

# Consequences of refining biological networks through detailed pathway information

From genes to proteoforms

---

Luis Francisco Hernández Sánchez

Thesis for the degree of Philosophiae Doctor (PhD)  
University of Bergen, Norway  
2022

UNIVERSITY OF BERGEN



# Consequences of refining biological networks through detailed pathway information

From genes to proteoforms

Luis Francisco Hernández Sánchez



Thesis for the degree of Philosophiae Doctor (PhD)  
at the University of Bergen

Date of defense: 28.11.2022

© Copyright Luis Francisco Hernández Sánchez

The material in this publication is covered by the provisions of the Copyright Act.

Year: 2022

Title: Consequences of refining biological networks through detailed pathway information

Name: Luis Francisco Hernández Sánchez

Print: Skipnes Kommunikasjon / University of Bergen

## **Scientific environment**

This dissertation is part of a PhD carried out at the University of Bergen, Faculty of Medicine, Department of Clinical Science as part of the Center for Diabetes Research, from mid-2016 through to mid-2022. The period included a research stay at the European Bioinformatics Institute in Hinxton, United Kingdom, and multiple trips abroad to international conferences. In addition to the scientific research, the PhD fellowship included 25% of duty work such as providing bioinformatic support and updating of websites for various research groups of the faculty.

The candidate was associated with NORBIS, the National Research School in Bioinformatics, Biostatistics and Systems Biology.

## Acknowledgements

In the first place, I would like to express deep gratitude to Marc for all the advice and guidance over the complete PhD period. Thank you for the intense effort and openness since before the admission process. I will always remember that he was the cornerstone that made this opportunity become a reality. Thank you also for the advice and guidance along the process, he helped me grow and learn beyond the scientific knowledge of the topic.

I want to thank all my co-supervisors and research group. Thanks to Pål and Stefan for their support and leadership of the Center for Diabetes Research. Thanks to you it was possible to open the doors necessary to materialize this PhD position and bring it all the way to the end. Thanks in particular to Harald for being good-temper and even in the discussions, it was invaluable help to count with such stable and reliable person during all the PhD.

Also want to thank my parents and sister for their love and care since forever. You set the foundations for my life and enabled me to get where I am. Thank you for the words of encouragement at the multiple phases of the degree. Thank you for believing that it was possible even before this was real.

I want to thank my former supervisors and professors at the Universidad Autónoma Metropolitana, especially to Francisco Zaragoza, Ma. Antonieta Ortega Rodríguez, Laura Chavez, Georgina Pulido and Rafael Bracho for motivating me by their good example of commitment and care for students' academic growth. Thank you for so actively supporting me to pursue studies abroad.

Thanks to all the friends and people around me which in one way or another illuminated my way and provided company in this journey. Thanks for those evenings going out for tacos and lifting the soul. Whether I had your company for days or years I will remember your smiles and the words of support when they were necessary.

May this project lead to more useful research for understanding of the Truth.

Luis Francisco Hernández Sánchez

Bergen, 31/08/2022



---

## Abbreviations

CC	Connected component
EWAS	Entity with accessioned sequence
LCC	Largest connected component
MS	Mass spectrometry
PTM	Post-translational modification
SCC	Smallest connected component
SNP	Single nucleotide polymorphism
PSI	Proteomics Standards Initiative



## Abstract

**Background:** Understanding biological mechanisms underlying the etiology of a disease is central to establishing new prevention and treatment strategies. Technological advances progressively enable identifying more biomolecules like metabolites and proteins in specific states and forms, referred to as proteoforms. These are however currently disregarded by computational methods and tools for data interpretation and analysis.

**Aims:** Include proteoforms and small molecules in the representation of pathways as biological networks to enable their query, and study the structural changes induced compared to gene-centric networks.

**Materials and methods:** Procedures for matching omics data to pathways based on proteoforms were designed with multiple levels of stringency using the Reactome knowledgebase. Rules were defined to construct interaction networks using proteoforms as elementary units, taking into consideration their topology using network metrics such as node degree, centrality and connectivity. Implementations were done in Java and Python.

**Results:** We provide an analysis of the current knowledge of biochemical pathways, and how the structure of its representation as a network influences the interpretation of the results from biomedical studies. Subsequently, we enable the construction and query of biological networks at the level of proteoforms. Finally, we show how changing the representation of networks from gene- to proteoform-centric networks and including small molecules influence the global and local structure of the network.

**Conclusion:** Providing a refined modeling of biochemical reactions, this thesis proposes the use of proteoforms as the fundamental elements when constructing biological networks. The consequences of this novel paradigm on the global structure of the network revealed implications for the interpretation of biomedical data sets. This thesis further highlights current limitations in the current knowledge on pathways, and the challenges posed by hyperconnected small molecules.

---

## List of publications

### *Paper I*

Burger B, **Hernández Sánchez LF**, Lereim RR, Barsnes H, Vaudel M. *Analyzing the Structure of Pathways and Its Influence on the Interpretation of Biomedical Proteomics Data Sets*. J Proteome Res. 2018 Nov 2;17(11):3801-3809.

<https://doi.org/10.1021/acs.jproteome.8b00464>

### *Paper II*

**Hernández Sánchez LF**, Burger B, Horro C, Fabregat A, Johansson S, Njølstad PR, Barsnes H, Hermjakob H, Vaudel M. *PathwayMatcher: proteoform-centric network construction enables fine-granularity multiomics pathway mapping*. Gigascience. 2019 Aug 1;8(8):giz088.

<https://doi.org/10.1093/gigascience/giz088>

### *Paper III*

**Hernández Sánchez LF**, Burger B, Castro Campos RA, Johansson S, Njølstad PR, Barsnes H, Vaudel M. *Extending protein interaction networks using proteoforms and small molecules*.

*Published preliminary results as preprint.*

<https://doi.org/10.1101/2022.09.06.506730>

*Reprints were made with permission from Oxford University Press GigaScience and Journal of Proteome Research.*

***Additional Paper I (not included)***

Binz PA, Shofstahl J, Vizcaíno JA, Barsnes H, Chalkley RJ, Menschaert G, Alpi E, Clauser K, Eng JK, Lane L, Seymour SL, **Hernández Sánchez LF**, Mayer G, Eisenacher M, Perez-Riverol Y, Kapp EA, Mendoza L, Baker PR, Collins A, Van Den Bossche T, Deutsch EW. *Proteomics Standards Initiative Extended FASTA Format*. J Proteome Res. 2019 Jun 7;18(6):2686-2692.

<https://doi.org/10.1021/acs.jproteome.9b00064>

---

# Table of contents

Scientific environment .....	2
Acknowledgements.....	3
Abbreviations .....	5
Abstract .....	6
List of publications .....	7
Table of contents .....	9
<b>1. Introduction .....</b>	<b>11</b>
1.1 <i>Systems medicine</i> .....	11
1.1.1 From organisms to molecules .....	11
1.1.2 From genes to proteins and proteoforms.....	12
1.1.3 Other molecules.....	16
1.1.4 Molecular variation between and within individuals.....	17
1.1.5 Linking molecular changes to phenotypes.....	18
1.1.6 From molecular interactions to pathways .....	19
1.2 <i>Omics data</i> .....	21
1.2.1 Genomics and transcriptomics .....	22
1.2.2 Proteomics, metabolomics, and lipidomics .....	23
1.2.3 Omics integration.....	25
1.3 <i>Functional knowledge: modelling and access</i> .....	26
1.3.1 Naming conventions for participants of biological systems .....	27
1.3.2 Representing and querying biological functions.....	29
1.3.3 The Reactome knowledgebase .....	30
1.4 <i>Functional analysis of omics data</i> .....	32
1.4.1 Network analysis.....	33
1.4.2 Pathway analysis.....	35
<b>2. Objective of the study.....</b>	<b>39</b>
<b>3. Materials and methods .....</b>	<b>40</b>
3.1 <i>Navigating the Reactome pathway data model</i> .....	40
3.2 <i>Constructing the graph representation of a pathway</i> .....	42

---

3.3	<i>Network topology analysis</i> .....	43
3.4	<i>Software implementation</i> .....	44
<b>4.</b>	<b>Results</b> .....	<b>46</b>
4.1	<i>Paper I: Modelling pathways as a biological network</i> .....	46
4.2	<i>Paper II: Matching of omics data to pathways</i> .....	48
4.3	<i>Paper III: Extending protein interaction networks using proteoforms and small molecules</i> .....	50
<b>5.</b>	<b>Discussion</b> .....	<b>52</b>
<b>6.</b>	<b>Conclusion</b> .....	<b>55</b>
<b>7.</b>	<b>Outlook</b> .....	<b>57</b>
<b>8.</b>	<b>Future work</b> .....	<b>59</b>
8.1	<i>Automatic annotation of pathways</i> .....	59
8.2	<i>Multidimensional functional analysis</i> .....	61
<b>9.</b>	<b>References</b> .....	<b>63</b>
<b>10.</b>	<b>Papers</b> .....	<b>74</b>

# 1. Introduction

## 1.1 Systems medicine

The aim of clinical research can be broadly defined as preventing, diagnosing, and handling human diseases. A disease condition is a *phenotype*, *i.e.*, an observable trait of an organism. For example, the disease *diabetes* is defined by The Centers for Disease Control and Prevention as a *chronic (long-lasting) health condition that affects how your body turns food into energy*. Systems medicine considers the human body as a system where biochemical, physiological, and environmental factors interact and lead to diverging phenotypes. Understanding how these factors result in specific phenotypes is one of the major goals of modern medicine.

### 1.1.1 From organisms to molecules

When studying biological systems, researchers model the components of the system, how they function and interact, such that the whole system can in turn be better understood<sup>1</sup>. The human body consists of multiple organs performing various interconnected tasks. Organs are themselves composed of different types of cells, with distinct identities and roles. Cells are the basic units of biological function and structure. There are about 200 cell types<sup>2</sup> and approximately 37 trillion cells in the human body<sup>3</sup>. As an example, the pancreas is an organ performing exocrine and endocrine functions. The exocrine part is composed of acinar cells producing digestive enzymes and ductal cells to form channels to the duodenum. The endocrine part of the pancreas is composed of alpha, beta, delta, pancreatic polypeptide, and epsilon cells at the islets of Langerhans. Beta cells produce insulin and alpha cells produce glucagon, both necessary for glucose processing and homeostasis<sup>4</sup>.

Human cells are themselves organized in cellular compartments. Each compartment contains its own set of molecules accomplishing different tasks contributing to the function and survival of the cell. Molecules in living organisms are carbon-based and composed of sets of atoms linked by covalent bonds in specific configurations<sup>5</sup>. The main molecule categories are nucleotides, amino acids, sugars, and fatty acids. These

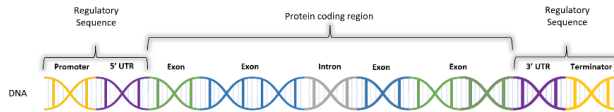
---

small molecules may bond with each other to form macromolecules. Both small and macromolecules are involved in metabolic processes as intermediate or end products. These processes compose the metabolism and make up the necessary biochemical reactions necessary to sustain life in an organism.

### 1.1.2 From genes to proteins and proteoforms

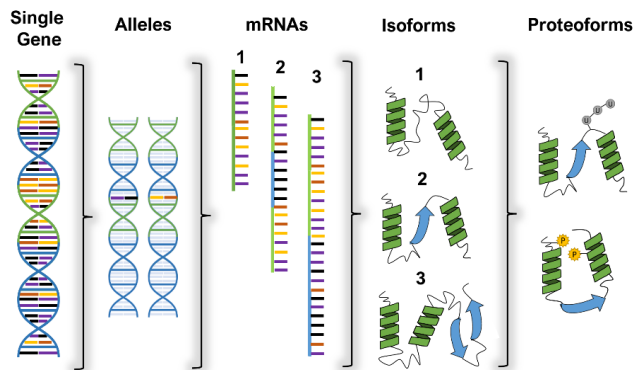
The observation that phenotypes show varying levels of heritability, *i.e.*, correlation between generations, sparked the study of the mode of inheritance<sup>6</sup>. Identifying the molecular mechanisms sustaining the inheritance of a trait holds the promise for better diagnostic, and possibly explaining the pathogenesis of diseases<sup>7-9</sup>. The essential information passed from parents to offspring is contained in their genetic material, composed of deoxyribonucleic acid (DNA). DNA is a polymer macromolecule composed of four nucleotide bases: adenine, cytosine, guanine, and thymine. Nucleotides are chained together by deoxyribose molecules in a specific order. The entire genetic material of an organism is referred to as its *genome*, forming a three billion base long code for humans<sup>10</sup>. Normally, the human genome is organized in 23 pairs of chromosomes, each pair being constituted by one maternal and one paternal chromosome. As opposed to a phenotype, the entire genetic makeup of an individual is called its *genotype*.

The DNA sequence contains discrete units of inheritance called genes<sup>7</sup>. Genes have specific start and end points, marked by sequence patterns called promoters and terminators (**Figure 1**). The region comprised between a promoter and a terminator is composed of an open reading frame (ORF) flanked by untranslated regions (UTR). Genes are used as a reference to synthesize other molecules and notably the proteins. For this, the DNA of a gene is first transcribed as a ribonucleic acid (RNA), a process named transcription. The RNA sequence is also called a *transcript*, and as for the genome, the entire set of transcripts in an organism or a sample is called the *transcriptome*.



**Figure 1:** General structure of a human gene. Composed in general by regulatory and protein coding regions. Regulatory regions contain specific starting and ending sequences: promoters, 5' UTR initiation sequence, and 3' UTR termination sequence. The protein coding region, also called open reading section, codes for mRNA transcripts that will be used to encode proteins. Formed by introns and exons used for alternative splicing.

During transcription, the bases of DNA are coded in a new chain of bases replacing thymine for uracil – the other three bases remain the same and are simply copied into the new strand. DNA hence stays untouched inside the nucleus of the cell at all times, and the transcripts exit the nucleus and serve as messengers for the information contained in the DNA. The newly made transcript is called precursor messenger RNA and is processed into a mature molecule by removing the non-coding parts of the sequence referred to as introns. The remaining parts are bound together and referred to as exons. This process, called splicing, results in that different mature RNA sequences can be obtained from the same gene. Such alternative splicing will eventually lead to different protein sequences called *isoforms* (**Figure 2**).



**Figure 2:** Multiple sources of variation can result in multiple proteoforms. Genetic sequence variation can lead to difference in gene sequences. Alternative splicing can lead to multiple transcripts and then protein isoforms. They in turn may be altered via post-translational modifications, yielding different proteoforms.



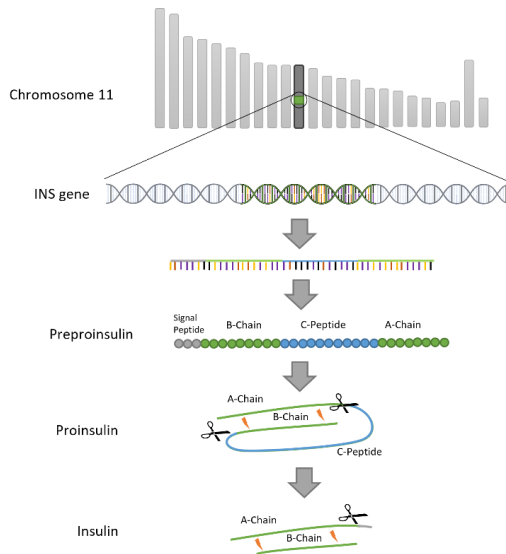
---

The information contained in the mature mRNA is subsequently translated into a chain of amino acids by ribosomes. Three nucleic acids, a codon, encode one amino acid. Translation starts and stops at so-called *start* and *stop* codons, and the sequence comprised between these two codons is called the coding sequence. Each amino acid, also called a residue, has a specific chemical composition and structure giving them unique properties with regards to polarity, size, pH, molecular weight, and hydrophobicity. When connecting multiple amino acids, the differences in their properties make the proteins adopt specific three-dimensional conformations.

Proteins have a very dynamic nature, in the sense that their sequence and shape can be tuned to achieve specific functions. Such structural changes of a protein are called post-translational modifications (PTMs)<sup>11</sup>. PTMs include proteolytic cleavages or addition of a chemical group to an amino acid residue, and cross-linking of the polypeptide chain<sup>12</sup>. In the case of the addition of a chemical group, it can be a single and simple group, like phosphorylation<sup>13</sup> or acetylation, or more complex structural modifications like polyubiquitinylation or glycosylation, where complex glycan groups are attached to the nitrogen atom side chain of the asparagine residue following specific sequence motifs<sup>14</sup>. PTMs regulate the function of proteins by activating or deactivating them, by changing their properties, structure, or binding to other molecules. Taking into account isoforms and PTMs, a protein can thus come in many forms, referred to as *proteoforms*<sup>15,16</sup>. PTMs produce an exponential growth in the diversity of proteoforms, and producing these proteoforms requires a maturation process where they are translated, modified, and conformed into their functional form<sup>11</sup>.

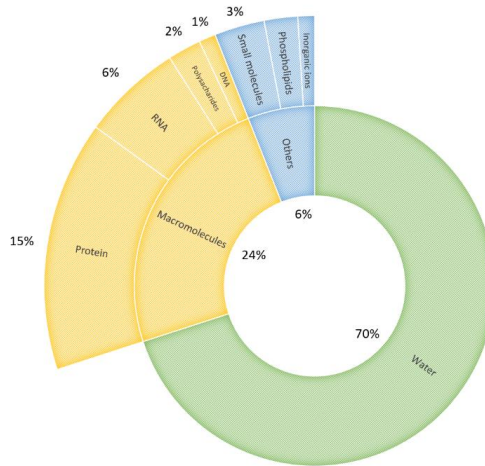
Let us take the maturation process of insulin as example; first a preproinsulin protein molecule is produced consisting of 110 amino acids. It is then transported outside the endoplasmic reticulum where a subsequence of amino acids at the beginning of the protein called the signal peptide is removed enzymatically to form proinsulin. The protein then folds into its 3D structure, such that two of its ends, called the A-chain and B-chain, are bound together with disulfide bonds, stabilizing the molecule. This molecule is then processed at the Golgi apparatus, cleaving of a subsequence of amino

acids called the C-peptide. Finally, the C-terminus of the protein sequence is removed to produce mature insulin<sup>17</sup>.



**Figure 3:** Insulin maturation process. From the INS gene on Chromosome 11 translated into preproinsulin, an amino acid chain cleaved into proinsulin and then further processed into insulin.

Proteins are the major functional constituents of cells (**Figure 4**), making up about half of the total dry mass of the cell with a total concentration of two to four million proteins per cubic micrometer (100 – 300 mg per ml)<sup>18</sup>. Like for the genome or the transcriptome, the entire set of proteins in a given organism or sample is called the *proteome*. To date, it is not possible to determine with certainty the number of participants in the proteome, but current estimates indicate potentially up to ~6 million proteoforms can be obtained<sup>19</sup> from the around 19,700 human protein coding genes<sup>20</sup>.



**Figure 4:** Molecular composition of an animal cell. The distribution is approximate by weight. H<sub>2</sub>O accounts for most of the total cell mass. Macromolecules make up for more than half of the total dry mass of the cell.

### 1.1.3 Other molecules

In addition to DNA-encoded participants of biological systems, many other types of molecules are necessary for the function of cells and organisms. Small molecules of a weight of 9 kDa or less are referred to as *metabolites*, and broadly cover organic compounds as well as circulating amino acids and short peptides. For example, adenosine triphosphate (ATP) is the most abundant energy carrier molecule in the cells. ATP hydrolysis releases energy to perform many metabolic processes such as synthesis of new molecules<sup>5</sup>. ATP also participates in phosphorylation reactions, as well as modification of amino acids during protein synthesis<sup>13</sup>.

Two additional broad classes of molecules of biological origin are lipids and sugars. Lipids are essential components of membranes and can also be involved in biochemical reactions. A well-known example of lipids is cholesterol, which helps increase the order and stability of cell membranes while keeping fluidity and diffusion rates through membranes<sup>21,22</sup>. Sugars are essential for the metabolism and a major source of energy; the most common sugar is glucose, which contributes in the production of ATP, synthesis of neurotransmitters and neuromodulators, among others functions<sup>23</sup>.

---

Finally, an often-overlooked class of participants of biological systems is inorganic elements, and notably metallic ions, which are essential to many biological reactions and represent a valuable source of information about living organisms<sup>24</sup>. Like for the genome, transcriptome, and proteome, the study of the entire compendium of metabolites, lipids, sugars, and metallic ions in a biological system is called the metabolome<sup>25,26</sup>, lipidome<sup>27</sup>, glycome<sup>28,29</sup>, and metallome<sup>24</sup>, respectively, and their mapping is a very active field of research.

#### **1.1.4 Molecular variation between and within individuals**

A fundamental source of molecular variation between individuals lies in differences in the genomic sequence. While the genome is very stable between cells from the same organism, a variety of differences can be observed between individuals<sup>30</sup>. Genomic variation can be broadly classified by their size: (i) chromosomal abnormalities concern changes involving entire chromosomes<sup>31</sup>; (ii) structural variation involves intra-chromosome changes that are large enough to alter the structure of a chromosome, typically over 1 kb and below 3 Mb, and include sequence deletions, duplications, insertions, inversions and translocations; (iii) single-nucleotide variant (SNV) or single-nucleotide polymorphism (SNP) involving the substitution of single nucleotides.

Another source of molecular variation lies in the regulation of gene expression, which can differ between individuals but also between the cells of the same individual. The most common modulators of gene expression are DNA methylation, which can turn the expression of a gene 'on' or 'off'; and histone modifications, which alter the accessibility of genetic material. When such variation yields to an inheritable trait that does not involve changes in the DNA sequence, it is referred to as epigenetic<sup>32,33</sup>.

As detailed above, gene transcription and translation can yield different isoforms of the same protein, and these isoforms can in turn be matured into different proteoforms. The nature of the proteoforms present in a cell differ greatly between cell types<sup>34</sup>. Beyond their nature, proteins also span a vast range of abundances, ranging from a few copies to about hundreds of millions of copies per gene in a single mammalian cell<sup>7</sup>. The

---

relative levels of a protein differ between organs<sup>35,36</sup>, cell types<sup>11,37</sup>, vary over time, and the difference is even more pronounced at the level of proteoforms<sup>11</sup>.

### 1.1.5 Linking molecular changes to phenotypes

Linking a biological entity to a phenotype enables its use as a diagnostic or prognostic marker. Such a marker is also called a biomarker, which can either be used alone or combined as panel<sup>38</sup>. For example, the C-peptide is a by-product of insulin production (**Figure 3**) that can be used to monitor insulin secretion in patients, and is therefore used as a marker to diagnose different types of diabetes<sup>39</sup>. Moreover, identifying the molecular mechanisms underlying a disease paves the way for better prevention, diagnosis, and therapy. Linking molecular variation (*e.g.*, genetic, epigenetic, or proteomics) to disease states (phenotypes) has therefore sparked a lot of interest in the research community<sup>6</sup> and constitutes a major goal in clinical research<sup>7</sup>. For example, in the case of diabetes, the regulation of glucose levels is impaired by the inability of the pancreas to produce insulin, or by insulin-sensitive tissues to consume glucose<sup>40</sup>. Clinical research has demonstrated that diabetes is associated with an array of environmental, metabolic, and genetic factors<sup>41</sup>. However, the complexity of biomedical systems, comprised of thousands of different molecules, changing, and interacting over time, spanning several orders of magnitudes of abundance levels, distributed across different tissues and organs, makes the linking of molecular changes to phenotypes overwhelmingly challenging.

Modern data interpretation approaches enabling the integration of large and heterogeneous data are seen as a promising way to better capture this complexity<sup>42</sup>. An objective of such approaches is to provide a representation of biomedical systems that can be interpreted mathematically and computationally across scales, from molecules to organisms, and from genotypes to phenotypes. A major challenge is then to model the biochemical reactions that cause or sustain a given phenotype<sup>5</sup>. Examples of such reactions include assembling or breaking up of compounds into smaller molecules that that serve as usable sources of energy (*e.g.* glycogen breakdown) or provide the cell with elementary building blocks (*e.g.* amino acids)<sup>5</sup>. Another important example is the

---

attachment of reactive groups, *e.g.*, a phosphorylation group, so that proteins have specific biochemical properties<sup>43</sup>.

In order to better describe biological systems, an important endeavor has been to chart all biochemical reactions and their participants. As early as 1965, a consolidated list of all known proteins was made publicly available by pioneer Prof. Margaret Dayhoff – regarded as one of the founding text for bioinformatics<sup>44</sup>. Subsequently, the sequencing of the human genome, *i.e.*, all genetic information contained in a human cell, through the human genome project<sup>10</sup> launched the systematic mapping of all human genes. Yet, sequencing the human genome revealed that the number of genes was dramatically lower than estimated<sup>45</sup>, and while it is known that proteoforms are key to achieving specific functions, molecular functions are generally constructed in a one-gene-one-protein paradigm<sup>15,16</sup>. Charting the number and identities of proteoforms in a biological system, and their function, is still a very active research field<sup>11</sup>.

For example, in the case of diabetes, the regulation of glucose levels is impaired by the inability of the pancreas to produce insulin, or by insulin-sensitive tissues to consume glucose<sup>40</sup>. Research has demonstrated that diabetes is associated with an array of environmental, metabolic, and genetic factors<sup>41</sup>.

### **1.1.6 From molecular interactions to pathways**

The participants of biological systems can interact in many ways. The term *interactome* refers to the complete set of interactions among molecules that occur in cells and are necessary to perform the processes essential to sustain life<sup>46</sup>. For example, molecules may form larger complex molecules binding to each other. Such complexes are formed by multiple copies of the same molecule or by diverse molecules. Hemoglobin is an example of a complex protein molecule that transports oxygen from the lungs to the rest of the body through the blood<sup>47</sup>. It is composed by four subunits, heme molecules bound to iron. In the complex form, when oxygen binds to one of the subunits there is a change in shape of the molecule, inducing a higher binding affinity to oxygen in the other subunits<sup>5</sup>. Such reactions require participants to be present in a specific form, *e.g.*, proteoforms for proteins, and in a given stoichiometry.

---

Another important class of reactions relates to the passing of a signal, *e.g.*, a phosphorylation group. Such reactions often occur in chains, allowing the passing of information between members of a biological system. Such a series of biochemical events leading to a particular biological function are called *pathways*<sup>5</sup>. In the context of a pathway, reactions take reactants (inputs) to synthesize products (output), activated by catalyzers, and regulated or inhibited by other molecules. Macromolecules or small molecules can both be input or output of reactions. For example, the assembly of the Origin recognition complex (ORC) can be modeled as a pathway. It consists of reactions that attach subunits to the complex, where the inputs are ORC proteins and ATP, and the outcome is the complex ORC(1-6), which binds in an ATP-dependent manner to sequence in a genome at which replication is initiated, also called the origin of replication<sup>48</sup>. Pathways that define general molecular process are referred to as canonical pathways, *e.g.*, the Wnt canonical pathway<sup>49,50</sup>, but there is no formal definition on what distinguishes canonical and non-canonical pathways. The main types of well-characterized pathways are metabolic, signal transmission, and regulatory pathways.

Metabolic pathways involve the production or change of components of the biological systems that are essential to preserve life. Typical examples are decomposition of compounds into simpler molecules or the construction of new molecules such as complexes. They are often a sequence of reactions catalyzed by proteins called enzymes and may achieve the transformation of one molecule to another through a series of intermediate steps, each catalyzed by an enzyme. For example, Glycolysis is a metabolic pathway which breaks down glucose by enzymes producing energy for the cell as ATP and pyruvic acid<sup>23,51</sup>.

Signal transmission or signal transduction pathways are the means for the cell to react to events and stimuli happening both inside and outside the cell<sup>5</sup>. This mechanism is used by biological systems to communicate. They heavily depend on receptor proteins, often located at the cell surface, and detect the presence of specific molecules nearby and transmit the signal *via* a biochemical reaction cascade, typically using protein phosphorylation, and resulting in a cellular response<sup>52</sup>. The proteins attaching

---

phosphate groups to other proteins to activate them or pass a signal are called kinases. Afterwards, other proteins called phosphatases can remove the attached phosphate group, hence reverting the modification. This enables the cell to activate or deactivate a protein without having to express a gene. For example, the PI-3-Kinase-Akt signaling pathway is activated by insulin, promoting the survival and growth of multiple cell types. They depend on the activation of Akt by phosphorylation<sup>5</sup>.

Regulatory pathways regulate quantitative metrics of the cell, *e.g.*, regulation of cell growth, proliferation, or regulation of gene expression. They commonly depend on extracellular factors which activate translation initiation. They also trigger the intake of nutrients and production of energy<sup>5</sup>. For example, the pathway involving mTOR regulates cell growth by means of an activated growth receptor and an activated Akt protein, phosphorylated by the PI-3 kinase<sup>5,53</sup>.

In conclusion, phenotypes are influenced by biological mechanisms, themselves dependent on the molecules available at the cellular level. Knowledge of elucidated mechanisms may provide better understanding, diagnostic, and intervention. Using the pathways inherent to biological systems, clinical researchers aim at building models that can pinpoint a specific set of reactions key to a given pathology. The ability to predict outcomes of biological pathways holds the promise to modify or even repair the cause of diseases<sup>54</sup>, for example, in the context of gene therapy<sup>55</sup>.

## 1.2 Omics data

In an attempt to characterize the different *omes* (genome, metabolome, proteome, *etc.*) in biological samples, various *omics* (genomics, metabolomics, proteomics, *etc.*) were established<sup>7,56,57</sup>. Strong research efforts are also invested in the combination of multiple omics to benefit from their complementarity and expand the capacity to understand biological systems. While omics data acquisition is not the primary focus of this thesis, the meaning, representation, and limitations of the data sets provided have a strong influence on functional analysis, notably impacting *Paper I* and *Paper II*, and will thus be introduced in the context of their use in downstream integrative



---

analyses. It is important to note that the omics technologies, whether they probe the genome, transcriptome, proteome, or metabolome, require reference knowledge to consolidate the result set, *e.g.*, a reference genome, proteome, or set of metabolites. The availability and quality of this reference knowledge varies over time and between species, and is continuously improved<sup>58,59</sup>. Missingness or errors in the reference can greatly influence the outcome of an analysis, and the results must be interpreted with this limitation in mind<sup>46,59</sup>. Note that this thesis focuses solely on the analysis of human samples and can thus take advantage of extensive amounts of reference knowledge.

### 1.2.1 Genomics and transcriptomics

The comprehensive analysis of the genome has been enabled by the discovery of polymerase chain reaction (PCR), which enables the amplification of the signal encoded in DNA<sup>60</sup>. Genomic results can present themselves in very different forms based on the type of analysis conducted. If the genome was sequenced, information will be available on nearly all the available DNA of the individual<sup>61</sup>. However, in most cases, sequencing the entire genome is expensive and not necessary. An alternative consists of sequencing only the protein-coding regions of the genome, referred to as the *exome*. Another solution to probe the genome at lower cost is *genotyping*, where specific sequence variants are targeted by specialized assays. In this approach, data is only obtainable for the variants available in the assay. For common variants, it is however possible to impute the most likely alleles present in the sample due to linkage disequilibrium<sup>62</sup>. This technique is notably very popular in genome-wide association studies<sup>63</sup> (GWAS). GWAS leverage large cohorts in order to study the association of sequence variation with phenotypic traits, *e.g.*, variants within introns of the *FTO* gene are associated with the body mass index (BMI) in humans<sup>64</sup>. Note that genotyping data can also be used to study structural variation<sup>65</sup>.

For genomic methods, the results are summarized in terms of sequence or structural variation, *e.g.* the set of common and rare SNPs found to be associated with a trait<sup>63</sup>. Functional analyses then summarize the data per pathway or gene cluster to provide biological context to the genomic variation<sup>66</sup>. Analysis methods rely on the current knowledge of genome structure, largely based on the international collaborative efforts

---

like the HapMap Project<sup>67</sup> and the Human Genome Project<sup>68</sup>. However, when using genome-wide data covering non-coding regions, linking variation to the actual effector gene(s) is a major challenge, making it difficult to explain why there is an association between a variant and a phenotype<sup>69</sup>. While in the majority of cases, the effector gene is the nearest gene<sup>70</sup>, the functional mode of action of variants can be much more complicated, and hence taking the nearest gene can lead to erroneous conclusions<sup>71</sup>. In the example of *FTO*, it has taken many years of research to uncover that the effect of the variants on BMI was most likely mediated through Iroquois-class homeodomain protein IRX-3 and IRX-5.

As for DNA, RNA can be analyzed from low sample amounts with high sequence coverage<sup>72</sup>. In biomedical sciences, transcriptomic analyses generally aim at inferring the abundance of transcripts in samples<sup>73</sup>, and the transcript abundance is in turn used as a proxy for gene expression. The transcriptomic data processed in functional analyses thus present themselves as a list of abundance estimates summarized per gene or transcript<sup>74</sup>.

### **1.2.2 Proteomics, metabolomics, and lipidomics**

Unlike for DNA, protein sequences cannot be amplified, and the large-scale analysis of proteins is therefore dependent on the sensitivity of the protocols and instrumentation. One way to detect a protein is to use antibodies that bind to the protein of interest which can then be detected and quantified with specialist assays. Antibody assays have been developed that can monitor a large number of proteins in biomedical samples<sup>75</sup>. Over the past decade, aptamers have been introduced as an alternative allowing the screening of thousands of proteins<sup>76</sup>. Both technologies provide protein abundances summarized per protein, in a one-gene-one-protein paradigm. Therefore, while they allow the screening of large numbers of samples at a moderate price, they do not allow distinguishing different proteoforms. Which proteoform is detected is in fact unclear based on the specificity of the antibodies used in this method, therefore concerns have been raised regarding the specificity of these techniques<sup>77,78</sup>.

---

An alternative that has lower sensitivity and is less scalable but allows the characterization of different proteoforms to the level of isoforms, sequence variants, and PTMs, is mass spectrometry<sup>18</sup> coupled to liquid chromatography (LC-MS)<sup>79</sup>. In LC-MS-based proteomics, biomolecules are ionized, and their mass-to-charge ratios measured as they fly through an electromagnetic field. The biomolecules can be measured in both targeted and untargeted manners, the latter attempting to cover the proteome as widely as possible.

Mass spectrometry-based proteomic approaches are generally categorized as either top-down, where entire proteins and proteoforms are analyzed, or bottom-up, where the proteins are first enzymatically digested into peptides, and the presence of a given protein or proteoform is inferred from the detected peptides<sup>80</sup>. Top-down approaches have the advantage that one can analyze entire intact proteins and proteoforms, but suffer from low proteome coverage, while bottom-up strategies enable detecting the product of thousands of different genes, but there is often ambiguity regarding the protein or proteoform that led to the detection of a given peptide<sup>81</sup>. Overall, bottom-up proteomics techniques help get a wider coverage of the proteome while top-down focuses on detailed characterization of the proteome as it can identify complex molecules and their components<sup>82</sup>.

Similar to transcriptomics, proteomics results are generally summarized per protein on a one-gene-one-protein basis. This presents the disadvantage that proteoform-level information is not available for functional analysis. This is further discussed in *Paper II*, and a standard to encode the information at the proteoform level was established as part of *Additional Paper I*<sup>83</sup>.

Metabolites and lipids<sup>27</sup> can also not be amplified, and the breadth and depth of the analysis therefore relies purely on analytical performance<sup>84</sup>. The two most widely encountered technologies to analyze these molecules in biomedical samples are LC-MS<sup>84</sup> and nuclear magnetic resonance (NMR)<sup>85,86</sup>. For LC-MS analyses, the analytes are processed in a similar way as in top-down proteomics. In NMR, the samples are inserted into a strong magnetic field and the frequency of resonance of their nuclear

---

spin enables distinguishing the different molecules in the samples. For both techniques, a list of quantified biomolecules is obtained, along with their abundance estimates.

### 1.2.3 Omics integration

To take advantage of the complementarity of the information provided by the different omics technologies, recent research has focused on the acquisition and integration of multiple omics data sets from the same sample<sup>56</sup>. For example, the large-scale combination of genomic and transcriptomic data has enabled the computation of genome-wide quantitative trait loci (QTL), highly valuable for the interpretation of GWAS association signals<sup>87</sup>. Similarly, combining genomics and metabolomics can shed light on the genetic influence on metabolisms and pathways<sup>88</sup>, while the integration of genomics and proteomics has been very valuable in multiple fields, providing a new layer of knowledge to medical and population genetic studies<sup>89-91</sup>. Altogether, integrating multiple omics data is a promising avenue to better understand biological systems. To achieve this, the ability to exchange data in the same or compatible formats is essential.

The combination of protein assays with genomic analyses can be particularly valuable given that proteins can in turn be used as biomarkers and are prime candidates for drug targets<sup>92</sup>. At the same time, integrating genomics and proteomics, often termed *proteogenomics*, has been central to recent advances in cancer research. For example, tracking how genomic structural and sequence variation is affecting gene expression, proteins, and modified protein levels provided an unprecedented view at the mechanisms sustaining cancer cell survival and proliferation<sup>90,93</sup>. Furthermore, the detection of peptides and proteins produced only by cancer cells holds the promise to design better immunotherapies and maybe even vaccines against cancer<sup>94</sup>.

Two proteins physically involved in a reaction are said to be interacting, and the central role of protein-protein interactions in biological processes have yielded great interest in the study of protein-protein interactions<sup>95</sup>. There are various experimental approaches for elucidating binary interactions among proteins. Some techniques devise large-scale assays to test interactions for hundreds or thousands of proteins at once,

---

while others test specific pairs of proteins in small-scale assays<sup>95</sup>. Results of both techniques can be aggregated and contribute drafting the complete interaction map with all proven interactions referred to as the *interactome*. *Yeast-two-hybrid* (Y2H)<sup>96</sup> is one of the most reliable and common techniques for detecting direct interactions among two proteins<sup>97</sup>. Other techniques use affinity purification coupled with mass spectrometry (AP-MS)<sup>98</sup>, which enable the characterization of protein complexes in multiple subcellular locations using the affinity of a bait protein to pull down direct and indirect interaction partners.<sup>95</sup> Protein-protein interactions can be aggregated in an interaction network for a more comprehensive view of the interactome. Networks can be studied programmatically using methods from graph theory and network science, facilitating the understanding of complex mechanisms by their network properties<sup>95</sup>.

### 1.3 Functional knowledge: modelling and access

The study of biological systems produces vast amounts of knowledge on their molecular entities and processes. Historically, researchers would consult textbooks, scientific papers, and domain experts to find the function of a biomolecule. The vast number of biological processes and their participants, as well as the ever-increasing technological and scientific throughput, especially with the advent of integrative omics approaches, made manual mining of functional knowledge rapidly intractable. New solutions are therefore constantly needed to aggregate previous and new knowledge on biological systems. Large databases were established as a solution to consolidate vast amounts of knowledge on biological functions. They respond to current challenges with the help of modern computational methods and technologies.

Often called *knowledgebases*, they focus on different objectives and data, and may focus on the entities themselves, or their relationships. Examples of databases focused on concrete types of entities are UniProt<sup>99</sup> which contains a comprehensive curated compendium of protein sequences; RefSeq<sup>100</sup> which contains a collection of sequences including genomic DNA, transcripts and proteins; Ensembl<sup>101</sup> which annotates genomes and provides multiple tools for genome browsing. Functional knowledgebases consolidate information on biological entities and the functions that

---

can be performed by those entities. Examples are WikiPathways<sup>58</sup> and Reactome<sup>59</sup> for pathways, BioGRID<sup>102</sup> for molecular interactions<sup>102</sup>, and OMIM<sup>103</sup> for a catalogue of human genes and their associated genetic disorders<sup>103</sup>.

These databases gather knowledge either by a manual effort to annotate the knowledge contained in the literature, also named curation, or with automatic mining and annotation. Manual curation is commonly done by expert curators and is by no means exhaustive. It may suffer from selection biases<sup>104</sup>, a problem that is further explored in *Paper I*. Automated annotation can be done through text mining of the literature<sup>105</sup>, or via reprocessing of public experimental data<sup>106</sup>.

In order to enable the computational query of data, biological entities and functions must be represented in standardized computer-readable formats. Depending on the omics field, there are multiple international organizations who manage standards for data representation. One example in proteomics is the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI), providing standards for proteomics-related formats. HUPO-PSI defines standards that facilitate data comparison, exchange, and verification. There are similar organizations for other omics such as the Global Alliance for Genomics and Health. Multiple factors need to be taken into account when designing standards, including storage space, access speed, or human readability. While this is achievable for simple entities, some biologically relevant details, *e.g.*, regarding isoforms or PTMs is often not available, not encoded, or cannot be queried, a problem central to *Paper II*. The representation of biological functions, which involve multiple entities with different roles is even more complex and designing a one-size-fits-all representation is very challenging – a problem discussed in all papers of the thesis.

### **1.3.1 Naming conventions for participants of biological systems**

In order to enable the naming of the same entity across resources and experiments, different nomenclatures, identifiers, controlled vocabularies, and ontologies have been developed. A *nomenclature* is a system for naming certain types of objects. One very common example is the HUGO Gene nomenclature<sup>107</sup>, which assigns unique character

---

strings as identifiers for genes discovered to contribute to a phenotype or function. They consist typically of 3 to 6 uppercase Latin letters and Arabic numbers, *e.g.*, *INS* for the gene encoding insulin. It is often used to encode and compare omics results to the level of genes. Even though the scientific community is putting tremendous efforts in the development and adoption of standards, different definition, standards, and naming schemes for the same molecules persist, and are challenging to harmonize<sup>108</sup>. Comparing data sets using HUGO gene names is for example challenging as there is not a one-to-one mapping between gene names and proteins or genes in knowledgebases.

An example of gene variant identifier system is the *rsID* defined by *dbSNP*<sup>109</sup>, a database for single nucleotide variants. It assigns an identifier composed of letters and numbers starting with the letters “rs” to genetic variation at a specific locus of the genome. For example, rs689 is a single nucleotide variant changing nucleotide A to G or T at an intron of the insulin gene, located at chromosome 11 locus 2160994. For protein sequences the most adopted identifier system is the UniProt<sup>99</sup> *accession number*. UniProt accession numbers are stable identifiers composed of six to ten letters and numbers. For example, P01308 is the accession number for insulin.

Multiple controlled vocabularies were established to standardize the representation of PTMs so that they are shareable and compatible across databases. Examples include the UniProt controlled vocabulary for post-translational modifications<sup>110</sup>, Unimod<sup>111</sup>, PSI-MOD<sup>112</sup>, and the RESID Database of Protein Modifications<sup>113</sup>. Of special interest is PSI-MOD as it is a community standard for representation of protein modification data maintained by the HUPO-PSI and used in Reactome pathway annotations. It is defined as an ontology for protein chemical modifications organized as a hierarchy of modification types, where there are 45 top categories and a total of 2,098 types, with a maximum of nine levels of depth in the hierarchy. Modifications contained in PSI-MOD are defined as covalent modifications, or changes that alter the measured molecular mass, of a peptide or protein amino acid residue.

---

The definitions of PTMs in one controlled vocabulary are often poorly translatable to another, and here again, one-to-one mappings do not exist, greatly challenging the representation of modified proteins. The increasing ability to detect and study specific proteoforms sparked the development of standards for the representation of proteoforms<sup>114</sup>. Recently, two standards have been established, Proforma<sup>115</sup> and PEFF (*Additional Paper I*)<sup>83</sup>. Due to the lack of available format at time of writing, we established our own representation of proteoforms as a string for *Paper II*.

### 1.3.2 Representing and querying biological functions

Multiple dedicated databases provide associations between biological entities and functions. Common examples of databases of this type associate genetic variants with phenotypes or diseases, like OMIM<sup>103</sup>, ClinVar<sup>116</sup> and PheGenI<sup>117</sup>. There are also databases with associations between genes or proteins such as BioGRID<sup>102</sup> or IntAct<sup>118</sup>. In an attempt to represent the entire interactome, these databases link all entities possibly interacting, with very limited biological context. Conversely, pathway knowledgebases base their representation of biological functions to the level of biological processes, with rich context and details on the biological function. Prominent examples are KEGG<sup>119</sup>, WikiPathways<sup>58</sup> and Reactome<sup>59</sup>. They contain rich information on consensual textbook biological pathways.

Some resources like WikiPathways rely on crowdsourcing, while KEGG and Reactome are manually curated by experts. The manual curation process in Reactome stores pathway data into a relational database where reactions and pathways are then modelled by making participating molecules instances of entity classes, and their interactions are modelled as relationships part of biochemical reactions. Pathway knowledge in Reactome is also accessible as a graph database based on Neo4j<sup>120</sup>, where access to data is much faster than with a relational database and complex queries can be formulated using Cypher, a language to query graph data.

Once biological functional knowledge is stored in an organized and standardized data model, it is necessary to provide user interfaces for manual or programmatic access, *i.e.*, a guided user interface (GUI) or an application programming interface (API)<sup>120,121</sup>.



---

APIs make it possible to programmatically query functions of thousands of entities at a time, study their interactions, and link to the relevant literature. Examples of modern knowledge REST-based APIs are the KEGG API and the Reactome Content Service API for direct data model queries to the Reactome graph database<sup>59</sup>. These interfaces allow *querying* knowledgebases, *e.g.*, returning the information relevant to a given set of entities. Depending on the limitations of the GUI or API, or internal representation, the data must be converted into a specific format before being queried. Often, these require representing the data as a list of gene or protein names or identifiers, which dramatically limits the representation of complex omics data.

Once requests are sent to the knowledgebase it is necessary to map the content of the query to the reference data contained in the system. This might involve conversion of identifiers of different naming conventions. It must also be decided which entities are considered equivalent, which can become challenging when mapping, *e.g.*, gene names to protein identifiers. This process of *data mapping* may be conducted internally, and is not always well documented, and can yield ambiguity in the results. This is particularly the case for proteoforms, which can suffer from missing isoform or PTM annotation, or ambiguity in the PTM localization – a problem for which we propose solutions in *Paper II*.

### 1.3.3 The Reactome knowledgebase

The Reactome knowledgebase is central to the work conducted in this thesis. Reactome contains biomolecular pathways with details on the steps and molecules performing them. Pathways annotated are related to signaling and metabolic cellular processes, as well as hereditary and acquired disease processes<sup>59</sup>. Reactome is free to access and use for research purposes. It has a manual curation process by experts who annotate data from peer-reviewed literature. There is a strong focus on human data, but it also features annotations for other species. The database is made available through a web user interface and programmatic tools which are free and open-source, meaning that the code executing operations to read from the database and analyze the data can be transparently inspected<sup>122</sup>. These bioinformatic tools make it possible to search, visualize, and perform over-representation analysis of sample data in the pathways.

---

Analysis results are visualized with a color scheme in the pathway explorer visual interface and also displayed as a table of numerical results, which can be used to interpret the biological state of a sample<sup>123</sup>.

Reactome also provides an open REST-based API called Content Service which allows programmatic access of the knowledgebase. Pathway knowledge in Reactome is stored as a relational database populated by the curators. To facilitate and speed up access to knowledge, a graph-based database version built on Neo4j is also available<sup>120</sup>. The complete database content can be downloaded and queried locally via Neo4j using the Cypher query language. This results in a very convenient way to explore the content of Reactome, relying on the expressive and intuitive characteristics of Cypher for writing complex graph traversing queries.

Biomolecular processes are represented by a large network of molecular transformations organized in a hierarchy of pathways. Each pathway encompasses a set of ordered events where molecules participate, leading towards molecular transformations of chemical compounds necessary for achieving biological goals. Molecular events are mostly reactions which have participant physical entity molecules performing the roles of reactants (input), products (outputs), catalyzers and regulators. Another type of events, named *ReactionLikeEvents*, are molecular transformations that are not necessarily a proper chemical reaction, such as polymerization, depolymerization, or failed reactions<sup>122</sup>. The majority of event participants are human proteins, which means Reactome effectively contains a systematic description of protein functions. The reactions between participants describe functional relationships among molecules, including both protein-protein and protein-metabolite.

As of Reactome version 78, there are annotations for 10,726 human proteins, covering about 52.5% of the predicted human protein-coding genes of Ensembl (release 104), participating in 13,890 reactions curated from more than 34,000 scientific publications<sup>59</sup>. Reactome contains functional annotation on 29,466 proteoforms combining all species. In comparison, the Proteoform Atlas<sup>34</sup> contains 5,705,254 proteoforms deriving from 100,687 proteins at time of writing. The large difference in

---

proteoform numbers shows that the vast majority of proteoforms do not have curated functional annotation as of now.

One key aspect of Reactome is the fact that when proteins need to be in a specific state for reactions to happen, this information is annotated in the graph database. Annotations include the minimal requirements of sequence variant (protein isoform) and a set of PTMs that a protein needs to carry for a certain task, possibly with sequence-level information. With this fine level of detail, it is possible to infer proteoforms by placing together the protein isoform and set of PTMs necessary to participate in a reaction with the annotated role. These requirements are minimal – a protein may carry a larger set of modifications when performing a reaction. The curation process of Reactome does not focus on the full characterization of proteoforms, but attempts annotating the minimal set of conditions for a proteoform to participate in a given reaction.

## 1.4 Functional analysis of omics data

Mapping omics data to functional knowledgebases is seen as a promising way to go from a list of identifiers to biological knowledge. Using functional knowledgebases as a roadmap to navigate large omics data, statistical procedures are employed to extract a signal from the noise. This procedure, often called *functional analysis* of omics data, can result in the identification of genes and proteins key to a given process, or help find the biological functions most likely to be affected by a given set of genes, proteins, or proteoforms. Instead of a procedure resulting in a definitive answer on the biological function causing a disease, functional analyses should rather be considered as providing a guide for the interpretation of complex data, and help generating hypothesis for follow-up studies.

By design, the performance of functional analyses strongly depends on the amount and quality of information consolidated in knowledgebases, which we analyzed in the case of Reactome in *Paper I*. Another important factor, the ability to match the experimental data to the knowledgebase, and the internal representation of the knowledgebase can

---

greatly influence the results of a database query, as we demonstrated in *Paper II* and *Paper III*. To extract relevant functional knowledge in relationship to a query, different statistical methods were developed, that allow answering different types of questions.

### 1.4.1 Network analysis

Network analyses model the interactome as a biological network, where participants of biological functions are represented as nodes, and nodes are connected when the participants can interact. The interactions used to build the network can be obtained from various sources: text mining, co-expression, inference from similar organisms, experimental interaction, or known biochemical pathways<sup>124</sup>. By using methods from the mathematical study of graphs, network-based functional analyses provide information that cannot be obtained without information on the connectedness of biological entities<sup>125</sup>.

General topological properties measured in networks are *network size* and *degree distribution*. The network size corresponds to the number of nodes and connections in the network, while the degree of a node is the number of connections to other nodes, *i.e.*, the number neighbor nodes directly connected. When a node has notably higher degree compared to the rest of the nodes of the network it may be referred as a *hub node*. Biological networks tend to follow a power-scale degree distribution meaning that most nodes have a low degree while a few nodes have very high degree<sup>95</sup>. Another property related to the connectivity in the network is *clustering*. It is a score measuring how often neighbors of a node are also neighbors with each other. Biological networks tend to have high clustering score, meaning that when two proteins interact, their neighbors tend to interact as well. This is due to the tendency of molecules to work together to achieve processes, for example, forming complexes.

Networks are composed of connected components, which are sections of the network where each node can reach any other node in the same component by traversing connections either directly or through intermediate nodes. According to the hypothesis that molecules participating in the same function cluster together, molecules with

---

shared functions tend to be in the same connected component. Hence, when the function of a certain protein is not known, its functional profile can be inferred from its neighbors in the interactome. The interactome annotated in Reactome has grown so much that it has become a single large entangled connected component, surrounded with many small, connected components of rare or novel proteins, as studied in *Paper I*<sup>126</sup>.

The calculation of the distance between nodes, defined as the minimal number of connections needed to be traversed on average to go from one node of the network to any other node, provides a measure of the functional distance between proteins. Biological networks are referred to as scale-free and “small-world”, meaning that two proteins can be connected by a small number of connections, notably thanks to the presence of large hyperconnected hubs. Alternative distance metrics like *diffusion state distance* were established to provide meaningful measures of distance despite the presence of large hubs<sup>127</sup>. Distance measures have proven useful, *e.g.*, partition the genes associated with a disease based on their mode of action, as done for type 2 diabetes<sup>128</sup> with a remarkable agreement with other independent approaches<sup>129</sup>.

Distance metrics enable clustering procedures to decide which entities are related functionally to others. Once a cluster of interest is identified, a technique called *guilty by association* proposes to investigate the neighbors the molecules in this cluster for their association with the disease<sup>130</sup>. This rationale is for example used to conduct gene set enrichment analyses accounting for the connectedness of genes in the interactome<sup>131</sup>. On the other hand, there are known limitations to clustering techniques like *ties in proximity problem*<sup>132</sup> when distances among many entities is the same. Proteoforms may help mitigate limitations by increasing the specificity of the nodes and changing the topology of the network as we studied in *Paper III*.

A specific case of protein-protein interaction is the binding of proteins as part of complexes<sup>133</sup>, where larger structures are required to perform functional tasks. Protein complexes have gained particular interest due to their central role in molecular biology, and because they strongly impact the abundance of proteins, which can be used to

---

monitor the functional status of cells<sup>94,134</sup>. Specific databases<sup>135</sup> and analysis tools are available to interpret data in terms of complexes<sup>136</sup>.

Another strategy is *guilty by profiling*<sup>104</sup> where it is assumed that molecules share function or properties when their expression or abundance is correlated. This can be used to enrich networks by accounting for co-expression<sup>137,138</sup> or co-abundance<sup>139</sup>. Networks are also seen as powerful tools to combine multiple omics data, allowing the comparison of samples across multiple biological layers simultaneously<sup>140,141</sup>. Hybrid networks aggregating vast amounts of heterogeneous data have also been developed as a powerful way to combine experimental and literature knowledge<sup>142</sup>.

Since the topological properties of biological networks greatly influence such analyses, they have been under constant study<sup>143</sup>. For example, on highly connected scale-free networks, estimating the distance between different points however becomes a challenge, and solutions have been proposed to reduce the influence of ubiquitously connected nodes<sup>127</sup> or to reduce the density of connections<sup>144</sup>. Changing the underlying representation of the biological network, *e.g.*, by accounting for isoform-specific interactions<sup>145</sup>, has also been shown to influence the architecture of biological networks. When the properties of the network change then functional results are influenced, this is further explored in *Paper I* and *Paper III*.

### 1.4.2 Pathway analysis

Pathway analysis aims at understanding the complex biological processes that underlie diseases. By mapping omics data to pathway knowledgebases, it can provide insights into the underlying causes of the disease and offer potential targets for therapy<sup>146</sup>. For this, statistical methods are required to identify key molecules or mechanisms driving the differences in expression or phenotype represented by data obtained from biological samples. An advantage over interaction network analyses is that pathway analyses can use the structure and rich annotation of pathway knowledgebases, but in turn suffer from limited coverage of nodes and interactions.

In its most simple form, pathway analysis consists of using a pathway knowledgebase as a very detailed reference map for the interpretation of biomedical data. For example,

---

to display observed protein abundances linked by common biological processes and investigate whether the observed values agree with known regulatory processes. Such investigation also allows identifying other participants of the pathway of interest that are of value but were not detected or did not pass significance thresholds of a large-scale analysis. The pathway participants can then be included in follow-up analyses, *e.g.*, through a validation experiment. This was, for example, used to test the inhibition or activation of the Notch signaling pathway in breast cancer, highlighting the potential value of this target for therapeutic intervention<sup>147</sup>.

Often, samples are analyzed against all known pathways in an agnostic fashion. The main method used is over-representation analysis, where the results of an experiment are queried against a pathway knowledgebase. Each pathway is then evaluated based on its coverage by the sets of molecules queried, and the statistical significance of this coverage is reflected as a  $p$ -value that evaluates the likelihood of matching this pathway by chance. Reactome calculates the statistical significance with a binomial test<sup>122</sup>. Pathways with the lowest likelihood to be covered by the queried molecules by chance have a very low  $p$ -value, meaning that the likelihood of finding a pathway with this many or more sample entities as participants by chance is very low. The  $p$ -value is then corrected for multiple hypothesis testing over all pathways using, for example, the Benjamini-Hochberg approach<sup>148</sup>, providing the user with a false discovery rate (FDR)<sup>122,146</sup>. The FDR provides an estimate of the share of random matches in the list of pathways of highest significance. The pathways passing a desired FDR threshold, typically 0.05, are deemed of particular interest for the analysis.

Examples of pathway over-representation analyses are ubiquitous in the biomedical literature, *e.g.*, to identify drug targets<sup>149</sup> or key genes and the pathways sustaining disease progression<sup>150</sup>. Sample data used to perform this type of pathway analysis commonly consist of sets of gene or protein identifiers selected by a previous analysis. Such methods are thus restricted to a paradigm where pathway information is summarized at the gene level, hence ignoring proteoform-level annotation. Consequently, pathways involving specific protein isoforms or modified proteins can

---

appear as artifactually covered. In *Paper II*, we implemented a method that makes use of the proteoform-level annotation of Reactome to refine over-representation analyses.

A major drawback of such analyses is that pathways are considered as independent units. Hence, they are scored separately, ignoring their hierarchical structure and the fact pathways form a complete interconnected network where they overlap, cross, and even cause each other<sup>151</sup>. Another drawback is that calculations give the same importance to all processes, ignoring rich information both in the input, e.g., abundance levels, and in the knowledgebase, e.g., types of interaction. Furthermore, the pathway topology is overlooked, and pathways will receive the same statistics independently of whether the matched entities belong to the same reaction or are completely disconnected.

Recent research in systems biology aimed to provide a better integration of biomedical data in the complex network of interactions formed by pathways. Some approaches include not only significantly differentially abundant entities, but include also those which change little but may work together with other molecules, e.g., to study differential gene expression profiles<sup>152</sup>. Other approaches propose to consider relationships between entities, their abundances across time, cell types or subcellular localization, and network topology<sup>153</sup>. One example is NGSEA, which calculates the enrichment score of gene sets with the expression of both individual genes and their neighbors in the functional network<sup>131</sup>.

Pathway analyses are highly dependent on the definition of pathways, which differ between knowledgebases and between versions. Some pathways have reached more consensus than others as a result of different curation guidelines and priorities. Curation introduces biases in the prioritization of pathways and participants to annotate and in their representation in the knowledgebase<sup>104</sup>. The manual curation and annotation of pathways also strongly limits the pace at which interactions can be annotated. As a result, pathway-derived interactomes are orders of magnitude smaller than networks obtained from high-throughput text mining or experimental procedures.





## 2. Objective of the study

The purpose of this work is to contribute to the understanding of biochemical pathways by studying their representation in the Reactome knowledgebase and refining their representation and query using proteoforms.

- **Study I:** Study the current state of pathway knowledge in Reactome and its representation as a network.
- **Study II:** Enable the query of Reactome and the building of networks using different levels of granularity, notably using proteoform representation and matching.
- **Study III:** Evaluate the effects of augmenting gene-centric biological networks using proteoform-specific information and small molecules.

---

## 3. Materials and methods

### 3.1 Navigating the Reactome pathway data model

We used Reactome<sup>59</sup> as reference database to get knowledge about pathways. Reactome was selected because it contains details on the pathway participant proteins which includes isoform and modification information. We used this to produce a representation of pathways at the level of proteoforms, perform pathway analysis using proteoforms, and also construct proteoform interaction networks. The large set of resources including the visual pathway browser, search web pages, and the wide programmatic access to Reactome made it best suited for this study. At the time of implementation, Reactome was the only knowledgebase which situated proteoforms in the context of their biological processes as pathways, in a form that can be queried computationally.

One of the challenges when performing pathway analysis and network analysis was selecting the basic entity level: gene, protein, or proteoform. In some pathway databases participants are represented by gene names, while in others they are represented by protein identifiers. Reactome has a data model that attempts at modelling biological process at a finer level, featuring multiple identifiers and rich information on the participants of reactions. The data model employed consists of objects with relationships among them. Each object is an instance of an object class. The class hierarchy is defined in the database schema. One of the most central classes is the *Physical Entity*. It represents all types of molecules participating in the pathways and reactions. They can be sequence entities polymers, multi-molecular complexes, drugs or even sets of entities. Each of the individual physical entity objects comprises its chemical structure, including potential covalent modifications, and the subcellular location where it can be found.

Subclasses of physical entities include *entities with accessioned sequences* (genes, RNA transcripts, proteins), *simple entities* (chemical elements, compounds, and small molecules like metabolites and molecular complexes. Each of these is an identified

---

molecule that has a reference to a specific external reference database. UniProtKB<sup>99</sup> accession numbers are used to identify proteins. Simple entities are referenced with ChEBI<sup>154</sup> identifiers. For genes, there is less standardization in the naming conventions, thus a physical entity object might have multiple gene names as a list, where some of them follow the HUGO gene nomenclature<sup>107</sup>. For our study we focused on the more unique protein accessions when possible and use the first element in the list of genes. Only for the cases when we searched for pathways with a specific gene name, was the complete list of gene names considered.

In order to perform pathway analysis with proteoforms we first had to identify which physical entities correspond to input entity set, whether they are genes, proteoforms, or something else. Then we investigated which pathways have them as participants. We call *matching* the process of saying that an input molecule in the sample corresponds to a particular entity in the database. Matching proteins was conducted using accessioned numbers. Accession numbers also allow aggregating all protein object instances independently of the subcellular location where the reaction is happening. Matching gene participants is done by gene name and may introduce errors since the naming is not standardized. Names may differ between databases and publications, so it may be necessary check all possible names to match them or even go around the name and check its protein products with their accession numbers to identify with certainty genes in the reference database. The lack of standardized gene names may lead into wrong matching or mismatching. Annotations of gene participants in Reactome may include a name list that tries to cover the most common ones.

For the matching of proteoforms it is also necessary to consider the isoform number and post-translational modification associated with physical entity objects. To represent a specific isoform of a protein Reactome annotates a physical entity instance and connects it to a reference entity object of the UniProt database which has the isoform identifier as a property. It is very important to check the specific isoform of the protein because they can differ in length and amino acids at certain positions of the sequence, which may affect the correct identification of the location of modifications. Therefore, a different physical entity instance is used every time a specific proteoform

---

participates in a reaction to enable having different isoforms or combinations of modifications in each case.

When a protein requires a modification at a specific amino acid of its sequence to perform a role in a reaction, Reactome annotates this by adding a relationship to a *Translational Modification* object with the coordinate (position) of the residue. The coordinate is a positive integer representing the index of the amino acid where the modification is located. In some cases, the coordinate is not annotated, which can indicate that the reference did not specify it or that the values were ambiguous. When the coordinate is missing it is set as “?” or simply blank. It is important to consider that protein modification localization is challenging, so tolerances are required when matching proteoforms.

Modification instances are also connected in the database model to *mod* objects which contain more properties of the modifications such as its type. Modification types requires the use of standard names using controlled vocabularies, such as PSI-MOD<sup>112</sup>. Modification types can be very generic or very particular, notably regarding their function or target, *e.g.*, *Phosphorylation* is very generic, while *O-acetyl-L-serine* (modification that effectively converts an L-serine residue to O-acetyl-L-serine) is very specific. Modification definitions follow a hierarchy with nested types and subtypes that must be considered to decide if a modification in the input proteoform set corresponds to the annotated modifications in the proteoforms of the database.

When querying Reactome, we also included the possibility to use SNPs, which we had to also match to the physical entities of Reactome. We used the Ensembl Variant Effect Predictor<sup>9</sup> (VEP) to connect SNPs with most likely affected proteins and used their accessions to search for pathways.

## 3.2 Constructing the graph representation of a pathway

A network consists of a set of nodes linked by connections. This is mathematically represented as a *graph*, composed of *vertices* connected by *edges*. In network science notation the graph is called *network*, vertices are *nodes*, and edges are *connections* or

---

*links*. In biological terms, connections are *interactions* between *molecules* which are the nodes. A pathway in the data model of Reactome is composed by a set of events, most of them are reaction events. Each event has participants that are molecules. Molecules can be proteins, metabolites, complexes, sets among others. Participants perform different roles in reactions such as input (reactant), output (product), regulator, and catalyzer. To build a graph representation of a pathway, we iterate its reactions, extract all participants with the role they perform, and connect the entities participating in the same reaction.

The participants containing a sequence and having a standardized identifier are called *EntityWithAccessionedSequence*, e.g., genes formed by nucleotide bases or proteins formed by sequences of amino acids. Genes have names from the HUGO nomenclature and proteins have accessioned numbers from UniProt. Other types of smaller molecules which can be compounds or elements are identified by the ChEBI database, a dictionary of molecular entities focused on small chemical compounds.

Other types of participants are complexes and entity set. Complexes are larger molecules composed of other small molecules or accessioned entities which come together to collectively perform a role in the reaction. They commonly do not have an external accession number, but they are identified with internal stable identifiers within Reactome.

The participants of reactions hence become the nodes of the network. We take as nodes only basic molecules, such as accessioned entities or chemical compounds. In the process of construction of the network, the complexes are decomposed into their participants until obtaining simple molecules. The members of complexes are connected to each other to represent their collective involvement in the reaction.

### 3.3 Network topology analysis

We constructed interaction networks from pathways with an entity type as the node type. For the analysis in *Paper III*, we implemented analysis pipelines in Python 3.10.2, using Jupyter notebooks. We built interaction networks from Reactome version 80

---

using different types of molecules and yielding multiple network models. We compute multiple topological characteristics of the networks, to evaluate their topological properties. We conduct the same analysis locally, considering each pathway separately.

Given the hypothesis that molecules related to a specific disease tend to interact with each other in biochemical reactions, it is assumed that they are located in the same region of the complete interactome network. Since the properties regarding the number of neighbors each entity has in the interaction network is highly used in network analyses, we studied how much the degree varies between the gene and proteoform interactome. Also, we investigate the prevalence of bottleneck nodes and links such as articulation points and bridges. The topological properties computed and how they are derived are detailed in *Paper III*.

To evaluate the robustness of different network representations, we employed a percolation analysis, as introduced for the study of the incomplete interactome by Menche *et al.*<sup>46</sup>. Percolation analyses measure the robustness of a network or subnetwork by iteratively removing nodes or connections. In such a procedure, networks featuring few or poorly connected nodes will collapse faster than well populated and connected networks. One considers that such networks are not suited for systematic analyses like pathway analyses.

In *Paper III*, we perform the percolation analysis of networks built from Reactome using different representations of pathways. With a repeated random sampling, we get an approximation to the subnetwork sizes that can be observed if a random number of connections is removed.

### 3.4 Software implementation

For the analyses performed in *Paper I* and *Paper II*, we implemented a Java application, *PathwayMatcher* and an *Extractor* tool to obtain all proteoforms and interactions among the entities in Reactome. *PathwayMatcher* is an application that serves as a proof-of-concept showing that Reactome can be queried using proteoforms. *PathwayMatcher* can generate interaction networks from specific queries or of all

---

known interactions composing Reactome. The process to build networks from Reactome is detailed in more detail in *Paper I*. When PathwayMatcher is used to search for pathways using proteoforms a *proteoform matching* procedure is performed meaning that proteoforms in the input sample must be matched to proteoforms contained in the reference database. More details can be found in *Paper II*. To the best of our knowledge, there are no other public proteoform matching methods published, we thus established rules and guidelines for this procedure, especially with regards to partial and approximate matching.

Further details for software implementation can be found on the wiki pages for the public repositories established as part of this thesis work:

<https://github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki>

<https://github.com/PathwayAnalysisPlatform/ProteoformNetworks/wiki>

<https://github.com/PathwayAnalysisPlatform/Extractor>

All of the code is available in public repositories hosted in GitHub:

<https://github.com/PathwayAnalysisPlatform/PathwayMatcher>

<https://github.com/PathwayAnalysisPlatform/Networks>

<https://github.com/PathwayAnalysisPlatform/ProteoformNetworks>

[https://github.com/PathwayAnalysisPlatform/ProteoformNetworks\\_resources](https://github.com/PathwayAnalysisPlatform/ProteoformNetworks_resources)

<https://github.com/PathwayAnalysisPlatform/MappingFiles>

<https://github.com/PathwayAnalysisPlatform/Extractor>



---

## 4. Results

### 4.1 Paper I: Modelling pathways as a biological network

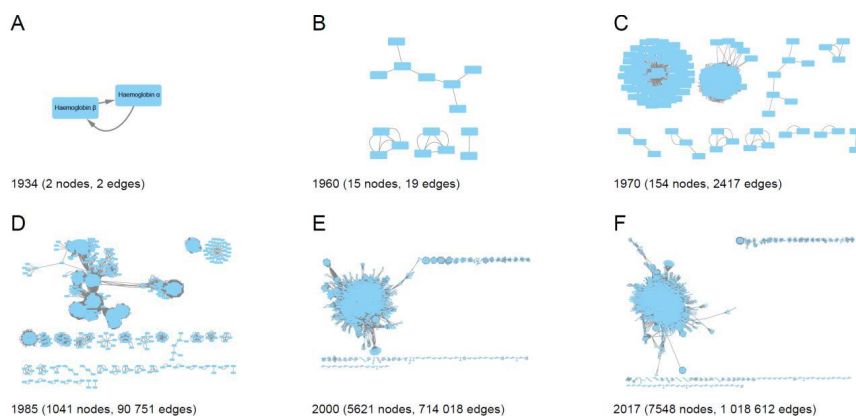
In this paper, the current knowledge on pathways was modelled as a network, showing how decades of biomedical research has shaped a complex representation of interconnected processes. The properties of pathway participants with respect to how many processes they are involved in, how they interact with each other, and how biases and missing information can influence the interpretation of proteomics data sets was explored. These insights shed light on the information underlying pathway analyses and provide critical information for researchers when interpreting the results of pathway analyses.

The current knowledge of Pathways can be modelled as a biological network where participants of the events composing the pathways are represented as nodes and their relationships are represented as connections. Representing pathways as biological networks enable their programmatic analysis. Pathway knowledge increased more rapidly as technologies improved (**Figure 5**). Protein networks are becoming larger and denser. This is due to the substantial increase of interactions among proteins rather than increase in proteins. There can be over a million interactions between only about ten thousand proteins. The result is a very densely connected network, with a slightly less connected periphery. The gain of additional data annotations comes at the price of increased complexity.

We found that the nested structure and hierarchy of pathways is important when analyzing omics data. Higher level pathways may be understandable, but often too generic to reach a conclusion. While more specific pathways are very specialized and understanding of the context where they are located is important. We also observed that parts of some pathways are used in multiple contexts, therefore it is necessary to consider how ubiquitous a subprocess is, to consider how meaningful results are.

The selection of proteins to be studied has a great influence on the results. If proteins are very well studied, it might be possible to get very detailed insights. While if proteins are not very well studied, or even not known to interact with any other protein at all, then they appear as isolated entities of the network with a very limited potential for functional analysis.

Another aspect of an entity is its ubiquity. Some proteins are very specific to certain processes, while others perform tasks related to a vast number of pathways. The selection of ubiquitous proteins as biomarkers may drastically reduce the potential to reach specific conclusions. Conversely, intersecting pathway and protein localization information might help identifying processes involved in diseases with shared etiology.



**Figure 5:** Protein network evolution. For each year, only proteins participating in a reaction published in or before that year according to Reactome are included. A) 1934: only one reaction is documented, B) 1960: four connected components containing 15 proteins in total, C) 1970: larger, very dense components start to appear, D) 1985: several very densely connected components, connected to each other in various degrees, E) and F) 2000 and 2017: a single very densely connected component, with slightly less well-connected periphery and a number of smaller components. See main text for further details.

---

## 4.2 Paper II: Matching of omics data to pathways

In *Paper I*, we underlined the lack of tools considering proteoforms when matching omics data to biological pathways. In *Paper II*, a novel bioinformatic tool called PathwayMatcher was implemented supporting the matching of omics data to the Reactome knowledgebase at three biological levels, namely for genes, proteins, or proteoforms. The results show how the use of detailed proteoform annotations improves the representation of biochemical reactions, how the connectivity of pathway participants is affected, and how the matching of omics data becomes more specific. By enabling the advanced matching of omics data to pathways, and connecting all pathway participants, PathwayMatcher provides the first proteoform-centric biological networks as an improved representation of biological pathways:

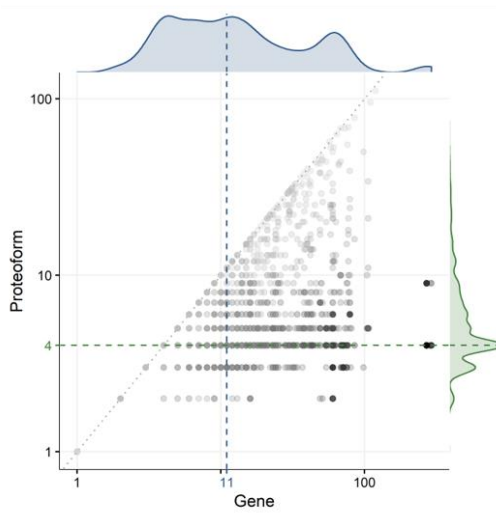
PathwayMatcher is the first software tool capable of selecting pathways related to a set of defined proteoforms as well as genes, proteins, or genetic variants. It proposes the first implementation of pathway analysis with proteoforms in the form of an overrepresentation analysis.

One of the main findings is that proteoforms participate in equal or fewer pathways and reactions than their protein or gene counterparts (**Figure 6**). This implies that proteoform searches are more specific when selecting candidate pathways for a set of sample entities. This is possible because the different isoform sequences and sets of post-translational modifications of proteoforms allow distinguishing between the instances where products of the same gene have different roles. As data becomes richer annotating more isoform and post-translational modification specific information, proteoform-based analysis will potentially be even more accurate and specific.

We designed and implemented procedures to decide if proteoforms in the input sample should be matched to proteoforms annotated in the database. We called this process proteoform matching. This is a challenging problem that requires careful consideration and knowledge of both pathway data model and sample acquisition. Reactome annotates the minimal set of required modifications for the protein to participate in a biochemical reaction. On the other hand, input proteoforms, might contain more, the

same, or less than the minimal set of annotations in Reactome. We defined different levels of stringency for the matching and allow the user to adjust to the level of stringency to the specificities of the analysis.

There was also a need to define a simple format to represent and read sets of proteoforms in a text file. For that, we created a simple representation standard for proteoforms that includes the isoform number and set of modifications represented with identifiers from the controlled vocabulary of the Proteomics Standard Initiative. There were no commonly adopted formats for proteoforms at the time. Alternative formats were published since then, such as PEF<sup>83</sup> and Proforma<sup>115</sup>, but it is worth noting that they are more complex than necessary for the application of pathway matching, the relevance of their implementation in PathwayMatcher is therefore unclear.



**Figure 6:** For all proteoform-specific participants, the number of pathways mapped using the proteoform versus gene is plotted in black. The density of the number of pathways mapped are indicated at the top (blue) and right (green) for gene and proteoform matching, respectively. The median number of pathways mapped is indicated with dashed lines.

---

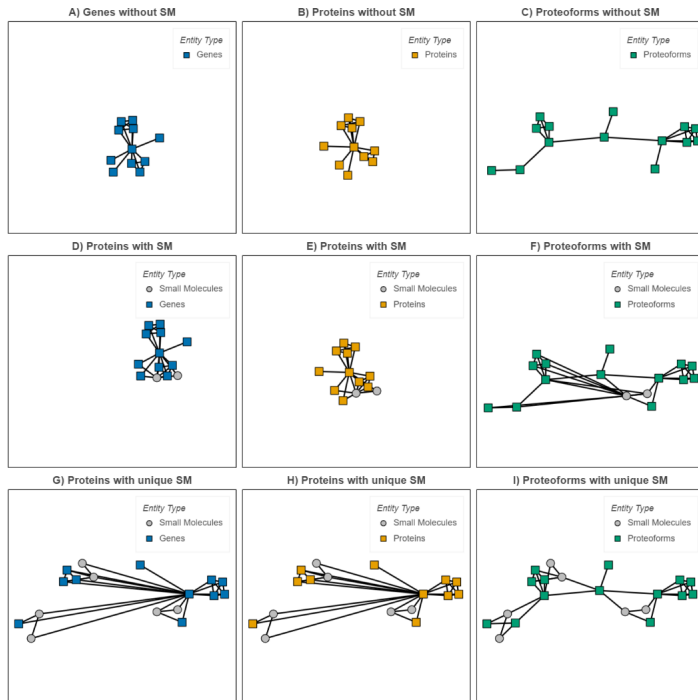
### 4.3 Paper III: Extending protein interaction networks using proteoforms and small molecules

In *Paper II*, we demonstrated that the modelling of pathways as biological networks can be changed by extending the elementary unit representing a gene or a protein using proteoform annotation. In *Paper III*, the consequences of extending the gene-centric paradigm to protein- and proteoform-centric networks was investigated. Furthermore, the implications of also including small molecules to such networks were explored. The results demonstrate how changing the representation of the networks induces changes in topology that can make certain nodes in the network essential for the interface between distinct biological processes, and thus potentially more clinically relevant.

We noticed remarkable differences in network topology when creating interactome networks for all pathways in Reactome, and when constructing interaction networks for single pathways one at a time (**Figure 7**). Single proteoforms may be better understood in an isolated pathway by looking at the connections with other entities. But interpreting biochemical interactions in these arbitrarily defined subnetworks overlook many connections that can be of importance in biological samples.

We propose including small molecules in certain types of studies when networks are sparse or lightly connected. Small molecules are not considered as the driving molecules achieving biological processes, but they are necessary in many reactions, and thus need to be considered as part of the connecting entities. They even help alleviate the disconnectedness, showing which possible inputs can be needed for the next reaction.

We quantified the changes in connectivity of the networks when changing gene- or protein-centric networks into proteoform centric networks, with different representations of small molecules. The network structure was generally enriched, but in some cases proteoform annotation is so generic that all proteoforms resulting from a single gene are connected to each other, therefore creating an unnecessary complexity, and artificially increasing the degree of the resulting nodes.



**Figure 7:** Interaction networks for Pathway "Receptor Mediated Mitophagy" (R-HSA-8934903) represented as the level of gene (left), protein (center), and proteoform (right), when not including small molecules (top), including them (middle), and including them but restricting their connections to single reactions (bottom).

---

## 5. Discussion

Proteoforms display strong potential to provide better understanding of molecular mechanisms given their wide variety and corresponding specificity. The most popular bottom-up proteomics methods, based on the digestion of protein molecules by proteolytic enzymes, suffer from the *protein inference problem*<sup>155,156</sup>. The issue is that identification of proteins with repetitive patterns cannot be done with certainty by assembling the identified peptide sequences. The problem becomes more acute for the identification of proteoforms with specific combinations of PTMs at specific locations<sup>11</sup>. Many molecular details are lost if the inference of the right location and combination of modifications cannot be done. Molecular details also very difficult to elucidate when sequences have crosstalk points<sup>34</sup>.

There are attempts to increase the certainty of the candidate proteins suggested by the assembly of peptides found by bottom-up proteomics. One example suggests using multiple proteases<sup>157</sup> to increase sequence coverage or get different combinations of peptides<sup>158</sup>, they expand the set of digestion protease enzymes to complement trypsin with LysC, ArgC, AspN, or GluC. Without the use of multiple enzymes, regions of the proteome are inaccessible for identification procedures. Other approaches propose the use of middle-down proteomics proteases which can yield on average larger peptides (> 6.3 kDa), such as the outer membrane protease T (OmpT) enabling the identification of isoforms and their respective post-translational modifications<sup>159</sup>.

When we talk about proteoforms in a pathway database we may not necessarily be talking about intact molecules identified with a top-down proteomics approach, but by more prevalent methods of bottom-up proteomics. One may question how reliably we are inferring the right proteoform with a specific combination of PTMs when the protein inference problem is one pitfall of bottom-up approaches. To answer that we must remember that inference of proteins from peptides is not a problem affecting all proteins and it is possible and certain to infer certain proteins. It gets more complicated for molecular structures which have a lot of repetition or complex formations using cross links dependent on cleavage of peptides and assembly from multiple components.

---

There are proteins with *unique* peptides<sup>160</sup> while other proteins result from the assembly of *shared* peptides, resulting in ambiguous protein of origin inferences.

Proteoform inference further suffers from the difficulty of confidently asserting the isoform, PTMs, and their location. It is a gradient of certainty which goes from very certain to impossible to be sure. Once a researcher proves experimentally that the protein in question has certain identity and a set of modifications, that annotation is curated into the pathway database independently of the proteomics approach used. It is not restricted to certain techniques. Naturally, the more complex the molecule, the more accurate top-down techniques have potential, especially if technology developments allow its intact identification without previous fractionation as part of the sample preparation. Therefore, the minimal requirements of a molecule annotated to participate in a pathway, including the set of modifications, can help improve pathway analyses, but the current state of our knowledge does not provide comprehensive information on the proteoform state of pathway participants. An alternative when mapping partial omics data to pathway knowledgebases and improve pathway inference, would be to map the sequences directly to the knowledgebase.

Current interactome and pathway knowledge is incomplete but sufficient to be used for functional analysis. There are efforts to study this incompleteness of the interactome<sup>46</sup>. Percolation analysis can be used to evaluate the robustness of the connections in the largest connected component of the interactome. This analysis translates into how robust the network is to the lack of knowledge of some nodes. If the network stays connected even without many nodes or links removed, then it means that we already have enough information about sufficient number of nodes to perform functional analysis on it, given that the network stays connected in the same way even with changes. We can build the interactome using genes, proteins, or proteoforms as nodes of the networks. As we have shown they show different structural properties. Percolation analysis can be done to interactomes of each type of entity and see which network is more vulnerable to the lack of knowledge. Given the distinct nature of the type of entities and the meaning of the links, different robustness is expected.



The definition of the interactome itself plays a key role in functional network analysis. In this work we considered a reference database which had multiple types of biomolecules – gene-derived but also small molecules. The protein annotation covered genes, proteins, isoform, and PTMs. Most commonly interaction networks focus on a single type of entities, typically proteins, *e.g.*, phosphorylation mechanisms in signaling networks<sup>13</sup> or protein-protein interactions using gene names or generic protein accession identifiers<sup>161</sup>. We advocate for the integration of multiple data sources to keep constructing rich interaction networks which include small molecules and proteoform knowledge but recommend that this integration is not done at the cost of information like PTMs, which are essential to the understanding of biological systems. Even though it is already challenging to integrate multiple sources of protein-protein interaction networks for the construction of a unified interactome<sup>162</sup>, the effort is necessary to provide better overview of molecular mechanisms, compensate topic biases of databases, and deliver unbiased and rich systematic network functional analysis.

There are already multiple interactome maps available, which try to systematically build the network based on experimental interactions<sup>163</sup>. Examples of available interactomes can be obtained from BioGrid, Intact, or from the Human Reference Protein Interactome Mapping Project<sup>164</sup>. They set great examples towards portraying the complete interactome, but much work remains until the complete map of the human biological system is defined<sup>165</sup>, for which extending the proteoform interactome is essential<sup>160</sup>.

---

## 6. Conclusion

In this thesis we first investigated the inside of biological pathway databases, one of the richest sources of recorded biological processes. We shed light on the consequences of how knowledge was acquired and annotated (curation) and organized (data model). With this analysis, we identified strengths and limitations that must be considered when performing pathway analyses, and help researchers better take advantage of the rich knowledge aggregated by decades of scientific research.

The thesis has also generated insights into what kind of molecular information is recorded in the Reactome pathway database, and based on this we propose the adoption of analysis methods that go beyond gene names and take advantage of the rich knowledge available. Proteoforms contain more biological details that enable specific matching of results to pathways and help conducting finer analyses. Proteoforms have the potential to narrow the number of hypothetically affected biological processes because they require the physical entity to have a specific set of modifications and sequence variations which the gene name or protein accession number cannot provide. Overall, there might be circumstances where certain types of molecules such as genes can be used for functional analysis and other circumstances where molecules such as proteoforms provide better insights at the cost of more data complexity.

We provide a proof-of-concept implementation of a mapping tool which takes as input sample data and searches reference pathways. Among the sample data we included proteoforms, showing that such searches are possible and that results can be better if the biological database contains these kinds of annotations, but also revealing challenges in the matching of proteoforms. We therefore also developed methods for matching sample entities with reference entities in the pathway database by comparing the sets of PTMs and isoforms of proteins. Entity mapping can be done with a strict or flexible comparison. Stringency increases the confidence that we refer to the same entities, while flexibility increases the number of matched entities and, thereby, the number of selected pathways matched to the sample. Matching flexibility increases sensibility, while stringency increases specificity. We show that for different types of

---

omics data, different levels of stringency are helpful, and it is up to the researcher to define the level of analysis best suited to the data analyzed.

We furthermore analyzed the topological properties of the interactome and other subnetworks with alternative network construction methods. The use of genes, proteins, and proteoforms as basic nodes of the network was considered and the consequences of including small molecules explored. The results clearly show that the topological characteristics of the networks are affected by the type of molecules modeled by the nodes, especially compared to the traditional approach of gene-centric interaction networks. These differences open the possibility to choose the most adequate type of network for different types of analysis. For example, when there are too many connections among nodes, proteoforms can help separate the connections across more nodes focused on more specific biological processes. Another situation is when the studied network is too sparse, having many nodes disconnected from each other, including small molecules can then help restore the connectivity of the network and see which molecules that are working with each other even if indirectly.

Proteome molecules describe in detail the participants of biological processes. Nevertheless, their identification is still a challenge, both because of the great variety and the very small difference between them that makes it very complicated to distinguish close yet different forms. Proteoforms can potentially have a combinatorial explosion of possibilities. There is currently a bloom in representation formats, and mapping, as well as analysis and interpretation methods. All can be improved to include proteoforms to improve our understanding of biological systems. Proteome annotations will continue to grow, making their impact in functional analysis more significant, and allowing the design of novel and finer methods for pathway analysis. Such methods combined with data from different omics technologies will allow reaching conclusions with more certainty and accuracy such that diseases can be diagnosed, prognosed, and treated in a better way.

---

## 7. Outlook

Proteomics technologies are improving considerably, resulting in the identification of more and more proteoforms<sup>81,166</sup>. This increase in proteoform data seems promising to build more specific interaction networks, where there are more connections among nodes that interact only when they have the right conformation or set of post-translational modifications. Currently, redundant connections are added to the network because experiments or databases do not distinguish the proteoform participating in a specific reaction part of a pathway, leading to generic proteins highly connected where specific proteoforms would distribute the connections. Future research will tell how much of the pathway knowledge can be refined to the proteoform level, but our results and recent research suggest that proteoforms will enable more specific annotations<sup>167</sup>.

Proteins are essential to understand the connection between genotype and phenotype, and proteoforms are the next step in elucidating the molecular details of the proteome. Not only identification of proteoforms is necessary, but also functional annotations must be included to depict functional associations between molecules and diseases. We heavily depend on databases being updated to contain proteoform data. In addition to new interactions, we need to improve the current annotation of reaction participants to include proteoform information. There are multiple efforts to annotate more proteoforms<sup>34</sup> along with their functions like the Proteoform Atlas<sup>34</sup> and the Human Proteoform Project<sup>160</sup> which has the ambitious goal to draft a reference set of all proteoforms produced by the human genome. Another example is the neXtProt knowledgebase<sup>168</sup> which is expanding the annotation of proteoforms, and is one of the early adopters of the proteoform format PEF<sup>83</sup>, including their cellular localization, protein function, and expression by tissue.

To expand the functional knowledge of molecules, there are approaches based on the known function of previously studied molecules. First case is when selecting more functional molecules by the guilty-by-association process, interacting partners are more specific, leading to a smaller more specific selection of candidate processes where those neighbors participate. On a second case, when comparing multiple disease

---

modules, it may happen that we assume that a set of entities shares biological function because they share some of the entities or connect to molecules in the other module, nevertheless, when adjusting the disease module networks to proteoforms, the nodes might not overlap anymore. As part of our study, we checked whether this happened using a set of disease modules and with the proteoform information in Reactome, but it was not clear that this happened already due to the lack of more specific proteoform annotation as well as lacking proteoform-disease annotations. Since no database associates sets of proteoforms to diseases, instead of gene or protein sets, we had to infer which proteoforms are equivalent to the set of genes associated to a disease through converting the identifiers and then discarding the proteoforms without a direct connection in the interaction. This leads to proteoform disease modules very similar in structure to the gene networks modules, but that potentially could differ if we knew the concrete proteoforms which really are associated to the disease. Therefore, better functional annotation of proteoforms could alleviate this type of network comparison, mitigating the proteoform network construction issue by using proteoform participants directly.

Regarding functional analysis of biological entities for understanding disease mechanisms, one can build gene network modules, *i.e.*, subnetworks of the interactome made of entities known to be associated with a particular disease. A common source of the associations is results from GWAS. These modules allow the transfer of disease mechanisms knowledge from one disease to another, by mapping the function of the participants. When there are diseases sharing entity nodes, there might be a functional overlap. Larger versions of the interactomes<sup>163</sup> become available as more molecular entities are identified, especially small molecules and proteoforms, and their functions are discovered<sup>166</sup>. An interesting challenge will be to adapt the functional analysis methods to include this new type of hybrid network, as we pioneered in the work of this thesis.

---

## 8. Future work

### 8.1 Automatic annotation of pathways

Interactome networks try to model the complex cellular processes where many types of molecules participate as nodes<sup>162</sup>. Proteins play an essential role given their large diversity and capacity to mediate biochemical reactions. Knowledge has grown to proteome level in recent years to include annotations of more than half of the proteins experimentally proven in UniProtKb<sup>99</sup> for human. Nevertheless, our current knowledge of the interactome is incomplete, it is not possible to know how many interactions happen in reality as part of cellular processes<sup>46</sup>. Databases annotate interactions for only a part of existing proteins. Databases may only include experimentally proven or predicted interactions<sup>118</sup>. Functional network analysis studies then should keep in consideration that the interactome is an incomplete model of reality that keeps growing as more molecules and interactions are discovered. One problem that may arise is that a study wants to discover the function of a set of molecules, *e.g.*, a protein set where some of them are not yet in the interactome.

One way to tackle the problem of missing molecules is to aggregate multiple independent reference databases<sup>118,163</sup>. Databases use different curation methods and focus on different levels of detail, therefore some databases include many more proteins and interactions<sup>163</sup>. For example, a pathway database might grow slower because it tries to include many details of participants using a manual curation process, while an interaction database might include many more interacting molecules using programmatic automation tools which ignore the functional context and only focus on an unsupervised data aggregation process. For example, on one hand we can consider Reactome<sup>59</sup> pathway database as a source of implicit interactions as part of reactions and on the other hand we can consider IntAct<sup>118</sup>, a database specialized on molecular interaction data curated from literature and direct user submissions. IntAct includes more than 1.18 million interactions between more than 118,000 molecules. This can mitigate the problem of missing proteins in Reactome. Although in practice, aggregating multiple database sources bears also the risk of combining data with

---

different levels of certainty, creating duplicate records of the same interaction, or adding errors by describing interactions in contradicting terminology leading to misunderstandings.

Even when database integration problems are solved, it is necessary to place proteins and their interactions in their functional context; not only list the set of isolated interactions. One way to make use of the functional knowledge of currently annotated molecules in the interactome is to connect newer proteins to the rest of the network by defining paths with the set of interactions aggregated from multiple databases. These paths would be composed of interactions with different levels of certainty, and they may have an arbitrary length. It would be necessary to devise algorithms for constructing these paths, with the most basic case being a direct connection from a protein in the interactome to a disconnected protein, *e.g.*, a path of length 1<sup>169</sup>. For proteins which cannot be connected directly there could be many possible paths, even with cycles and redundant ways to reach them. Furthermore, we highlight that adding small molecules can greatly alleviate the problem of disconnected entities. Given the complexity and number of interactions this process may not be performed manually. There is also no clear approach to define the new paths because they would be, in other terms, extensions of pathways or mechanisms to automatically annotate pathways. These methods would then be defining what makes a pathway a pathway.

Automatic extension of pathways requires certain level of consensus among researchers, but it is not completely clear what characteristics are necessary to promote a sequence of interactions into a pathway. There are modern machine learning methods that attempt to discover patterns in sequential data<sup>170</sup>, they can later be used as a reference to predict sequences that conform to the learned pattern<sup>171</sup>. Recurrent neural networks are one of them<sup>172-174</sup>. Given a sequence of elements, we can try to predict what element will come next. Other machine learning methods receive a set of elements and decide what label fits to them. These methods can potentially decide which elements come next if we try to extend a pathway or certain unclassified regions of the interactome. Another possibility is that we arbitrarily construct paths from interactome proteins to disconnected proteins and then decide functional labels for each alternative

path built. Neural networks involving sequence patterns are promising candidates for such tasks.

The advantage of these methods would be the faster automatic extension of pathways and construction of candidate new regions of the interactome which provide functional context for more recent proteins not yet manually annotated in databases. These methods require wide characterization of the proteins and their interactions. Their properties would help train the models and the more information can be condensed into the sequences the more elements the neural networks have to predict the new paths. Nevertheless, it remains challenging to decide which properties of the interactions and the molecules themselves can be useful to decide if a node will connect to others. It is necessary to make assumptions that characteristics inherent to the proteins can tell which are their binding partners before and after in the sequence. It is also needed to predict only from the interaction partners available in the same subcellular location and the correct abundance, which is often not available.

Another challenge is providing certainty of how likely it is that a series of interactions could become a pathway. Even deciding what kind of score could project that certainty. In any case, a clear benefit would be an extended interactome which includes more of the interest entities and allow the identification of closest or most related pathways, by pointing out specific regions of the interactome.

## 8.2 Multidimensional functional analysis

Pathways consist of a sequence of reactions. Each reaction may take different time to happen. A first reaction could be the activation of a receptor, which happens instantaneously after binding to the ligand; followed by the receptor leading to the activation of an intracellular signaling cascade. The downstream event can remain active for some minutes such that it achieves cellular response. For example, they may induce or repress the expression of certain genes for a given time until the abundance of the protein desired products is adequately regulated.



There are rare situations where pathways are annotated with time or abundance information which is often necessary to understand biological processes. In some databases reactions have annotated the stoichiometry coefficients, like in Reactome, but such properties are rarely found to a level that allows including them pathway analyses. Functional analysis including these other dimensions and many more could lead to more accurate and sensible results when selecting biological processes. On the other hand, it requires much more extensive databases and experiments, and more complex computational models. With the massive increase in computing capacity and data interpretation, it may be possible to start tackling these complex problems that seemed like science fiction only a few decades ago.

---

## 9. References

- 1 Auffray, C., Chen, Z. & Hood, L. Systems medicine: the future of medical genomics and healthcare. *Genome Medicine* **1**, 2, doi:10.1186/gm2 (2009).
- 2 Macosko, Evan Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).
- 3 Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Annals of Human Biology* **40**, 463-471, doi:10.3109/03014460.2013.807878 (2013).
- 4 Muraro, Mauro J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems* **3**, 385-394.e383, doi:<https://doi.org/10.1016/j.cels.2016.09.002> (2016).
- 5 Jinek, M., Chylinski, K. & Fonfara, I. Alberts, B., Johnson, A., Lewis, J. *et al.* (2014). *Molecular Biology of the Cell*, 6e. New York: Garland Science. *An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications* **27**, 1043-1149 (2020).
- 6 Abbott, S. & Fairbanks, D. J. Experiments on Plant Hybrids by Gregor Mendel. *Genetics* **204**, 407-422, doi:10.1534/genetics.116.195198 (2016).
- 7 Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics* **21**, 630-644, doi:10.1038/s41576-020-0258-4 (2020).
- 8 Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580-585, doi:10.1038/ng.2653 (2013).
- 9 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).
- 10 Bentley, D. R. The Human Genome Project—An Overview. *Medicinal Research Reviews* **20**, 189-196, doi:[https://doi.org/10.1002/\(SICI\)1098-1128\(200005\)20:3<189::AID-MED2>3.0.CO;2-#](https://doi.org/10.1002/(SICI)1098-1128(200005)20:3<189::AID-MED2>3.0.CO;2-#) (2000).
- 11 Aebersold, R. *et al.* How many human proteoforms are there? *Nature Chemical Biology* **14**, 206-214, doi:10.1038/nchembio.2576 (2018).
- 12 Kluger, R. & Alagic, A. Chemical cross-linking and protein-protein interactions—a review with illustrative protocols. *Bioorganic Chemistry* **32**, 451-472, doi:<https://doi.org/10.1016/j.bioorg.2004.08.002> (2004).
- 13 Robles, M. S., Humphrey, S. J. & Mann, M. Phosphorylation Is a Central Mechanism for Circadian Control of Metabolism and Physiology. *Cell Metabolism* **25**, 118-127, doi:<https://doi.org/10.1016/j.cmet.2016.10.004> (2017).
- 14 Li, Y., Tran, A. H., Danishefsky, S. J. & Tan, Z. in *Methods in Enzymology* Vol. 621 (ed Arun K. Shukla) 213-229 (Academic Press, 2019).
- 15 Smith, L. M. & Kelleher, N. L. Proteoforms as the next proteomics currency. *Science* **359**, 1106-1107, doi:10.1126/science.aat1884 (2018).
- 16 Smith, L. M. *et al.* Proteoform: a single term describing protein complexity. *Nature Methods* **10**, 186-187, doi:10.1038/nmeth.2369 (2013).

- 
- 17 Tokarz, V. L., MacDonald, P. E. & Klip, A. The cell biology of systemic insulin function. *J Cell Biol* **217**, 2273-2289, doi:10.1083/jcb.201802095 (2018).
- 18 Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355, doi:10.1038/nature19949 (2016).
- 19 Ponomarenko, E. A. *et al.* The Size of the Human Proteome: The Width and Depth. *International Journal of Analytical Chemistry* **2016**, 7436849, doi:10.1155/2016/7436849 (2016).
- 20 Adhikari, S. *et al.* A high-stringency blueprint of the human proteome. *Nature Communications* **11**, 5301, doi:10.1038/s41467-020-19045-9 (2020).
- 21 Mouritsen, O. G. & Zuckermann, M. J. What's so special about cholesterol? *Lipids* **39**, 1101-1113, doi:<https://doi.org/10.1007/s11745-004-1336-x> (2004).
- 22 Luo, J., Yang, H. & Song, B.-L. Mechanisms and regulation of cholesterol homeostasis. *Nature Reviews Molecular Cell Biology* **21**, 225-245, doi:10.1038/s41580-019-0190-7 (2020).
- 23 Diemel, G. A. Brain Glucose Metabolism: Integration of Energetics with Function. *Physiological Reviews* **99**, 949-1045, doi:10.1152/physrev.00062.2017 (2019).
- 24 Morel, J.-D. *et al.* The mouse metallomic landscape of aging and metabolism. *Nature Communications* **13**, 607, doi:10.1038/s41467-022-28060-x (2022).
- 25 Doerr, A. Global metabolomics. *Nature Methods* **14**, 32-32, doi:10.1038/nmeth.4112 (2017).
- 26 Beger, R. D. A Review of Applications of Metabolomics in Cancer. *Metabolites* **3**, 552-574 (2013).
- 27 Han, X. Lipidomics for studying metabolism. *Nature Reviews Endocrinology* **12**, 668-679, doi:10.1038/nrendo.2016.98 (2016).
- 28 Pothukuchi, P. *et al.* Translation of genome to glycome: role of the Golgi apparatus. *FEBS Letters* **593**, 2390-2411, doi:<https://doi.org/10.1002/1873-3468.13541> (2019).
- 29 Rillahan, C. D. & Paulson, J. C. Glycan Microarrays for Decoding the Glycome. *Annual Review of Biochemistry* **80**, 797-823, doi:10.1146/annurev-biochem-061809-152236 (2011).
- 30 Jonsson, H. *et al.* Differences between germline genomes of monozygotic twins. *Nature Genetics* **53**, 27-34, doi:10.1038/s41588-020-00755-1 (2021).
- 31 Corinne Tarantino, M. *Chromosomal Aberrations What Are They, Causes, and More*, <<https://www.osmosis.org/answers/chromosomal-aberrations>> (2022).
- 32 Holliday, R. Epigenetics: A Historical Overview. *Epigenetics* **1**, 76-80, doi:10.4161/epi.1.2.2762 (2006).
- 33 Feinberg, A. P. The Key Role of Epigenetics in Human Disease Prevention and Mitigation. *New England Journal of Medicine* **378**, 1323-1334, doi:10.1056/NEJMr1402513 (2018).
- 34 Melani, R. D. *et al.* The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. *Science* **375**, 411-418, doi:10.1126/science.aaz5284 (2022).
- 35 Jiang, L. *et al.* A Quantitative Proteome Map of the Human Body. *Cell* **183**, 269-283.e219, doi:<https://doi.org/10.1016/j.cell.2020.08.036> (2020).

- 
- 36 Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular systems biology* **15**, e8503-e8503, doi:10.15252/msb.20188503 (2019).
- 37 Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics* **13**, 227-232, doi:10.1038/nrg3185 (2012).
- 38 López De Maturana, E. *et al.* Challenges in the Integration of Omics and Non-Omics Data. *Genes* **10**, 238, doi:10.3390/genes10030238 (2019).
- 39 Jones, A. G. & Hattersley, A. T. The clinical utility of C-peptide measurement in the care of patients with diabetes. *Diabetic Medicine* **30**, 803-817, doi:<https://doi.org/10.1111/dme.12159> (2013).
- 40 Röder, P. V., Wu, B., Liu, Y. & Han, W. Pancreatic regulation of glucose homeostasis. *Experimental & molecular medicine* **48**, e219-e219, doi:10.1038/emm.2016.6 (2016).
- 41 Flannick, J., Johansson, S. & Njølstad, P. R. Common and rare forms of diabetes mellitus: towards a continuum of diabetes subtypes. *Nature Reviews Endocrinology* **12**, 394-406, doi:10.1038/nrendo.2016.50 (2016).
- 42 Bersanelli, M. *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* **17**, doi:10.1186/s12859-015-0857-9 (2016).
- 43 Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nature Biotechnology* **38**, 365-373, doi:10.1038/s41587-019-0344-3 (2020).
- 44 Chang, M., Dayhoff, M., Eck, R. & Sochard, M. Atlas of protein sequence and structure. (1965).
- 45 Pertea, M. *et al.* CHES: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology* **19**, 208, doi:10.1186/s13059-018-1590-2 (2018).
- 46 Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601, doi:10.1126/science.1257601 (2015).
- 47 Ouellette, R. J. & Rawn, J. D. in *Organic Chemistry (Second Edition)* (eds Robert J. Ouellette & J. David Rawn) 929-971 (Academic Press, 2018).
- 48 Siddiqui, K. & Stillman, B. ATP-dependent Assembly of the Human Origin Recognition Complex \*<sup>2666</sup>. *Journal of Biological Chemistry* **282**, 32370-32383, doi:10.1074/jbc.M705905200 (2007).
- 49 Ge, X. & Wang, X. Role of Wnt canonical pathway in hematological malignancies. *Journal of Hematology & Oncology* **3**, 33, doi:10.1186/1756-8722-3-33 (2010).
- 50 Yang, K. *et al.* The evolving roles of canonical WNT signaling in stem cells and tumorigenesis: implications in targeted cancer therapies. *Laboratory Investigation* **96**, 116-136, doi:10.1038/labinvest.2015.144 (2016).
- 51 Chandel, N. S. Glycolysis. *Cold Spring Harbor Perspectives in Biology* **13**, a040535 (2021).
- 52 Lim, W., Mayer, B. & Pawson, T. *Cell signaling*. (Taylor & Francis, 2014).

- 
- 53 Zoncu, R., Efeyan, A. & Sabatini, D. M. mTOR: from growth signal integration to cancer, diabetes and ageing. *Nature Reviews Molecular Cell Biology* **12**, 21-35, doi:10.1038/nrm3025 (2011).
- 54 Tabery, J., Monika Piotrowska, and Lindley Darden. in *The Stanford Encyclopedia of Philosophy* (ed Edward N. Zalta) (Metaphysics Research Lab, Stanford University, 2021).
- 55 Mamcarz, E. *et al.* Lentiviral Gene Therapy Combined with Low-Dose Busulfan in Infants with SCID-X1. *N Engl J Med* **380**, 1525-1534, doi:10.1056/nejmoa1815408 (2019).
- 56 Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nature Reviews Genetics* **19**, 299-310, doi:10.1038/nrg.2018.4 (2018).
- 57 Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology* **18**, 83, doi:10.1186/s13059-017-1215-1 (2017).
- 58 Slenter, D. N. *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research* **46**, D661-D667, doi:10.1093/nar/gkx1064 (2017).
- 59 Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Research* **50**, D687-D692, doi:10.1093/nar/gkab1028 (2021).
- 60 Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333-351, doi:10.1038/nrg.2016.49 (2016).
- 61 Aronson, S. J. & Rehm, H. L. Building the foundation for genomics in precision medicine. *Nature* **526**, 336-342, doi:10.1038/nature15816 (2015).
- 62 Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).
- 63 Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research* **27**, e1608, doi:<https://doi.org/10.1002/mpr.1608> (2018).
- 64 Dina, C. *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. *Nature Genetics* **39**, 724-726, doi:10.1038/ng2048 (2007).
- 65 Smajlagić, D. *et al.* Population prevalence and inheritance pattern of recurrent CNVs associated with neurodevelopmental disorders in 12,252 newborns and their parents. *European Journal of Human Genetics* **29**, 205-215, doi:10.1038/s41431-020-00707-7 (2021).
- 66 Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology* **30**, 1095-1106, doi:10.1038/nbt.2422 (2012).
- 67 Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789-796, doi:10.1038/nature02168 (2003).
- 68 Rood, J. E. & Regev, A. The legacy of the Human Genome Project. *Science* **373**, 1442-1443, doi:10.1126/science.ab15403 (2021).
- 69 McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356-369, doi:10.1038/nrg2344 (2008).

- 
- 70 Fauman, E. B. & Hyde, C. An optimal variant to gene distance window derived from an empirical definition of cis and trans protein QTLs. *BMC Bioinformatics* **23**, 169, doi:10.1186/s12859-022-04706-x (2022).
- 71 Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Research* **47**, e3-e3, doi:10.1093/nar/gky837 (2018).
- 72 Burgess, D. J. Spatial transcriptomics coming of age. *Nature Reviews Genetics* **20**, 317-317, doi:10.1038/s41576-019-0129-z (2019).
- 73 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63, doi:10.1038/nrg2484 (2009).
- 74 Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211-220, doi:10.1038/s41586-021-03634-9 (2021).
- 75 Correa Rojo, A. *et al.* Towards Building a Quantitative Proteomics Toolbox in Precision Medicine: A Mini-Review. *Frontiers in Physiology* **12**, doi:10.3389/fphys.2021.723510 (2021).
- 76 Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541, doi:doi:10.1126/science.abj1541 (2021).
- 77 Baker, M. Reproducibility crisis: Blame it on the antibodies. *Nature* **521**, 274-276, doi:10.1038/521274a (2015).
- 78 Joshi, A. & Mayr, M. In Aptamers They Trust. *Circulation* **138**, 2482-2485, doi:doi:10.1161/CIRCULATIONAHA.118.036823 (2018).
- 79 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207, doi:10.1038/nature01511 (2003).
- 80 The, M. & Käll, L. Focus on the spectra that matter by clustering of quantification data in shotgun proteomics. *Nature Communications* **11**, 3234, doi:10.1038/s41467-020-17037-3 (2020).
- 81 Schaffer, L. V. *et al.* Identification and Quantification of Proteoforms by Mass Spectrometry. *PROTEOMICS* **19**, 1800361, doi:10.1002/pmic.201800361 (2019).
- 82 Fornelli, L. *et al.* Top-down proteomics: Where we are, where we are going? *J Proteomics* **175**, 3-4, doi:10.1016/j.jprot.2017.02.002 (2018).
- 83 Binz, P.-A. *et al.* Proteomics Standards Initiative Extended FASTA Format. *Journal of Proteome Research* **18**, 2686-2692, doi:10.1021/acs.jproteome.9b00064 (2019).
- 84 Alseekh, S. *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nature Methods* **18**, 747-756, doi:10.1038/s41592-021-01197-1 (2021).
- 85 Würtz, P. *et al.* Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies. *American Journal of Epidemiology* **186**, 1084-1096, doi:10.1093/aje/kwx016 (2017).
- 86 Soininen, P., Kangas, A. J., Würtz, P., Suna, T. & Ala-Korpela, M. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in

- Cardiovascular Epidemiology and Genetics. *Circulation: Cardiovascular Genetics* **8**, 192-206, doi:doi:10.1161/CIRCGENETICS.114.000216 (2015).
- 87 Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330, doi:doi:10.1126/science.aaz1776 (2020).
- 88 Fang, S., Holmes, M. V., Gaunt, T. R., Smith, G. D. & Richardson, T. G. An atlas of associations between polygenic risk scores from across the human phenome and circulating metabolic biomarkers. *medRxiv*, 2021.2010.2014.21265005, doi:10.1101/2021.10.14.21265005 (2021).
- 89 Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics* **53**, 1712-1721, doi:10.1038/s41588-021-00978-w (2021).
- 90 Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55-62, doi:10.1038/nature18003 (2016).
- 91 Mani, D. R. *et al.* Cancer proteogenomics: current impact and future prospects. *Nature Reviews Cancer*, doi:10.1038/s41568-022-00446-5 (2022).
- 92 Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nature Metabolism* **2**, 1135-1148, doi:10.1038/s42255-020-00287-2 (2020).
- 93 Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387, doi:10.1038/nature13438 (2014).
- 94 Johansson, H. J. *et al.* Breast cancer quantitative proteome and proteogenomic landscape. *Nature Communications* **10**, 1600, doi:10.1038/s41467-019-09018-y (2019).
- 95 Lage, K. Protein–protein interactions and genetic diseases: The interactome. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1842**, 1971-1980, doi:<https://doi.org/10.1016/j.bbadis.2014.05.028> (2014).
- 96 Fields, S. & Song, O.-k. A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245-246, doi:10.1038/340245a0 (1989).
- 97 Braun, P. Interactome mapping for analysis of complex phenotypes: Insights from benchmarking binary interaction assays. *PROTEOMICS* **12**, 1499-1518, doi:<https://doi.org/10.1002/pmic.201100598> (2012).
- 98 Gavin, A.-C., Maeda, K. & Kühner, S. Recent advances in charting protein–protein interaction: mass spectrometry-based approaches. *Current Opinion in Biotechnology* **22**, 42-49, doi:<https://doi.org/10.1016/j.copbio.2010.09.007> (2011).
- 99 Consortium, T. U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480-D489, doi:10.1093/nar/gkaa1100 (2020).
- 100 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733-D745, doi:10.1093/nar/gkv1189 (2015).
- 101 Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Research* **45**, D635-D642, doi:10.1093/nar/gkw1104 (2016).
- 102 Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* **30**, 187-200, doi:10.1002/pro.3978 (2021).

- 
- 103 Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research* **47**, D1038-D1043, doi:10.1093/nar/gky1151 (2018).
- 104 Rolland, T. *et al.* A Proteome-Scale Map of the Human Interactome Network. *Cell* **159**, 1212-1226, doi:10.1016/j.cell.2014.10.050 (2014).
- 105 Grissa, D., Junge, A., Oprea, T. I. & Jensen, L. J. Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration. *Database* **2022**, doi:10.1093/database/baac019 (2022).
- 106 Volders, P.-J. *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Research* **41**, D246-D251, doi:10.1093/nar/gks915 (2012).
- 107 Bruford, E. A. *et al.* Guidelines for human gene nomenclature. *Nature Genetics* **52**, 754-758, doi:10.1038/s41588-020-0669-3 (2020).
- 108 Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310-315, doi:10.1038/s41586-022-04558-8 (2022).
- 109 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308-311, doi:10.1093/nar/29.1.308 (2001).
- 110 Farriol-Mathis, N. *et al.* Annotation of post-translational modifications in the Swiss-Prot knowledge base. *PROTEOMICS* **4**, 1537-1550, doi:<https://doi.org/10.1002/pmic.200300764> (2004).
- 111 Creasy, D. M. & Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *PROTEOMICS* **4**, 1534-1536, doi:<https://doi.org/10.1002/pmic.200300744> (2004).
- 112 Montecchi-Palazzi, L. *et al.* The PSI-MOD community standard for representation of protein modification data. *Nature Biotechnology* **26**, 864-866, doi:10.1038/nbt0808-864 (2008).
- 113 Garavelli, J. S. The RESID Database of Protein Modifications as a resource and annotation tool. *Proteomics* **4**, 1527-1533, doi:10.1002/pmic.200300777 (2004).
- 114 Drown, B. S. *et al.* Mapping the Proteoform Landscape of Five Human Tissues. *Journal of Proteome Research*, doi:10.1021/acs.jproteome.2c00034 (2022).
- 115 LeDuc, R. D. *et al.* ProForma: A Standard Proteoform Notation. *Journal of Proteome Research* **17**, 1321-1325, doi:10.1021/acs.jproteome.7b00851 (2018).
- 116 Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Research* **48**, D835-D844, doi:10.1093/nar/gkz972 (2019).
- 117 Ramos, E. M. *et al.* Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *European Journal of Human Genetics* **22**, 144-147, doi:10.1038/ejhg.2013.96 (2014).
- 118 Orchard, S. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* **42**, D358-D363, doi:10.1093/nar/gkt1115 (2013).



- 
- 119 Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research* **49**, D545-D551, doi:10.1093/nar/gkaa970 (2020).
- 120 Fabregat, A. *et al.* Reactome graph database: Efficient access to complex pathway data. *PLOS Computational Biology* **14**, e1005968, doi:10.1371/journal.pcbi.1005968 (2018).
- 121 Nightingale, A. *et al.* The Proteins API: accessing key integrated protein and genome information. *Nucleic Acids Research* **45**, W539-W544, doi:10.1093/nar/gkx237 (2017).
- 122 Fabregat, A. *et al.* Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **18**, 142, doi:10.1186/s12859-017-1559-2 (2017).
- 123 Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Research* **44**, D481-D487, doi:10.1093/nar/gkv1351 (2016).
- 124 Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**, D605-D612, doi:10.1093/nar/gkaa1074 (2021).
- 125 Barabási, A. L. & Pál3sfai, M. *Network Science*. (Cambridge University Press, 2016).
- 126 Burger, B., Hernández Sánchez, L. F., Lereim, R. R., Barsnes, H. & Vaudel, M. Analyzing the Structure of Pathways and Its Influence on the Interpretation of Biomedical Proteomics Data Sets. *Journal of Proteome Research* **17**, 3801-3809, doi:10.1021/acs.jproteome.8b00464 (2018).
- 127 Cao, M. *et al.* Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks. *PLOS ONE* **8**, e76339, doi:10.1371/journal.pone.0076339 (2013).
- 128 Fernández-Tajes, J. *et al.* Developing a network view of type 2 diabetes risk pathways through integration of genetic, genomic and functional data. *Genome Medicine* **11**, 19, doi:10.1186/s13073-019-0628-8 (2019).
- 129 Udler, M. S. *et al.* Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLOS Medicine* **15**, e1002654, doi:10.1371/journal.pmed.1002654 (2018).
- 130 Oliver, S. Guilt-by-association goes global. *Nature* **403**, 601-602, doi:10.1038/35001165 (2000).
- 131 Han, H., Lee, S. & Lee, I. NGSEA: network-based gene set enrichment analysis for interpreting gene expression phenotypes with functional gene sets. *bioRxiv*, 636498, doi:10.1101/636498 (2019).
- 132 Arnau, V., Mars, S. & Marín, I. Iterative Cluster Analysis of Protein Interaction Data. *Bioinformatics* **21**, 364-378, doi:10.1093/bioinformatics/bti021 (2005).
- 133 Andersen, C. B. F. *et al.* Structure of the haptoglobin–haemoglobin complex. *Nature* **489**, 456-459, doi:10.1038/nature11369 (2012).

- 
- 134 Gonçalves, E. *et al.* Widespread Post-transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. *Cell Systems* **5**, 386-398.e384, doi:<https://doi.org/10.1016/j.cels.2017.08.013> (2017).
- 135 Meldal, B. H M. *et al.* Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Research* **47**, D550-D558, doi:10.1093/nar/gky1001 (2018).
- 136 Chalabi, M. H., Tsiamis, V., Käll, L., Vandin, F. & Schwämmle, V. CoExpresso: assess the quantitative behavior of protein complexes in human cells. *BMC Bioinformatics* **20**, 17, doi:10.1186/s12859-018-2573-8 (2019).
- 137 Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* **302**, 249-255, doi:doi:10.1126/science.1087447 (2003).
- 138 Drăghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C. & Krawetz, S. A. Global functional profiling of gene expression ☆ ☆ This work was funded in part by a Sun Microsystems grant awarded to S.D., NIH Grant HD36512 to S.A.K., a Wayne State University SOM Dean's Post-Doctoral Fellowship, and an NICHD Contraception and Infertility Loan to G.C.O. Support from the WSU MCBI mode is gratefully appreciated. *Genomics* **81**, 98-104, doi:[https://doi.org/10.1016/S0888-7543\(02\)00021-6](https://doi.org/10.1016/S0888-7543(02)00021-6) (2003).
- 139 Gupta, S., Turan, D., Tavernier, J. & Martens, L. The online Tabloid Proteome: an annotated database of protein associations. *Nucleic Acids Research* **46**, D581-D585, doi:10.1093/nar/gkx930 (2018).
- 140 Pai, S. *et al.* netDx: interpretable patient classification using integrated patient similarity networks. *Molecular Systems Biology* **15**, e8497, doi:<https://doi.org/10.15252/msb.20188497> (2019).
- 141 Sonawane, A. R., Weiss, S. T., Glass, K. & Sharma, A. Network Medicine in the Age of Biomedical Big Data. *Frontiers in Genetics* **10**, doi:10.3389/fgene.2019.00294 (2019).
- 142 Santos, A. *et al.* A knowledge graph to interpret clinical proteomics data. *Nature Biotechnology* **40**, 692-702, doi:10.1038/s41587-021-01145-6 (2022).
- 143 Schmidt, H. H. H. W. & Menche, J. The regulatory network architecture of cardiometabolic diseases. *Nature Genetics* **54**, 2-3, doi:10.1038/s41588-021-00994-w (2022).
- 144 Dobay, M. P., Stertz, S. & Delorenzi, M. Context-based retrieval of functional modules in protein-protein interaction networks. *Briefings in Bioinformatics* **19**, 995-1007, doi:10.1093/bib/bbx029 (2018).
- 145 Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164**, 805-817, doi:<https://doi.org/10.1016/j.cell.2016.01.029> (2016).
- 146 García-Campos, M. A., Espinal-Enríquez, J. & Hernández-Lemus, E. Pathway analysis: state of the art. *Frontiers in physiology* **6**, 383 (2015).
- 147 Krishna, B. M. *et al.* Notch signaling in breast cancer: From pathway analysis to therapy. *Cancer Letters* **461**, 123-131, doi:<https://doi.org/10.1016/j.canlet.2019.07.012> (2019).

- 
- 148 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
- 149 Folger, O. *et al.* Predicting selective drug targets in cancer through metabolic networks. *Molecular Systems Biology* **7**, 501, doi:<https://doi.org/10.1038/msb.2011.35> (2011).
- 150 Yi, Y., Fang, Y., Wu, K., Liu, Y. & Zhang, W. Comprehensive gene and pathway analysis of cervical cancer progression. *Oncol Lett* **19**, 3316-3332, doi:10.3892/ol.2020.11439 (2020).
- 151 Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101-113, doi:10.1038/nrg1272 (2004).
- 152 Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
- 153 Ihnatova, I., Popovici, V. & Budinska, E. A critical comparison of topology-based pathway analysis methods. *PLOS ONE* **13**, e0191154, doi:10.1371/journal.pone.0191154 (2018).
- 154 Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research* **44**, D1214-1219, doi:10.1093/nar/gkv1031 (2016).
- 155 Huang, T., Wang, J., Yu, W. & He, Z. Protein inference: a review. *Briefings in Bioinformatics* **13**, 586-614, doi:10.1093/bib/bbs004 (2012).
- 156 Nesvizhskii, A. I. & Aebersold, R. Interpretation of Shotgun Proteomic Data. *Molecular & Cellular Proteomics* **4**, 1419-1440, doi:10.1074/mcp.R500012-MCP200 (2005).
- 157 Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *Journal of proteome research* **9**, 1323-1329 (2010).
- 158 Miller, R. M. *et al.* Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data. *Journal of Proteome Research* **18**, 3429-3438, doi:10.1021/acs.jproteome.9b00330 (2019).
- 159 Wu, C. *et al.* A protease for 'middle-down' proteomics. *Nature Methods* **9**, 822-824, doi:10.1038/nmeth.2074 (2012).
- 160 Smith, L. M. *et al.* The Human Proteoform Project: Defining the human proteome. *Science Advances* **7**, eabk0734, doi:10.1126/sciadv.abk0734 (2021).
- 161 Cafarelli, T. M. *et al.* Mapping, modeling, and characterization of protein-protein interactions on a proteomic scale. *Current Opinion in Structural Biology* **44**, 201-210, doi:<https://doi.org/10.1016/j.sbi.2017.05.003> (2017).
- 162 Melkonian, M., Juigné, C., Dameron, O., Rabut, G. & Becker, E. Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases. *Bioinformatics* **38**, 1685-1691, doi:10.1093/bioinformatics/btac013 (2022).

- 
- 163 Luck, K., Sheynkman, G. M., Zhang, I. & Vidal, M. Proteome-Scale Human Interactomics. *Trends in Biochemical Sciences* **42**, 342-354, doi:<https://doi.org/10.1016/j.tibs.2017.02.006> (2017).
- 164 Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402-408, doi:10.1038/s41586-020-2188-x (2020).
- 165 Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509**, 575-581, doi:10.1038/nature13302 (2014).
- 166 Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annual review of analytical chemistry (Palo Alto, Calif.)* **9**, 499-519, doi:10.1146/annurev-anchem-071015-041550 (2016).
- 167 Bludau, I. *et al.* Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nature Communications* **12**, 3810, doi:10.1038/s41467-021-24030-x (2021).
- 168 Zahn-Zabal, M. *et al.* The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Research* **48**, D328-D334, doi:10.1093/nar/gkz995 (2020).
- 169 Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biology* **11**, R53, doi:10.1186/gb-2010-11-5-r53 (2010).
- 170 Gan, W., Lin, J. C.-W., Fournier-Viger, P., Chao, H.-C. & Yu, P. S. A survey of parallel sequential pattern mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **13**, 1-34 (2019).
- 171 Lipton, Z. C., Berkowitz, J. & Elkan, C. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019* (2015).
- 172 Yu, Y., Si, X., Hu, C. & Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation* **31**, 1235-1270, doi:10.1162/neco\_a\_01199 (2019).
- 173 Graves, A. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850* (2013).
- 174 Salehinejad, H., Sankar, S., Barfett, J., Colak, E. & Valaee, S. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078* (2017).



---

## 10. Papers



# Paper I














# **Paper II**



TECHNICAL NOTE

# PathwayMatcher: proteoform-centric network construction enables fine-granularity multiomics pathway mapping

Luis Francisco Hernández Sánchez <sup>1,2,3</sup>, Bram Burger <sup>4,5</sup>, Carlos Horro <sup>4,5</sup>, Antonio Fabregat <sup>3</sup>, Stefan Johansson <sup>1,2</sup>, Pål Rasmus Njølstad <sup>1,6</sup>, Harald Barsnes <sup>4,5</sup>, Henning Hermjakob <sup>3,7</sup> and Marc Vaudel <sup>1,2,\*</sup>

<sup>1</sup>K.G. Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Children's Hospital, Haukeland University Hospital, 5021 Bergen, Norway; <sup>2</sup>Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, P.O. Box 1400, 5021 Bergen, Norway; <sup>3</sup>European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom; <sup>4</sup>Proteomics Unit, Department of Biomedicine, University of Bergen, Postbox 7804, 5020 Bergen, Norway; <sup>5</sup>Computational Biology Unit, Department of Informatics, University of Bergen, P.O. Box 7803, 5020 Bergen, Norway; <sup>6</sup>Department of Pediatrics, Haukeland University Hospital, 5021 Bergen, Norway and <sup>7</sup>Beijing Proteome Research Center, National Center for Protein Sciences Beijing, No. 38, Life Science Park Road, Changping District, 102206 Beijing, China

\*Correspondence address. Marc Vaudel, Children's Hospital, Haukeland University Hospital, 5021 Bergen, Norway. E-mail: [marc.vaudel@uib.no](mailto:marc.vaudel@uib.no)  <http://orcid.org/0000-0003-1179-9578>

## Abstract

**Background:** Mapping biomedical data to functional knowledge is an essential task in bioinformatics and can be achieved by querying identifiers (*e.g.*, gene sets) in pathway knowledge bases. However, the isoform and posttranslational modification states of proteins are lost when converting input and pathways into gene-centric lists. **Findings:** Based on the Reactome knowledge base, we built a network of protein-protein interactions accounting for the documented isoform and modification statuses of proteins. We then implemented a command line application called PathwayMatcher ([github.com/PathwayAnalysisPlatform/PathwayMatcher](https://github.com/PathwayAnalysisPlatform/PathwayMatcher)) to query this network. PathwayMatcher supports multiple types of omics data as input and outputs the possibly affected biochemical reactions, subnetworks, and pathways. **Conclusions:** PathwayMatcher enables refining the network representation of pathways by including proteoforms defined as protein isoforms with posttranslational modifications. The specificity of pathway analyses is hence adapted to different levels of granularity, and it becomes possible to distinguish interactions between different forms of the same protein.

**Keywords:** pathway; posttranslational modification; network; proteoform

Received: 17 December 2018; Revised: 3 June 2019; Accepted: 30 June 2019

© The Author(s) 2019. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

## Findings

In biomedicine, molecular pathways are used to infer the mechanisms underlying disease conditions and identify potential drug targets. Pathways are composed of series of biochemical reactions, of which the main participants are proteins, that together form a complex biological network. Proteins can be found in various forms, referred to as proteoforms [1]. The different proteoforms that can be obtained from the same gene/protein depend on the individual genetic profiles, on sequence cleavage and folding, and on posttranslational modification (PTM) states [2]. Proteoforms can carry PTMs at specific sites, conferring each proteoform unique structure and properties [2]. Notably, many pathway reactions can only occur if all or some of the proteins involved are in specific posttranslational states.

However, when analyzing omics data, both input and pathways are summarized in a gene- or protein-centric manner, meaning that the different proteoforms and their reactions are grouped by gene name or protein accession number, and the fine-grained structure of the pathways is lost. One can therefore anticipate that proteoform-centric networks provide a rich new paradigm to study biological systems. But while gene networks have proven their ability to identify genes associated with diseases [3], networks of finer granularity remain largely unexplored.

Here, we present PathwayMatcher, an open-source standalone application that considers the isoform and PTM status when building protein networks and mapping omics data to pathways from the Reactome database. Reactome [4] is an open-source curated knowledge base consolidating documented biochemical reactions categorized in hierarchical pathways and notably includes isoform and PTM information for the proteins participating in reactions and pathways.

As an example of the complexity of hierarchical pathway information, we provide a graph representation of *Signaling by NOTCH2* from Reactome (Fig. 1). This pathway is a subpathway of the pathways *Signaling by NOTCH* and *Signal Transduction*. It is composed of two subpathways (*NOTCH2 Intracellular Domain Regulates Transcription* and *NOTCH2 Activation and Transmission of Signal to the Nucleus*), comprising 32 and 54 reactions, yielding 28 and 141 edges, respectively. The 31 participants of the *Signaling by NOTCH2* pathway are also involved in reactions in other pathways, between themselves and with 2,055 other proteins, resulting in 6,525 external edges. Note that in this pathway, Cyclic AMP-responsive element-binding protein 1 (coded by *CREB1*) is phosphorylated at position 46 (labeled as *CERB1.P* in Fig. 1) and Neurogenic locus notch homolog protein 2 (coded by *NOTCH2*) is found in 3 forms (unmodified and with two combinations of glycosylation, labeled as *NOTCH2*, *NOTCH2.Gly1*, and *NOTCH2.Gly2*, respectively).

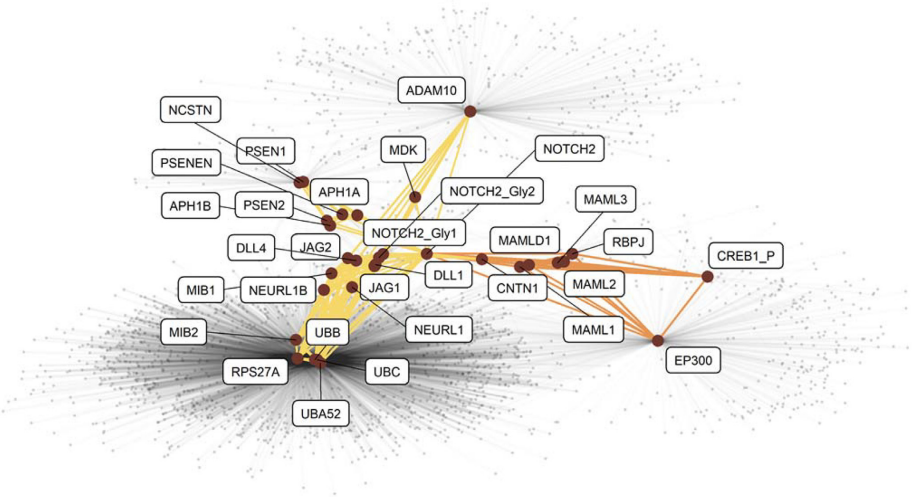
The amount of information available on reactions involving modified proteins has dramatically increased during the past two decades (Fig. 2), with 3,947 and 5,631 publications indexed in Reactome (version 64 at time of writing) describing at least one reaction between modified proteins or between a modified and an unmodified protein, respectively. To harness this vast amount of knowledge, we built a network representation of pathways that we refer to as *proteoform-centric*, where protein isoforms with different sets of PTMs are represented with different nodes, in contrast to *gene-centric* networks, where one node is used per gene name or protein accession. In this representation, two proteoforms are connected if they participate in the

same reaction. Note that proteoforms can participate in reactions both individually and as part of a set or complex. Furthermore, they can have four different roles: input, output, catalyst, or regulator.

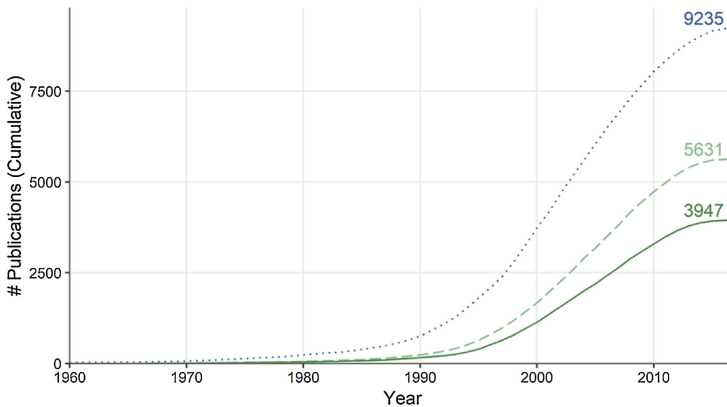
The fundamental difference between gene- and proteoform-centric networks is illustrated in Fig. 3, showing the graph representation of interactions with the protein *cellular tumor antigen p53* (P04637) from the *TP53* gene. In a gene-centric paradigm (Fig. 3A), 221 nodes are connected to a single node, making 220 connections; while in a proteoform-centric network (Fig. 3B), 227 proteoforms connect to 23 proteoforms coded by *TP53*, making 414 connections. Note that the proteoforms coded by *TP53* are themselves involved in reactions, making 24 *TP53-TP53* connections. In this example, the proteoform-centric network thus presents more nodes and connections than the gene-centric network, with visible structural differences in the network organization. We hypothesize that the proteoform-centric network paradigm depicted in Fig. 3B provides a rich map that will enable navigating biomedical knowledge to a higher level of detail, to better assess the effect of perturbations and identify drug targets more specifically.

PathwayMatcher allows the user to tune the granularity of the network representation of pathways by representing nodes as (i) gene names, (ii) protein accession numbers, or (iii) proteoforms and supports the mapping of multiple types of omics data: (i) genetic variants, (ii) genes, (iii) proteins, (iv) peptides, and (v) proteoforms. Genetic variants are mapped to proteins using the Ensembl Variant Effect Predictor [5], gene names are mapped to proteins using the UniProt identifier mapping [6], and peptides are mapped to proteins using PeptideMapper [7]. If a peptide maps to different proteins, all possible proteins are considered for the search and protein inference must be conducted *a posteriori* [8]. If peptides are modified, they are mapped to the proteoforms presenting compatible PTM sets. Proteins are mapped to the pathway network using their accession, while proteoforms are mapped by comparing their protein accession, isoform number, and PTM set. A schematic representation of the PathwayMatcher matching procedure is shown in Fig. 4. More details on the mapping procedure, formats, and settings can be found in the Methods section and in the online documentation ([github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki](https://github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki)). For more information on how the pathway representation is constructed from the different external resources, please consult the Methods section and the online documentation ([github.com/pathwayanalysisplatform/pathwaymatcher/tree/master/src/main/java/extractor](https://github.com/pathwayanalysisplatform/pathwaymatcher/tree/master/src/main/java/extractor)).

PathwayMatcher produces three types of output: (i) the result of the matching, listing all possible reactions and pathways linked to the input; (ii) the results of an overrepresentation analysis; and (iii) networks in relationship with the input. The overrepresentation analysis is performed on the pathways matching and follows the first generation of pathway analysis methods [9], i.e., a *P*-value for each pathway in the reference database is calculated using a binomial distribution followed by Benjamini-Hochberg correction [10] (in a similar way as performed by the Reactome online analysis tool [4]). If the input can be mapped to proteoforms, the overrepresentation analysis is conducted using a proteoform-centric representation of pathways, using proteins otherwise. The exported networks represent the internal and external connections that can be drawn from the input, where internal connections connect two nodes from the input list, and external connections connect one node from the input list to any



**Figure 1.** Graph representation of the Signaling by NOTCH2 pathway as extracted from the Reactome database. Participating proteins are displayed as large dark red dots labeled with their canonical gene name. Posttranslational modifications (PTMs) are indicated with suffixes in the label. A connection between two dots indicates a documented interaction between the two proteins in the given pathway. Connections belonging to the subpathways NOTCH2 intracellular domain regulate transcription and NOTCH2 activation and transmission of signal to the nucleus are displayed in orange and yellow, respectively. The interactions involving these proteins in other pathways are displayed with light gray connections in the background.



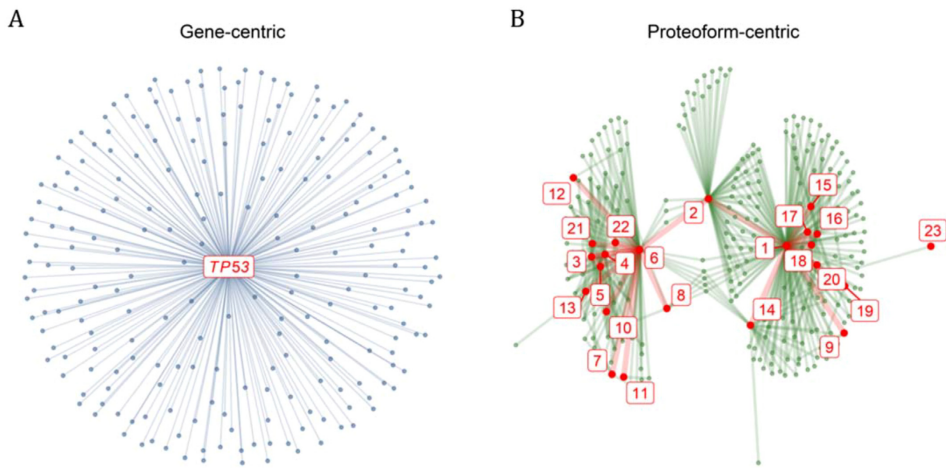
**Figure 2.** The cumulative number of publications indexed in Reactome documenting at least one reaction between two proteins with PTMs (solid dark green line), between one protein with PTMs and one without (dashed light green line), and two proteins without PTMs (dotted blue line), counting all publications with a year earlier than or equal to the x-axis value. The number of publications in each category at time of writing is indicated to the right.

node not in the input. The user can select to export these networks using nodes defined as genes, proteins, or proteoforms. Connections between nodes in the network are annotated with information on whether they participate as complex or set and their role in the reaction.

As displayed in Fig. 5A, 68% of the pathways present at least one proteoform-specific participant, i.e., with isoform or PTM annotation. The number of pathways containing a given gene product or proteoform is displayed in Fig. 5B, showing how using proteoforms allows distinguishing pathways more specifically than genes, with a median of 4 pathways matched per proteoform compared to 11 pathways per gene. When the input can be

mapped to proteoforms, PathwayMatcher can restrict the search for reactions and pathways to those that specifically involve proteins in the desired form, hence reducing the number of possible connections for a given node in the resulting network. Conversely, the proteoform-centric network representation allows identifying interactions between multiple proteoforms originating from the same gene or protein, resulting in new connections compared to a gene-centric representation.

Figure 5C shows that the number of connections per proteoform is lower than the number of connections for the respective gene for most proteoforms, varying from a 300-fold decrease to a 10-fold increase. Interestingly, plotting the number of connec-



**Figure 3.** Gene-centric versus proteoform-centric representation. (A) Graph representation of the genes involved in reactions (through their corresponding proteins) with (the corresponding proteins of) TP53, with a single node per gene. TP53 is represented with a red label at the center and genes coding proteins involved in reactions with TP53 are represented with smaller blue dots at the periphery connected to the TP53 gene with blue lines. (B) Graph representation of the proteins involved in a reaction with gene products of TP53, distinguishing isoforms and posttranslationally modified proteins as different proteoforms. The proteoforms coded by TP53 and the proteoforms involved in a reaction with them are represented with large red and small green dots, respectively. The proteoforms coded by TP53 are numbered according to Table 1. The connections between proteoforms coded by TP53 are displayed with thick red lines and connections with other proteoforms with thin green lines.

tions of a proteoform in gene-centric or proteoform-centric networks shows that the largest gene-centric hubs, corresponding to 5 genes, decompose into 127 proteoforms that do not outlie the distribution of the number of connections in the proteoform network (Fig. 5D). Conversely, a group of 484 densely connected outliers emerges from 44 genes.

In order to fully benefit from the gain in specificity of the proteoform-representation of pathways, it is necessary to exactly match the representation of proteoforms in Reactome. Any mismatch between the input data and the database would result in a loss of sensitivity. In practice, such mismatches can result from an incomplete proteoform representation in Reactome, where only the minimal set of modifications necessary to perform a reaction is annotated. Conversely, input data can present unresolved isoform, missing modifications, or inaccurate localization, especially in the case of bottom-up proteomics [11]. Since the size of the proteoform network is unknown to date, the effect of missing annotations in the database is not directly quantifiable.

To estimate the sensitivity of the matching, we mapped the phosphoproteome from Ochoa et al. [12] to Reactome using PathwayMatcher: among the 10,588 accessions representing phosphoproteins, 5,519 (52%) could be matched to an accession in Reactome, while among the 116,258 phosphosites reported, only 654 (<1%) could be matched exactly in Reactome. Accession matching is equivalent in terms of sensitivity and specificity to a gene-centric representation of pathways, while strict proteoform matching, requiring exact isoform and modification set, maximizes specificity at the cost of sensitivity.

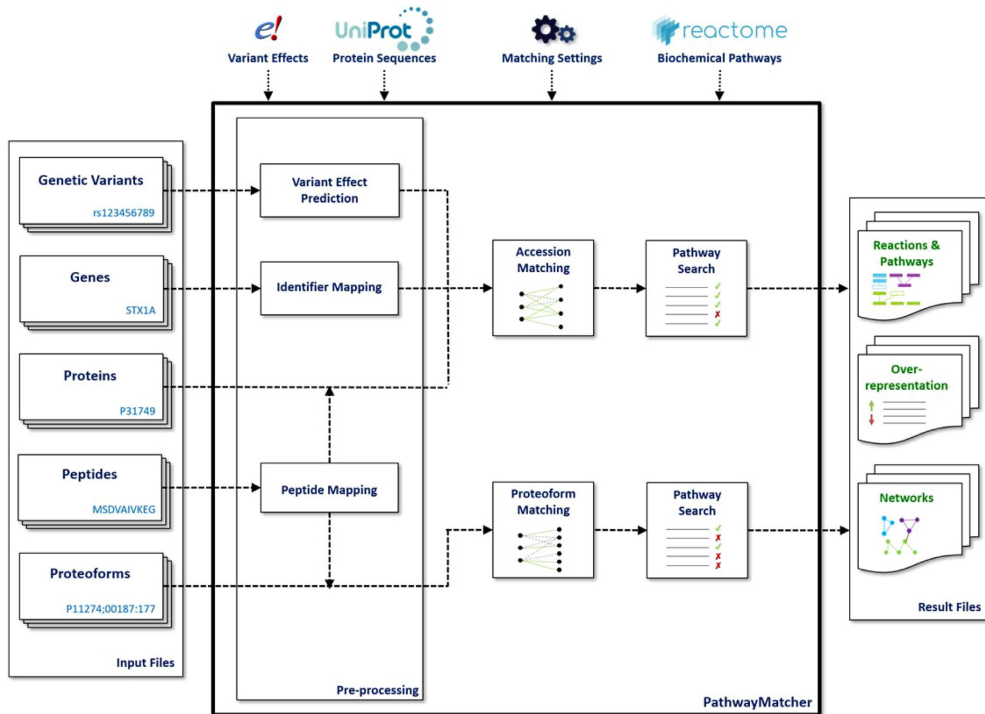
In order to mitigate the sensitivity loss while maintaining specificity, we implemented multiple types of matching that present different levels of stringency, as detailed in the methods: (i) *One*, (ii) *One without PTM types*, (iii) *Superset*, (iv) *Superset without PTM types*, (v) *Subset*, (vi) *Subset without PTM types*, and (vii) *Strict*. Table 2 lists the share of phosphosites that can be

matched to a proteoform in Reactome when querying the accession with a phosphorylation at the given site, and only at this site, with a tolerance of 5 amino acids. There, one can see that increasing the stringency of the matching dramatically reduces the sensitivity. Since both Reactome and the list of phosphosites represent a minimal set of modifications, the *Strict* matching is overly selective, while *Accession* and *Superset* include reactions where the proteins are not modified.

*Subset* and *One* represent the coverage of the input by Reactome. Here, *Subset* and *One* are equivalent because the input consists of single phosphosites. In a data set containing combinations of phosphosites, *Subset* would match proteoforms taking phosphosite combinations into account, while *One* would represent any proteoform with at least one matching phosphosite. The increased number of matches without PTM type can be imputed to mismatching PTM identifiers or the presence of other PTMs at the input sites or at neighboring positions.

To illustrate the difference induced by each matching type on the proteoform matching, we calculated the percentage of proteoforms matched with selected example proteoforms. In Fig. 6, we present two example proteoforms, one from insulin (P01308) and one from mitogen-activated protein kinase kinase 7 (MAP3K7). Insulin and MAP3K7 have five and seven different proteoforms annotated in Reactome, four and six of them with PTM annotation, respectively. By design, the *Strict* matching type matches only the original proteoform while the accession matching matches all proteoforms. The other matching types allow balancing between the two stringencies and display varying levels of specificity for those proteoforms. The results show that relaxing the stringency of the matching rapidly induces a loss in specificity due to the similarity of the different proteoforms of a given gene or protein.

Furthermore, we randomly selected proteoforms in Reactome and altered them by changing the type and localization of the PTMs to simulate mismatching or missing information,



**Figure 4.** Schematic representation of the PathwayMatcher matching procedure. Input of various types is modeled as sets of proteins or proteoforms based on the annotation of isoforms and PTMs. Proteins and proteoforms are then mapped to Reactome based on user settings. Matched reactions and pathways, the results of an overrepresentation analysis, and subnetworks generated from the input are exported as text files.

and the altered proteoforms were matched to Reactome; see details in the Methods section. In this setup, the share of altered proteoforms that can be recovered using the different matching types, referred to as *Original* matches, provides an estimate of the matching sensitivity in case of incomplete or mismatching proteoform definition. Conversely, the share of other proteoforms matching despite not being originally selected, referred to as *Other* matches, provides an estimate of the error rate, the complement of specificity.

Fig. 7 shows the percentage of proteoforms that matched at least one proteoform in the database separated on matching type. As expected, accession matching displays the highest sensitivity at the lowest specificity, while the *Strict* and *Subset* matching display the highest specificity at the lowest sensitivity. The *Superset* matching presented low sensitivity and low specificity, while the *One* matching presented a balance between specificity and sensitivity. Finally, the matching with no types presented similar trends but with almost maximum sensitivity and lower specificity. Together, these results show how relaxing the matching stringency allows balancing between sensitivity and specificity, and they demonstrate the importance of accurate proteoform definition in both the input and the reference knowledge base.

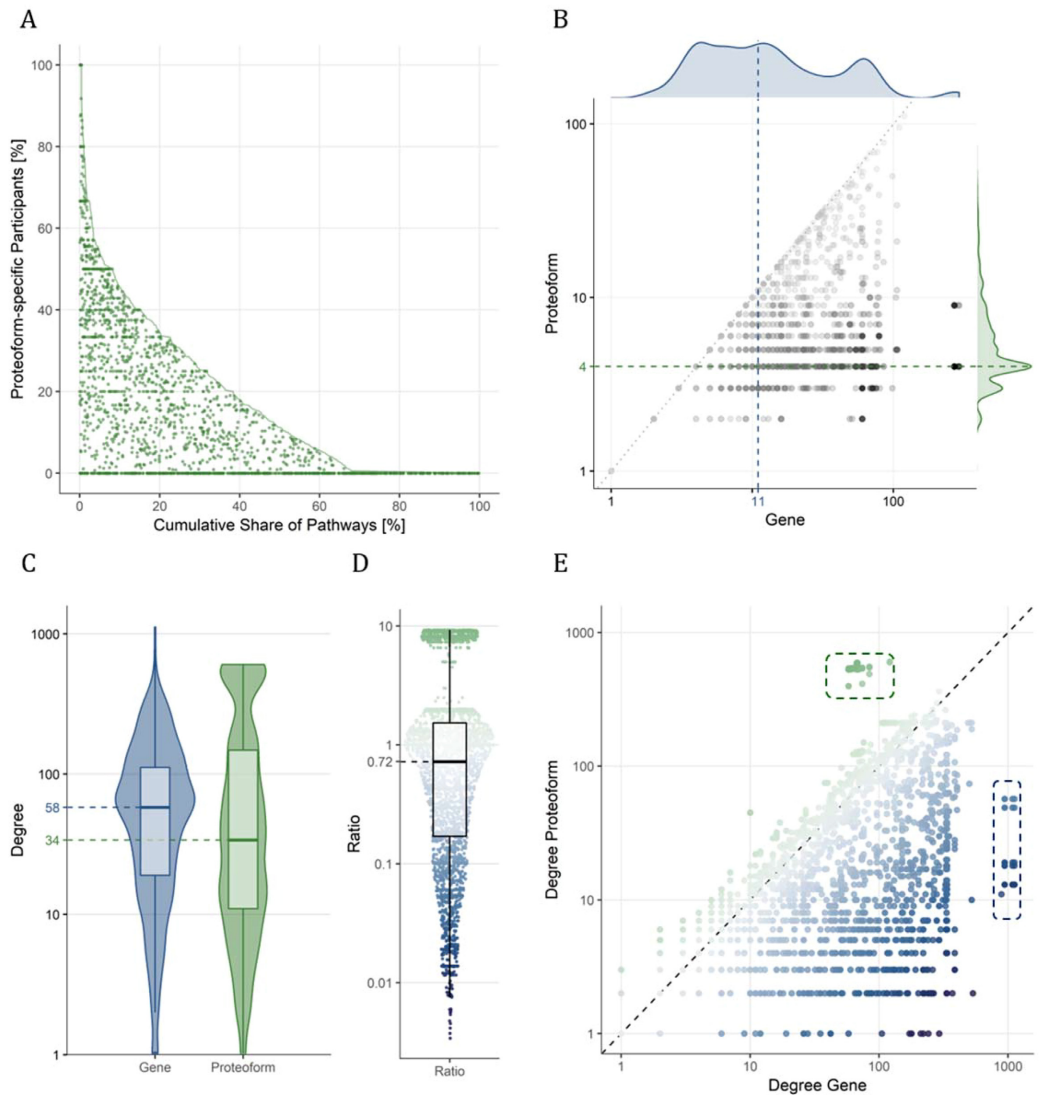
Through its paradigm shift, PathwayMatcher hence provides a fine-grained representation of pathways for the analysis of omics data. However, this comes at the cost of increased complexity: gene-centric networks comprise a limited

number of nodes, approximately 20,000 for humans, whereas in a proteoform-centric paradigm, the human network is expected to have several million nodes [13]. With the current version of Reactome, building the gene- and proteoform-centric networks results in 9,759 and 12,775 nodes with 443,229 and 672,047 connections, respectively. We classified the nodes into two categories, canonical or specific gene products, depending on whether or not they represent the unmodified canonical isoform of a protein according to UniProt. Within the proteoform network, 432,169 connections between 9,694 nodes link two canonical gene products, 95,539 connections between 7,734 nodes involved one canonical and one specific gene product, and 2,806 nodes with 144,339 connections involved two specific gene products. More summary statistics on the underlying network can be found in the wiki of the PathwayMatcher repository.

In addition to the increased size of the underlying network, matching proteoforms requires comparing isoforms and sets of modifications, possibly with tolerance and wildcards for the modification definition and localization, which is computationally much more intensive than simply comparing identifiers. Fig. 8 shows the performance of PathwayMatcher benchmarked against public data sets of (A) genetic variants, (B) proteins, (C) peptides, and (D) proteoforms.

For the proteins and proteoforms, the processing time increased linearly related to the query size with a small slope, making it possible to search all available proteins within a few seconds. As expected, protein identifiers provided the fastest re-





**Figure 5.** Prevalence of proteoforms in pathways. (A) The share of proteoform-specific participants in a pathway (i.e., proteins that are annotated with isoform and/or PTM information) is plotted against the cumulative share of pathways, going from the highest share of proteoforms to the lowest. The cumulative share of pathways is displayed with a solid green line. The share of proteoform-specific participants in each pathway is plotted with a green dot with a jitter on the x-axis between zero and the solid line. (B) For all proteoform-specific participants, the number of pathways mapped using the proteoform versus gene is plotted in black. The density of the number of pathways mapped is indicated at the top (blue) and right (green) for gene and proteoform matching, respectively. The median number of pathways mapped is indicated with dashed lines. (C) The violin and box plots of the degree, i.e., number of connections, for the proteoform-specific participants in a gene-centric or proteoform-centric network are plotted to the left (blue) and right (green), respectively. (D) The ratio of degrees, proteoform over gene, is plotted with a blue-gray-green gradient with the box plot overlaid in black. (E) The degree of the proteoform-specific participants in the proteoform-centric network is plotted against the degree in the gene-centric network. Dots are colored with a blue-gray-green gradient corresponding to the ratio in D. Outliers of high degree in the gene-centric but not in the proteoform-centric network are indicated with blue dashes to the right. Outliers of high degree in the proteoform-centric but not in the gene-centric network are indicated with green dashes to the top. Note that base 10 logarithmic scales are used for the axes in B, C, D, and E.

sponse time, while proteoforms were the second fastest. Mapping peptides took approximately 30 seconds more, corresponding to the indexing time of the protein sequences database by

PeptideMapper [7], after which the time increased linearly in a similar fashion as for proteins. For the genetic variants, an extra mapping step is required to map possibly affected proteins, adding additional computing time. The overall mapping time for

**Table 1.** Proteoforms of Figure 3B

#	Isoform	Modifications
1	Canonical	None
2	Canonical	pS15
3	Canonical	pS15 pS20 aceK120 aceK382
4	Canonical	pS15 pS20 aceK382
5	Canonical	pS15 pS20 aceK120
6	Canonical	pS15 pS20
7	Canonical	pS15 pS20 dimethR335 dimethR337 methR333
8	Canonical	pS15 pS20 ubiK
9	Canonical	pS15 pS33 pS46
10	Canonical	pS15 pS20 pS269 pT284
11	Canonical	pS15 pS20 methK370
12	Canonical	pS15 pS20 methK372
13	Canonical	pS15 pS20 methK382
14	Canonical	ubiK
15	Canonical	pS315
16	Canonical	pT55
17	Canonical	pS15 pS392
18	Canonical	pS37
19	Canonical	dimethK373
20	Canonical	sumoK386
21	Canonical	pS15 pS20 pS46
22	Canonical	pS15 pS20 pS392
23	Canonical	dimethK370 dimethK382

Only the canonical isoforms are annotated to date, as indicated in the second column. The posttranslational modification status is indicated in the third column with modification short name and modification site when annotated. Abbreviations: aceK, N6-acetyl-L-lysine; dimethK, N6, N6-dimethyl-L-lysine; dimethR, symmetric dimethyl-L-arginine; methK, N6-methyl-L-lysine; methR, omega-N-methyl-L-arginine; pS, O-phospho-L-serine; pT, O-phospho-L-threonine; ubiK, ubiquitinated lysine; sumoK, sumoylated lysine.

**Table 2.** Share of the phosphosites from Ochoa et al. [12] matching to Reactome using different matching types

Matching Type	Share of Phosphosites Matched
Accession	57.44%
Superset without PTM types	56.38%
Superset	56.33%
One without PTM types	6.01%
Subset without PTM types	6.01%
One	1.27%
Subset	1.27%
Strict	0.15%

Proteoforms were constructed by adding a phosphorylation at the given site, and only at this site, and were queried against Reactome. The percentage of proteoforms matched is provided in the second column. A tolerance of 5 amino acids was used on the modification site. More details on this analysis can be found in the Methods section.

a million single-nucleotide polymorphisms (SNPs) was less than a minute, which is acceptable compared to the other steps of a variant analysis pipeline. Note that the processing time was very reproducible across runs, where minor variation is only noticeable using genetic variants, resulting in very thin ribbons in Fig. 8B-D.

In conclusion, PathwayMatcher is a versatile application enabling the mapping of several types of omics data to pathways in reasonable time and can readily be included in bioinformatic workflows. It is important to underline that PathwayMatcher maps experimental data to pathways in a systematic and unbi-

ased fashion, i.e., it collects all pathways containing at least one of the participant proteins or proteoforms of the input data and does not perform any filtering or biological inference. Through this process, it attempts at minimizing the prevalence of false negatives by considering all the possible pathways annotated in the reference database. It can, however, not control for missing annotation, i.e., what is not annotated in the knowledge base is not considered.

Furthermore, although PathwayMatcher implements an overrepresentation analysis module, we recommend that users rather interpret the results of the matching and the resulting networks using the systems biology method that best suits the experiment and biomedical context. Based on generic pathways, PathwayMatcher is not developed as a mechanism inference or validation tool, but as a hypothesis generation tool, helping to navigate large data sets and guide experiments to uncover biological processes relevant to specific research questions.

Thanks to the fine-grained information available in Reactome, PathwayMatcher supports refining the pathway representation to the level of proteoforms. To date, only a fraction of the several million expected proteoforms [13] has annotated interactions, but as the understanding of protein interactions continues to increase and the ability to identify and characterize them in samples progresses, proteoform-centric networks will surely become of prime importance in biomedical studies. Notably, the effect of genetic variation on genes, transcripts, and proteins is currently only partially resolved for a fraction of the genome. The rapid development of this field will make it possible to identify biological functions affected by variants within the human network. Refining its representation to the level of proteoforms will allow pinpointing more precisely reactions and pathways, and hence increase our ability to understand biological mechanisms and potentially identify druggable targets.

## Methods

### Implementation

PathwayMatcher is implemented in Java 8.0.

### Availability

PathwayMatcher is freely available at [github.com/PathwayAnalysisPlatform/PathwayMatcher](https://github.com/PathwayAnalysisPlatform/PathwayMatcher) under the permissive Apache 2.0 license. It is also possible to use PathwayMatcher as a Docker image: [hub.docker.com/r/lfhs/pathwaymatcher](https://hub.docker.com/r/lfhs/pathwaymatcher). PathwayMatcher can be obtained from the Bioconda channel of the Conda [14] package manager at [bioconda.github.io/recipes/pathwaymatcher/README.html](https://bioconda.github.io/recipes/pathwaymatcher/README.html). Finally, PathwayMatcher is available as a Galaxy [15] tool in the Galaxy ToolShed [16] at [toolshed.g2.bx.psu.edu/view/galaxyp/reactome.pathwaymatcher](https://toolshed.g2.bx.psu.edu/view/galaxyp/reactome.pathwaymatcher), where it can be readily integrated into analysis workflows. PathwayMatcher has also been installed into the public European Galaxy instance, [usegalaxy.eu](https://usegalaxy.eu), making it possible to use the application without requiring any local configuration and just providing valid input files and options. The complete URL for the online tool is listed in reference [17].

Upon installation, PathwayMatcher can be used from the command line to query Reactome using various types of omics data. Either the "jar" file is run directly using Java or the Docker image is instantiated to a container. Detailed information on implementation, installation, usage, and format specifications is available in the online documentation at [github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki](https://github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki).

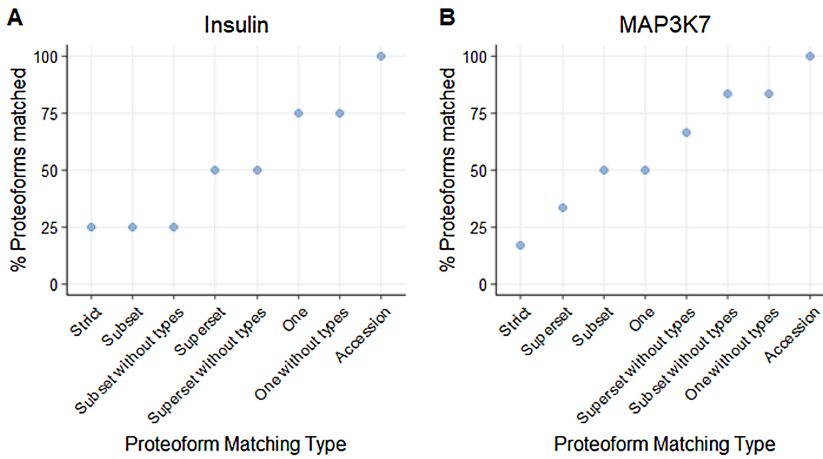


Figure 6. Two examples of proteoforms showing the proteoform matching results for each matching type. (A) Proteoform P01308; MOD:00087:53;MOD:00798:31;MOD:00798:43, from insulin (P01308), is matched against all modified proteoforms of insulin in Reactome. (B) Proteoform O43318;MOD:00047:184;MOD:00047:187, from “mitogen-activated protein kinase kinase kinase 7” (MAP3K7), is matched against all modified proteoforms of MAP3K7 in Reactome.

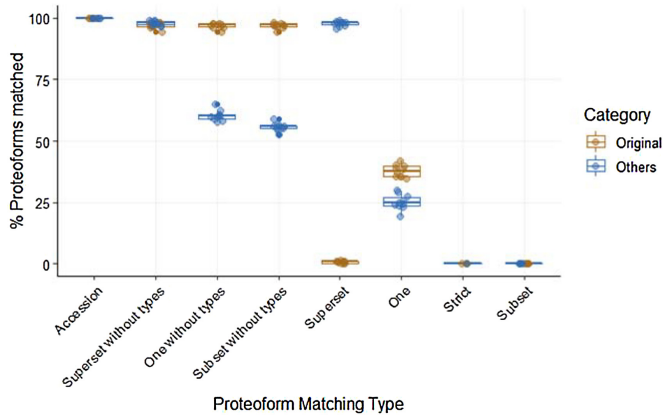


Figure 7. Percentage of proteoforms with at least one proteoform match in the database with each matching criterion. The total candidate proteoforms available are separated in two categories, *Original* and *Others*. *Original* is the proteoform in the database that was modified for the sampling, while *Others* are the proteoforms that share the same protein accession.

### Input and output

Detailed and updated documentation of the input and output can be found in the online documentation at [github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki](https://github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki).

As schematized in Fig. 9, a simple representation is used for proteoforms: (i) the UniProt protein accession and (ii) the set of PTMs separated by a semicolon “;”. The protein accession can include the isoform number specified with a dash “-”. The PTM set contains each PTM separated by a comma “,”. Each PTM is specified using a modification identifier and a site, separated by a colon “:”.

Note that the order of PTMs does not affect the search. The PTM identifier is a 5-digit identifier from the PSI-MOD Protein Modification [18]. The site is an integer specifying the 1-based

index of the modified amino acid on the sequence as defined by UniProt. The modification site field is mandatory, and ? or null indicates that the position is not known.

It is common to write the identifiers for the PTM types with the prefix “MOD:” before the 5 digits of the ontology term. PathwayMatcher also allows the user to write the identifier without the prefix. PathwayMatcher also allows querying all proteoforms modified at a given site using the “00000” wildcard for modification type combined with a matching type that does not consider the modification types such as *One without types* or *Subset without types*. For more details, see the Proteoform Matching subsection.

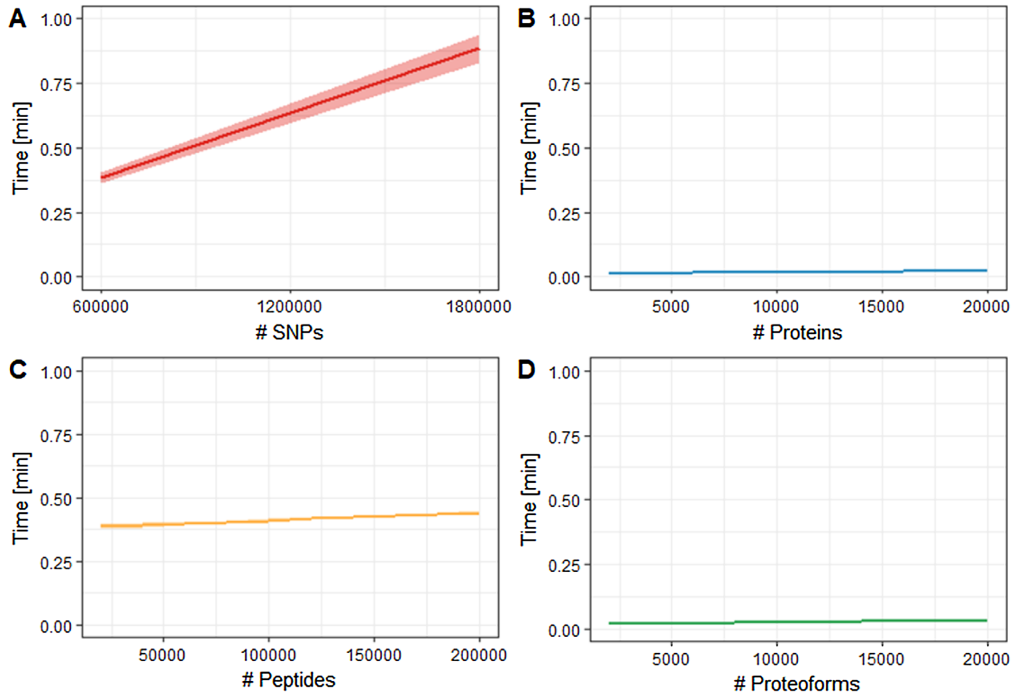


Figure 8. Performance of PathwayMatcher using (A) genetic variants as single-nucleotide polymorphisms (SNPs), (B) proteins, (C) peptides, and (D) proteoforms. Time in minutes is plotted against input size. The mean is displayed as a solid line and the 95% range as a ribbon (only visible in (A) due to the high reproducibility in other cases).

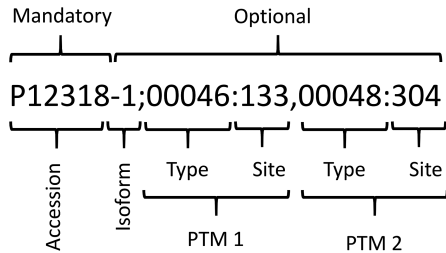


Figure 9. Example of proteoform notation, composed of a protein accession, an isoform number, and a set of PTMs.

### Posttranslational modifications in the Reactome data model

The Reactome object model specifies physical entities (e.g., complexes, proteins, and small molecules) and proteins are annotated using unique identifiers. These entities participate in reactions in specific cellular compartments. They can also be connected to multiple instances of *Translational Modification* objects, which contain a specific coordinate on the protein sequence and an identifier following the PSI-MOD ontology [18]. The portion of physical entities referring to proteins is associated with another class of objects as reference entities, which contain protein annotations in external databases such as UniProt [19]. Therefore,

a proteoform is represented as a physical entity associated with a set of modifications for specific processes at a specific subcellular location. Each modification has a PSI-MOD ontology identifier as type and an integer coordinate for the site in the peptide sequence where the modification occurs. The coordinate can be ? or null when the site is not known. Reactome annotates 127 different protein modifications for humans, of which Fig.10 Reference source not found. displays the most frequent.

### Proteoform matching

Searching pathways using gene names or protein accessions solely requires mapping a string of characters between the input and the knowledge base. In order to map the proteoforms to reactions and pathways, it is necessary to decide if the proteoforms in the input are equivalent to the proteoforms annotated in the reference database, Reactome, taking into account the protein accession, isoform information, and the set of PTMs. Two proteoforms can have all, some, or none of these elements in common. We defined a set of criteria to match two proteoforms, one from the input and another from the reference database. First, identical protein accession and isoform numbers are required for a match: either both proteoforms are from the canonical isoform (e.g., P31749) or from the same isoform (e.g., P31749-3). Then, the PTMs carried by each proteoform are compared using the modification type and the modification site on the protein sequence. For 2 PTMs to match, their modification type as defined by the PSI-MOD ontology [18] needs to be identical and

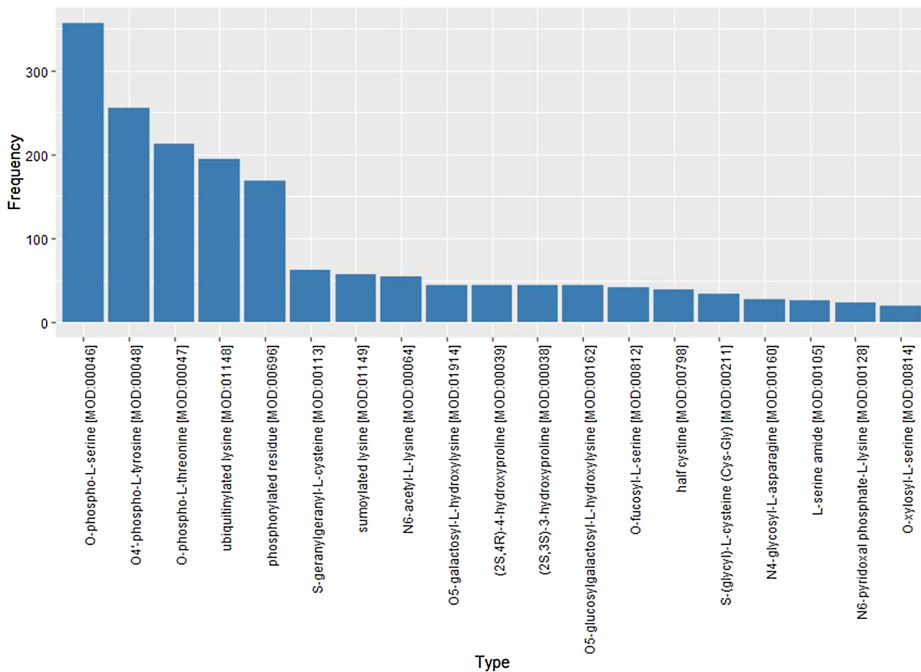


Figure 10. Prevalence of the different PTM annotations in Reactome. PTM labels are extracted from the Reactome database and the number of proteins annotated with the PTM is displayed for each label. If a protein is carrying multiple instances of the PTM, the PTM is counted only once.

Table 3. Posttranslational modification coordinates criteria for comparison

Input	Reference	Margin	Matched	Comment
17	17	0	Yes	Equal
16	17	0	No	Out of margin
7	13	5	No	Out of margin
8	13	5	Yes	In margin
19	13	5	No	Out of margin
0	2	5	No	Input in margin, but 0 is not a valid coordinate
-1	2	5	No	Input in margin but negative
?, empty, null	c	k	Yes	Input is less specific
c	?, empty, null, -1	k	Yes	Input is more specific
?, empty, null	?, empty, null, -1	k	Yes	Equally unspecific
Negative int, zero	Any	k	No	Negative or zero input are invalid

This table compares the value of a PTM coordinate of an input Proteoform with the value of a PTM coordinate in a reference proteoform. The letter k represents any positive integer.

the distance between their sites must be below a user-provided margin, as detailed in Table 3.

#### PTM

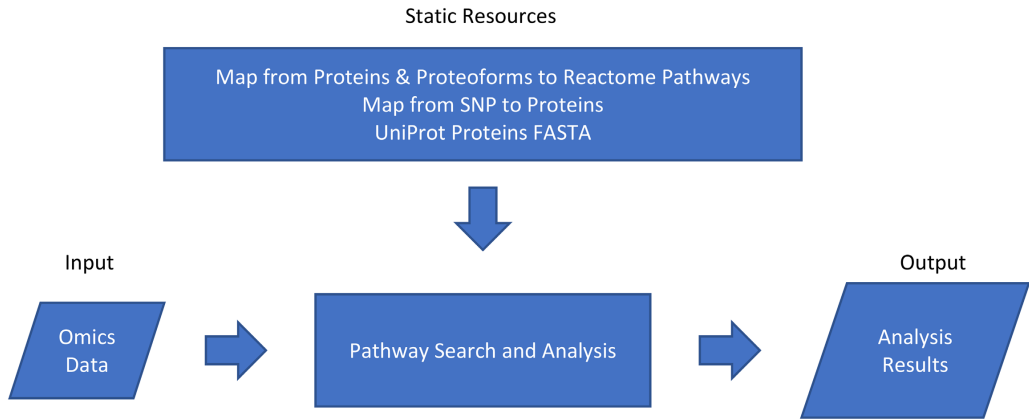
Different matching types are implemented in Pathway-Matcher for the PTM sets:

- **Strict:** the input and reference proteoforms have the same number of PTMs and every PTM of the input proteoform matches a PTM in the reference proteoform.
- **Superset:** every PTM of the reference proteoform matches a PTM of the input proteoform, but some PTMs in the input

proteoform may not match PTMs in the reference proteoform.

- **Subset:** every PTM of the input proteoform matches a PTM of the reference proteoform, but some PTMs of the reference proteoform may not match PTMs in the input proteoform.
- **One:** at least one PTM of the input proteoform matches a PTM of the reference proteoform.

In addition, *Superset without PTM types*, *Subset without PTM types*, and *One without PTM types* are identical to *Superset*, *Subset*, and *One*, respectively, but do not account for modification type in PTM matching. Finally, note that for the *Strict* matching, the



**Figure 11.** PathwayMatcher general overview. The program takes the user input in the form of omics data files and the reference pathways from the database as input. It then executes the search and analysis algorithm to create a resulting list of output files.

PTMs match when their sites are exactly identical and no margin is allowed: either both are the same positive integer or both are null or ?.

For details and examples to run PathwayMatcher with the different matching criteria, see the online documentation ([github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki/Proteoform-matching](https://github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki/Proteoform-matching)).

Additional considerations:

- Negative, zero, or floating-point values are invalid as sequence coordinates in the input.
- The margin to compare the coordinates must be a positive integer.

### Sensitivity analysis

In order to estimate the prevalence of missing annotation in Reactome, we evaluated the matching power of each matching type of PathwayMatcher using a reference list of 116,258 phosphosites obtained from Ochoa et al. [12]. Each phosphosite was transformed into a proteoform, which had the same protein accession and a single PTM at the given site. The PTM accession number 00046, 00047, or 00048 was used if the phosphorylated amino acid reported was a serine, a threonine, or a tyrosine, respectively. Each of the proteoforms with a single phosphorylation was matched against all proteoforms available in Reactome using PathwayMatcher. The share of phosphosites yielding a match for each matching type is available in Table 2.

Subsequently, we evaluated the robustness of each matching type by selecting sets of proteoforms from Reactome, altering them, and matching them back.

First, we selected the proteins that had multiple proteoforms with at least one PTM (1,364 proteins). Then, we gathered all those posttranslationally modified proteoforms and altered them: (1) for the proteoforms with one or more PTMs, the type of the first PTM was replaced by “00000” and modification sites were increased by 5 positions; (2) for the proteoforms with two or more PTMs, the site of the second PTM was moved as well.

Then, we took ten samples of 300 altered proteoforms and matched them to proteoforms in Reactome using PathwayMatcher. For each matching type, we calculated the percentage

proteoforms in the sample that matched any proteoform in the database.

The results for all ten samples are shown in Fig. 7, where we split the matching of the original sample proteoforms and other candidate proteoforms.

### Mapping omics data to pathways

The input is mapped to proteins or proteoforms to find the reactions where the input entities are participants (Fig. 11). The input is mapped to proteins when data types without PTMs or specific translation products are specified; otherwise, a mapping to proteoforms is used. When one type of data yields multiple results due to ambiguity (e.g., a SNP or peptide mapping multiple proteins), all the possibilities are included in the search entities.

When a list of SNPs is provided, mapping from the Ensembl Variant Effect Predictor [5] is used to find the possibly affected proteins. When peptides are provided, their sequence is mapped to UniProt protein identifiers [6] using PeptideMapper [7] and possible proteoforms are constructed. When proteins or proteoforms are available, PathwayMatcher maps them to reactions and pathways using data structures embedded in the PathwayMatcher jar file. These data structures are extracted from the Reactome Neo4j graph database [19] and serialized. All mapping files are available in a dedicated repository: [github.com/PathwayAnalysisPlatform/MappingFiles](https://github.com/PathwayAnalysisPlatform/MappingFiles).

In addition, we made it possible for the user to generate new mapping files as detailed in the PathwayMatcher repository ([github.com/PathwayAnalysisPlatform/PathwayMatcher/tree/master/src/main/java/extractor](https://github.com/PathwayAnalysisPlatform/PathwayMatcher/tree/master/src/main/java/extractor)). PathwayMatcher can then be executed with the new set of mapping files as provided by the user.

### Overrepresentation analysis

The matching of each entity to a given pathway is modeled as a Bernoulli trial with two possible outcomes: success or failure, depending on whether the protein or proteoform is a participant of a reaction in the pathway. Trials are considered independent from each other, meaning that the outcome of previous trials does not affect the next. Finally, the probability of success is cal-

culated by the proportion of choosing a protein in a pathway over the total number of possible proteins, and therefore the probability is constant over all trials.

First, we search all the input entities (proteins or proteoforms) across all the pathways and count how many of them were found in each pathway. The number of entities found in a pathway is taken as the number of successful trials. Then, with the binomial probability distribution, we calculate how likely it would be to get a result equal to or more extreme than the current result (the same number or more proteins or proteoforms in the pathway), given that the input (proteins or proteoforms) was randomly selected [9].

This is done using the cumulative distribution function for the binomial distribution, which calculates the probability of getting at most  $k$  successes out of  $n$  trials, with a probability  $p \in [0,1]$ , where  $X$  is a random variable following the binomial distribution, as detailed in Equation 1.

$$F(k, n, p) = \Pr(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i} \quad (1)$$

For each pathway,  $p$  is set to the ratio between the number of total proteins or proteoforms in the pathway and the total possible entities in the database,  $n$  is the number of proteins or proteoforms in the input sample,  $k$  is the number of proteins successfully mapped in the pathway, and  $X$  is the number of entities found in the current pathway after the search.

Finally, given that the  $P$ -value requires the calculation of the probability of an equal or more extreme result, we use the complement of Equation 1 to calculate the probability of getting at least  $k$  successful trials out of  $n$ , as stated in Equation 2.

$$\Pr(X \geq k) = 1 - \Pr(X \leq k-1) \quad (2)$$

The calculations for proteins or proteoforms are similar but are performed separately depending on the input. If the input consists of protein accessions, the number of participants is calculated by only considering proteins. On the other hand, for the proteoform input, the number of entities in the pathways and the database are the participant proteoforms.

## Performance benchmark

The performance of PathwayMatcher was evaluated using data sets of different sizes obtained from sampling publicly available resources:

- Proteins: human complement of the UniProtKB/Swiss-Prot database (release 2017.10)
- Peptides: ProteomeTools [20] as available in PRIDE [21], data set PXD004732, release date January 23, 2017
- Genetic variants: variants from the human assembly GRCh37.p13
- Proteoforms: annotated proteoforms in Reactome Graph database version 62

Performance testing was done using a standard desktop computer (Intel® Core™ i7-6600U CPU @ 2.60 GHz with 2 cores using 64-bit Windows 10 with Java SE 1.8.0.144 on SSD). Details and code are available at [github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki/Performance](https://github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki/Performance).

## Metrics and figures

The metrics presented in this article were obtained by querying the Reactome graph database directly [22]. The queries used can be found in the online documentation at [github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki/queries](https://github.com/PathwayAnalysisPlatform/PathwayMatcher/wiki/queries).

The figures in this article were built in R version 3.4.1 (2017-06-30)—“Single Candle” (r-project.org) using the following packages: ggplot2, ggrepel, igraph, scico, grid, purr, dplyr, graphlayouts, and gtable. The R scripts used to build the figures are available in the tool repository at [github.com/PathwayAnalysisPlatform/PathwayMatcher.Publication/tree/master/R](https://github.com/PathwayAnalysisPlatform/PathwayMatcher.Publication/tree/master/R).

## Availability of supporting source code and requirements

**Project name:** PathwayMatcher

**Project home page:** [github.com/PathwayAnalysisPlatform/PathwayMatcher](https://github.com/PathwayAnalysisPlatform/PathwayMatcher)

**Operating system(s):** Platform independent

**Programming language:** Java

**Other requirements:**

**License:** Apache 2.0

**RRID:** SCR.01 6759

## Availability of Supporting Data

Snapshots of our code and other supporting data are available in the GigaScience repository, GigaDB [23].

## Declarations

### List of abbreviations

PTM: posttranslational modification.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## Funding

LFHS, SJ, PRN, and MV are supported by the European Research Council and the Research Council of Norway. BB, CH, and HB are supported by the Bergen Research Foundation. HB is also supported by the Research Council of Norway. LFHS, SJ, PRN, and MV are supported by the European Research Council and by the Research Council of Norway. This work has been supported by National Institutes of Health BD2K grant (U54 GM114833) and National Human Genome Research Institute at the National Institutes of Health Reactome grant (U41 HG003751).

## Authors' contributions

LFHS did most of the programming, testing, and documentation and wrote the manuscript. BB and CH contributed with programming, testing, documentation, ideas, and manuscript writ-

ing. AF, SJ, and PRN contributed with ideas and manuscript writing. HB contributed with ideas, supervised the work, and wrote the manuscript. HH contributed with project design, manuscript writing, and supervised the work. MV contributed with project design, programming, documentation, and testing; supervised the work; and wrote the manuscript. All authors participated in the preparation of the manuscript.

## Acknowledgments

The authors thank the Reactome curators for their massive curation effort and the guidance in interpreting the annotations. The authors thank the Galaxy community, especially Dr. Björn Grüning, for their indefectible support.

## References

- Smith LM, Kelleher NL; The Consortium for Top Down P. Proteoform: a single term describing protein complexity. *Nat Methods* 2013;**10**(3):186–7.
- Seet BT, Dikic I, Zhou M-M, Pawson T. Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* 2006;**7**:473.
- Menche J, Sharma A, Kitsak M, et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* 2015;**347**:6224.
- Fabregat A, Jupe S, Matthews L, et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res* 2018;**46**(D1):D649–D55.
- McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol* 2016;**17**(1):122.
- The UniProt C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**(D1):D158–D69.
- Kopczynski D, Barsnes H, Njolstad PR, et al. PeptideMapper: efficient and versatile amino acid sequence and tag mapping. *Bioinformatics* 2017;**33**(13):2042–4.
- Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 2005;**4**(10):1419–40.
- García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: state of the art. *Front Physiol* 2015;**6**:383.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol* 1995;**57**:289–300.
- Schaffer LV, Millikin RJ, Miller RM, et al. Identification and quantification of proteoforms by mass Spectrometry. *Proteomics* 2019;**19**(10):1800361.
- Ochoa D, Jarnuczak AF, Gehre M, et al. The functional landscape of the human phosphoproteome. *bioRxiv* 2019; doi:10.1101/541656.
- Aebersold R, Agar JN, Amster IJ, et al. How many human proteoforms are there? *Nat Chem Biol* 2018;**14**:206.
- Grüning B, Dale R, Sjödin A, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 2018;**15**(7):475–6.
- Afgan E, Baker D, Batut B, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 2018;**46**(W1):W537–W44.
- Blankenberg D, Von Kuster G, Bouvier E, et al. Dissemination of scientific software with Galaxy ToolShed. *Genome Biol* 2014;**15**(2):403.
- [https://usegalaxy.eu/?tool\\_id=toolshed.g2.bx.psu.edu%2Frepos%2Fgalaxy%2Freactome\\_pathwaymatcher%2Freactome\\_pathwaymatcher](https://usegalaxy.eu/?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fgalaxy%2Freactome_pathwaymatcher%2Freactome_pathwaymatcher), Accessed July 24, 2019
- Montecchi-Palazzi L, Beavis R, Binz PA, et al. The PSI-MOD community standard for representation of protein modification data. *Nat Biotechnol* 2008;**26**(8):864–6.
- Natale DA, Arighi CN, Blake JA, et al. Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. *Nucleic Acids Res* 2017;**45**(D1):D339–D46.
- Zolg DP, Wilhelm M, Schnatbaum K, et al. Building Proteome-Tools based on a complete synthetic human proteome. *Nat Methods* 2017;**14**:259.
- Vizcaino JA, Csordas A, del-Toro N, et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* 2016;**44**(D1):D447–D56.
- Fabregat A, Korninger F, Viteri G, et al. Reactome graph database: efficient access to complex pathway data. *PLoS Comput Biol* 2018;**14**(1):e1005968.
- Hernández Sánchez LF, Burger B, Horro C, et al. Supporting data for "PathwayMatcher: proteoform-centric network construction enables fine-granularity multi-omics pathway mapping" GigaScience Database 2019. <http://dx.doi.org/10.5524/100621>.





# **Paper III**



# Extending protein interaction networks using proteoforms and small molecules

*Luis Francisco Hernández Sánchez<sup>1,2</sup>, Bram Burger<sup>1,2,3,4</sup>, Rodrigo Alexander Castro Campos<sup>5</sup>, Stefan Johansson<sup>1,2</sup>, Pål Rasmus Njølstad<sup>1,6</sup>, Harald Barsnes<sup>3,4,†</sup>, and Marc Vaudel<sup>1,7,†,\*</sup>*

<sup>1</sup> Center for Diabetes Research, Department of Clinical Science, University of Bergen, Norway

<sup>2</sup> Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway

<sup>3</sup> Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway

<sup>4</sup> Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway

<sup>5</sup> Departamento de Sistemas, Universidad Autónoma Metropolitana Azcapotzalco, Mexico City, Mexico

<sup>6</sup> Department of Pediatrics, Haukeland University Hospital, Bergen, Norway

<sup>7</sup> Department of Genetics and Bioinformatics, Health Data and Digitalization, Norwegian Institute of Public Health, Oslo, Norway

† These authors jointly supervised the work

\* To whom correspondence should be addressed

Biological network analysis is used to interpret modern high-throughput biomedical data sets in terms of biological functions and pathways. However, the results greatly depend on the topological characteristics of the underlying network, commonly composed of nodes representing genes or proteins that are connected by edges when interacting. In this study, we build biological networks accounting for small molecules, protein isoforms and post-translational modifications. We highlight how these change the global structure of the network and how the connectedness of pathway-based networks is altered. Our findings highlight the importance of carefully crafting the networks for network analysis to better represent the reality of biological systems.

**Keywords:** proteoforms, post-translational modification, biological networks, pathways, functional analysis

## BACKGROUND

---

Biological networks are a promising way to interpret modern biomedical data at scale<sup>1</sup>. They allow the study of molecular patterns at both local and global scale, and hence provide a systemic view on molecular processes. For example, by modeling the interactome — the entire collection of biological interactions — Menche *et al.* identified disease modules, and studied their topological properties and pairwise relationships<sup>2</sup>. An overlap between disease modules would then indicate a functional relationship, hinting at shared mechanisms and possible common drug targets.

The fundamental building blocks of a biological network are the interactions between biological entities, with the entities themselves represented by nodes and their interactions by connections<sup>3</sup>. The entire collection of interactions in a biological system is called the interactome. The main participants of the interactome are proteins, represented in biological networks by the name of the gene encoding them. A relationship between proteins can be inferred from multiple sources: text mining, co-expression, physical interaction, or from literature knowledge on the functions of proteins<sup>4</sup>. Such networks have proved to be very useful for understanding biological mechanisms<sup>5-7</sup>. For example, gene network approaches have been used for analyzing functions of genes associated to different types of cancer<sup>8,9</sup>.

Based on a given interactome, network analyses attempt to extract knowledge concerning specific sets of proteins. For example, *guilty by association* procedures assume that proteins colocalizing in the network are functionally related<sup>8</sup>. Similarly, diffusion models estimate the effects of gene alterations towards their neighborhood<sup>10</sup>. By design, such network analysis methods rely heavily on network structural properties such as the number of neighbors per node or the number of connections between groups of nodes<sup>1</sup>. It is then vital to carefully choose what the nodes and connections represent, such that any inference from the network mirrors the reality of biological systems.

In practice, as a result of genetic variation, RNA splicing, and post-translational modification (PTM), a gene can yield many distinct forms of a protein, called proteoforms<sup>11</sup>. For most proteins, the different isoforms of a gene share less than 50 % of interactions<sup>12</sup>. For example, *Bcl-2* has two isoform products Bcl-xl and Bcl-xs resulting from alternative splicing. Bcl-xl, which contains the BH1 and BH2 domains, is responsible for programmed cell death, while Bcl-xs lacks both domains, therefore contributing to the opposite function<sup>13</sup>. One can legitimately anticipate an even higher specificity when including PTMs. However, this information is lost when creating biological networks using gene names as sole descriptor of the protein.

Another source of information lost in the construction of gene-centric networks is the role of small molecules, which play essential roles in biological systems, *e.g.*, metabolites participating as reactants, catalyzer, or inhibitor of reactions. For example, adenosine triphosphate (ATP) and guanosine triphosphate (GTP) are essential metabolites needed as energy sources. ATP hydrolysis provides the energy for protein transport in the mitochondria, for binding and releasing the newly synthesized polypeptide molecules from the *hsp70* chaperone proteins<sup>14</sup>.

Previously, we have demonstrated that it is possible to leverage the rich information contained in the Reactome pathway knowledgebase to refine the representation of biological networks by accounting for proteoform-specificity of biological reactions<sup>15</sup>. Here, we demonstrate how changing the type of node from gene to proteoform influences the structure of the obtained networks. In addition, we study how

the inclusion of small molecules affects the representation of the network. Together, our results show that changing the representation biological networks can help refine the modeling of biological processes, but that the limited information of proteoform-specific interaction still impairs the application of such approaches at scale.

## RESULTS

---

### INCREASED SIZE OF THE INTERACTOME

A recent estimate for the human genome lists approximately 47,000 genes, of which approximately 19,000 are coding for proteins<sup>16</sup>. The estimated number protein products resulting from alternative splicing is around 70,000 isoforms<sup>17</sup>. The total number of functional proteoforms remains unknown, but estimates are in the millions depending on how proteoforms are defined<sup>18</sup>. Changing the representation of a network from a gene-centric to a proteoform-centric paradigm should therefore result in a network several orders of magnitude larger. Based on isoform and post-translational modification information from the Reactome knowledgebase v80 for *Homo Sapiens*, we can represent 14,246 distinct proteoforms participating in 13,806 reactions (see methods for details). These 14,246 proteoforms represent 11,074 proteins linking to 10,976 gene names, making 1.3 proteoforms per gene on average. We constructed a network based on all pathways in Reactome by connecting entities when they participate in the same reaction. Building the network based on proteoforms instead of genes yields 3,270 (+29.8 %) additional nodes and 224,207 (+61.2 %) additional connections. Thus, while the proteoform annotation provides enough information to substantially increase the size of the network, only few proteoforms are annotated functionally.

The genes with the highest number of proteoforms annotated are *UBC*, *H3C1*, and *H3C15*, with 55, 52, and 48 proteoforms respectively, participating in diverse pathways and located in multiple subcellular compartments (**Supplementary Table 1**). *UBC*, for example, has products mostly ubiquitinated or with crosslinks between L-lysine residues and glycine at multiple locations of the sequence, generating a high number of proteoforms representing different combinations of post-translational modifications. *HLA-A* and *HLA-B* are also genes with high numbers of proteoforms, not due to splicing variants or PTMs, but because there are multiple protein accessions linked to them, 36 and 21, respectively. For these examples, the proteoform representation of biological interactions will be completely different compared to a gene-centric network.

Reactome contains also small molecules annotated as participants of human reactions. Extending the gene- and proteoform-centric networks with small molecules increases the number of nodes by 2,057, representing an increase of 18.7 % and 14.4 %, respectively (**Supplementary Table 2**). Adding small molecules creates 85,282 and 91,476 new connections, corresponding to an increase of 23.3 % and 15.5 % for the gene- and proteoform-centric networks, respectively. However, this creates situations where small molecules ubiquitous in biochemical reactions, like H<sub>2</sub>O or ATP, connect most of the network. To take the influence of small molecules into account without distorting the network globally, we introduced the possibility for small molecules to connect pathway participants within but not

between reactions. The number of new connections then becomes 442,004 and 457,127, corresponding to an increase of 120.7 % and 77.4 %, for the gene- and proteoform-centric networks, respectively.

### INTERCONNECTED PROTEOFORMS ALTER THE DEGREE DISTRIBUTION

The connectivity of a node in a network is measured by the number of connections, also called the degree. Without accounting for small molecules, 143,255 (24.3 %) of the connections in the proteoform-centric network represent connections where proteoform-level information is available for both nodes, while 101,907 (17.3 %) and 345,253 (58.5 %) of the connections present no proteoform annotation for one or both nodes, respectively. Since proteoforms are specific forms of a protein<sup>11</sup>, changing a gene-centric network into a proteoform-centric representation can be seen as distributing the protein-protein interactions between new, more specific, nodes. Thus, intuitively, proteoform nodes are expected to have a smaller degree than the gene that encodes them. As we previously described<sup>15</sup>, the majority of proteoforms indeed present a degree lower than their genes in the gene-centric network. This picture is however complicated by proteoform-proteoform interactions and, as detailed in **Figure 1** and **Supplementary Table 3**, at the scale of the entire network, the degree is increased when taking proteoforms into account.

This is for example the case for collagen-related genes such as *COL7A1*, *COL3A1*, and *COL6A3*, which present a much higher degree in the proteoform-centric than in the gene-centric network: 606 vs. 121, 547 vs. 67, and 546 vs. 66, respectively. These collagen nodes are expanded to a wide variety of proteoforms as they become multiply modified by sequential reactions. For example, in the pathway *Collagen biosynthesis* a reaction converts *collagen lysines* to *5-hydroxylysines*, and diverse *COL7A1* gene products are input and output of the reaction. In a gene-centric network, this reaction is modeled as a single *COL7A1* gene node, while in the proteoform-centric network, the input nodes *COL7A1*, *3x4Hyp-COL7A1*, and *3x4Hyp-3Hyp-COL7A1* are connected to the output nodes *5Hyl-COL7A1*, *3x4Hyp-5Hyl-COL7A1*, and *3x4Hyp-3Hyp-5Hyl-COL7A1*, yielding nodes with higher detail of information but also with higher degree than in the gene-centric network. Other nodes that consequently have their degree increased do not necessarily have proteoform-level annotation, such as *PLOD3*, which has its degree increased by an order of magnitude (from 46 to 529), simply because it participates in reactions with multiple collagen gene products, therefore connecting to many proteoforms. The examples of genes with highest increase in degree between gene- and proteoform-centric networks are listed in **Supplementary Table 4**.

To evaluate the local vs. global effect of introducing proteoforms in the network, we evaluated the degree of nodes per pathway (**Supplementary Table 5**). Then, the average degree per proteoform, 14.1, was slightly lower than per gene, 14.3 (-1.4 %). It therefore appears that the increase in degree observed for the whole network is not due to within-reaction or within-pathway connections, but rather between-pathway connections between proteoforms and other proteins. This highlights the importance of between-pathway connections and how picturing canonical pathways as separate entities distorts the reality of the interactome. We further evaluated whether the robustness of the network was altered by introducing proteoforms using a percolation analysis<sup>2</sup>. Both gene- and proteoform-level interactomes showed similar percolation curves, with a slightly better robustness for the proteoform network **Figure 2**.

As detailed in **Supplementary Table 3**, when extending the gene- and proteoform-centric networks with small molecules, the average degree of accessioned entity nodes increased from 66.7 to 72.9 (+9.2 %) and from 82.9 to 88.1 (+6.2 %), respectively. As previously introduced, the ubiquitousness of small molecules however produces hyperconnected nodes with up to 3,473 and 4,141 connections in the gene-centric and proteoform-centric networks, respectively, while the most connected genes and proteoforms present 1,290 and 1,520 connections, respectively. Restricting small molecules to reaction-specific relationships allows considering the local function of small molecules without creating such hyperconnected nodes: the average degree of nodes is increased to 99 and to 109 for the gene- and the proteoform-centric networks, respectively, while the maximal degree increases to 2,361 and 2,376, and the maximal degree of small molecules remains 304 in both networks.

### PROTEOFORMS MODIFY THE LAYOUT OF CONNECTED COMPONENTS

*Connected components* are the maximal subnetworks in which all nodes of the component can reach each other through a path (**Figure 3**). The *Largest Connected Component (LCC)* of a network is the component with the highest number of nodes. In our analysis of Reactome, gene- and proteoform-centric networks showed similar relative size of the LCC (**Supplementary Table 6**). Given the hypothesis that nodes involved in the same biological function are connected in the pathway network, one expects that they should belong to the same connected component. Connected components can further be separated into subnetwork modules based on the topology of the network or based on their association with specific functions or diseases. Functional studies comparing such modules study the overlap between diseases to identify common molecular mechanisms or drug targets, and transfer knowledge of one module to the other. Thereby, nodes shared between biological processes have been suggested to be of particular interest for the study of disease mechanisms and treatment<sup>2</sup>.

The proteoform interactome extends gene nodes into multiple proteoform nodes. Proteoforms resulting from variation of a single gene, called a proteoform family, may participate in disjoint sets of reactions in the network. If gene nodes are represented by multiple proteoforms participating in separate reactions or pathways, the overlap will only be observable at the gene level and not at the proteoform level. In other words, proteoforms from a single gene may be split over different modules and even different connected components. In this case, modules would intersect in the gene-centric representation of the network, but not in the proteoform-centric representation, where the different modules would be disconnected.

We found 497 proteins where at least one proteoform of the family participates in a biochemical reaction where the other members of the family are not involved. Identifying such a proteoform in a sample therefore provides pathway-specific information that is lost in a gene-centric representation, as in that case all reactions and pathways where any of the family members participate become indistinguishable. As an example, the human protein Peroxiredoxin-5 (P30044) has isoforms P30044-1 located at the Mitochondrial Matrix, and P30044-2 in the Cytosol. They differ in sequence, the second one missing the first 52 amino acids, and participate in separate reactions in different subcellular locations: “*PRDX5 reduces peroxynitrite to nitrite using TXN2*” and “*PRDX1,2,5 catalyzes TXN reduced + H2O2 => TXN oxidized + 2H2O*” respectively. In this case, a proteoform-centric module representation would distinguish the mitochondrial from the cytosol reaction, connecting them through the translocation and processing of P30044 into P30044-1, while a gene-centric representation would make both reactions indistinguishable.



## SMALL MOLECULES REDUCE THE PREVALENCE OF ISOLATED COMPONENTS AND NODES

Adding nodes representing small molecules considerably increases the percentage of nodes part of the LCC, from 85% to 98% in proteoform interactomes. Conversely, adding reaction-unique nodes for small molecules, rather than once for the whole interactome, prevents merging connected components when small molecules are the only nodes shared between reactions. By design, the number of connected components using reaction-unique small molecules is then greater than or equal to the number of connected components obtained when using small molecules, as displayed in Table 6, and consequently the LCCs are smaller. At the other end of the scale, some pathways contain proteins performing multiple roles in a reaction but not connected to other proteins, leading to isolated nodes only connected to themselves in the network. This may happen when different isoforms or proteoforms of the same protein participate in the reaction with different roles, resulting in the gene-centric representation being a single node interacting with itself while the proteoform-centric representation would show a module composed of multiple nodes. We found 1,665 and 1,696 isolated nodes for the gene- and proteoform-centric networks, respectively, showing an overall stable number of isolated nodes. For example, the reactions sustaining Vitamin B1 (thiamin) metabolism (**Figure 4**) yield isolated nodes that stay isolated even in the proteoform-centric representation.

Adding small molecules reduces the number of isolated nodes to 164 and 171, respectively. Among the 10,976 accessioned entity nodes, 2,789 (25 %) are connected only through small molecules. When considering 1,119 pathways, 226 displayed less isolated nodes when considering small molecules. Conversely, for 39 pathways there were more isolated nodes when adding small molecules. When the studied network is sparse or with many disconnected nodes, it thus becomes useful to include small molecules. They show indirect ways to reach one gene from another through reactions, yet the relevance of connecting two distant proteins by a small molecule can be questioned. Reaction-unique small molecules provide a balance between reducing the number of isolated nodes while not connecting nodes across different pathways. They allow connecting otherwise isolated nodes through a path that alternates between accessioned entities and reaction-relevant small molecules, while preserving the disconnection of pathways and components.

## ARTICULATION POINTS AND BRIDGES

Articulation points and bridges are respectively nodes and connections that, if removed, break one connected component into two or more components (**Figure 3**). They are thus important members of the network, maintaining the connection between otherwise disconnected clusters of nodes. The higher the prevalence of articulation points and bridges, the less robust the network. We therefore investigated whether the prevalence of bridges and articulation points changed from a gene-centric to a proteoform-centric representation. However, as detailed in **Supplementary Table 8** and **Supplementary Table 9** for articulation points and bridges, respectively, adding proteoform annotation does not substantially change the share of articulation points (from 2.43 % to 2.46 % of nodes). Articulation points in the gene-centric network either stay articulation points in the proteoform-centric network or become more connected due to the multiplicity of proteoform nodes in a proteoform family. Therefore, proteoforms do not yield to more isolated nodes but may create more connected components (**Supplementary Table 6**). This indicates that, although proteoform annotation increases the connectivity in the network, it is mainly through within-component connectivity.

Given the ubiquitous nature of some small molecules, which participate in many pathways across many contexts, and the increased connectivity that they induce, as observed in the previous sections, it can be anticipated that they create new connections between connected components. Indeed, adding small molecules reduced the prevalence of bridges and articulation points. In proteoform-centric networks they reduce from 351 (2,46 %) to 254 (1,56 %). In the network extended with small molecules, 40 % of articulation points were small molecules and 60 % accessioned entities (**Supplementary Table 8**). Conversely, when restricting the role of small molecules to single reactions, the number of bridges was tripled. Thus, adding reaction-specific single molecules improved the connectivity of the network through single reactions, that are biologically more specific, but less robust. Small molecules do not perform biological processes on their own, they need to interact with accessioned entities. Therefore, they are rarely the only shared node between steps of pathways, resulting in being articulation points less frequently. Adding reaction-specific small molecules also has the effect of increasing the percentage of proteoforms that are articulation points and increasing the percentage of bridges going out of small molecule nodes, from 3.55 % to 10.70 % (**Supplementary Table 9**).

We investigated the changes in prevalence and nature of bridges and articulation points at the level of pathways. **Supplementary Table 10** and **Supplementary Table 11** detail the averaged values among all pathways in Reactome considered individually. The share of articulation points considering pathways one by one is slightly higher than when considering the complete interactome, highlighting how interactomes aggregate pathways, overlapping the connections of nodes in different contexts. Once again, accessioned entities are more often articulation points than small molecules, demonstrating their key role in biological processes. Nevertheless, small molecules still represent one third of articulation points. Even per pathway, the tendency of small molecules to reduce the percentage of bridge connections is clear, confirming the important role of small molecules for the connectivity of the network at both local and global levels.

Bridges connect more than twice as often accessioned entities rather than small molecules, and the prevalence increases when studying per pathway than for the complete interactome. This increase can be interpreted as the connections of proteoforms conveying more unique information, whereas small molecules may connect more diverse types of other molecules. Reaction-unique small molecules are expected to be articulation points more frequently than regular small molecules, but no difference was found on average. Reaction-unique small molecules increase the total number of articulation points by increasing the percentage of accessioned entity nodes that are articulation points. This is due to the smaller average node degree of the reaction-unique small molecules, compared to the regular small molecules. Hence, when they connect to an accessioned entity, they may convert that accessioned entity into an articulation point.

## CONCLUSIONS

---

Interaction networks are a useful representation of biological processes, *e.g.*, to study if biological entities are functionally related. We demonstrate multiple ways to build interaction networks from the Reactome pathway knowledgebase, resulting in networks with very different general and local properties. Such differences in network topology are likely to influence the biological interpretation of experimental data. In particular, we explored the possibility of adding proteoforms and small molecules,

which are usually not considered when building interaction networks, despite playing essential roles in biological processes.

We show that extending the representation of proteins using isoforms and post-translational modifications has an impact on the structure of biological networks, but that this information is only available for a subset of the proteins. This results in highly connected interfaces between single proteins with no proteoform annotation and proteoforms from the same family. We also demonstrate how small molecules such as metabolites alter the structure of the networks. Including them can help connecting sparse areas of the network, but it can also result in highly connected nodes with little to no biological relevance. As a compromise, we propose to take advantage of the annotation of pathway databases to restrict the interactions of such molecules.

With this study we further compared the results of topological analyses conducted to the level of the entire network and when considering one pathway at a time. Each approach yielded different results, showing how local and global properties of the network differ. This highlighted how the arbitrary representation of pathways may alter the perception of the connectedness of biological entities, hiding inter-pathway connections. Overall, our results point towards the importance of using the rich information contained in pathway databases to contextualize network analyses while also highlighting the difficulty to provide an unbiased representation of interaction networks, both locally and globally.

## DISCUSSION

---

This study investigates the impact of changing the network representation through the inclusion of proteoforms and small molecules. We base these findings on the Reactome knowledgebase, which contains rich information on biological pathways. Due to the high level of detail on biochemical reactions required to build such networks, functional annotation on proteoforms and small molecules is still scarce. The rapid pace in increase of functional knowledge indicates that such analyses will become increasingly powerful. As the interactome becomes more connected, refining its representation using the rich information available in pathway knowledgebases represents a promising avenue to tease apart densely connected functional regions.

Factors such as analytical challenges, research interest, and literature curation lead to some proteins or pathways to be better annotated than others. The better annotated pathways give a more detailed representation of the biological processes, while understudied proteins or pathways have a much less mature representation or even remain undiscovered. Such biases have a strong influence on the representation of the biological processes involved, and dramatically alter the ability to conduct refined studies such as proteoform-level network analyses. The disparity in biological functional knowledge is a strong limitation of the field, yielding to a network where some processes yield densely connected subnetworks of proteoforms and small molecules, while others are only represented by sparse disconnected gene names – when any information is available at all.

Technologies to identify proteoforms and small molecules are improving constantly but integrating these biological entities in pathways at scale poses numerous challenges. It is therefore important to develop new biological network analysis approaches that can handle the heterogeneity in pathway annotation without losing the rich information gathered by the scientific community. One can envision

that such approaches will be generalizable to hybrid networks combining pathway knowledgebases with interaction networks derived from experiments or text mining.

Constructing an interaction network using refined information like proteoforms or small molecules is even more challenging using multiple sources of data. Functional annotations often refer only to gene or protein accessions<sup>19</sup>, hence overlooking post-translational regulatory mechanisms central to many biological processes. The broad adoption of proteoforms in the representation of biological processes is essential to generalize the approaches presented in this study, and hence allow the refinement of the representation of biological processes, which will eventually provide biomedical researchers with more powerful tools to interpret their data.

## Acknowledgments

---

This work was supported by the Research Council of Norway (project #301178 to M.V.) and by the Bergen Research Foundation (project #BFS2016REK02 to H.B.B. and H.B.).

## Competing interests

---

The authors declare no competing interests.

## METHODS

---

Reference knowledge to conduct the analysis was obtained from the Reactome graph database (version 80). The database dump file ([reactome.org/download-data](https://reactome.org/download-data)) was loaded and run using Neo4j Desktop 1.4.2 to Neo4j Graph Database Manager 4.4.5. The analysis scripts were implemented using Python 3.10.2 organized as Jupyter notebooks. They communicated with the database management system using the Neo4j Python Driver Manual version 4.4.2.

All the code used to construct the networks and replicate the topological analysis is publicly available at the public repository: [github.com/PathwayAnalysisPlatform/ProteoformNetworks](https://github.com/PathwayAnalysisPlatform/ProteoformNetworks)

The Reactome graph database data model is organized as nodes and relationships with properties and labels. The main nodes we used were *Event* nodes, which involve the transformation of input nodes to output nodes in one or multiple steps. We queried for two types of event nodes: *Pathway* and *ReactionLikeEvents*. *ReactionLikeEvents* convert input entities to output entities in one step, while *Pathways* group sets of *ReactionsLikeEvents*. Each event has participant molecules which perform roles of input (reactant), output (product), regulators and catalyzers (enzymes). The data model represents events occurring in sequence by annotating the output of the first event as input of the second event.

Participants of reactions are physical entities, which typically are of two types: accessioned sequences entities (genes, transcripts, or proteins) or small molecules (metabolites, water, etc.). Accessioned sequences stand for those molecules which have a standard identifier for each sequence pattern,

typically nucleotide-based sequences (DNA pieces like genes) or amino acid-based sequences (proteins). Genes are annotated with HUGO gene nomenclature identifiers<sup>16</sup>, while proteins have UniProt<sup>20</sup> accession numbers. Small molecules also have unique identifiers from the Chemical Entities of Biological Interest (ChEBI) database<sup>21</sup>. These refer to the chemical element or compound rather than a sequence molecule.

When this information is available, accessioned sequence participants are annotated with their isoform and the minimal set of post-translational modifications necessary to perform their role in the biological event. Combining the set of modifications and isoform sequence, we built a theoretical proteoform state in which the gene products need to be present to participate in a given reaction. Participants of events may also be entity sets or complexes. Entity sets stand for groups of entities which may be used almost interchangeably with the desired role in the biological event. For example, multiple proteins may interchangeably perform the same role in a reaction, *e.g.*, catalyzing a reaction. Complexes are the conjunction of multiple molecules into a single unit. The members of the complex may be of all other types of participants, *i.e.*, accessioned sequences, proteins, metabolites or even complexes.

We constructed gene- and proteoform-centric interaction network representations of the Pathways in Reactome by taking participating entities of reactions as nodes of the network, as in previous studies<sup>3,15</sup>. For the gene-centric representation, all physical entities associated with a given gene and each of its associated UniProt protein accessions are represented by a single node, *i.e.*, merging all protein products, isoforms and proteoforms into one node. For the proteoform-centric representation, we represent each proteoform with a separate node. We take the associated protein accession, the isoform, and set of post-translational modifications annotated to represent a single proteoform. Then, all physical entities yielding the same isoform with the same sequence modification combinations are represented by a single node. For both the gene- and proteoform-centric networks we constructed two alternative networks which additionally considered the small molecule participants of reactions; the first alternative adds a single node for each small molecule, the second alternative adds a node for each small molecule to every reaction in which the given small molecule participates.

Once nodes are defined, we set a connection between two nodes when they perform a role in the same reaction, such as input and output. To construct a complete interactome we process all pathways with all their respective reactions to obtain their nodes and connections. We do not repeat nodes, but instead aggregate their connections obtained from each pathway. The resulting network contains all annotated genes or proteoforms for humans in the pathway database.

Networks were represented using the Networkx library version 2.7.1 for Python. The library allowed the calculation of size, articulation points, bridges, and connected components. The robustness of the network was calculated through percolation analysis resulting in a percolation curve, which shows the average size of the LCC, called *giant component*, when random nodes or connections are removed. There is usually a point when the size of the LCC collapses rapidly, that represents the percolation threshold, indicating the average size of modules that can be observed in the network<sup>2</sup>.

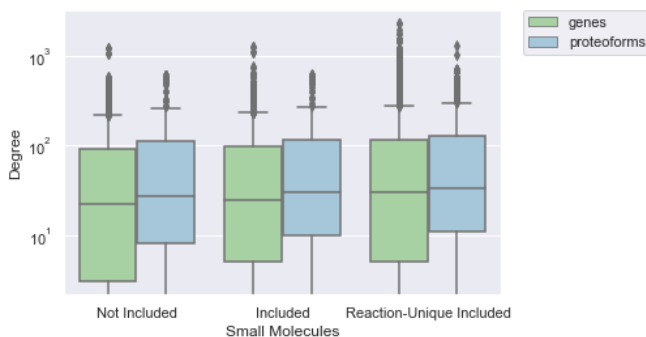
## REFERENCES

---

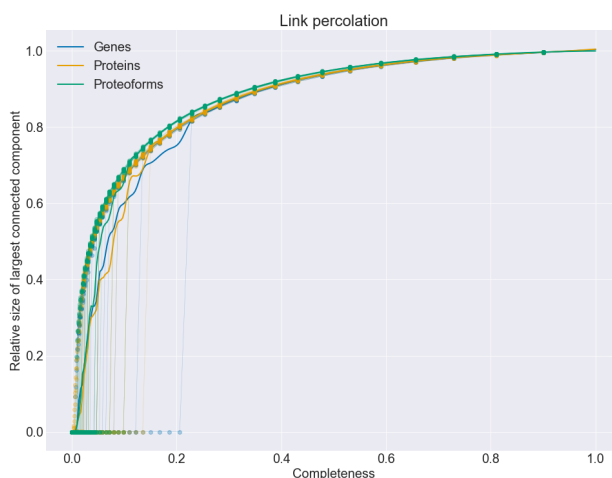
- 1 Sonawane, A. R., Weiss, S. T., Glass, K. & Sharma, A. Network Medicine in the Age of Biomedical Big Data. *Frontiers in Genetics* **10** (2019). <https://doi.org/10.3389/fgene.2019.00294>
- 2 Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015). <https://doi.org/10.1126/science.1257601>
- 3 Burger, B., Hernández Sánchez, L. F., Lereim, R. R., Barsnes, H. & Vaudel, M. Analyzing the Structure of Pathways and Its Influence on the Interpretation of Biomedical Proteomics Data Sets. *Journal of Proteome Research* **17**, 3801-3809 (2018). <https://doi.org/10.1021/acs.jproteome.8b00464>
- 4 Fernández-Tajes, J. *et al.* Developing a network view of type 2 diabetes risk pathways through integration of genetic, genomic and functional data. *Genome Medicine* **11**, 19 (2019). <https://doi.org/10.1186/s13073-019-0628-8>
- 5 Dimitrakopoulos, C. *et al.* Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **34**, 2441-2448 (2018). <https://doi.org/10.1093/bioinformatics/bty148>
- 6 Reyna, M. A. *et al.* Pathway and network analysis of more than 2500 whole cancer genomes. *Nature Communications* **11** (2020). <https://doi.org/10.1038/s41467-020-14367-0>
- 7 Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505-509 (2017). <https://doi.org/10.1038/nature22366>
- 8 Pathway and network analysis of cancer genomes. *Nature Methods* **12**, 615-621 (2015). <https://doi.org/10.1038/nmeth.3440>
- 9 Wu, G. & Stein, L. A network module-based method for identifying cancer prognostic signatures. *Genome Biology* **13**, R112 (2012). <https://doi.org/10.1186/gb-2012-13-12-r112>
- 10 Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology* **18**, 507-522 (2011). <https://doi.org/10.1089/cmb.2010.0265>
- 11 Smith, L. M. *et al.* Proteoform: a single term describing protein complexity. *Nature Methods* **10**, 186-187 (2013). <https://doi.org/10.1038/nmeth.2369>
- 12 Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164**, 805-817 (2016). <https://doi.org/10.1016/j.cell.2016.01.029>
- 13 Schwerk, C. & Schulze-Osthoff, K. Regulation of Apoptosis by Alternative Pre-mRNA Splicing. *Molecular Cell* **19**, 1-13 (2005). <https://doi.org/10.1016/j.molcel.2005.05.026>
- 14 Jinek, M., Chylinski, K. & Fonfara, I. Alberts, B., Johnson, A., Lewis, J. *et al.* (2014). *Molecular Biology of the Cell*, 6e. New York: Garland Science. *An Introduction to Molecular Biotechnology: Fundamentals, Methods and Applications* **27**, 1043-1149 (2020).
- 15 Sánchez, L. F. H. *et al.* PathwayMatcher: proteoform-centric network construction enables fine-granularity multiomics pathway mapping. *GigaScience* **8** (2019). <https://doi.org/10.1093/gigascience/giz088>
- 16 Tweedie, S. *et al.* Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research* **49**, D939-D946 (2020). <https://doi.org/10.1093/nar/gkaa980>
- 17 Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Research* **45**, D635-D642 (2016). <https://doi.org/10.1093/nar/gkw1104>
- 18 Aebersold, R. *et al.* How many human proteoforms are there? *Nature Chemical Biology* **14**, 206-214 (2018). <https://doi.org/10.1038/nchembio.2576>

- 19 Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402-408 (2020). <https://doi.org/10.1038/s41586-020-2188-x>
- 20 Consortium, T. U. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480-D489 (2020). <https://doi.org/10.1093/nar/gkaa1100>
- 21 Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research* **44**, D1214-1219 (2016). <https://doi.org/10.1093/nar/gkv1031>

## FIGURES

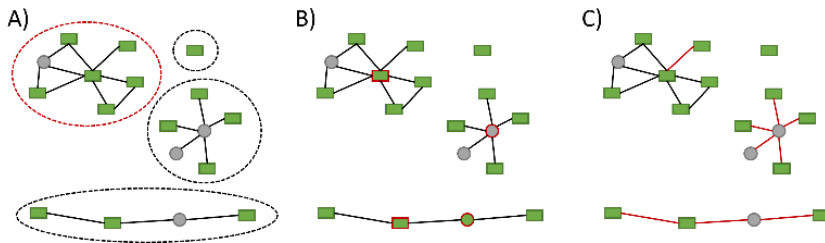


**Figure 1** Node degree distribution in the different interactomes depending on how small molecules are considered. Left: only gene or proteoform nodes and not including small molecules as nodes. Center: small molecules included, one node for each. Right: “reaction-unique” small molecule nodes included, adding one separate node for each reaction where the small molecule participates (“reaction-unique”).

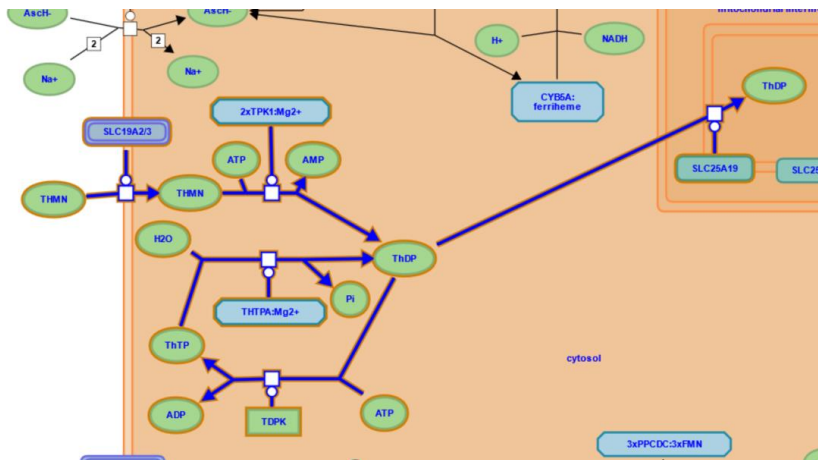


**Figure 2** Approximations of link percolation curves for gene, protein, and proteoform interactome networks without small molecules. Relative size of the LCC (y axis) is the number of nodes with relatively to the original number of nodes in the complete

network. Dots represent measures at each step of the percolation. Soft colored lines connect measures of each replicate, starting with completeness 1.0 (complete interactome) and iteratively removing connections until completeness is 0. Intense lines show average tendencies. Completeness (x axis) is the share of original connections kept after removing random connections.



**Figure 3** Illustration of graph theory concepts using hypothetical networks with proteoform nodes (green rectangles) and small molecule nodes (gray circles). A) *Connected components* of the network, each one surrounded with a dotted line. Largest connected component highlighted with red dotted line. B) *Articulation points*, nodes highlighted with a red border. C) *Bridges*, connections highlighted with red lines.



**Figure 4** Section of Reactome Pathway diagram of "Vitamin B1 (thiamin) metabolism" (R-HSA-196819). White squares represent reactions composing the pathway. Green ovals are small molecules. Green rectangles represent protein molecules. Blue rectangles with chopped corners are complex molecules. Dark blue arrows show relationship between reactants (inputs) and products (outputs) of the reactions following the direction of the arrow from input to output. Molecules connected with a white circle on the arrow represent catalyst molecules.



## SUPPLEMENTARY TABLES

gene	proteoforms	proteins	num proteins	num proteoforms
<b>UBC</b>	{P0CG48,01148:63, P0CG48,01148:239, P0CG48,011...	{P0CG48}	1	55
<b>H3C1</b>	{P68431,00064:null,00083:5, P68431,00064:null,...	{P68431}	1	52
<b>H3C15</b>	{Q71DI3,00085:5, Q71DI3,00083:37, Q71DI3,00085...	{Q71DI3}	1	48
<b>HLA-B</b>	{Q95365,, Q9MY60,, Q04826,, P30466,, P30464,, ...	{P30461, P30495, P18464, P30475, P30486, Q3161...	36	37
<b>HLA-A</b>	{P30450,, P10314,, P16190,, P30455,, P10316,, ...	{P30512, P18462, P04439, P30457, P16188, P3045...	21	23
<b>TP53</b>	{P04637,00046:15,00046:33,00046:46, P04637,011...	{P04637}	1	23
<b>UBB</b>	{P0CG47,01148:11, P0CG47,, P0CG47,00134:76, P0...	{P0CG47}	1	19
<b>RUNX2</b>	{Q13950-1,00046:294,00046:312, Q13950-1,00046...	{Q13950}	1	16
<b>RB1</b>	{P06400,00047:356, P06400,00046:788, P06400,00...	{P06400}	1	16
<b>PPP1R1B</b>	{Q9UD71,00047:75, Q9UD71,00046:102,00046:137,0...	{Q9UD71}	1	16
<b>H4C1</b>	{P62805,00064:17, P62805,00084:21, P62805,0006...	{P62805}	1	15
<b>SHC1</b>	{P29353-1,00048:349,00048:350,00048:427, P2935...	{P29353}	1	15
<b>HLA-C</b>	{P30501,, P30508,, Q07000,, Q29963,, Q95604,, ...	{P30508, P10321, P30510, Q95604, P30504, Q0700...	14	14
<b>H3-3A</b>	{P84243,00047:12,00083:10, P84243,, P84243,000...	{P84243}	1	14
<b>HLA-DRB1</b>	{Q29974,, P13760,, P13761,, Q9TQE0,, P01911,, ...	{Q95IE3, Q30134, P20039, P13760, P04229, Q3016...	13	13
<b>ERBB2</b>	{P04626,00048:1023,00048:1112,00048:1139,00048...	{P04626}	1	13
<b>FGFR2</b>	{P21802,00048:466,00048:586,00048:588,00048:65...	{P21802}	1	13
<b>KRAS</b>	{P01116-2,01116:186, P01116-1,00111:185, P0111...	{P01116}	1	13
<b>COL11A2</b>	{P13942,01914:null, P13942,00037:null,00039:nu...	{P13942}	1	12
<b>COL27A1</b>	{Q8IZC6,00037:null,00039:null, Q8IZC6,00037:nu...	{Q8IZC6}	1	12

**Supplementary Table 1** Top 20 genes with most proteoforms products participating in Reactome pathways. For each gene the number of protein and proteoform products is shown, along with some protein and proteoform examples. Proteins are represented by their UniProt Accessions; Proteoforms by the protein isoform variant and a set of post translational modifications; and each modification as a PSIMOD identifier paired with an integer coordinate indicating its location on the protein sequence. If no localization information is known, "null" replaces the localization.

		Interactions	Nodes
	Entity Level		
<b>Not Included</b>	<b>genes</b>	366208	10976
	<b>proteoforms</b>	590415	14246
<b>Included</b>	<b>genes</b>	451490	13033
	<b>proteoforms</b>	681891	16303
<b>Reaction-Unique Included</b>	<b>genes</b>	808212	40575
	<b>proteoforms</b>	1047542	43845

**Supplementary Table 2** Sizes of six alternative interactome networks resulting from combining entity level (genes, proteoforms) and three options to consider small molecule nodes. Sizes are shown as number of connections (interactions) and number of nodes.

		Q1 AE	Q2 AE	Q3 AE	Q4 AE	Avg AE	Q1 SM	Q2 SM	Q3 SM	Q4 SM	Avg SM
	Entity Level										
<b>Not Included</b>	<b>genes</b>	3.00	22.00	91.00	1,243.00	66.73	0.00	0.00	0.00	0.00	0.00
	<b>proteoforms</b>	4.00	23.00	109.00	1,474.00	82.89	0.00	0.00	0.00	0.00	0.00
<b>Included</b>	<b>genes</b>	5.00	25.00	98.00	1,290.00	72.88	4.00	7.00	20.00	3,473.00	50.09
	<b>proteoforms</b>	6.00	26.00	116.00	1,520.00	88.06	4.00	7.00	20.00	4,141.00	53.10
<b>Reaction-Unique Included</b>	<b>genes</b>	5.00	30.00	115.00	2,361.00	99.23	2.00	4.00	10.00	304.00	17.81
	<b>proteoforms</b>	6.00	30.00	128.00	2,376.00	108.99	2.00	4.00	10.00	304.00	18.32

**Supplementary Table 3** Descriptive summary statistics on the node degree for the different interactome networks resulting from combining entity level (genes, proteoforms) and three options to consider small molecule nodes. Degree values are shown in

separate columns for the two types of nodes: accessioned entities (AE) and small molecules (SM). Columns show quartiles of node degree values separating the lowest 25% as Q1, median at 50% as Q2, value setting the top 75% as Q3 and the maximum value as Q4.

	Proteform Degree	Gene	Gene Degree
<b>Proteform</b>			
Q9UBT2:00211:173	9	UBA2	6
P68431:00064:null,00083:28	10	H3C1	348
Q17RW2:00037:null,00038:null,00039:null	544	COL24A1	66
Q14594-1;	85	NR1I3	52
P39059:00037:null,00038:null	536	COL15A1	67
Q7L5L4:00047:12,00047:35	13	MOB1B	8
P60484:01148:13,01148:289	7	PTEN	34
P56524:01149:559	39	HDAC4	58
Q14155-1;	2	ARHGEF7	142
P24941:00048:15	5	CDK2	158
P50542-1;	65	PEXS	67
P22607:00048:577,00048:647,00048:648,00048:724,00048:760,00048:770	30	FGFR3	91
Q92614:00048:599,00048:768,00048:955,00048:969	27	MYO18A	37
P05161:00134:157	25	ISG15	60
P36957:00127:110	14	DLST	13
Q96J84:00076:14,00078:49,00078:53,00078:370	3	PIWIL1	3
P11216:00128:681	91	PYGB	92
P19419:00046:324,00046:383,00046:389,00046:422,00047:336	8	ELK1	7
Q9BXL7:00046:552,00046:845	2	CARD11	14
Q07092:00038:null,00039:null,01914:null	538	COL16A1	63

**Supplementary Table 4** Degree comparison of nodes from the gene and proteoform interactomes. Each row shows a proteoform with its source gene, both with their respective degree.

		Min. AE	Avg. AE	Max. AE	Min. SM	Avg. SM	Max. SM
<b>Small Molecules</b>	<b>Entity Level</b>						
<b>Not Included</b>	<b>genes</b>	5.12	15.72	24.33	0.00	0.00	0.00
	<b>proteoforms</b>	5.07	17.72	28.16	0.00	0.00	0.00
<b>Included</b>	<b>genes</b>	6.01	17.55	27.31	7.23	12.34	18.58
	<b>proteoforms</b>	5.83	19.40	30.96	7.80	13.56	20.96
<b>Reaction-Unique Included</b>	<b>genes</b>	6.30	20.44	34.27	3.74	8.66	14.63
	<b>proteoforms</b>	6.06	21.91	37.47	3.86	9.12	15.93

**Supplementary Table 5** Descriptive summary statistics on the node degree per pathway. Values for each pathway are taken from six networks resulting from combining entity level (genes, proteoforms) and three options to consider small molecule nodes. Degree values are shown in separate columns for the two types of nodes: accessioned entities (AE) and small molecules (SM). Values refer only to pathways where proteoforms annotated with isoform or post-translational modifications participate.

		Num. CCs	Size of LCC	Relative size of LCC	Average size of CCs	Size of SCC	Num. isolated nodes
Small Molecules	Entity Level						
<b>Not Included</b>	<b>genes</b>	1774	8967	0.82	6.19	1	1665
	<b>proteofoms</b>	1819	12091	0.85	7.83	1	1696
<b>Included</b>	<b>genes</b>	194	12792	0.98	67.18	1	164
	<b>proteofoms</b>	202	16053	0.98	80.71	1	171
<b>Reaction-Unique Included</b>	<b>genes</b>	2445	30153	0.74	16.60	1	1160
	<b>proteofoms</b>	2480	33107	0.76	17.68	1	1167

**Supplementary Table 6** Descriptive summary statistics on the connected components (CCs) for the interactome networks resulting from combining entity level (genes, proteofoms) and three options to consider small molecule nodes. Largest Connected Component (LCC) and Smallest Connected Component (SCC) sizes are evaluated using the number of nodes. Relative size of the LCC represents the fraction of nodes in the complete network that are also members of the LCC.

		Avg. Rel. Size of LCC	Avg. Rel. Size of CCs	Avg. Rel. Size of SCC	Avg. Rel. Number of isolated nodes
Small Molecules	Entity Level				
<b>Not Included</b>	<b>genes</b>	0.91	0.83	0.77	0.08
	<b>proteofoms</b>	0.91	0.81	0.76	0.08
<b>Included</b>	<b>genes</b>	0.83	0.72	0.66	0.20
	<b>proteofoms</b>	0.83	0.71	0.64	0.18
<b>Reaction-Unique Included</b>	<b>genes</b>	0.83	0.71	0.64	0.09
	<b>proteofoms</b>	0.83	0.70	0.62	0.09

**Supplementary Table 7** Descriptive summary statistics on connected components (CCs) per pathway. Values for each pathway are taken from six alternative networks resulting from combining entity level (genes, proteofoms) and three options to consider small molecule nodes. Relative size refers to the fraction of nodes in the complete network that are also members of a connected component. Values are an average of the values per pathway, considering only pathways where proteofoms annotated with isoform or post-translational modifications participate.

		Art. Points	% Art. Points	AE	% AE	SM	% SM
Small Molecules	Entity Level						
<b>Not Included</b>	<b>genes</b>	267	2.43	267	2.43	0	0.00
	<b>proteofoms</b>	351	2.46	351	2.46	0	0.00
<b>Included</b>	<b>genes</b>	244	1.87	143	1.10	101	0.77
	<b>proteofoms</b>	254	1.56	151	0.93	103	0.63
<b>Reaction-Unique Included</b>	<b>genes</b>	2012	4.96	1600	3.94	412	1.02
	<b>proteofoms</b>	2050	4.68	1638	3.74	412	0.94

**Supplementary Table 8** Descriptive summary statistics on the prevalence of articulation points for the networks resulting from combining entity level (genes, proteofoms) and three options to consider small molecule nodes. Columns show the total number of articulation points in each network, the percentage of nodes in the network that are articulation points, then by node type: accessioned entities (AE) and small molecules (SM).

		Bridges	% Bridges	% Bridges out of AE	% Bridges out of SM
Small Molecules	Entity Level				
Not Included	genes	511	0.14	5.41	0.00
	proteoforms	582	0.10	4.81	0.00
Included	genes	523	0.12	4.49	3.53
	proteoforms	527	0.08	3.52	3.55
Reaction-Unique Included	genes	3444	0.43	5.87	11.07
	proteoforms	3356	0.32	4.63	10.79

Supplementary Table 9 Descriptive summary statistics on the prevalence of bridges for the networks resulting from combining entity level (genes, proteoforms) and three options to consider small molecule nodes. Columns show the total number of bridges, the percentage of connections in each network that are bridges, the percentage of connections of a node that are bridges, then by node type: accessioned entities (AE) and small molecules (SM).

		Art. Points	% Art. Points	AE	% AE	SM	% SM
Small Molecules	Entity Level						
Not Included	genes	0.74	3.03	0.74	3.03	0.00	0.00
	proteoforms	1.02	3.92	1.02	3.92	0.00	0.00
Included	genes	0.86	2.86	0.58	2.05	0.28	0.80
	proteoforms	0.91	2.65	0.62	1.88	0.29	0.78
Reaction-Unique Included	genes	1.67	3.68	1.45	3.31	0.23	0.37
	proteoforms	1.71	3.42	1.49	3.06	0.23	0.36

Supplementary Table 10 Descriptive summary statistics on the prevalence of articulation points per pathway. Values for each pathway are taken from six networks resulting from combining entity level (genes, proteoforms) and three options to consider small molecule nodes. Values are averaged per pathway, considering only pathways where proteoforms annotated with isoform or post-translational modifications participate. Columns show the total number of articulation points in each network, the percentage of nodes in the network that are articulation points, then by node type: accessioned entities (AE) and small molecules (SM).

		Bridges	% Bridges	% Bridges out of AE	% Bridges out of SM
Small Molecules	Entity Level				
Not Included	genes	2.03	0.10	10.94	0.00
	proteoforms	2.28	0.10	11.14	0.00
Included	genes	1.43	0.04	6.66	2.74
	proteoforms	1.42	0.03	5.43	2.38
Reaction-Unique Included	genes	3.17	0.05	7.72	7.45
	proteoforms	3.08	0.04	6.37	6.91

Supplementary Table 11 Descriptive summary statistics on the prevalence of bridges per pathway. Values of each pathway are taken from six networks resulting from combining entity level (genes, proteoforms) and three options to consider small molecule nodes. Values are averaged per pathway, considering only pathways where proteoforms annotated with isoform or post-translational modifications participate. Columns show the total number of bridges, the percentage of connections in each network that are bridges, the percentage of connections of a node that are bridges, then by node type: accessioned entities (AE) and small molecules (SM).



Graphic design: Communication Division, UIB / Print: Skjipes Kommunikasjon AS



[uib.no](http://uib.no)

ISBN: 9788230866375 (print)  
9788230844878 (PDF)