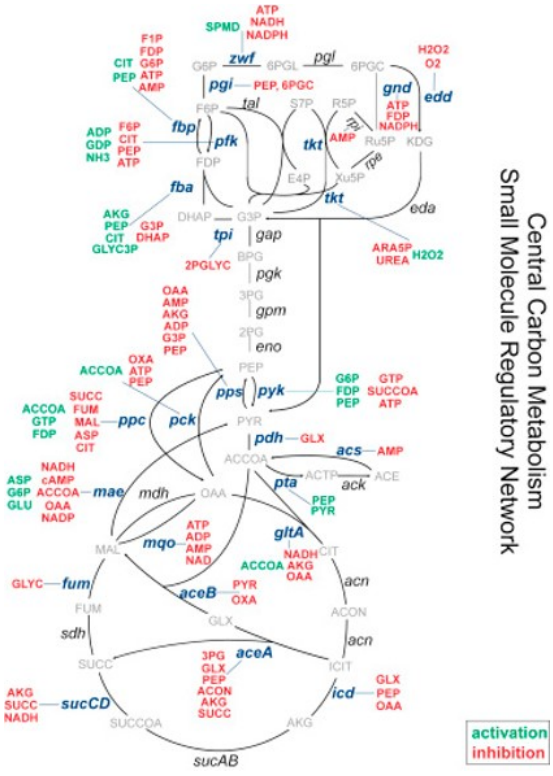


Figure 1.9: The small-molecule regulatory network of *E. coli* central carbon metabolism, assembled by Reznik *et.al.* [27].

small molecules than the average metabolic enzyme. Especially the enzymes of upper and terminal glycolysis and those branching the citric acid cycle are very heavily regulated. This structure suggests a non-random distribution of regulatory interactions, which can be explained by the conservation of resources accomplished by feedback inhibition [27].

The authors highlight that having a proper understanding of enzyme activation and inhibition is important for improving the accuracy of metabolic models, as well as it can also facilitate the engineering of new metabolic pathways, and improve our knowledge of how and why metabolic abnormalities affect health and disease. This alternative approach could cover a larger proportion of metabolism than previously developed approaches, could strengthen the evidence for poorly documented interactions, and also help uncover the role of regulatory metabolites and enzymes in relation to other processes that constitute metabolic control [27]. While it is unsure whether this approach can be used to study an organism on a genome scale, these factors are strong motives for conducting further research on allosteric interaction networks. Incentives such as these are the subject of the following Section.

1.4 Motivation and project aim

The studies described in Subsection 1.1.2 highlight the role of allosteric regulation as an essential mechanism to maintain metabolic homeostasis by regulating a large selection of cellular processes [4, 11]. However, as was demonstrated by the work of Reznik *et.al.* described in the previous Section [27], many interactions between proteins and metabolites are still missing from existing metabolic maps [4]. The work by Machado *et.al.*, illustrated in Figure 1.10, is another example of an extended metabolic model which shows that including allosteric interactions in the core metabolism model of *E. coli* highly increases its complexity [28], further strengthening the concept of allostery as a key phenomenon to understanding biological systems and diseases [6].

Due to the importance of allostery in metabolic regulation, there are several advantages associated with increased allosteric knowledge. For example, as described by Daran-Lapujade *et.al.* [2], many attempts have been made at increasing the fermentative capacity of *Saccharomyces cerevisiae* via genetic engineering, but so far all have failed. In their analysis of the glycolytic enzyme regulation in *S. cerevisiae*, Daran-Lapujade *et.al.* found that the regulation of glycolysis is in fact not mainly exerted at the level of gene expression, but that it rather resides in the interactions of these enzymes with their environment. This discovery may provide clarification as to why such engineering attempts have previously failed, and also suggests that metabolic engineers face a greater challenge to achieve the goal of enhanced fermentative capacity in yeast than what has been previously anticipated [2]. As was also highlighted by Reznik *et.al.*, uncovering the secrets of the protein-metabolite interactome can thus contribute to more realistic and accurate metabolic models that not only have purposes relevant to metabolic engineering, but to all domains of systems biology.

While these hypothetical advantages of increased allosteric knowledge motivate the efforts of the research community to reveal the secrets of the allosteric interactome, there are several obstacles standing in the way of such discoveries happening at a satisfyingly high pace. Some of these obstacles are method-specific, as those described in Section 1.2, while some are relevant to the general case of an allosteric study.

One of the main challenges associated with studying allosteric interactions is the vast space of possible protein-metabolite interactions. Any metabolite could in theory interact with any protein, and molecules also affect each others' ability to bind proteins [29]. Pairing every small molecule in combination with each other and with every protein results in an infinite

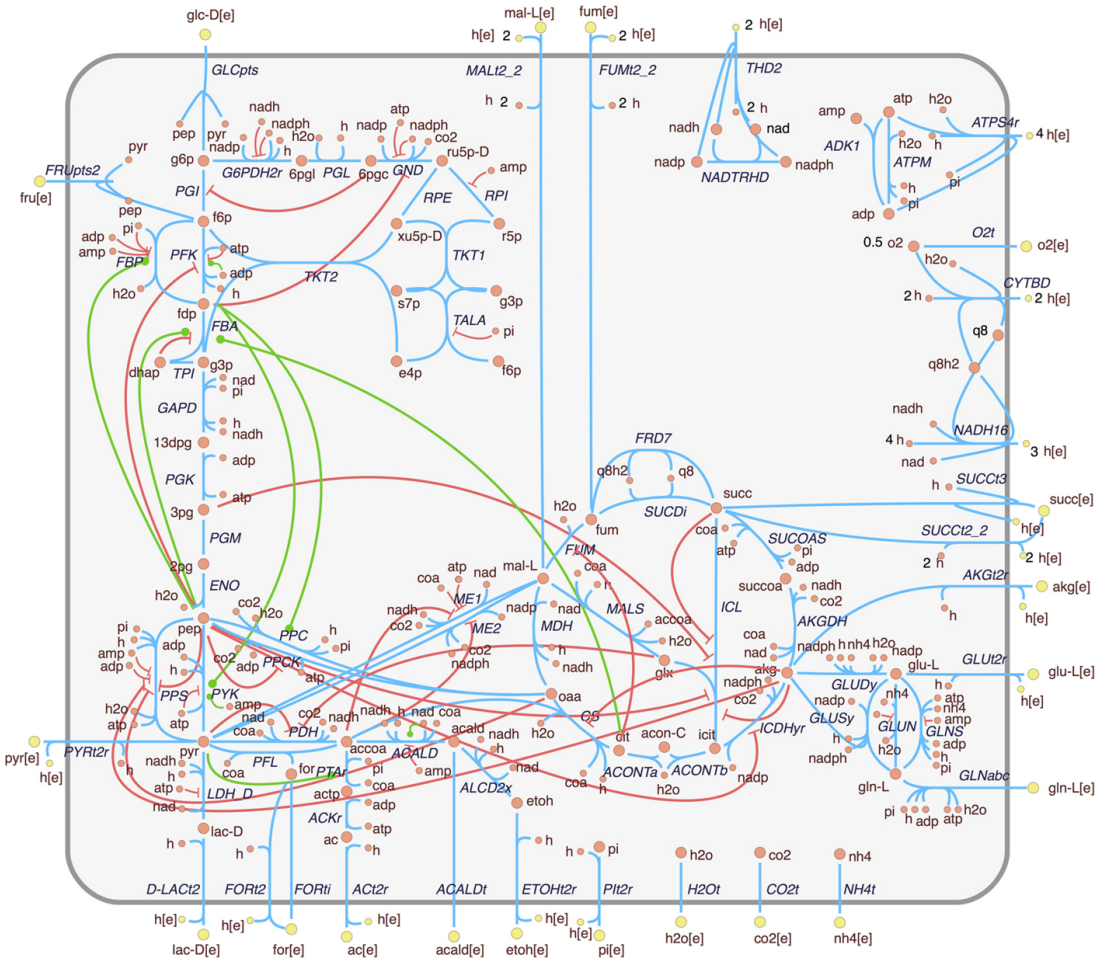


Figure 1.10: Model of the *E. coli* core metabolism, extended with allosteric interactions. Activating and inhibiting interactions are illustrated by green and red edges, respectively [28].

number of putative interactions, whose evaluation is practically impossible to execute. Additionally, all protein surfaces are potential allosteric sites, and with both enzymes and non-catalytic proteins as possible allosteric targets [7], this immense interaction space poses a challenge that cannot be tackled by either experimental or computational approaches that are not designed for the purpose of large-scale studies.

Two other major problems in discovering allosteric effectors that are more specific to the case of experimental detection, are low-affinity interactions and unknown chemical composition of the interacting metabolites [8]. Since many of the molecules that can be sensed by allosteric proteins are normally present in high concentrations, the interaction must be of low affinity for the proteins to sense small changes in the metabolites. However, as mentioned in relation to the approaches described in Subsection 1.2.1, this type of low-affinity interaction is difficult to discover experimentally. Additionally, unlike active-site structures, allosteric sites are not typically conserved between proteins [8]. While this does aid in the application of allosteric effectors as therapeutic drugs [16, 17], it also implies that there are no constraints on the chemistry of the effector molecules [8], thereby providing no contribution to the downsizing of the broad allosteric interaction space.

The problems related to the discovery of allosteric interactions are especially relevant to the application of experimental approaches, thereby encouraging the use of computational methods instead. The approaches described in Subsection 1.2.2 provide evidence of the potential for studying allostery computationally, and as demonstrated by the work of Lu *et.al.*

[3], the prediction power provided by computational methods enables the identification of allosteric sites from protein sequence. Even though these technological developments facilitate the identification of novel protein-metabolite interactions, many existing approaches seem to be adapted to the prediction of sites in and interactions involving only single proteins. While this may be useful for the study of individual reactions and smaller pathways, systems biologists are typically interested in information on a larger scale, preferably spanning the entire genomic content. So far, most systematic studies of allosteric interactions are limited to either central carbon metabolism or other metabolic subsystems, and to our knowledge, there have been no efforts to thoroughly map the PMIs of either entire organisms or universally.

With these motivational factors in mind, this project will combine knowledge from the fields of biochemistry and bioinformatics in order to evaluate the potential of predicting allosteric interactions from genome sequences. By utilizing data on known protein-metabolite interactions, connections will be drawn between two factors: features annotated to the protein that is subject of regulation, and the metabolite responsible for the regulating behavior with the associated mode of regulation.

Documented protein-metabolite interactions, both activating and inhibiting, will be collected for all available species. Protein structural and sequence features, such as domains and conserved sites, for the enzymes participating in these interactions, will then be retrieved as represented by their respective InterPro IDs [30]. Potential patterns between protein features and protein-metabolite interactions may then be used to deduce the presence of the same interactions in organisms displaying the same genomic content. Possible uses of these predictions include the improved accuracy of systems biology models as a result of implementing currently undocumented interactions, increased competence on topics related to metabolic engineering, and the discovery of novel drug targets.

2 Methods

This Chapter describes the work conducted in this thesis, whose workflow is illustrated in Figure 2.1. The first Section of this Chapter describes part 1 of the diagram, which consists of downloading data on allosteric interactions from the BRENDA database [31], processing this information, and performing various statistical analyses in order to evaluate the general knowledge of allostery. The second Section regards part 2 of the diagram, where the prepared data from part 1 is continuously exploited in order to associate allosteric interactions with protein families, domains, and sequence features retrieved from The Universal Protein Resource (UniProt) [32]. The topics described include how enrichment analysis was conducted, and the assessment of evidences of statistically associated features and interactions.

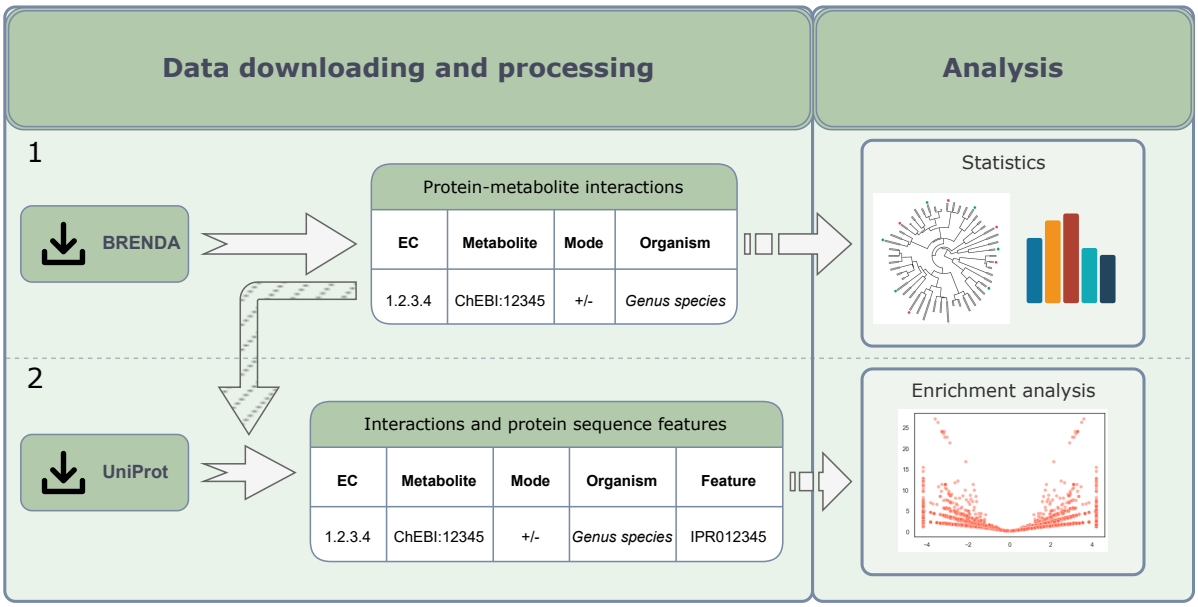


Figure 2.1: Workflow for this thesis.

The code utilized for the completion of these tasks is available at the GitHub repository created for this project, at `https://github.com/elinsroed/predicting-allostery`. The scripts are written in Python (v. 3.8.6) using Jupyter Notebook (v. 6.4.12) [33], while Microsoft Excel (v. 2208) [34] and Python's Pandas package (v. 1.2.2) [35, 36] are utilized to organize the data. Matplotlib (v. 3.6.3) [37] and Seaborn (v. 0.12.0) [38] are used for the plotting of most figures. An overview and description of the scripts and datafiles associated with the current work are given in Table A1 (Appendix A, p. 89).

2.1 Creating an allosteric interactions database

The first main part of the work consisted of establishing a knowledge database of allosteric interactions. This work consisted of downloading allosteric interaction data from the publicly available database BRENDA [31], standardizing and filtering that data in order to create an organized datafile consisting of the interactions that are of the highest interest, and then analyzing this data with the purpose of gaining a better understanding of our allosteric knowledge.

2.1.1 Data download and cleaning

The source of allosteric data utilized in this project is the BRENDA enzyme database [31]. BRENDA classifies enzyme information with respect to the EC nomenclature, which groups enzymes based on their performance in either the same or related enzymatic functions [31, 39]. The raw dataset of activating compounds and inhibitors from BRENDA was downloaded by searching the database for ”Activating Compounds” and ”Inhibitors”, respectively. More specifically, the compounds were accessed by utilizing the database’s ”Classic view” option, and searching within the ”Activating Compounds” and ”Inhibitors” groups under ”Reaction & Specificity”. The searches were conducted without specifying parameters in order to retrieve all available information, and the box for showing organism-specific information was ticked off in order to include the name of the organism for which the interaction was documented. The retrieved information included the interactions’ respective EC number, recommended name (protein), activating/inhibiting compound, commentary, organism, and primary accession number. The complete results were then downloaded as a CSV file and imported into Microsoft Excel, resulting in the file ”Allosteric_interactions_BRENDA.xlsx” (Supplementary information 1, App. A, p. 89) containing separate spreadsheets for the activators and inhibitors.

Upon viewing the downloaded information, it became evident that processing of the data was necessary before further analysis could be performed. The addressed issues include lacking standardization of metabolite and organism ID, and the presence of many unwanted entries, such as metabolites listed as ”additional information”, interactions documented in viruses, and interactions involving inorganic and extracellular compounds. The following two Subsections will describe the standardization and filtering of the allosteric data that was performed. The notebook used for conducting this work, ”BRENDA_data.ipynb”, can be found in Supplementary information 6, while all utilized datafiles are available in Supplementary information 1 and 2 (App. A, p. 89).

2.1.1.1 Standardizing metabolite and organism IDs

A preliminary look at the data downloaded from BRENDA made it apparent that metabolites were referred to by many different synonyms. For example, isopropanol, with Chemical Entities of Biological Interest (ChEBI) ID 17824 [40], is in addition to isopropanol also referred to as 2-propanol, propan-2-ol, 2-hydroxypropane, and isopropyl alcohol, to mention a few. In order to ensure accurate analysis, these metabolite names were standardized by ChEBI ID. ChEBI is a database and ontology of biologically interesting chemical entities that is frequently used as a source of unique identifiers for compounds [40]. The standardization of metabolites was performed using a file containing the name, ChEBI ID, and International Chemical Identifier (InChI) [41] string for all metabolites documented in BRENDA. The file, named ”brenda_compounds.tsv”, was downloaded by conducting an empty search for ligands in the BRENDA database under ”Classic view” - ”Reaction & Specificity” - ”Ligands” [31]. The metabolite names were mapped to ChEBI ID which was further used as the metabolites’ identifier in the remaining work.

Furthermore, some of the organism names included strain names and other details that were unnecessary for the purposes of this current work. The organism names were therefore standardized by limiting the name length to two, causing different strains and subspecies of the same organism to be classified together. This was done on basis of the assumption that there is no strain-level variation for the presence of allosteric regulation.

2.1.1.2 Filtering the data

Removing illogical entries

The data from BRENDA contained several entries that were illogical for the purpose of this work. These entries included interactions where the metabolites were listed as ”additional information” rather than as actual metabolites, and interactions documented with different viruses as the originating organism. The entries with ”additional information” as the metabolite ID and ”virus” in the organism ID were therefore removed from the dataset.

Intracellular compounds

As satisfactory filtering for allosteric interactions in BRENDA was not possible, quite a big proportion of the retrieved regulatory interactions were not of an allosteric character. Many of the interactions were competitive, and quite a large group were also extracellular. In order to increase the proportion of allosteric and metabolic interactions, these entries were therefore removed by filtering for and only keeping interactions involving intracellular compounds.

This part of the filtering process was achieved by downloading a dataset with all metabolites participating in models from the BiGG database [42]. The file was retrieved as ”bigg_models_metabolites.txt” from ”Data Access” at BiGG’s website [42], and is available as ”bigg_models_metabolites.csv” (Supplementary information 2, App. A, p. 89). In order to identify the intracellular compounds, ChEBI IDs were extracted from the ’database_links’ column of the BiGG dataset. The intracellular interactions from BRENDA were then isolated by filtering the data for interactions involving metabolites whose ChEBI ID was present in the list of BiGG ChEBI IDs.

Organic compounds

Primary analysis of the original, unfiltered data showed that many of the documented interactions involved inorganic compounds. Despite these compounds being biologically and metabolically relevant, they are typically not allosteric effectors, but rather cofactors or other forms of inhibiting metabolites. To ensure an interaction dataset consisting of mostly allosteric interactions, a final filtering was performed to remove these compounds.

Filtering of the inorganic compounds required the distinction between organic and non-organic compounds. There are several recorded definitions for an organic compound. One of these states that organic compounds are compounds ”in which one or more atoms of carbon are covalently linked to atoms of other elements, most commonly hydrogen, oxygen, or nitrogen” [43]. There are however several exceptions to this rule, including the compounds carbon dioxide and cyanides which are classified as inorganic even though such bonds are present [43]. Albeit, it is definite that all organic compounds will contain carbon, and from experience, most biologically relevant organic compounds will also contain either hydrogen, oxygen, or both. In order to ensure the most complete filtering, several combinations of determining elements were used and assessed for the filtering of the interaction data.

The filtering process was conducted by checking for the presence of carbon, hydrogen, and oxygen in the metabolites’ chemical formula, given by the InChI string that was associated with the metabolites in Subsection 2.1.1.1. InChI is the International Chemical Identifier developed under IUPAC, and contains the chemical content of a compound [41]. The data was filtered in three separate rounds using [C], [C,H], and [C,H,O] as the determining factors to decide which approach was most suitable for isolating the interactions with or-

ganic metabolites. Primary analysis showed that [C] and [C,H] resulted in the remnant of inorganic compounds such as cyanide, while [C,H,O] gave seemingly complete filtering. Also, there seemed to be little difference in the most frequent metabolites among the [C,H] and [C,H,O] groups, and there was only a slightly higher number of both activators and inhibitors in the [C,H] group, indicating that the typical biologically relevant compound contains not only carbon and hydrogen, but also oxygen. The data filtered using [C,H,O] as the determining factor was therefore used for further work.

A final downsizing of the dataset was also conducted by removing duplicate interactions, whose presence was probably due to several people documenting the same interaction. The separate datasets of activating and inhibiting interactions were then concatenated, resulting in a dataset of protein-metabolite interactions involving only organic and intracellular compounds from non-viral organisms. These interactions are documented in ”BRENDA_interactions_intracellular.txt” (Supplementary information 1, App. A, p. 89).

2.1.2 Data analysis

In order to gain a better understanding of the state of our allosteric knowledge, various analyses were conducted on the data from BRENDA that was prepared as described in the previous Subsection (Sec. 2.1.1). These analyses include statistics on the different elements of the data, the creation of frequency distributions for these elements, and the visualization of an interactive network for the most highly regulated enzymes. The Jupyter Notebook containing the script used for performing these analyses, ”BRENDA_analysis.ipynb”, is available in Supplementary information 9, while the utilized datafiles can be found in Supplementary information 1 and 2 (App. A, p. 89).

Firstly, statistics on the data were generated by counting the number of interactions, proteins, metabolites, activations, inhibitions, activators, and inhibitors present in the dataset. The utilized identifiers were EC number for proteins and ChEBI ID for metabolites, while activations and inhibitions were identified by the mode of interaction, given by a ”+” or a ”-”, respectively. Furthermore, frequency distributions of the metabolites, divided into activators and inhibitors, reactions, and organisms were created to further examine the general documentation of allosteric data. The top ten metabolites and proteins in the activator, inhibitor, and enzyme groups were also extracted to evaluate their biological functions and relevance, while the top ten organisms, ranged by number of documented interactions, were extracted to assess the documentation of metabolic data in different species. Additionally, a scatter plot of the activating versus inhibiting metabolites was created with the purpose of evaluating whether frequent activators are also frequent inhibitors, and vice versa.

Lastly, using the Python package Pyvis (v. 0.3.1) [44], a network of the top ten enzymes and their most frequent interacting metabolites was created in order to visualize and more easily evaluate the connections of this group. The most frequent metabolic regulators were defined as those interacting with either two or more of the top ten enzymes, and the network is visualized in the HTML file ”network.html” (Supplementary information 10, App. A, p. 89).

2.1.3 Conservation of allosteric interactions

As mentioned in the introduction, the prediction and discovery of novel allosteric interactions are made more difficult by allosteric sites not typically being conserved between proteins [8]. As the conformation of the allosteric sites affects which effector molecules bind and what effect they have on the protein, it is possible that this low degree of conservation might

influence the conservation of allosteric interactions across species as well. As a mode of evaluating whether this is the case, this part of the current work has the aim of mapping a selection of allosteric interactions onto a phylogenetic tree spanning the three taxa of life, Archaea, Bacteria, and Eukaryota.

In order to create a phylogenetic tree, a phylogeny file must first be built. Using a prepared file of the top 100 most documented species in the interaction data from BRENDA, the tool PhyloT [45] was utilized for creating a Newick file for the phylogeny of these species. PhyloT builds the file by retrieving the genomes of the wanted species from the Taxonomic Database of the National Library of Medicine (NCBI) [45, 46]. Among the top 100 species in BRENDA were seven species whose genomes could not be retrieved. These species included subspecies without specification and a species listed as ”Mammalia”. Because of this, only 93 species were included in the tree generated by PhyloT, and the remaining six unique organisms had to be manually added. These were ’Pseudomonas sp.’, ’Rattus sp.’, ’Bacillus sp.’, ’Synechocystis sp.’, ’Streptomyces sp.’, and ’Arthrobacter sp.’. As a branch of the Mammalian species was already present in the tree, these organisms were simply identified in order to properly map interactions in later steps.

For the manipulation of the Newick tree file created with PhyloT, Python’s package ETE3 (v. 3.1.2) [47] was utilized. The notebook used for this manipulation, ”create_phylotree.ipynb”, is available in Supplementary information 8 (App. A, p. 89). The organisms were added by manually researching their phylogeny in the NCBI Taxonomic Database [46], and then using ETE3’s built-in functions to traverse and search the tree for the closest relatives of these species. When the closest precursor was found, the organisms were added as children to these nodes. The results from this work was a Newick phylogenetic tree file of 99 organisms, available as ”tree_final.nw” (Supplementary information 4, App. A, p. 89).

The interactions to be mapped were selected based on their frequency of documentation among the species in the tree. The top ten documented EC numbers were first identified, and the top metabolic regulator for each EC number was subsequently determined based on the number of interactions between the relevant protein and metabolite with the correspondent mode (activation/inhibition).

For annotating the tree with the chosen allosteric interactions, the web tool Interactive Tree Of Life (iTOL) [48] was utilized. iTOL allows the user to upload a file, for example in Newick format, for visualization of the tree, and then to upload several datasets for annotating the tree with different symbols based on defined labels. Several different formats of such files are accepted, but the one most appropriate for the work in this project is the binary annotation file. This type of file contains information about the dataset, such as the type and label of the dataset, and what type of separator is used for separating different variables. The file also contains information on the different labels to be mapped, including their name, and what color and shape each label should be denoted by. The data itself is represented by a binary matrix, where each row contains one organism and one number for each label, as shown in the example matrix in Table 2.1. The different possible values are 1, 0, and -1, resulting in a colored symbol, an empty symbol, and no symbol in the tree, respectively. For solving this current task, interactions being present in an organism were represented by the number 1, and interactions not being present were represented by the number -1. The different interactions were distinguished by different symbol colors, resulting in a tree where each documented interaction is visible as a colored square at the end of an organism’s branch, as shown in Figure 2.2 for the species of the Mammalian class with three interactions.

Table 2.1: Example of binary annotation matrix, for mapping documented protein-metabolite interactions.

Field labels	Int 1	Int 2	Int 3
Homo_sapiens	1	1	1
Rattus_norvegicus	1	-1	1
Bos_taurus	-1	1	-1
Mus_musculus	-1	1	-1
Sus_scrofa	1	1	1
Oryctolagus_cuniculus	1	-1	-1
Ovis_aries	-1	1	-1
Rattus_sp.	1	-1	1
Equus_caballus	-1	-1	1
Canis_lupus	1	1	1
Cricetulus_griseus	1	-1	-1
Cavia_porcellus	-1	-1	-1
Mesocricetus_auratus	-1	-1	1

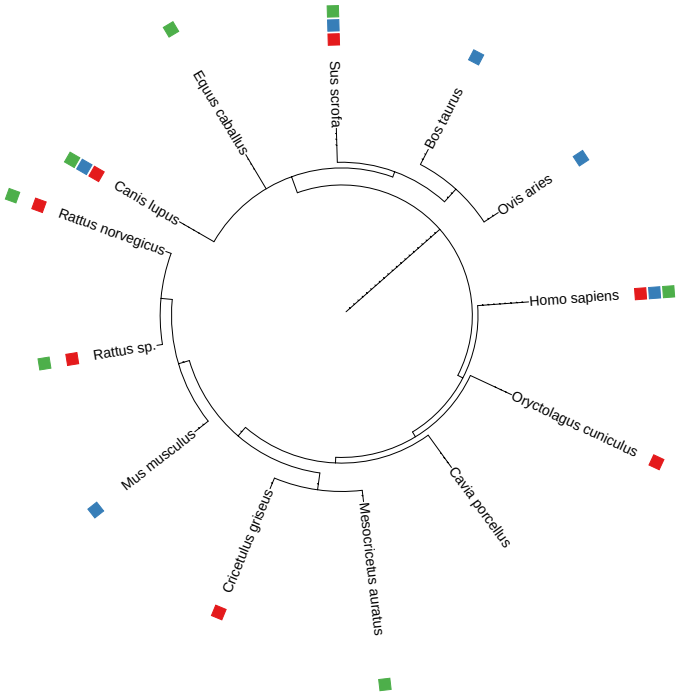


Figure 2.2: Example of phylogenetic tree, mapped with documented protein-metabolite interactions.

2.2 Predicting interactions from protein features

The database of allosteric interactions that was constructed in the first part of the work in this thesis provides basic allosteric knowledge of numerous species. The parts of this Section will describe the work performed to exploit this knowledge for the purpose of predicting protein-metabolite interactions based on information of protein families, domains, and sequence features. This work consisted of retrieving protein information for all EC number and organism couples present in the data, performing enrichment analysis using Fisher's exact test for determining the statistical significance of feature-interaction associations, analyzing the results from enrichment analysis by investigating the biological relevance of

associated interactions and features, and lastly, mapping and validating predicted interactions to the phylogenetic tree created in 2.1.3 to evaluate whether this approach can be used for predicting allostery from protein structure.

2.2.1 Downloading protein annotations

The publicly available databases UniProt and InterPro [32, 30] contain structural annotations for most of the proteins that are documented in BRENDA, and UniProt entries are additionally reported with cross-references to InterPro accessions [32, 30]. The UniProt database [32] was therefore used to retrieve protein data for the top hundred organisms. The script used for downloading protein data is available in the notebook "download_features.ipynb" (Supplementary information 7, App. A, p. 89), while all utilized datafiles can be found in Supplementary information 1, 2, and 3 (App. A, p. 89).

Classifying proteins by structure

Proteins can be classified into distinct groups based on several different properties and behaviors. For example, proteins can be classified based on chemical and structural properties such as solubility, or based on their biological functions [49]. However, proteins may have similar solubilities despite being both structurally and functionally different, and they may also display a range of functions that proposes their classification into several groups [49]. Additionally, not all proteins have known functions, such as those that have been recently discovered and require functional clarification. Another strategy for classifying proteins that has been applied in more recent years is classification based on the proteins' structural and sequential properties. Identifying such properties of novel proteins allow scientists to predict their biological function without conducting any additional experimental work besides determining the protein's amino acid sequence and structural conformation [50]. The functional properties of a protein are in part determined by what molecules it binds and how that binding affects the protein's dynamics, for example the effect of allosteric regulators. The protein structure may therefore be used as a determinant of allosteric effector binding, and traits utilized for such structural classification include protein families, domains, and sequence features [50].

Proteins grouped together due to common evolutionary origin constitute the same protein family [50]. In InterPro, a source of protein sequence information and tools for performing functional protein analysis [30], entries are classified into family and homologous superfamily based on similar functions, sequence, and structure and only similar structure, respectively [30]. Proteins of the same superfamily are more distantly related and display lower sequence similarity than proteins of the same family [50].

While proteins within the same families typically display similar functions, protein domains can be present in proteins with very different functions [50]. A protein domain is defined as a distinct either functional, structural, or sequence unit in a protein, that is usually responsible for the conduction of a specific function or interaction which affects the protein's overall biological role [30, 50]. Domains are not constricted to specific biological contexts and can be found in proteins belonging to the same or different protein families [50].

The third type of protein classifier is sequence features. Sequence features are similar to domains as they confer specific functions to the protein that affects its overall role, but they are much smaller, usually only a few amino acids, and often reside within domains. Different types of sequence features are active sites, binding sites, post-translational modification (PTM) sites, and repeats [50]. InterPro also documents a type of feature referred to as conserved site [30]. Active sites and binding sites contain conserved amino acids involved

in catalytic activity and binding of molecules or ions, respectively [30, 50]. Conserved sites are similar to these, but the conserved residues may not have a documented function [30]. PTM sites consist of residues that are modified after protein translation, for example by acetylation or methylation, while repeats are repeated amino acid sequences within a protein that may possess binding or structural properties [50].

As mentioned, UniProt contains protein structure, sequence, and feature information about all proteins in BRENDA. UniProt is a resource for protein sequence and annotation data that collects information from several external sources in order to preserve the UniProt databases, including the UniProt Knowledgebase (UniProtKB) of functional information of proteins [32]. One of these external sources is the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI), which provides the InterPro service [32, 30]. Information on InterPro entries for protein characteristics annotated to a protein in UniProt is available under the 'Family and domain databases' category of the 'Family & Domains' section, where cross-references to other databases are listed. The InterPro entries have IDs denoted by 'IPR', followed by a six-digit number, and each ID represents one unique InterPro entry of one of the types that were mentioned above, namely family, homologous superfamily, domain, active site, binding site, conserved site, PTM, or repeat. For the sake of simplicity, the term 'feature' will be used to describe all structural and sequential classifiers.

Retrieving protein information

For downloading the protein information for the top hundred organisms in the BRENDA data from UniProt, Python's Bioservices package (v. 1.10.4) [51] was first used. The utilized function, `get_df(search_string)`, retrieves a dataframe of all data that can be accessed in the web interface of UniProt, including information on properties such as binding sites, active sites, and domains for the protein that is searched. The proteins of interest in this analysis were obtained from the file of allosteric interactions created in 2.1.1. As the interactions are organism-specific, UniProt entries of the proteins were searched for by creating strings containing information on both the EC number and organism; `'EC:12345 AND Homo sapiens AND reviewed:true'`. The search strings also specified that only information on reviewed entries was to retrieved, as these have been verified by the team at UniProt [32]. By default, the `get_df()` function only returns information on the top ten entries. To ensure an as accurate analysis as possible, and despite a slightly longer run time, this limit was increased to 100 entries. The result of this process, with only relevant data columns included, is recorded in "protein_features_uniprot.txt" (Supplementary information 3, App. A, p. 89). As the UniProtKB database is continuously updated, it is to be noted that a new retrieval attempt might return other results than what was achieved in this work.

Examining the retrieved protein data made it clear that protein features were described by the type and name of feature in string format, rather than by unique identifiers. As these feature names are not necessarily standardized for all entries and therefore more difficult to work with for associating a feature with an interaction, finding other sources of protein information was appealing. This work was therefore continued by utilizing the SeqIO module of the Biopython package (v. 1.80) [52] and the Python urllib package (v. 1.26.12) [53] to extract data from the URLs for every protein whose UniProt entry ID had been retrieved and was documented in the file "protein_features_uniprot.txt" (Supplementary information 3, App. A, p. 89). By this approach, the InterPro IDs that are cross-referenced to every UniProt entry could be downloaded and associated with the respective protein-organism pair, resulting in the file "protein_feature_interpro.txt" (Supplementary information 3, App.

A, p. 89) which contains a list of InterPro IDs for every unique combination of UniProt entry ID, organism, and EC number.

Lastly, in order to associate the retrieved InterPro IDs with allosteric interactions, the retrieved data was merged with the prepared file from Subsection 2.1.1, after filtering the interactions file for the top hundred organisms. The data was further reorganized to get one unique interaction and feature on every row, which resulted in a file of unique associations between interactions, denoted by EC number, organism, metabolite and mode, and InterPro IDs. As the InterPro accession IDs are not very descriptive, they were mapped to entry type and name by merging the dataframe with a file of basic InterPro entry information, "entry.list.txt" (Supplementary information 2, App. A, p. 89), which was downloaded from the "Download" section on the InterPro website [30]. The association of interactions with features and mapping of InterPro IDs to types and names resulted in the files "features_interactions_merged.txt" and "features_interactions_merged_types.txt", respectively (Supplementary information 3, App. A, p. 89).

2.2.2 Enrichment analysis

Enrichment analysis is a type of analysis that is performed with the purpose of identifying enriched sets of a type of variable among a large population. It is typically used in gene expression analyses for identifying over- or under-expressed sets of genes in a sample [54], but it can in principle be applied to any case. Enrichment analysis requires the application of a statistical test to calculate the degree of enrichment and its statistical significance [54], for example, Fisher's exact test. The following Subsections detail how enrichment analysis was conducted using Fisher's exact test, and describe the evaluation and validation of statistically significant associations between features and interactions. The work of the two first parts of the enrichment analysis, described in Subsection 2.2.2.1 and 2.2.2.2, was performed utilizing the notebook "enrichment_analysis.ipynb", while the notebook "explore_predictions.ipynb" contains the main scripts used for assessing predicted interactions (Supplementary information 11, App. A, p. 89), which is described in the final Subsection (2.2.2.3) of this Chapter. The datafiles used for the work described in this Section are available in Supplementary information 1, 2, 3, and 5 (App. A, p. 89).

2.2.2.1 Fisher's exact test

In order to investigate whether there are any statistically significant associations between protein features and protein-metabolite interactions, Fisher's exact test was applied. The Fisher's exact test is a type of statistical test that can be used for determining whether there is an association between two categorical variables [55], with the null hypothesis being that there is no association between the rows and columns of a 2x2 contingency table [56]. In the case of the current work, where the aim is to associate protein features and interactions, the null hypothesis is defined as follows: there is no association between the presence of a specific protein feature, given by its InterPro ID, and a specific protein-metabolite interaction, given by the interacting metabolite and the mode of interaction (+/-).

Table 2.2 shows an example of a 2x2 contingency table for a certain protein feature and a certain interaction. The number of cases in which both the feature and interaction are present provides the value in the upper left cell (fi), the number of cases in which only the feature is present provides the value of the upper right cell ($\tilde{fi}$)), the number of cases in which only the interaction is present provides the value of the lower left cell ($f\tilde{i}$), and lastly, the number of cases in which neither the feature nor the interaction is present provides the

value in the lower right cell ($\tilde{fi}$). As there are several thousands of different interactions and features, the number of combinations in which neither specific variable is present will always be the highest.

Table 2.2: An example of a 2x2 table of categorical variables, for statistical testing of the association between an interaction and a protein feature using Fisher's exact test.

	Protein feature present (f)	***Protein feature not present ($\tilde{f}$)***
Interaction present (i)	8	75
Interaction not present ($\tilde{i}$)	102	11 573

Fisher's exact test produces two values; the odds ratio (OR) and the p-value. The OR shows how many more times positive cases occur than negative cases [57], which in the current point of issue are constituted by the cases where either both the feature and the interaction is present (fi) or neither the feature nor the interaction is present ($\tilde{f}\tilde{i}$), and those where only the feature or the interaction is present ($f\tilde{i}$, $\tilde{f}i$), respectively. The OR is calculated as shown in Eq. 1 [57, 58].

$$OR = \frac{fi/\tilde{f}i}{f\tilde{i}/\tilde{f}\tilde{i}} \tag{1}$$

The OR can be perceived as a measure of association between an exposing variable, the feature, and an outcome, the interaction [58]. For a more intuitive interpretation, the OR value can also be normalized using $\log_{10}$-transformation (log(OR)). The OR will in this project be interpreted as explained by Eq. 2-4 [58]:

$$OR = 1 \Leftrightarrow log(OR) = 0 \Rightarrow \textit{presence of interaction is not affected by presence of feature} \tag{2}$$

$$OR > 1 \Leftrightarrow log(OR) > 0 \Rightarrow \textit{presence of interaction is associated with higher presence of feature} \tag{3}$$

$$OR < 1 \Leftrightarrow log(OR) < 0 \Rightarrow \textit{presence of interaction is associated with lower presence of feature} \tag{4}$$

Fisher's exact test can be applied either two-sided or one-sided, where the one-sided version has two alternatives; less and greater. With a two-sided hypothesis test, one tests the null hypothesis of the OR being equal to 1 (OR = 1), while the one-sided version tests whether the OR is equal to or greater than 1 (OR $\geq$ 1) and equal to or less than 1 (OR $\leq$ 1) for the less and greater variant, respectively [57]. The example in Table 2.2 gives an OR of 12.10 with associated p-value of 1.00e^{-6} using a two-sided test, which implies that the probability of the OR being 1 is 0.0001%. For determining whether this is a statistically significant score, the p-value must be compared to a confidence threshold α [59]. Choosing a 95% confidence interval, which implies that 95% of the calculated intervals upon repeating the estimation process with random samples from the same distribution is expected to contain the true value [60], gives a threshold of 0.05 (5%). As the p-value of 1.00e^{-6} is less than 0.05, this indicates that there is a significant association between the feature and the interaction. Had the p-value been equal to or above 0.05, the null hypothesis of the OR being equal to 1 could not have been rejected, and the specific feature could thereby not be associated with that specific interaction. Likewise for the one-sided version, the p-value must be below the threshold (α) set by the confidence interval for the null hypothesis to be rejected [57].

As this type of analysis involves not only one, but several separate statistical tests, the resulting p-values should be multiple testing corrected. This type of correction is performed in order to adjust the statistical confidence measures based on the number of tests performed, and thereby reduce the number of false-positives [59]. Adjustment can be conducted using false discovery rate (FDR) estimation. The FDR is a measure of the number of incorrect associations among all that are accepted, defined as the rate of false positives within the group of accepted associations (p-value < α) [61]. The FDR can further be used to calculate the adjusted p-value, also referred to as the q-value. Equally as the p-value, the q-value is also a measure of the significance of an association, but it is calculated in terms of the FDR instead of the false positive rate. This means that while a threshold of 5% for the p-value implies that 5% of the truly not significant associations are considered significant, the same threshold for the q-value entails that among all significant associations, 5% of these are truly null [62].

The results from Fisher's exact test for a larger group of exposure-outcome-pairs can be visualized by plotting the OR-values and q-values as a scatter plot, with log(OR) on the x-axis and log(q-value) on the y-axis. As for the confidence values, a threshold can be set for the OR-value as a measure of how strong the association between the exposure and outcome must be in order for the association to be scientifically interesting. This limit could for example be $\log_{10}(OR) > 1$, which in the current case insinuates that the association is only regarded if the odds of the interaction being present is at least 10 times higher when the feature is present. These positively associated features and interactions will be visible as dots in the top right of the scatter plot. The negatively significantly associated pairs will be to the left of the -1 mark on the x-axis, while the null-associations lie closer to the x-axis around the 0 mark. Due to their resemblance to volcano outbursts, these types of plots are often referred to as volcano plots.

To test the null hypothesis of protein features and protein-metabolite interactions not being associated, two-sided Fisher's exact test was applied to the data of protein features generated in Subsection 2.2.1 ("features_interactions_merged_types.txt", Supplementary information 3, App. A, p. 89) utilizing the SciPy statistical functions module (v. 1.6.1) [63, 64]. The resulting p-values were adjusted by FDR-correction utilizing the Statsmodels module (v. 0.13.5) [65, 66], and OR-values of infinite magnitude were limited to a threshold outside of the OR range. Adjusted p-values (q-values) and OR-values were then transformed on a negative and positive $\log_{10}$-scale, respectively, utilizing Python's Numpy package (v. 1.20.1) [67, 68]. For simplicity, -log(q-value) will be denoted as log(q) where relevant. Analysis of these results demonstrated that no or close-to-no features were negatively associated with interactions. The enrichment analysis was therefore repeated utilizing the greater one-sided variant, and the resulting data, reported in the file "fishers_test_results.txt" (Supplementary information 5, App. A, p. 89) for all associations and in the file "predicted_interactions.txt" (Supplementary information 5, App. A, p. 89) for only statistically significant associations (log(OR) > 1, q-value < 0.05, further denoted 'predictions'/'predicted interactions'), was further used for the work that will be detailed in the following Subsections.

2.2.2.2 Statistically associated features and interactions

The data of associated protein features and protein-metabolite interactions that was generated in Subsection 2.2.1 included protein features of eight different types: active site, binding site, conserved site, domain, family, homologous superfamily, PTM, and repeat. In order to visualize the results produced from Fisher's exact test in the previous Subsection (Sec. 2.2.2.1), the log(OR)- and log(q)-values from "fishers_test_results.txt" (Supplemen-

tary information 5, App. A, p. 89) were plotted in 16 different volcano plots separated by type of feature and by activating and inhibiting interactions. The plots were created utilizing shared x- and y-axes for easier comparison between the groups, and to aid in the evaluation of the effect and significance of different features in the prediction of allosteric interactions from protein structure. The statistical values from Fisher's exact test and the volcano plots were further used to identify highly associated feature-interaction pairs, which were further explored to evaluate whether the association can be biologically explained. This evaluation was performed manually by researching the biological function of the relevant metabolite and protein feature, and assessing their connection.

Additionally, two UpSet-plots comparing the predicted interactions, reported in "predicted_interactions.txt" (Supplementary information 5, App. A, p. 89), for each feature type were created using Python's UpSetPlot package (v. 0.8.0) [69, 70]. UpSet plots are a way of visualizing the relationship between larger number of sets, and are therefore a good alternative to Venn diagrams when working with multiple data groups [69]. Generating an UpSet plot requires the creation of a dataframe of counts for different subsets and boolean values that denote whether the individual sets are part of the subset in question. Python's UpSetPlot package has a function for constructing such a dataframe from existing dataframe columns, which was used to create plots from dataframes of feature-predicted activating and inhibiting interactions. The features were mapped from InterPro ID to type using the file "entry_list.txt" (Supplementary information 2, App. A, p. 89). For better visualization, only groups containing 6 and 10 or more entries were included for activations and inhibitions, respectively. The purpose of generating these plots was to further elucidate the relationship between different feature types, and possibly deduce whether any of the types are excessive in the prediction of PMIs.

Lastly, histograms of the number of predicted interactions for each individual feature, defined by unique InterPro ID, and histograms of the number of features associated with each predicted interaction were created for every distinctive type of protein feature. This was done with the purpose of evaluating the importance of specific features for predicting allosteric interactions, and in order to investigate whether any interactions are more highly predicted in terms of the number of protein features they are predicted by. Separate histograms were made for the activating and inhibiting interactions, resulting in four different histograms for each feature type.

2.2.2.3 Predicting interactions

After conducting enrichment analysis, the statistically significant associations were used to predict allosteric interactions that had not been previously reported. This was accomplished by mapping predicted interactions onto the phylogenetic tree that was created in Subsection 2.1.3. In order to annotate interactions to all organisms and not only to those for which the mapped EC numbers were already documented in BRENDA, protein features for all annotated EC numbers for all organisms present in the phylogenetic tree were downloaded utilizing the notebook "download_features.ipynb" (Supplementary information 7, App. A, p. 89). The data retrieval was performed by the same approach as described in Subsection 2.2.1, only searching for combinations of the top hundred organisms and the ten EC numbers mapped in the tree. The results from this search are available in the file "features_for_ECs_in_tree.txt" (Supplementary information 3, App. A, p. 89).

For determining predicted interactions for the organisms in the phylogenetic tree, the notebook "explore_predictions.ipynb" (Supplementary information 11, App. A, p. 89) was utilized. The dataframe from the file of predicted interactions and their associated features

from Subsection 2.2.2.1, ”predicted_interactions.txt” (Supplementary information 5, App. A, p. 89), was filtered for the annotated interactions using the prepared file ”interactions.csv” (Supplementary information 4, App. A, p. 89), which contains information on all organisms and their annotated protein-metabolite interactions from the tree created in Subsection 2.1.3. To establish which features could be used as predictors of each interaction, the features in this dataframe were grouped to create a new dataframe where each interaction (ChEBI : Mode) is assigned a list of predicting features.

For each feature retrieved for an EC number-organism pair, from ”features_for_ECs_in_tree-.txt” (Supplementary information 3, App. A, p. 89), each group of interaction-associated features was iterated through to find matches between the retrieved and the predicting features. The organisms with matching features, and who weren’t already documented to have the interaction, were added to an interaction-specific list of predicted organisms. A dataframe containing both documented and predicted organisms for each interaction (EC number : ChEBI ID : Mode) was then used to create a new binary annotation file (see Subsec. 2.1.3) such that the predicted interactions are represented by a value of 0 in the matrix, resulting in an empty symbol at the respective label and species, as illustrated in Table 2.3 and Figure 2.3, respectively.

Furthermore, as a mode of validation of these organism-specific predicted interactions, literature was searched with the aim of finding potential experimental proof of these feature-interaction connections. As BRENDA requires scientists to upload data themselves, there is a genuine possibility that novel interactions are discovered but simply not documented in BRENDA. If that is the case, it would function as a validation of the computational prediction. Finding literature that states the opposite of what is suggested by the predictions would also function as a mode of detecting potential false-positive predictions. The validation process was conducted by utilizing the notebook ”explore_predictions.ipynb” (Supplementary information 11, App. A, p. 89) to create a file of the interactions that were predicted to take place in the organisms of the phylogenetic tree based on the presence of associated protein features, named ”predicted_interactions_to_be_validated.txt” (Supplementary information 5, App. A, p. 89), and searching for literature containing information on these organism-specific interactions. The literature search itself was conducted with the help of the co-supervisor of this thesis, Elisa Márquez-Zavala. The resulting evidence is reported in the file ”evidence_of_predictions.csv” (Supplementary information 5, App. A, p. 89), which includes article information such as the title, abstract, and DOI for each article that was found for every organism-specific interaction. These evidences were investigated manually.

Table 2.3: Example of binary annotation matrix, for mapping documented and predicted protein-metabolite interactions.

Field labels	Int 1	Int 2	Int 3
Homo_sapiens	1	1	1
Rattus_norvegicus	1	0	1
Bos_taurus	0	1	-1
Mus_musculus	-1	1	0
Sus_scrofa	1	1	1
Oryctolagus_cuniculus	1	0	-1
Ovis_aries	0	1	-1
Rattus_sp.	1	0	1
Equus_caballus	0	-1	1
Canis_lupus	1	1	1
Cricetulus_griseus	1	0	-1
Cavia_porcellus	-1	0	0
Mesocricetus_auratus	-1	0	1

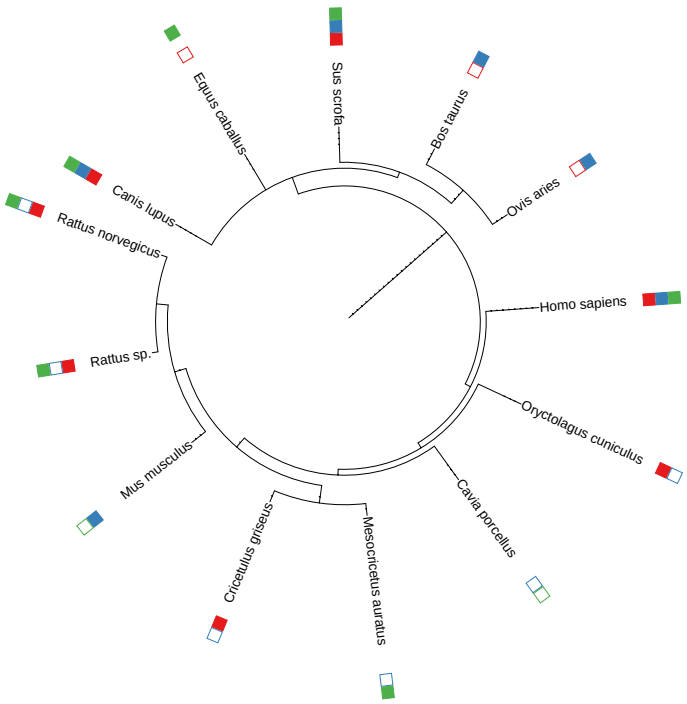


Figure 2.3: Example of phylogenetic tree, mapped with documented and predicted protein-metabolite interactions.

3 Results

This Chapter will present and describe the results from the work performed in this thesis, following the structure given in the Chapter of Methodology (Chapter 2). The first Section is dedicated to examining the content of the assembled allosteric database. This examination includes various statistical analyses, a biological review of frequent allosteric regulators and proteins, and assessing the conservation of allosteric interactions across evolutionary distant and related species. These analyses are followed by the second Section which gives a short description of the results from protein feature retrieval, before going into depth of the results from enrichment analysis where all retrieved data is exploited for drawing connections between the level of protein structure and the level of metabolic regulation by an allosteric mechanism. This final Subsection will provide an overview of the identified connections, present some of these associations in more detail to investigate their biological validity, and demonstrate how predicted interactions affect the allosteric conservation among species.

3.1 Creating an allosteric interactions database

The creation of an allosteric interactions database exploited data on activating and inhibiting compounds retrieved from the BRENDA database [31] for all available organisms and interactions. The first of the following Subsections summarises the state and statistics of this assembled database after processing. The second Subsection presents information on the different contents of the data, while the third and last Subsection will regard the phylogenetic conservation of allosteric interactions as suggested by the available information.

3.1.1 Data download and cleaning

Filtering for organic and intracellular compounds resulted in a file of organism-specific protein-metabolite interactions whose data statistics are displayed in Table 3.1. The number of interactions, including activations and inhibitions, is identified by the number of interactions between unique proteins and unique metabolites, which are identified by EC number and ChEBI ID, respectively. The number of unique, not organism-specific interactions is included in parentheses.

Table 3.1: Statistics on the interaction data from BRENDA, after filtering for organic and intracellular compounds. Proteins are identified by EC number and metabolites by ChEBI ID. Parentheses denote the number of unique, not organism-specific interactions.

Data type	Number
Interactions	32 535 (18 854)
Proteins	3 097
Metabolites	1 002
Activations	6 931 (4 096)
Inhibitions	25 604 (14 758)
Activators	479
Inhibitors	985

An observation made from these statistics is that there is a higher content of inhibitors and inhibiting interactions than of activators and activating interactions. More specifically, the number of inhibitors is double the number of activators, and 78.7% of all interactions are inhibitory. A possible reason for this relationship is that the inhibiting interactions include both allosteric and competitive inhibitors. Although it would have been beneficial to separate and filter out competitive inhibitors, there was not enough information available in BRENDA to classify the type of interaction. The possible reasons for this observed relationship will be elaborated upon in Chapter 4 (p. 59).

3.1.2 Data analysis

Figure 3.1 shows the result from frequency distribution analysis of the activating metabolites, with the number of compounds on a $\log_{10}$-transformed scale. The majority of the activators participate in only a few interactions, as represented by the high bars to the left of the plot. More specifically, 448 of the activators (93.5%) participate in 50 interactions or less, 353 activators (73.7%) in 10 interactions or less, and lastly, 145 activators (30.3%) participate in only 1 interaction. To compare, only 12 compounds take part in 100 or more activating interactions, clearly showing that only a few metabolites regulate very high numbers of proteins in an activating manner.

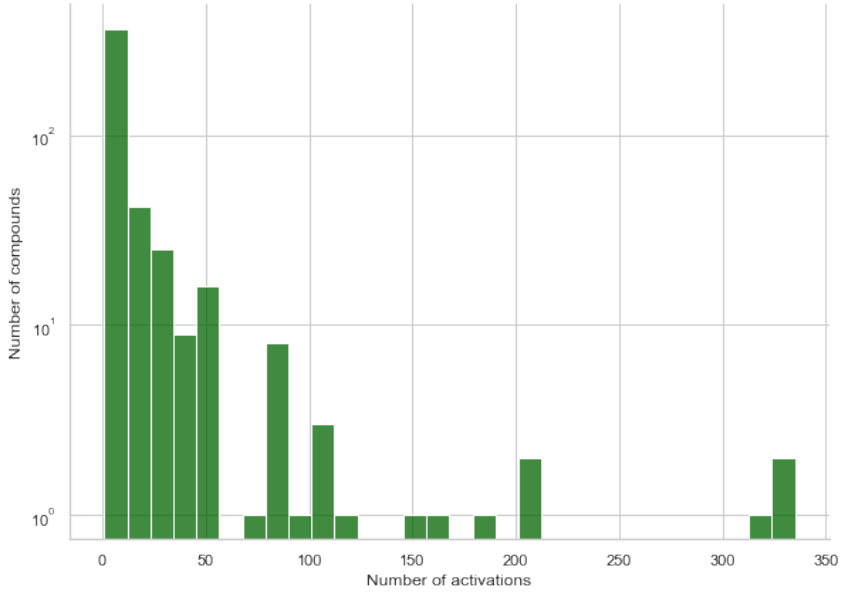


Figure 3.1: Frequency distribution of activators. The frequency of each metabolite was calculated by counting its number of occurrences in activating interactions, which are denoted by *Mode:* +. The number of compounds is shown on a $\log_{10}$-transformed scale.

As the same protein-metabolite interaction may occur in several different organisms, the prepared data contains duplicated interactions. This was accounted for by analyzing the frequency distribution of activators participating in only unique interactions as well, with results illustrated in Figure 3.2. The general distribution has not changed; the majority of activators still participate in only a few interactions, while a few compounds display high activating behavior. One observation made from these two distributions, however, is that the top ten regulatory compounds were somewhat different. When not accounting for duplicated documentation, adenosine triphosphate (ATP) was the top activating metabolite, with glutathione and cysteine as second and third most frequent, respectively. After removing duplicated interactions these metabolites had been rearranged, making glutathione the top

activator and ATP the third most frequent.

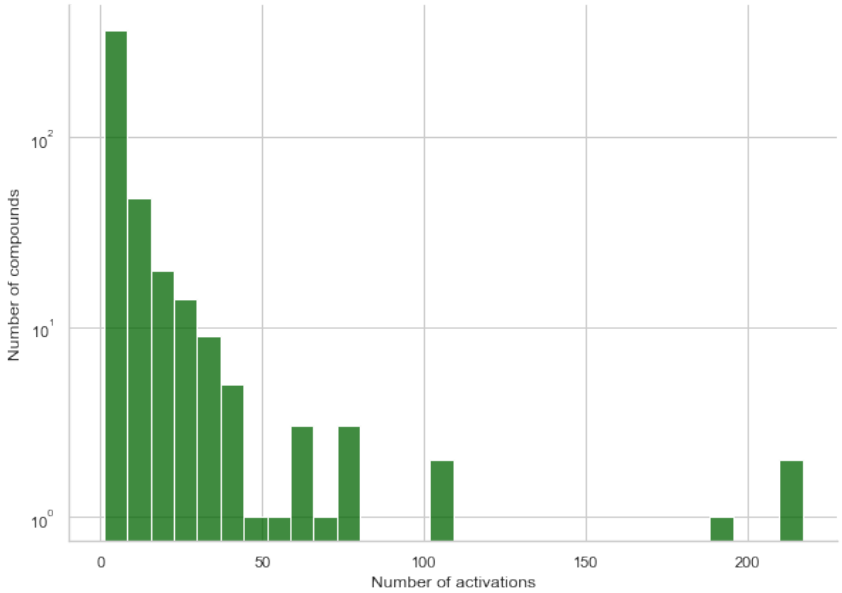


Figure 3.2: Frequency distribution of activators; unique interactions. The frequency of each metabolite was calculated by counting its number of occurrences in unique activating interactions, which are denoted by *Mode:* +. The number of compounds is shown on a $\log_{10}$-transformed scale.

Table 3.2 summarizes the metabolic functions of the top ten most common activators, not accounting for multiple documentation. These metabolites include the nucleotides ATP, adenosine diphosphate (ADP), adenosine monophosphate (AMP), and guanosine triphosphate (GTP), the antioxidants glutathione, cysteine, and ascorbate, the alcohol ethanol, the essential cofactor pyridoxal phosphate (PLP), and the phosphorous sugar compound fructose 1,6-bisphosphate. The metabolic pathways regulated by these metabolites include central carbon metabolism, degradation and biosynthesis of glycogen, amino acid catabolism, and synthesis of nucleotides.

The result from frequency distribution analysis of the inhibiting metabolites is displayed in Figure 3.3, with the number of compounds shown on a $\log_{10}$-transformed scale. Similar as for the activating compounds, the majority of inhibitors participate in only a few interactions, while the minority participate in either an intermediate or low number of inhibitions. 871 of the inhibitors (88.4%) participate in 50 interactions or less, 616 (62.5%) in 10 interactions or less, and 207 (21.0%) in 1 interaction. Furthermore, 61 (6.2%) and 24 (2.4%) of these compounds take part in 100 and 200 inhibitions or more, respectively, while 1 metabolite (0.1%), which was identified as ATP, participates in over 1 000 inhibitions. Although the inhibiting metabolites show a trend of participating in more interactions than what is observed for the activators, these results still indicate that compounds tend to interact with only a few proteins in a regulating matter. However, the results also imply the existence of metabolites that partake in a remarkably high number of protein-metabolite interactions, and who thereby seem to exert a high level of control over several cellular processes.

As for the activating compounds, the frequency distribution of inhibitors is affected by the presence of duplicated interactions. The frequency distribution of inhibitors participating in only unique interactions is displayed in Figure 3.4, which shows a similar pattern between the number of inhibitions and the number of compounds as previously observed. Unlike the activating compounds, the top ten inhibitors were only slightly different after accounting for duplicated interactions; ATP was still identified as the top inhibitor, followed by ADP and AMP.

Table 3.2: Top ten activators and their biological functions.

Metabolite	Type of compound	Biological functions
ATP	Nucleotide [71]	An energy carrier [72], higher concentration indicates a high energetic state [73, p. 582]. Allosterically activates pyruvate dehydrogenase kinase which is part of the pyruvate dehydrogenase complex. The kinase inactivates the complex causing a reduced flux through the citric acid cycle [73, p. 640].
Glutathione	Tripeptide [74]	Functions as a redox buffer and an antioxidant. It is for example used to remove toxic peroxides formed under aerobic conditions, and functions as a reducing agent in deoxyribonucleotide synthesis [73, p. 885]. Glutathione has also been shown to allosterically activate the master virulence regulator PrfA of *Listeria monocytogenes* [75], and it functions as a cofactor for several enzymes [76].
Cysteine	Amino acid [73, p.689]	Cysteine has strong antioxidant properties [77], functions as a sulfide donor [78], and functions as an acid-base catalyst in the active site of enzymes [73, p.197].
D-fructose 1,6-bisphosphate	Sugar phosphate	Fructose 1,6-bisphosphate is an intermediate of glucose metabolism, formed in one of the steps of glycolysis [79]. It can be further transformed in late glycolysis and the citric acid cycle for production of energy and other metabolic building blocks [79], but it also has a role as a metabolic regulator by affecting the activity of several enzymes involved in central carbon metabolism [80]. Its regulatory mechanisms include allosteric activation of pyruvate kinase and glucose-1-phosphate adenylyltransferase [32].
Ascorbate	Vitamin [81]	A vitamin that functions as a coenzyme for several enzymes, and has anti-oxidant and anti-inflammatory effects [81]. Has been shown to enhance the enzymatic activity of TET-enzymes, which play important roles in different biological and pathological processes such as regulation of DNA demethylation, gene transcription, embryonic development, and oncogenesis [82].
AMP	Nucleotide	Higher cellular concentration of AMP indicates a low energetic state. Allosterically activates AMP-activated protein kinase (AMPK) which increases glucose transport, activates glycolysis and fatty acid oxidation, and suppresses energy-requiring processes. Also allosterically activates phosphofructokinase-1 (PFK-1) to increase flux through glycolysis, glycogen phosphorylase to increase glycogen breakdown, and pyruvate dehydrogenase to increase flux through the citric acid cycle [73, p. 582, 592, 609, 640].
ADP	Nucleotide	Higher cellular concentration of ADP indicates a lower energetic state. ADP allosterically activates phosphofructokinase-1 (PFK-1), causing an increased flux through glycolysis. ADP also activates citrate synthase of the citric acid cycle and glutamate dehydrogenase which functions in amino acid catabolism. In addition, higher concentrations of ADP also activate other enzymes of the citric acid cycle and the respiratory chain [73, p. 582, 592, 640, 681, 743], all causing higher energy production.
Ethanol	Alcohol	Affects the function of several neurotransmitter-gated ion channels, including allosteric activation of the glycine receptor [83].
GTP	Nucleotide	GTP can be utilized for nucleic acid synthesis, as an energy source for protein synthesis and gluconeogenesis, and as a signaling molecule [84]. Regulatory functions of GTP include activation of phosphoenolpyruvate carboxylase [85] and allosteric activation of cytidine-5'-triphosphate (CTP) synthase (CTPS) [86], and it may also allosterically activate argininosuccinase [87].
Pyridoxal phosphate	Coenzyme [88]	Pyridoxal phosphate (PLP), the catalytically active form of vitamin B_6, is an essential cofactor for several different classes of enzymes [89]. PLP is the cofactor of all aminotransferases, which catalyze the first step in the catabolism of most amino acids [73, p. 679].

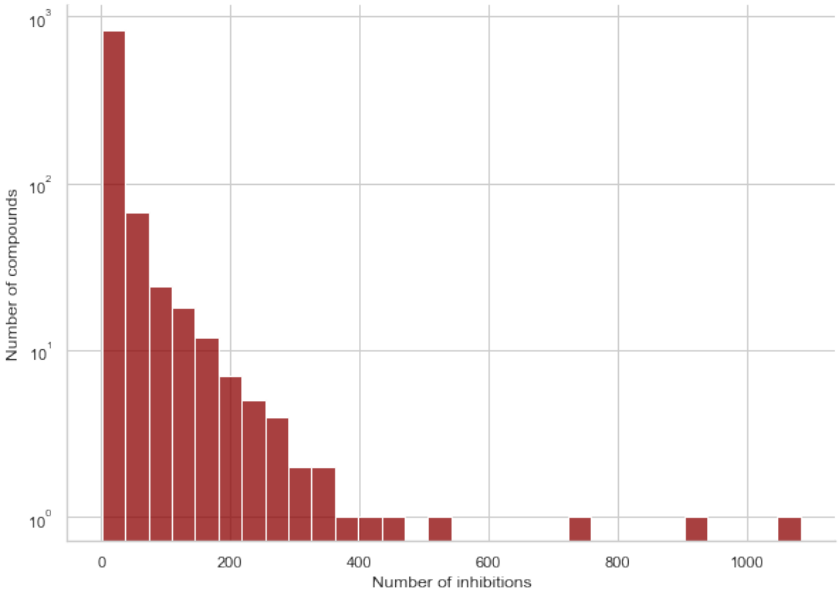


Figure 3.3: Frequency distribution of inhibitors. The frequency of each metabolite was calculated by counting its number of occurrences in inhibitory interactions, which are denoted by *Mode:* -. The number of compounds is shown on a $\log_{10}$-transformed scale.

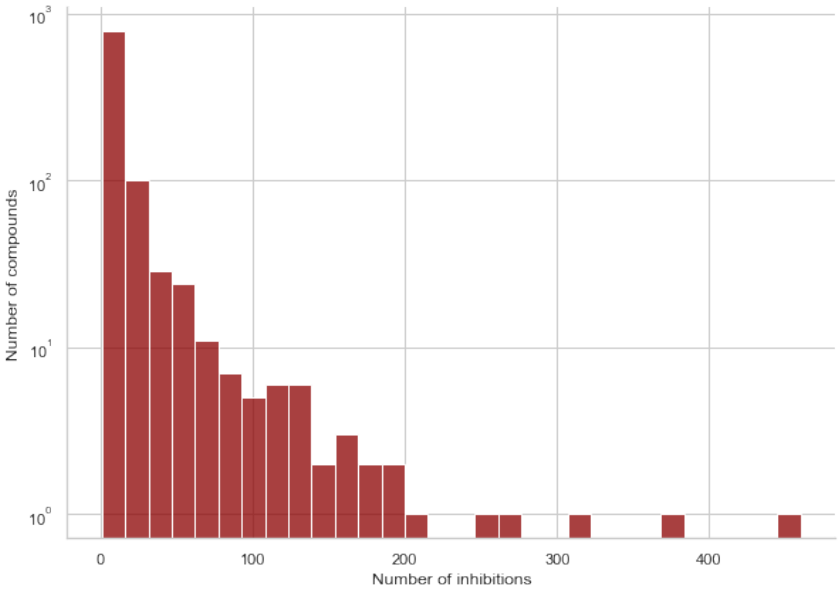


Figure 3.4: Frequency distribution of inhibitors; unique interactions. The frequency of each metabolite was calculated by counting its number of occurrences in unique inhibitory interactions, which are denoted by *Mode:* -. The number of compounds is shown on a $\log_{10}$-transformed scale.

An overview of the top ten most common inhibitors and their functions in metabolism is given in Table 3.3. As for the activators, these compounds include the energy carriers ATP, ADP, and AMP [72, 73], the amino acid cysteine [73, p. 689], and the tripeptide glutathione [74]. In addition to these, the group also contains the nitrogenous compound urea [90], the coenzymes nicotinamide adenine dinucleotide (NADH) and nicotinamide adenine dinucleotide phosphate (NADP+) [73, p. 522], the organic acid citrate, and the sugar-compound glucose. The cellular functions governed by these compounds are mainly constituted by the same pathways that were previously mentioned for the activators.

Table 3.3: Top ten inhibitors and their biological functions.

Metabolite	Type of compound	Biological functions
ATP	Nucleotide [71]	An energy carrying molecule [71, 72], whose higher concentration indicates a high cellular energetic state [73, p. 582]. ATP regulates cellular respiration by negative feedback which lowers the rate of energy production [91]. Enzymes that are allosterically inhibited by ATP include phosphofructokinase-1 and pyruvate kinase in the glycolytic pathway, pyruvate dehydrogenase, citrate synthase, and isocitrate dehydrogenase of the citric acid cycle [73, p. 592, 640], and glycogen phosphorylase of the glycogen catabolic pathway [32].
ADP	Nucleotide	Higher cellular concentration of ADP is an indicator of a lower energetic state [73, p. 582]. ADP allosterically inhibits glycogen phosphorylase, causing suppression of glycogen breakdown [32].
AMP	Nucleotide	Higher cellular concentration of AMP indicates a low energetic state. AMP allosterically inhibits fructose 1,6-bisphosphatase to reduce flux through gluconeogenesis [73, p. 582, 592], and inhibits glucose-1-phosphate adenylyltransferase which is part of the glycogen biosynthetic pathway [32]. AMP thereby leads to a reduced flux through energy-requiring processes.
Cysteine	Amino acid [73, p. 689]	Cysteine has strong antioxidant properties [77], functions as a sulfide donor [78], and functions as an acid-base catalyst in the active site of enzymes [73, p. 197]. It is also an irreversible inhibitor of histidine ammonia lyase [92], which catalyzes the first step of the histidine-degradation pathway [93].
Urea	Nitrogenous compound [90]	Urea is formed from ammonia via the urea cycle as the final end product of protein metabolism [90]. Urea has been shown to cause the denaturation of proteins by both direct and indirect effects [94].
NADH	Coenzyme [95]	NADH is an important coenzyme in redox reactions of metabolism, including reactions of the glycolytic pathway and citric acid cycle. A high NADH to NAD^+ ratio inhibits several enzymes involved in cellular respiration, including pyruvate dehydrogenase, citrate synthase, isocitrate dehydrogenase, and alfa-ketoglutarate dehydrogenase of the citric acid cycle [96], and beta-hydroxyacyl-CoA dehydrogenase of fatty acid oxidation [73, p. 522, 640, 661]. Thus, high levels of NADH cause lower flux through energy-producing pathways.
Glutathione	Tripeptide [74]	Functions as a redox buffer and an antioxidant. It is for example used to remove toxic peroxides formed under aerobic conditions, and functions as a reducing agent in deoxyribonucleotide synthesis [73, p. 885]. Effects include inhibition of γ-glutamyl-cysteine synthetase by a non-allosteric mechanism [97], and protection of cells from immunological cell damage via inhibition of antibody-antigen binding and suppression of complement activation [98]. Glutathione also imposes feedback inhibition upon its own biosynthetic enzymes [76].
Citrate	Organic acid	High concentrations of citrate indicates a sufficient level of energy-yielding metabolism by oxidation of fats and proteins [73, p. 592]. Citrate inhibits several enzymes involved in cellular respiration, and thereby reduces flux through energy-producing pathways. The regulating mechanisms of citrate include the exertion of negative feedback on glycolysis by inhibiting phosphofructokinase 1 and fructose-2,6-bisphosphatase, and on the citric acid cycle by inhibiting pyruvate dehydrogenase and succinate dehydrogenase [99].
Glucose	Sugar	Glucose functions as an energy source for the cell and a precursor to many metabolic intermediates. Glucose allosterically inhibits glycogen phosphorylase [73, p. 534, 609], which participates in the degradation of glycogen [100].
NADP+	Coenzyme [101]	NADP+ is an important coenzyme in redox reactions of metabolism, including reactions of the pentose phosphate pathway [73, p. 522, 565]. It has also been shown to be a negative regulator of ADP-ribosylation [102], a type of modification in which ADP-ribose is transferred from NAD^+ to a substrate [103].

The result from frequency distribution analysis of the targeted enzymes is shown in Figure 3.6, with number of enzymes shown on a $\log_{10}$-transformed scale. As can be seen from the plot, the distribution follows a similar pattern as the distributions for the regulating compounds, where most enzymes are subject to regulation by only a few interactions. The highest regulated enzyme, identified as glucose-1-phosphate adenylyltransferase, is very prominent by participating in a number of 926 reported protein-metabolite interactions. This is almost three times as many interactions as the second-highest regulated protein, which was identified as pyruvate kinase. Furthermore, in total six enzymes (0.19%) are documented with 200 or more interactions, 30 enzymes (0.97%) with 100 or more interactions, and 116 (3.75%) are documented with 50 or more metabolic interactions. To compare, as many as 2 357 (76.1%) proteins are documented with 10 interactions or less, and almost a fourth of the enzymes, 767 (24.8%) to be exact, participate in only one documented protein-metabolite interaction.

Accounting for the multiple documentation of interactions resulted in the distribution depicted in Figure 3.5. This consideration caused quite a drastic change in the enzyme distribution; the highest number of interactions was reduced from 926 to 78, and glucose-1-phosphate adenylyltransferase, which descended to number twelve on the list of enzyme targets, was replaced as the top regulated enzyme by pyruvate kinase. Additionally, the slope of the distribution is a lot less steep than previously, and the percentage of proteins that participate in an intermediate number of interactions is now much higher. Even though this does indicate that the differences in the regulatory behavior of proteins are less drastic than what was initially observed, combining these results with those from frequency distribution analysis of the metabolites implies a biased distribution of allosteric interactions where a few metabolites and enzymes exhibit high regulatory activity. This matter will be further explored in a later paragraph of this Subsection, and in Chapter 4 (p. 59).

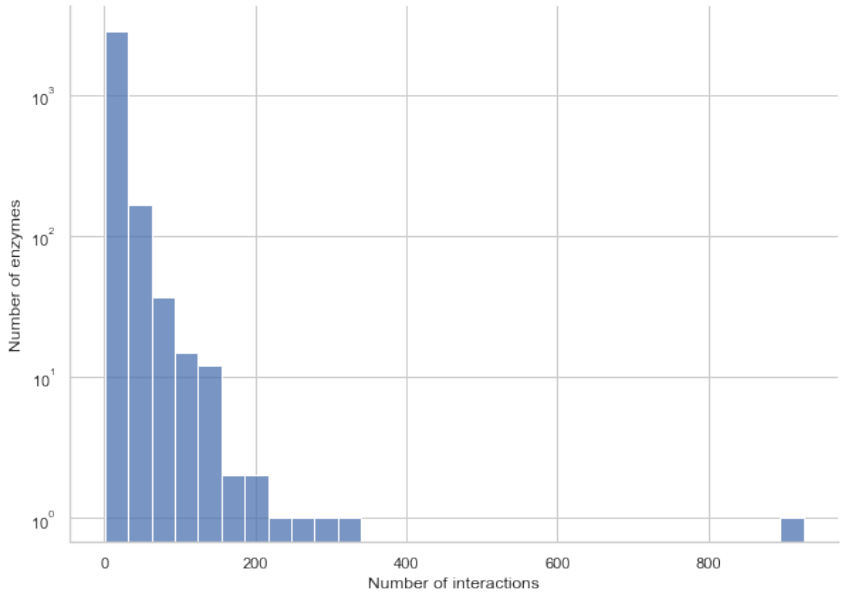


Figure 3.5: Frequency distribution of regulated enzymes. The frequency was calculated by counting the number of interactions, determined by metabolite ChEBI ID and mode, each enzyme was subjected to. The number of enzymes is shown on a $\log_{10}$-transformed scale.

The top ten regulated enzymes and their functions in metabolism are shown in Table 3.4. These include key enzymes of central carbon metabolism pathways, such as pyruvate kinase, 6-phosphofructokinase, and citrate synthase, and enzymes involved in glycogen metabolism. The top regulated pathways are thereby coherent between the regulating

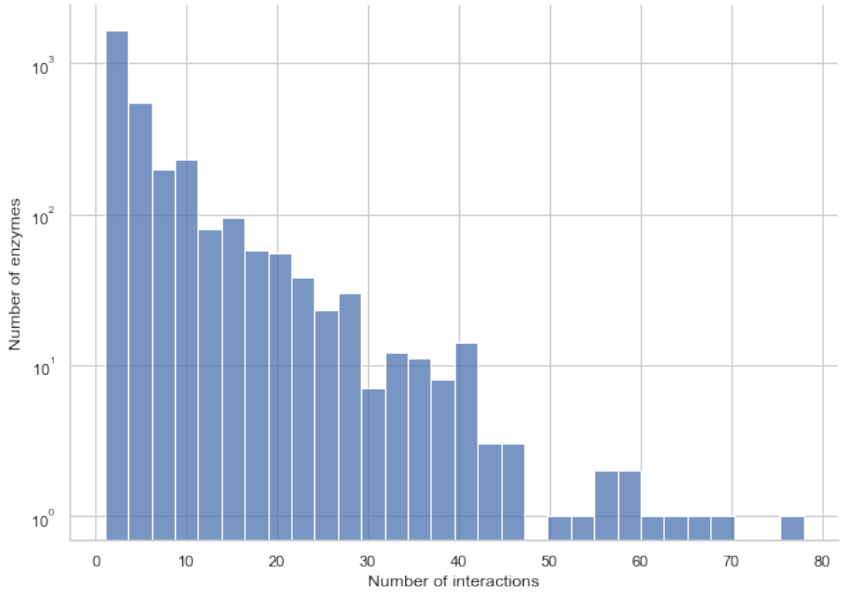


Figure 3.6: Frequency distribution of regulated enzymes; unique interactions. The frequency was calculated by counting the number of unique interactions, determined by metabolite ChEBI ID and mode, each enzyme was subjected to. The number of enzymes is shown on a $\log_{10}$-transformed scale.

metabolites and the regulated proteins.

In addition to the frequency distributions, a network of the top ten regulated enzymes and their interacting metabolites was generated (see Figure B1, App. C, p. 90). The metabolites included in the network are those that interact with either two or more of the enzymes in the chosen group. An observation made from viewing this network is that most of the metabolites interact with more than two proteins, in both an activating and inhibiting matter. Furthermore, all enzymes are subjected to both activating and inhibiting regulation from several different compounds. The network is thereby quite heavily connected, indicating a well-structured, non-random regulation of the pathways that are represented by the included enzymes and compounds. As mentioned, these are mainly constituted by central pathways such as glycolysis and the citric acid cycle, and both synthesis and degradation of glycogen. The distribution of metabolic regulatory activity will be further assessed in Chapter 4 (p. 59).

Figure 3.7 shows the result from frequency distribution analysis of the organisms that were documented in the data, with number of organisms on a $\log_{10}$-transformed scale. Similar as for the other frequency distributions, only a few organisms have a higher number of documented interactions. For instance, only seven organisms (0.29%) had 500 or more documented interactions, while 2 024 (82.4%) and 818 (33.3%) were documented with 10 or fewer and 1 interaction, respectively. The organisms with the ten highest number of documented interactions were *Homo sapiens, Rattus norvegicus, Escherichia coli, Bos taurus, Saccharomyces cerevisiae, Mus musculus, Sus scrofa, Oryctolagus cuniculus, Arabidopsis thaliana*, and *Gallus gallus*. These are all organisms that are either frequently used as model organisms, or that are of special interest to science due to their relevance for human medical development. These results thereby indicate a biased research focus, causing a lower documentation rate for the less interesting organisms. This matter will be further discussed in Chapter 4 (p. 60).

Lastly, Figure 3.8 and Figure 3.9 display the scatter plots of inhibitors versus activators with respect to the number of inhibitions or activations in which each metabolite partici-

Table 3.4: Top ten regulated enzymes and their biological functions.

Enzyme	Function
Glucose-1-phosphate adenylyltransferase	Functions in glycogen and starch biosynthesis in bacteria and plants, respectively [104]. Is allosterically regulated by metabolites depending on the type of organism. Bacterial allosteric activators include fructose-1,6-bisphosphate, hexanediol 1,6-bisphosphate, NADPH, and pyridoxal phosphate, and inhibitors are AMP and $MgCl_2$. In plants, the enzyme is activated by 3-phosphoglycerate and inhibited by orthophosphate [32].
Pyruvate kinase	Catalyzes the conversion of phosphoenolpyruvate and ADP to pyruvate and ATP in the last step of glycolysis. Is allosterically activated by either fructose-1,6-bisphosphate or AMP and other sugar phosphates [105, 106, 107], and inhibited by ATP, among others [73, p. 595].
Beta-glucosidase	Hydrolyses glycosidic bonds to release terminal glucosyl residues from glycosides and oligosaccharides. Has several competitive inhibitors [108].
6-phosphofructokinase	Catalyzes the first committing step of glycolysis, where D-fructose 6-phosphate is phosphorylated to fructose 1,6-bisphosphate by ATP. Is allosterically activated by ADP, AMP, and fructose 2,6-bisphosphate, and inhibited by ATP and citrate [32].
Tyrosinase	Catalyzes the initial and rate-limiting step in a cascade of reactions leading to melanin production from tyrosine [32].
Phosphoenolpyruvate carboxylase	Carboxylates phosphoenolpyruvate to oxaloacetate, which is further processed in the citric acid cycle [109], in plants and bacteria [110]. It is especially important in plants, for the fixation of atmospheric CO_2 in photosynthesis [109]. PEP carboxylase is allosterically regulated by many effectors, depending on the organism [110]. Activators include acetyl-CoA, fructose 1,6-bisphosphate, GTP and glucose-6-phosphate [85, 110]. Is inhibited by aspartate and malate [111, 110].
Glycogen phosphorylase	Catalyzes the rate-limiting step in glycogen catabolism, and thus has a central role in maintaining cellular and organismal glucose homeostasis. Uses pyridoxal 5'-phosphate as a cofactor and is allosterically activated by AMP and inhibited by ATP, ADP, and glucose-6-phosphate [32].
Glutamine synthetase	Catalyzes ATP-dependent conversion of glutamate and ammonia to glutamine. Complete and partial inhibitors include glutamine, glycine, alanine, and AMP [32].
L-lactate dehydrogenase	Catalyzes the conversion of L-lactate to pyruvate. Is inhibited by pyruvate [32].
Citrate synthase	Key enzyme in the citric acid cycle, where it catalyzes the condensation of acetyl-CoA with oxaloacetate to form citrate [112]. Is allosterically inhibited by NADH [32]. Depending on the cell type, succinyl-CoA, NADH, ATP, long-chain fatty acyl-CoA, and citrate are negative allosteric modulators of the enzyme [113].

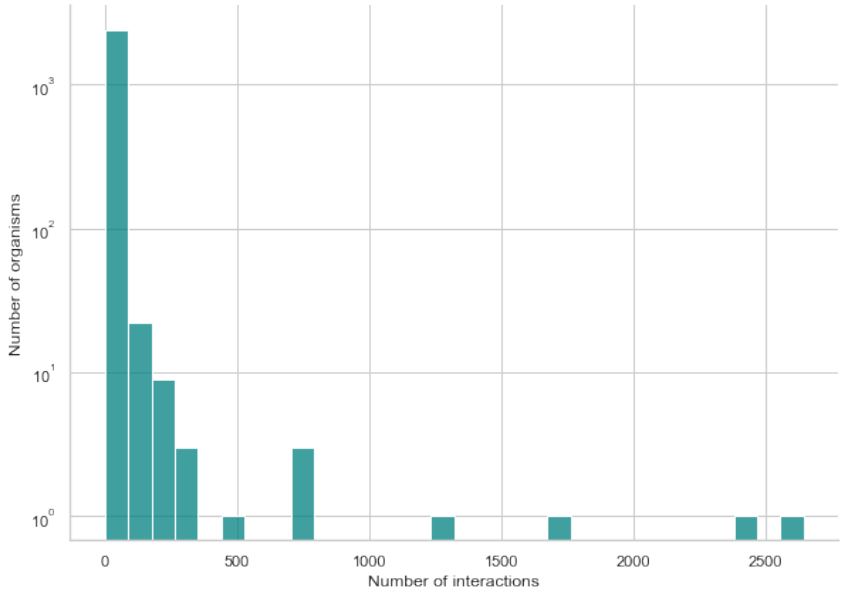


Figure 3.7: Frequency distribution of organisms. The frequency was calculated by counting the number of unique interactions, determined by EC number, metabolite ChEBI ID, and mode, documented for each organism. The number of organisms is shown on a $\log_{10}$-transformed scale.

pates, for reported and unique interactions, respectively. The names of the most easily distinguished metabolites are denoted next to their corresponding data point. The intersections of activating and inhibiting metabolites consisted of 462 compounds. This means that only 17 metabolites were uniquely activators, while the unique inhibitors constituted 523 compounds.

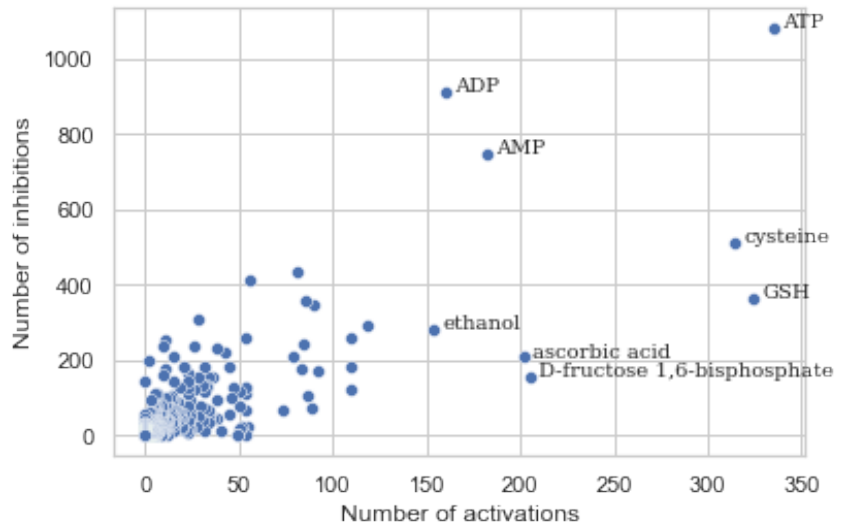


Figure 3.8: Scatter plot of activators versus inhibitors. The plot was created using the activation and inhibition frequencies of each metabolite that were computed for the generation of Figure 3.1 and 3.3.

An observation made from these results is that most of the metabolites cluster in the lower left corner of the plots, while only a few are located further down the axes. Although some compounds appear to only act as either activators or inhibitors, the majority do seem to participate in both types of interaction. One compound, ATP, is located to the top right corner with some distance from the rest of the groups, and is thereby easily distinguishable

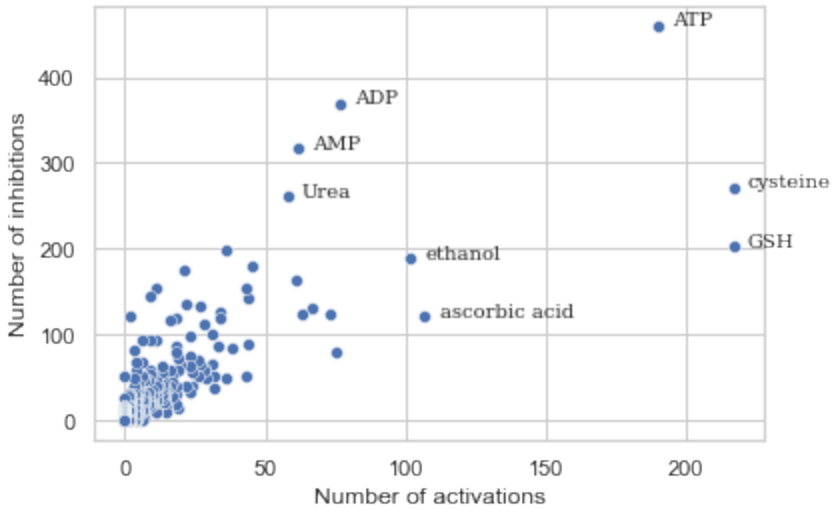


Figure 3.9: Scatter plot of activators versus inhibitors; unique interactions. The plot was created using the activation and inhibition frequencies of each metabolite that were computed for the generation of Figure 3.2 and 3.4.

as the overall top regulatory metabolite. The results from this analysis contribute to the implication that most regulatory metabolites take part in a low or intermediate number of interactions, while also showing that the majority functions as both activators and inhibitors. These results also elucidate the central role of certain metabolites, including ATP, ADP, AMP, and fructose 1,6-bisphosphate, in metabolic control. Chapter 4 will further review these conclusions in light of a biased research focus.

In summary, analysis of the data in the assembled database shows that the metabolic network of allosteric interactions is characterized by high connectivity, and that there exists a biased regulatory consideration that is skewed towards central pathways. While most metabolites and enzymes take part in only a few interactions, there is an exception to this trend that constitutes key metabolic pathways such as central carbon and glycogen metabolism. For these pathways, there is observed a high clustering among the interacting enzymes and compounds, which may be an indication that these pathways are effectively more regulated due to their importance and central role in metabolism. However, it may also indicate that these pathways are simply more studied, which is a relevant topic for the case of the organisms as well. Analysis of the organismal content of the prepared data did indeed imply a skewed research focus in which typical model organisms are well studied, while species that are less important from a research view remain fairly poorly studied. The documentation of allosteric interactions among different organisms is a matter that will be further explored in the following Subsection.

3.1.3 Conservation of allosteric interactions

The phylogenetic tree annotated with the ten most frequent unique EC numbers and their most frequently associated metabolite among the top 100 documented species is shown in Figure 3.10. The three taxa of life are separated by different colors; Archaea in green, Bacteria in purple, and Eukaryota in pink. Each individual interaction is denoted by a uniquely colored square symbol, as described in Table 3.5. The tree in rectangular form with labels for the interactions is available in Figure C1 (see App. C, p. 91).

As can be seen from the figure, most of the interactions are spread throughout the entire tree, and no clear clustering of specific interactions can be observed in any taxonomic group.

Table 3.5: Annotated interactions overview; colors of square symbols used for annotating protein-metabolite interactions to the phylogenetic tree as shown in Figure 3.10.

Color	EC number	Enzyme	Metabolite	Mode
	1.3.5.1	Succinate dehydrogenase	Malonate	-
	2.2.1.6	Acetolactate synthase	L-valine	-
	2.7.1.1	Hexokinase	Glucose-6-phosphate	-
	2.7.1.11	Phosphofructokinase-1	Citrate	-
	2.7.1.30	Glycerol kinase	α-glycerophosphate	-
	2.7.1.40	Pyruvate kinase	Fructose 1,6-bisphosphate	+
	2.7.2.4	Aspartate kinase	L-threonine	-
	2.7.7.27	Glucose-1-phosphate adenylyltransferase	3-phosphoglycerate	+
	3.1.3.11	Fructose-1,6-bisphosphatase	AMP	-
	6.4.1.1	Pyruvate carboxylase	Acetyl-CoA	+

Figure 3.10: Phylogenetic tree mapped with allosteric interactions. Taxa are indicated as Archaea in green, Bacteria in purple, and Eukaryota in pink. See Table 3.5 for information on the mapped interactions and their respective label colors. The tree was created using the iTOL (v. 5) online tool [48].

Furthermore, several of the interactions are present in species belonging to different taxa. For example, both *Homo sapiens* and *Escherichia coli*, who are quite distant from each other phylogenetically, are documented with five common interactions; inhibition of succinate dehydrogenase by malonate (red), inhibition of phosphofructokinase-1 by citrate (purple), inhibition of glycerol kinase by α-glycerophosphate (orange), activation of pyruvate kinase by fructose 1,6-bisphosphate (dark pink), and inhibition of fructose-1,6-bisphosphatase by AMP (grey). All of these enzymes are either directly a part of or related to the different parts of central carbon metabolism, including glycolysis, gluconeogenesis, and the citric acid cycle [32, 105, 114, 115, 116]. The conservation of these interactions, as well as the general case of allosteric conservation, will be further discussed in Chapter 4 (p. 61).

While there cannot be observed any clear clustering of interactions in the tree, there does appear to be an abundance of documented interactions within certain families and species. For example, the Mammalia group, which is part of Eukaryota and branches species from *Rattus norvegicus* to *Canis lupus*, seems to have a higher percentage of documented interactions than other parts of the tree. Additionally, species such as the bacterium *Escherichia coli* and the yeast *Saccharomyces cerevisiae* are documented with all but two and one of the mapped interactions, respectively. These results are in concordance with the findings from frequency distribution analysis of the organisms, which stated that several of these same species were among the best documented organisms in the data from BRENDA.

Furthermore, some of the interactions appear to only be present in specific taxa or groups. One of these is inhibition of hexokinase by glucose-6-phosphate (green), which only occurs in Eukaryota. Hexokinases from different species differ in molecular mass and specificity. Highly specific hexokinases are mainly found in bacteria and unicellular eukaryotes, while non-specific are found in higher eukaryotes. Literature claims that inhibition by glucose-6-phosphate is restricted to the 100-kDa hexokinases, which are mainly found in vertebrates [117]. Additional studies have however shown that some plant hexokinases are sensitive to inhibition by glucose-6-phosphate under certain conditions [118], and glucose-6-phosphate has been declared a possible inhibitor of the *Saccharomyces cerevisiae* hexokinase as well [119].

Three interactions are present in almost all kingdoms, except Animalia (Metazoa). These are inhibition of acetolactate synthase by valine (blue), inhibition of aspartate kinase by threonine (brown), and activation of glucose-1-adenylyltransferase by 3-phosphoglycerate (light pink). Acetolactate synthase participates in the synthesis of the branched-chain amino acids valine, leucine, and isoleucine [120], while aspartate kinase is involved in the biosynthetic pathway leading from aspartate to the formation of homoserine. These two pathways occur only in plants and microorganisms [121, 120], which explains the absence of these interactions in animals. Activation of glucose-1-adenylyltransferase by 3-phosphoglycerate is only documented for kingdoms of the bacterial taxa, as well as the kingdom of green plants (Viridiplantae) [46]. As described in Table 3.4, this enzyme is important for the biosynthesis of glycogen and starch that occurs in bacteria and plants, respectively [104]. The synthesis of glycogen is conducted by a different pathway in organisms such as yeast and mammals [122], which explains the distribution of this PMI.

Lastly, even though the interactions are somewhat well-distributed trough out all three main taxa, there are several gaps present at organisms that lack interactions documented in their nearby relatives. Looking at the Mammalia group again, the species *Cricetulus griseus* (Chinese hamster) and *Equus caballus* (horse) lack four of the interactions that are frequently documented in other mammalian organisms, while *Mesocricetus auratus* (golden hamster) and *Cavia porcellus* (domestic guinea pig) [46] lack three. Additionally, *Candida albicans*, which is the closest relative of *Saccharomyces cerevisiae*, has no documentation

of any of the annotated interactions, and several interactions are also missing from nearby neighbors of *Escherichia coli*. Even though these gaps might be present due to scientifically correct reasons, meaning that the interactions are in fact not possessed by these species, they might also be caused by biased and incomplete research. This issue of inadequate allosteric documentation will be of importance in later Subsections of this Chapter as well, and will also be discussed in Chapter 4.

3.2 Predicting interactions from protein features

The database of allosteric interactions that was created in the first part of this project was merged with structural information on the documented proteins retrieved from the UniProt and InterPro databases. The following Subsections have the intention of describing the results from making these connections, and start with assessing the content of this now augmented database. The second Subsection presents the volcano plots visualizing associated protein features and protein-metabolite interactions, before regarding these associations from a biological view. The final part of this Subsection will display the previously disclosed phylogenetic tree after modification based on the new information, and assess the findings from validation of the novel predictions.

3.2.1 Downloading protein annotations

The retrieval of information on the proteins documented in the processed data returned protein features of eight different types: active site, binding site, conserved site, domain, family, homologous superfamily, PTM, and repeat. Figure 3.11 shows the frequency distribution of all protein features, as defined by their InterPro ID, with the number of features on a $\log_{10}$-transformed scale. The total number of unique features was 5 827. Similar to the relationship observed for the previously displayed frequency distributions (see Subsection 3.1.2), most features have a low or intermediate number of documentations, while only a few are documented many times. One feature, identified as NAD(P)-binding domain superfamily, is easily distinguishable as the most common, with almost 1 200 documentations. In contrast, 966 features (16.6%) were documented only one time, while 3 845 (66.0%) and 5 464 (93.8%) features were documented 10 and 50 or fewer times, respectively. These results thereby demonstrate a trend where the average protein feature is present in a low number of different proteins, while a few dominating features exist and can be found in a very high number of enzymes.

Table 3.6 displays the top ten documented protein features, given by their InterPro ID, type, and name. These features were determined by computing the number of occurrences of each feature in the data downloaded from UniProt and merged with the BRENDA data. As can be seen in Table 3.6, all of the top ten features were of the type homologous superfamily, indicating that this type of protein feature is typically well documented in the relevant protein databases. Furthermore, three of the features are related to the same class of enzymes, namely pyridoxal phosphate-dependent transferases. Pyridoxal phosphate (PLP) is an active form of vitamin B_6 whose dependent enzymes are very versatile catalysts, and the reactions controlled by these enzymes are typically involved in the biosynthesis of amino acids and derivatives. PLP-dependent transferases have also been identified as important drug targets [123], and include the mammalian aspartate aminotransferase and bacteric tryptophan synthase [124]. The best-documented feature, NAD(P)-binding domain superfamily, represents the superfamily of NAD- and NADP-binding domains that can be found in a variety of different enzymes, including several dehydrogenases [30].

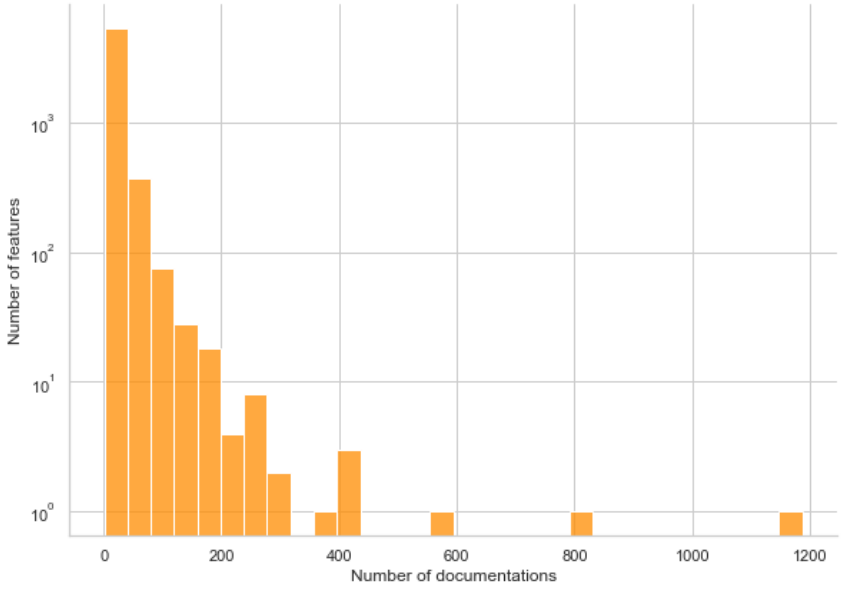


Figure 3.11: Frequency distribution of protein features. The frequency was calculated by counting the occurrence of each protein feature, defined by its InterPro ID. The number of features is shown on a $\log_{10}$-transformed scale.

Table 3.6: Top ten protein features, given by their InterPro ID, type and name. The quantity of each feature was determined by the number of occurrences in the downloaded data.

InterPro ID	Type	Name
IPR036291	Homologous superfamily	NAD(P)-binding domain superfamily
IPR027417	Homologous superfamily	P-loop containing nucleoside triphosphate hydrolase
IPR013785	Homologous superfamily	Aldolase-type TIM barrel
IPR029044	Homologous superfamily	Nucleotide-diphospho-sugar transferases
IPR015421	Homologous superfamily	Pyridoxal phosphate-dependent transferase, major domain
IPR015424	Homologous superfamily	Pyridoxal phosphate-dependent transferase
IPR015422	Homologous superfamily	Pyridoxal phosphate-dependent transferase, small domain
IPR036188	Homologous superfamily	FAD/NAD(P)-binding domain superfamily
IPR011009	Homologous superfamily	Protein kinase-like domain superfamily
IPR011004	Homologous superfamily	Trimeric LpxA-like superfamily

3.2.2 Enrichment analysis

Figure 3.12 shows the volcano plots of associated protein features and protein-metabolite interactions identified by enrichment analysis, for all eight feature types. The significance thresholds in this analysis are set to 1.0 for $\log_{10}$(OR) and 0.05 for FDR-corrected p-values (q-values). Among 57 908 associated features and interactions, 32 276 associations were considered significant. The feature types with the highest number of observed significantly associated interactions from Figure 3.12 are domain, family, and homologous superfamily. The three different types of protein sites, namely active site, binding site, and conserved site, also have several statistically significant associations. The last two feature types, PTM and repeat, appear to have few associated interactions above the statistically significant thresholds. The curves for these two groups are relatively flat compared to the rest of the

plots, and their graphs also contain quite few datapoints. These results thereby indicate that feature types such as domains, families, superfamilies, and protein sites are more important than PTMs and repeats regarding the prediction of allosteric interactions.

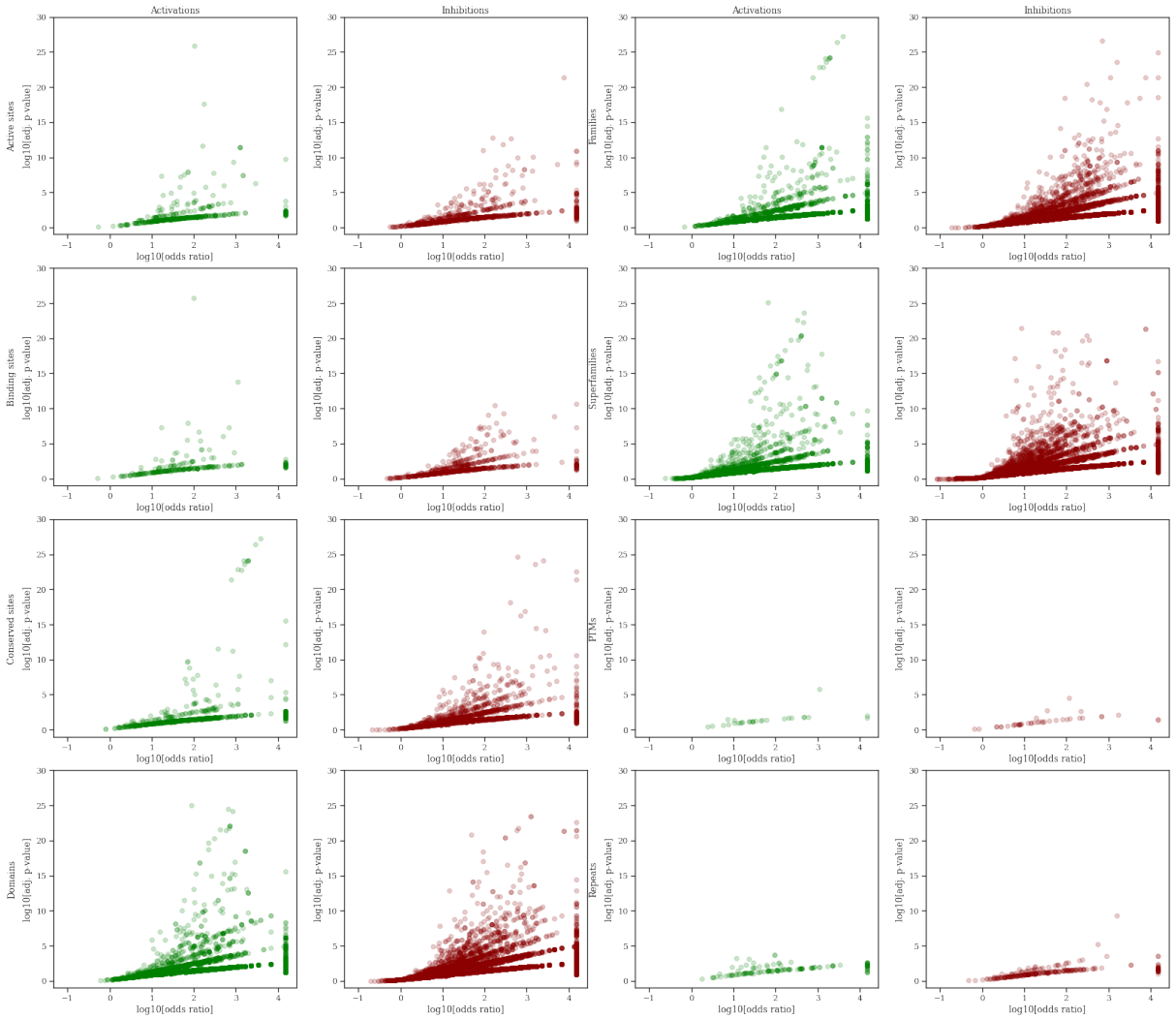


Figure 3.12: Associations of protein features and protein-metabolite interactions, for eight different types of protein features (active site, binding site, conserved site, domain, family, homologous superfamily, PTM, repeat). OR and FDR-corrected p-values are shown on a $\log_{10}$-transformed scale, and the OR of any association shown here with $\log_{10}(\text{OR}) > 4$ was downsized from infinite magnitude.

The importance of the different feature types in predicting allosteric interactions was further assessed by comparing the sets of interactions that were statistically associated with each type, and evaluating their exclusivity regarding predicted interactions. This comparison was visualized by generating two UpSet plots for activating and inhibiting interactions, which are illustrated in Figure D1 (see App. D, p. 93). Observations made from these plots include an extensive overlap of predicted interactions between all feature types. The features conserved site, domain, family, and homologous superfamily have the greatest intersection for both interaction groups, while the second greatest intersection is among domain, family, and homologous superfamily. These features also constitute the largest sets of predictions. Furthermore, domain and family are the only two feature types that predict interactions not predicted by any other feature type. This observation, and the large sizes of these groups, indicate that domain and family are the two most essential feature types for this kind of analysis.

The next Subsections are dedicated to a closer review of the statistically significant associated protein features and protein-metabolite interactions identified in the enrichment analysis. Firstly, specific associated features and interactions are thoroughly examined in order to deduce whether these associations can be explained by a biologically valid motive. Secondly, the work with the phylogenetic tree in Subsection 3.1.3 (Fig. 3.10, p. 46) is continued by mapping the cases in which these interactions are predicted by the presence of a protein feature, for the sake of evaluating whether these predicted interactions fill empty gaps left by the documented data.

3.2.2.1 Statistically associated features and interactions

Of the 32 276 feature-interaction associations that were considered to be significant, a further 7 315 of these were deemed exclusive (infinite OR, downsized to 15 000). Table 3.7 displays a selection of the most highly associated metabolite interactions and protein features among each feature type, together with possible biological explanations for these associations. Features denoted by a * represent those that were among the top five interaction-predicting features in the histograms illustrating the type-separated frequency distributions of interaction-predicting protein features, showed in Figure E1 (see App. E, p. 94-95).

Other highly predicting features identified from Figure E1 included the active sites of protein-tyrosine phosphatase (IPR016130) and histidine acid phosphatase (IPR033379), the PLP-binding site of class I aminotransferases (IPR004838), the conserved cysteine active site of aldehyde dehydrogenase (IPR016160), the deoxynucleoside kinase domain (IPR031314) and family (IPR002624), two superfamilies of domains that are found in aldehyde dehydrogenases (IPR016162, IPR016163), and one superfamily of domains found in both aldehyde and histidinol dehydrogenases (IPR016161). Figure E2 (see App. E, p. 96-98) was used to identify interactions associated with high numbers of protein features, of which some are activation by cyclic AMP (cAMP), activation by estrogen, activation by ATP, inhibition by glucose, inhibition by quercetin, activation by pyruvate, and inhibition by doxorubicin.

An observation made from Table 3.7 is that several of the feature-interaction associations represent the same protein-metabolite interactions. Examples of these include the activating allosteric interaction between cAMP and cAMP-dependent protein kinases, the inhibiting effect of the substrate ribulose 1,5-diphosphate on Rubisco, and allosteric feedback inhibition of threonine on aspartate kinase. Two other repetitive interactions were activation of glucose-1-phosphate adenylyltransferase (also called ADP-glucose pyrophosphorylase) by 3-phosphoglycerate, which is a known activator of this enzyme, and the inhibition of phosphoglucose isomerase by gluconate 6-phosphate, which is known to regulate the enzyme in a competitive matter.

While there exist biological explanations for a considerable number of the described statistically significant associations, many of these represented interactions are not of an allosteric character. In addition to the non-allosteric interactions mentioned above, the results also included covalent bindings, such as biotin to the biotin-binding site, interactions between enzymes and their cofactors, such as the binding of thiamin diphosphate, and several activations and inhibitions whom were either not classified or could not be proved by the reviewed literature. Albeit, the analysis also identified multiple known allosteric interactions. These include, in addition to those previously mentioned, activation of pyruvate kinase by fructose 1,6-bisphosphate, inhibition of PEP carboxylase by aspartic acid, and specificity regulation of ribonucleotide reductase by deoxyribonucleoside triphosphates such as dTTP, dGTP, and dATP. These results thereby imply that even though the analysis did identify biologically

relevant protein feature-metabolite associations, their represented interactions are not all of an allosteric character. This matter will be further discussed in Chapter 4.

Table 3.7: A selection of the most highly associated interactions and features, separated by feature type. A possible biological explanation is given for each association. The adjusted p-values (q) are given by the exponential. Features among the top five interaction-predicting features of their respective type are denoted by *.

Active sites					
Interaction	**InterPro ID**	**Name**	**OR**	**q**	**Biological explanation**
Activation: cAMP	IPR008271*	Serine/threonine-protein kinase, active site	102	10^{-26}	Serine/threonine-protein kinases phosphorylates serine and/or threonine residues on protein substrates, altering their function [125]. cAMP dependent protein kinase (PKA) is a type of serine-threonine kinase [126] that is activated by cAMP in a dynamic and allosteric way [32], which may thereby explain this association.
Inhibition: ribulose 1,5-diphosphate	IPR020878	Ribulose bisphosphate carboxylase, large chain, active site	$7.6e3$	10^{-22}	Ribulose bisphosphate carboxylase (Rubisco) catalyzes the carboxylation step in the Calvin cycle and the oxygenation step in photorespiration. Ribulose 1,5-diphosphate is its substrate [127], which has been shown to inhibit enzyme activation by CO_2/Mg^{2+} [128].
Activation: fructose 1,6-biphosphate	IPR018209	Pyruvate kinase, active site	174	10^{-18}	Pyruvate kinase catalyzes the conversion of phosphoenolpyruvate and ADP to pyruvate and ATP in glycolysis. It is allosterically activated by fructose 1,6-bisphosphate [105], which is a key intermediate and regulatory molecule in carbon metabolism [79, 80].
Inhibition: aspartic acid	IPR033129 IPR018129	Phosphoenolpyruvate (PEP) carboxylase, His active site Phosphoenolpyruvate (PEP) carboxylase, Lys active site	$15e3$ $15e3$	10^{-11} 10^{-11}	PEP carboxylase catalyses the carboxylation of PEP to oxaloacetate [129], and is allosterically inhibited by aspartic acid [111].
Activation: ATP, dATP, dGTP, dCTP, dTTP	IPR030475*	Ribonucleotide reductase small subunit, active site	$15e3$ $1.3e3$ $1.3e3$ $1.3e3$ 863	10^{-10} 10^{-12} 10^{-12} 10^{-12} 10^{-10}	Ribonucleotide reductases synthesize deoxyribonucleoside triphosphates (dNTPs) from ribonucleoside di- or triphosphates (NTPs) for DNA replication. ATP, dATP, dGTP and dTTP are known allosteric effectors that ensure specificity for various substrates [130, 131].
Binding sites					
Interaction	**InterPro ID**	**Name**	**OR**	**q**	**Biological explanation**
Activation: cAMP	IPR017441*	Protein kinase, ATP binding site	99.6	10^{-26}	Protein kinases regulate cellular processes by phosphorylating amino acid residues in protein substrates [132]. cAMP dependent protein kinase (PKA) is a kinase [126] whose activity is stimulated by cAMP, which may thereby explain this association.
Activation: biotin	IPR001882	Biotin-binding site	$1.1e3$	10^{-14}	Biotin binds covalently to a lysine residue in this binding site [30].
Inhibition: NADPH	IPR006184	6-phosphogluconate-binding site	$15e3$	10^{-11}	This binding site is found in the C-terminal of 6-phosphogluconate dehydrogenase [30], which catalyzes the conversion of 6-phosphogluconate to ribulose 5-phosphate. The conversion reduces $NADP^+$ to NADPH [133], which inhibits the enzyme [134].
Inhibition: oxalosuccinate	IPR018136	Aconitase family, 4Fe-4S cluster binding site	$15e3$	10^{-8}	Aconitase catalyzes the conversion of citrate to isocitrate in the citric acid cycle. The enzyme is iron-dependent, with iron present within a Fe-S cluster [135]. Oxalosuccinate is an intermediate in the subsequent reaction, where isocitrate is oxidized to α-ketoglutarate [73, p. 628]. Inhibition of aconitase by oxalosuccinate could thus be explained as a negative feedback mechanism.

Table 3.7: A selection of the most highly associated interactions and features (continued).

Conserved sites

Interaction	InterPro ID	Name	OR	q	Biological explanation
Activation: 3-phosphoglycerate, sedoheptulose 1,7-DP, 2-deoxyribose 5-P, hydroxypyruvate, 2-keto-3-deoxy phosphoglu-conate, phosphoenolpyruvate α-ketobutyrate. Inhibition: trehalose phos-phate.	IPR005836	ADP-glucose pyrophospho-rylase, conserved site	$3.9e3$ $2.9e3$ $1.9e3$ $1.9e3$ $1.9e3$ $1.5e3$ $1.7e3$ $1.6e3$	10^{-28} 10^{-27} 10^{-25} 10^{-25} 10^{-25} 10^{-25} 10^{-25} 10^{-24}	ADP-glucose pyrophosphorylase (also called glucose-1-phosphate adenylyltransferase [31]) is an allosterically regulated enzyme that catalyzes the synthesis of ADP-glucose from glucose-1-phosphate and ATP as part of starch and glycogen biosynthesis [136, 137]. Many of these enzymes are activated by glycolytic metabolites [137], and also regulated by other intermediates of the major carbon assimilatory pathway [138]. The glycolytic metabolite 3-phosphoglycerate is a known activator of the enzyme [138], while phosphoenolpyruvate is also an intermediate of glycolysis and may thus be a putative activator [73, p. 535].
Inhibition: threonine	IPR018042	Aspartate kinase, conserved site	$15e3$	10^{-23}	Aspartate kinase catalyzes the first step in the synthesis of aspartate-derived amino acids such as threonine, lysine, and methionine [139]. The enzyme is known to be regulated by the end-products through feedback inhibition [140], including allosteric inhibition by threo-nine [139].
Inhibition: D-gluconate 6-phosphate	IPR018189	Phosphoglucose isomerase, conserved site	$15e3$	10^{-22}	Phosphoglucose isomerase interconverts glu-cose 6-phosphate and fructose 6-phosphate, and is thus a key part of glycolysis and glu-coneogenesis [141]. D-gluconate 6-phosphate is a competitive inhibitor [142].

Domains

Interaction	InterPro ID	Name	OR	q	Biological explanation
Activation: cAMP	IPR000719	Protein kinase domain	89.6	10^{-26}	See *Binding sites - Activation: cAMP - IPR017441*.
Activation: 3-phosphoglycerate, sedoheptulose 1,7-DP, 2-keto-2-deoxy phosphoglu-conate, hydroxypyruvate, 2-deoxy-D-ribose 5-P, α-ketobutyrate	IPR005835	Nucleotidyl transferase do-main	652 832 720 720 720 672	10^{-25} 10^{-25} 10^{-23} 10^{-23} 10^{-23} 10^{-22}	Nucleotidyl transferases transfer nucleotides between compounds. The specific domain is found in enzymes that transfer nucleotides to phosphosugars [30], and has been an-notated to subunits of the enzyme glucose-1-phosphate adenylyltransferase which is al-losterically activated by 3-phosphoglycerate in plants [32]. It is plausible that this domain is annotated to other transferases that have con-nections to the additional metabolites.
Inhibition: glycerol 3-phosphate	IPR018485 IPR018484	Carbohydrate kinase, FGGY, C-terminal Carbohydrate kinase, FGGY, N-terminal	$1.2e3$ $1.2e3$	10^{-24} 10^{-24}	Proteins of this family carry out ATP-dependent phosphorylation of sugar substrates. They in-clude glycerol kinase [143], which phosphory-lates glycerol to glycerol 3-P and is allosterically inhibited by fructose 1,6-bisphosphate [144]. A BRENDA entry claims that glycerol 3-P is a competitive inhibitor of the enzyme [31], but no evidence of this was found.
Inhibition: threonine	IPR001341	Aspartate kinase	$15e3$	10^{-23}	See *Conserved sites - Inhibition: threonine - IPR018042*.
Inhibition: D-gluconate 6-phosphate	IPR035476 IPR035482	Phosphoglucose isomerase, SIS domain 1 Phosphoglucose isomerase, SIS domain 2	$15e3$ $15e3$	10^{-22} 10^{-22}	See *Conserved sites - Inhibition: D-gluconate 6-phosphate - IPR018189*. The enzyme is com-prised of two domains that are both SIS domain folds [30].

Table 3.7: A selection of the most highly associated interactions and features (continued).

Families					
Interaction	**InterPro ID**	**Name**	**OR**	**q**	**Biological explanation**
Activation: 3-phosphoglycerate, sedoheptulose 1,7-DP, hydroxypyruvate +++	IPR011831	Glucose-1-phosphate adenylyltransferase	$3.9e3$ $2.9e3$ $1.9e3$	10^{-28} 10^{-27} 10^{-25}	See *Conserved sites - Inhibition: 3-phosphoglycerate - IPR005836*.
Inhibition: glycerophosphate	IPR005999	Glycerol kinase	$15e3$	10^{-25}	See *Domains - Inhibition: glycerol 3-phosphate - IPR018485*.
Inhibition: D-gluconate 6-phosphate	IPR001672	Phosphoglucose isomerase (PGI)	$15e3$	10^{-22}	See *Conserved sites - Inhibition: D-gluconate 6-phosphate - IPR018189*.
Inhibition: threonine	IPR005260	Aspartate kinase, mono-functional class	$15e3$	10^{-19}	See *Conserved sites - Inhibition: threonine - IPR018042*. The specific entry represents a subclass of aspartate kinases that are mostly lysine-sensitive.
Superfamilies					
Interaction	**InterPro ID**	**Name**	**OR**	**q**	**Biological explanation**
Inhibition: ribulose 1,5-diphosphate	IPR036376 IPR036422	Ribulose bisphosphate carboxylase, large subunit, C-terminal domain superfamily RuBisCO large subunit, N-terminal domain superfamily	$7.6e3$ $7.6e3$	10^{-22} 10^{-22}	See *Active sites - Inhibition: ribulose 1,5-diphosphate - IPR020878*.
Inhibition: D-gluconate 6-phosphate	IPR023096	Phosphoglucose isomerase, C-terminal	$15e3$	10^{-17}	See *Conserved sites - Inhibition: D-gluconate 6-phosphate - IPR018189*. This specific superfamily is not found in archaeal proteins [30].
Activation: cAMP	IPR011009	Protein kinase-like domain superfamily	66.5	10^{-26}	See *Binding sites - Activation: cAMP - IPR017441*. This superfamily represents the protein-kinase-like domain and other structurally similar domains [30].
Activation: thiamine diphosphate	IPR029061	Thiamin diphosphate-binding fold	470	10^{-24}	The entry represents the thiamin diphosphate-binding fold found in enzymes such as pyruvate dehydrogenases and phosphoketolases [30]. Pyruvate dehydrogenase catalyzes the conversion of pyruvate to acetyl-CoA, utilizing thiamine diphosphate (TPP) as a cofactor [32, 145, 146].

3.2.2.2 Predicted interactions

Expanding the phylogenetic tree from Subsection 3.1.3 with the protein feature-predicted interactions identified in enrichment analysis resulted in the tree depicted in Figure 3.13. As previously, the taxa Archaea is shown in green, Bacteria in purple, and Eukaryota in pink, and the overview of the color-mapped interactions can be found in Table 3.5. The tree in rectangular form with labels for the interactions is available in Appendix C, Figure C2 (p. 92).

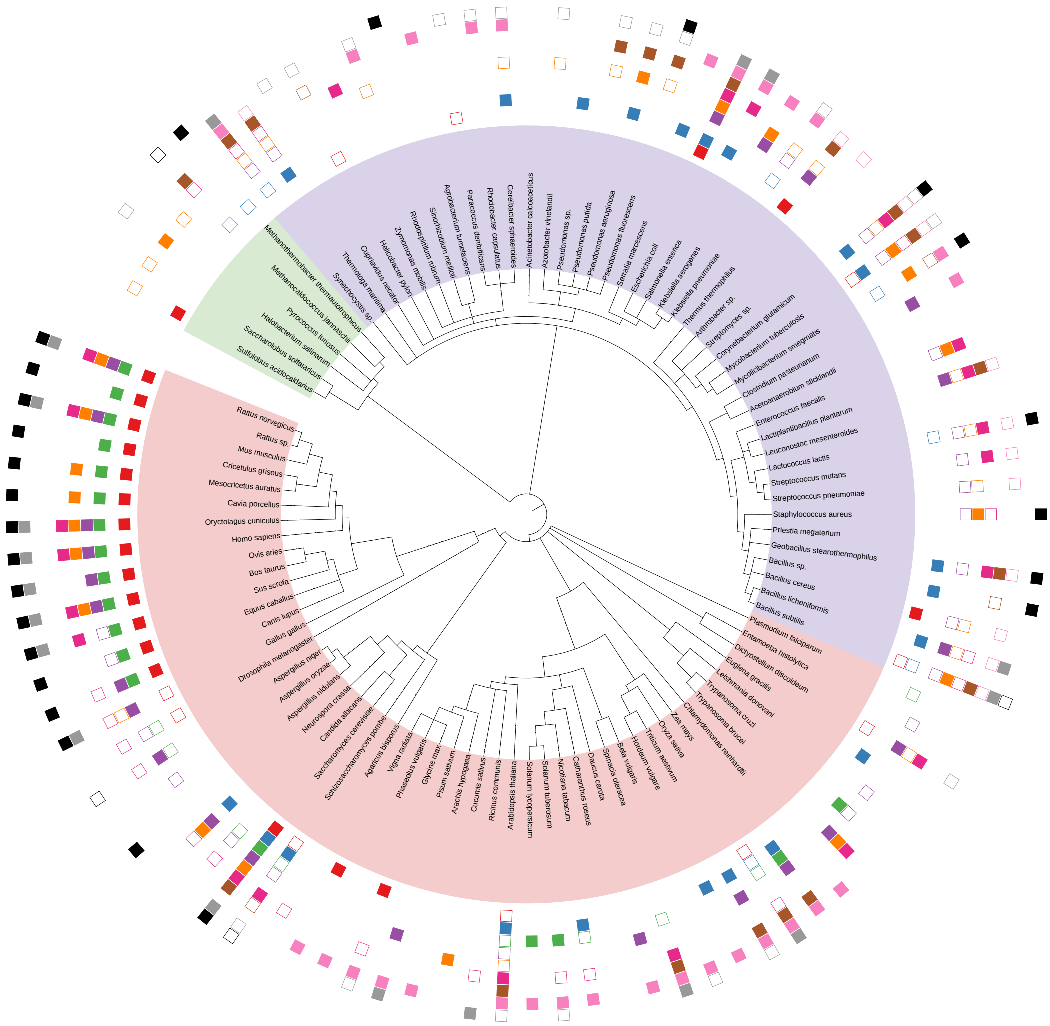


Figure 3.13: Phylogenetic tree with allosteric interactions: documented (filled squares) and predicted (empty squares). Taxa are indicated as Archaea in green, Bacteria in purple, and Eukaryota in pink. See Table 3.5 for information on the mapped interactions and their respective label colors. The tree was annotated using the iTOL (v. 5) online tool [48].

The new phylogenetic tree contains a total of 346 interactions, of which 132 are predicted and 214 are documented. These newly added interactions appear to not have changed the general clustering pattern that was previously observed, in which interactions are spread throughout all taxa. Furthermore, the predictions seem to fill up some of the gaps quite well. For example, the interactomes of *Equus caballus*, *Sus scrofa* and *Canis lupus* are expanded with one of the interactions documented for their relatives, and three mammalian protein-metabolite interactions were predicted for the closest neighbor of mammals, *Gallus gallus*. Moreover, a few organisms went from zero interactions to three and four. These include

for example *Candida albicans*, the closest relative of *Saccharomyces cerevisiae*, *Drosophila melanogaster*, which is a common model organism, and *Streptococcus pneumoniae*. The interactome of another close relative of *S. cerevisiae*, *Schizosaccharomyces pombe*, was expanded from only two interactions to include all but one of those documented for its neighbor. Additionally, the bacteria *Synechocystis sp.*, *Mycobacterium tuberculosis*, and *Bacillus subtilis* went from having three interactions to seven, eight, and nine, respectively, while four interactions were added to the interactomes of the bacteria *Lactococcus lactis*, *Thermotoga maritima*, and *Klebsiella pneumoniae*, and the eukaryote *Oryza sativa*. Lastly, *Arabidopsis thaliana*, which is a typical model organism for plants, went from four to nine annotated interactions.

Regarding the validation of predictions, the literature search for experimental proof of the organism-specific predicted interactions resulted in a total of 10 articles related to two individual protein-metabolite interactions. Due to time constraints on the execution of the project, this was a non-exhaustive search that required little effort and time, and only articles that were not already reported by BRENDA were retrieved for validation. The obtained evidence provided information that both validated and rejected the predicted interactions, but it also contained non-conclusive and non-relevant literature. This non-relevant information regarded the inhibition of hexokinase by glucose 6-phosphate (EC 2.7.1.1, green) in *Plasmodium falciparum*. The evidence obtained for this interaction describes a study investigating the characteristics of the *Toxoplasma gondii* enzyme, which is 44% identical to that of *P. falciparum* [147]. The study states that the hexokinase is not inhibited by the metabolite in question [147], but a quick literature search for the hexokinase of *P. falciparum* revealed that the enzyme is in fact inhibited by its product glucose 6-phosphate [148], which thereby confirms the validity of this prediction.

Of the predicted interactions for which relevant literature was obtained, two predictions related to the activation of pyruvate kinase by fructose 1,6-bisphosphate (EC 2.7.1.40, dark pink) were true positives. These confirmations were related to predictions for *Mycobacterium tuberculosis* and *Neurospora crassa*. The evidence obtained for *M. tuberculosis* describes a study of allosteric regulation of *M. tuberculosis* pyruvate kinase by metabolites, which concludes that fructose 1,6-bisphosphate activates the enzyme [149]. The evidence obtained for *N. crassa* consisted of four articles describing studies of the relevant enzyme, of which all either mentioned or confirmed the activating effect of fructose 1,6-bisphosphate on the *N. crassa* pyruvate kinase [150, 151, 152, 153].

The false-positive predictions that were revealed by the validation process were all related to the activation of pyruvate kinase by fructose 1,6-bisphosphate (EC 2.7.1.40, dark pink). According to literature, fructose 1,6-bisphosphate, AMP and other sugar phosphates are common activators of bacterial pyruvate kinase, and the enzyme may be classified into the pyruvate kinases that are activated by fructose 1,6-bisphosphate and those activated by AMP or another sugar phosphate, such as ribose 5-phosphate [106, 107]. The obtained evidences show that the pyruvate kinases of the bacteria *Thermotoga maritima* [106], *Bacillus licheniformis* [107], *Staphylococcus aureus* [154], and *Synechocystis* (*Synechocystis sp.*) [155] are of the AMP-activated and not fructose 1,6-bisphosphate-activated types, thereby disproving these predictions.

Additionally, it was previously mentioned that the inhibition of hexokinase by glucose-6-phosphate (green) only occurs in higher eukaryotes (see Subsection 3.1.3, p. 46) [117]. However, the predictions in Figure 3.13 indicate that this interaction is present in several eukaryotic microorganisms as well, and, as was mentioned above, confirming evidence was found for this prediction in the parasite *P. falciparum* [46]. The matter of false-positive predictions and this case of seemingly illogical interaction annotation will be further discussed

in Chapter 4 (p. 68).

Overall, analyzing the retrieved structural features for the regulated proteins of the top hundred organisms in the assembled database revealed that the feature types homologous superfamily, family, and domain are most abundant. Associating these features with metabolite-interactions by enrichment analysis resulted in 57 908 associations, of which 32 275 were deemed statistically significant. Comparing the feature types regarding their predicting ability further identified family and domain as the most important interaction-predicting features, and literature reviews revealed that the associations represented several protein-metabolite connections that could be biologically justified. Lastly, annotating predicted interactions to the phylogenetic tree that was disclosed in Subsection 3.1.3 resulted in improved conservation of protein-metabolite interactions. The conducted validation process of these predictions confirmed the organism-specific presence of three interactions, while four were revealed to be false positives.

4 Discussion

This Chapter will review and discuss the results presented in the previous Chapter. The biological interpretation of the content and analysis of the assembled allosteric interactions database will be expanded upon by comparison with related work, and connections between the protein feature and metabolite-interaction level will be further reviewed with the purpose of explaining and assessing the accuracy of the applied approach.

4.1 Creating an allosteric interactions database

In this Section, results generated during the first main part of this conducted work will be regarded. Firstly, the results from analyzing the assembled database of allosteric interactions are discussed and compared to the work performed by Reznik *et.al.* [27]. Furthermore, this Section will evaluate the contents of the utilized database in view of diverse and precise research, as well as assess the conservation of the acquired allosteric interactions on a taxonomic level.

4.1.1 Comparison with previous work

The analysis performed in this project has a lot in common with the work conducted by Reznik *et.al.* for reconstructing a small-molecule regulatory network [27], which was described in Subsection 1.3. These common components include the investigation of frequently regulating metabolites and frequently regulated enzymes, and the identification of highly regulated and controlled metabolic pathways.

From Table 3.2 and 3.3, metabolites such as ATP, ADP, AMP, GTP, and cysteine were identified as frequent regulators of protein activity in the current analysis. These results are in concordance with the conclusions of Reznik *et.al.*, which identified these compounds, among others, as frequent regulators of *E. coli* metabolic pathways [27]. Additionally, ATP was identified as the overall most frequent regulator in both analyses, with participation in 57 different interactions in *E. coli* [27] and 650 unique interactions in the data utilized in the current work. These results indicate that ATP possesses high regulatory control, which makes sense from a biological view considering the central role of ATP as the main cellular energy-carrier [72].

Another similarity between the results of Reznik *et.al.* and this current analysis is the percentage of inhibitory interactions. Despite the higher numbers of identified interactions, enzymes, and metabolites in the current work, which are caused by the higher number of considered organisms, the percentage of inhibitory interactions reported by Reznik *et.al.* for *E. coli* is only 5.3% higher than the percentage of inhibitions in the currently assembled data. This indicates that inhibitory interactions are more common than activating interactions, and that this relationship exists regardless of organism. Another remark to support this implication is derived from Figures 3.1, 3.2, 3.3 and 3.4, which show that the number of interactions rendered by inhibitors is higher than for activators. These results make sense considering the widespread use of feedback inhibition mechanisms to control metabolite levels in pathways such as glycolysis and amino acid metabolism. Additionally, even though filtering was conducted to remove interactions conducted in a competitive manner, viewing the data did make it apparent that this filtering was not 100% successful. It is, therefore, a great possibility that the final dataset contains a significant proportion of competitive inhibitors, which is the case for the dataset utilized by Reznik *et.al.* as well.

An interesting observation regarding the relationship between activating and inhibiting metabolites was made from Figure 1.8 and the data utilized for creating the scatter plots in Figures 3.8 and 3.9; 14 metabolites were identified as uniquely activators in *E. coli* [27], and only 17 uniquely activators were identified from the data assembled in this work. This implies that expanding the analysis of protein-metabolite interactions from *E. coli* alone to all available organisms increased the number of unique activators by only 3 metabolites. To compare, this expansion increased the number of inhibitors by over 300 metabolites. Moreover, the interactive network in Figure B1 (p. 90) shows that most metabolites regulate enzymes by both activating and inhibiting mechanisms, and the few compounds that only regulate by one mode usually do so by inhibition. These findings demonstrate that very few compounds act solely as activators on a global basis, and further emphasize the dominance governed by inhibiting effectors on metabolic regulatory networks.

Several of the highly regulating metabolites identified in Tables 3.2 and 3.3 and Figures 3.8 and 3.9 were found to be important for the regulation of central pathways such as glycolysis, the citric acid cycle, gluconeogenesis, and glycogen metabolism. Additionally, many of the identified highly regulated enzymes act as catalysts in these pathways, making these results coherent with the conclusions of Reznik *et.al.* [27]. The interactive network in Figure B1 (App. A, p. 90) further displays the high connectivity among these top regulated enzymes. While the results from frequency distribution analysis indicate a somewhat poorly connected metabolic interaction network due to few interactions posed by the majority of metabolites, the hubs exemplified in Figure B1 make the interactive network highly interconnected. These findings thereby indicate that enzymes of central metabolic pathways such as glycolysis are devoted more regulatory attention than the average protein.

Overall, the results from data analysis and the work by Reznik *et.al.* [27] imply that the most frequent regulatory metabolites and targeted enzymes are biased towards central pathways. Reznik *et.al.* explains that this non-random distribution of regulatory interactions may be explained by the conservation of resources accomplished by feedback inhibition [27], such as that imposed by ATP on the glycolytic enzymes 6-phosphofructokinase and pyruvate kinase, and on citrate synthase of the citric acid cycle [32, 113, 73, p. 71]. However, as these pathways are so essential for the general function and life of a cell, they are also of the highest interest to researchers of the life sciences. This has encouraged the conduction of excessive research on these pathways, making them generally more well-studied than other less central parts of metabolism. This means that even though the observed pattern of regulatory attention may be caused by effectively higher regulation of central pathways, one cannot reject the possibility of other pathways being just as highly regulated. This topic of a biased research focus will continue to be of importance in further discussions of this Chapter.

4.1.2 Common model organisms are well documented

The frequency distribution of documented organisms in Figure 3.7 (p. 44) implies that the top ten documented organisms consist of common model organisms and mammalian species. For example, *Homo sapiens* was the species with most documented interactions, followed by *Rattus norvegicus* and *Escherichia coli*. While these results might be interpreted as an indication of a higher allosteric interaction content in the listed species, these observations are more likely due to a biased research focus rather than differences in the allosteric interactome. Naturally, more research is being performed on species that are frequently used as model species, such as *E. coli*, mouse and rat (*Mus musculus*, *R. norvegicus*), and *Arabidopsis thaliana*. Additionally, developing new and improved technologies and treatments within the field of medicine will always be of the utmost importance to the human

race. This causes a great abundance of research and documented interactions for *H. sapiens* and related species.

The interaction-annotated phylogenetic tree in Figure 3.10 (p. 46) further demonstrates the trend of skewed documentation. Observations include considerable documentation of interactions in the kingdom of Mammalia and a high number of interactions documented for *E. coli* and the model-yeast *Saccharomyces cerevisiae*. Several archaeal species, on the other hand, display much more incompleteness in the allosteric content, which supports the assumption that PMIs of model organisms are more commonly discovered and reported. However, there are some observations that don't quite follow this pattern. For example, *Drosophila melanogaster*, which is frequently used for research, had no documented interactions in Figure 3.10, and *Chlamydomonas reinhardtii*, a green alga which is claimed to be a well-established model [156], only possessed one annotated interaction. Furthermore, *A. thaliana*, which was established as one of the top ten documented species, had fewer annotated interactions than other model organisms, and also fewer than its plant-relative *Zea mays*.

The differences in allosteric content observed in this analysis might imply that species such as *A. thaliana* simply don't possess the missing interactions. However, these discrepancies might also be indicators of lacking documentation. It is for example unlikely that none of the ten annotated interactions occur in *D. melanogaster* and that only one occurs in *C. reinhardtii*, especially considering the documentations of their neighbors. If these findings happen to be the result of lacking research, that might cause an inaccurate image of allosteric conservation. Also, absent protein-metabolite interactions may result in false or imprecise conclusions of studies that utilize these species due to the lack of possible important metabolic mechanisms and constraints. The issue of absent documentation for possible naturally present interactions will be further discussed later in this Chapter.

4.1.3 Allosteric interactions are well conserved

As described in Subsection 3.1.3, the interaction-annotated phylogenetic tree displayed in Figure 3.10 (p. 46) indicates a taxonomic-wise well-conserved allosteric interactome. This is made apparent from the spread of interactions across different taxa, families, and species. For example, five interactions were common between *Homo sapiens* and *Escherichia coli*. These two species belong to two different taxa, namely Eukaryota and Bacteria, respectively. The presence of allosteric interactions in two such genetically distant species is an indication that these interactions are conserved on the taxa level, and are therefore likely to be present in relatives of these species as well. For *H. sapiens*, four close neighbors are documented with all of the relevant interactions, and several others are documented with either one, two, or three of these PMIs. On the other hand, none of *E. coli*'s nearest neighbors are shown to possess all of the common interactions, while a few relatives have one or two relevant documentations. As the presence of these interactions across taxa indicates universal conservation, it is likely that the absence of these PMIs from other species, especially the species closely related to the organisms in question, is due to a lack of documentation or research rather than biological reasons. These findings thereby support the hypothesis formulated in Subsection 4.1.2.

4.1.4 BRENDA reports uncertain information

As has been acknowledged several times in the previous Subsections of this Chapter, missing interactions from the interaction-annotated phylogenetic tree in Figure 3.10 (p. 46) might be a biological inaccuracy caused by incomplete documentation. However, the tree

also contains a few interactions for which no evidence could be obtained in either UniProt or other literature, and that thus appear to be wrongfully annotated. For example, Figure 3.10 (p. 46) and Table 3.5 (p. 46) show that the activation of glucose-1-phosphate adenylyltransferase by 3-phosphoglycerate (light pink) is present in bacteria and plants. As described in Table 3.4 (p. 43), this enzyme participates in the biosynthetic pathway of glycogen and starch in bacteria and plants, respectively, and it is allosterically regulated by organism-dependent metabolites. While 3-phosphoglycerate is reported as the main activator of the plant enzyme, reviewed literature claims that the only bacterial species that possess this regulation are oxygenic photosynthetical cyanobacteria [157]. Despite this fact, BRENDA contains several documentations of this interaction's presence in other bacteria as well, including *E. coli*, *Salmonella enterica* and *Clostridium pasteurianum* [31]. Many of these BRENDA entries don't specify the mode of activation, namely if the activation occurs by an allosteric mechanism, and quite a large number of these organism-specific PMIs are also denoted as pH-dependent. As much of the data reported in BRENDA originates from studies performed in vitro, the conditions under which these interactions take place might be different from in vivo conditions. This might result in the detection and documentation of interactions that don't generally occur in natural biological systems, and that are thus not reported by studies performed in vivo. Additionally, it has frequently been claimed in this Chapter that a lack of research is an essential issue in the field of allosteric regulation. Therefore, it is possible that the entries in BRENDA do in fact represent correct information, but that enough research has not been performed to further validate these interactions and document them in additional databases such as UniProt.

Moreover, Figure 3.10 (p. 46) and Table 3.5 (p. 46) imply that the hexokinase of several eukaryotes is inhibited by glucose-6-phosphate (green). As was noted in Subsection 3.1.3, hexokinases from different species differ in molecular mass and specificity for both substrates and regulators, and inhibition by glucose-6-phosphate is mainly restricted to hexokinases of vertebrates [117]. Studies do however indicate that some plant hexokinases and the *Saccharomyces cerevisiae* hexokinase may be sensitive to glucose-6-phosphate inhibition [118, 119], which thereby explains the presence of this interaction in species such as *Zea mays* and *S. cerevisiae*. One contradiction concerning this interaction regards the kinetoplast *Trypanosoma cruzi* [46]. As indicated by Figure 3.10, BRENDA reports that the *T. cruzi* hexokinase is inhibited by glucose-6-phosphate. However, more recently published literature than what was referred to by BRENDA states that no such inhibition takes place [158], which means that BRENDA essentially reports wrongful evidence according to the current knowledge. This type of inaccuracy is to be expected when utilizing BRENDA because it is only possible to add a discovery to a database, not a negative validation. From these observations, it appears as if the data in BRENDA is somewhat outdated and affected by an automatic and naive approach to the collection of information. As BRENDA is utilized as the main source of information in the current work, this constitutes a possible weakness of the current approach. This issue can also be regarded as a weakness of the general documentation of allosteric interactions, as true allosteric interactions are not distinguished from weak, competitive, or environment/condition-dependent interactions.

4.2 Predicting interactions from protein features

This Section will address different aspects related to the second main part of the conducted work, beginning with discussing the predictive ability of different protein structure features. The approach for predicting allosteric interactions from protein feature data will then be assessed and reviewed in light of the validity and correctness of feature-interaction associations and predicted interactions, and the effect of predictions on the allosteric conservation

discussed in the previous Section will be considered.

4.2.1 Domain and family are most important for predicting PMIs

The retrieved data of feature information for the proteins in the assembled allosteric database contained 5 827 unique features classified by InterPro into one of eight different feature types: active site, binding site, conserved site, domain, family, homologous superfamily, PTM, and repeat. Table 3.6 (p. 49) shows that all top ten documented features were of the type homologous superfamily. Furthermore, the feature types domain, family, and homologous superfamily constitute the groups that contained the highest numbers of statistically significant feature-interaction associations (see Fig. D1, p. 93). In order to gain a better understanding of the connection between protein features and metabolite-interactions, this Subsection will try to uncover the reasons behind these observations.

Firstly, what will cause a specific feature type to be associated with many interactions? The current analysis was performed by associating each regulated protein in the database with its annotated protein features, and then each of these features was associated with the interaction in question. For a protein that is highly regulated, each annotated feature will be associated with several interactions, while a barely regulated protein might yield the association of just one interaction for each feature. Furthermore, features that are annotated to a high number of proteins will likely be associated with more interactions than if they were only annotated to one or two proteins. This means that features annotated to several highly regulated proteins will be associated with more interactions than features annotated to few, relatively poorly regulated proteins.

While the number of associated interactions for a feature is related to the number of proteins to which the feature is annotated, the annotation of this feature is further dependent on factors related to the characteristics of its feature type. For example, superfamilies are the most generic type of feature because they not only encompass other, less general features such as domains, but also because the proteins of a superfamily are only similar by structure. Proteins of a family, on the other hand, are similar by both function and sequence, as well as structure, which considerably reduces the number of compatible proteins. Thus, the number of proteins within a family will likely be lower than the number of proteins within a superfamily, causing a higher relative documentation rate of superfamilies.

The abundance of superfamily as an interaction-associated feature type is made even more clear when examining the top ten documented protein features listed in Table 3.6 (p. 49). As mentioned, all of these features were superfamilies, and three of them were related to PLP-dependent transferases. The catalysts that depend on PLP (pyridoxal phosphate) are typically very versatile, and the PLP-dependent transferases include, among others, mammalian aspartate aminotransferase and bacteric tryptophan synthase [124]. As this group of transferases is comprised of proteins from at least two distinct taxa of life, that means that the proteins of these three superfamilies are most likely quite prevalent, producing a high documentation rate. The abundance of these superfamilies is further clarified by one of the highly predicting features that were identified from the histograms in Figure E1 (p. 94-95), namely the PLP-binding site of class I aminotransferases. The proteins annotated with this binding site are probably also annotated with either one or several of the superfamilies related to PLP-dependent transferases. For a feature to predict an interaction, it must be associated with that interaction via the proteins to which it is annotated. These results thereby indicate that the proteins annotated with PLP-dependent transferase-related features are subject to regulation by many different interactions in many different species, causing a high number of documentations.

A similar case of documentational abundance was found for the NAD(P)-binding domain superfamily. This was the overall best documented protein feature, and it represents protein domains to which the common cofactors NAD and NADP bind. As these cofactors are essential for the life of all cells, the magnitude and ubiquity of the proteins possessing such domains is a probable reason for the high number of documentations of this specific feature. Additionally, the NAD(P)-binding domain superfamily is found in many dehydrogenases [30], and the histograms in Figure E1 (p. 94-95) identified several different features related to this enzyme class as frequent predictors. These included the conserved cysteine active site of aldehyde dehydrogenase, two superfamilies of domains found in aldehyde dehydrogenases, and one superfamily of domains found in both aldehyde and histidinol dehydrogenases. Due to the same reasoning applied to the case of PLP-dependent transferase superfamilies, the high documentation rate of this feature in the current dataset appears perfectly valid.

As previously mentioned, Figure D1 demonstrates that the feature types domain, family, and superfamily are those that are statistically significantly associated with the highest numbers of interactions. Similar as for superfamilies, it is plausible that domains and families are more frequently annotated to proteins than less generic feature types such as active sites and binding sites. This is because the more generic features might be easier to detect and characterize as their identification does not require as much information about the protein as features that constitute more specific sequential units. However, a high annotation frequency means that these features are most likely annotated to proteins with allosteric effectors just as often as they are present in proteins that do not have allosteric effectors. This type of behavior is accounted for by Fisher's exact test, as both the numerator and denominator of Eq. 1 (see Subsec. 2.2.2.1, p. 28) increase. Therefore, these results indicate that the feature types domain, family, and superfamily are more often statistically significantly associated with metabolite-interactions due to factors other than their frequency, and that the connection between interactions and protein structure is stronger for these feature types.

Despite the prevalence of superfamilies in the protein feature data and as an interaction-predictor, they were shown to be excessive for the prediction of allosteric interactions. As was established in Subsection 3.2.2, Figure D1 (p. 93) shows that the only two feature types which predicted interactions not covered by any others are domain and family. This means that these two feature types together provide all information supplied by superfamilies and the remaining feature types. These observations thereby indicate that even though superfamilies are well documented and thus provide a substantial amount of information about the interactions of proteins, they are not unique contributors. Unique interactions are instead predicted by features belonging to the types of domain and family, which further implies that these two categories are the most important feature types to assess when predicting protein-metabolite interactions from protein structure.

4.2.2 Biological validity of protein feature - interaction associations

Among 57 908 associations between 5 827 unique protein structural features and 13 737 metabolite-interactions, the conducted enrichment analysis identified 32 276 statistically significant associations. Further 7 315 of these associations were deemed exclusive. The exclusive associations are constituted by situations where there are no occurrences of the feature being present when the interaction is not present, which results in an infinite odds ratio due to the division by zero in Eq. 1 (see Subsec. 2.2.2.1, p. 28). For plotting purposes, the odds ratio of these incidents was reduced to 15 000 (15e3 in Table 3.7, p. 52-54), corresponding to log(OR) > 4 in Figure 3.12 (p. 50). Of the associated features

and interactions assessed in Table 3.7, twelve were exclusive, while the remaining 33 were non-exclusively significant. As the exclusive association of a feature with an interaction indicates a strong connection between the two variables, it is of interest to determine whether these exclusive associations are in fact more accurate than those that are merely highly significant. This Subsection is therefore devoted to further examination of the biological validity of the associations reviewed in Table 3.7, with the additional aim of elucidating the biological accuracy of the utilized approach in general.

As indicated by the biological explanations given in Table 3.7, two of the evaluated exclusive associations represent actual allosteric protein-metabolite interactions. These were inhibition of PEP carboxylase by aspartic acid, and the inhibition of aspartate kinase by threonine. The latter interaction was represented by several feature-interaction associations, of which all were exclusive, while two exclusive associations were evaluated for inhibition of PEP carboxylase by aspartic acid. Additionally, literature confirms that NADPH acts as an inhibitor of 6-phosphogluconate dehydrogenase, and that gluconate 6-phosphate is a competitive inhibitor of phosphoglucose isomerase. While only one association was evaluated for NADPH-inhibition of 6-phosphogluconate dehydrogenase, several exclusive associations were found concerning phosphoglucose isomerase inhibition. The exclusive association of the ribonucleotide reductase small subunit active site with ATP is also biologically valid, as ATP is one of several allosteric effectors which ensures substrate specificity of the enzyme.

While these biologically justifiable findings indicate a strong accuracy of exclusive associations, two feature-interaction associations represent PMIs that could not be readily explained. These were inhibition of aconitase by oxalosuccinate and inhibition of glycerol kinase by glycerophosphate. Even though the structure of the biological pathways in which these enzymes take part suggests that the metabolites might exert negative feedback, no such evidence was found. This means that these feature-interaction associations could not be certainly verified by the regulatory mechanisms of their related enzymatic entities according to existing literature. Nevertheless, it cannot be excluded that these associations might still represent the bindings of metabolites to proteins. As previously stated, lacking research on the protein-metabolite interactome is an essential issue, and it might therefore be possible that the associations detected in this analysis represent PMIs that have simply not been adequately studied.

In addition to the feature-interaction associations regarded above, several non-exclusive associations could also be validated by the regulatory mechanisms of the enzymes to which the features are related. For example, different features specific to ADP-glucose pyrophosphorylase (glucose-1-phosphate adenylyltransferase) were statistically associated with activation by 3-phosphoglycerate, which is a documented allosteric activator of the enzyme. Furthermore, the active site of the ribonucleotide reductase small subunit was statistically associated with several documented substrate-specific allosteric activators of ribonucleotide reductase, and the active site of pyruvate kinase was statistically associated with activation by its allosteric modulator fructose 1,6-bisphosphate. While several of the non-exclusive feature-interaction associations could not be explained with confidence based on characteristics of the enzymes related to the features, it is still, as for the exclusive associations, possible that these associations represent PMIs that take place either naturally or under in vitro conditions. As this analysis did not consider all associated features and interactions, it is from these findings difficult to determine whether exclusive associations are generally more accurate than non-exclusive ones. No definite conclusion can therefore be drawn regarding whether exclusive associations represent true allosteric behavior more frequently than non-exclusive, significant associations.

Another noteworthy metabolite-interaction that was found to be statistically associated with protein features is cAMP-activation. As implied by Table 3.7, this interaction was associated with a high number of protein features related to the protein class of protein kinases. This class includes the cyclic-AMP dependent protein kinase that depends on cAMP for activity stimulation [126], which thus explains the association of cAMP with general protein kinase features. One of these cAMP-activation-associated features, the protein kinase-like domain superfamily, was also identified as one of the overall top ten documented protein features (see Table 3.6, p. 49), which provides an indication that general protein kinase features are well documented. This may explain why cAMP-activation is associated with several, and not just one or two, protein kinase features. The overall frequent association of cAMP-activation with protein features is further clarified by Figure E2 (see App. E, p. 96-98), which recognized a high number of predicting features for this interaction among several different feature types.

While the conducted review identified many biologically valid protein feature-interaction associations, there were also several suggested effectors for which no certain biological connection could be found to their associated proteins. This problem might have been caused by the issue discussed in Subsection 4.1.4, namely the uncertain quality of entries reported in BRENDA. As was previously discussed, BRENDA contains several questionable documentations, including reported interactions of proteins and metabolites that are either inconsistent with other studies, condition-dependent, or of a weak nature. This problem affects the results of this part of the analysis because features may be associated with interactions even though they are naturally or biologically unrelated. This complicates the correct association of protein structural features with metabolite-interactions, and the identification of seemingly unrelated proteins and interactions implies that the approach should be improved in order to be used confidently as the basis for predicting protein-metabolite interactions. Such improvements could for example include a more sophisticated approach to associating protein features with metabolite-interactions, in which only those that are biologically relevant are used for predictions. On the other hand, the utilized approach did, as mentioned, also identify several true protein-metabolite interactions, and thus demonstrates a biologically significant relationship between protein structure and metabolite-interaction which can be exploited to predict allosteric interactions.

Additionally, the identified feature-interaction associations represent several protein-metabolite interactions that do occur in in vivo systems, but are not of an allosteric character. As was mentioned in the Chapter of Methods, filtering for purely allosteric interactions in BRENDA was not possible. Despite removing extracellular and inorganic compounds from the dataset during its compilation, the analysis in Subsection 3.1.2 revealed that a great number of cofactors and competitive inhibitors remained present. Due to these findings, it was expected that the feature-interaction association would result in the association of protein features with interactions of cofactors and competitive inhibitors. While this issue does not affect the ability of this project to demonstrate a biologically significant relationship between protein structure and metabolite-interaction, the filtering process should be improved to remove such interactions if the current approach is to be used as the basis for a method of predicting allosteric regulation.

Lastly, the predictions of metabolite-interactions from protein features included a few interactions that are coherent with the findings from the analysis of the protein-metabolite interaction data. As was mentioned in Subsection 3.2.2.1, the histograms in Figure E2 (see App. E, p. 96-98) identified several metabolite-interactions that were predicted by high numbers of features. These interactions included activation by ATP and pyruvate and inhibition by glucose. As was established in Subsections 3.1.2 and 4.1.1, the results from

this current work and that conducted by Reznik *et.al.* identified ATP as the most common regulator. Reznik *et.al.* also recognized pyruvate as a common effector molecule [27], and the network in Figure B1 (see App. B, p. 90) indicates a central role of glucose as a metabolic regulator as well. The high association rate of these interactions with protein features is thus coherent with their role as frequent protein modulators, which suggests that the approach accomplished to detect, at least to some degree, a logical connection between protein structure and metabolite-interaction. The abundance of associated protein features for these central metabolic regulators may also be interpreted as further evidence of the non-random distribution of both enzyme regulation and research focus that has been repeatedly suggested in this Chapter.

4.2.3 Predicted interactions closed phylogenetic gaps

The expanded phylogenetic tree in Figure 3.13 (p. 55) contained 132 predicted interactions, which was 82 interactions less than the number of documentations. With the assumption that these predictions are biologically accurate, this high number of predictions provides an indication of a statistically significant relationship between protein features and metabolite interactions which can be exploited to predict allosteric protein-metabolite interactions from protein structure. On the other hand, had the approach not yielded any predictions, that would have suggested that there exists no connection between the presence of protein structural features and the regulation of proteins by metabolites. This could have meant that the protein structure characteristics utilized in the current work are too generic, but as the approach did yield predictions, no such conclusions can be drawn.

Furthermore, the annotated predicted interactions appear to not have changed the general clustering pattern that was observed in Figure 3.10 (p. 46). Interactions are still evenly spread throughout all taxa, and if any change was made, annotating the predictions actually caused interactions to be even more spread than previously. For example, the interactions defined as inhibition of fructose-1,6-bisphosphatase by AMP (grey), activation of pyruvate kinase by fructose 1,6-bisphosphate (dark pink), and inhibition of acetolactate synthase by valine (blue) were not present in Archaea in Figure 3.10. However, due to the prediction of these interactions in the species *Halobacterium salinarum*, *Methanocaldococcus jannaschii*, and *Methanothermobacter thermautotrophicus* and *Methanocaldococcus jannaschii*, respectively, Archaea was shown to also possess these three PMIs. The annotation of predicted interactions to the phylogenetic tree thus resulted in a higher conservation of protein-metabolite interactions, indicating that distant species are more closely related in terms of their protein-regulatory mechanisms than what is represented by documented research.

As was described in Subsection 3.2.2.2, mapping predicted interactions to the phylogenetic tree closed some of the gaps that were mentioned in Subsection 3.1.3 and discussed in Subsection 4.1.3. For example, the interactome of *Arabidopsis thaliana* was expanded from only four to nine out of the top ten interactions that were selected as a case study. This new interactome better represents *A. thaliana* as a common model organism, as in concordance with the results described in Subsections 3.1.2 and 4.1.1. A similar case is found for the model organism *Drosophila melanogaster*, which advanced from zero to four annotations. Furthermore, two close relatives of the model yeast *Saccharomyces cerevisiae* were predicted to possess three and eight of the interactions documented for *S. cerevisiae*, which displays a clear case of gap-filling in related species. Other such examples are found in the phylum Chordata, which includes mammals and *Gallus gallus*, and in the bacterial taxa as well. These findings indicate that the absence of interactions from certain species might in fact be caused by lacking documentation, and thus support the previously stated

hypothesis of both biased and incomplete research focus. However, it was revealed in Subsection 3.2.2.2 that not all of the predicted interactions are biologically valid. This issue of false positive predictions will be addressed in the following Subsection.

4.2.4 Interactions were both correctly and falsely predicted

The utilized approach resulted in the prediction of at least three biologically correct organism-specific protein-metabolite interactions. As was validated by the obtained evidence described in Subsection 3.2.2.2, the pyruvate kinase of both the bacterium *Mycobacterium tuberculosis* and the eukaryote *Neurospora crassa* is in fact activated by fructose 1,6-bisphosphate. Additionally, although the article that was obtained as evidence did not describe the relevant enzyme, it was found that the hexokinase of *Plasmodium falciparum* is inhibited by its product glucose 6-phosphate. These true positive predictions function as further indicators of the validity of the current approach, and suggest that the utilized structural protein features have potential as predictors of allosteric interactions. However, in addition to these validations, the obtained evidence also prompted the identification of four false positive predictions. The remains of this Subsection are devoted to attempting to uncover what might cause such false positive predictions to occur, as well as to consider this issue in relation to the confidence of the current approach.

The results from the conducted validation process showed that all false positive predictions were related to the same interaction, namely activation of pyruvate kinase by fructose 1,6-bisphosphate (dark pink in Fig. 3.13). The obtained literature states that the pyruvate kinases of the organisms in question are of the AMP-activated and not fructose 1,6-bisphosphate-activated types, which implies that even though these specific pyruvate kinases are not inhibited by fructose 1,6-bisphosphate, there exist several pyruvate kinases that are. The main reason for these false positive predictions most likely resides in the utilization of EC numbers as the protein identifier. EC numbers represent catalytic reactions and not specific enzymes, and some reactions may be catalyzed by more than one unique protein. In some cases, these enzymes display the same regulatory behaviors, but there also exist cases where the different proteins of the same catalytic function are regulated by different mechanisms. Because the proteins documented in BRENDA are possibly not properly standardized by name, EC number was chosen as the simplest and most accurate identifier. However, this makes it impossible to distinguish between what specific proteins are actually regulated by the metabolite in question, as in the case of AMP- and fructose 1,6-bisphosphate-activated pyruvate kinases. This essentially means that modulators are associated with all proteins related to an EC number, despite possible differences in their regulation. While such an assumption of universal regulators for an EC number might be valid for interactions of some enzymes, it may result in false positive predictions in cases where proteins related to the same EC number are regulated by different mechanisms.

Furthermore, it is possible that the false positive predictions of pyruvate kinase inhibition were based on association with features that are too generic to be used confidently. Feature types such as superfamilies are more generic than the others in the sense that they encompass proteins of only similar structure. This implies that two proteins of the same superfamily might be entirely unrelated in terms of biological function and small-molecule regulation. Domains, on the other hand, are more directly related to specific biological roles, as is also the case of features such as active sites and binding sites. This means that interactions associated with certain superfamilies might be specific to only parts of the group of proteins to whom this feature is annotated, while interactions associated with domains are more likely to be directly related to a biological function or characteristic that determines or is connected to a specific regulatory behavior. Therefore, it is plausible that

some feature types might be too generic to be utilized as predictors with confidence, simply because they are not sufficiently specific in terms of their annotated proteins and their characteristics.

The issues of non-protein- and non-interaction-specific associations described in this Subsection can be argued as a consequence of the general research focus of this thesis. Focusing on conservation at the level of phylogenetic taxa rather than the molecular level provides less detail about the connection between protein structure and metabolite-interaction, which again reduces the confidence of predictions regarding prevalent proteins that are regulated by multiple organism-dependent metabolites. However, focusing on the taxa level also reduces the complexity of the analysis. As was elaborated upon in the introduction, a great deal of the problem with studying allostery is the low throughput of existing methods relative to the vast interaction space that is to be uncovered. By studying protein-metabolite interactions on a genome scale for a larger group of organisms, this relationship between throughput and interaction space is much more equalized. Utilizing relatively simple protein characteristics such as features as the predicting variable also makes the approach more applicable and easier to execute on a larger scale than existing methods, at least compared to those that focus mainly on single protein-metabolite interactions. However, the prediction of false positive interactions does imply that there is a certain inaccuracy to the approach. While a payoff between accuracy and simplicity is always to be expected with these types of analyses, it would be possible to improve the current method by applying stricter demands to the utilized features. As was previously established, the main inaccuracy is likely related to different regulatory mechanisms of taxa-wise conserved proteins related to the same EC numbers. Thus, if one could identify features responsible for the binding of specific metabolites for proteins whose regulation is highly organism-dependent, the prediction of these interactions could be restricted to those features. Nevertheless, the identification of experimentally proven organism-dependent interactions does provide evidence that general protein features have the potential as allosteric interaction-predictors, implying that the approach's simplicity is not an overall weakness.

In regard to false positive predictions, it should also be acknowledged that the current approach does not attempt at predicting allosteric interactions. Due to time limitations, this project was conducted utilizing a naive approach of predicting protein-interaction associations based on threshold values in the volcano plots of Figure 3.12. The cases of predicted interactions discussed in this Subsection are only part of a small case study that was performed as a mode of demonstrating the potential of utilizing protein features to predict allosteric interactions. This type of naive approach is bound to result in both false positive and false negative associations due to issues with the inaccuracy of associations, and these issues would have been better handled by a method based on machine learning that was trained to have better accuracy. This will be elaborated upon in Chapter 5.

Lastly, the inhibition of hexokinase by glucose-6-phosphate (green) was predicted for several different eukaryotic species. As previously acknowledged, hexokinase is a common protein that exists in different sizes and with different specificities. Literature implies that only vertebrate 100-kDa hexokinase is inhibited by glucose-6-phosphate [159], but BRENDA documents the interaction for species of other groups as well. Similar to the case of pyruvate kinase, the prediction of this interaction in non-vertebrate species might be due to the presence of similar catalytical, hexokinase-specific, or generic structural features. However, this interaction was as mentioned also predicted for the parasite *Plasmodium falciparum*, and this prediction was later confirmed by additional literature reviews. In fact, the *P. falciparum* hexokinase is declared a 55.3-kDa protein with 26% identity to that of humans [148], and should according to certain literature thereby not be inhibited by glu-

cose 6-phosphate [159]. However, the results from another study imply that there is no difference between the species with 50- and 100-kDa hexokinases in relation to inhibition by glucose 6-phosphate [117]. It thus appears as if this regulatory behavior is not limited to higher eukaryotes after all, but might be more widespread than anticipated. As most of the reviewed articles about the regulation of this specific enzyme were published several decades ago, this might be an indication that previous research is somewhat outdated. This matter will be part of the topic addressed in the following Subsection.

4.2.5 Poor documentation and research of allostery

An observation made about the literature that has been reviewed in the entirety of this thesis is that the majority of the utilized articles were published either before or in the early 2000s. Much of the existing allosteric and enzymatic research appears to have been performed several decades ago, and there seems to be a significant shortage of compulsory research efforts on these topics. In fact, the general amount of literature available for allosteric regulation seem sparse compared to other fields of biology, including topics such as transcriptional regulation. This assumption is supported by the results of the validation process described in Subsection 3.2.2.2, which consisted of only ten articles concerning the prediction of two protein-metabolite interactions in seven individual organisms. The file of predicted interactions used to conduct the search included ten interactions in 60 individual organisms. When also considering the fact that four of the retrieved articles described the same organism-specific interaction and that one of the retrieved articles was irrelevant to its case, little relevant information was obtained relative to the potential. Despite the high simplicity of this search, these results, in combination with the biased research focus that has been proposed in several previous sections of this thesis, is further indication of not only biased research within the field of allostery, but also of poor attention paid towards discovering and studying regulatory protein-metabolite interactions in general. As was also argued in Subsection 4.1.4, these circumstances might promote outdated and misleading scientific conclusions due to the use of incomplete and inaccurate metabolic models.

5 Conclusion and Outlook

The aim of this thesis, as was described in Chapter 1, was to evaluate the potential of predicting allosteric interactions from genome sequences. In order to achieve this objective, data on protein-metabolite interactions was first obtained and assembled into a standardized database. Following analysis of its contents, structural features annotated to the proteins of the database were retrieved, and connections were drawn between these features and interactions with these proteins in terms of the regulatory metabolite and mode. This current Chapter will summarize the efforts and highlight the most essential results and conclusions from each part of the work, followed by descriptions of possible further uses and advancements of the developed approach.

The utilized protein-metabolite interaction data was retrieved from the publicly available enzymatic database BRENDA. After filtering for intracellular and organic compounds, this assembled database contained 32 535 organism-specific interactions among 3 097 proteins and 1 002 metabolites. 18 854 of these interactions were unique, and 78.7% of the non-unique interactions were inhibitory. Comparing the contents of this database with findings from the study of small-molecule regulation in *E. coli* by Reznik *et.al.* indicated a general abundance of inhibitory regulation regardless of the considered organism. These findings are likely caused by the widespread use of competitive inhibition and feedback regulation in metabolism in order to maintain condition-suitable levels of resources.

Furthermore, the contents of the assembled database were analyzed by studying frequency distributions of the activators, inhibitors, targeted enzymes, and documented organisms. The results from this work showed that the typical metabolite and enzyme only participate in a few interactions, while the few effector molecules and proteins that are highly regulatory and regulated are essential for central pathways such as glycolysis, gluconeogenesis, the citric acid cycle, and glycogen/starch metabolism. These results are indications of two possible hypotheses; central pathways are either subject to more regulatory attention due to the conservation of resources by feedback mechanisms, or they are subject to more research due to their essential role in the maintenance of cellular life. Similar patterns were observed for the documented organisms, where the skewed distribution of interactions reported indicated a biased research focus towards typical model organisms and species related to *Homo sapiens*.

With the purpose of assessing the conservation of allosteric interactions on the level of phylogenetic taxa, ten well-reported interactions from the assembled database were annotated to a phylogenetic tree of the top documented organisms. This work resulted in a tree of 99 individual organisms annotated with a total of 214 interactions. Important takeaways from this result included well conservation of protein-metabolite interactions across taxa, and the presence of interaction-gaps at several species whose relatives were documented with well-conserved interactions. These observations are further indicators of a biased research focus toward central species, which implies that missing interactions may represent lacking documentation or research rather than lacking biological regulation.

To evaluate the potential of utilizing genome sequence as a predictor of allosteric regulation, protein structural features represented by InterPro IDs were retrieved from UniProt for all EC numbers documented for the top hundred organisms. These features were associated with their EC numbers' respective metabolite-interaction, and enrichment analysis was performed to find statistically significant relationships between features and interactions. This resulted in 32 276 statistically significant associations, of which 7 315 were exclusive. Investigating the overlap of predicted interactions of the eight different feature types

showed that domain and family are the two most important interaction-predicting features as they produced both many and unique predictions. Moreover, several feature-interaction associations were proven to be biologically valid, but no definite indications were found of exclusive predictions being more accurate than non-exclusive, significant associations.

Predicted interactions identified from enrichment analysis were annotated to the previously created phylogenetic tree in order to assess whether such predictions are capable of closing observed gaps in conserved interactions. This work resulted in 132 predicted interactions, which gave the impression of enhanced conservation of allostery due to a more evenly spread distribution of interactions and several closed phylogenetic gaps. These results were further indicators of the previously suggested issue of biased research, and also of lacking documentation of allostery. Additionally, the number of predicted interactions in the tree as well as the biological validity of associations indicate that the utilized protein structure features are valid predictors of protein-metabolite interactions. Searching for experimental proof of predicted interactions did however reveal a few false positive predictions that might be the result of associations based on generic features in terms of their relation to enzyme function or widespread association.

Lastly, the literature obtained for this work indicates an abundance of outdated research and highlights the need for newer efforts within the field of allosteric regulation. The utilized database of enzymatic information also displays a need for quality control of its entries, as findings of possibly biologically inaccurate information indicate a somewhat naive approach to information retrieval and documentation. In addition to the lack of documentation that has previously been suggested, these circumstances might promote outdated and misleading scientific conclusions which could be resolved by better allosteric knowledge.

If this project was given a few more months, it would have been of high interest to develop a machine learning tool that exploits the statistically significant relationship between protein structure features and protein-metabolite interactions to predict allostery from protein structure. While the current approach proposes the existence of such a relationship, it would have been beneficial to have a tool that directly suggests metabolic regulators for a protein given its annotated structural features. Moreover, as the conducted work also resulted in false positive predictions for the considered case study, it could be advantageous to investigate these connections more thoroughly to possibly determine factors that cause the prediction of these protein-metabolite interactions to be more challenging and inaccurate. This could either confirm or reject the hypotheses suggested in Subsection 4.2.4, or uncover alternative causes for these false positives. Knowledge about these factors could further be used to improve the method, and solutions could thereby be implemented in the machine learning tool to reduce the uncertainty of predictions.

Furthermore, as was mentioned introductory-wise, improved allosteric knowledge would likely contribute to increased accuracy of metabolic models. Therefore, it could be interesting to include some of the predicted interactions in a genome scale metabolic model for one of the species that, according to this analysis, is less studied allostery-wise. The utilized organism could for example be *Bacillus subtilis*, which only had three documentations but six additional predicted interactions. In order to assess the effect of protein-metabolite interactions on the metabolic flux, flux balance analysis could be run and compared between original and interaction-extended models. The allosteric interactions could be implemented in the analysis as additional constraints on the enzymatic activity, for example by using the concentration of essential effector molecules as additional constraining variables. Results from these analyses could then be used to deduce the effect of including regulatory protein-metabolite interactions, and possibly lead to a more accurate metabolic model of the organism in question.

References

1. Cobb M. 60 years ago, Francis Crick changed the logic of biology. PLoS Biol 2018; 15. DOI: 10.1371/journal.pbio.2003243.
2. Daran-Lapujade P, Rossell S, Gulik W van, Luttik M, Groot M de, Slijper M, Heck A, Daran J, Winde J de, Westerhoff H, Pronk J, and Bakker B. The fluxes through glycolytic enzymes in Saccharomyces cerevisiae are predominantly regulated at post-transcriptional levels. Proceedings of the National Academy of Sciences 2007; 104:15753–8. DOI: 10.1073/pnas.0707476104.
3. Lu S, Shen Q, and Zhang J. Allosteric Methods and Their Applications: Facilitating the Discovery of Allosteric Drugs and the Investigation of Allosteric Mechanisms. Accounts of Chemical Research 2019; 52:492–500. DOI: 10.1021/acs.accounts.8b00570.
4. Link H, Kochanowski K, and Sauer U. Systematic identification of allosteric protein-metabolite interactions that control enzyme activity *in vivo*. Nature Biotechnology 2013; 31:357–61. DOI: 10.1038/nbt.2489.
5. Cooperman B. Allosteric Regulation. *Encyclopedia of Biological Chemistry (Second Edition)*. Ed. by Lennarz W and Lane M. Second Edition. Waltham: Academic Press, 2013 :71–4. DOI: 10.1016/B978-0-12-378630-2.00001-3.
6. Liu J and Nussinov R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. PLOS Computational Biology 2016; 12:1–5. DOI: 10.1371/journal.pcbi.1004966.
7. Goodey N and Benkovic S. Allosteric regulation and catalysis emerge via a common route. Nature chemical biology 2008; 4:474–82. DOI: 10.1038/nchembio.98.
8. Lindsley J and Rutter J. Whence cometh the allosterome? Proceedings of the National Academy of Sciences 2006; 103:10533–5. DOI: 10.1073/pnas.0604452103.
9. Chubukov V, Uhr M, Le Chat L, Kleijn R, Jules M, Link H, Aymerich S, Stelling J, and Sauer U. Transcriptional regulation is insufficient to explain substrate-induced flux changes in Bacillus subtilis. Molecular Systems Biology 2013; 9:709. DOI: 10.1038/msb.2013.66.
10. Kochanowski K, Sauer U, and Chubukov V. Somewhat in control—the role of transcription in regulating microbial metabolic fluxes. Current Opinion in Biotechnology 2013; 24:987–93. DOI: 10.1016/j.copbio.2013.03.014.
11. Diether M and Sauer U. Towards detecting regulatory protein–metabolite interactions. Current Opinion in Microbiology 2017; 39:16–23. DOI: 10.1016/j.mib.2017.07.006.
12. Pai Y, Lomenick B, Hwang H, Schiestl R, McBride W, Loo J, and Huang J. Drug Affinity Responsive Target Stability (DARTS) for Small-Molecule Target Identification. *Chemical Biology: Methods and Protocols*. Ed. by Hempel J, Williams C, and Hong C. New York, NY: Springer New York, 2015 :287–98. DOI: 10.1007/978-1-4939-2269-7_22.
13. Gholizadeh E, Rezaei tavirani M, Emadi A, Karbalaei R, and Khaleghian A. A New Drug Discovery Approach Based on Thermal Proteome Profiling to Develop More Effective Drugs. Middle East Journal of Rehabilitation and Health Studies 2021; 8. DOI: 10.5812/mejrh.113533.
14. Roelofs K, Wang J, Sintim H, and Lee V. Differential radial capillary action of ligand assay for high-throughput detection of protein-metabolite interactions. Proceedings of the National Academy of Sciences 2011; 108:15528–33. DOI: 10.1073/pnas.1018949108.
15. Diether M, Nikolaev Y, Allain F, and Sauer U. Systematic mapping of protein-metabolite interactions in central metabolism of Escherichia coli. Molecular Systems Biology 2019; 15. DOI: 10.15252/msb.20199008.

16. Song K, Liu X, Huang W, Lu S, Shen Q, Zhang L, and Zhang J. Improved Method for the Identification and Validation of Allosteric Sites. Journal of chemical information and modeling 2017; 57. DOI: 10.1021/acs.jcim.7b00014.
17. Sheik Amamuddy O, Veldman W, Manyumwa C, Khairallah A, Agajanian S, Oluyemi O, Verkhivker G, and Tastan Bishop Ö. Integrated Computational Approaches and Tools for Allosteric Drug Discovery. International Journal of Molecular Sciences 2020; 21. DOI: 10.3390/ijms21030847.
18. Freire E. Can allosteric regulation be predicted from structure? Proceedings of the National Academy of Sciences 2000; 97:11680–2. DOI: 10.1073/pnas.97.22.11680.
19. Hardy J and Wells J. Searching for new allosteric sites in enzymes. Current Opinion in Structural Biology 2004; 14:706–15. DOI: 10.1016/j.sbi.2004.10.009.
20. Süel G, Lockless S, Wall M, and Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nature structural biology 2003; 10:59–69. DOI: 10.1038/nsb881.
21. Guarnera E and Berezovsky I. Structure-Based Statistical Mechanical Model Accounts for the Causality and Energetics of Allosteric Communication. PLOS Computational Biology 2016; 12:1–27. DOI: 10.1371/journal.pcbi.1004678.
22. Shen Q, Wang G, Li S, Liu X, Lu S, Chen Z, Song K, Yan J, Geng L, Huang Z, Huang W, Chen G, and Zhang J. ASD v3.0: unraveling allosteric regulation with structural mechanisms and biological networks. Nucleic Acids Research 2015; 44:D527–D535. DOI: 10.1093/nar/gkv902.
23. Huang W, Wang G, Shen Q, Liu X, Lu S, Geng L, Huang z, and Zhang J. ASBench: Benchmarking sets for allosteric discovery. Bioinformatics (Oxford, England) 2015; 31. DOI: 10.1093/bioinformatics/btv169.
24. Huang W, Lu S, Huang z, Liu X, Mou L, Luo Y, Zhao Y, Liu Y, Chen Z, Hou T, and Zhang J. Allosite: A method for predicting allosteric sites. Bioinformatics (Oxford, England) 2013; 29. DOI: 10.1093/bioinformatics/btt399.
25. Li S, Shen Q, Su M, Liu X, Lu S, Chen Z, Wang R, and Zhang J. Alloscore: a method for predicting allosteric ligand–protein interactions. Bioinformatics 2016; 32:1574–6. DOI: 10.1093/bioinformatics/btw036.
26. Liu X, Lu S, Song K, Shen Q, Ni D, Li Q, He X, Zhang H, Wang Q, Chen Y, Li X, Wu J, Sheng C, Chen G, Liu Y, Lu X, and Zhang J. Unraveling allosteric landscapes of allosterome with ASD. Nucleic Acids Research 2019; 48:D394–D401. DOI: 10.1093/nar/gkz958.
27. Reznik E, Christodoulou D, Goldford J, Briars E, Sauer U, Segrè D, and Noor E. Genome-scale architecture of small molecule regulatory networks and the fundamental trade-off between regulation and enzymatic activity. Cell reports 2017; 20:2666–77. DOI: 10.1016/j.celrep.2017.08.066.
28. Machado D, Herrgård M, and Rocha I. Modeling the Contribution of Allosteric Regulation for Flux Control in the Central Carbon Metabolism of E. coli. Frontiers in Bioengineering and Biotechnology 2015; 3. DOI: 10.3389/fbioe.2015.00154.
29. Kell D. Metabolites do social networking. Nature chemical biology 2011; 7:7–8. DOI: 10.1038/nchembio.505.
30. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto B, Salazar G, Bileschi M, Bork P, Bridge A, Colwell L, Gough J, Haft D, Letunić I, Marchler-Bauer A, Mi H, Natale D, Orengo C, Pandurangan A, Rivoire C, Sigrist C, Sillitoe I, Thanki N, Thomas P, Tosatto S, Wu C, and Bateman A. InterPro in 2022. Nucleic Acids Research 2022; 51:D418–D427. DOI: 10.1093/nar/gkac993.
31. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, and Schomburg D. BRENDA, the ELIXIR core data resource in 2021:

new developments and updates. Nucleic Acids Res. 2021; 49:D498–D508. DOI: 10.1093/nar/gkaa1025.
32. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 2020; 49:D480–D489. DOI: 10.1093/nar/gkaa1100.
33. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, and Willing C. Jupyter Notebooks – a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by Loizides F and Schmidt B. IOS Press. 2016 :87–90
34. Corporation M. Microsoft Office Excel. Version 2208. [software]. Released 1987; Updated February 14th 2023. Available from https://www.microsoft.com/nb-no/microsoft-365/excel.
35. The pandas development team. pandas-dev/pandas: Pandas 1.2.2. Version 1.2.2. [software]. Zenodo. Released December 25th 2009; Updated February 9th 2021. DOI: 10.5281/zenodo.4524629.
36. McKinney W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*. Ed. by Walt S van der and Millman J. 2010 :56–61. DOI: 10.25080/Majora-92bf1922-00a.
37. Hunter J. Matplotlib: A 2D graphics environment. Computing in Science & Engineering 2007; 9:90–5. DOI: 10.1109/MCSE.2007.55.
38. Waskom M. seaborn: statistical data visualization. Journal of Open Source Software 2021; 6:3021. DOI: 10.21105/joss.03021.
39. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, and Bourne P. The Protein Data Bank. Nucleic Acids Research 2000; 28:235–42. DOI: 10.1093/nar/28.1.235. Available from http://www.rcsb.org/.
40. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, and Steinbeck C. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research 2015; 44:D1214–D1219. DOI: 10.1093/nar/gkv1031.
41. Heller S, McNaught A, Pletnev I, Stein S, and Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. 2015; 7. DOI: 10.1186/s13321-015-0068-4.
42. King Z, Lu J, Dräger A, Miller P, Federowicz S, Lerman J, Ebrahim A, Palsson B, and Lewis N. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Research 2015; 44:D515–D522. DOI: 10.1093/nar/gkv1049.
43. Britannica, The Editors of Encyclopedia. organic compound. *Encyclopedia Britannica*. [Internet]. Updated October 19th 2022; Accessed December 9th 2022. Available from https://www.britannica.com/science/organic-compound.
44. Unpingco J. pyvis. Version 0.3.1. [software]. Released May 15th 2018; Updated November 11th 2022. Available from https://pypi.org/project/pyvis/0.3.1/. Released May 15th 2018; Updated November 11th 2022.
45. Letunic I. PhyloT. Version 2022.3. [software]. Available from https://phylot.biobyte.de/.
46. Schoch C, Ciufo S, Domrachev M, Hotton C, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan J, Sun L, Turner S, and Karsch-Mizrachi I. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020; 2020. baaa062. DOI: 10.1093/database/baaa062.
47. Huerta-Cepas J, Serra F, and Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution 2016; 33:1635–8. DOI: 10.1093/molbev/msw046.

48. Letunic I and Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research 2021; 49:W293–W296. DOI: 10.1093/nar/gkab301.
49. Koshland D and Haurowitz F. protein. *Encyclopedia Britannica*. [Internet]. Updated October 26th 2022; Accessed March 29th 2023. Available from https://www.britannica.com/science/protein.
50. Sangrador-Vegas A and Mitchell A. Protein classification: An introduction to EMBL-EBI resources. 2011. DOI: 10.6019/tol.prc.2011.00001.1.
51. Cokelaer T. bioservices. Version 1.10.4. [software]. Released February 27th 2013; Updated October 5th 2022. Available from https://pypi.org/project/bioservices/1.10.4/.
52. Cock P et.al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009; 25:1422–3. DOI: 10.1093/bioinformatics/btp163.
53. Petrov A. urllib3. Version 1.26.12. [software]. Released November 24th 2008; Updated August 22th 2022. Available from https://pypi.org/project/urllib3/.
54. Lopez A. How to perform a gene set enrichment analysis (GSEA). [Internet]. Updated September 8th 2016; Accessed March 29th 2023. Available from https://www.biobam.com/gene-set-enrichment-analysis/?cn-reloaded=1.
55. Warner P. Testing association with Fisher's Exact test. BMJ Sexual & Reproductive Health 2013; 39:281–4. DOI: 10.1136/jfprhc-2013-100747.
56. Freeman J and Campbell M. The analysis of categorical data: Fisher's exact test. Scope 2007; 16:11–2. Available from: https://www.researchgate.net/publication/237336173_The_analysis_of_categorical_data_Fisher's_exact_test.
57. Cansiz S. Interpretation of Odds Ratio and Fisher's Exact Test. [Internet]. Updated December 11th 2020; Accessed January 12th 2023. Available from https://towardsdatascience.com/interpretation-of-odds-ratio-and-fishers-exact-test-c6dde394d204.
58. Szumilas M. Explaining odds ratios. Journal of the Canadian Academy of Child and Adolescent Psychiatry 2010; 19. PMID: 20842279.
59. Noble W. How does multiple testing correction work? Nature Biotechnology 2009; 27. DOI: 10.1038/nbt1209-1135.
60. Hazra A. Using the confidence interval confidently. Journal of thoracic disease 2017; 9:4125. DOI: 10.21037/jtd.2017.09.14.
61. Aggarwal S and Yadav A. False Discovery Rate Estimation in Proteomics. *Statistical Analysis in Proteomics*. Ed. by Jung K. New York, NY: Springer New York, 2016 :119–28. DOI: 10.1007/978-1-4939-3106-4_7.
62. Storey J and Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 2003; 100:9440–5. DOI: 10.1073/pnas.1530509100.
63. SciPy. scipy.stats.fisher_exact. [Internet]. Accessed March 14th 2023. Available from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html\#scipy.stats.fisher_exact.
64. Virtanen P, Gommers R, Oliphant T, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt S, Brett M, Wilson J, Millman K, Mayorov N, Nelson A, Jones E, Kern R, Larson E, Carey C, Polat İ, Feng Y, Moore E, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero E, Harris C, Archibald A, Ribeiro A, Pedregosa F, van Mulbregt P, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 2020; 17:261–72. DOI: 10.1038/s41592-019-0686-2. Available from https://rdcu.be/b08Wh.

65. Seabold S and Perktold J. statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*. 2010. DOI: 10.25080/MAJORA-92BF1922-011.
66. Statsmodels. statsmodels.stats.multitest.fdrcorrection. [Internet]. Accessed March 14th 2023. Available from https://www.statsmodels.org/stable/generated/statsmodels.stats.multitest.fdrcorrection.html\#statsmodels.stats.multitest.fdrcorrection.
67. Oliphant T. numpy. Version 1.20.1. [software]. Released March 14th 2006; Updated February 7th 2021. Available from https://pypi.org/project/numpy/.
68. Harris C, Millman K, Walt S van der, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith N, Kern R, Picus M, Hoyer S, Kerkwijk M van, Brett M, Haldane A, Río JF del, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, and Oliphant T. Array programming with NumPy. Nature 2020; 585:357–62. DOI: 10.1038/s41586-020-2649-2.
69. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, and Pfister H. UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics 2014; 20:1983–92. DOI: 10.1109/TVCG.2014.2346248.
70. Nothman J. UpSetPlot. Version 0.8.0. [software]. Released February 21st 2018; Updated January 5th 2023. Available from https://pypi.org/project/UpSetPlot/.
71. Britannica, The Editors of Encyclopedia. adenosine triphosphate. *Encyclopedia Britannica*. [Internet]. Updated March 20th 2023; Accessed April 3rd 2023. Available from https://www.britannica.com/science/adenosine-triphosphate.
72. Biologydictionary.net Editors. Adenosine Triphosphate (ATP). Biology Dictionary. [Internet]. Updated October 4th 2019; Accessed November 28th 2022. Available from https://biologydictionary.net/atp/.
73. Nelson D and Cox M. Lehninger Principles of Biochemistry. 7th ed. New York: W. H. Freeman and Company, 2017
74. National Center for Biotechnology Information. PubChem Compound Summary for CID 124886, Glutathione. Pubchem. [Internet]. Accessed April 11th 2023. Available from https://pubchem.ncbi.nlm.nih.gov/compound/Glutathione.
75. Reniere M, Whiteley A, Hamilton K, John S, Lauer P, Brennan R, and Portnoy D. Glutathione activates virulence gene expression of an intracellular pathogen. Nature 2015; 517:170–3. DOI: 10.1038/nature14029.
76. Meister A. Biosynthesis and Functions of Glutathione, an Essential Biofactor. Journal of Nutritional Science and Vitaminology 1992; 38:1–6. DOI: 10.3177/jnsv.38.Special_1.
77. Oshimura E and Sakamoto K. Chapter 19 - Amino Acids, Peptides, and Proteins. *Cosmetic Science and Technology*. Ed. by Sakamoto K, Lochhead R, Maibach H, and Yamashita Y. Amsterdam: Elsevier, 2017 :285–303. DOI: 10.1016/B978-0-12-802005-0.00019-7.
78. Hell R and Wirtz M. Molecular Biology, Biochemistry and Cellular Physiology of Cysteine Metabolism in Arabidopsis thaliana. Arabidopsis Book 2011; 9:e0154. DOI: 10.1199/tab.0154.
79. Stringer J and Xu K. Possible mechanisms for the anticonvulsant activity of fructose-1,6-diphosphate. Epilepsia 2008; 49:101–3. DOI: 10.1111/j.1528-1167.2008.01849.x.
80. Kirtley M and McKAY M. Fructose-1, 6-bisphosphate, a regulator of metabolism. Molecular and Cellular Biochemistry 1977; 18:141–9. DOI: 10.1007/BF00280279.
81. Xu Y, Wu Y, Xiong Y, Tao J, Pan T, Tan S, Gao G, Chen Y, Abbas N, Getachew A, Zhuang Y, You K, Yang F, and Li Y. Ascorbate protects liver from metabolic disorder through inhibition of lipogenesis and suppressor of cytokine signaling 3 (SOCS3). Nutrition & Metabolism 2020; 17. DOI: 10.1186/s12986-020-0431-y.

82. Yin X and Xu Y. Structure and Function of TET Enzymes. *DNA Methyltransferases - Role and Function*. Ed. by Jeltsch A and Jurkowska RZ. Vol. 945. Cham: Springer International Publishing, 2016 :275–302. DOI: 10.1007/978-3-319-43624-1_12.
83. Kirson D, Todorovic J, and Mihic S. Positive allosteric modulators differentially affect full versus partial agonist activation of the glycine receptor. J Pharmacol Exp Ther. 2012; 342:61–70. DOI: 10.1124/jpet.112.191486.
84. National Center for Biotechnology Information. PubChem Compound Summary for CID 135398633, Guanosine-5'-triphosphate. PubChem. [Internet]. Accessed November 29th 2022. Available from https://pubchem.ncbi.nlm.nih.gov/compound/guanosine-triphosphate.
85. Caspi R, Billington R, Keseler I, Kothari A, Krummenacker M, Midford P, Ong W, Paley S, Subhraveti P, and Karp P. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res48 2020; 48. DOI: 10.1093/nar/gkz862.
86. Bearne S, Guo C, and Liu J. GTP-Dependent Regulation of CTP Synthase: Evolving Insights into Allosteric Activation and NH3 Translocation. Biomolecules 2022; 12:647. DOI: 10.3390/biom12050647.
87. Rochovansky O. On the role of substrate and GTP in the regulation of argininosuccinase activity. Journal of Biological Chemistry 1975; 250:7225–30. DOI: 10.1016/S0021-9258(19)40932-0.
88. National Center for Biotechnology Information. PubChem Compound Summary for CID 1051, Pyridoxal phosphate. PubChem. [Internet]. Accessed April 28th 2023. Available from https://pubchem.ncbi.nlm.nih.gov/compound/Pyridoxal-phosphate.
89. Mukherjee T, Hanes J, Tews I, Ealick S, and Begley T. Pyridoxal phosphate: Biosynthesis and catabolism. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 2011; 1814. Pyridoxal Phosphate Enzymology:1585–96. DOI: 10.1016/j.bbapap.2011.06.018.
90. National Center for Biotechnology Information. PubChem Compound Summary for CID 1176, Urea. PubChem. [Internet]. Accessed December 2nd 2022. Available from https://pubchem.ncbi.nlm.nih.gov/compound/Urea.
91. Biologydictionary.net Editors. Allosteric inhibition. Biology Dictionary. [Internet]. Updated November 29th 2020; Accessed November 28th 2022. Available from https://biologydictionary.net/allosteric-inhibition/.
92. Weber K and Rétey J. On the nature of the irreversible inhibition of histidine ammonia lyase by cysteine and dioxygen. Bioorganic & Medicinal Chemistry 1996; 4:1001–6. DOI: 10.1016/0968-0896(96)00091-0.
93. Baedeker M and Schulz G. Structures of two histidine ammonia-lyase modifications and implications for the catalytic mechanism. European Journal of Biochemistry 2002; 269:1790–7. DOI: 10.1046/j.1432-1327.2002.02827.x.
94. Bennion B and Daggett V. The molecular basis for the chemical denaturation of proteins by urea. Proceedings of the National Academy of Sciences of the United States of America 2003; 100:5142–7. DOI: 10.1073/pnas.0930122100.
95. National Center for Biotechnology Information. PubChem Compound Summary for CID 439153, 1,4-Dihydronicotinamide adenine dinucleotide. PubChem. [Internet]. Accessed April 28th 2023. Available from https://pubchem.ncbi.nlm.nih.gov/compound/1_4-Dihydronicotinamide-adenine-dinucleotide.
96. Sun P, Liu Y, Ma T, and Ding J. Structure and allosteric regulation of human NAD-dependent isocitrate dehydrogenase. Cell Discovery 2020; 6. DOI: 10.1038/s41421-020-00220-7.
97. Richman P and Meister A. Regulation of gamma-glutamyl-cysteine synthetase by nonallosteric feedback inhibition by glutathione. Journal of Biological Chemistry 1975; 250:1422–6. DOI: 10.1016/S0021-9258(19)41830-9.

98. Zhang Z, Zhang X, Fang X, Niimi M, Huang Y, Piao H, Gao S, Fan J, and Yao J. Glutathione inhibits antibody and complement-mediated immunologic cell injury via multiple mechanisms. Redox Biology 2017; 12:571–81. DOI: 10.1016/j.redox.2017.03.030.
99. Iacobazzi V and Infantino V. Citrate - new functions for an old metabolite. Biological chemistry 2014; 395. DOI: 10.1515/hsz-2013-0271.
100. Agius L. Role of glycogen phosphorylase in liver glycogen metabolism. Molecular Aspects of Medicine 2015; 46. From Glucose to glycogen and Back: Festschrift in honor of Bill Whelan 90the birthday:34–45. DOI: 10.1016/j.mam.2015.09.002.
101. National Center for Biotechnology Information. PubChem Compound Summary for CID 5885, Triphosphopyridine nucleotide. PubChem. [Internet]. Accessed April 28th 2023. Available from https://pubchem.ncbi.nlm.nih.gov/compound/Triphosphopyridine-nucleotide.
102. Bian C, Zhang C, Luo T, Vyas A, Chen S, Liu C, Kassab M, Yang Y, Kong M, and Yu X. NADP+ is an endogenous PARP inhibitor in DNA damage response and tumor suppression. Nature communications 2019; 10:693. DOI: 10.1038/s41467-019-08530-5.
103. Lüscher B, Ahel I, Altmeyer M, Ashworth A, Bai P, Chang P, Cohen M, Corda D, Dantzer F, Daugherty M, Dawson T, Dawson V, Deindl S, Fehr A, Feijs KH, Filippov D, Gagné J, Grimaldi G, Guettler S, Hoch N, Hottiger M, Korn P, Kraus W, Ladurner A, Lehtiö L, Leung A, Lord C, Mangerich A, Matic I, Matthews J, Moldovan G, Moss J, Natoli G, Nielsen M, Niepel M, Nolte F, Pascal J, Paschal B, Pawłowski K, Poirier G, Smith S, Timinszky G, Wang Z, Yélamos J, Yu X, Zaja R, and Ziegler M. ADP-ribosyltransferases, an update on function and nomenclature. The FEBS Journal 2022; 289:7399–410. DOI: 10.1111/febs.16142.
104. Kleczkowski L, Villand P, Lönneborg A, Olsen O, and Lüthi E. Plant ADP-Glucose Pyrophosphorylase -Recent Advances and Biotechnological Perspectives (A Review). Zeitschrift für Naturforschung C 1991; 46:605–12. DOI: 10.1515/znc-1991-7-817.
105. Israelsen W and Vander Heiden M. Pyruvate kinase: Function, regulation and role in cancer. Semin Cell Dev Biol. 2015; 43:43–51. DOI: 10.1016/j.semcdb.2015.08.004.
106. Johnsen U, Hansen T, and Schönheit P. Comparative analysis of pyruvate kinases from the hyperthermophilic archaea Archaeoglobus fulgidus, Aeropyrum pernix, and Pyrobaculum aerophilum and the hyperthermophilic bacterium Thermotoga maritima: unusual regulatory properties in hyperthermophilic archaea. Journal of Biological Chemistry 2003; 278:25417–27. DOI: 10.1074/jbc.M210288200.
107. Tanaka K, Sakai H, Ohta T, and Matsuzawa H. Molecular Cloning of the Genes for Pyruvate Kinase of Two Bacilli, Bacillus psychrophilus and Bacillus licheniformis, and Comparison of the Properties of the Enzymes Produced in Escherichia coli. Bioscience, Biotechnology, and Biochemistry 1995; 59:1536–42. DOI: 10.1271/bbb.59.1536.
108. Ketudat Cairns J and Esen A. β-Glucosidases. Cellular and molecular life sciences : CMLS 2010; 67:3389–405. DOI: 10.1007/s00018-010-0399-2.
109. Lepiniec L, Vidal J, Chollet R, Gadal P, and Crétin C. Phosphoenolpyruvate carboxylase: structure, regulation and evolution. Plant Science 1994; 99:111–24. DOI: 10.1016/0168-9452(94)90168-6.
110. O'Leary M. Phosphoenolpyruvate carboxylase: an enzymologist's view. Annual Review of Plant Physiology 1982; 33:297–315. DOI: 10.1146/annurev.pp.33.060182.001501.
111. Kai Y, Matsumura H, and Izui K. Phosphoenolpyruvate carboxylase: three-dimensional structure and molecular mechanisms. Archives of Biochemistry and Biophysics 2003; 414:170–9. DOI: 10.1016/S0003-9861(03)00170-X.

112. Wiegand G and Remington S. CITRATE SYNTHASE: Structure, Control, and Mechanism. Annual Review of Biophysics and Biophysical Chemistry 1986; 15:97–117. DOI: 10.1146/annurev.bb.15.060186.000525.
113. Bhagavan N and Ha C. Chapter 12 - Carbohydrate Metabolism I: Glycolysis and the Tricarboxylic Acid Cycle. *Essentials of Medical Biochemistry (Second Edition)*. Ed. by Bhagavan N and Ha C. Second Edition. San Diego: Academic Press, 2015 :165–85. DOI: 10.1016/B978-0-12-416687-5.00012-9.
114. Huang S and Millar A. Succinate dehydrogenase: the complex roles of a simple enzyme. Current Opinion in Plant Biology 2013; 16. Physiology and metabolism:344–9. DOI: 10.1016/j.pbi.2013.02.007.
115. Dipple K, Zhang Y, Huang B, McCabe L, Dallongeville J, Inokuchi T, Kimura M, Marx H, Roederer G, Shih V, et al. Glycerol kinase deficiency: evidence for complexity in a single gene disorder. Human genetics 2001; 109:55–62. DOI: 10.1007/s004390100545.
116. Timson D. Fructose 1, 6-bisphosphatase: getting the message across. Bioscience reports 2019; 39. DOI: 10.1042/BSR20190124.
117. Cárdenas M, Cornish-Bowden A, and Ureta T. Evolution and regulatory role of the hexokinases. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1998; 1401:242–64. DOI: 10.1016/S0167-4889(97)00150-X.
118. Claeyssen É and Rivoal J. Isozymes of plant hexokinase: Occurrence, properties and functions. Phytochemistry 2007; 68:709–31. DOI: 10.1016/j.phytochem.2006.12.001.
119. Gao H and Leary J. Multiplex inhibitor screening and kinetic constant determinations for yeast hexokinase using mass spectrometry based assays. Journal of the American Society for Mass Spectrometry 2003; 14:173–81. DOI: 10.1016/S1044-0305(02)00867-X.
120. Shimizu T, Nakayama I, Nagayama K, Miyazawa T, and Nezu Y. Acetolactate Synthase Inhibitors. *Herbicide Classes in Development: Mode of Action, Targets, Genetic Engineering, Chemistry*. Ed. by Böger P, Wakabayashi K, and Hirai K. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002 :1–41. DOI: 10.1007/978-3-642-59416-8_1.
121. Ohshida T, Koba K, Hayashi J, Yoneda K, Ohmori T, Ohshima T, and Sakuraba H. A novel bifunctional aspartate kinase-homoserine dehydrogenase from the hyperthermophilic bacterium, Thermotoga maritima. Bioscience, Biotechnology, and Biochemistry 2018; 82:2084–93. DOI: 10.1080/09168451.2018.1511365.
122. Wilson W, Roach P, Montero M, Baroja-Fernández E, Muñoz F, Eydallin G, Viale A, and Pozueta-Romero J. Regulation of glycogen metabolism in yeast and bacteria. FEMS Microbiology Reviews 2010; 34:952–85. DOI: 10.1111/j.1574-6976.2010.00220.x.
123. Eliot A and Kirsch J. Pyridoxal Phosphate Enzymes: Mechanistic, Structural, and Evolutionary Considerations. Annual Review of Biochemistry 2004; 73:383–415. DOI: 10.1146/annurev.biochem.73.011303.074021.
124. Mozzarelli A and Bettati S. Exploring the pyridoxal 5′-phosphate-dependent enzymes. The Chemical Record 2006; 6:275–87. DOI: 10.1002/tcr.20094.
125. Fulcher L and Sapkota G. Functions and regulation of the serine/threonine protein kinase CK1 family: Moving beyond promiscuity. Biochemical Journal 2020; 477:4603–21. DOI: 10.1042/BCJ20200506.
126. Abel T and Nguyen P. Chapter 6 Regulation of hippocampus-dependent memory by cyclic AMP-dependent protein kinase. *Essence of Memory*. Ed. by Sossin W, Lacaille J, Castellucci V, and Belleville S. Vol. 169. Elsevier, 2008 :97–115. DOI: 10.1016/S0079-6123(07)00006-4.
127. Andersson I, Knight S, Schneider G, Lindqvist Y, Lundqvist T, Bränden C, and Lorimer G. Crystal structure of the active site of ribulose-bisphosphate carboxylase. Nature 1989; 337:229–34. DOI: 10.1038/337229a0.

128. Jordan D and Chollet R. Inhibition of ribulose bisphosphate carboxylase by substrate ribulose 1,5-bisphosphate. Journal of Biological Chemistry 1983; 258:13752–8. DOI: 10.1016/S0021-9258(17)43982-2.
129. Nimmo H. The regulation of phosphoenolpyruvate carboxylase in CAM plants. Trends in Plant Science 2000; 5:75–80. DOI: 10.1016/S1360-1385(99)01543-5.
130. Nordlund P and Reichard P. Ribonucleotide reductases. Annu. Rev. Biochem. 2006; 75:681–706. DOI: 10.1146/annurev.biochem.75.103004.142443.
131. Reichard P. Ribonucleotide reductases: Substrate specificity by allostery. Biochemical and Biophysical Research Communications 2010; 396. Recent Progress in Molecular Sciences: Reviews from Karolinska Institutet at its 200-year Anniversary:19–23. DOI: 10.1016/j.bbrc.2010.02.108.
132. Krebs E. Protein Kinases* *Support of the National Institutes of Arthritis and Metabolic Diseases, NIH, U. S. Public Health Service (AM 12842), the Muscular Dystrophy Association of America, and the American Heart Association is acknowledged. *Current Topics in Cellular Regulation*. Ed. by Horecker B and Stadtman E. Vol. 5. Academic Press, 1972 :99–133. DOI: 10.1016/B978-0-12-152805-8.50010-1.
133. Toews M, Kanji M, and Carper W. 6-Phosphogluconate dehydrogenase. Purification and kinetics. Journal of Biological Chemistry 1976; 251:7127–31. DOI: 10.1016/S0021-9258(17)32951-4.
134. Moritz B, Striegel K, Graaf A de, and Sahm H. Kinetic properties of the glucose-6-phosphate and 6 phosphogluconate dehydrogenases from Corynebacterium glutamicum and their application for predicting pentose phosphate pathway flux in vivo. European Journal of Biochemistry 2000; 267:3442–52. DOI: 10.1046/j.1432-1327.2000.01354.x.
135. Robbins A and Stout C. The structure of aconitase. Proteins: Structure, Function, and Bioinformatics 1989; 5:289–312. DOI: 10.1002/prot.340050406.
136. Figueroa C, Asencion Diez M, Ballicora M, and Iglesias A. Structure, function, and evolution of plant ADP-glucose pyrophosphorylase. Plant Molecular Biology 2022; 108:307–23. DOI: 10.1007/s11103-021-01235-8.
137. Ballicora M, Iglesias A, and Preiss J. ADP-Glucose Pyrophosphorylase, a Regulatory Enzyme for Bacterial Glycogen Synthesis. Microbiology and Molecular Biology Reviews 2003; 67:213–25. DOI: 10.1128/MMBR.67.2.213-225.2003.
138. Ballicora M, Iglesias A, and Preiss J. ADP-glucose pyrophosphorylase: a regulatory enzyme for plant starch synthesis. Photosynthesis research 2004; 79:1–24. DOI: 10.1023/B:PRES.0000011916.67519.58.
139. Dumas R, Cobessi D, Robin A, Ferrer J, and Curien G. The many faces of aspartate kinases. Archives of Biochemistry and Biophysics 2012; 519:186–93. DOI: 10.1016/j.abb.2011.10.016.
140. Kato C, Kurihara T, Kobashi N, Yamane H, and Nishiyama M. Conversion of feedback regulation in aspartate kinase by domain exchange. Biochemical and Biophysical Research Communications 2004; 316:802–8. DOI: 10.1016/j.bbrc.2004.02.122.
141. Chou C, Sun Y, Meng M, and Hsiao C. The crystal structure of phosphoglucose isomerase/autocrine motility factor/neuroleukin complexed with its carbohydrate phosphate inhibitors suggests its substrate/receptor recognition. The Journal of biological chemistry 2000; 275:23154–60. DOI: 10.1074/jbc.m002017200.
142. Jeffery C, Bahnson B, Chien W, Ringe D, and Petsko G. Crystal structure of rabbit phosphoglucose isomerase, a glycolytic enzyme that moonlights as neuroleukin, autocrine motility factor, and differentiation mediator. Biochemistry 2000; 39:955–64. DOI: 10.1021/bi991604m.

143. Zhang Y, Zagnitko O, Rodionova I, Osterman A, and Godzik A. The FGGY carbohydrate kinase family: insights into the evolution of functional specificities. PLoS computational biology 2011; 7:e1002318. DOI: 10.1371/journal.pcbi.1002318.
144. Ormö M, Bystrom C, and Remington S. Crystal Structure of a Complex of Escherichia c oli Glycerol Kinase and an Allosteric Effector Fructose 1, 6-Bisphosphate. Biochemistry 1998; 37:16565–72. DOI: 10.1021/bi981616s.
145. Walsh D, Cooper R, Denton R, Bridges B, and Randle P. The elementary reactions of the pig heart pyruvate dehydrogenase complex. A study of the inhibition by phosphorylation. eng. Biochemical journal 1976; 157:41–67. DOI: 10.1042/bj1570041.
146. Patel M and Roche T. Molecular biology and biochemistry of pyruvate dehydrogenase complexes1. The FASEB Journal 1990; 4:3224–33. DOI: 10.1096/fasebj.4.14.2227213.
147. Saito T, Maeda T, Nakazawa M, Takeuchi T, Nozaki T, and Asai T. Characterisation of hexokinase in Toxoplasma gondii tachyzoites. International Journal for Parasitology 2002; 32:961–7. DOI: 10.1016/S0020-7519(02)00059-0.
148. Harris M, Walker D, Drew M, Mitchell W, KD, Schroeder C, Flaherty D, Weiner W, Golden J, and Morris J. Interrogating a Hexokinase-Selected Small-Molecule Library for Inhibitors of Plasmodium falciparum Hexokinase. Antimicrobial Agents and Chemotherapy 2013; 57:3731–7. DOI: 10.1128/AAC.00662-13.
149. Snášel J and Pichová I. Allosteric regulation of pyruvate kinase from Mycobacterium tuberculosis by metabolites. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 2019; 1867:125–39. DOI: 10.1016/j.bbapap.2018.11.002.
150. Kapoor M, O'Brien M, and Braun A. Modification of the regulatory properties of pyruvate kinase of Neurospora by growth at elevated temperatures. Canadian Journal of Biochemistry 1976; 54:398–407. DOI: 10.1139/o76-058.
151. Kapoor M. Subunit Structure and some Properties of Pyruvate Kinase of Neurospora. Canadian Journal of Biochemistry 1975; 53:109–19. DOI: 10.1139/o75-017.
152. Kapoor M and Bishop M. An immunological study of the interaction of ligands with pyruvate kinase of Neurospora crassa. Canadian Journal of Biochemistry 1982; 60:771–6. DOI: 10.1139/o82-095.
153. O'brien M and Kapoor M. Studies of the structure-function relationships of Neurospora pyruvate kinase: denaturation by urea and the effect of ligands on the refolding and renaturation process. International Journal of Biochemistry 1980; 11:107–16. DOI: 10.1016/0020-711x(80)90242-6.
154. Zoraghi R, See R, Gong H, Lian T, Swayze R, Finlay B, Brunham R, McMaster W, and Reiner N. Functional analysis, overexpression, and kinetic characterization of pyruvate kinase from methicillin-resistant Staphylococcus aureus. Biochemistry 2010; 49:7733–47. DOI: 10.1021/bi100780t.
155. Knowles V and Plaxton W. From Genome to Enzyme: Analysis of Key Glycolytic and Oxidative Pentose-Phosphate Pathway Enzymes in the Cyanobacterium Synechocystis sp. PCC 6803. Plant and Cell Physiology 2003; 44:758–63. DOI: 10.1093/pcp/pcg086.
156. Calatrava V, Tejada-Jimenez M, Sanz-Luque E, Fernandez E, Galvan A, and Llamas A. *Chlamydomonas reinhardtii*; a Reference Organism to Study Algal-Microbial Interactions: Why Can't They Be Friends? Plants (Basel, Switzerland) 2023; 12:788. DOI: 10.3390/plants12040788.
157. Ferretti M, Hussien R, Ballicora M, Iglesias A, Figueroa C, and Asencion Diez M. The ADP-glucose pyrophosphorylase from Melainabacteria: a comparative study between photosynthetic and non-photosynthetic bacterial sources. Biochimie 2022; 192:30–7. DOI: 10.1016/j.biochi.2021.09.011.

158. Cáceres A, Portillo R, Acosta H, Rosales D, Quiñones W, Avilan L, Salazar L, Dubourdieu M, Michels P, and Concepción J. Molecular and biochemical characterization of hexokinase from Trypanosoma cruzi. Molecular and Biochemical Parasitology 2003; 126:251–62. DOI: 10.1016/S0166-6851(02)00294-3.
159. Ureta T, Medina C, and Preller A. The evolution of hexokinases. Archivos de biologia y medicina experimentales 1999; 20:343–57. DOI: 10.1042/BST027A056A.

Appendices

Appendix A: Supplementary information

This appendix contains information about the supplementary material that is associated with this work. The material is available at the GitHub repository created for this thesis (`https://github.com/elinsroed/predicting-allostery`), and consists of Python scripts (v. 3.8.6) written in Jupyter Notebook (v. 6.4.12) [33], raw datafiles downloaded from relevant databases, and the results from the analyses conducted in this project. Table A1 provides a description of the different files found in the repository, and their paths for easier access.

Table A1: Supplementary information overview; description of the content available at the thesis' GitHub repository.

No.	Description	Path
1	Raw and cleaned allosteric data from BRENDA	datafiles\interactions
2	Supporting files for mapping and filtering interaction and feature data	datafiles\support
3	Files of protein sequence features	datafiles\features
4	Files used for creating and annotating the phylogenetic tree	datafiles\phylotree
5	Files of results from enrichment analysis and validation	datafiles\results
6	Script for cleaning interaction data from BRENDA	data download and cleaning\BRENDA_data.ipynb
7	Script for downloading protein sequence features	data download and cleaning\download_features.ipynb
8	Scripts for creating and annotating phylogenetic tree	analysis\data_analysis\phylogenetic_tree
9	Script for analysing interaction data	analysis\data_analysis\BRENDA_analysis.ipynb
10	HTML file of interactive network	analysis\data_analysis\network.html
11	Scripts for conducting enrichment analysis and assessing results	analysis\enrichment_analysis

Appendix B: Network of protein-metabolite interactions

This appendix contains the network of protein-metabolite interactions for the top ten regulated enzymes and metabolites interacting with either two or more of the proteins in this subgroup, created using Python's pyvis package (v. 0.3.1) [44]. The network is displayed in Figure B1, with enzymes shown as orange squares, metabolites as blue dots, and interactions shown as green (activating) and red (inhibiting) edges. Its interpretation is given in Section 3.1.2.

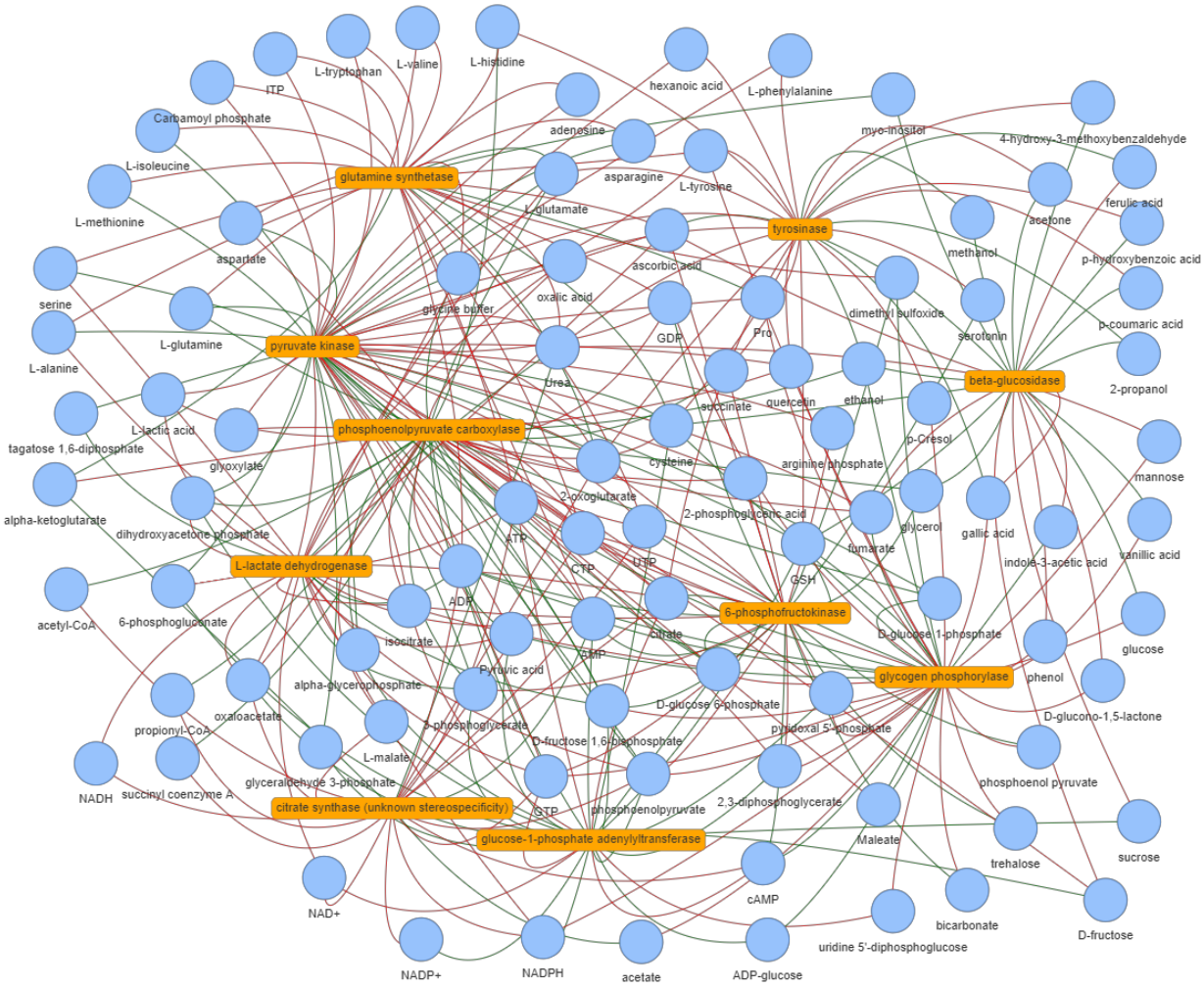


Figure B1: Network of protein-metabolite interactions, consisting of the top ten regulated enzymes (orange squares) and metabolites (blue dots) interacting with two or more enzymes in this subgroup of proteins. Activating interactions are shown by green edges and inhibiting interactions are shown by red edges. This network was created and visualized using Python's pyvis package (v. 0.3.1). [44]

Appendix C: Interaction-annotated phylogenetic trees

This appendix contains the phylogenetic trees mapped with allosteric interactions in rectangular format with labels for the interactions. Figure C1 displays the phylogenetic tree annotated with only documented interactions, while Figure C2 shows the phylogenetic tree annotated with documented (filled squares) and predicted (empty squares) interactions. The interpretation of the trees is given in Subsection 3.1.3 and 3.2.2.2, respectively.

Figure C1: Phylogenetic tree mapped with allosteric interactions, rectangular. Taxa are indicated as Archaea in green, Bacteria in purple, and Eukaryota in pink. Each mapped interaction is presented by a single column, with the interaction label *(EC number : Metabolite name : Mode)* as column header. The tree was annotated using the iTOL (v. 5) online tool [48].

Figure C2: Phylogenetic tree with allosteric interactions, rectangular: documented (filled squares) and predicted (empty squares). Taxa are indicated as Archaea in green, Bacteria in purple, and Eukaryota in pink. Each mapped interaction is presented by a single column, with the interaction label *(EC number : Metabolite name : Mode)* as column header. The tree was annotated using the iTOL (v. 5) online tool [48].

Appendix D: Interaction-predicting features

This appendix contains the results of the comparison of interactions predicted by every feature type (active site, binding site, conserved site, domain, family, homologous superfamily, PTM, repeat), that was conducted with the purpose of assessing the importance of these features in predicting protein-metabolite interactions. The overlaps of predicted interactions were plotted as UpSet plots, separated by activating and inhibiting interactions, which are displayed in Figure D1. The interpretation of these plots is given in Subsection 3.2.2.

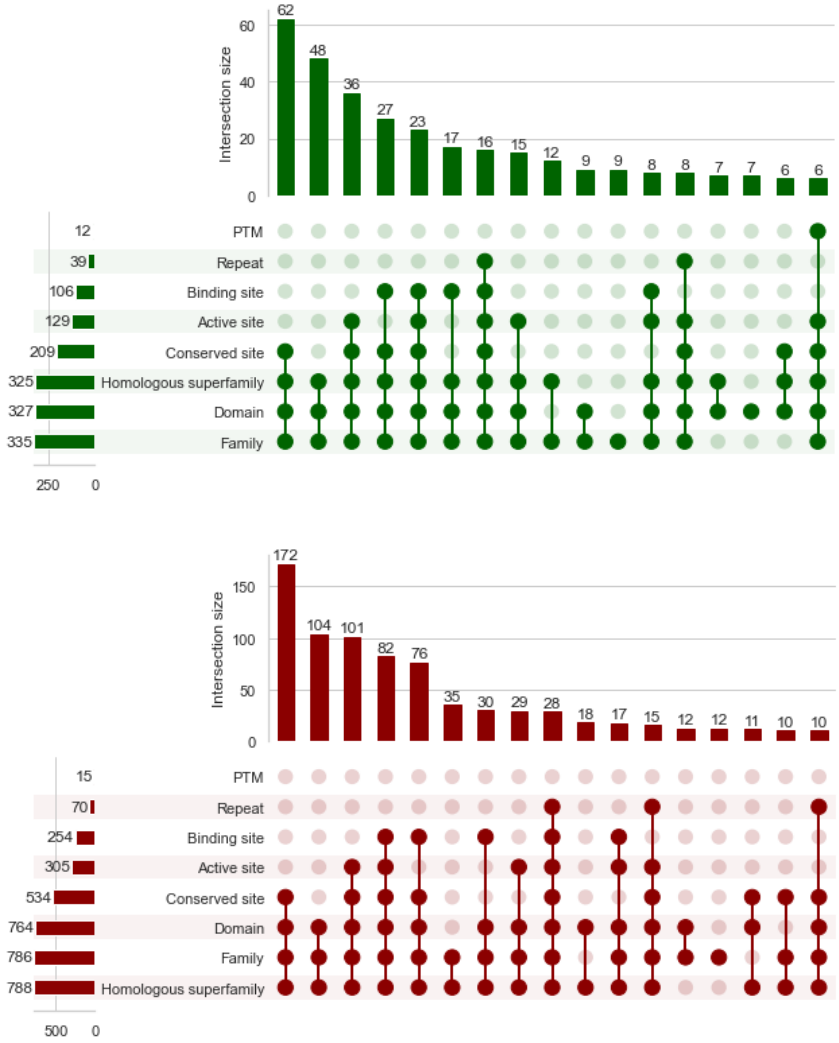


Figure D1: Overlaps of protein-metabolite interactions predicted by eight protein feature types: activating interactions (green) and inhibiting interactions (red). The vertical bars represent the number of predicted interactions in the intersection of the groups highlighted by a dot in the diagram below, while the horizontal bars represent the total amount of interactions predicted by the respective feature type. The bars are sorted by size, and only groups with 6 and 10 or more entries are included for activating and inhibiting interactions, respectively.

Appendix E: Statistically significantly associated features and interactions

This appendix contains histograms illustrating the frequency distributions of interaction-predicting protein features and protein feature-predicted interactions that were identified utilizing the results from enrichment analysis (Subsections 2.2.2 and 2.2.2.2). Figures E1 and E2 show the number of statistically significant associated interactions for each individual protein feature, and the number of statistically significant associated protein features for each individual interaction, respectively. The plots are separated by protein feature type and mode of interaction (activation/inhibition), and a count limit, which is specified for each individual plot, has been set with the purpose of bettering the visualization.

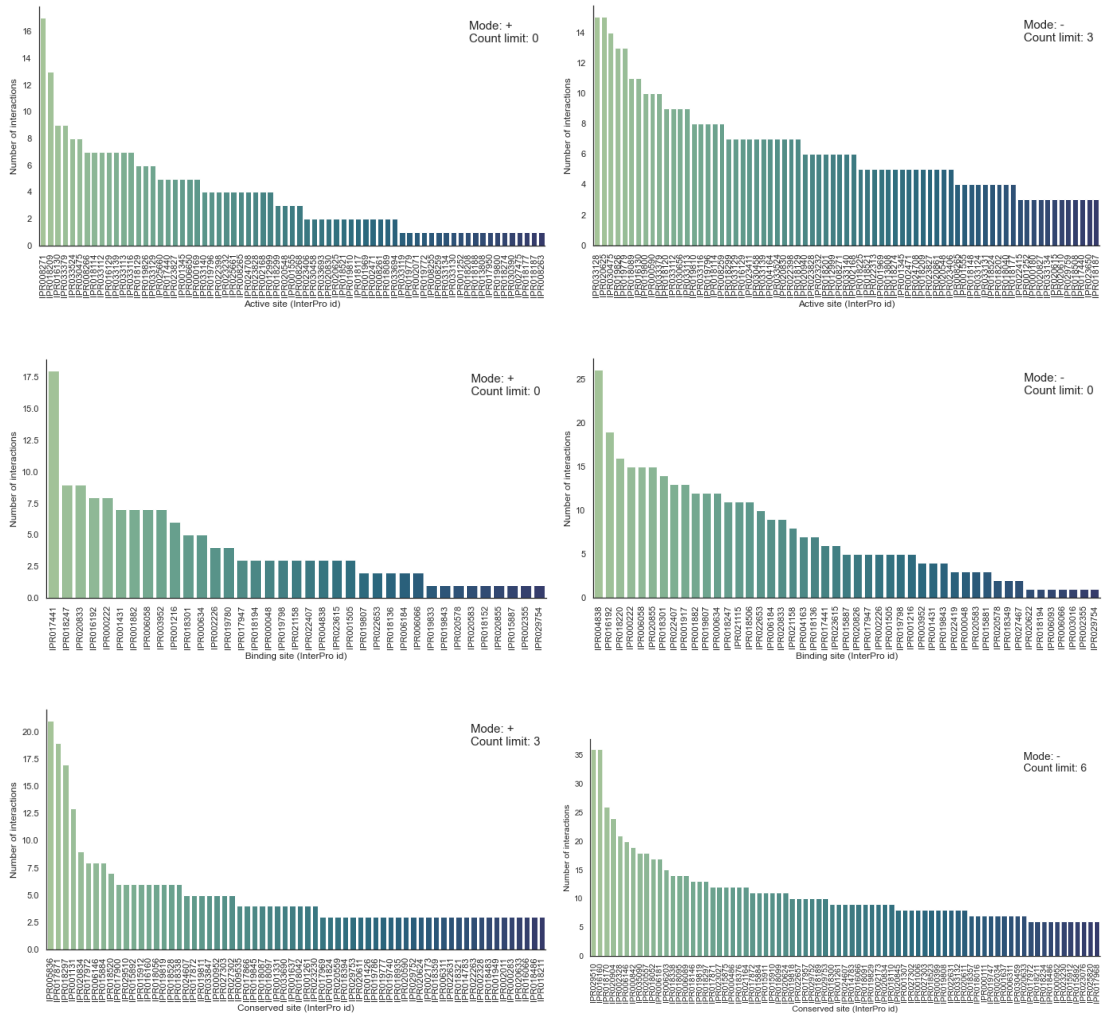


Figure E1: The number of associated interactions for each protein feature, separated by type of protein feature and activating (+) and inhibiting (-) interactions. The mode of interaction and count limit for the plots are indicated in the top right corner, while feature type is denoted by the x-axis label.

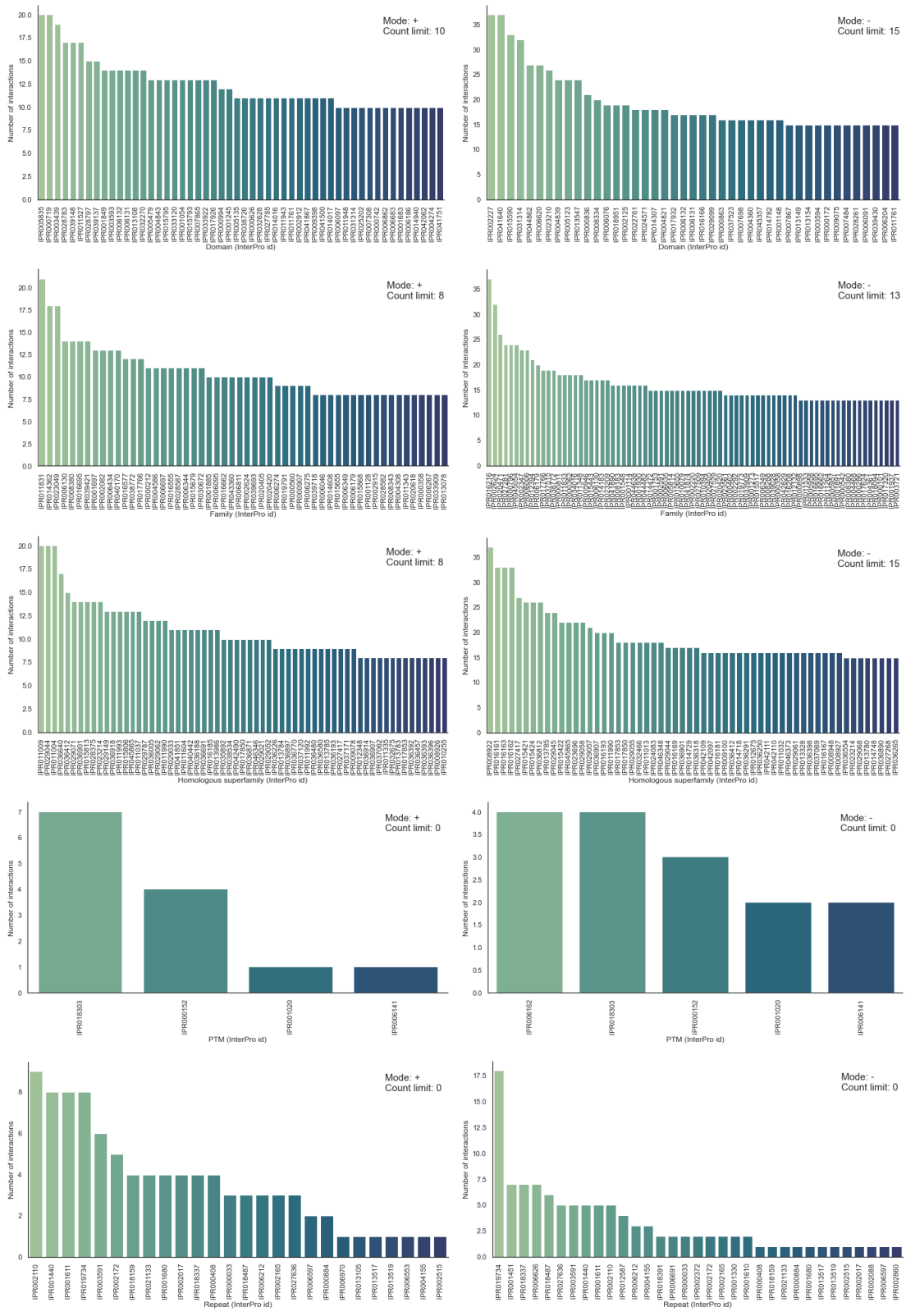


Figure E1: The number of associated interactions for each protein feature (continued).

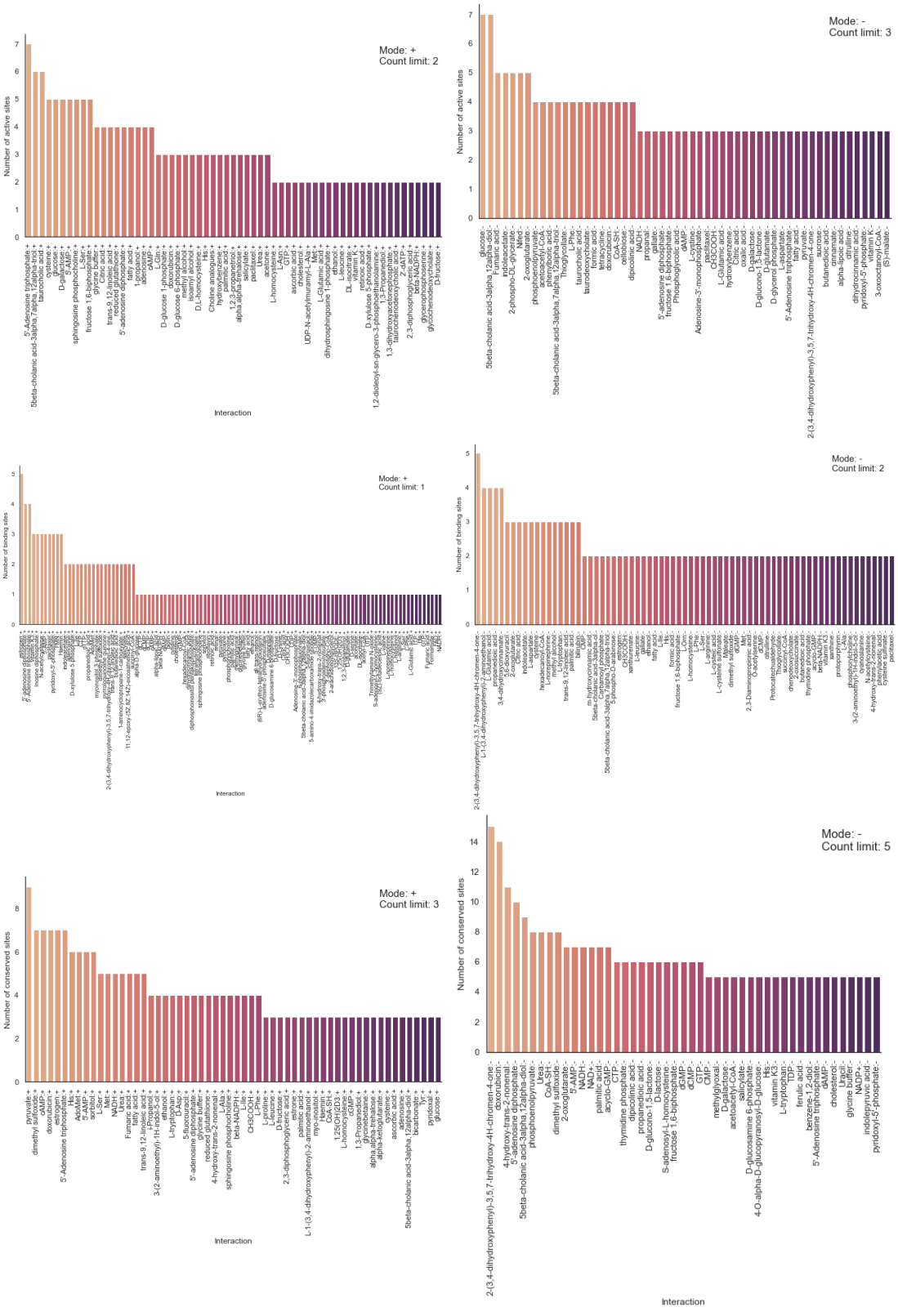


Figure E2: The number of protein features associated with each interaction, separated by type of protein feature and activating (+) and inhibiting (-) interactions. The mode of interaction and count limit for the plots are indicated in the top right corner, while feature type is denoted by the y-axis label.

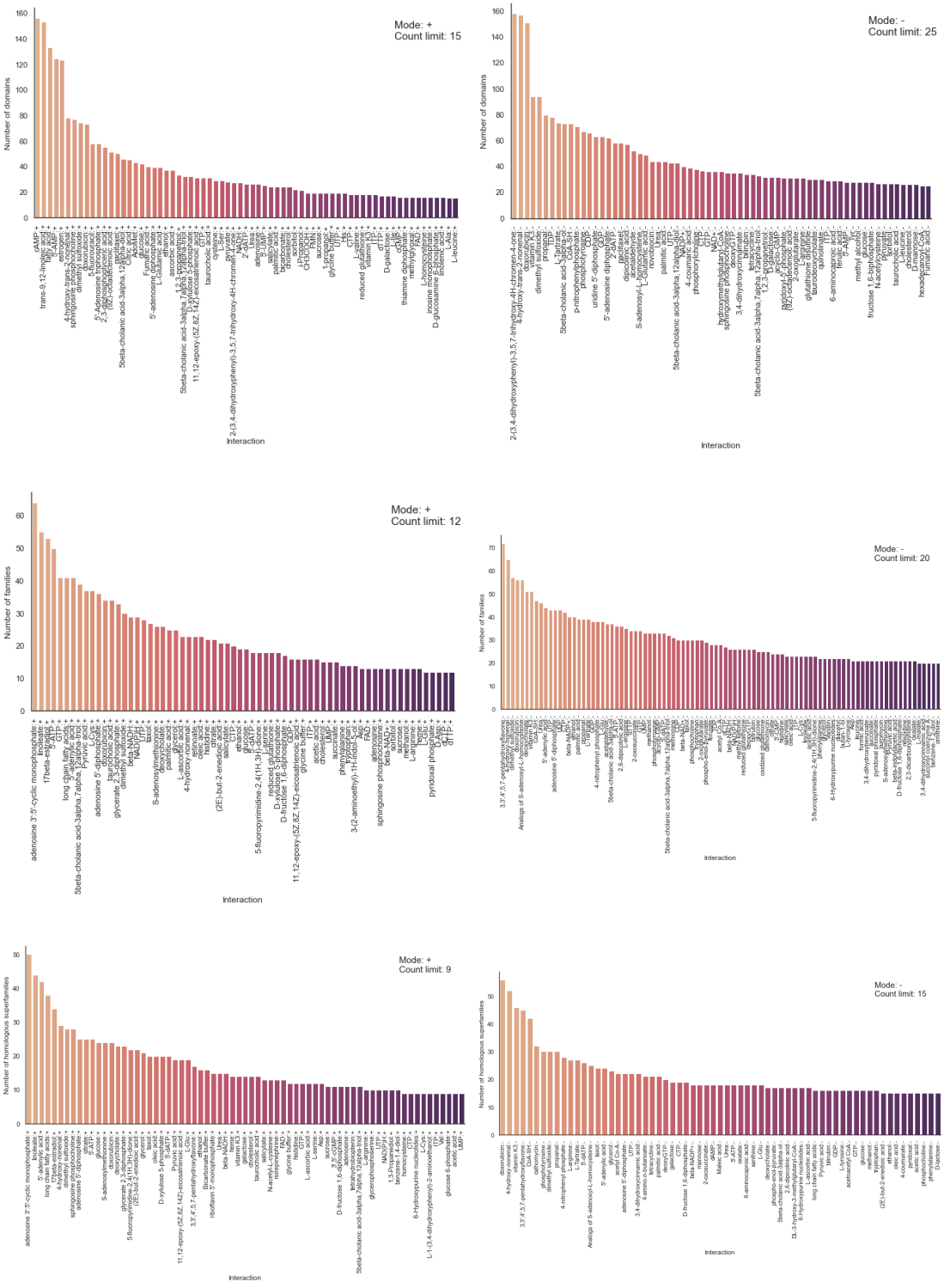


Figure E2: The number of protein features associated with each interaction (continued).

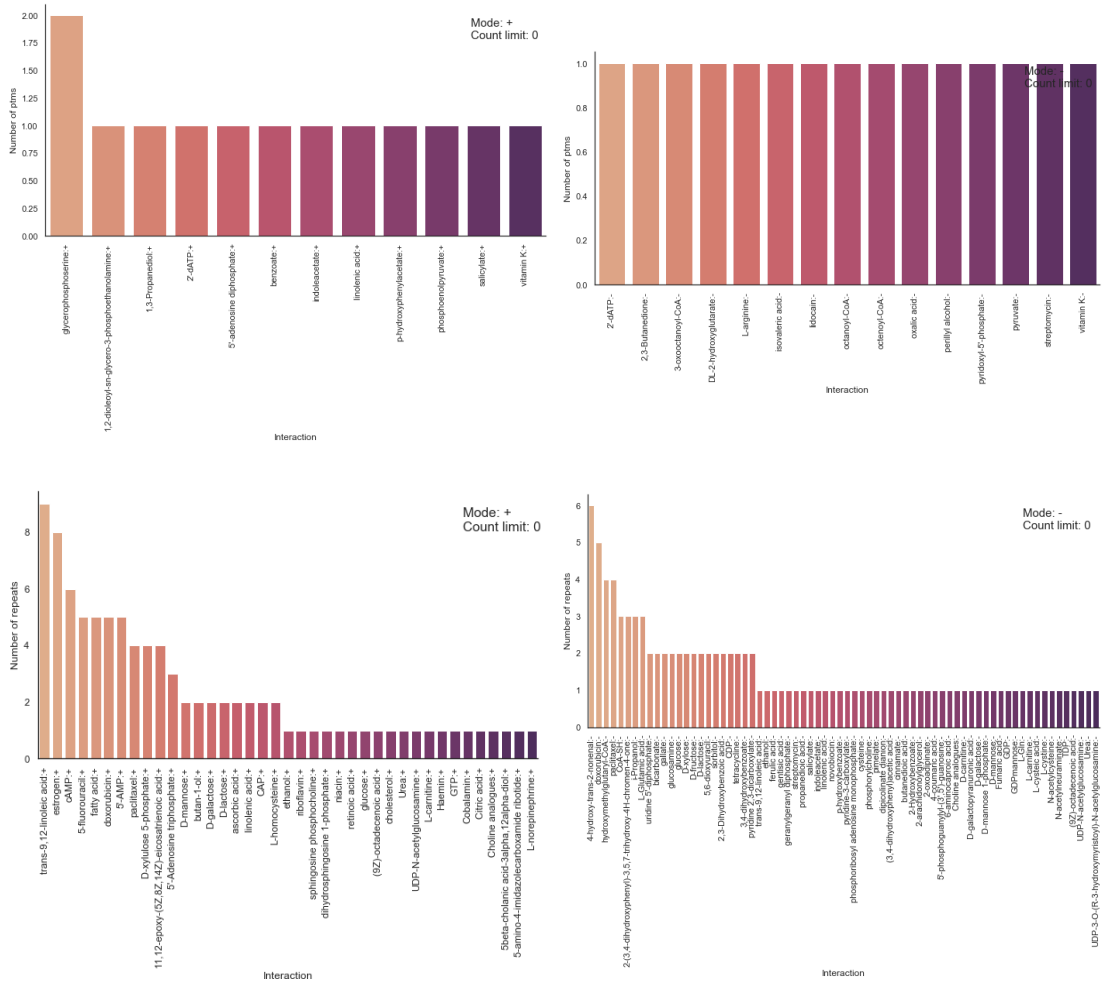
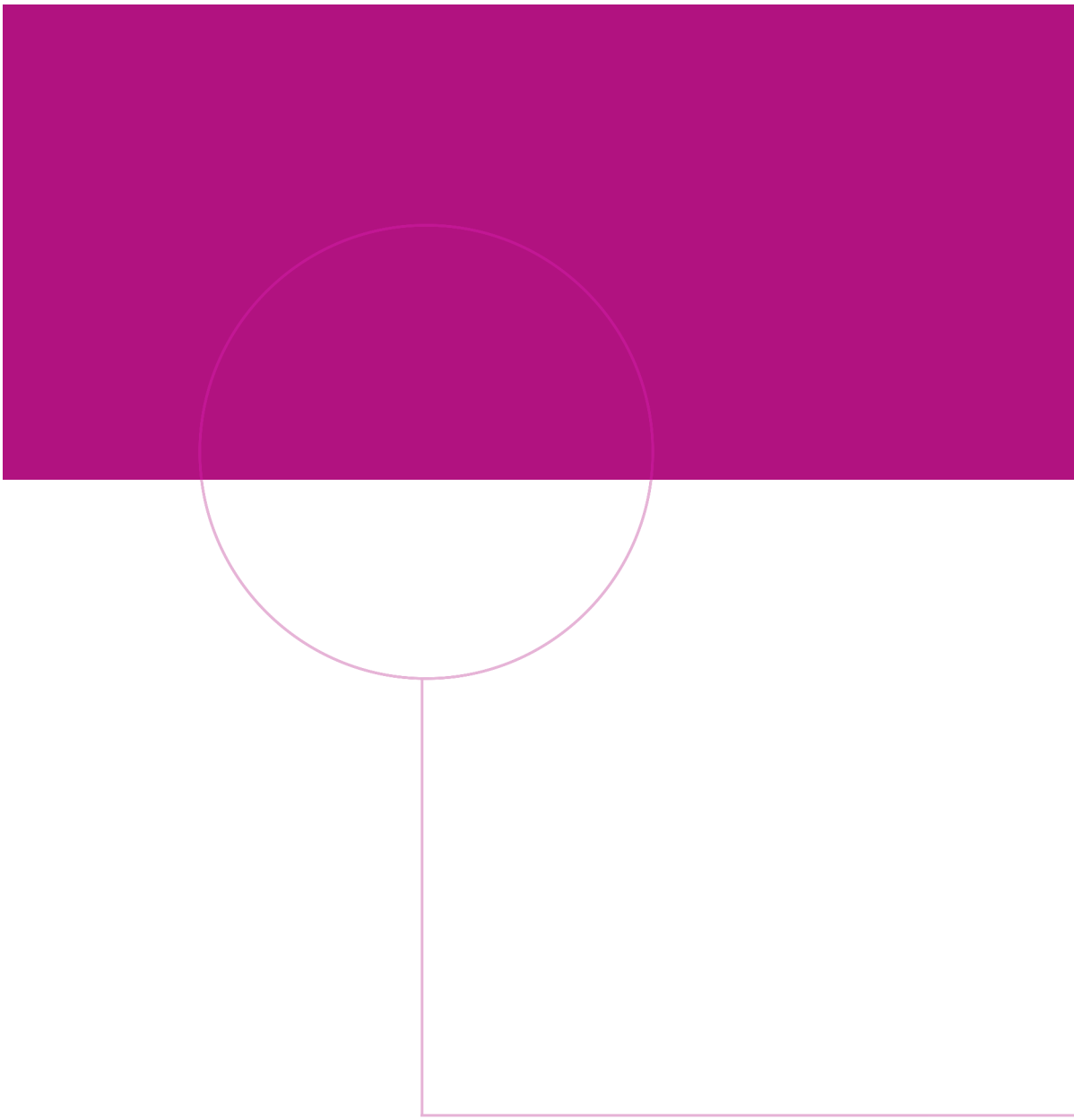


Figure E2: The number of protein features associated with each interaction (continued).



NTNU

Norwegian University of
Science and Technology