Bridging the knowledge gap in multigenic pathogenicity predictions

Leveraging integrated networks and machine learning to explain predictions in oligogenic diseases

Renaux, Alexandre

Bridging the Knowledge Gap in Multigenic Pathogenicity Predictions

Leveraging integrated networks and machine learning to explain predictions in oligogenic diseases

Thesis submitted by Alexandre RENAUX

in fulfilment of the requirements of the PhD Degree in Sciences (ULB - “Doctorat en Sciences”) and in Sciences (VUB)
Academic year 2023-2024

Supervisors: Professor Tom LENAERTS (Université Libre de Bruxelles)

Machine Learning Group

and Professor Ann NOWÉ (Vrije Universiteit Brussel)

Artificial Intelligence Lab

Thesis jury:

Gianluca BONTEMPI (Université Libre de Bruxelles, Chair)
Bart BOGAERTS (Vrije Universiteit Brussel, Secretary)
Guillaume SMITS (Université Libre de Bruxelles)
Catharina OLSEN (Vrije Universiteit Brussel)
Catia PESQUITA (University of Lisbon)
Kris LAUKENS (University of Antwerp)

La Région et l'Europe investissent dans votre avenir ! - Het Gewest en Europa investeren in uw toekomst!

This PhD thesis has been conducted and written under the supervision of prof. dr. Tom Lenaerts (Université Libre de Bruxelles) and prof. dr. Ann Nowé (Vrije Universiteit Brussel).

Thesis abstract

English

Recent years have seen significant progress in medical genetics, with enhanced access to sequenced genomic data and the evolution of computational methods for analysing genetic variation. These advances have improved our understanding of Mendelian ’one gene - one phenotype’ genetic models, yet the transition to the study of oligogenic diseases, where a small number of genes are involved, remains a considerable challenge.

With the rising reports of clinical oligogenic cases, resources and machine learning tools have been developed to leverage this data. Nevertheless, despite their high accuracy, these predictors can be viewed as “black-boxes” due to their limited interpretability. This complexity poses challenges for medical professionals in validating these predictions and restricts their understanding of potential underlying disease mechanisms. Our research aimed to tackle these limitations using structured background knowledge to provide additional context to predictions.

Our first key contribution is a web platform designed to allow geneticists to filter and analyse patient-level variant data with predictors designed for oligogenic diseases. This platform allows in-depth exploration of predictions, including feature contribution analyses, gene pathogenicity networks and gene module mapping to biological networks.

Taking a step further towards enhanced explainability, we extended our research to the design of a whitebox machine learning approach that could provide both predictions and meaningful explanations based on background integrated knowledge. Our second contribution is a biological knowledge graph that blends data from known oligogenic diseases with multi-scale biological networks, emphasising the importance of diverse information for understanding oligogenic complexity.

Our third contribution builds on this, presenting an interpretable model based on path semantics between gene pairs. This model, capable of learning and applying rules for oligogenic interactions, offers a novel method for geneticists to validate predictions and hypothesise about the causal mechanisms of oligogenic diseases.

In conclusion, this research shows how background knowledge can enhance the explainability of predictions for pathogenic genetic interactions. Our analysis platform gives geneticists the necessary tools to understand oligogenic predictions more effectively. Additionally, our biological knowledge graph opens new avenues for investigating the intricate relationships within oligogenic diseases. Finally, using our rule-based approach in tandem with these resources can improve prediction validation and facilitate the generation of mechanistic hypotheses, advancing our understanding of oligogenic diseases.

Français

Au cours des dernières années, des progrès notables ont été réalisés dans le domaine de la génétique médicale, facilités par un accès accru aux données de séquençage génomique et l'évolution de méthodes computationnelles d'analyse de variants génétiques. Ces progrès ont amélioré notre compréhension des maladies génétiques dites monogéniques. Toutefois, les maladies oligogéniques, qui résultent de l'interaction d'un nombre limité de gènes, continuent de poser un défi majeur en termes d'analyse et d'interprétation.

De nombreuses méthodes d'apprentissage automatique ont été développées pour prédire ces interactions génétiques. Cependant, bien qu'elles soient précises, elles génèrent souvent des résultats difficiles à interpréter pour les généticiens. La complexité de ces prédictions rend ardue la compréhension des mécanismes moléculaires en jeu, ainsi que la validation des résultats obtenus. Pour remédier à ce défi, nous avons exploré différentes approches qui exploitent les réseaux de connaissances biomédicales pour améliorer l'interprétabilité des approches prédictives dans ce domaine.

L'une des contributions de cette thèse est la conception d'une plateforme destinée aux généticiens permettant la prédiction et l'analyse de combinaisons de variants génétiques pour un patient. Elle offre notamment une analyse en réseaux des gènes pathogéniques, et met en lumière des connections entre modules de gènes et réseaux biologiques.

Pour aller plus loin dans la compréhension de ces relations, nous avons conçu un graphe de connaissances biologiques intégrant des données relatives aux maladies oligogéniques avec de multiples réseaux biologiques et ontologies biomédicales. Ce graphe hétérogène permet l'exploration des interactions complexes entre gènes et constitue une nouvelle ressource pour le développement de nouvelles méthodes. Sur cette base, nous avons élaboré un modèle d'apprentissage automatique interprétable. Il tire parti du graphe de connaissance pour identifier des interactions pathogéniques, tout en expliquant ses prédictions. Cet outil fournit aux généticiens une nouvelle approche pour valider les prédictions et élaborer des hypothèses sur les mécanismes sous-jacents des maladies oligogéniques.

En conclusion, notre approche combinée d'apprentissage automatique et d'exploitation de réseaux de connaissances offre des outils prédictifs plus interprétables pour l'étude des maladies oligogéniques, contribuant ainsi à la recherche en génétique médicale.

Nederlands

De laatste jaren is er aanzienlijke vooruitgang geboekt in de medische genetica, met verbeterde toegang tot genetische data en de ontwikkeling van nieuwe analysemethoden. Deze vooruitgang leidde tot verschillende ontdekkingen in monogenetische ziekten, maar het vermogen om meer complexe ziekten met varianten in twee of meer genen (oligogenetische ziekten) aan te pakken, bleef beperkt.

De toename in klinische oligogenetische casussen heeft toegelaten om deze gegevens te verzamelen en intelligente methoden te ontwikkelen die deze gegevens benutten. Ondanks hun hoge nauwkeurigheid, voorzien deze methoden weinig inzage in redenen achter de voorspellingen. Deze beperkte interpreteerbaarheid vormt een uitdaging voor medische professionals aangezien deze methoden weinig inzicht in de specifieke biologische entiteiten en betrokken relaties geven die licht kunnen werpen op onderliggende ziektemechanismen.

Deze thesis is gericht op het aanpakken van deze beperkingen met behulp van gestructureerde achtergrondkennis in de vorm van biologische netwerken en kennisgrafieken, die extra context bieden voor voorspellingen.

Onze eerste belangrijke bijdrage is de creatie van een platform dat oligogenetische voorspelling annoteert met de achterliggende biologische en cellulaire kennis. Door dit platform is een diepgaande verkenning van de voorspellingen mogelijk, inclusief analyses van kenmerkbijdragen, gen-pathogeniteitsnetwerken en het in kaart brengen van genenmodules in biologische netwerken. Dit hulpmiddel wordt ondertussen regelmatig gebruikt door internationale genetische experten in de analyse van verschillende zeldzame aandoeningen.

Om verder dit interpreteerbaarheidsprobleem op te lossen, werd dit thesisonderzoek uitgebreid naar het ontwerp van een whitebox machine learning-benadering die zowel voorspellingen als zinvolle verklaringen kan bieden op basis van geïntegreerde achtergrondkennis.

Onze tweede bijdrage in die context, is het ontwerp van een biologische kennisgrafiek die gegevens van bekende oligogenetische ziekten combineert met biologische netwerken op meerdere moleculaire en cellulaire niveaus, waarbij het belang wordt benadrukt van de diversiteit in informatie voor het begrijpen van de oligogenetische complexiteit.

De derde wetenschappelijke bijdrage bouwt hierop voort en presenteert een interpreteerbare voorspeller op basis van de semantiek van de paden tussen oligogenetische genenparen in de kennisgrafiek. Dit model, dat in staat is om regels betreffende oligogenetische interacties te leren en toe te passen, biedt een nieuwe methode voor genetici om voorspellingen te valideren en hypothesen te formuleren over de causale mechanismen van oligogenetische ziekten.

In conclusie, deze thesis laat zien hoe moleculaire en cellulaire achtergrondkennis, vertaald naar kennisgrafieken, de interpretatie van voorspellingen rond pathogene genetische interacties kan verbeteren en het begrip betreffende oligogenetische ziekten kan vergroten.

Acknowledgments

These doctoral years have been immensely enriching both professionally and personally. Despite the uncertainty, frustration, and disappointments, this PhD journey has allowed me to have a deeper understanding of research and science and to learn invaluable skills and personal lessons. The success of this PhD is not solely my own; it is also attributed to many people who have supported, guided, and inspired me along the way. I would like to take this opportunity to express my gratitude to them.

First and foremost, I would like to express my gratitude to my two thesis supervisors, Prof. Tom Lenaerts and Prof. Ann Nowé, for giving me the opportunity to undertake this PhD. I thank Prof. Tom Lenaerts for his kind guidance, his constant encouragements, and for ensuring I had the necessary resources to complete my journey. I extend my thanks to Prof. Ann Nowé for her helpful feedbacks and oversight throughout this process. I also thank Matthieu Defrance for being a part of my advisory committee.

To my thesis jury for agreeing to evaluate my work. To Prof. Catia Pesquita and Prof. Kris Laukens for taking the time to critically review my work. To Prof. Gianluca Bontempi for his advice and constructive feedbacks. To Prof. Guillaume Smits and Prof. Bart Bogaerts for being part of my advisory committee and closely following my research. And to Dr. Catharina Olsen for her positive remarks which, along with Prof. Guillaume Smits, highlighted the enthusiasm my research could spark in clinical research.

I thank Prof. Michael Cochez and Prof. Ilaria Tiddi from the Vrije Universiteit Amsterdam for their expertise on knowledge graphs which significantly enhanced my research, and for their guidance. Their remote scientific support and collaboration, especially during the COVID lockdown, helped me stay motivated.

I am grateful to the entire oligogenic team for the collaborative and supportive spirit, the fun times at conferences and all the hangouts. I thank Dr. Sofia Papadimitriou for her invaluable scientific assistance, encouragement, and support during my moments of doubt. I thank Charlotte Nachtegael for her insightful advice, ingenious tool acronym ideas inspired by beer, and her inspiring commitment to science communication. I also thank Barbara Gravel for the rich scientific discussions, her assistance in supervising master's students, and her cheerful attitude. I am grateful to Nassim Versbraegen for his expertise in managing complex computational tasks and his consistent optimism. I thank Emma Verkinderen for her enthusiasm and dedication – she did an excellent job in ensuring the continuity of the ORVAL platform. I thank Simon Boutry for his constructive feedbacks and positivity. I also want to acknowledge and thank Chloé Terwagne and Inas Bosch, who successfully defended their master's theses. Guiding them through their academic journey was a genuine pleasure.

I also thank all my colleagues from IB^2 and the Machine Learning Group for creating a pleasant working environment and organising numerous social events. Special thanks to Elias Fernandez for organising many of these wonderful moments outside the university, Edoardo Giuili for his moral support and positive energy in the lab, and Jelena Grujic for organising events and DataBeers that brought us together. I am also grateful to all members of the PhD and PostDoc Society of ULB, especially the peer-writing circle, for reigniting my writing motivation in the final year. It was a pleasure to be a part of this community and to interact with doctoral students from all faculties of ULB.

Thank you to those who inspired me before I embarked on this PhD journey and helped me learn bioinformatics. Special thanks go to Hélène Dauchel for guiding me to the Bioinformatics Master's program in Rouen, to David Vallenet for mentoring me during my time at Genoscope, and to all the amazing friends and colleagues I made during my master's program, at Genoscope/LABGeM, and at EBI/UniProt.

To my dear friends, whether close or far geographically: Océane for her moral support and passionate conversations, Maria and Rafael for encouraging me to stay motivated and cheering me up, Juliette for her companionship during travels and the fun moments she brought into my life, Victor for lifting my spirit and his encouragements and Loïc for his advice and sharing laughs with me.

My deepest gratitude to Rodrigo, my partner, for his love and support. He provided me with much-needed breaks from my PhD through walks in nature, travels, and culinary discoveries. Even during the most challenging times, his kindness, patience and advice kept me grounded.

Lastly, I thank my entire family for their support. To my parents, who have always encouraged me to forge my own path and instilled in me the essential values of perseverance, curiosity, and openness. To my sister, nieces, and nephew for the moments of joy they brought into my life. And to my grandparents for their kindness and reassurance throughout this PhD journey.

Table of Contents

Acronyms and abbreviations

Abbreviation	Full name
1KGP	1000 Genomes Project
ARBOCK	Association Rule learning Based on Overlapping Connections in Knowledge graphs
AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under the ROC Curve
BOCK	Biological networks and Oligogenic Combinations as a Knowledge graph
BA	Balanced Accuracy
BSR	Blast Score Ratio
CADD	Combined Annotation Dependent Depletion
CAR	Class Association Rule
CNV	Copy Number Variant
CS	Classification Score (VarCoPP)
DAG	Directed Acyclic Graph
DE	Differential Evolution
DE	Digenic Effect
DIDA	Digenic Disease Database
DiGePred	DiGenic Predictor
DS	Decision Set
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GDI	Gene Damage Index
GO	Gene Ontology
GWAS	Genome-Wide Association Studies
HPO	Human Phenotype Ontology

Abbreviation	**Full name**
HTS	High-Throughput Sequencing
IC	Information Content
INDEL	genomic Insertion-Deletion
KG	Knowledge Graph
KGE	Knowledge Graph Embedding
KR	Knowledge Representation
LoF	Loss-Of-Function
MAF	Minor Allele Frequency
MCC	Matthew's correlation coefficient
ML	Machine Learning
MP	Metapath
MVC	Model-View-Controller
OLIDA	Oligogenic Disease Database
ORVAL	Oligogenic Resource for Variant Analysis
OWL	Web Ontology Language
PPI	Protein-Protein Interaction
PRS	Path Reliability Score
RDF	Resource Description Framework
RF	Random Forest
RVIS	Residual Variant Intolerance Score
RWR	Random Walk with Restart
SD	Standard Deviation
SNP	Single-Nucleotide Polymorphism
SNV	Single-Nucleotide Variant
SS	Support Score (VarCoPP)
TP	True Negative
TP	True Positive
TPR	True Positive Rate
URI	Uniform Resource Identifier

Abbreviation	Full name
UUID	Universally Unique Identifier
VarCoPP	Variant Combination Pathogenicity Predictor
VCF	Variant Call Format

Notations

Symbol	Definition
	General machine-learning context
ϵ	Residual: difference between model predictions ($f(X)$) and actual observations (Y)
$\mathcal{D}$	Set of training instances or data points
$\mathcal{D}_i$	Single training instance or data point
D	Set of disease-causing gene pairs
D_i	Specific disease-causing gene pair
l	Label or class
l_D	Disease-causing label
N	Set of neutral gene pairs
N_i	Specific neutral gene pair
v	Vector of features
$p(l)$	Likelihood of a data point being assigned the label l
X	Feature space: a matrix of feature values for each data point
Y	Target space: a vector of variables representing the outcomes or target values for each data point
	Rule mining and rule-based learning context
α	Relative importance of the positive instance coverage (Weighted set cover algorithm)
θ	Set of Path Reliability Score (PRS) thresholds for a rule metapaths
$\mathcal{I}$	Collection of items in transactional data
$C_R(\mathcal{D})$	Coverage of a rule set R over training instances $\mathcal{D}$
$C_r(\mathcal{D})$	Coverage of a rule r over training instances $\mathcal{D}$
i	Single item in transactional data
k	Size of an itemset

Symbol	Definition
L	Set of itemsets
$maxlen$	Maximum size of an itemset (Apriori algorithm)
$minsup$	Minimum support of an itemset (Apriori algorithm)
$P_r(D_i)$	Number of paths returned for a pathogenic gene pair when matching rule r
R	Set of association rules
r	Single association rule
$w(R)$	Weight of a rule set R (Weighted set cover algorithm)
$w_m(r)$	Marginal weight of a rule (Weighted set cover algorithm)
X	Antecedent of an association rule
Y	Consequent of an association rule which represents the predicted class or outcome.
Networks, ontologies and knowledge graph context	
$\mathcal{E}$	All entities in a graph
$\mathcal{ET}$	All entity types in a knowledge graph
$\mathcal{G}$	Network or a graph
$\mathcal{K}$	Knowledge graph
$\mathcal{R}$	All relationships in a graph
$\mathcal{RT}$	All relationship types in a knowledge graph
$A(t)$	Set of terms including t and all its ancestors
e	Single entity (or node) in a graph
E_j	Entity type in a knowledge graph
$FI(t)$	Functional information for a term t
G_S	Source gene
G_T	Target gene
$g(t)$	Set of genes associated to a term t and its descendants
$g(T)$	Set of genes associated with all terms of type T
(h, r, t)	Triple in a graph: head, relation, tail
$I(t)$	Set of instances associated with the term t and its descendants

Symbol	**Definition**
M	Metapath (type of path in a knowledge graph)
M_i	Single metapath
$M(D_i)$	Set of metapaths for a given pathogenic gene pair
$N(u)$	Set of neighbours of a node
$p(t)$	Probability of encountering an annotation to the term t or its descendants for a given corpus of annotations
P	Path in a network
$\|P\|$	Path length computed as the number of edges
$PRS(P)$	Path reliability score for a specific knowledge graph path P
R_i	Relationship type in a knowledge graph
r	Relationship (or edge) in a graph
T	All term of a specific ontology (or all semantically defined entities of a specific type in a knowledge graph)
t	A term in an ontology (or a semantically defined entity in a knowledge graph)
$T(g)$	All terms of a specific ontology T associated to a gene g
t_s	Subterm or descendant term
$w(e)$	Edge-associated weight

Chapter 1

Introduction

The field of medical genetics has been witnessing remarkable advancements, largely facilitated by access to genomic data and computational methods to analyse this data. This has led to an enhanced understanding of Mendelian genetics, where a single gene corresponds to a single phenotype. However, the transition to the study of oligogenic diseases, which involve a small number of genes, has been challenging.

Machine learning tools developed to predict pathogenic gene interactions in the context of oligogenic diseases have demonstrated impressive accuracy. Yet, they often lack interpretability, limiting their applicability in clinical settings where predictions need to be assessed against established knowledge and validated evidence.

This thesis presents new approaches to address these limitations by combining structured background knowledge in the form of biological networks and knowledge graphs with predictive methods. This not only provides an enriched context to genetic predictions but also aids in unveiling potential mechanisms driving diseases.

The introductory chapter begins by discussing the fundamentals of human genetics and genome structure (Section 1.1). We then introduce the complexities of rare genetic diseases, their characteristics, and the role of genetic variants in these conditions (Section 1.2. Further, we present the oligogenic disease model, highlighting the concept of epistasis and the related resources and predictive approaches (Section 1.3). This is followed by an exploration of the role and challenges of interpretable machine learning in biomedicine (Section 1.4). Lastly, we discuss the power of graph-based knowledge representations in contextualising and interpreting biomedical data (Section 1.5).

1.1 Deciphering the human genome

1.1.1 Genetics and DNA: from discovery to sequencing

Behind all human physiological processes, billions of cells work together to form tissues, adapt to the environment, and manage the intricate systems that keep us alive. Central to these cellular functions is DNA, or deoxyribonucleic acid, the molecule responsible for carrying genetic information and ensuring the continuity of life.

The field of genetics, which studies heredity and variations in inherited traits, traces its origins to Gregor Mendel's groundbreaking experiments in the mid-19th century. These experiments laid the foundation for the essential laws of inheritance. However, it was not until the late 19th century that Friedrich Miescher discovered nucleic acids. These were later conceptualised as DNA by Albrecht Kossel, who also identified their primary components, including the bases adenine (A), thymine (T), guanine (G), and cytosine (C) [1]. Building upon the foundation set by Mendel, Eduard Garrod in 1902 postulated a clear link between inherited traits and specific biochemical processes within cells.

The discovery of DNA's double helix structure in 1953 by James Watson, Francis Crick, Maurice Wilkins and Rosalind Franklin, profoundly transformed our understanding of genetics [2]. This structure highlighted complementary base pairing, a foundational principle ensuring the accurate replication of DNA and led to the elucidation of the genetic code, the sequence of nucleotides that dictates the biological functions and hereditary traits of organisms.

In the late 1950s and early 1960s, the central dogma of molecular biology was established. It outlines the fundamental processes of DNA replication, transcription, and translation. DNA replication ensures the faithful duplication of genetic information [3], while transcription and translation govern the flow of this information from DNA to RNA, and subsequently to protein [4, 5, 6]. Together, these processes underscored the role of DNA as the blueprint of life.

The development of DNA sequencing techniques has revolutionised our ability to read the genetic code. The first widely used method, Sanger sequencing, was developed in the 1970s by Frederick Sanger [7]. It involves generating DNA fragments of varying lengths and detecting specific nucleotides at the end of each fragment to read a sequence.

In the early 21st century, next-generation sequencing (NGS) technologies emerged, offering faster and more cost-effective ways to sequence DNA [8]. Unlike Sanger sequencing, which sequences one DNA fragment at a time, NGS technologies can sequence millions of fragments simultaneously, greatly increasing the speed of data generation.

The Human Genome Project, an international scientific research project with the goal of determining the sequence of nucleotide base pairs that make up human DNA, marked the beginning of the genomic era [9]. This era is characterised by a focus on the genome, the complete set of genes or genetic material present in a cell or organism. The project, completed in 2003, provided the first comprehensive map of the human genome, opening up new avenues for genetic research.

In the years following the completion of the Human Genome Project, the field of genomics has continued to evolve rapidly. The advent of high-throughput sequencing technologies has made it possible to sequence an individual's entire genome in a matter of days, a task that took the Human Genome Project over a decade to accomplish. As a result, individuals can have their entire genomes sequenced, providing valuable insights into various aspects of their genetic makeup.

The vast amount of sequencing data fuels groundbreaking research, allowing scientists to investigate the connections between genetic variations and diseases. Furthermore, it equips healthcare professionals to diagnose genetic disorders, identify disease-causing mutations, and customise treatment plans based on a patient's genetic profile. These innovations pave the way to a personalised medicine [10].

1.1.2 Structure and organisation of the human genome

The human genome is located within the nucleus of our cells and is made up of approximately 3 billion nucleotides. These nucleotides are organised into 23 pairs of chromosomes, with each individual inheriting one set from each parent [11].

This genome is more than just a sequence of nucleotides; it has a complex structure with various regions. These include genes, intergenic regions, enhancers, promoters, and other non-coding elements.

At the heart of the genome are genes or loci. These are the main functional units, encoding products essential for numerous biological processes. The expression of these genes follows a two-step process [6]. First is the transcription, where a DNA strand is copied into RNA. Among the different types of RNA, messenger RNA (mRNA) is a crucial functional molecule. Indeed, in the second step, translation, this mRNA is used to synthesise sequences of amino acids known as peptides, which, after folding, can form functional proteins. It is also worth noting that beyond serving as templates for protein synthesis, certain RNA molecules play regulatory roles, influencing gene expression without being translated into proteins [12].

These proteins are fundamental to the cell's function and maintenance. They serve roles ranging from catalysing biochemical reactions as enzymes, providing structural support, to facilitating intercellular communication and defending against pathogens. It is worth noting that while our genome has around 20,000 genes that encode proteins, many genes encode non-coding RNA molecules which can interact with proteins and regulate various cellular processes [13].

A gene has both coding and non-coding parts. The coding parts, called exons, are joined together to form RNA. The non-coding parts, known as introns, along with other regulatory regions, are either not included in the transcription or are removed before translation. Furthermore, through a process called alternative splicing, different exons can be included or excluded from the final mRNA transcript. This allows a single gene to produce multiple alternative transcripts [14].

The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure, and histone modification in the human genome [15]. These data have enabled the assignment of biochemical functions for 80% of the genome, particularly outside of the well-studied protein-coding regions [15]. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation.

1.1.3 Genetic variations and their implications in human traits

Every individual has a unique genetic makeup due to the presence of genetic variants within their DNA sequence. These genetic variations are essential for the diversity observed within human populations and play an important role in natural selection, evolution, and the development of traits. Variants contribute differently to an individual's fitness *i.e.* the ability to reproduce and survive. While some might reduce fitness and become rare due to negative selection, others that enhance fitness persist through positive selection [16].

To understand these variations, genomes are often compared to a standard reference genome. The differences identified in this comparison are termed variants or mutations. The collection of genetic variants in an individual is called its *genotype*, which significantly influences observable traits or *phenotype*.

Variant types

Genetic variants can vary in size, from single-nucleotide variants (SNVs) and small insertions and deletions (INDELs), which together account for 99.9% of variants, to structural variants (SVs) that affect larger segments of the genome, including multiple genes. These SVs can take various forms, such as large deletions, copy-number variants (CNVs), and large insertions or inversions [17, 18].

Origins and zygosity of variants

Humans are diploid organisms with two sets of chromosomes in their somatic cells. For each position in the DNA sequence, there can be two alleles: one inherited from each parent. The wild-type or major allele represents the most commonly observed trait in the population. In contrast, the less frequent allele is called the variant or mutation [11].

The zygosity of a specific locus in the genome is described as:

- *Heterozygous*: When the two alleles differ, with one being the wild-type and the other a variant.
- *Homozygous*: When both alleles are identical.

- *Hemizygous*: Used for genes on the X and Y chromosomes in males, signifying only one allele's presence.

"De novo" variants arise uniquely in an individual and are not inherited from either parent. These emerge either in a parent's germ cells or in the fertilised egg. They can be especially impactful since they have not undergone evolutionary selection and often have more significant consequences for an individual's fitness than inherited variants [11].

Variant effect and gene inheritance patterns

Variants are also classified according to their effect on the encoded protein sequences and their locations [11]. The effect can be:

- *Silent*, when the change in the DNA sequence within a protein-coding portion of a gene does not affect the sequence of the protein's amino acids. Such mutations usually result from synonymous mutations in the coding part of the genome, but they can also occur in non-coding areas.
- *Missense*, when a non-synonymous mutation within a protein-coding portion of a gene results in the substitution of a different amino acid in the resulting protein. This change can have minor or major effects, depending on the location of the variant within the gene.
- *Nonsense*, when a mutation introduces a stop codon where there was previously a codon specifying an amino acid. This premature stop codon results in the production of a truncated, and likely nonfunctional, protein.
- *Frameshift*, when insertions or deletions (indels) mutations disrupt the gene's reading frame, usually leading to severe effects on protein function.

Variants causing a gene to lose its associated function, usually by severely altering the encoded protein's structure, are termed loss-of-function (LoF) variants [19]. The effect of a variant is influenced by the gene's inheritance pattern. Recessive genes need both mutated alleles for the associated trait to manifest, while dominant genes can show their trait even with only one wild-type allele present. However, certain genes become non-functional with just one loss-of-function variant allele, a phenomenon termed haploinsufficiency [11].

Variant consequences

On a phenotypical level, variants can impact fitness in diverse ways [20]:

- *Neutral*: They neither benefit nor harm the fitness and make up a large proportion of variants.
- *Beneficial*: In certain environments, they provide advantages that enhance an individual's survival and reproductive capacity.
- *Pathogenic or Disease-causing*: These have detrimental effects, often leading to diseases. Their impact can be profound, especially if they affect gene translation, function, or regulation.

Variant frequency in human populations

The Genome Aggregation Database (GnomAD) consortium has reported 229.9 million high-quality genetic variations identified from the genetic sequences of 141,456 humans worldwide [21]. Among these variants, 14.9 million are located in the exome, the regions of the genome known to encode proteins. This large-scale data aggregation has facilitated the definition of variant frequency metrics such as the minor allele frequency (MAF) and the categorisation of variants into rare variants (shared by less than 1% of the population), variants of low frequency (between 1%-5%), and common variants or polymorphisms. The majority of variants in a single genome are common, with just 40,000 to 200,000 variants (14%) in a typical genome having a frequency <0.5% [18].

Projects and resources to link genotype to phenotype

Over the past 25 years, the field has been revolutionised by rapid advances in high-throughput sequencing (HTS) technologies and analytical methods. These advances have enabled the analysis of large patient cohorts and the creation of publicly accessible catalogs of genotypic and phenotypic data. Major community-driven projects such as HapMap [22], the 1000 Genomes Project [18] and the UK10K project [23] have made significant contributions to these advancements. These resources have provided an unprecedented view of human genetic variation, laying the foundation for understanding the genetic basis of human traits and diseases.

1.2 Unraveling the causes of rare genetic disorders

While most of our genetic variations are benign or contribute to our personal traits, a few mutations in our genome can alter molecular functions or biological processes, leading to genetic diseases. The range of genetic disorders is wide, but rare diseases – affecting only a small percentage of the population – often present the most significant diagnostic and treatment challenges.

1.2.1 Characteristics of rare genetic diseases

Genetic diseases are broadly classified into two categories: rare and common genetic diseases. Rare genetic diseases, also known as orphan diseases, are those that affect a small population. As defined by the European Union policies, these are life-threatening or chronic disorders affecting fewer than 5 per 10,000 people [24].

Approximately 10,000 rare diseases have been identified to date, with 80% having a genetic origin. Around 300-400 million people worldwide suffer from a rare disease (1 in 20 people) [25] with approximately 50% of them being children [26].

Conversely, common genetic diseases are those affecting a larger proportion of the population, typically more than 1 in 100 people. These diseases have a complex aetiology, with both genetic and environmental factors playing a role in their development. Heart disease, diabetes, and certain types of cancer serve as examples of common genetic diseases.

Despite their low prevalence, rare diseases constitute a significant health burden due to their overall numbers and often severe impact on patients' lives. Only half of these 10.000 diseases have a resolved genetic etiology [27, 28]. Challenges with diagnosis and treatment, combined with their heterogeneity, contribute to this burden, reinforcing the need for comprehensive genetic research in this area [27, 29].

1.2.2 The contribution of genetic variants to diseases

A key aim of medical genetics is to understand how genetic variants contribute to phenotypic variations that lead to genetic diseases. Phenotypic consequences of genetic variants can range from benign to severe, depending on the function of the gene in which the variant occurs and the nature of the variant itself. Variants leading to a loss of function of a gene often result in more severe phenotypes compared to those that only slightly alter the function of the gene product.

In rare diseases, variants often exert a profound effect on gene function, often manifesting as a loss-of-function. Despite their low frequency in the general population due to their rarity, these variants can profoundly affect individuals who possess them, leading to distinct and often severe phenotypic presentations. In contrast, common diseases arise from an interplay of environmental factors and numerous genetic variants, each contributing modestly to the overall effect [30].

When studying genetic diseases through population statistics, two important concepts in this context are penetrance and expressivity. *Penetrance* refers to the proportion of individuals carrying a specific variant of a gene that also expresses an associated trait. In other words, it is the probability that a genetic variant will result in a disease. *Expressivity*, on the other hand, refers to the degree to which a trait is expressed in an individual. A genetic variant can have complete penetrance but variable expressivity, meaning that all individuals with the variant will show some form of the trait, but the severity or characteristics of the trait can vary significantly among individuals [31]. For example, Huntington's disease is caused by a specific genetic variant that has complete penetrance, meaning that everyone who inherits the variant will eventually develop the disease if they live long enough. However, the age at onset and progression of the disease can vary widely among individuals, demonstrating variable expressivity [32].

Efforts are ongoing to understand the relationship between genetic variants and disease manifestations. ClinVar [33] and OMIM (Online Mendelian Inheritance in Man) [34] are notable databases that aggregate and make publicly available information on the associations between human variations and diseases. For detailed information on rare diseases, Orphanet [35] is a valuable resource gathering information on rare diseases and their associated genetic variants. Additionally, multiple projects such as the Deciphering Developmental Disorders (DDD) [36] and the Undiagnosed Diseases Network [37], aim to diagnose individuals with a focus on unidentified genetic conditions, while making the collected data available to researchers. These combined efforts are enhancing our understanding of genetic diseases, which is critical for improving diagnosis and treatment.

1.2.3 Approaches for pathogenic variant identification

Building on the necessity to understand the phenotypic consequences of genetic variants, various statistical and computational methods have emerged to identify potential disease-associated genetic variants. These approaches leverage different techniques ranging from machine-learning to association studies.

The first category of computational methods encompasses variant pathogenicity predictors [38]. Early approaches, such as SIFT [39], primarily relied on sequence conservation across related proteins to predict the potential deleterious impact of amino acid substitutions. Many contemporary methods employ machine-learning techniques, training on datasets of variants known to be associated with diseases or considered benign. For instance, PolyPhen-2 [40] integrates sequence conservation and structural features in a machine-learning model. CADD [41] further broadens the scope by incorporating a diverse array of features, including genetic, molecular, evolutionary, and structural characteristics, enabling it to score a more extensive variety of variants. The widespread adoption of these tools in both research and clinical contexts underscores illustrates their critical role in modern genetic analysis.

A different approach, variant or gene prioritisation methods, rank variants based on their relevance to a specific disease. These methods often use the "guilt-by-association" principle, assuming that the most probable candidates are linked with genes or other biological entities previously associated with a given disease [42, 43, 44, 45].

While successful in detecting strong monogenic variants, these methods have limited performance in situations where a complex genetic pattern is present, as they are not explicitly developed to identify pathogenic mutations in multiple genes.

In addition, strategies based on Genome Wide Association Studies (GWAS) principles have seen widespread adoption. GWAS compares genetic variants across large cohorts of affected individuals and controls, aiming to identify variants statistically associated with the disease in question. This approach has been effective in identifying the genetic factors contributing to common diseases influenced by multiple genes.

However, the application of GWAS principles to rare diseases poses unique challenges. Primarily, the requirement for large patient cohorts in GWAS is a significant challenge for rare diseases due to the limited access to patient data. Moreover, the high level of statistical noise associated with rare disease data, due to their heterogeneity, can further complicate the analysis. A critical aspect of GWAS is its ability to robustly detect common variants (Minor Allele Frequency, MAF>5%), but it struggles with rare variants due to low statistical power in single-variant association tests [46]. To tackle this issue, recent advancements have introduced methods such as rare-variant collapsing. This strategy involves grouping rare and low-frequency variants at the gene or pathway level, facilitating simultaneous evaluation of the effects of multiple variants [47, 48].

1.2.4 The continuum of genetic diseases

Notwithstanding these advancements, establishing the direct contribution of individual variants for rare human disorders remains challenging, and many patients with suspected rare genetic diseases are left without a definitive diagnosis. Some genetic disorders display unique features: they might have *incomplete penetrance*, where known disease-causing mutations are found in healthy individuals; exhibit *phenotypic variability*, where the range of symptoms cannot be solely attributed to known pathogenic mutations; or show *locus heterogeneity*, with mutations at different genetic locations leading to a similar disease phenotype [26].

One of the early conclusions from GWAS studies was that, for many traits, the most significant genetic associations only explained a minor portion of the genetic variance. This was true even when combining common variants with modest effect sizes. Collectively, these findings indicated limited impacts on overall population variance and predictive power [49]. This gap in understanding is often referred to as the "missing heritability problem" [26].

These difficulties are questioning the traditional Mendelian view of a one-to-one association between a single pathogenic mutation and its phenotypic consequence [50]. Deviations from Mendelian expectations have led to the discovery of more complex genetic underpinnings of disease [51, 52], proving that genetic models involving the interaction between several variants and genes, causing or modulating the development of diseases, need to be considered [53, 54]. Conversely, the polygenic model, involving a contribution of many common variants in multiple genes, has been proved multiple times to be insufficient to explain some genetic disorders [55].

Geneticists have been trying to separate diseases according to their observed pattern of inheritance as:

- *Monogenic diseases*, caused by variants that occur in a single primary gene that accounts for the large amount of penetrance.

- *Oligogenic diseases*, caused by variants in a few genes [56] that can explain the phenotypes of some patients and their unaffected relatives more clearly than the genotypes at one locus alone.

- *Polygenic diseases*, caused by multiple variants in many genes, such as diabetes and coronary heart disease. The distinction between oligogenic and polygenic diseases remains unclear.

- *Complex or multifactorial diseases*, caused by a combination of genetic and environmental factors, such as the lifestyle, dietary habits, exposure to toxins or birth defects.

Studies of human genetic disorders have traditionally followed a reductionist paradigm by separating rare and complex disorders but it has been hypothesised that the cause of genetic diseases form a continuum with unclear boundaries from strong monogenic to multifactorial causes [54] (Figure 1.1).

One emerging hypothesis anticipates that a significant proportion of the 'missing heritability' is attributable to low-frequency variants with intermediate penetrance effects, which have been largely ignored by conventional gene-discovery approaches [49]. This highlights the need of a conceptual bridge between strong monogenic and polygenic models that would better explain some rare diseases.

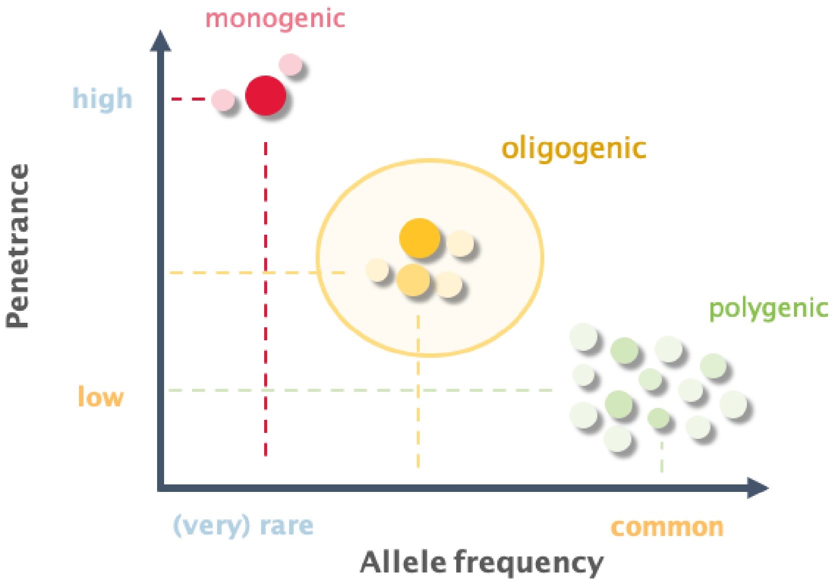


Figure 1.1. Continuum of genetic diseases according to the frequency and penetrance of their causal variants. Monogenic diseases are attributable to very low-frequency genetic variants with large penetrance effects on a single gene. On the other side of the spectrum, polygenic diseases are attributable to many common variants with modest effect size in multiple genes. The oligogenic disease model is a bridge between these extreme disease models, involving the contribution of low frequency variants with intermediate penetrance effects in a few number of genes. Figure adapted from McCarthy, et al. (2008) [49].

1.3 The oligogenic disease model

Oligogenic diseases, those influenced by a few genes, serve as a critical link between monogenic diseases and the complex, polygenic disorders, providing a more comprehensive genetic model to interpret disease aetiology [57].

1.3.1 A shift away from the monogenic model

Over time, evidence for oligogenic patterns of inheritance has emerged for a variety of diseases, including those initially considered monogenic, such as phenylketonuria or hereditary non-syndromic deafness [58, 59]. Even within the same disease, one can find evidence of monogenic inheritance, effects of genetic modifiers, or oligogenic inheritance [57]. For instance, while cystic fibrosis is often considered an exemplary case of monogenic inheritance, it is now known that its phenotypic expression can be significantly modulated by variants in modifier genes [57, 60].

Genetic modifiers play complex roles in disease presentation. They can influence the phenotype established by mutations at the primary disease locus in several ways. This influence can lead to variable expressivity or reduced penetrance of specific disease features. For instance, a modifier can cause earlier disease onset or accelerate progression, alter the clinical presentation to be milder, or even potentiate certain disease features, leading to novel disease forms [56, 61].

Apart from the role of modifiers, disease etiology can also involve interactions between genes, where the combined effect of multiple genes leads to the disease phenotype. An illustrative example is retinitis pigmentosa, a degenerative disorder causing progressive loss of vision caused by mutations in two loci: the photoreceptor-specific genes ROM1 and peripherin/RDS, in which only double heterozygotes develop retinitis pigmentosa [62].

Oligogenic inheritance have also been suggested in several neurodevelopmental disorders, including epileptic encephalopathy, intellectual disability, autism, and schizophrenia, showcasing the complexity of genetic contributions to these conditions [63].

1.3.2 Oligogenic inheritance: From digenic to complex models

The simplest form of oligogenic inheritance, the digenic model, emerges when genotypes at two loci explain the disease phenotype more accurately than a single locus [64].

Numerous examples in the literature demonstrate the significance of digenic models in understanding disease mechanisms. Bardet-Biedl syndrome, characterised by various clinical features such as pigmentary retinal dystrophy, polydactyly, obesity, developmental delay, and renal defects, exemplifies the complex interplay of multiple genes in disease manifestation [65, 66].

Multiple genes have been implicated in conditions leading to hearing loss, and both monogenic and digenic inheritance patterns have been documented. This complexity makes the study of digenic models particularly relevant. For instance, Usher syndrome, a condition marked by a combination of hearing loss and visual impairment, showcases the cooperative effect of two genes in determining the disease phenotype [67].

Digenic inheritance can take various forms. According to the literature, there are three different classes of digenic models [59, 64, 68] (Figure 1.2):

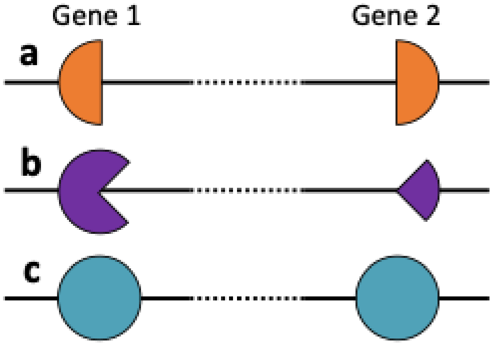


Figure 1.2. The three digenic effect models as described in [59, 64, 68]. (a) True Digenic Effect ; (b) Composite or 'Monogenic + Modifier' Effect ; (c) Dual Molecular Diagnosis Effect.

- *True digenic cases*, where the patient will only manifest the disease when variants on two separate genes are co-inherited.
- *Composite cases or "Monogenic + Modifier"*, where a variant in a primary gene establishes a diagnosis, and a second variant in a modifier gene alters the phenotype.
- *Dual molecular diagnosis cases*, where two separate genes carry variants, each following a classic Mendelian mode of inheritance, and are independently segregated in the same individual, leading to two different diagnoses.

Although most studies focused on this digenic form, it is worth noting that more complex oligogenic disease models involving more than two genes are increasingly being recognised. Such scenarios could involve the synergy of multiple genes, or one primary gene and multiple modifier genes [69, 70, 71].

1.3.3 Molecular mechanisms behind oligogenic disorders

Although there are now several examples of genetically established oligogenic diseases, few of the conducted studies have functionally characterised the genes involved, resulting in a poor understanding of the molecular basis of oligogenicity [53]. The challenging question to answer is how two mutant alleles at two different loci can act in conjunction to cause or exacerbate the same disease phenotype. Since mutations contribute to diseases by affecting the expression, the function, and/or the interaction of gene products, understanding the molecular interactions in cells is thus essential to elucidating these oligogenic mechanisms [57].

In general terms, the effect of mutations on the genetic background in which they occur is known as epistasis or genetic interaction. Fisher defined epistasis as the statistical deviation from the additive combination of two loci in their effects on a phenotype [72]. Since then, the term epistasis has also been adopted in a more mechanistic context, although it is most commonly defined as the deviation from the expected outcome when combining mutations [73]. Epistatic interactions can be both alleviating (i.e leading to a better phenotypic outcome, referred to as positive or antagonistic epistasis) or aggravating (i.e leading to a worst outcome, also called negative or synergistic epistasis) [74].

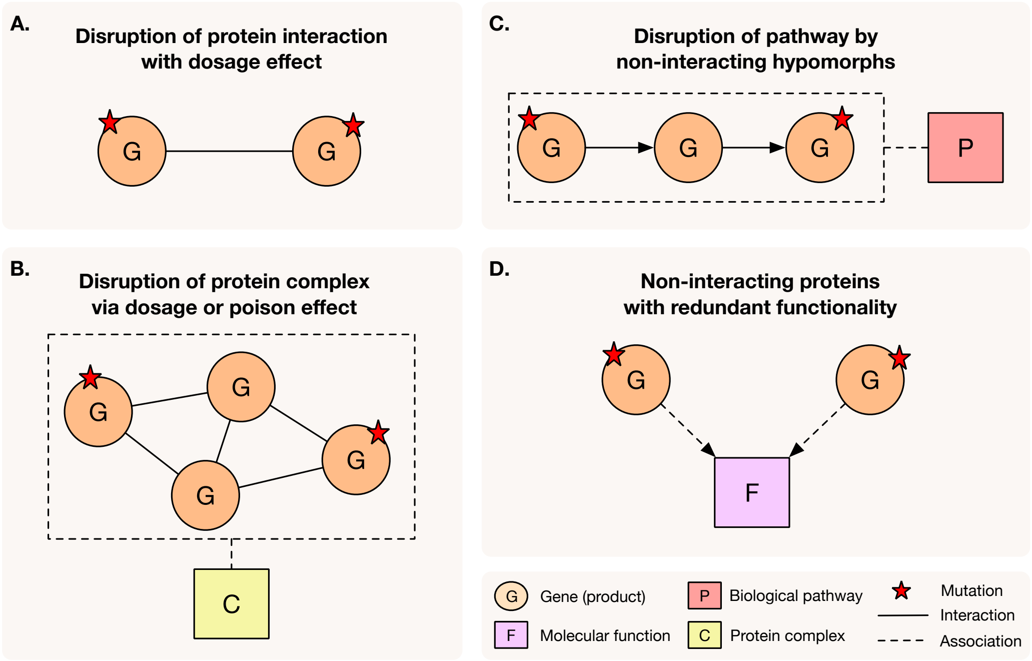

A.
Disruption of protein interaction with dosage effect
G
G
C.
Disruption of pathway by non-interacting hypomorphs
G
G
G
P
B.
Disruption of protein complex via dosage or poison effect
G
G
G
G
C
D.
Non-interacting proteins with redundant functionality
G
G
F
G Gene (product)
P Biological pathway
Mutation
Interaction
Association
F Molecular function
C Protein complex

Figure 1.3. Illustration of some possible mechanisms behind oligogenic diseases. Systematic and simplified categorisation of epistatic mechanisms leading to oligogenic diseases based on [53, 57]. Mutations causing a decrease in dosage (expression or function) or a complete loss-of-function (red stars) affect pairs of genes which cause a functional disruption through different types of interactions. **(A)**: Mutations in two genes coding for interacting proteins prevent their binding and functionality.**(B)**: A decrease in one protein's function, coupled with another's detrimental mutation, prevents the formation of a protein complex or reduce the number of functional complexes. **(C)**: Hypomorphic mutations of non-interacting genes involved in different step of a process lead to an additive loss of signal **(D)**: Mutations in two proteins with similar functions deplete the system's backup capability, leading to a disease state. Note that these mechanisms do not represent an exhaustive list of all possible gene interactions that could cause oligogenic diseases.

We illustrate schematically in Figure 1.3 some possible epistatic mechanisms behind oligogenic disorders based on the models proposed by N. Katsanis, J. Robinson and J.

Badano [53, 57]. Two major types of mechanisms were described:

1. *Nonallelic noncomplementation*, where mutations in two interacting proteins cause or exacerbate the disease phenotype, further subdivided in two models:

   - *The dosage model*: a concomitant decrease in dosage (expression or function) of two different interacting proteins is necessary to cause a pathogenic phenotype.
   - *The poison model*: A mutant protein disrupts or "poisons" a multimeric protein complex to which it normally binds. Although this disruption might not be enough to cause visible pathogenesis on its own, the presence of another mutation in an interacting protein amplifies the dysfunction, leading to a pathogenic phenotype.

   An example of nonallelic noncomplementation can be observed for the genes $ROM1$ and RDS in causing Retinitis Pigmentosa. Both proteins form homodimers, which combine to create tetrameric complexes vital for retinal photoreceptor integrity. Digenic RDS mutations hinder RDS-RDS homocomplex formation, while null $ROM1$ mutations further reduce functional complexes, leading to photoreceptor degeneration [62].

2. *Noninteracting noncomplementation*, where mutations in two non-interacting proteins can also give rise to an oligogenic disorder. This model has been observed in two different scenarii:

   - *Non-interacting hypomorphs*: Each mutated protein contributes quantitatively to the dysfunction of a biological pathway resulting in an additive reduction of its activity beyond a certain threshold that results in a disease phenotype.
   - *Non-interacting proteins with redundant functionality*: the mutated proteins have a redundant function making a certain signaling pathway resilient to the disruption of a single protein. Loss-of-function mutations in both proteins reduce the signal below a critical threshold, resulting in a disease phenotype.

   Examples of noninteracting noncomplementation can be found for the genes $H6PDH$ and $11\beta - HSD1$, which proteins act sequentially in a shared pathway. $H6PDH$

regenerates NADPH in the endoplasmic reticulum, determinant for $11\beta - HSD1$'s activity. Mutations in each protein cumulatively disrupt the pathway, causing a disease when $11\beta - HSD1$ activity drops below a threshold [75]. Other examples include the $Mitf$ and $Tfe3$ genes sharing a similar function as transcription factors for osteoclast development. Individually, their mutation does not cause osteoporosis due to the backup from the other redundant protein. However, the combined loss of the two genes causes severe osteoporosis [76].

Since the reported oligogenic cases are often exhibiting evidence of negative epistasis [77, 78, 79], having a way to detect or predict the involved molecular mechanisms would greatly advance the investigation of digenic disorders and might assist in cases where the effects of genes are not completely understood [53, 78].

1.3.4 A central repository for oligogenic disease information

In past years, studies and clinical reports have highlighted the role of oligogenic inheritance in disease manifestation. This influx of evidence emphasised the need for specialised databases and resources to centralise and organise the data.

The Digenic Disease Database (DIDA) [68] was a pioneering effort in consolidating data on digenic diseases. The database not only cataloged the genes and genetic variants implicated in these diseases but also included associated phenotypic information, inheritance patterns, and relevant literature references.

Building on the success and utility of DIDA, the Oligogenic Disease Database (OLIDA) [71] was established to broaden the scope to include oligogenic diseases. OLIDA offers comprehensive data on the genes, genetic variants, associated phenotypes, and inheritance patterns of these diseases.

Unique to OLIDA is its curation score, a measure of data quality and reliability. This score is determined based on the completeness of the information, the data source, and the strength of the evidence supporting the oligogenic nature of the disease.

The establishment of OLIDA also prompted the development of guidelines for reporting oligogenic cases [80], ensuring data collected is consistent, reliable, and useful for further research.

These databases, particularly OLIDA, are vital for machine learning models. The performance of these models is heavily influenced by the quantity and quality of the data they are trained on. Thus, OLIDA, with its well-curated oligogenic disease data, plays a significant role in the development of robust and accurate machine learning models.

1.3.5 Predicting digenic variant and gene combinations

Building on the foundation laid by these databases, various predictive strategies have been developed. The earliest approaches focused on predicting the type of digenic effect (DE predictor) [81, 82], providing an associated probability for each class. These efforts were followed by the development of the Variant Combination Pathogenicity Predictor (VarCoPP) [83], a machine learning method that can predict potential disease-causing bi-locus variant combinations. These approaches leveraged multi-level features at the variant, gene, and combination level and have demonstrated satisfactory performance in cross-validation settings and on new independent data.

Other techniques have been devised based on digenic case data towards similar objectives. For example, OligoPVP [84] uses phenotype information and the connectivity in a protein-protein interaction network to prioritise disease-causing variant combinations. Another machine-learning approach, DiGePred [85], exclusively uses gene-level features and shared information between genes from functional networks and phenotype data to differentiate disease-causing gene pairs from neutral ones.

Despite their strengths, these predictive methods provide limited explainability to help geneticists understand the rationale behind any given prediction. This is mainly due to the complexity of the machine-learning models employed (*e.g.* large Random Forest in the case of VarCoPP, DE Predictor and DiGePred) and the abstract nature of features they employ, often derived from integrated bioinformatics scores (*e.g.* CADD scores [86], HIPred score [87], ...). Additionally, they provide little to no information about the molecular associations and functional patterns driving the disease, which could indicate compensatory and synergistic mechanisms [88, 89, 90, 91]. Finally, it has been observed that these approaches can produce numerous false positives when analysing patient exomes, due to the large number of variant combinations to be considered.

1.4 Interpretable machine-learning in biomedicine

While all the predictive methods discussed previously hold potential to uncover pathogenic variant and gene combinations, they offer limited explanations of their predictions and are often referred as *black-box*. This limits the ability of domain experts, geneticists and researchers, to validate predictions and better understand the patterns they are based on.

The use of techniques to probe these black-box models, the development of fully interpretable predictors and the choice of more readily interpretable features represent promising solutions in addressing these challenges.

1.4.1 Concepts and techniques of interpretable machine-learning

Introduction of key concepts

Machine learning (ML) has become a predominant aspect of modern bioinformatics, illustrated by methods like CADD [86], DeepVariant [92], and AlphaFold [93]. Fundamentally, supervised ML involves the application of algorithms that discern patterns within datasets, subsequently constructing a model that can generalise these patterns to make predictions on novel, previously unseen data. Given the predictive nature of supervised ML, understanding the decision-making processes of these models, especially in genomics research and the broader biomedical field, has grown in importance.

Interpretability in machine learning, particularly in supervised models, is an essential quality that facilitates the evaluation and understanding of a model's decision-making process and outcomes. It is particularly important for several reasons, including debugging, uncovering biases, gaining trust, ensuring fairness, accountability, and gaining insights [94, 95]. Specifically, it can help machine learning practitioners to identify and address model inaccuracies and biases. For end-users, it offers comprehensible explanations that inform decision-making and may lead to the discovery of new insights [96, 97].

There is a distinction between interpretability and explainability, though they are frequently used interchangeably. *Interpretability* is concerned with understanding the inner mechanics and decisions of a model. In contrast, *explainability* aims to convey these mechanisms in a manner that is understandable for a specific audience [97]. This simplification can be likened to scientific modelling, where approximate models are built to provide a broad understanding of how systems behave, rather than capturing their full behaviour [98].

We can also differentiate two types of explanations: *global* explanations provide an overall understanding of the model's behaviour, while *local* explanations give insights into individual predictions. While global explanations can be useful to debug a model and analyse potential biases, local explanations are particularly useful for domain experts who need to understand the specific reasoning behind each prediction [96, 99].

Models and techniques to interpret predictions

For domain experts focused on specific research questions, the ability to understand individual predictions is essential. This understanding can be facilitated through the use of inherently interpretable machine-learning models or by applying post-hoc interpretability techniques (see Table 1.1 for some examples).

Interpretable models, like logistic regression, decision trees, and rule-based models, have the advantage of being intrinsically transparent. Their decision-making processes are generally straightforward and can be directly inspected and understood. For instance, decision trees split data based on feature values and can be visualised and followed from root to leaf, allowing for a clear path of reasoning [95, 100].

Conversely, model-agnostic or post-hoc interpretability techniques come into play when dealing with models that do not offer inherent transparency. Support Vector Machines (SVM) and Neural Networks (NN), for instance, are complex models whose decision boundaries or internal weights might not directly convey the reasoning behind predictions. In these scenarios, post-hoc techniques, like LIME (Local Interpretable Model-Agnostic Explanations) [101] or SHAP (SHapley Additive exPlanations) [102], can be used to derive interpretable insights or visualisations from these otherwise opaque models [103].

Nevertheless, explanations provided by post-hoc interpretability techniques are often not reliable, and can be misleading for the end-user. The key issue of methods such as SHAP and LIME, is that they show additive local representations, while complex models are usually non-additive. Therefore, these methods often do not include all nuances of a model, such as interactions, and therefore turn out to be too imprecise and misleading [104]. Some researchers have advocated for directly using models that are inherently interpretable, as they provide their own explanations, which are faithful to what the model actually computes [95].

Finally, the interpretability of a model's predictions is intrinsically tied to the interpretability of the features it uses. Even the most interpretable model cannot provide insights if the features themselves are difficult to understand [101]. Providing explanations based on this type of features might lead to unintuitive, or even misleading results [98]. It is therefore important, especially in the biomedical field, to use features that are readily understandable by domain experts or to find good approximations of these features when providing explanations.

1.4.2 Applications and challenges in biomedical genomics

With the growing availability of large datasets and advanced sequencing technologies, machine learning has become an important tool in biomedical genomics for both research and diagnosis. However, alongside its utility in identifying complex patterns in data, the challenge of providing interpretable results remains a significant concern [112].

Specific requirements in the field

The level and type of explainability required in the biomedical field might differ between a clinician, a researcher or even a patient. Tailoring explanations to the specific needs of the end-user can be challenging, but we can distinguish several aspects that are particularly relevant in the biomedical field, especially in medical genomics.

Firstly, ethical considerations are primordial in the biomedical field in general. Machine-learning methods should aim for transparency and accountability in medical decision-making, to ensure that medical decisions or research directions are considered based on carefully evaluated and justified criteria, that can be traced back, validated or corrected by human experts [113].

Additionally, when it comes to uncovering genomics data, geneticists and researchers often seek explanations that provide mechanistic insights, relating the predictions to un-

Type	Technique	Advantages	Limitations
Interpretable Models	Logistic regression [105]	Interpretable coefficients, output probability	Linear decision boundary
	Decision Tree [100]	Captures non-linearity, interpretable structure	Overfitting, instability
	Rule-Based Models [106]	Clear decision-making	Complexity, may miss complex patterns
	k Nearest Neighbor [107]	Simplicity, non-parametric	Sensitive to noise and high-dimensionality
	Naive Bayes Classifier [108]	Robust, return probability estimates	Strong feature independence assumption
Post-hoc interpretability techniques	Local Surrogate (LIME) [101]	Local fidelity, sparse explanations	Approximation, unstable
	Counterfactual Explanations [109]	Actionable insights, model-agnostic	Requires realistic counterfactuals, computation cost
	Shapley Values (SHAP) [102]	Unified measure of feature importance	Challenging interpretation, computation cost
	Anchors [110]	Local fidelity	Requires sufficient data, computation cost
	Individual Conditional Expectation (ICE) [111]	Local and detailed insights	No interactions, computation cost

Table 1.1. Interpretable models and techniques for explaining classifier predictions. This table presents an overview of machine learning models and post-hoc explanation techniques that are designed to enhance the interpretability of classifier predictions. Each entry is categorised by type, with a brief summary of its key advantages and limitations.

derlying biological causes [114]. Such explanations can, for instance, help them understand the complex relationships between genetics and diseases [114, 115]. These insights are particularly important in the biomedical field as they can help generate testable hypotheses for designing experiments and trials [116].

Application examples

In biomedical genomics, explainable machine learning methods are increasingly receiving attention. One commonly used approach for explaining predictions is feature importance, which linear models employ to identify potential causal variants in Single Nucleotide

Polymorphism (SNP) arrays. Penalised regression techniques such as lasso and elastic net are frequently used for this purpose, especially in Genome-Wide Association Studies (GWAS) [117].

Local linear approximators, especially SHAP, have been applied in biomarker identification. SHAP was for example used to find CpG loci in DNA methylation experiments, predicting variables such as cell type, age, and smoking status [118].

Rule lists, though less common, have seen some significant applications. Using sample compression theory combined with recursive partitioning, rule-based classifiers linking genotypes to phenotypes were developed [119]. Similarly, a rule mining procedure was employed to uncover gene expression patterns in obese subjects from DNA microarray data [120]. In a similar vein, rule-based machine learning was used to analyse gene expression measurements in Autistic Spectrum Disorder individuals, revealing a notable co-predictive mechanism between two genes [121].

Open challenges

Despite the potential benefits of explainable machine learning, several challenges remain across different application fields. One of the main challenges is the trade-off between explanation fidelity and complexity versus approximation and simplicity. While complex explanations may be more accurate, they can also be more difficult for domain experts to understand [95].

Additionally, while it is possible to link prediction to the model features, understanding how users should interpret explanations to make testable hypotheses remains an open challenge [96, 116].

In general, domain expert preferences, and especially those of geneticists and biomedical researchers, are largely understudied. This makes it difficult to design models and explanations that meet the needs of the end users [122]. Therefore, the collaboration between machine-learning practitioners and domain experts is essential to ensure that the features and explanations provided are both meaningful and helpful.

Lastly, while interpretable models can provide valuable insights beyond post-hoc explainability techniques, they often do not perform as well as more complex, less interpretable models [95]. This trade-off between accuracy and interpretability is a significant challenge when analysing large volume of data, which is often the case in the field of biomedical genomics.

1.5 Supporting biomedical discoveries with graph-based knowledge representations

In biomedical research, graph-based knowledge representations such as biological networks, biomedical ontologies, and knowledge graphs, are critical for providing context to complex molecular profiles and predictions. They aid in organising, interpreting, and validating a vast array of information, including molecular signatures associated with diseases. This section will explore the significant contribution of these knowledge representations to biomedical discoveries, focusing on their ability to enhance prediction understanding and validation, and to provide a meaningful context for disease-associated molecular data [123].

1.5.1 Network-based knowledge representations

Definition and types of knowledge representations

Knowledge representations (KR) are formal languages or structures used to encode information or knowledge about the world in a way that a computer can process. In the context of biomedical research, the three prominent forms of KR are biological networks, biomedical ontologies, and knowledge graphs [124, 125] (Figure 1.4).

Biological networks are graphical representations of biological systems, where nodes represent biological entities (*e.g.* genes, proteins, metabolites) and edges, which can be directed or undirected, represent relationships or interactions between these entities (*e.g.* protein-protein interactions, gene regulatory interactions, metabolic reactions) [126]. These networks can be used to model and analyse complex biological systems, and to predict the behaviour of these systems under different conditions.

Biomedical ontologies are structured vocabularies or classification systems that provide

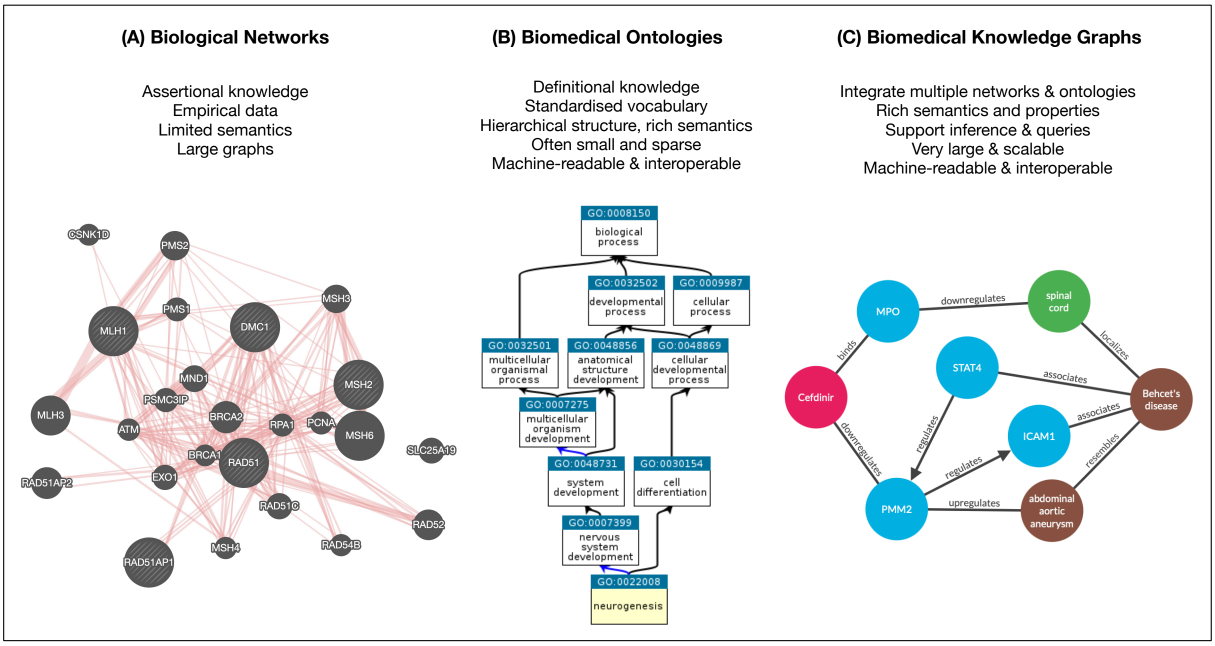

(A) Biological Networks
Assertional knowledge
Empirical data
Limited semantics
Large graphs
CSNK1D
PMS2
MSH3
PMS1
MLH1
DMC1
MND1
MSH2
PSMC3IP
MLH3
BRCA2
ATM
RPA1
PCNA
MSH6
SLC25A19
BRCA1
RAD51
RAD51AP2
EXO1
RAD51C
RAD52
RAD54B
MSH4
RAD51AP1
(B) Biomedical Ontologies
Definitional knowledge
Standardised vocabulary
Hierarchical structure, rich semantics
Often small and sparse
Machine-readable & interoperable
GO:0008150 biological process
GO:0032502 developmental process
GO:0009987 cellular process
GO:0032501 multicellular organismal process
GO:0048856 anatomical structure development
GO:0048869 cellular developmental process
GO:0007275 multicellular organism development
GO:0048731 system development
GO:0030154 cell differentiation
GO:0007399 nervous system development
GO:0022008 neurogenesis
(C) Biomedical Knowledge Graphs
Integrate multiple networks & ontologies
Rich semantics and properties
Support inference & queries
Very large & scalable
Machine-readable & interoperable
MPO
downregulates
spinal cord
binds
localizes
STAT4
associates
Cefdinir
Behcet's disease
associates
regulates
ICAM1
downregulates
regulates
resembles
PMM2
upregulates
abdominal aortic aneurysm

Figure 1.4. Comparative overview of graph-based knowledge representations in biomedicine. This diagram presents three types of knowledge representations commonly used in biomedicine: (A) Biological Networks, illustrated by a Protein-Protein Interaction network from GeneMania based on a selected subset of genes; (B) Biomedical Ontologies, exemplified by an ancestor chart from the Gene Ontology (GO) accessed via QuickGO, with "isA" relationships indicated by black arrows and "partOf" relationships by blue arrows; and (C) Biological Knowledge Graphs, represented by a subset of Hetionet, a comprehensive knowledge graph that integrates biomedical information from 29 prominent bioinformatics networks and ontologies, showing relationships between a compound and a disease. Each section provides key characteristics and an illustrative example of the corresponding knowledge representation type.

a common language for describing and organising knowledge in the biomedical domain [127]. They define concepts (*e.g.* diseases, genes, biological processes) and relationships between these concepts (*e.g. is-a*, *part-of*, *associated-with*) in a systematic way. These concepts are often described in a hierarchical manner which allows for the organization of knowledge from general to specific, enabling efficient data retrieval, integration, and reasoning [127]. Examples of biomedical ontologies include the Gene Ontology (GO) [128], the Human Phenotype Ontology (HPO) [129], and the Disease Ontology (DO) [130].

While these two types of representation can easily be distinguished, *Knowledge graphs* are a broader type of KR that integrate information from multiple sources into a single, unified graph-based model [131]. They represent knowledge as a graph, where nodes represent different types of entities (*e.g.* genes, diseases, drugs) and edges model different kinds of relationships between these entities (*e.g.* gene-disease associations, drug-target interactions). Unlike traditional networks, knowledge graphs excel at consolidating diverse data types and sources and capture the richness of semantics, properties, and contexts associated with each entity and relationship, enabling semantic querying and facilitating the discovery of hidden patterns and indirect associations through transitive relationships [132].

Importance and challenges in the biomedical field

Biomedical research generates vast amounts of data that are often complex and heterogeneous. The integration of prior knowledge represented as biological networks, ontologies, and knowledge graphs, can provide valuable information to guide the analysis and interpretation of this data. These representations can capture known relationships between biological entities, such as protein-protein interactions, gene regulatory networks, and metabolic pathways [123]. Incorporating this prior knowledge into machine learning models can help to identify potential causative factors and mechanisms, increase explainability, and facilitate the generation of new hypotheses [114, 123, 133].

These structured representations offer unique and shared advantages for integrating, organising, querying, exchanging, and visualising diverse biological data. By providing a structured and standardised way to represent and share knowledge, they facilitate interoperability between different databases and tools, and enable the reuse of data and knowledge across different studies and applications [134].

However, the use of knowledge representations also comes with several challenges. These representations can be very dense, making it difficult to filter out the most relevant information. They can also be noisy or incomplete, as they depend on the quality and completeness of the data they are built from [135, 136]. Furthermore, integrating different knowledge representations can be challenging due to differences in their structure, semantics, and coverage [136]. Despite these challenges, the potential benefits of using knowledge representations in biomedical research far outweigh the difficulties, making them an indispensable tool in this field.

1.5.2 A tool for contextualising and interpreting biomedical data

The interpretation of biomedical data is a challenging task, given the complexity of biological systems. Knowledge representations serve as invaluable tools in this process, providing structured frameworks to contextualise experimental data or predictive outputs. This contextualisation, sometimes also referred as functional interpretation, enable the extraction of meaningful insights from complex datasets, enhancing our understanding of underlying biological mechanisms [123].

Network Analysis and visualisation

Biological networks offer a structured representation of biological systems, capturing interactions among genes, proteins, and metabolites. Through network analysis, computational methods are used to study these connections and derive valuable insights [126, 134].

These analyses facilitate the identification of nodes with high connectivity, commonly referred to as "hubs". Such nodes often correspond to genes or proteins that play central roles in multiple biological processes. Conversely, nodes that frequently act as intermediaries in the shortest paths between other nodes, often termed "bottlenecks", may mediate interactions between distinct biological pathways, such as metabolic and signalling pathways [126].

Furthermore, clustering algorithms can be used to identify communities or modules within networks [137]. Here, a community or module is defined as a set of nodes that have more connections among themselves than with nodes outside of the set. These identified sets can correspond to real-world biological structures, such as protein complexes or metabolic pathways, and can provide valuable insights into the organisation of the biological system being studied.

Network visualisation tools enable domain experts to explore the network's structure in detail. They are usually provided with a range of options for filtering, rearranging, and analysing specific regions or modules in the network, facilitating the extraction of biological insights [138].

Enrichment analysis and semantic similarity

From knowledge representations, two main methodologies have emerged to contextualise biomedical data, using the structured information they offer.

Enrichment analysis, on the one hand, identifies over- or under-represented features in datasets, using external biological information to link data to prior knowledge. Its aim is to reveal biological themes, such as processes, cellular components, or disease associations, within gene or protein sets. Gene Set Enrichment Analysis (GSEA) apply this principle to pinpoint gene sets over-represented at the extremes of a ranked list, suggesting coordinated regulation in a given condition [139].

Semantic similarity, on the other hand, involves comparing ontology terms to facilitate the interpretation of complex biological data. By evaluating the functional resemblance between genes or proteins based on their annotations, this methodology uncovers valuable insights within specific biological contexts [140]. This type of analysis has been used for instance to validate gene function predictions, assess the quality of predicted interaction, to align biological pathways or to generate functionally meaningful network subsets [141].

Both methodologies bridge biomedical data and predictions to biological knowledge, enabling researchers to uncover the mechanisms behind observed changes in various biomedical contexts.

Contextualisation and explainability in machine-learning

Contextualisation serves a distinct yet complementary role to interpretability and explainability in machine learning applications. While interpretability focuses on understanding the internal mechanics of a model (*i.e.* how features and parameters interact to produce a decision), explainability aims to make these decisions understandable to end-users, often by relating them to domain-specific knowledge [97, 98].

Contextualisation based on biomedical knowledge aligns more closely with the aim of explainability, as it acts as an external layer that can help in interpreting the model's raw predictions in terms of biological or clinical relevance. For instance, a machine learning model might predict that a particular set of genes is associated with a disease. Contextualisation, would involve mapping these genes onto known biological pathways or existing disease ontologies to provide a richer, more understandable explanation of the prediction. This is particularly useful in biomedical genomics, where the complexity and volume of data can make raw predictions difficult to interpret or validate [112].

However, while contextualisation improves a model's explainability by associating its predictions with domain-specific knowledge, it does not directly offer insights into the model's internal decision-making mechanisms. Unlike interpretability, which aims to elucidate the contributions or interaction of features to a model's predictions, contextualisation functions externally, by providing domain-relevant hypotheses that may diverge from the model's actual decision process unless specific techniques are employed [114].

Leveraging knowledge graphs for predictions and insights

The growing adoption of heterogeneous networks and knowledge graphs (KGs) has been accompanied by an increasing interest in harnessing these integrated networks for predictive methodologies and the extraction of novel insights. KGs, with their semantically-rich and diverse relationships show a great potential in biomedical applications. For instance, integrated networks within KGs have demonstrated superior performance in prioritising novel disease-gene associations [142, 143, 144]. Additionally, KGs offer a valuable resource for mitigating the challenges of data scarcity in the study of rare diseases [145].

However, predictive methods applied to KGs often necessitate the transformation of the original network into a homogeneous structure [146] or its embedding into a latent space [147]. Such transformations can lead to potential information loss and a decrease in interpretability [148].

In response to these challenges, researchers have proposed to use KGs to enhance the explainability of machine learning systems and to provide more coherent explanations [133, 149]. A particularly promising approach involves exploring paths within KGs. Given their structured nature, KGs facilitate the tracing of paths between interrelated entities, offering a rich contextual background for machine learning models. This context, which captures intricate interaction patterns and dependencies, augments the explainability of predictions. Such path exploration is foundational to multi-hop reasoning, a technique which consists in navigating through multiple interconnected entities within a KG to deduce answers [133, 150]. One application of this approach is the PoLo method (Path-Based Reasoning Over Learned Ontologies) [151], which integrates reinforcement learning for KG exploration and has shown efficacy to answer complex questions in the biomedical domain.

Path-centric approaches, especially in biological networks, have proven effective in inferring potential causal mechanisms supporting high-throughput experimental data. Such methodologies can assist in constructing mechanistic models grounded in proteomics profiles and in interpreting transcriptional changes [152, 153]. Applied to KGs, this approach have also shown its effectiveness in generating additional insights. For instance, the RPath method [154] employs reasoning over paths in a KG, using transcriptomic data as a guide, to prioritise drugs for specific diseases. Simultaneously, it identifies targeted proteins along these paths.

In predictive modelling, path information has also been employed for rule inference within KGs [155, 156, 157], facilitating the prediction of novel facts with enhanced explainability [158, 159]. Notably, these methods abstract path information by cataloging sequences of node and edge types, termed metapaths [160, 161]. Metapaths have been employed to generate features from KGs for various classification tasks. For instance, metapath features extracted from Hetionet, a comprehensive biomedical KG, have been employed in gene-disease prioritisation [162] and drug repurposing [163]. These applications underscore the potential of harnessing paths within KGs to enhance both the explainability and effectiveness of machine learning in the biomedical field.

Chapter 2

Research objectives and outline

2.1 Problem definition

Recent advances in genomic sequencing have improved our ability to find and link genetic variations to a variety of human traits. This has led to important progress in understanding the genetics of human diseases and in the field of precision medicine (Section 1.1). Computational and statistical methods have played an important role in identifying variants linked to human genetic disorders.

Despite this progress, the traditional idea that one gene corresponds to one disease, known as the Mendelian model, has come into question. Many diseases show incomplete penetrance, variable expressivity, and locus heterogeneity, adding to the "missing heritability problem" [30] (Section 1.2). As a result, new models like the oligogenic model have gained attention. This model suggests that a disease can be caused by multiple variants across a few genes, offering a more complete view of human genetics (Section 1.3).

The recognition of oligogenic diseases has encouraged an influx of clinical studies presenting evidence of this type of inheritance in various disorders, including diseases traditionally classified as monogenic. This has led to the development of machine-learning approaches to predict the pathogenicity of variant and gene combinations, which have demonstrated their utility. However, a primary limitation of these methods is their lack of interpretability. The complexity of the models, along with their reliance on abstract features – many derived from advanced bioinformatics techniques – complicates the task for geneticists seeking to understand the rationale behind specific predictions. This interpretability gap impedes geneticists and researchers from effectively validating these predictions based on their domain expertise, subsequently reducing their trust in the model's outputs (Section 1.4).

Adding to this challenge is the problem of elucidating the functional relationships and molecular associations underpinning oligogenic diseases. Functional studies and meta-analyses indicate that the origins of oligogenic diseases often stem from intricate, synergistic gene interactions, a phenomenon known as epistasis [53, 74, 81]. These associations can span both direct and long-range interactions across various biological layers. Yet, the specific causal mechanisms for many oligogenic diseases remain elusive. While structured background knowledge, such as biological networks and biomedical ontologies, offers a promising avenue to explore these patterns, their manual exploration can be arduous for geneticists (Section 1.5). Moreover, the diverse nature of epistatic mechanisms presents its own modelling challenges.

Consequently, there is a pressing need to move beyond current methods. We should develop innovative approaches and tools that not only encapsulate the predictive characteristics of oligogenic diseases but also shed light on the potential biological patterns driving these predictions. Ideally, these methods should harness existing knowledge and incorporate interpretable machine learning techniques, transforming the complexity of predictive patterns into concrete biological insights.

2.2 Research question and objectives

Having defined the problem and the context in which it resides, we are led to the central research question of this thesis:

> **Research Question**
>
> How can we effectively leverage prior biological knowledge to design and enhance predictive approaches for oligogenic diseases, shedding light into their potential causal mechanisms?

Guiding hypotheses

To navigate our way through this research question, we articulate several guiding hypotheses.

Firstly, we propose that merging predicted gene pathogenic interactions with an extensive background of biological knowledge can enhance both the interpretability of predictions and their assessment by geneticists.

Secondly, we hypothesise that machine learning predictors have a significant potential to capture intricate patterns underlying pathogenic gene interactions in oligogenic diseases. By employing interpretable methods, these patterns could be unveiled, paving the way for a systemic understanding of these diseases.

Finally, building on the previous hypotheses, we put forth the idea that relationships captured as paths in a biological knowledge graph could shed light on potential causal mechanisms in pathogenic gene interactions. Explanations based on these paths could contribute significantly to the understanding of these mechanisms.

Research objectives

Informed by our research question and guiding hypotheses, we set the following research objectives:

1. **Post-hoc interpretability and contextualisation**: Building on established machine-learning predictors, our goal is to develop post-hoc interpretability strategies while also placing predictions in the context of existing biological knowledge. By doing so, we aim to equip geneticists with integrated tools to analyse patient-specific variant data and derive meaningful insights from prediction outcomes.

2. **Knowledge graph for oligogenic diseases**: To facilitate a deeper understanding of the intricate gene relationships in oligogenic diseases, we seek to construct a biological knowledge graph, integrating information from known oligogenic diseases with various biological networks and ontologies. The resulting resource should enable complex gene relationship queries and prove its utility in inferring pathogenic gene interactions.

3. **Knowledge-driven interpretable model**: Using the biological knowledge graph as a foundation, we aim to develop a machine learning method that can clearly explain pathogenic gene interactions. This predictive model should highlight and present patterns from the knowledge graph in a manner that is both comprehensible and verifiable by geneticists and researchers.

By advancing towards these objectives, our overarching goal is to combine prior biological knowledge with interpretable machine learning, offering geneticists and researchers a suite of tools to better understand the biological mechanisms behind oligogenic diseases.

2.3 Thesis outline

In Chapter 3, we lay down key concepts, resources, and tools underpinning this thesis. We start with bioinformatics resources vital for variant interpretation and continue to gene and protein annotations, detailing their functional roles. Subsequently, we examine biological networks and biomedical ontologies, presenting their structure, key concepts, computational measures, and evaluation techniques. The chapter further presents core machine learning concepts and interpretability techniques. This enables us to introduce oligogenic databases and related predictors in that field. Lastly, we focus on knowledge graphs, detailing their construction process, the importance of metapaths, and the fundamentals of knowledge graph embeddings.

In Chapter 4, we detail ORVAL (Oligogenic Resource for Variant Analysis), a web platform designed to provide additional biological context and interpretability to variant combination predictors. ORVAL assists in predicting and analysing variant combinations. Central to the platform's utility is its user-friendly interface, ease of accessibility, and integrated tools that furnish a comprehensive analysis of submitted patient data. This chapter details the technical aspects of the constructed platform, including the architecture of its components, and an annotation database offering extensive genetic information of human variation. We further discuss how the platform filters and annotates variant data, predicts variant combinations, and provides diverse views for exploring and interpreting the results. Additionally, we showcase how prior knowledge is incorporated to contextualise pathogenic gene modules. We conclude by emphasising ORVAL's impact on genetic research following its public release and its ongoing enhancements, aiming for richer genetic annotations and improved predictive accuracy.

In Chapter 5, we present BOCK (Biological networks and Oligogenic Combinations as a Knowledge graph), a unique knowledge graph developed as part of this research. This chapter articulates the development and application of BOCK in oligogenic disease research. BOCK, by integrating multiple public biological network resources and curated oligogenic disease information from published clinical cases, provides a comprehensive, semantically-rich representation capturing the complexity and diversity of relationships relevant to oligogenic diseases. This chapter explores BOCK's structure and formats, the selection and integration of source databases, flexible querying, and information retrieval capabilities. Furthermore, we discuss the identification and selection of relevant oligogenic gene pairs and the comprehensive network analysis of genes and gene pairs involved in oligogenic diseases. We finally highlight preliminary results on the application of a state-of-the-art KG embedding technique to predict the pathogenicity of gene pairs.

In Chapter 6, we describe the development ARBOCK, an interpretable machine learning approach that leverages the rich information within BOCK to predict and interpret pathogenic gene interactions. ARBOCK aims to fill the interpretability gap by deciphering complex relational patterns between pathogenic gene pairs and offering meaningful explanations for these predictions. This chapter examines the concept of metapaths in creating interpretable predictions and introduces a framework that combines these metapaths as rules for link prediction. We further investigate the discovered patterns associated with pathogenic gene pairs and evaluate the predictive performance of our model. Moreover, we demonstrate the inherent interpretability of this model and showcase its ability to explain predicted pathogenic gene pairs using subgraphs from BOCK. This new approach can provide geneticists and researchers a new way to validate predictions using background knowledge and for generating mechanistic hypotheses, thereby enriching our understanding of oligogenic diseases.

In Chapter 7, the final part of this thesis, we provide a thorough reflection on the research undertaken. We highlight the concrete developments and the scientific contributions made to the field. We also openly discuss the challenges we have encountered throughout the research process, as well as the limitations of our current work. Finally, we present potential areas for future research and exploration, based on our findings and observations.

2.4 Scientific publications and communications

Peer-reviewed journal publications

- Renaux, A., Papadimitriou, S., Versbraegen, N., Nachtegael, C., Boutry, S., Nowé, A., Smits, G., Lenaerts, T. (2019). ORVAL: A novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Research*. 47(W1):W93-W98. DOI: `https://doi.org/10.1093/nar/gkz437`.

- Renaux, A., Terwagne, C., Cochez, M., Tiddi, I., Nowé, A., Lenaerts, T. (2023). A Knowledge Graph approach to predict and interpret Disease-causing Gene Interactions. Forthcoming in *BMC Bioinformatics*.

- Versbraegen, N., Gravel, B., Nachtegael, C., Renaux, A., Verkinderen, E., Nowé, A., Lenaerts, T., and Papadimitriou, S. (2023). Faster and more accurate pathogenic combination predictions with VarCoPP 2.0. *BMC Bioinformatics*, 24(1), 179. DOI: `https://doi.org/10.1186/s12859-023-05291-3`

- Laan, M., Kasak, L., Timinskas, K., Grigorova, M., Venclovas, C., Renaux, A., Lenaerts, T., and Punab, M. (2021). NR5A1 c.991-1G > C splice-site variant causes familial 46,XY partial gonadal dysgenesis with incomplete penetrance. *Clinical endocrinology*, 94(4), 656666. DOI: `https://doi.org/10.1111/cen.14381`

Journal publications in peer-review

- Gravel B., Renaux A., Papadimitriou S., Smits G., Nowé A. and Lenaerts T. (2023) Prioritization of oligogenic variant combinations in whole exomes. [in peer-review (*Bioinformatics*)].

Oral presentations at scientific conferences

- A knowledge graph approach for interpretable prediction of pathogenic genetic interactions - *Student Council Symposium* of the *European Conference on Computational Biology (ECCB)*. **15 mn Talk** | Poster: `https://doi.org/10.7490/f1000research.1119234.1` [Sitges, Spain - 09/2022]
- Towards oligogenic disease prediction with ORVAL: a web-platform to uncover pathogenic variant combinations - *Intelligent Systems for Molecular Biology / European Conference on Computational Biology (ISMB/ECCB)*. **20 mn Talk** | Poster: `https://doi.org/10.7490/f1000research.1117298.1` [Basel, Switzerland - 07/2019]
- Deciphering oligogenic diseases with ORVAL - *Belgian Society for Human Genetics (BeSHG) meeting*. **15 mn Talk** [Brussels, Belgium - 03/2020]

2.5 Supervised master theses

- Bosch, I. Knowledge graph embeddings for the prediction of pathogenic gene pairs. *MSc. thesis. MSc. in Bioinformatics and Modelling - Université Libre de Bruxelles (2023)*, 87. – Thesis co-supervised by Renaux, A., Gravel, B., Lenaerts, T.
- Terwagne, C. Critical assessment of predictive methods in Bioinformatics: Use-case study on predicting gene pairs involved in digenic diseases. *MSc. thesis. MSc. in Bioinformatics and Modelling - Université Libre de Bruxelles* (2021), 96. – Thesis co-supervised by Renaux, A., Lenaerts, T.

2.6 Scientific open data and softwares

Open data

- **BOCK - Biological networks & Oligogenic Combinations as a Knowledge graph** [164]
  - Availability: `https://doi.org/10.5281/zenodo.7185679`
  - Type: Knowledge Graph – RDF, GraphML, TSVs – License: CC-BY-NC 4.0
  - Developers: *Renaux A.* [Conceptualisation & integration]

Softwares

- **ORVAL - Oligogenic Resource for Variant Analysis** [165]
  - Availability: `https://orval.ibsquare.be`
  - Type: Web platform – Sources not distributed – License: CC-BY-NC 4.0
  - Developers: *Renaux A.* [Conceptualisation & development of the fully functional web platform. Maintenance until Nov 2021] ; *Verkinderen E.* [Maintenance from Dec. 2021, implemented VarCoPP 2.0, new database integration pipeline, hg38 update, cloud deployment] ; *Papadimitriou S.* [User documentation]; *Versbraegen N.* [Initial database and initial S-plot] ; *Nachtegael C.* [Initial gene network].
- **ARBOCK - A rule-based classifier based on KG paths: application in pathogenic gene interaction prediction** [164]
  - Availability: `https://github.com/oligogenic/ARBOCK`
  - Type: Open-source Python library and Python Notebook – License: MIT
  - Developers: *Renaux A.* [Conceptualisation & development]

2.7 Fundings

This thesis was supported by the European Regional Development Fund (ERDF) and the Brussels-Capital Region-Innoviris within the framework of the Operational Programme 2014-2020 through the ERDF-2020 project ICITY-RDI.BRU [27.002.53.01.4524] and an Innoviris Joint R&D project Genome4Brussels [2020 RDIR 55b].

We also thank *Foundation 101 Genomes* (`f101g.org`) for fruitful collaboration, creative exchange and scientific support.

Chapter 3

Material and Methods

This chapter sets out the foundational concepts, resources, and tools necessary for understanding the results in subsequent chapters of this thesis.

In the first sections, we focus on bioinformatics resources and predictive tools for single variant interpretation (Section 3.1). Following this, we provide an overview of gene and protein annotations and identifiers, exploring different mapping methods and functional annotations (Section 3.2).

Building upon these annotations, we introduce the basic notions to understand biological networks, outlining key network concepts, relevant network types, and measures of similarity and distance (Section 3.3). We also present ontologies and their application in the biomedical field, detailing their structure, key resources, and methods for evaluating information (Section 3.4).

We then explain fundamental concepts of machine learning and focus on specific techniques of frequent pattern mining and interpretability techniques (Section 3.5). The next section highlights the direct application of machine-learning in the context of oligogenic diseases, as we provide an overview of relevant databases from which we obtain our training dataset and existing prediction models (Section 3.6).

Finally, in Section 3.7, we turn our attention to knowledge graphs – a knowledge representation that can integrate both networks and ontologies discussed earlier. We outline the steps involved in constructing a knowledge graph, explore the importance and usage of metapaths and querying techniques, and acquaint the reader with the fundamentals of knowledge graph embeddings.

3.1 Bioinformatics resources for variant interpretation

Genetic variants are responsible for most of the diversity in individual traits and can be identified by comparing an individual's genome to a reference sequence, known as the reference genome. Genotyping, the process of determining an individual's genetic variation by analysing their DNA sequence, is becoming increasingly common in both research and clinical settings. The key problem has now shifted to the accurate interpretation of the detected genetic variants with respect to their functional and clinical significance. This challenge is being addressed with the help of biomedical databases that integrate large amounts of variant data as well as bioinformatics methods for inferring metrics that facilitate their interpretation.

3.1.1 Variant acquisition and related formats

Genetic variants from patients are typically derived from sequencing data, which undergoes several processing steps before being represented in a standardised format. A brief overview of the sequencing and post-processing steps is as follows [166, 167]:

1. *Library preparation and sequencing*: Depending on the sequencing modality, DNA fragments of interest are collected. Modalities include whole genome sequencing (WGS), whole-exome sequencing (WES), and targeted sequencing panels.
2. *Quality control and read alignment*: Raw sequencing data undergoes quality assessment, filtering out potential artefacts [166]. High-quality reads are then aligned to a reference human genome using alignment tools such as BWA [168].
3. *Variant identification*: This step, often termed variant calling, employs tools to identify genetic variants, such as HaplotypeCaller from the Genome Analysis Tool Kit (GATK) [169], which are then represented in a standardised format.

On average, whole-exome sequencing identifies approximately 25,000 variants per individual [170], while up to 5 million variants are reported on average by whole-genome sequencing [18].

The representation of genetic variants is influenced by the genome assembly used, such as *hg19/GRCh37* or *hg38/GRCh38* [167]. The reference genome assembly serves as the foundation for variant representation, with variants annotated concerning it. Therefore, it is important to use the appropriate genome assembly when processing variant data, as distinct assemblies might yield different variant positions or representations.

Genetic variants can be represented and identified in various ways in databases and in the literature. One of the most common representations is based on genomic coordinates, which describe the position of the variant in the genome, relative to a reference genome. Other representations include the change in the transcribed RNA molecule (cDNA change) and the change in the amino acid sequence of the protein. The Human Genome Variation Society (HGVS) [171] nomenclature system facilitates the standardised description of sequence variants at genomic, transcript, and protein levels.

Variants are typically exchanged in the variant call format (VCF) [172]. The VCF file comprises a header section detailing the format and the encapsulated data. This is followed by a data section, a tab-separated list where each line represents a single variant and its associated details. The first eight columns of the data section are mandatory, encompassing details like chromosome, position, reference allele, and alternate allele. Subsequent columns can contain variant-specific information, such as functional predictions or population frequency data. Given its widespread adoption in genetic research, the VCF format is compatible with numerous tools and pipelines.

3.1.2 Human sequencing and genotyping projects

Over the years, several large projects have been launched with the ambition of creating a comprehensive catalogue of human genetic variations.

One of the early projects was the International HapMap project [173], which aimed to develop a haplotype map of the human genome. A haplotype refers to a set of genetic variants that tend to be inherited together. By understanding these common patterns of human genetic variation, the HapMap project sought to enhance our understanding of the relationships between genotype and phenotype. Following this project, rapid advancements in sequencing technologies have enabled the collection of vast amounts of information on human genetic variation and its implications for disease.

The 1000 Genomes Project [18] was a landmark project, launched in 2007 and finalised in 2015, that used deep exome sequencing and low-coverage genome sequencing to obtain public information on the genetic variations of 2504 healthy individuals from 26 different populations across major human population groups. The project established the first high-quality resource of its kind and provided valuable insights into the differences in genetic variation among populations.

Following these early successes, multiple similar initiatives were launched in different countries. For instance, the UK10K project [23] sequenced the genomes of 6000 individuals from the UK with rare disorders, including neurodevelopmental diseases and obesity, as well as the genomes of 4000 control individuals.

The large amount of genetic variant data collected from these different projects enabled a more accurate estimation of variant frequencies over populations as well as their effects and phenotypic consequences.

3.1.3 Human genetic variant databases

The aggregation of population-level genetic datasets with various conditions has opened up new avenues of exploration to better understand the significance and consequence of genetic variants. These projects have also significantly contributed to the development of public variant databases, providing valuable resources for researchers in the field of human genetics.

Amongst these databases, the Genome Aggregation Database (GnomAD) [21] offers data on over 200,000 genomes from multiple ethnic groups, helping researchers and clinicians to evaluate the rarity and potential pathogenicity of specific genetic variations and to guide the interpretation of genetic test results. One of the key metrics provided by GnomAD is the Minor Allele Frequency (MAF), which quantifies the prevalence of a particular variant in the population. MAF is calculated by dividing the number of instances of the minor allele (the less frequent variant) by the total number of alleles observed in the sample population. By comparing the MAF of a genetic variant to specific thresholds, researchers can determine if a variant is rare or common, thus aiding in the evaluation of its potential clinical significance.

Some public resources provide aggregated information about genetic variations, such as Ensembl [174] and dbSNP [175]. They offer information on the location, identity, and frequency of genetic variants, as well as functional annotations for specific genomic regions.

Other databases are specialised in providing a relationship between genetic variation and disease. For example, the Online Mendelian Inheritance in Man (OMIM) [176] was one of the first resources of its kind, providing information on genetic disorders and their associated genes. The Clinical Variation (ClinVar) database [33] was established more recently and provides information on the relationship between genetic variation and disease, drawing from various sources and including clinician-submitted data.

3.1.4 Predicted variant pathogenicity

The large amount of genetic variant data generated by modern sequencing technologies calls for efficient, automated methods to evaluate the impact of genetic variants. Bioinformatics tools have been developed to address this need, using various approaches to predict variant consequences and aid researchers identify disease-associated variants.

The Combined Annotation-Dependent Depletion (CADD) predictor [86] is a widely used tool in genetic research and clinical genetics to evaluate the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome. CADD provides two types of scores: a raw score, which represents the relative rank of a variant, and a Phred-like score, which is a logarithmic transformation of the raw score. The raw score is particularly informative for comparing variants across multiple datasets.

CADD is based on a machine-learning model trained with a diverse set of functional and non-functional variants. The functional variants were derived from several sources, including ClinVar [176], while the non-functional variants were generated through in silico mutagenesis. This training dataset allowed CADD to effectively predict the deleteriousness of genetic variants across the human genome, incorporating a wide range of features, such as conservation scores and gene expression levels.

In addition to providing scores, CADD also offers annotations for the variants, such as gene identifiers, amino acid changes, and functional consequences. The CADD scores have been widely adopted in clinical and bioinformatics settings, facilitating the prioritisation of variants for further investigation and supporting clinical decision-making.

3.2 Gene and protein annotations and identifiers

Genetic variants are commonly associated with distinct genes. To streamline this association, public databases provide mappings and identifiers. Moreover, both genes and their corresponding proteins can be further characterised using various scores and enriched with functional annotations.

3.2.1 Mappings into gene and protein identifiers

In data integration and annotation stages, it is essential to attribute variants to specific transcripts and genes accurately. Ensembl [174] is a comprehensive genomic database that can help determining if a variant is within a transcript's boundaries or a defined distance upstream or downstream. Some variants are close to the transcription start site (TSS) or transcription termination site (TTS) of a gene, while others are in distant regulatory regions. Variants in these remote locations can be challenging to attribute due to the potential influence of distant regulatory elements on genes.

To manage the complexity of genes with multiple transcripts, Ensembl provides a canonical transcript for each gene[1]. This canonical transcript is the most biologically relevant representation for that gene, especially when multiple alternative transcripts map to the same genomic coordinates [20].

For consistent gene naming, the HUGO Gene Nomenclature Committee (HGNC) [177] offers official gene names and symbols, ensuring uniformity in gene naming across the scientific community.

Regarding proteins, many transcripts code for proteins with a UniProt identifier. UniProt [178] is a central database for protein information, detailing protein sequences, functions, and interactions. A single gene can produce multiple protein isoforms due to

[1] Ensembl canonical transcript: https://www.ensembl.org/info/genome/genebuild/canonical.html

processes like alternative splicing, which should be taken into account when providing gene to protein mappings.

3.2.2 Predicted gene characteristics

A range of bioinformatics scores have been developed to assess gene characteristics and predict the potential impact of any deleterious variant on gene function, providing valuable insights into the characteristics of human genes.

The Gene Damage Index (GDI) [179] is a gene-level approach that prioritises exome variants based on cumulative mutational damage. In contrast, the Residual Variation Intolerance Score (RVIS) [180] quantifies the intolerance of genes to functional genetic variation by comparing the actual number of common functional variants to the expected number based on a gene's total mutational burden.

To provide more accurate predictions of pathogenicity by incorporating inheritance modes, the Inheritance-mode specific pathogenicity prioritisation (ISPP) [181] combines multiple biological features and scores. This method specifically considers the inheritance modes of genes, such as autosomal dominant, autosomal recessive, or X-linked, to better prioritise genes potentially associated with pathogenic variants.

Haploinsufficiency refers to the situation when a single functional copy of a gene is insufficient to maintain normal function, leading to disease or abnormal phenotype. The P(HI) score [182] and the HIPred score [87] assess haploinsufficiency by integrating various genomic features using statistical and machine learning-based predictive approaches, respectively.

Gene essentiality is another important aspect to consider, as it describes the critical importance of a gene for an organism's survival or normal development. The P(rec) score [19] and the Essential in mouse feature [183] tackle this issue using statistical methods and experimental data from mouse knockout studies, respectively.

In addition to these scores, the ratio of non-synonymous to synonymous substitution rates (dN/dS) [184] serves as a measure of selective pressure on genes. This ratio offers insight into the evolutionary forces acting upon genes and can thus provide an indication of their functional importance.

To facilitate the retrieval and analysis of these various scores, resources such as the dbNSFP database [185] consolidate numerous gene-level features and provide mappings to gene identifiers, such as Ensembl identifiers.

3.2.3 Functional gene and protein annotations

Functional gene and protein annotations from various resources play a crucial role in understanding the biological mechanisms and characteristics of genes and proteins. These annotations can be focused on the genetic level, reflecting associations with specific genes, or the protein level, detailing the characteristics and interactions of the proteins they encode.

It is important to note that some of the resources discussed here also describe underlying biological networks (more details provided in Section 3.3) or can be formally described as an ontology (more detailed provided in Section 3.4).

Protein domains and families are functionally and structurally conserved regions within proteins that can be identified using predictive models known as signatures. *InterPro* is a database that integrates protein families, domains, and important sites from various specialised databases [186]. By classifying proteins based on their domains and families, InterPro provides insights into their functions and evolutionary relationships.

Protein complexes are groups of proteins that physically interact and work together to perform specific biological functions. *CORUM* is a manually curated database that provides a comprehensive repository of experimentally characterised protein complexes in mammals [187].

Biological pathways represent series of molecular interactions and reactions that occur within a cell to accomplish specific biological processes. *Reactome* is a widely-used curated and peer-reviewed pathway database containing signalling and metabolic molecules and their relationships organised into reactions, biological pathways, and processes [188].

Gene Ontology (GO) is a widely used resource that provides a controlled vocabulary for cellular components, molecular functions, and biological processes [128, 189]. GO also provides gene annotation data, linking genes to the ontological terms.

Human Phenotype Ontology (HPO) is a resource that defines a controlled vocabulary for human phenotypes and provides annotations for genes associated with these phenotypes [190].

In summary, functional annotations provide essential information about the roles and characteristics of genes and proteins, and enable researchers to investigate their functions within the context of biological systems.

3.3 Biological networks

Biological networks offer a structured representation of complex interactions that exist within a biological system. In this section, we introduce topological metrics in network analysis, present relevant biological network resources for our work as well as similarity and distance measures that were evaluated.

3.3.1 Introduction to networks and their topology

Network structure

A network or a graph $\mathcal{G}(\mathcal{E}, \mathcal{R})$ can be described as a set of entities $\mathcal{E}$, represented as vertices (or nodes), and a set of relationships $\mathcal{R}$, represented as edges, that connect these entities. In biological networks, entities can represent proteins, genes, or cells, while relationships can represent the interaction or association between these biological entities.

Network topology

Network topology quantitively describes the configuration of nodes and edges within a network and the patterns that emerge from these arrangements [191]. A number of metrics can be employed to analyse network topology, including:

- **Degree Centrality**: This measure quantifies the number of connections or interactions a node has within the network. Nodes with a significantly higher number of edges, commonly known as “hubs”, usually play essential roles within biological networks.

- **Closeness Centrality**: This metric quantifies how close a node is to all other nodes in the network. It is calculated as the reciprocal of the sum of the shortest path lengths from the node in question to all other nodes in the network. In the context of biological networks, nodes with high closeness centrality are often critical for rapid information or signal transmission across the network.
- **Graph Density**: This is a measure of the overall level of connectivity within the network. High graph density indicates that a large proportion of all possible connections between nodes are actual connections.
- **Clustering Coefficient**: This measure estimates how closely the neighbours of a given node are interconnected with each other. A high clustering coefficient suggests that a node's neighbours also tend to be neighbours with each other, creating a local "clique" or cluster.

3.3.2 Biological network types and resources

Biological information can be represented as biological networks, which are complex structures consisting of nodes and edges. In biology, nodes represent biological entities, and edges represent interactions or associations between them. These networks are valuable resources for understanding the underlying mechanisms of diseases. We describe here multiple types of networks representing different types of biological relationships between genes and proteins.

Protein-Protein Interaction (PPI) networks provide information on the physical interactions between proteins. These interactions can be direct, where proteins form stable or transient complexes, or indirect, mediated by post-translational modifications or other molecular intermediaries. Three databases providing PPI networks, mentioned in our research, are *Mentha*, *ComPPI* and *STRING*. *Mentha* [192] is a protein-protein integration database that favours precision over comprehensiveness and focuses on experimentally determined protein interactions. Additionally it provides edge confidence scores based on experimental evidence. *ComPPI* [193] is a cellular compartment-specific database of proteins and their interactions from multiple integrated databases, enabling an extensive, compartmentalised protein-protein interaction network. *STRING* [194] is another resource that, besides other types of biological relationships, includes physical protein-protein interactions integrated from multiple source databases.

Gene coexpression networks are built by identifying pairs of genes that show similar expression patterns across samples. They can provide insights into the functional relationships between genes and their potential roles in diseases. *TCSBN* [195] is a database of tissue-specific co-expression networks generated from 46 normal tissues, with scores calculated based on the Pearson correlation coefficient. They have been generated using data from the *GTEx* project [196], a comprehensive resource that catalogs gene expression patterns across various human tissues. Another resource, *CoExpressDB* [197], is a gene co-expression database for animals that offers co-expressed gene lists and gene networks for comparison across different species and platforms.

Sequence homology networks capture the relationships between genes or proteins based on their sequence similarity. The STRING database [194], beside other types of biological relationships, contains pre-calculated pairwise protein sequence alignment bitscore obtained with the BLAST algorithm [198].

3.3.3 Similarity and distance measures

In network analysis, particularly in the context of biological networks, calculating the similarity or distance between two nodes, for example a pair of genes, can help understanding their relatedness. These metrics can also be used to infer unknown relationships.

Neighbour-based similarity measures

The first type of measure calculates similarity based on the shared neighbours between two nodes, which can inform in the short-range similarity. This can particularly useful for identifying gene pairs with many common interaction partners, which can be predictive of their direct interactions [193].

We provide in Table 3.1 a summary of such neighbour-based similarity measures.

Similarity measure	Definition	Formula
Total Neighbours	Union of common neighbours of two nodes	$\lvert N(x) \cup N(y) \rvert$
Preferential Attachment	Likely connectivity based on the product of the degrees of two nodes	$\lvert N(x) \rvert \cdot \lvert N(y) \rvert$
Jaccard Index	Size of intersection over size of union of the neighbours of two nodes	$\frac{\lvert N(x) \cap N(y) \rvert}{\lvert N(x) \cup N(y) \rvert}$
Adamic Adar	A measure of similarity between two nodes based on their shared neighbours, with greater weight given to less common neighbors	$\sum_{u \in N(x) \cap N(y)} \frac{1}{\log \lvert N(u) \rvert}$

Table 3.1. Common neighbour-based similarity measures in networks. This table illustrates several common measures used to calculate node similarity based on shared neighbours in a network. The measures are detailed along with their mathematical formula where $N(u)$ represents the set of neighbours of a node u, while x and y are the two nodes for which the measure is being calculated.

Path and walk-based measures

Other similarity and distance measures consider long-range relationships in the form of paths or walks in the network. Such measures include the Shortest Path Length (SPL) and the Random Walk with Restart (RWR). These measures offer the advantage of capturing indirect associations between nodes.

The SPL, a metric used for distance measurement in network analysis, quantifies the minimal topological distance between two nodes within the network [199]. In the context of this research, the SPL is formally defined as the minimum number of edges that must be traversed to establish a path from one node to another.

The Random Walk with Restart (RWR) is an advanced network analysis technique that simulates the behaviour of a hypothetical walker traversing the network, initiating from a specified starting node. At each iteration, the walker either moves to an adjacent node or returns to the starting node with a predefined probability α [200].

Mathematically, it can be represented as:

$$r = (1 - \alpha)Ar + \alpha r_0 \tag{3.1}$$

where r is the steady-state probability vector, indicating the likelihood of the walker being at each node after multiple iterations, α is the restart probability, determining the chance of the walking returning to the starting node at each step, A is the adjacency matrix representation of the network, capturing the connections between nodes, and r_0 is the initial probability vector, set such that the starting nodes have a probability of 1 and all other nodes have a probability of 0.

The strength of RWR lies in its ability to integrate both local and global network topology. This feature enables the algorithm to uncover nuanced relationships and associations that may not be readily apparent through the application of simpler, local topology metrics [200].

In the field of genomics, RWR is frequently employed to prioritise candidate genes associated with specific diseases. The algorithm starts its random walk from nodes representing genes already known to be associated with the disease under study. Once the steady-state is reached, the resulting probabilities are used to prioritize other genes. Genes that rank highest are considered more likely to have a functional relationship with the disease in question [201].

3.4 Biomedical ontologies

Biomedical ontologies play a crucial role in the organisation and integration of biological data, providing a structured framework for annotating and connecting diverse types of biological information. This section introduce ontologies and present the related exchange formats, biomedical resources and similarity measures relevant for our work.

3.4.1 Ontology structure and related formats

Ontologies, in the context of bioinformatics, are formal, explicit specifications of shared concepts [202]. They play a crucial role in data annotation, integration, analysis, and interpretation, offering a controlled vocabulary of terms and relationships.

Ontology structure and semantics

Formally, an ontology is a graph, with the distinction that relationships in ontologies convey semantically rich, directed relationships that often form a hierarchy, resembling a Directed Acyclic Graph (DAG).

The hierarchical structure of ontologies is based on the subsumption relationship between terms. In this hierarchy, a term subsumes all the terms that fall under it, forming a parent-child relationship between terms. This relationships is often represented as “isA” or ‘subClassOf”. Therefore terms in an ontology can be described in relation with their parent and child concepts. The term with children but no parent concepts, at the top of that hierarchy, is designed as the root term while terms with parent concepts but no children are referred as leaves [127, 203].

Ontologies leverage the semantics inherent in relationships to facilitate reasoning and inference, enabling the extraction of novel knowledge from existing information. For example, if “e_1 is a subclass of e_2” (denoted as e_1 isA e_2) and “e_2 is a subclass of e_3” (e_2 isA e_3), an ontology allows the inference that “e_1 is a subclass of e_3" (e_1 isA e_3).

Semantic web principles and formats

Ontologies can be represented using various formats. The Resource Description Framework (RDF) is typically used for encoding and sharing data in a machine-readable form over the web. The Web Ontology Language (OWL), on the other hand, is more often used to define the model or structure of data, providing a robust language for ontologies that require intricate relationships and high expressivity [204]. Additionally, the Open Biomedical Ontologies (OBO) format is another standard for ontology representation, specifically tailored for the life sciences domain. It offers a simpler syntax than OWL, making it more accessible for certain applications, while still supporting the integration of diverse biological data types and their relationships [127].

3.4.2 Biomedical ontology resources

Biological concepts are often organised using ontology structures, which offer a standardised vocabulary for researchers across disciplines and enable a hierarchical representation of these terms.

Here, we focus on three widely-used biomedical ontologies used in this work: Reactome, the Gene Ontology (GO), and the Human Phenotype Ontology (HPO).

Gene Ontology (GO) is a widely used resource that offers a controlled vocabulary for describing the molecular functions, cellular components, and biological processes associated with genes and gene products [128, 189]. The GO ontology is organised into three distinct but interconnected sub-ontologies: molecular function, cellular component, and biological process. Each sub-ontology is arranged hierarchically, with more general terms at higher levels and more specific terms at lower levels. Key relationships in the GO ontology include “is-a”, “part-of”, “regulates” and “occurs in”. GO annotations connect genes and gene products to the appropriate ontology terms.

Human Phenotype Ontology (HPO) is a resource that defines a controlled vocabulary for human phenotypes and provides annotations for genes associated with these phenotypes [190]. The HPO ontology is organised hierarchically, with more general terms representing broad phenotypic categories and more specific terms describing detailed phenotypic features. The primary relationships in the HPO ontology are “is-a” and “part-of” which capture the hierarchical organisation of phenotypic features. HPO also links genes to the associated phenotypes and includes connections to other disease-related ontologies, such as the Online Mendelian Inheritance in Man (OMIM) [176] and Orphanet [205].

Reactome is a comprehensive and peer-reviewed pathway database that provides an ontology-based representation of biological pathways, including both signaling and metabolic processes [188]. The Reactome ontology is organised hierarchically, with higher-level terms representing general biological processes and lower-level terms specifying more detailed molecular reactions and interactions. The ontology includes relationships such as "is-a", "part-of" and "has-part", which capture different aspects of the relationships between molecular entities, events, and their components. Reactome also annotates genes and proteins involved in these pathways, linking molecular entities to their respective reactions and processes.

These ontologies are commonly used as cross-reference in other reference biological databases such as UniProt [178] and Ensembl [174] and for the development of numerous computational methods [206, 207].

3.4.3 Evaluating information and similarity in ontologies

When analysing biomedical terminologies, encountered for example in gene-associated pathways or patient-associated phenotypes, some questions might arise such as *"How specific or broad is this term?"* and *"Are these two terms related, and if so, to what extent?"*. This section explores two key concepts: Information Content (IC) and Semantic Similarity Measures (SSM) that provides practical tools leveraging ontologies and their annotations to answer these questions.

Accessing term specificity with the information content

Information Content (IC) is a concept derived from information theory, introduced by Claude Shannon. It quantifies the amount of information carried by an event, with the understanding that the more surprising or rare the event, the higher its information content [208].

In the context of ontologies, IC was first introduced by Resnik to measure the specificity of a term [209]. The IC of a term t is given by:

$$IC(t) = -\log(p(t)) \tag{3.2}$$

where $p(t)$ is the probability of encountering an annotation (for example the number of genes associated) to the term t or its descendants ($t_s \sqsubseteq t$) in a given corpus of annotations. The more specific a term (i.e., the fewer annotations it has), the higher its IC.

In bioinformatics, for example, one might use the frequency of a gene annotation to a term in a large gene annotation database, such as Gene Ontology Annotations (GOA), to estimate this probability.

Generally, this probability $p(t)$ is computed as:

$$p(t) = \frac{|I(t \;\cup\; (\bigcup_{t_s \sqsubseteq t} t_s))|}{|I(T)|} \tag{3.3}$$

Where, $I(t)$ denotes the set of instances or annotations associated with the term t and its descendants, and $I(T)$ refers to the set of instances or annotations associated with any term in the entire ontology.

Semantic similarity between ontology terms

Semantic Similarity Measures (SSMs) quantify similarity between ontology terms, aiding data mining and analysis. These SSMs are broadly classified into edge-based, node-based, and hybrid methods [141].

Edge-based metrics, such as Wu and Palmer's measure [210], derive similarity from the path length and ontology tree depth. Conversely, node-based metrics like Resnik's measure [209], use term properties like Information Content (IC), specifically leveraging the Most Informative Common Ancestor (MICA). Lastly, hybrid methods combine both strategies for a nuanced similarity understanding.

The SimGIC similarity measure, in contrast to the measures mentioned above, is a groupwise measure that considers the set of terms that annotate a gene, rather than individual term pairs. For two genes g_1 and g_2, SimGIC is defined as the ratio of the sum of the IC of the terms in the intersection of their annotation sets, to the sum of the IC of the terms in the union of their annotation sets [140].

$$\text{SimGIC}(g_1, g_2) = \frac{\sum_{t \in T(g_1) \cap T(g_2)} IC(c)}{\sum_{c \in T(g_1) \cup T(g_2)} IC(c)} \tag{3.4}$$

While many SSMs have been proposed, they share common core elements which have been unified under a theoretical framework [211], enabling their comparison, selection, and the development of new measures. Based on this framework, Harispe et al. proposed SimcGIC, a pairwise measure, extending the groupwise SimGIC. This measure quantifies the semantic similarity between two terms u and v as follows:

$$\text{SimcGIC}(u,v) = \frac{\sum_{c\in A(u)\cap A(v)} IC(c)}{\sum_{c\in A(u)} IC(c) + \sum_{c\in A(v)} IC(c) - \sum_{c\in A(u)\cap A(v)} IC(c)} \tag{3.5}$$

where A(t) represents the set of terms including t and all its ancestors.

3.5 Machine learning: from predictions to insights

In this section, we will discuss the machine learning concepts and techniques that are used in this research. We will also discuss techniques for explaining predictive results and how frequent patterns in data can be leveraged to create interpretable machine learning models.

3.5.1 Machine learning concepts and techniques

Machine Learning (ML) is a field at the intersection of computer science and statistics, focused on algorithms that can learn from and make predictions based on data. The foundation of ML is in designing and implementing models that can be trained on data, enabling them to make accurate predictions or gain insights into unknown or unseen data.

Data representation and feature selection

In ML, data is represented as a set of input features, corresponding to characteristics or attributes of data points. For instance, in a study of disease susceptibility, the input features could include genetic variants, demographic information, and environmental factors. The set of features form a feature vector, commonly denoted as X. The number of features in this vector is referred to as its dimensionality [212].

Features can take several forms: they can be numerical (*e.g.*, age) or categorical (*e.g.*, sex). They can be directly derived from the original extracted data or be transformed with techniques of feature engineering, that can combine multiple attributes using computational methods or domain-specific transformations [212].

Feature selection is the process of identifying and selecting a subset of input features that are most relevant to the task at hand. This can improve the model's performance, reduce the complexity of the model, and increase interpretability by eliminating irrelevant or redundant features. This process can be guided by various criteria such as domain knowledge, the correlation between the feature and a target variable, or the increase in model predictive performance when the feature is included. Techniques like Recursive Feature Elimination (RFE) can be used to systematically select these features [213].

Supervised vs. Unsupervised

In this research, we focus on two fundamental types of ML: supervised and unsupervised learning [214].

In supervised learning, models are trained on labeled data where each data point has a known target variable. The models learn from the input features and their corresponding targets to establish relationships and make predictions for new, unseen instances. Examples of such models include logistic regression [215] and Random Forest (RF) [216], detailed next.

Unsupervised learning, on the other hand, focuses on analysing unlabelled data to discover patterns, structures, or relationships without specific guidance. It is particularly useful for exploratory data analysis, understanding the underlying structure of the data or even for generating novel features. Clustering techniques, such as K-means [217], and association rule mining algorithms such as the Apriori algorithm [218], are examples of unsupervised learning methods.

Supervised classification algorithms

In the paradigm of supervised learning, the learning algorithm is presented with a set of feature vectors along with their corresponding target values. The task of the learning algorithm is to find a mapping function from the inputs to the output space that can accurately predict the target value for new input data. This function should be able to generalise from the training data to unseen situations.

A supervised learning algorithm can be formalised as an optimisation process searching for a target function (f) that maps input variables (X) to an output variable (Y). Mathematically, it can be represented as: $Y = f(X) + \varepsilon$, where ε is an error term, minimised by the algorithm.

The nature of the output variable Y determines the type of supervised learning task. If Y represents categorical labels, the task is termed classification. Conversely, if Y denotes continuous values, the task is a regression. In the research presented in this thesis, the emphasis is on classification tasks, specifically binary classification, where Y assumes one of two values: positive (commonly denoted as 1) or negative (typically represented as 0). In many applications, including ours, the positive class is the primary focus for identification.

Two commonly used classification algorithms are *Logistic Regression* and *Random Forest*, which were applied in this research.

Logistic Regression [215] is a statistical model used for binary classification problems. It uses a logistic function to model a binary dependent variable, making it suitable for problems where the outcome is either one of two possible classes. Despite its simplicity, Logistic Regression can be effective, especially when the dimensionality of the data is high. A key feature of Logistic Regression is its ability to output probabilities, providing a measure of certainty around its predictions. This makes it not only a classifier but also a probabilistic model.

Random Forest [216] is an ensemble learning method that uses a collection of decision trees to make predictions (often employed with a classification objective). Each decision tree in the ensemble can be described as a flowchart-like structure, where each internal node represents a split on a feature (e.g., age ≤ 50, age > 50), and each leaf node corresponds to a class label. In Random Forest, each tree is trained on a random subset of the training data, which introduces diversity into the model and helps avoid overfitting. The final prediction is made by aggregating the votes from all the trees in the ensemble, typically using a "majority voting" approach.

The training algorithm for Random Forest involves growing decision trees based on measures of information gain and using bootstrapping to introduce randomness. Key parameters, such as the depth of the trees and the number of trees in the forest, can greatly influence the model's performance.

Predictive performance evaluation

Evaluating the performance of a machine learning model is a critical step when developing a predictive model. It helps in understanding the model's ability to generalise to unseen data and provides insights into the balance between bias and variance. *Bias* refers to the error that arises from the assumptions made by a model to simplify the problem, which can lead to underfitting. *Variance*, on the other hand, refers to the error due to the model's sensitivity to fluctuations in the training set, which can lead to overfitting [219].

When considering the binary classification models evaluated in this research, and considering that these models output a classification probability for the positive class, we can consider the following evaluation process:

1. **Train/Test split**: To assess a model's performance, data is typically partitioned into a training set and a test set. The training set is used to train the model, while the test set serves as new, unseen data for evaluation. This basic split, however, can sometimes lead to evaluations that are sensitive to the particular split.
2. **Cross-validation**: To mitigate the sensitivity of a single train/test split, cross-validation is employed. In this technique, the dataset is divided into k subsets. The model is then trained on $k - 1$ of these subsets and tested on the remaining one. This process is iteratively performed k times, ensuring each subset serves as the test set exactly once. The performance metrics from each iteration are then averaged, providing a more robust evaluation of the model's performance [220]. Two types of curves and metrics can be used to evaluate the model performance detailed next.
3. **ROC Curve and AUROC**: The Receiver Operating Characteristic (ROC) curve is a graphical representation that plots the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) for various threshold values [221]. The Area Under the ROC Curve (AUROC) provides an aggregate measure, indicating the model's ability to distinguish between positive and negative classes across all thresholds. An AUROC of 0.5 indicates that the model does not perform better than random.
4. **Precision-Recall Curve and AUPRC**: The Precision-Recall (PR) curve plots Precision against Recall for different thresholds, focusing on the model's performance concerning the positive class [222]. The Area Under the Precision-Recall Curve

(AUPRC) offers a summary measure of this performance. Its interpretation must consider the class imbalance in the evaluated dataset.

5. **Optimal threshold determination**: To achieve the best balance between sensitivity and specificity, an optimal classification threshold is determined, often using the ROC curve (averaged over multiple folds). The point on the curve that maximises the geometric mean is often chosen as the optimal threshold for the classifier.
6. **Evaluation on the test set**: Once the optimal threshold is identified from cross-validation, the model is applied to the test set and only instances predicted over the selected thresholds are considered predicted as positive. This ensures an unbiased evaluation of the model's generalisation capability on unseen data.
7. **Performance metrics calculation**: Based on the chosen threshold, various metrics such as precision, recall, and balanced accuracy can be computed to provide a comprehensive understanding of the model's performance.

Handling imbalanced data

Class imbalance is a prevalent issue in machine learning where the number of observations belonging to one class is significantly lower than those belonging to the other classes. This imbalance can lead to models that have poor predictive performance, specifically on the minority class, as the model tends to be biased towards the majority class [223].

Several strategies have been proposed to tackle this issue. Some models have been designed to natively handle class imbalance, such as the *Balanced Random Forest* model, a variant of Random Forests. In this model, each tree in the ensemble is trained on a data sample where the class distribution is balanced, which can be for example achieved by under-sampling the majority class in each bootstrap sample, giving equal importance to both classes [224].

Another common strategy to handle class imbalance is cost-sensitive learning, where the weights of the classes are adjusted based on their frequencies in the input data. This adjustment effectively increases the cost of misclassifying the minority class, making the model more sensitive to it [225]. This strategy is applicable to many machine learning algorithms.

When evaluating models trained on imbalanced data, stratified cross-validation is recommended. This variant of cross-validation maintains the same ratio of classes in each fold as in the full dataset, providing a more accurate estimate of the model's performance on the minority class [220].

3.5.2 Optimisation techniques in machine learning

We present here a some computational optimisation strategies, essential in machine learning and data mining.

Cost function and the greedy approach

In combinatorial optimisation, a cost function $C(\mathbf{x})$ quantifies the quality or suitability of a solution x. Similar to how natural selection assesses an organism's fitness for survival and reproduction, the cost function evaluates how well a solution addresses a given problem [226].

Optimisation algorithms use heuristics to efficiently traverse a vast search space of potential solutions, aiming to identify the solution with the optimal (often minimal) cost. In the context of machine learning, these algorithms are employed to determine a combination of parameter values that best aligns with the data.

One type of optimisation is the greedy optimisation strategy, favoured for its computational efficiency, makes the best local choice at each step, aiming for a global optimum. However, its short-sightedness, not considering the broader solution space, can result in local optima traps, leading to a sub-optimal solution [226].

The differential evolution algorithm

The differential evolution (DE) algorithm is a population-based optimisation techniques. It improves a set of candidate solutions iteratively through the processes of mutation, crossover, and selection [227].

- **Mutation** adjusts existing solutions, introducing variability within the population.
- **Crossover** combines elements from two solutions, creating a new potential solution.

- **Selection** retains the best-performing solutions based on their fitness, ensuring they progress to the next iteration.

The advantages of DE over other techniques include its ability in handling high-dimensional search spaces and its resilience against getting trapped in local optima. DE also emphasises real-valued solutions and its distinct mutation method, often outperforming other methods in numerical optimisation tasks [227].

The DE algorithm is as follows:

Algorithm 1: Differential Evolution Algorithm

```
Input: Population size $N$, maximum number of iterations $T$, mutation factor $F$,
       crossover rate $CR$
Output: Best solution $\mathbf{w}$
Initialize a population of $N$ solutions randomly;
Set iteration count $t = 0$;
while $t < T$ do
    foreach solution $\mathbf{w}$ in the population do
        Select three random solutions $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ distinct from $\mathbf{w}$;
        Generate a trial solution $\mathbf{v}$ by applying mutation: $\mathbf{v} = \mathbf{a} + F \cdot (\mathbf{b} - \mathbf{c})$;
        Perform crossover between $\mathbf{w}$ and $\mathbf{v}$ with probability $CR$ to produce a
          trial vector $\mathbf{u}$;
        Evaluate the fitness of $\mathbf{w}$ and $\mathbf{u}$ using the cost function;
        if fitness($\mathbf{u}$) is better than fitness($\mathbf{w}$) or a random value $< CR$ then
            Replace $\mathbf{w}$ with $\mathbf{u}$;
        end
    end
    Increment $t$ by 1;
end
Return the best solution $\mathbf{w}$ based on fitness;
```

The effectiveness of DE depends on the appropriate tuning of parameters like the mutation factor F and the crossover rate CR, which should be empirically assessed.

3.5.3 Discovering association rules with frequent pattern mining

In the context of machine learning, frequent pattern mining is an unsupervised approach aiming to find elements that frequently co-occur in a dataset. It aims to efficiently summarise the characteristics of complex data. Frequent pattern mining has proven its value in bioinformatics, to identify biologically relevant patterns [228].

Transactions, items and frequent itemsets

In frequent itemset mining, we consider transactional data, denoted by $\mathcal{D}$, which consist of a collection of transactions.

Transactions t can be seen as unique instances qualified with an identifier *tid*, associated with a set of items or *itemset*.

An *itemset* is a subset $X = \{i_1, \ldots, i_k\}$ of items $i \subseteq \mathcal{I}$, containing k items.

The *support* of an itemset X quantifies how often X appears in $\mathcal{D}$. If X appears more than a predetermined threshold *minsup*, it's termed a *frequent itemset* for that threshold.

To draw a simple analogy, we can think of items as products in a store and transactions as the set of products a customer buys during a visit. Frequent itemsets, in this case, would be popular combinations of products customers tend to buy together.

The Apriori algorithm

The Apriori algorithm [218] is a foundational method in frequent itemset mining, aiming to identify itemsets that frequently appear within a dataset. The algorithm operates iteratively, expanding the size of the itemsets in each cycle. Its efficiency is rooted in the anti-monotonicity property: if an itemset is infrequent, its supersets will also be infrequent, allowing the algorithm to prune a significant portion of the search space.

The Apriori algorithm has been further optimised by representing each itemset by its list of transaction IDs (*tid-list*), eliminating the need for repeated dataset scans [229].

The Apriori algorithm can be outlined as:

Algorithm 2: The Apriori algorithm

```
Input   : Dataset $\mathcal{D}$, Min. support (minsup), Max. itemset length (maxlen)
Output: Frequent itemsets up to length maxlen
$L[1]$ = GenerateL1($\mathcal{D}$);
$k = 2$;
while $k \leq$ maxlen and $L[k-1]$ is not empty do
    $C$ = GenerateCandidates( $L[k-1]$);
    $L[k]$ = FilterCandidates($C$, $L[1]$, minsup);
    $k = k + 1$;
return All L up to maxlen
```

More specifically, the key steps and optimisations are:

- **GenerateL1(D)**: Constructs tid-lists for each item in $\mathcal{D}$. It returns the set $L[1]$ containing 1-itemsets that meet the minsup threshold.
- **GenerateCandidates(L[k-1])**: Forms candidate itemsets C of size k by combining frequent $k-1$-itemsets from $L[k-1]$.
- **FilterCandidates(C, L[1], minsup)**: For each candidate in C, it first ensures the anti-monotonicity property by verifying that all of its $k-1$ subsets are frequent (present in $L[k-1]$). To determine the frequency of the candidate, and as part of the optimisation from [229], it intersects the tid-lists of the items forming the candidate, specifically focusing on the tid-list of the item in the candidate with the minimum support from $L[1]$. Only candidates that meet the minsup threshold are added to $L[k]$.

Closed Itemset Principle

The Apriori technique is known to discover many redundant patterns. One technique to reduce that redundancy is to consider only *closed itemsets* [230]. A k-itemset is termed as a closed itemset if there exists no superset ($k+1$-itemset) with the same support in the dataset. Using the closed itemset principle allows for the elimination of redundant patterns, as any non-closed itemset's information is already represented by its corresponding closed itemset.

Class association rules

Using frequent itemset mining technique, it is possible to mine *association rules*. Association rules represent knowledge in the form of logical if-then statements expressed as:

$$X \Rightarrow Y$$

These rules consist of a body or antecedent (if part) and a head or consequent (then part). We focus here on a specific type of association rules, class association rules (CAR) [231], where X, the antecedent is an itemset and the consequent, denoted as Y, represents the predicted class or outcome.

Rule interestingness measures

The interestingness of association rules can be assessed based on a wide variety of measures. The two most widely used constraints are: support and confidence [228].

The support of a rule represents the proportion of transactions in the dataset that contain both the antecedent and the consequent. It is a crucial measure as a rule with a low support might be a random occurence. It is calculated as:

$$\text{support}(X \Rightarrow Y) = \frac{\text{frequency}(X \cup Y)}{|\mathcal{D}|} \tag{3.6}$$

The confidence of a rule measures the proportion of transactions containing the antecedent that also contain the consequent. It can assess the reliability of the inference made by a rule. It is defined as:

$$\text{confidence}(X \Rightarrow Y) = \frac{\text{frequency}(X \cup Y)}{\text{frequency}(X)} \tag{3.7}$$

3.5.4 Interpretability in machine-learning

Feature contribution analysis

Feature contribution analysis helps understand the importance of different features in the decision-making process of a predictive model. It quantifies the influence of each feature on the model's predictions, more precisely their relative importance [232].

This technique is particularly useful for interpreting the predictions of complex models like random forests. It provides a post-hoc explanation of the model's predictions by decomposing each prediction into a sum of contributions from each feature [232]). The *treeinterpreter* [2] package, a Python tool, implements this method. It outputs a set of contribution values for each feature, for each prediction. These values indicate the amount that each feature contributed to the prediction, with positive values indicating an increase in the prediction value, and negative values indicating a decrease.

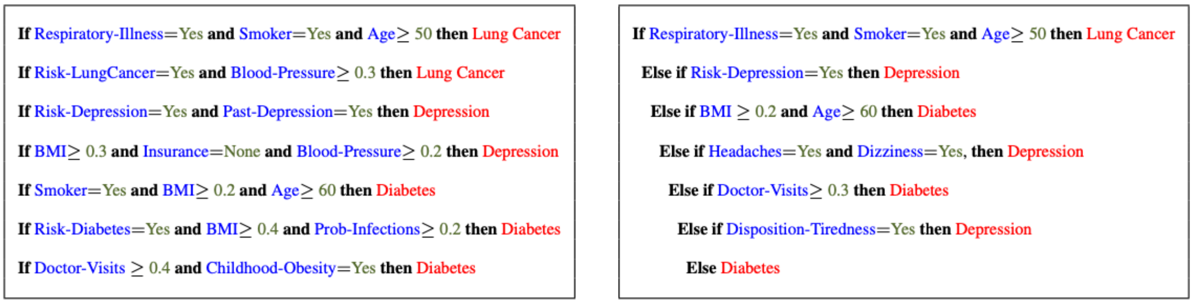
[2] Ando Saabas, "treeinterpreter", GitHub, `https://github.com/andosa/treeinterpreter`

However, the method assumes that the contributions of individual features are additive, which may not be the case in models where features interact with each other. Additionally, the accuracy of the feature contribution estimate can depend on factors such as the complexity of the model and the correlation between features [104].

Rule-based associative classifiers

Rule-based classifiers are considered the most interpretable type of models in machine-learning. They are often referred as *white-box* models, offering both global and local interpretability. The underlying rules, which can be explored and returned as explanations, can capture complex feature interactions associated to a particular class label or decision.

Associative classification is a supervised machine learning approach that leverages association rules (see Subsection 3.5.3) to build classifiers. It is an alternative to traditional rule-based models such as those derived from decision trees (e.g., RIPPER [106]). This approach allows for a more exhaustive exploration of the feature space, capturing complex relationships and avoiding the limitations of greedy approaches [233, 234].



If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age≥ 50 **then** Lung Cancer
If Risk-LungCancer=Yes **and** Blood-Pressure≥ 0.3 **then** Lung Cancer
If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression
If BMI≥ 0.3 **and** Insurance=None **and** Blood-Pressure≥ 0.2 **then** Depression
If Smoker=Yes **and** BMI≥ 0.2 **and** Age≥ 60 **then** Diabetes
If Risk-Diabetes=Yes **and** BMI≥ 0.4 **and** Prob-Infections≥ 0.2 **then** Diabetes
If Doctor-Visits ≥ 0.4 **and** Childhood-Obesity=Yes **then** Diabetes

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age≥ 50 **then** Lung Cancer
Else if Risk-Depression=Yes **then** Depression
Else if BMI ≥ 0.2 **and** Age≥ 60 **then** Diabetes
Else if Headaches=Yes **and** Dizziness=Yes, **then** Depression
Else if Doctor-Visits≥ 0.3 **then** Diabetes
Else if Disposition-Tiredness=Yes **then** Depression
Else Diabetes

Figure 3.1. Illustration of a decision set (left) versus a decision list (right) obtained from a medical dataset. Decision sets have independent rules, enhancing interpretability. In contrast, order matters in decision lists as each rule depends on preceding ones not being true. Credit figure: Lakkaraju et al. [235]

Decision Sets are particularly valuable in domains where interpretability is important. These classifiers are based on a compact unordered set rule, each interpreted in a disjunctive manner (*i.e* OR relationship) [235]. Unlike decision lists, which are structured with if-then-else statements, decision sets do not follow a hierarchical organisation (Figure 3.1). This non-hierarchical structure ensures that each rule in the decision set must be an accurate predictor in isolation, enhancing the interpretability and understandability of the model [235, 236].

A decision set R is composed of rules $\{r_1, ..., r_n\}$, each of the form $(X \Rightarrow Y)$, where X is an antecedent and Y the class label. The assignment of class labels by the model works as follows:

- An instance meeting one X is labeled Y.
- If no X is met, a default label is assigned (*e.g.* the majority class [235])
- Multiple X matches invoke a tie-breaker (*e.g.* the most accurate rule label [235])

The task of finding a compact rule set that best approximate the dataset is a challenging combinatorial optimisation problem. Earlier methods, such as CBA [231], used a sequential covering approach, but these often stop at locally optimal solutions. To counter this shortcoming, recent advances have framed the task as a weighted set cover problem [155] or submodular function maximization [235].

The problem of identifying an optimal rule set can be addressed via the weighted set cover algorithm, as demonstrated in the RUDIK method [155]. Central to this approach is the marginal weight of a rule. Given a set of rules R and an external rule r, the marginal weight, $w_m(r)$, is calculated as:

$$w_m(r) = w(R \cup \{r\}) - w(R)$$

where $w(R)$ represents the total weight of a candidate rule set R.

In RUDIK, considering a binary classification task, $w(R)$ is implemented as a weighted average between the positive and negative coverage of all rules in the candidate rule set.

The algorithm operates in a greedy manner, continuously selecting rules based on their marginal weight, as detailed below:

The rule selection procedure is as follows:

Algorithm 3: Rule selection via Greedy Weighted Set Cover

```
1. Start with all candidate rules $R$ and an empty solution set $R'$.
2. Compute the marginal weights $w_m(r)$ for all rules $r$ in $R$
3. In each iteration:
    • Choose the rule $r$ from $R$ with the lowest marginal weight.
    • If $w_m(r) < 0$, include $r$ in the solution $R'$ and remove $r$ from $R$.
4. Terminate the algorithm when any of the following conditions are met:
    • All rules in $R$ are part of the solution $R'$.
    • All positive instances are covered by $R'$.
    • No remaining rule in $R$ has a negative marginal weight.
```

In the RUDIK implementation, the algorithm aims to maximise positive rule coverage and minimise negative coverage. Compared to the sequential covering method, this approach is less myopic as it reconsiders the entire candidate rule set at each step, in comparison to the solution set. However, due to its greedy nature, it does not guarantee the discovery of an optimal rule set.

3.6 Oligogenic disease resources and predictive tools

3.6.1 The Digenic and Oligogenic Disease Databases

The Digenic Disease Database (DIDA) [68] was a pioneering resource in the field of oligogenic disease studies, providing valuable information on digenic variant combinations associated with human genetic diseases. Since its launch in 2015, DIDA has been widely consulted and used as a training dataset for various machine learning methods aiming to predict and understand the cause of digenic diseases [81, 82, 83, 85]. However, DIDA's limitations, including its narrow focus on digenic cases, required an adaptation of its architecture to accommodate information on other oligogenic cases or combinations involving CNVs and other variants. More importantly, the emergence of guidelines in reporting causative variants for genetic diseases highlighted the need for a serious re-evaluation of the original criteria for the inclusion of oligogenic combinations in the database, emphasising the necessity for objective evaluation metrics reflecting the quality and strength of different types of evidence, both genetic and functional, supporting their causality.

OLIDA (Oligogenic Disease Database) [71] is an improved and comprehensive database that expands upon DIDA, covering oligogenic diseases caused by mutations in multiple genes and incorporating CNVs, single-nucleotide variations, and small insertions/deletions (indels). Through the curation of 262 scientific articles, OLIDA includes 916 oligogenic variant combinations linked to 159 genetic diseases, involving 1,974 distinct variants in 757 distinct genes. Compared to DIDA, which included 52 genetic diseases, OLIDA now features 191 combinations with variants in more than two genes (up to 17) and 62 combinations involving 31 distinct CNVs.

The curation protocol for OLIDA involved a rigorous process of screening scientific articles and extracting relevant information based on specific criteria, including re-evaluating the original criteria for the inclusion of oligogenic combinations. Confidence scores are assigned for each oligogenic combination based on structured criteria reflecting the level of evidence supporting the causality of the combination for its associated disease.

Out of the 916 combinations present in OLIDA, 38% have a FINALmeta score of 1 or higher, indicating strong evidence supporting their causality. Among these, 133 combinations are linked with all three types of evidence for oligogenicity (familial, statistical, and functional) (Figure 3.2). Over-represented diseases in OLIDA include Kallman syndrome (10%), amyotrophic lateral sclerosis (10%), isolated anencephaly (8%), and normosmic congenital hypogonadotropic hypogonadism (8%). More than half of the diseases in OLIDA (59%) are linked with only one or two associated oligogenic combinations.

The OLIDA database aims to be a user-friendly, accessible resource for researchers in the field of oligogenic diseases, adhering to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. It provides a comprehensive and up-to-date repository of oligogenic variant combinations and their associated diseases via a website provides easy access to search for specific diseases or genes and allows users to download data in various formats.

Additionally, the standardised curation process that was designed and followed to integrated variant combinations in OLIDA highlighted several recurring issues concerning

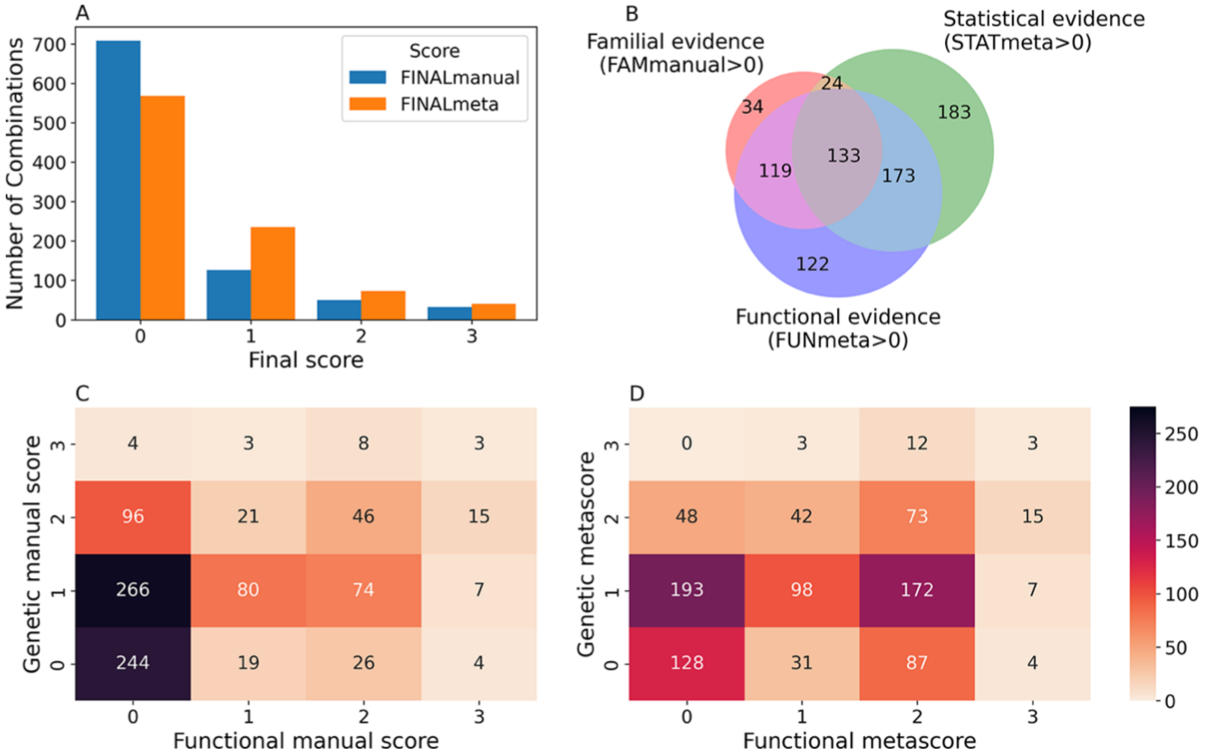


Figure 3.2. Confidence scores and types of evidence present in the OLIDA combinations. (A) Distribution of the FINALmanual and FINALmeta scores. (B) Venn diagram of the number of oligogenic combinations carrying a score of 1 or higher in the different main types of evidence metascores. The 130 oligogenic combinations whose FAMmanual, STATmeta and FUNmeta scores are all 0 are not shown in this diagram. (C) Heatmap of the number of combinations and their confidence functional and genetic scores based on the evidence collected via manual curation (Manual scores) only and (D) when adjusted using the external database information (Meta scores). The genetic score here represents the maximum score between the FAMmanual and STAT (manual or meta for plots a and b, respectively) and the functional score is the FUN (manual or meta for plots a and b, respectively). **Credit: original figure from Nachtegael & Gravel et al. 2022 [71]**

the reporting and pathogenicity assessment of oligogenic cases. These mainly concern the absence of strong evidence that refutes a monogenic model and the lack of a proper genetic and functional assessment of the joint effect of the involved variants. These observations led to the proposition of standards and guidelines on how these oligogenic/multilocus variant combinations should be reported in order to provide high-quality data and supporting evidence to the scientific community [80].

Overall, OLIDA represents a significant step towards a better understanding of the causes of oligogenic diseases, providing high-quality information for researchers and clinicians.

3.6.2 The Variant Combination Pathogenicity Predictor

The VarCoPP method [83] is a machine-learning predictor based on an ensemble of Random Forests, that can differentiates pathogenic from neutral bilocus variant combinations. The method incorporates 11 features at the variant, gene, and gene-pair levels. Variant-level features include the Combined Annotation Dependent Depletion (CADD) [86] scores of the four alleles of a bilocus combination. Gene-level features consider the recessiveness and haploinsufficiency probabilities for each gene, while the gene-pair level feature is a metric of biological relatedness based on protein-protein interaction data. It has been originally trained on the pathogenic variant combinations present in DIDA against a large subset of variant data derived from control individuals of the 1000 Genomes Project (1KGP) [18].

The VarCoPP method assigns a "pathogenic" or "neutral" label to each variant combination based on a majority vote among individual classifiers. It also provides two prediction scores: the *Classification Score (CS)*, which is the median probability of pathogenicity across all predictors, and the *Support Score (SS)*, which is the percentage of predictors agreeing on the pathogenic label.

The method has shown high precision and sensitivity, with 88% accuracy in cross-validation settings. However, it can yield an important number of false positives in full-exome analysis, necessitating additional filtering steps. Despite this, VarCoPP represents a significant advancement in multivariant pathogenicity predictions, offering a more informative and accurate approach than solely relying on monogenic variant pathogenicity scores.

While this thesis integrates and presents results for the initial VarCoPP model (see Chapter 4 - Section 4.4), a more advanced iteration, VarCoPP 2.0, has since been developed, following the same principles as the original VarCoPP, with updated training data from OLIDA, novel features and a simpler model. The evolution of the model and the role played by knowledge graph developed during this thesis (see Chapter 5) are further mentioned in Chapter 4 - Subsection 4.6.1.

3.6.3 The Digenic Effect Predictor

The Digenic Effect Predictor [81, 82] is a machine-learning method that predicts the type of digenic effect for a pathogenic digenic variant combination. It distinguishes between three classes of pathogenic variant combinations: True Digenic, Monogenic + Modifier, and Dual Molecular Diagnosis (see Subsection 1.3.2).

This predictor is based on the Random Forest (RF) algorithm trained on 240 pathogenic variant combinations, including 90 True Digenic and 75 Monogenic+Modifier combinations from the Digenic Diseases Database (DIDA) [68], and 75 Dual Molecular Diagnosis combinations from the work of Posey et al [51]. The training dataset comprises single nucleotide variations and small insertions/deletions.

The Digenic Effect Predictor is based on various biological features including variant-level features, such as CADD raw scores, gene-level attributes such as the gene recessiveness probabilities and gene-pathway information. The predictor can provide, for each variant combination, predicted probabilities for each digenic effect class. The final digenic effect class is determined based on the highest probability among the three classes.

The method shows some limitations in terms of predictive performance, especially for distinguishing True Digenic from Modifier cases (specificity and sensitivity below 0.52), a limitation that could be due to the limited data available. Nevertheless, it can serve as a valuable tool in predicting and understanding the digenic effects of variant combinations, providing insights into the mechanisms and clinical manifestation of these combinations in the context of digenic diseases.

3.6.4 The Digenic Gene pair Predictor

The DiGePred method is a random forest machine learning model following the same logic as VarCoPP, but for the identification of gene pairs associated with digenic diseases, as opposed to variant combinations.

It is trained on a dataset consisting of positive and negative gene pairs. The positive set comprises 140 known digenic disease gene pairs from the Digenic Diseases Database (DIDA), while the negative set consists of putative non-digenic gene pairs generated by the authors using negative controls, based on unaffected relatives of the Undiagnosed

Diseases Network (UDN) [37] cohort.

DiGePred incorporates various features reflecting gene properties and interactions. These features encompass protein family and domain similarity, evolutionary history and constraint, network and pathway information, and phenotype-related attributes such as phenotype similarity and the number of phenotypes for each gene.

The performance of the DiGePred classifier was evaluated through cross-validation and an independent test set. It achieved an AUROC of 0.97 and AUPRC of 0.75 in cross-validation. On their held-out test set, the classifier achieved an AUROC of 0.96 and an AUPRC of 0.69.

3.7 Knowledge graph integration and link prediction

Knowledge Graphs (KG) provide a structured way to integrate and represent data from different sources; this section explores their definitions, their construction and formats, the concepts of metapaths, how they can be queried and a representation learning technique known as KG embedding.

3.7.1 Definition and applications

A KG is a structured representation of data, where diverse entities (nodes) are interconnected through various relationships (edges). This structure, which is inherently heterogeneous, encapsulate rich semantics, often incorporating varied resources, including ontologies, to provide a more comprehensive understanding of the information.

More formally, a knowledge graph can be denoted as $\mathcal{K}$, with $\mathcal{E}$ and $\mathcal{R}$ representing the sets of entities and relations in $\mathcal{K}$, respectively.

Knowledge graphs have found utility in a variety of applications, including search engines, recommendation systems, and knowledge discovery, due to their ability to facilitate data integration and interoperability, and support inference and reasoning [124, 237].

3.7.2 Knowledge graph construction and formats

The construction of a knowledge graph involves the integration of data from multiple sources and the representation of this data in a structured format.

Multi-level data resource integration

The integration of data from multiple sources into a knowledge graph involves several key steps:

- **Schema Definition**: The schema, defining the types of entities and relationships, is the first step in data integration. It accommodates specific types of data and can be updated or expanded as new data types are added.
- **Data Parsing**: After schema definition, existing networks and ontologies are parsed to extract relevant data. The parsing methods can vary depending on the data sources and the schema.
- **Data Cleaning and Preprocessing**: Prior to integration, the data is checked for errors or inconsistencies, and necessary corrections or adjustments are made. This ensures the quality and consistency of the integrated data.
- **Semantic Integration**: The parsed data is semantically integrated into the knowledge graph according to the schema. This may involve mapping entities and relationships from the original data to their counterparts in the schema. In some cases, entities may be collapsed to simplify the graph structure, as done in Hetionet [162, 163] and PrimeKG [238].
- **Identity Resolution**: Identity resolution, the final step in data integration, resolves ambiguities in entity identities. This is crucial when integrating data from multiple sources. Registries, such as Ensembl gene stable identifiers [3] and identifiers.org [239] can be used to define unique Uniform Resource Identifiers (URI) [204].

[3] Ensembl gene stable identifiers: `https://www.ensembl.org/info/genome/stable_ids/index.html`

- **Evaluation of the Integrated Data**: Post-integration, the results are evaluated to ensure the success of the integration process. This could involve checking the knowledge graph for errors or inconsistencies, or assessing its performance in specific tasks or applications.
- **Maintenance and Update of the Knowledge Graph**: This ongoing process involves adding new data, updating existing data, and making necessary adjustments to the schema or the data integration processes.

Representations and formats

The representation and storage of knowledge graphs rely on specific formats and databases:

- **RDF (Resource Description Framework)**[4]: a standard model for data interchange on the Semantic Web, uses URIs to identify resources and describe them with RDF triples. This model integrates heterogeneous data sources and provides globally unique identifiers for resources in a knowledge graph.
- **OWL (Web Ontology Language)**[5]: enables the formal description of the knowledge graph schema – entities, relationships, and constraints within a knowledge graph. The use of RDF and OWL ensures compliance with Semantic Web standards, promoting interoperability [204].
- **GraphML**[6]: an XML-based file format, represents graph structures and their associated data. It supports various graph types and data types for node and edge attributes, facilitating the exchange and processing of graph data across software tools and libraries [240].
- **Neo4j**[7]: a graph database management system, supports efficient storage, retrieval, and traversal of graph data. Its Cypher query language allows for expressive querying of the graph, extracting complex patterns and relationships. Neo4j's capabilities for data integration, visualisation, querying, software integration, and deployment,

[4] RDF: https://www.w3.org/RDF/

[5] OWL: https://www.w3.org/OWL/

[6] GraphML: http://graphml.graphdrawing.org

[7] Neo4J: https://neo4j.com

along with its support for various graph algorithms and analytics, make it a versatile platform for diverse applications.

Python-based libraries such as NetworkX [8] and graph-tool [9] offer extensive functions for the creation, manipulation, and study of the structure and dynamics of complex networks. They also support the encoding of node and edge properties, making them suitable for the manipulation of knowledge graphs.

These tools provide options for representing, storing, and manipulating knowledge graphs, allowing researchers to select the most suitable ones for their specific needs.

3.7.3 Paths and metapaths in knowledge graphs

Paths and metapaths play a crucial role in KGs, particularly in tasks such as link prediction and inference. They provide a way to capture and quantify the relationships between entities in the graph.

In a knowledge graph $\mathcal{K}(\mathcal{E}, \mathcal{R})$, where $\mathcal{E}$ and $\mathcal{R}$ denote the set of entities (nodes) and relationships (edges) respectively, a path $P(e_1, e_n)$ between two entities e_1 and e_n is defined as a sequence $e_1 \xrightarrow{r_1} e_2 \xrightarrow{r_2} ..., e_n$ such that $e_i \in \mathcal{E}$, $r_i \in \mathcal{R}$, and r_i is the relationship connecting e_i and e_{i+1}.

In KGs, entity and relationships are typed. The set of entity and relationship types is denoted $\mathcal{ET}$ and $\mathcal{RT}$ respectively. These types are often referred to as metanodes and metaedges. For more granularity, especially when relationship types are ubiquitous (e.g., `linkedTo`), metaedges are often represented along with their connected metanodes (e.g., `Phenotype-linkedTo-Disease`).

A metapath is an oriented sequence of node types and edge types, starting from a source entity and ending at a target entity. Formally, a metapath M in a knowledge graph can be defined as a sequence:

$$M = E_1 \xrightarrow{R_1} E_2 \xrightarrow{R_2} ... \xrightarrow{R_m} E_n \tag{3.8}$$

[8] Networkx: https://networkx.org

[9] Graph-Tool: https://graph-tool.skewed.de

where $E_i \in \mathcal{ET}$ represents a specific metanode and $R_i \in \mathcal{RT}$ represents a specific metaedge. Metapaths capture the semantics of node connections and have been leveraged to compute the similarity between nodes and cluster them [241], in link prediction tasks [242] and representation learning [243], with applications in various fields. They have been in particular applied for fact validation tasks [244] and used to make recommendation systems explainable [245].

Metapaths as logical rules and rule antecedent

Metapaths can be formalised using first-order logic, expressing relationships between node types as predicates between two variables, chained as a conjunctive expression. For example, the metapath $M = E_x \xrightarrow{R_1} E_1 \xrightarrow{R_2} E_2 \xrightarrow{R_3} E_y$ can be represented as a conjunction of logical atoms: $R_1(E_x, E_1) \wedge R_2(E_1, E_2) \wedge R_3(E_2, E_y)$, where each $R_i(E_i, E_{i+1})$ is an atom, composed of a predicate R_i and two variable terms E_i and E_{i+1}. Metapaths can therefore be seen as logical queries and employed to query any knowledge graph, as shown in the next section.

This logical representation is also employed in rule mining approaches on KGs, such as the AnyBURL method [156], where discovered Horn rules can take the form: $R_1(E_x, E_1) \wedge \ldots \wedge R_m(E_n, E_y) \implies R_l(E_x, E_y)$. Here, the consequent of the rule, $R_l(E_x, E_y)$, represents a KG relationship type R_l that can be inferred between any node pair (x,y) satisfying the rule antecedent. Rule systems in KGs often leverage edge-labelled paths to discover predictive rules. While some of these approaches can also discover non-path structures and usually incorporate constants (*i.e* concrete entities) in their rules [157], metapaths can be seen as a restricted version of these rules' antecedents (referred as *path rule* in [156]), focusing solely on the sequence of abstract relationships between entities in the knowledge graph.

Metapath compact representation

To represent metapaths in a compact and readable way, a short sequence of characters can be assigned to each metanode and metaedge, respectively in upper and lowercase, as proposed by *Himmelstein et al.* [162, 163]. For example, the metapath `Gene-associated-`

`Phenotype-associated-Gene-upregulates->Gene` can be represented concisely with the notation: `GaPaGur>G`. In ambiguous cases, the directionality of the relationships can be explicitly represented in the metapath using the `<` or `>` symbol.

Metapath scoring

Finally, when extracting metapaths between two given nodes, it is possible to calculate an associated metapath weight. Many weighting strategies have been proposed, attempting to represent the underlying path information, including: *Path Count (PC)*, weighting a metapath with the count of its path instances and *Path Constrained Random Walk (PCRW)* [246], estimating the probability for a walk starting on the source node to reach the target node by following the metapath constraint.

3.7.4 Knowledge graph querying

KGs, with their rich semantic structure, offer a unique opportunity for complex querying and analysis. The ability to traverse the graph following specific types of relationships (edges) between specific types of entities (nodes) allows for the extraction of meaningful and contextually relevant information.

Querying a KG involves specifying a pattern of nodes and edges, and finding all subgraphs of the knowledge graph that match this pattern. The complexity of the query can range from simple lookups of nodes or edges, to more complex queries that involve multiple nodes and edges, and may include conditions on the types or properties of the nodes and edges. Query languages such as SPARQL for RDF-based knowledge graphs [247] and Cypher for property graph databases like Neo4j [248] are commonly used.

In the Cypher query language, an example of KG query can be:

```
MATCH p=(:Gene{name:'CFTR'})-[:interacts]-(i:Gene)-[:associated]->(:
   ↪Disease{name:'Cystic fibrosis'})
WHERE i.essential = true
RETURN p
```

This query would return all paths in the KG where the gene *CFTR* interacts with another essential Gene ($essential = true$), itself associated to the disease *Cystic fibrosis.*

A Cypher [248] query is usually composed of the following parts:

- `MATCH`: used to specify the pattern of nodes and edges to match in the graph. Nodes are represented in parentheses and edges in square brackets. Curly brackets are used to specify specific properties on entities. Variables can be attributed to any element (for instance i and p, respectively attributed to a Gene and a path).
- `WHERE`: used to specify conditions on any element declared above.
- `RETURN`: used to specify what information to retrieve from the matched pattern.

These elements form the basis of querying in Cypher and can be combined in various ways to perform complex queries on a knowledge graph.

3.7.5 Knowledge graph embedding

Knowledge graph embedding (KGE), also referred to as knowledge representation learning (KRL), plays an important role in the domain of knowledge representation learning, providing a compact, continuous vector space representation of entities and relations. These representations enable machines to harness the richness of structured knowledge in tasks like link prediction, relation extraction, and recommendation systems [249]. In this section, we present the methodologies and frameworks that were used in this study.

Knowledge graph representation learning

The goal of representation learning in KGs is to learn an accurate vectorial representation for both entities $\mathcal{E}$ and relationships $\mathcal{R}$. A knowledge graph $\mathcal{K}$ can be expressed as the list of its edges, in the form of triples (h, r, t), where the head and tail entities $h, t \in \mathcal{E}$ are connected via a relationship $r \in \mathcal{R}$.

To train a KG embedding model, two types of triples are sampled from the KG $\mathcal{K}$:

- Positive triples: which correspond to actual relationships in $\mathcal{K}$, in the form (h, r, t).

- Negative or corrupted triples: which do not exist in $\mathcal{K}$. These triples are synthetically produced by corrupting either the head or tail entities in a positive triple with a random entity h', leading to either (h, r, t') or (h', r, t).

In KG embedding tasks, the goal is generally to optimise a scoring function $f(h, r, t)$ such that positive triples receive higher scores and negative triples get lower scores.

In a typical training process, the KG is split into two distinct triple sets: training and testing. A commonly used loss function during training is the Margin Ranking Loss function [250], which is employed to optimise model parameters.

For each positive triple t^+ in the graph, a corresponding negative triple t^- is produced. The Margin Ranking Loss function assigns a real value to the pair as follows:

$$L(t^+, t^-) = \max(0, \lambda + (f(t^-) - f(t^+)),$$

where λ denotes the margin parameter. This loss function penalizes pairs where the difference between the scores of positive and negative triples is less than the margin.

Evaluation Metrics for KG Embeddings

The performance of embedding models in graph completion tasks is often assessed using metrics like *hits@10*. For each positive triple $t^+ = (h, r, t) \in \mathcal{K}_{test}$, negative triples in the forms (h', r, t) and (h, r, t') are generated, excluding any negative triple in $\mathcal{K}_{test}$. The rank of t^+ is its position in the list of scores, ordered in descending order, relative to its negative counterparts [249, 250]. The *hits@10* metric, given by:

$$\text{hits@10} = \frac{|\{t \in \mathcal{K}_{test} : rank(t) \leq 10\}|}{|\mathcal{K}_{test}|}$$

quantifies the proportion of positive triples ranked within the top 10 positions. A value closer to 1 signifies the model's proficiency in discerning true triples from corrupted ones.

PyKEEN: A Python Library for Knowledge Graph Embeddings

The Python KnowlEdge EmbeddiNgs (PyKEEN) library [250] is an open-source Python package designed to train and evaluate KG embedding models. PyKEEN provides a selection of embedding models, loss functions, regularisers, training and evaluation methods.

DistMult: A KG Embedding Model

KG embedding methods can be broadly categorised into translational-based models, matrix factorisation-based models, and neural-based models. Translational models represent relationships as spatial translations between entity points in the embedding space. Matrix factorisation models use linear algebra to break down matrices capturing entity-relation interactions. Meanwhile, neural models harness deep neural networks to adaptively learn intricate relationships in the data [158, 251].

Among these, matrix factorisation models represent entity-relation interplays in a lower-dimensional form. DistMult, introduced by Yang et al. [252], is a standout example. Building on the RESCAL model [253], DistMult simplifies representation by using diagonal matrices for relations, ensuring efficient computation and scalability for vast knowledge graphs.

The scoring function employed by DistMult to assess the compatibility of entities with a relation is:

$$f(h, r, t) = \mathbf{h}^\top \mathrm{diag}(\mathbf{r})\mathbf{t} = \sum_{i=0}^{d-1} \mathbf{r}_i \cdot \mathbf{h}_i \cdot \mathbf{t}_i.$$

This function essentially multiplies the corresponding elements of the entity and relation vectors, producing a cumulative score that indicates the fit of the head and tail entities with the relation.

Note that DistMult's scoring function inherently assumes symmetrical relations between entities, making it less suitable for anti-symmetric relations [252].

Combining embedding vectors

In the context of link prediction tasks, it is often necessary to combine the embedding vectors of two entities to infer the potential existence of a link.

A simple vector transformation is the concatenation of the two vectors, which results in a final vector of dimension $2d$ where d is the dimensionality of the initial vectors.

For more compact and computationally efficient transformations, *embedding-aggregating* *transformations* can be considered, which preserve the original vectors' dimensionality. These transformations were proposed by Grover and Leskovec in [254], and are detailed in Table 3.2.

Name	Transformation
Average	$(\mathbf{v}_1 + \mathbf{v}_2)/2$
Hadamard	$\mathbf{v}_1 \odot \mathbf{v}_2$
L1	$\lvert \mathbf{v}_1 - \mathbf{v}_2 \rvert$
L2	$(\mathbf{v}_1 - \mathbf{v}_2)^2$

Table 3.2. Overview of embedding-aggregating transformations. In these transformations, proposed by [254], $\mathbf{v}_1$ and $\mathbf{v}_2$ denote the embeddings of two nodes in a graph. The Average transformation computes the element-wise average of the two embeddings. The Hadamard transformation computes the element-wise product of the two embeddings, while the L1 transformation computes the element-wise absolute difference, and the L2 transformation computes the element-wise squared difference. Each transformation results in a new embedding that represents the aggregate information from $\mathbf{v}_1$ and $\mathbf{v}_2$.

With these transformations, the combined embeddings can be used directly in a binary classifier, such as a logistic regression model, to predict the existence of a link. The selected transformation and the binary classifier form a complete model for the link prediction task.

Chapter 4

ORVAL: A platform for the exploration and interpretation of oligogenic predictions

In the post-genomic era, the vast amount of genetic data available to clinicians and researchers has underscored the importance of bioinformatics predictive approaches. These tools are essential for predicting and prioritising variants and genes that could serve as potential diagnosis biomarkers for genetic diseases.

Detecting multi-locus patterns in oligogenic diseases presents a increased challenge due to the combinatorial explosion. Yet, the Variant Combination Pathogenicity Predictor (VarCoPP) [83] and the Digenic Effect Predictor [82] have demonstrated their effectiveness in predicting pathogenic variant combinations and their effects. However, these methods often yield results that can be challenging to interpret, limiting their accessibility to domain experts. The lack of a straightforward process to transition from patient variant data to predictive insights further restricts their broad adoption by researchers and clinicians.

Addressing this challenge required a system that integrates interpretability techniques and background knowledge to provide meaningful insights. This was crucial for ensuring that the predictive methods developed for oligogenic diseases were not only accurate but also comprehensible and actionable for domain experts.

With this objective in mind, we introduced ORVAL (Oligogenic Resource for Variant Analysis). ORVAL is a web platform designed to bridge the gap between complex machine learning predictions and domain-specific insights. It allows clinicians and researchers to directly submit patient data, annotate them, apply filters and predict the potential pathogenicity and effect of individual oligogenic variant combinations. More importantly, it incorporates analytical tools to contextualise these predictions within background knowledge, such as pathways, protein-protein interactions, and cellular locations. In doing so, ORVAL transforms raw predictions into actionable insights, making these predictive approaches more accessible and interpretable.

In this chapter, we will describe the technical aspects of this original platform including the architecture of its different components (Section 4.1). One crucial component of ORVAL is an annotation database, providing comprehensive genetic information of human variation ; we will therefore provide an overview of its model and data integration process (Section 4.2).

We will then dive into the general workflow of variant data processing, from the submission to the final results presented to the user (Section 4.3).

In Section 4.4, we will present all different types of predictive results and visualisations provided by ORVAL, underscoring its ability to contextualise predictions at different biological level. We will then demonstrate the usefulness of these analyses on independent oligogenic studies and the impact the ORVAL platform had on genetic research since its public availability (Section 4.5).

Finally, we will show the last improvements of the platform, demonstrating its continuous expansion towards more accurate models and precise annotation data as well as its future technological transfer (Section 4.6).

The ORVAL platform has been published in *Nucleic Acids Research* [165] and is available online at: `https://orval.ibsquare.be`.

4.1 The ORVAL platform architecture

The ORVAL platform was conceptualised as a web application available publicly, able to interact with a genetic annotation database and machine-learning models and providing a user-friendly interface with dynamic visualisations to the end-user. It was envisioned as a platform that could grow and evolve with the availability of new models, data sources and analyses.

This web platform was designed following good software practices with a multi-tier and modular architecture, which comprises server-side services and a web application that processes user requests through the backend and displays web pages via the frontend. Figure 4.1 provides an overview of the key web architecture components, services, and their interdependencies within the system.

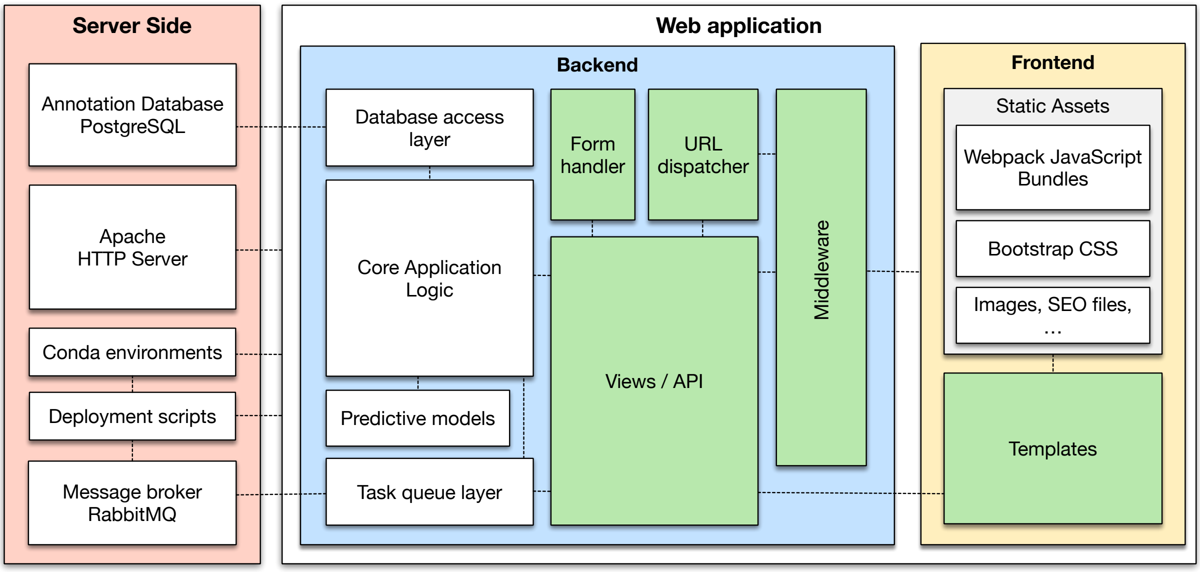


Figure 4.1. General overview of the ORVAL platform web architecture and key components. Interdependencies and communications channels between components and services are represented as dotted lines.

4.1.1 The Django Model-View-Controller architecture

The ORVAL web application is built using the Django web framework [1], which implements a Model-View-Template (MVT) architecture, a variation of the Model-View-Controller (MVC) pattern. Django promotes a clean, pragmatic design with a focus on rapid development.

We represent the Django framework key components integrated into the web application architecture as green boxes in Figure 4.1 and detailed next.

Note that the **Model** component of the Django architecture, abstracting the database interactions via an Object-Relational-Mapping (ORM) was replaced, in ORVAL, by a custom database access layer, to enable the fine-tuning of query performances via native SQL queries.

Additionally, Django provides entry point scripts to launch the web application and for communicating with the HTTP server via the Web Server Gateway Interface (WSGI). It also comes with a settings file containing all the configuration of the project.

4.1.2 Server side: infrastructure, environments and deployment

The ORVAL web platform is hosted on two servers: one for production (accessible by users) and another for internal testing (development). The production server is a GNU/Linux Ubuntu 16.04 x86_64 machine with 125 GB RAM, 40 CPUs, and 9.1 TB of storage.

An Apache HTTP server was configured for serving the web application through the Web Server Gateway Interface (WSGI) and for serving static files. It also handles server logs, security settings (SSL configuration, request body size limits, IP restrictions), and virtual host configurations.

Additionally, a message broker, RabbitMQ[2] was installed and configured to handle messages coming from the asynchronous task processing layer of the application and storing them in its message queue.

[1] Django Web Framework: https://www.djangoproject.com/

[2] RabbitMQ message broker: https://www.rabbitmq.com

To orchestrate all underlying and depending services, deployment bash scripts were developed. They can automatise the control of services such as the HTTP Server, job workers, message broker, javascript bundler and the activation of Conda[3] environments for loading Python dependencies.

4.1.3 Backend: handling requests, fetching data and core logic

The web application of ORVAL handles and processes user requests via its backend.

Users requests are first going through different modules implemented via the Django framework. The *middleware* is composed of different modules each responsible of different tasks such as caching, maintaining security against malicious injections, handling errors and user sessions. An *URL dispatcher* maps incoming URLs to determine the appropriate view to handle a request. A *form handler* facilitates the generation, parsing and validation of the variant data submission form. The *views* handle the processing of requests by calling functions from other application components and generating structured responses. Most of the views also support REST (Representational State Transfer) GET requests, to enable the automation of some processes such as checking a submitted job status or fetching results in the JSON format.

The views call functions from the core application logic either directly or indirectly via a *task queue layer*. Task queues are used as a mechanism to distribute work across threads. This layer has been implemented using the Celery[4] python package, able to create multiple dedicated workers monitoring a queue for new tasks to perform. We configured Celery to communicate with the message broker RabbitMQ, able to mediate incoming messages coming from the client, deliver the messages to the workers and handle cases where all workers are busy and messages need to be queued. We created three different queues with different level of priorities, to handle the main pipeline, fast-running analyses and job status updates through emails. Additionally, we implemented the Flower utility, enabling the execution monitoring of all jobs submitted. This task queue layer ensures

[3] Conda environment: https://anaconda.org/anaconda/conda

[4] Celery task queue: https://pypi.org/project/celery/

the high availability, concurrent access and scaling of jobs processed by the application.

The *core application logic* encompasses data pre-processing, filtering, annotation and post-processing. It leverages machine-learning models such as VarCoPP [83] and the DE Predictor [81, 82], cached in memory, to make predictions or generate explanations, which are then transformed into memory-efficient data structures for further analysis.

The *database access layer* provides access to the PostgreSQL annotation database. This layer has been implemented via native SQL calls to handle complex, multi-step SQL queries, enabling the fast retrieval of information.

4.1.4 Frontend: user interface and dynamic data exploration

The platform interface has been designed in a modular way by using Django's templates, responsible for the structure and layout of web pages, using a mix of HTML and Django's template language. Additionally, we used the Bootstrap framework [5], which ensures a responsive web design that adapts to various screen sizes and devices.

To handle user interactions, data visualisation and asynchronous behaviours, JavaScript modules have been developed in JavaScript ES6. Charts and networks were designed with the D3.js library [6]. JavaScript modules and library dependencies were bundled with the Webpack technology [7] and we used the Babel JavaScript transcompiler [8] for maximised cross-browser compatibility.

In summary, the ORVAL platform has been carefully designed using a combination of technologies such as Django, Apache, PostgreSQL, RabbitMQ, and various JavaScript and Python libraries to create a robust, efficient, and user-friendly web platform.

[5] Bootstrap web framework: https://getbootstrap.com/
[6] D3.js visualisation library: https://d3js.org/
[7] Webpack module bundler: https://webpack.js.org/
[8] Babel transcompiler: https://babeljs.io/

4.2 The oligogenic annotation database

The annotation database is a key component of ORVAL and the general software ecosystem developed by our research team. Its goal it to facilitate the integration and fast retrieval of relevant information on human genetics across different biological levels: variants, genes, proteins, interactions and pathways. The type of information it integrates, coming from public biological databases or as a storing of pre-calculated metrics, has been chosen according to the needs of predictors integrated in ORVAL.

4.2.1 Database architecture and data model

The annotation database infrastructure is managed by PostgreSQL, a free and open-source relational database management system (RDBMS). As of April 2023, the annotation database contains 896 GB of data and can handle up to 250 concurrent accesses. Regular automated dumps are performed to save backups.

The data stored in this database is structured according to an entity-relation schema, presented in Figure 4.2. The information is stored in different tables to avoid redundancy and to facilitate data integration.

To enable the fast retrieval of information and the joining of data across tables, some table fields were defined as primary key or indexed. Genetic variants were uniquely identified by keys made of the concatenation of chromosome, position, reference allele and alternative allele, separated by a comma, in the *variant* and *exac* table. Gene and protein level information were also uniquely defined respectively by their Ensembl [174] gene identifier and their UniProt accession [178].

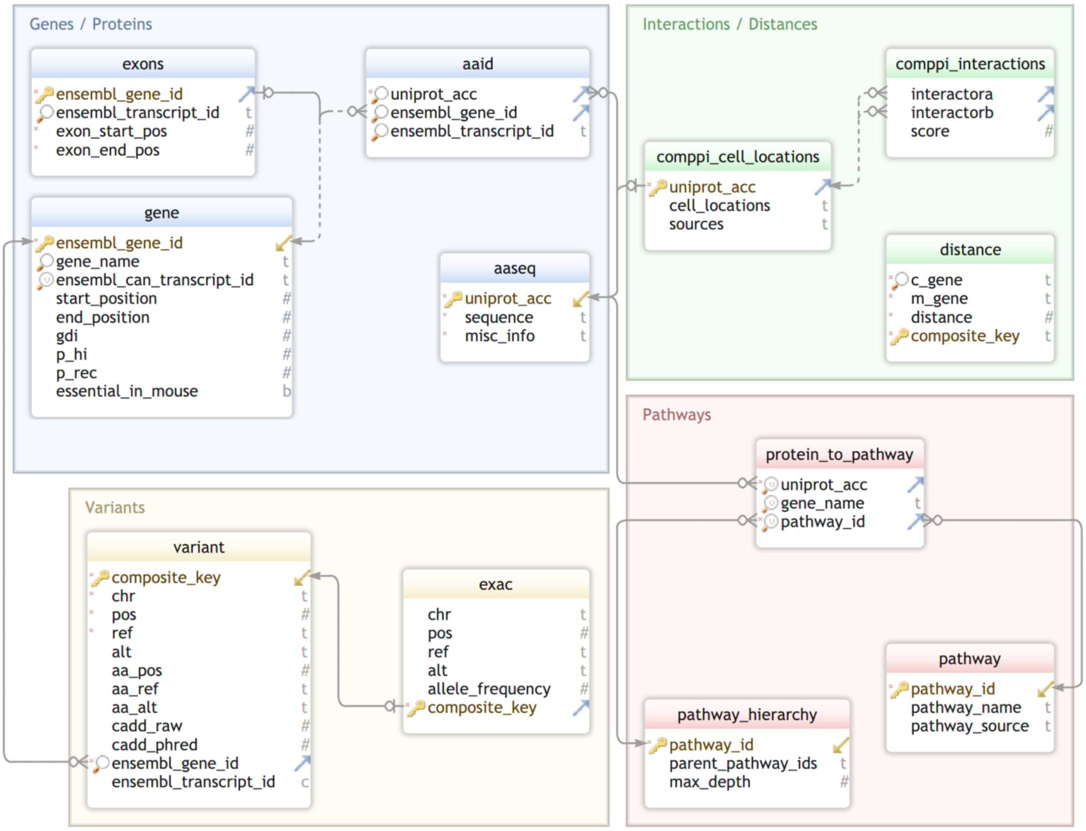


Figure 4.2. Database schema of the annotation database. Tables are represented grouped by the type of biological data they contain. Arrows represent foreign key relationships between tables for specific fields. The type of symbol at the base of the arrow represents the arity of relationships: $> o$ = zero-to-many ; $|o$ = zero-to-one.

4.2.2 Data integration pipeline

The annotation database contains data integrated from multiple public bioinformatics databases or resulting from pre-computed results. These resources include pre-calculated variant pathogenicity predictions and annotations (see Material and Methods - Subsection 3.1.4), variant population frequency databases (see Material and Methods - Subsection 3.1.3), reference databases providing unique identifiers and sequences for genes and proteins (see Material and Methods - Subsection 3.2.1), various collections of features characterising genes (see Material and Methods - Subsection 3.2.2) and a diverse range of functional annotation resources (see Material and Methods - Subsection 3.2.3). To consolidate heterogeneous data sources into a unified database model, an integration pipeline has been designed, consisting of several stages: data retrieval, data pre-processing and consolidation, and data mapping into the database model, all of which work together to ensure the integrity and reliability of the integrated data. We summarise the integration from sources to database tables in Table 4.1.

Source	**Database tables**
dbNSFP [185]	gene
UniProt [178]	aaseq, aaid
CADD [86]	variant
GnomAD / ExAC [21]	exac
ComPPI [255]	comppi_cell_locations, comppi_interactions
Biological distance (HGCS) [256]	distance
KEGG [257]	protein_to_pathway, pathway
Reactome [188]	pathway_hierarchy, protein_to_pathway, pathway
Ensembl [174]	exons, aaid, variant
HGNC [177]	gene, variant, protein_to_pathway

Table 4.1. Summary of mappings operated from data sources into the annotation database tables during data integration process.

Gene data integration

In our integration pipeline, Ensembl [174] gene identifiers are used to uniquely identify genes in the *gene*, *variant*, *exons* and *aaid* table. Additionally, the canonical transcript information, provided by Ensembl, enables the retrieval of canonical information at the protein / pathway level for a given gene. The HUGO Gene Nomenclature Committee (HGNC) [177] resource plays a critical role in maintaining the correctness and consistency of gene names throughout the data integration process. (see Material and Methods - Subsection 3.2.1).

The dbNSFP database [185] serves as a resource for integrating features at the gene level. This database combines multiple gene characteristics derived from bioinformatics tools and resources, such as the haploinsufficency index (p_hi) [182], the Gene Damage Index (gdi) [179] and the essentiality in mouse (essential_in_mouse) [183] (see Material and Methods - Subsection 3.2.2).

Variation data integration

The variation data integration pipeline uses two different sources: the annotated variant file provided by CADD [86] (see Material and Methods - Subsection 3.1.4) and variant frequencies provided by GnomAD [21] (formerly ExAC) (see Material and Methods - Subsection 3.1.3). For both sources, variants are uniquely identified with a composite key generated by concatenating the chromosome, genomic coordinate, reference and alternative allele of each variant.

Multiple gene identifiers might be associated with the same variant in CADD due to the variant's location in overlapping or adjacent genes. Therefore, to enable a one-to-one mapping from variant to gene, a systematic approach was employed, guided by a set of domain-based rules applied with a precedence order to decide which association to retain or discard. The first rules filter and retain variant-to-gene associations having an association with a known Ensembl gene identifier, gene name and associated to a canonical transcript. Subsequently, records are selected based on annotation type, functional consequences, and biotype information. To further refine the selection, the rules consider factors such as gene and transcript coordinates, proximity to splicing sites, and gene naming conventions. The low priority rules, used only if the previous rules do not resolve possible ambiguities, are based on transcript and gene lengths, strand orientation, and gene name length to select a final disambiguated association.

Interactions and distances integration

Protein-protein interactions are integrated from the ComPPI database [255] (see Material and Methods - Subsection 3.3.2). This resource can provide both the information on protein interactions, originally from nine sources, and the sub-cellular location in which the interacting proteins are located from eight datasets. We integrated this information along with scores asserting the confidence of each interaction. For each protein entry, multiple cellular locations are associated with a confidence score as well as a list of source databases for these associations.

The biological distances between pairs of genes are integrated from the Human Gene Connectome Server (HGCS). This metric is based on the distance between related proteins in a protein-protein interaction network from STRINGdb [258]. To enable a faster retrieval of multiple gene pair distances at once, a composite key is created from the concatenation of gene names in alphabetical order.

Pathways integration

The pathway information is integrated from two resources: Reactome [188] and KEGG [257] and linked to the associated UniProt accessions and gene names. Considering the hierarchical nature of Reactome, we integrated in the *pathway_hierarchy* table, for each pathway id, the list of parent pathways ids as well as the depth in the hierarchy (see Material and Methods - Subsection 3.2.3 and Subsection 3.4.2).

4.3 Submission and variant processing pipeline

4.3.1 Overview of the pipeline

The ORVAL web platform consists of a submission form, where users can submit genetic variant data along with filtering criteria, and a variant processing pipeline that first generates variant combinations, annotates them with variant, gene and combination level information, and then predicts which variant combinations may potentially be associated with the disease. The candidate digenic predictions are then used to rank gene pairs and build an interactive oligogenic network that can be further explored. All these description levels are enhanced by known cross-references. Figure 4.3 summarises the workflow and the components of ORVAL.

Job queuing and data privacy

ORVAL manages its variant processing pipeline with a secure asynchronous queuing system where jobs get assigned an Universally Unique Identifier (UUID), for every submission. Users can access a job page to track the status of their submissions and bookmark them for later use. It is also possible to provide an email address to be informed when a job has been completed. A maximum of 5 concurrent jobs per user can be submitted in parallel.

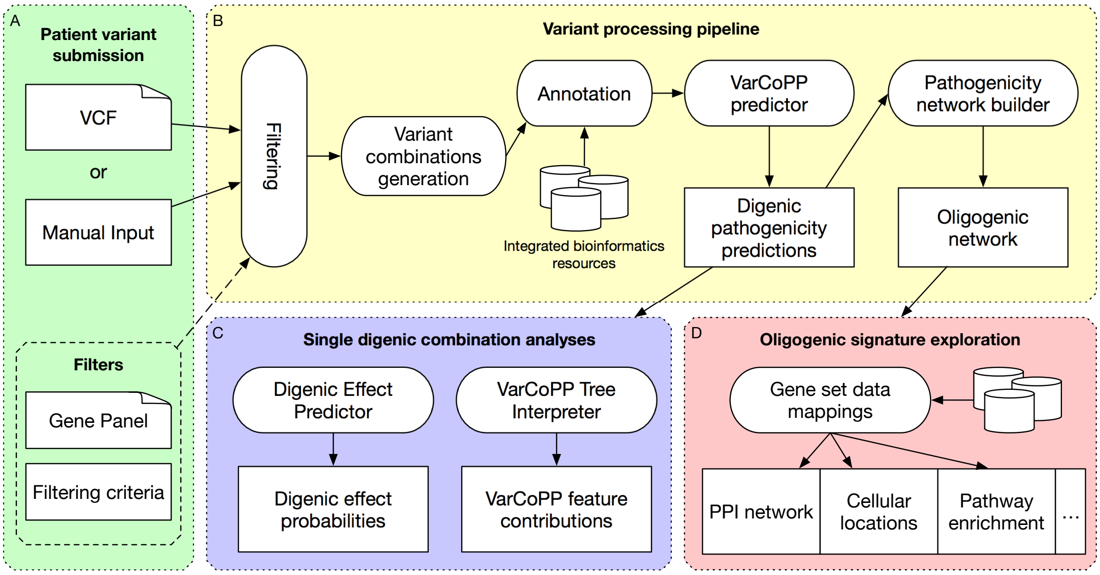

A
Patient variant submission
VCF
or
Manual Input
Filters
Gene Panel
Filtering criteria
B
Variant processing pipeline
Filtering
Variant combinations generation
Annotation
Integrated bioinformatics resources
VarCoPP predictor
Digenic pathogenicity predictions
Pathogenicity network builder
Oligogenic network
C
Single digenic combination analyses
Digenic Effect Predictor
Digenic effect probabilities
VarCoPP Tree Interpreter
VarCoPP feature contributions
D
Oligogenic signature exploration
Gene set data mappings
PPI network
Cellular locations
Pathway enrichment
...

Figure 4.3. ORVAL flowchart highlighting the major components of the platform. **(A)** Users can submit variants using a Variant Call Format (VCF) file or a tab-delimited variant list. The variants can be filtered with some predefined criteria or by using a gene panel. **(B)** Once submitted, variants are processed by a pipeline first applying the selected filters, then generating all di-, tri- and tetra-allelic variant combinations and annotating them using public bioinformatics resources, then predicting which variant combinations may be disease-causing with the VarCoPP predictor. Finally, the disease-causing variant combinations are aggregated at the gene level to build an oligogenic network. **(C)** By selecting a specific digenic variant combination, users can run a predictor to know the DE probabilities and can get an interpretation of the VarCoPP prediction based on its features. **(D)** It is also possible to interact with the oligogenic network to filter and explore specific oligogenic signatures. A dedicated page shows how the selected gene set maps with multiple cross-references to give an insight into the biological context.

In terms of submitted data privacy, the analysis results are provided to the user via a unique private link and are accessible for seven days starting from the time of submission. No input data nor user information is stored. All predicted data are permanently erased from our servers after the seven days period with an automatic procedure. Additionally, all communication with the ORVAL platform is secured via HTTPS. The submitted data is processed by the RabbitMQ queue system, which has been configured for SSL/TLS to secure the data during network transit.

4.3.2 Patient variant data submission

The ORVAL platform accepts a list of variants from a single individual as input. These variants can be entered manually or provided in a Variant Calling File (VCF) (compressed or not) (see Material and Methods - Subsection 3.1.1). Users can also choose to apply filtering options that discard variants based on a given threshold of Minor Allele Frequency (MAF), obtained from the ExAC database (see Material and Methods - Subsection 3.1.3), their genomic and exonic positions or their synonymous effect. Applying these filters is highly recommended to ensure that the remaining variants are in accordance with the variant types that were used to train the predictive methods integrated in ORVAL. Users can also provide a gene panel that will be used to restrict the analysis to only the genes of interest (Figure 4.3.A).

The ORVAL pipeline follows the processing steps described in Figure 4.3.B. This processing is divided in chunks and tasks are handled by an asynchronous task queue as described in Subsection 4.1.3. While a job is running, users can access a page indicating the status of the running job.

4.3.3 Variant annotation and prediction of combinations

All variants are annotated based on the integrated genetic annotation database (see Section 4.2). These annotations are: the variant CADD score [86], the protein sequence from UniProt [178], the gene recessiveness and haploinsufficiency probabilities from the dbNSFP database [259], the Gene Damage Index (GDI) [179] that provides the susceptibility of a gene to disease, and the Biological Distance [256] that shows the biological relatedness between any two genes based on PPI information.

After annotation, ORVAL creates all possible (i.e. bi-allelic, tri-allelic and tetra-allelic) combinations of variants occurring in gene pairs and applies the Variant Combination Pathogenicity Predictor (VarCoPP), a pre-trained machine-learning model, to predict the pathogenicity of each variant combination (see Material and Methods - Subsection 3.6.2).

Each prediction comes with two predictive scores (i.e. a Support Score (SS) and a Classification Score (CS)) whose previously defined thresholds determine whether a variant combination is predicted as potentially disease-causing or neutral. These scores are also assigned confidence labels providing a clear signal to identify the potentially most relevant pathogenic combinations.

Additionally, a gene pathogenicity network is computed using all (Gene$_1$, edge, Gene$_2$) triples where at least one predicted disease-causing variant combination (variant$_1$, variant$_2$)

exists with variant$_1$ $\in$ Gene$_1$ and variant$_2$ $\in$ Gene$_2$. Edges in the gene pathogenicity network are then weighted by the maximum pathogenicity score, considering all underlying variant combinations.

4.3.4 Processing pipeline outcome

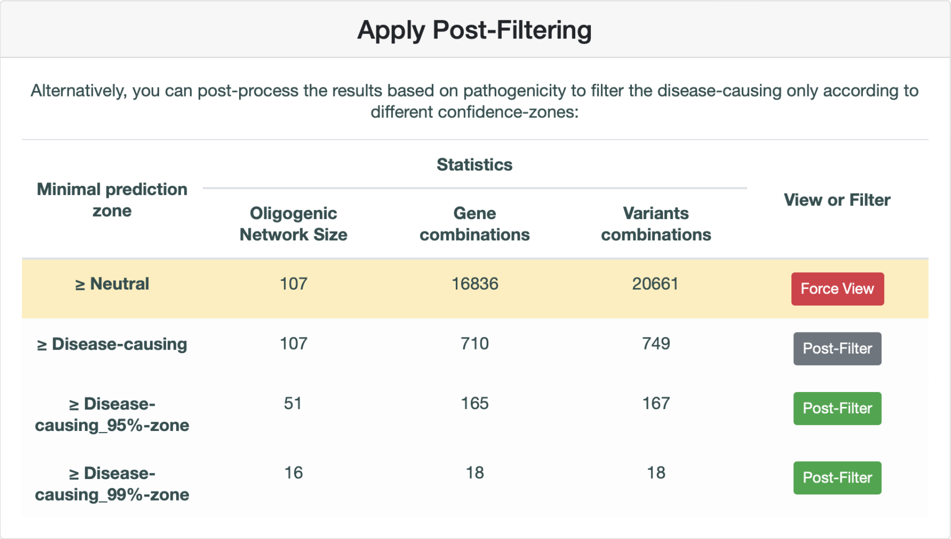
There are multiple possible outcomes for a submitted job:

1. If a job cannot be completed due to an error, a page is loaded with the error message, an email is sent to the user if this email was provided and an error report is sent to the administrator for troubleshooting.

2. If a job finishes successfully but the resulting gene pathogenicity network contains more than 100 genes, then the user lands on an intermediate page where it is optionally possible to filter results according to the predicted pathogenicity high-confidence zones (Figure 4.4).

3. If a job finishes successfully and yields less than 100 genes in the gene pathogenicity network, then the user is redirected to a result page described in Section 4.4 and, if provided, an email is sent to the user.

The results of a job remain available to the user for 7 days. After that period, all results are permanently erased from the server.



Apply Post-Filtering

Alternatively, you can post-process the results based on pathogenicity to filter the disease-causing only according to different confidence-zones:

Minimal prediction zone	Statistics			View or Filter
	Oligogenic Network Size	Gene combinations	Variants combinations	
≥ Neutral	107	16836	20661	Force View
≥ Disease-causing	107	710	749	Post-Filter
≥ Disease-causing_95%-zone	51	165	167	Post-Filter
≥ Disease-causing_99%-zone	16	18	18	Post-Filter

Figure 4.4. Post-filtering page proposing to explore pathogenic combinations with higher stringency Screenshot showing an example of the post-filtering page appearing if the results of a job contain more than 100 genes in the gene pathogenicity network. To avoid the loading and visualisation of too many combinations on the result page, users can choose to retain only variant combinations above a certain pathogenicity prediction zone. For each post-filtering option, the user can know in advance the size of the filtered gene pathogenicity network and the number of gene and variant combinations. After obtaining the filtered results, the user can return to this page to choose another post-filtering option to its convenience. Note that, while not recommended, it is still possible to visualise all predicted results without any post-filtering.

4.4 Exploring predictive results in ORVAL

Once variants have been processed and the predictions are available, the main results are presented to the user. In order of appearance, these results comprise: the gene pathogenicity network inferred from predicted pathogenic variant combinations, a summary statistics of all gene pairs with aggregated statistics and, finally, a detailed perspective on all predicted variant combinations that were found in the patient's data.

Note that ORVAL organises these results within a dynamic interface that provides guidance, at every step, on how to use the tool and interpret the results. Help buttons with summarised guidance are present in all result panels, while warning messages provide information on how to tackle exceptional issues that may arise during data submission or exploration of results. The documentation page of ORVAL contains a standalone in-depth guide discussing the data submission, filtering and annotation process of the users data, as well as of the predictive methods and the exploration of the results with case examples.

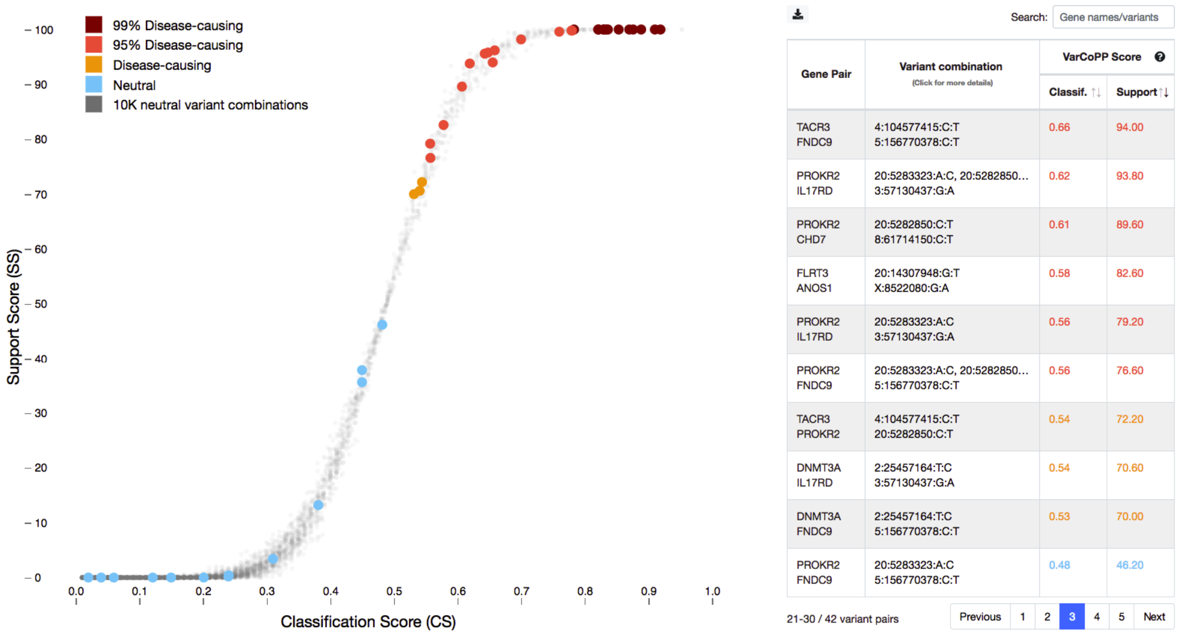
Additionally, every table can be downloaded as tab-separated values (TSV) format, while the oligogenic network (Figure 4.11), module network and PPI network can be downloaded in the GraphML format, so they can be easily imported in network analysis programs. All objects can be downloaded in the form they were originally obtained or after the application of any post-hoc filters (e.g. gene and edge filters or custom search query). The S-plot figure is available publication-ready in PNG format.

In the following section, we will present all types of results proposed by the platform starting from the more granular view, at the variant combination level, and finishing with the pathogenicity gene network view.

4.4.1 Analysing predicted variant combinations

In the most detailed view of the result page, all digenic variant combination predictions, as predicted by VarCoPP [83] (see Material and Methods - Subsection 3.6.2)), can be visualised in the form of an interactive S-plot (Figure 4.5) based on the two pathogenicity scores provided by VarCoPP. Each point in the curve is a digenic variant combination whose colour represents its pathogenicity confidence. A dynamic summary table next to the S-plot provides a complete list of all visualised combinations, ordered from high to low pathogenicity scores. Each combination is linked to additional detailed information.

By selecting a specific digenic combination in the S-plot or the summary table, the detailed information page opens, presenting details on the contributing predictive factors of VarCoPP, the Digenic Effect prediction (provided that the combination is predicted to potentially be pathogenic) and other useful annotations specific to the selected variant combination.



Gene Pair	Variant combination (Click for more details)	VarCoPP Score	
		Classif.	Support
TACR3 FNDC9	4:104577415:C:T 5:156770378:C:T	0.66	94.00
PROKR2 IL17RD	20:5283323:A:C, 20:5282850... 3:57130437:G:A	0.62	93.80
PROKR2 CHD7	20:5282850:C:T 8:61714150:C:T	0.61	89.60
FLRT3 ANOS1	20:14307948:G:T X:8522080:G:A	0.58	82.60
PROKR2 IL17RD	20:5283323:A:C 3:57130437:G:A	0.56	79.20
PROKR2 FNDC9	20:5283323:A:C, 20:5282850... 5:156770378:C:T	0.56	76.60
TACR3 PROKR2	4:104577415:C:T 20:5282850:C:T	0.54	72.20
DNMT3A IL17RD	2:25457164:T:C 3:57130437:G:A	0.54	70.60
DNMT3A FNDC9	2:25457164:T:C 5:156770378:C:T	0.53	70.00
PROKR2 FNDC9	20:5283323:A:C 5:156770378:C:T	0.48	46.20

21-30 / 42 variant pairs

Figure 4.5. An S-plot representing the classification of all digenic combinations as being neutral (in blue) or potentially disease-causing (from orange to dark red) depending on the predicted VarCoPP Classification Score (CS) and Support Score (SS). The table on the right shows the gene pair, variants and prediction scores.

The pathogenicity predictor feature contribution analysis (Figure 4.6) aims to explain the decision made by the predictor. It shows to the user, in the form of box plots, the relative contribution of each feature used by the VarCoPP model for either the positive (red colour) or the neutral (blue colour) class. This analysis is run in the background using the *treeinterpreter* library[9]. Note that this type of post-hoc explanation technique assumes an additive effect of features and do not capture complex interactions between them [95, 104]. Additionally, this type of explanation does not provide a global understanding of the model behaviour. Nevertheless, it can provides some idea on the main contributing factors explaining the prediction of interest (see Material and Methods - Subsection 3.5.4).

[9] Ando Saabas, “treeinterpreter”, GitHub, `https://github.com/andosa/treeinterpreter`

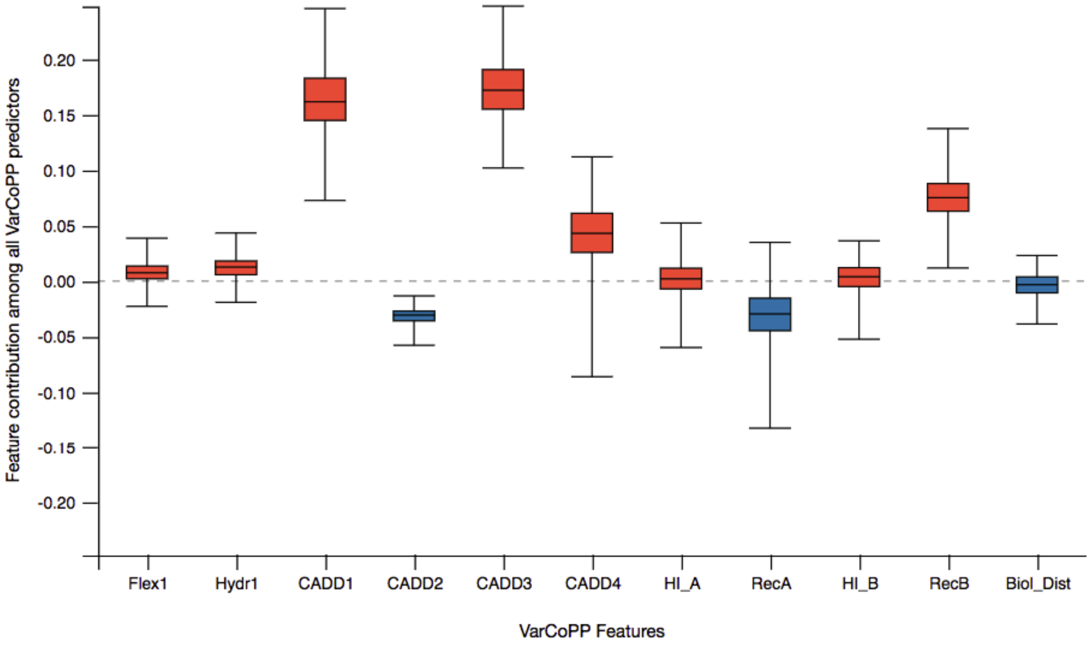


Figure 4.6. Feature contribution analysis, represented as a boxplot chart. It displays, for a specific predicted digenic combination, the relative contribution of each predictive features into the disease-causing class (in red) or neutral (in blue). Note that this analysis is performed with the *treeinterpreter* method, specifically designed for interpreting Random Forest models.

The Digenic Effect predictor assesses the likelihood that a given digenic variant combination falls into one of three distinct classes: *True Digenic*, where both variants must be co-inherited for the disease to manifest; *Monogenic + Modifier*, where a primary gene variant establishes the diagnosis and a secondary variant in a modifier gene alters the phenotype; and *Dual Molecular Diagnosis*, where each gene variant follows a classic Mendelian inheritance pattern and coexists in the same individual, resulting in a composite phenotype influenced by both conditions [82] (see Material and Methods - Subsection 3.6.3). To aid in interpretation, the predicted probabilities for each class are visualized through a radar plot (Figure 4.7). This feature allows users to interpret and better understand the complex genetic interactions at play while also informing on subsequent clinical or research decisions.

Additionally, biological annotations associated with the digenic combination are provided at the bottom of the detailed information page with cross-references to other bioinformatics resources. The users can get information at the gene level (e.g. gene name, Ensembl gene ID [260], recessiveness and haploinsufficiency probability), variant level

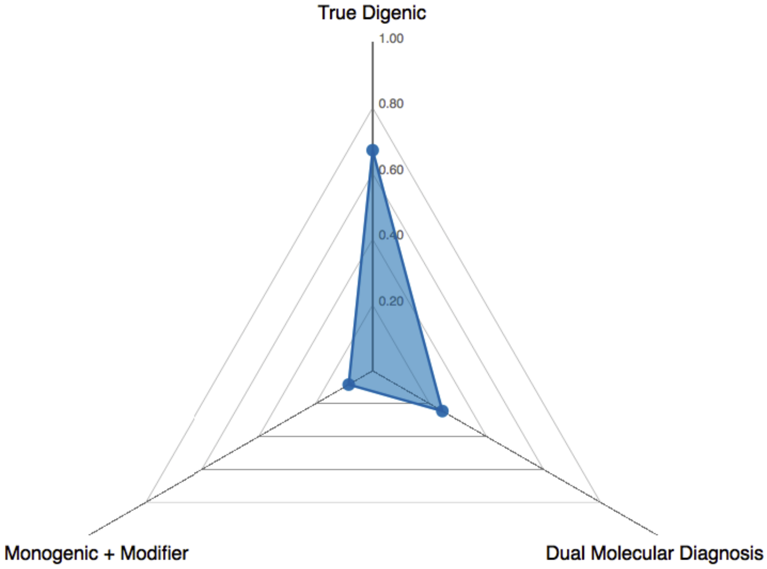


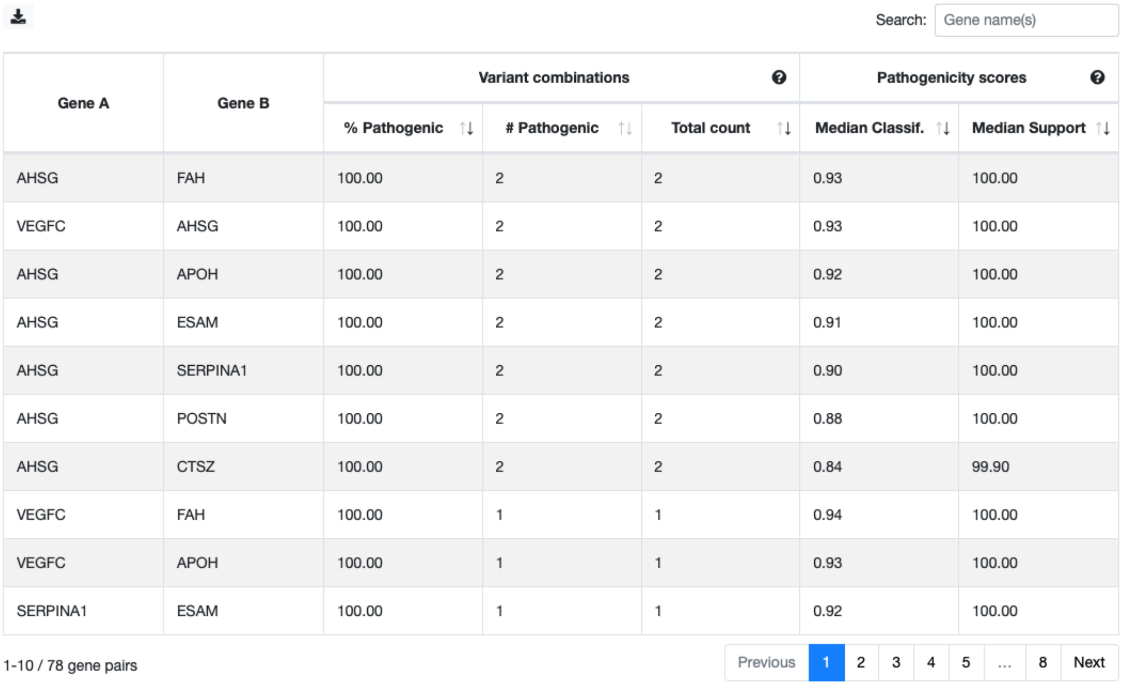
Figure 4.7. A radar plot, displayed for a specific digenic combination, showing the probabilities for each class of digenic effect predicted by the Digenic Effect Predictor [82]. In the showcased results, the highest probability correspond to the True Digenic class.

(e.g. zygosity, dbSNP ID, allele frequency) and gene pair level (e.g. information on the biological distance of the two genes).

4.4.2 Gene pair aggregation and ranking

The Gene Pair Ranking panel (middle of the Results page, Figure 4.8) aggregates the information generated for each variant combination at the level of the gene pair, providing an insight on the pathogenicity of the gene pairs in the data. This information is displayed in a table that includes summary statistics per gene pair, such as the percentage and number of predicted candidate disease-causing variant combinations, as well as the median pathogenicity scores. The gene pairs are initially ranked according to their percentage of candidate pathogenic combinations, followed by the median pathogenicity scores.

Note that we purposely proposed different way of reordering the table, considering our limited knowledge of collapsing strategies of variant combinations and burden scores. With further validation of these approaches, a more robust ranking could be proposed.



Gene A	Gene B	Variant combinations			Pathogenicity scores	
		% Pathogenic	# Pathogenic	Total count	Median Classif.	Median Support
AHSG	FAH	100.00	2	2	0.93	100.00
VEGFC	AHSG	100.00	2	2	0.93	100.00
AHSG	APOH	100.00	2	2	0.92	100.00
AHSG	ESAM	100.00	2	2	0.91	100.00
AHSG	SERPINA1	100.00	2	2	0.90	100.00
AHSG	POSTN	100.00	2	2	0.88	100.00
AHSG	CTSZ	100.00	2	2	0.84	99.90
VEGFC	FAH	100.00	1	1	0.94	100.00
VEGFC	APOH	100.00	1	1	0.93	100.00
SERPINA1	ESAM	100.00	1	1	0.92	100.00

1-10 / 78 gene pairs

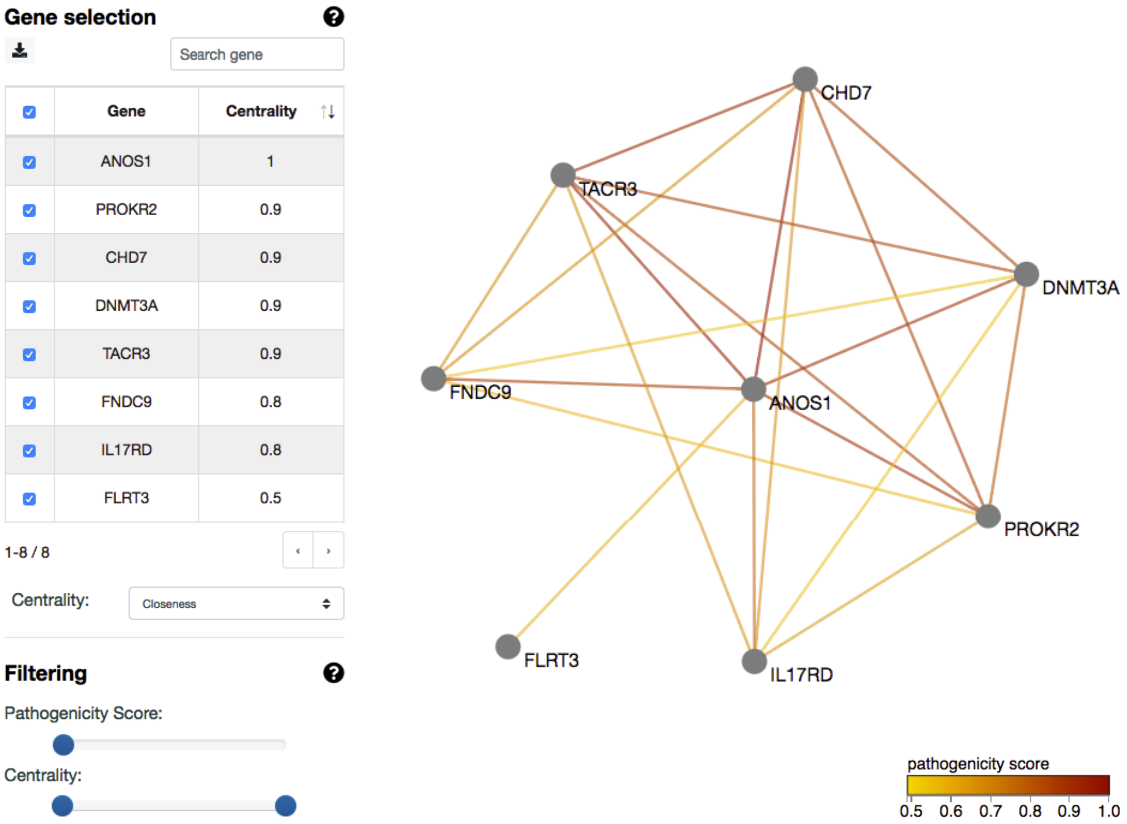
Figure 4.8. A gene pair table provides statistics on all gene pairs corresponding to all variant combinations. This table provides statistics on the percentage and absolute number of pathogenic variant combinations for each gene pair, and the median pathogenicity scores provided by VarCoPP (i.e. the Support Score and the Classification Score).

4.4.3 From predictions to contextualised oligogenic networks

The gene pathogenicity network

As a top-level aggregation, a dynamic visualization of a *gene pathogenicity network* is presented to the user (see Figure 4.9). In this network, nodes symbolise genes, and edges connect gene pairs with at least one predicted pathogenic variant combination. The colour of edges varies depending on the *Pathogenicity Score*, calculated as the maximum VarCoPP Classification Score (CS) considering all variant combinations associated with a gene pair.

All genes in the network are listed in a table along with precomputed centrality measures: degree and closeness centrality (see Material and Methods - Subsection 3.3.1). *Degree centrality* serves as a local measure, indicating a gene's frequent involvement in pathogenic interactions within its immediate network neighbourhood. A high degree centrality suggests that the gene could be a key player in multiple pathogenic combinations. On the other hand, *closeness centrality* is a more global measure, capturing a gene's av-



Gene selection

Search gene

☑	Gene	Centrality
☑	ANOS1	1
☑	PROKR2	0.9
☑	CHD7	0.9
☑	DNMT3A	0.9
☑	TACR3	0.9
☑	FNDC9	0.8
☑	IL17RD	0.8
☑	FLRT3	0.5

1-8 / 8

Centrality: Closeness

Filtering

Pathogenicity Score:

Centrality:

Figure 4.9. An interactive oligogenic network built from all gene pairs having at least one predicted candidate combination. The edges are coloured based on a pathogenicity score (highest Classification Score (CS) for a pair). The genes can be filtered out manually or based on their centrality (degree or closeness). The edges can be pruned based on the pathogenicity score.

erage proximity to all other pathogenic genes in the network. A high closeness centrality could indicate the gene's broader involvement across the network. Users can use this table to selectively remove genes based on these local and global topology metrics.

Additionally, two types of dynamic filters are proposed to the user:

- **A pathogenicity score filter**: setting the minimum threshold of pathogenicity score to include an edge in the network
- **A centrality filter**: setting the minimum threshold for the chosen centrality measure to include a node in the network

Users can also select specific genes in the network to obtain further information. This action opens a side panel that shows information about the gene and about the set of genes in the same connected component, called an *Oligogenic Module* in ORVAL. This side panel shows some module-relevant metrics, such as the size, the graph density (see Material and Methods - Subsection 3.3.1), the average pathogenicity score, as well as a summarised pathway information for the involved genes.

This panel also enable the user to contextualise the selected gene module of interest within protein-protein interactions, protein cellular locations and biological pathways. This take the form of a separate page for each analysed gene module, which analyses are presented next.

Contextualisation in protein-protein interactions

A protein-protein interaction network is built from the set of proteins belonging to the selected gene module, using the ComPPI database [255] (Material and Methods - Subsection 3.3.2).

The resulting network is visualised as an interactive circle-shaped network where the proteins corresponding to the gene module and external proteins directly interacting with them are represented (Figure 4.10). To limit the size of the network, these external proteins are represented only if they interact with at least two proteins of the selected gene module. The cellular location of every protein in the network is represented as an interactive pie chart that can be used to highlight the proteins from a specific location in the network.

This level of contextualisation can serve as a way to validate the plausibility of these genetic interactions and can assist geneticists in understanding potential epistatic effect that could be caused by the direct and indirect interactions between proteins encoded by the pathogenic gene module.

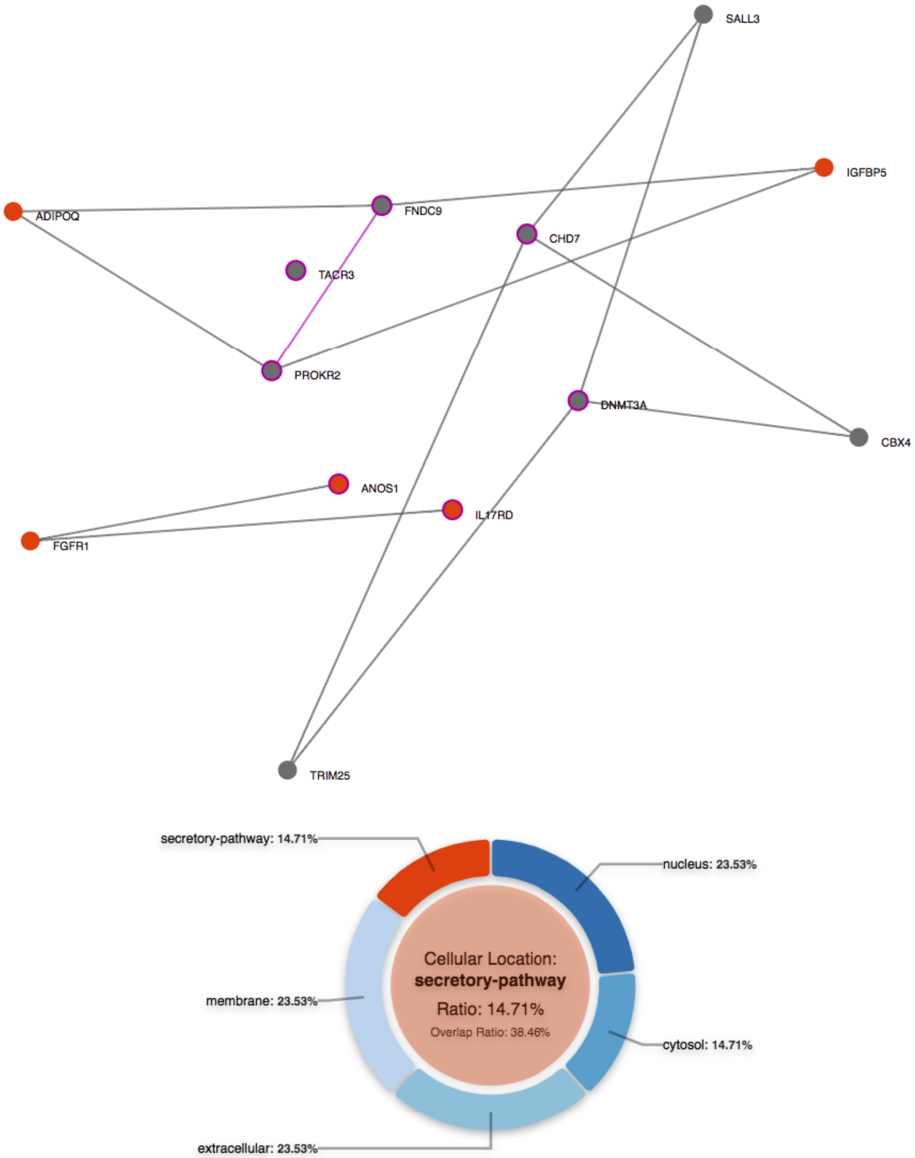


Figure 4.10. A protein-protein interaction (PPI) network where the central nodes circled in purple represent the proteins from a selected oligogenic module and the external nodes are the first-level interactors. Direct interactions (e.g. FNDC9-PROKR2) are coloured in purple. A pie chart showing the protein cellular locations is used to highlight the corresponding nodes in the network (here, secretory pathway).

Contextualisation in pathway information

The biological pathways associated with the gene module can be explored in the *Pathway Mappings* view. This view provides a mapping of the module's genes into pathway terms from the Reactome ontology (see Material and Methods - Subsection 3.4.2). It is represented graphically as a Tree-map, i.e. a plot where boxes represent nested pathways according to their hierarchical level in the pathway hierarchy and whose size is determined by the number of genes they contain. Detailed information about the genes involved in each pathway level is also shown in a dynamic table.

This pathway mapping provides geneticists with an insight on the biological pathways that could potentially be affected when genes of the predicted gene module are disrupted. The hierarchical relationship of Reactome enable to capture this relationship at different levels of granularity. This, in turns, facilitate the interpretation of the predicted results and their validation in comparison to domain knowledge and what is known about the patient suspected disease or symptoms.

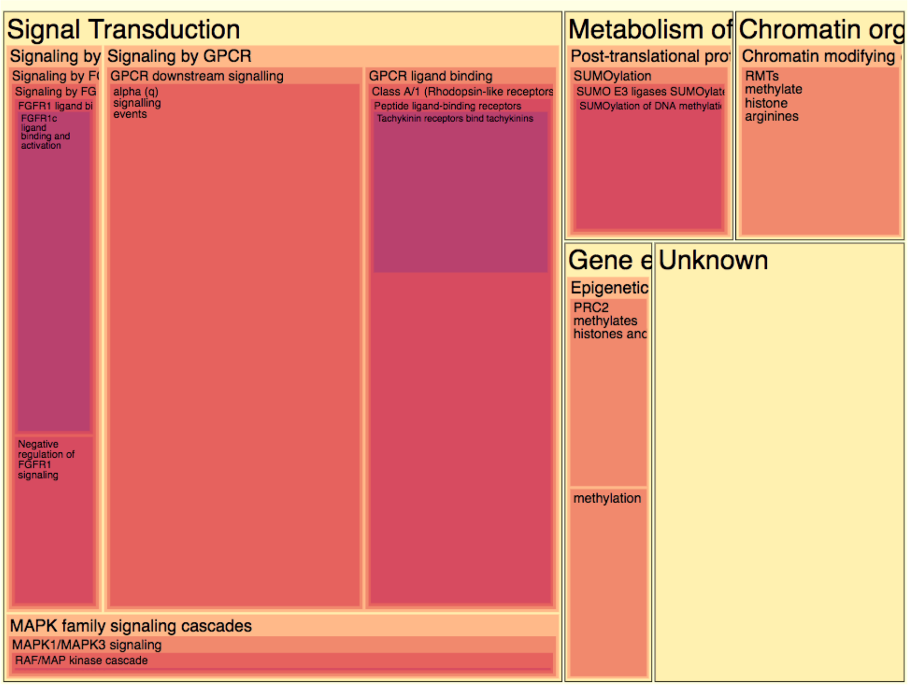


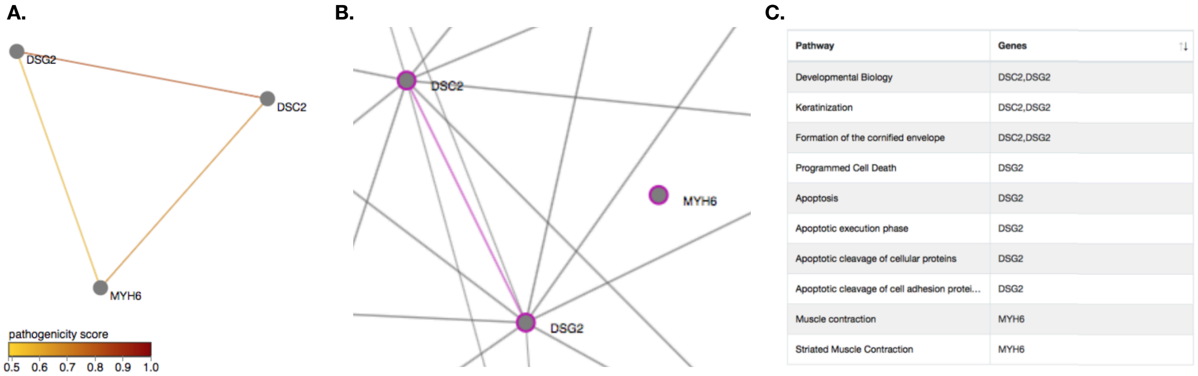
Figure 4.11. A Tree-map representing the Reactome ontology sized proportionally to the number of mapped genes from the oligogenic module and colour according to the level on the ontology hierarchy. On this example, the most represented pathways are part of the Signal Transduction pathway hierarchy.

4.5 Validation and application in genetic studies

4.5.1 Evaluation on independent oligogenic studies

To illustrate ORVAL's relevance for geneticists and clinicians, we briefly discuss here the results for some recently published cases that are associated with diseases having high genetic and phenotypic heterogeneity and show indications of oligogenicity. These cases were not part of the training data underlying the predictive models of ORVAL.

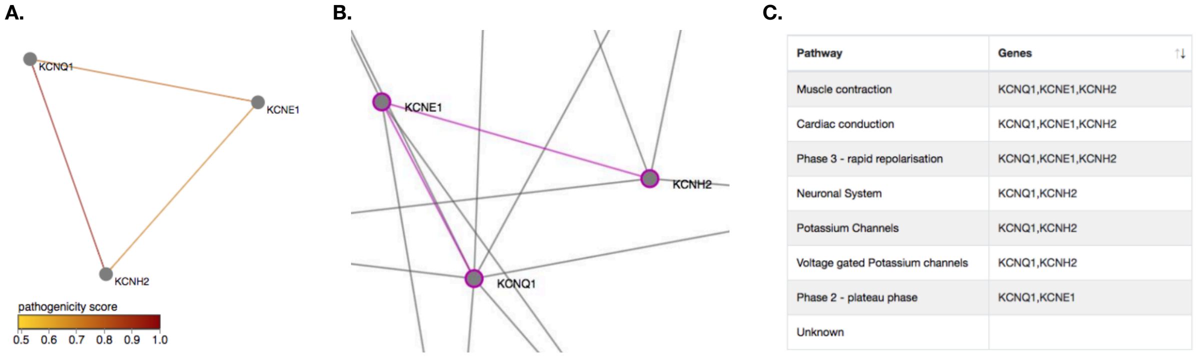
ORVAL supported the suspicions of oligogenicity for a patient with mild hypertrophic cardiomyopathy, carrying three potentially causative variants in the genes: MYH6, DSC2 and DSG2 [261]. All variant combinations were predicted as candidates with high confidence, creating a trigenic oligogenic network. The integrated PPI network informed about the physical connection between proteins DSC2 and SSG2, while the pathway treemap illustrated the involvement of the genes DSC2 and DSG2 in cell apoptosis, and of MYH6 in muscle contraction, further supporting that they can contribute in different phenotypes that can be blended in an individual: arrhythmogenic cardiomyopathy and hypertrophic cardiomyopathy, respectively (Figure 4.12).



Pathway	Genes ↑↓
Developmental Biology	DSC2,DSG2
Keratinization	DSC2,DSG2
Formation of the cornified envelope	DSC2,DSG2
Programmed Cell Death	DSG2
Apoptosis	DSG2
Apoptotic execution phase	DSG2
Apoptotic cleavage of cellular proteins	DSG2
Apoptotic cleavage of cell adhesion protei...	DSG2
Muscle contraction	MYH6
Striated Muscle Contraction	MYH6

Figure 4.12. Main ORVAL insights for the patient involved in mild hypertrophic cardiomyopathy from [261]. **A.** The predicted gene pathogenicity network. The gene pair DSG2-DSC2 has a higher pathogenicity score based on the variant combination predictions by VarCoPP for that pair, as depicted by the darker edge colour. **B.** The protein-protein interaction (PPI) provided at the oligogenic module page, containing the first-level interactors of the module proteins. DSC2 and DSG2 are directly interacting, while MYH6 does not share any first-level interactors with those proteins. **C.** The pathway mappings of the three genes involved in the oligogenic network

Moreover, ORVAL supported the oligogenic hypothesis for a patient with congenital long QT syndrome (LQTS), carrying variants in three LQTS-associated genes: KCNQ1, KCNH2 and KCNE1 [262]. These genes created, again, a trigenic oligogenic network in ORVAL. We could support with high confidence the author's suspicions for the pathogenicity of the gene pair KCNQ1 and KCNH2, as it obtained the highest median pathogenicity score. The specific combination KCNH2:p.K897T and KCNE1:p.G38SK by itself is neutral, supporting the modifier effect these two genetic variants have on the phenotype (Figure 4.13).



Pathway	Genes ↑↓
Muscle contraction	KCNQ1,KCNE1,KCNH2
Cardiac conduction	KCNQ1,KCNE1,KCNH2
Phase 3 - rapid repolarisation	KCNQ1,KCNE1,KCNH2
Neuronal System	KCNQ1,KCNH2
Potassium Channels	KCNQ1,KCNH2
Voltage gated Potassium channels	KCNQ1,KCNH2
Phase 2 - plateau phase	KCNQ1,KCNE1
Unknown	

Figure 4.13. Main ORVAL insights for the patient involved in congenital Long-QT (LQTS) syndrome from [262]. **A.** The predicted gene pathogenicity network. The gene pair KCNQ1 - KCNH2 has a higher pathogenicity score based on the variant combination predictions by VarCoPP for that pair, as depicted by the red edge colour. **B.** The protein-protein interaction (PPI) provided at the oligogenic module page, containing the first-level interactors of the module proteins. KCNE1 connects with an indirect interaction the genes KCNH2 and KCNQ1. **C.** The pathway mappings of the three genes involved in the oligogenic network. All of them are regulating the muscle, and more specifically cardiac, contraction, and the flow of potassium in the cellular membrane channels.

These two examples demonstrate the potential of the ORVAL platform in predicting pathogenic variant combinations while providing at the same time meaningful interpretations for their results through the contextualisation of the predicted pathogenic gene networks. This equips geneticists and researchers with the necessary tools for assessing these predictions and formulate hypotheses on the potential causal mechanisms underlying the disease physiopathology.

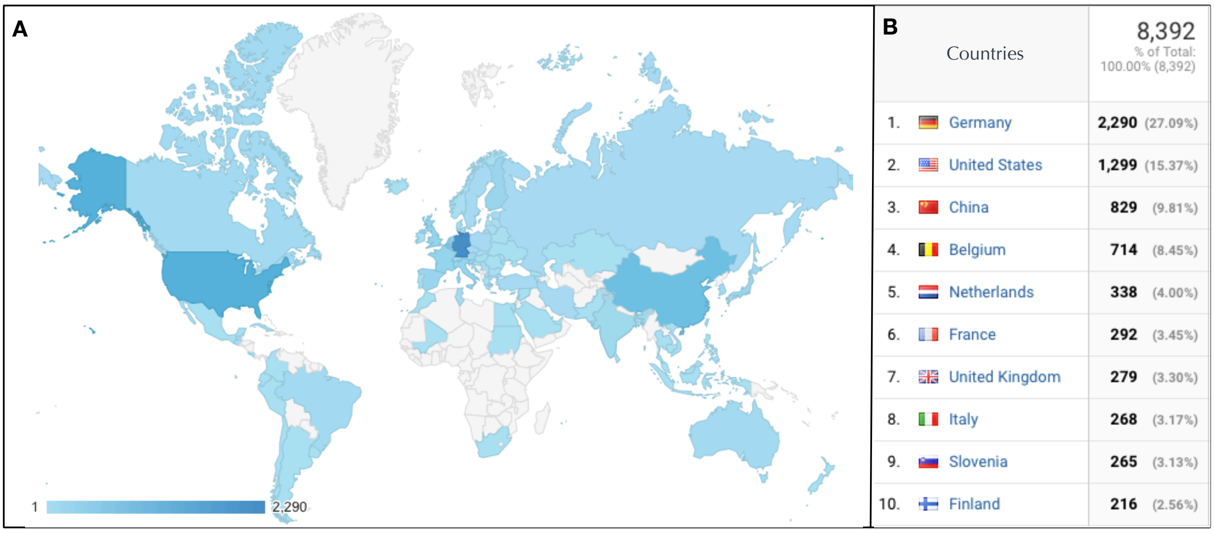
4.5.2 Impact on genetic research

The ORVAL platform has been used across the world since it has been made publicly made available in January 2019. Its publication in May 2019 [165] in Nucleic Acids Research and the talks given at the joint International Conference on Intelligent Systems for Molecular

Biology (ISMB) / European Conference on Computational Biology (ECCB) in July 2019 as well as at the 20th annual Belgian Society for Human Genetics in 2020, helped ORVAL to receive more national and international recognition.

Global usage of the platform

The communication around ORVAL led researchers across the world to analyse their own data with ORVAL. We report in Figure 4.14 the global usage of ORVAL since its inception and list the top 10 countries with the most users.



	Countries	8,392 % of Total: 100.00% (8,392)
1.	Germany	2,290 (27.09%)
2.	United States	1,299 (15.37%)
3.	China	829 (9.81%)
4.	Belgium	714 (8.45%)
5.	Netherlands	338 (4.00%)
6.	France	292 (3.45%)
7.	United Kingdom	279 (3.30%)
8.	Italy	268 (3.17%)
9.	Slovenia	265 (3.13%)
10.	Finland	216 (2.56%)

Figure 4.14. Usage of the ORVAL platform across the world since its inception (period of January 2019 - May 2023). (A) Worldmap highlighting countries using ORVAL. (B) Top 10 ranking of countries using ORVAL according to the number of unique users.

Discoveries and validation in clinical studies

Moreover, the ORVAL publication has been cited 32 times between May 2019 and May 2023 (only considering journal publications). These citations comprise studies where ORVAL was used to analyse genomics data, to discover potential pathogenic combinations or to validate them. Table 4.2 provides an overview of the 13 clinical studies using ORVAL, detailing the specific use and investigated disease. Additionally, 19 papers cite ORVAL as methodological comparison or as perspective for future studies (Table A.1).

Amongst the clinical studies using ORVAL (Table 4.2), we can note that four of them reported results from the gene pathogenicity network analysis [267, 269, 273, 274].

Ref.	Year	Usage	Disease	Details
[263]	2020	Validation	Cancer predisposition syndromes	1/6 patients with 1 predicted variant combination
[264]	2021	Validation	Partial gonadal dysgenesis	2 variant combinations correctly predicted.
[265]	2021	Discovery	Autism Spectrum Disorder	Prioritisation of 6 variants and 6 genes to study
[266]	2021	Validation	Early-onset dilated cardiomyopathy	3 predicted combinations with 3 genes
[267]	2021	Discovery	Odontochondrodysplasia	High confidence variant combinations and gene pathogenicity network use.
[268]	2021	Discovery, Validation	Bardet-Biedl syndrome	High confidence variant combinations predicted
[269]	2023	Discovery	Syncope caused by myocardial bridge	Gene pathogenicity network revealing 12 gene pairs
[270]	2022	Discovery	Hypospadias	Predicted combinations could not be linked to the case's phenotype
[271]	2022	Validation	Bardet-Biedl syndrome	Oligogenic effect supported in 44% of families by ORVAL and DiGePred
[272]	2022	Validation, Discovery	Escobar syndrome	Novel variants analysed and three pathogenic variant pairs identified
[273]	2022	Discovery	Susceptibility to mycobacterial infection	5/9 patients with pathogenic gene networks (Figure 4.15)
[274]	2023	Discovery	Differences of sex development	Pathogenic gene networks and interactions predicted for 3 patients
[275]	2023	Discovery	Congenital hydrocephalus	one True Digenic combination identified in one patient

Table 4.2. Research papers citing the ORVAL platform. All scientific publications citing ORVAL up to April 2023 are listed. The context in which ORVAL is cited is provided as: Discovery (using ORVAL to discover variant combinations to investigate) ; Validation (using ORVAL to validate known pathogenic variant combinations). Additionally, 19 papers cite ORVAL as methodological comparison or as perspective for future studies (Table A.1)

In Figure 4.15, the gene pathogenicity networks reported by Varzari et al. [273] are depicted. This study analysed nine patients with mycobacterial infection susceptibility, using whole-exome sequencing to identify candidate genes and sequence variants for susceptibility to mycobacterial infection.

The analysis identified 12 heterozygous variants, two of which were novel, in eight known MSMD-causing genes and found pathogenic or likely-pathogenic variants in 15 new candidate genes. Seven out of nine patients investigated had identified variants that occurred concomitantly in two or more different candidate genes. The ORVAL platform was then used to explore their potential digenic or oligogenic impact in increased predisposition to mycobacterial disease. ORVAL was used for both exploration and validation

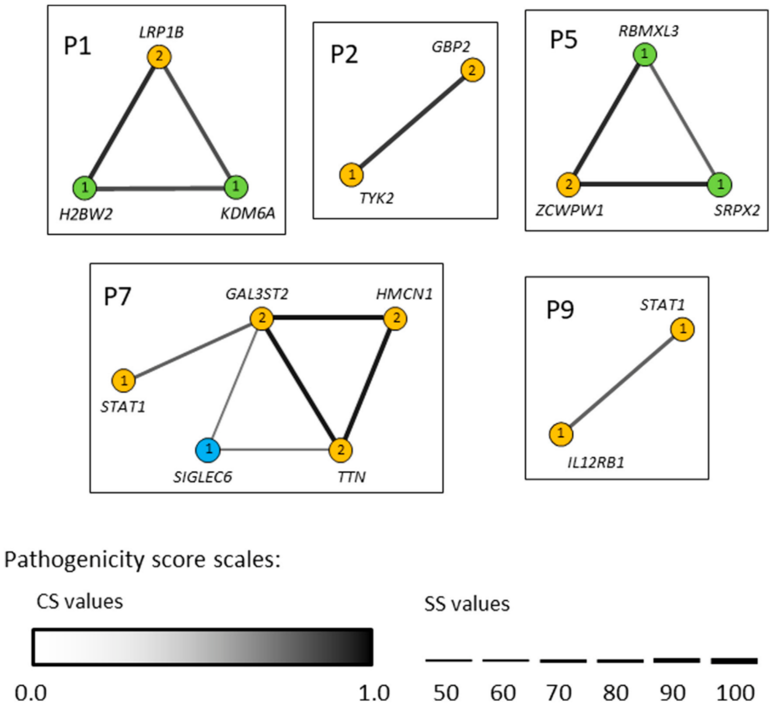


Figure 4.15. Gene networks showing the epistatic interactions between the genes in which rare variants have been identified as linked to the susceptibility to mycobacterial infection [273]. Nodes are individual genes and edges represent pair-wise interactions between genes/variants. Numbers inside the nodes denote number of variants. Yellow, blue and light-green nodes represent heterozygous, homozygous and hemizygous variants, respectively. Edge color intensity is proportional to the pathogenicity classification score (CS), whereas edge thickness is proportional to the pathogenicity support score (SS) for a combination of variants. In case two variants within one gene (i.e., compound heterozygotes), the highest pathogenicity scores computed for variant combinations are depicted. Only disease-causing variant combinations/networks (cutoff >0.532 for CS and ≥50% for SS) are illustrated. Credit Figure and Legend: Varzari et al. [273].

of known variant/gene combinations.

ORVAL's gene pathogenicity network conducted a comprehensive analysis of gene interactions, identifying potential pathogenic variant combinations in multiple patients. The authors discussed the GBP2-TYK2 combination (in P2) involving genes from the IL-12/IL-23/IFN-γ axis and the H2BW2-KDM6A combination (in P1) related to chromatin remodelling pathways. These findings align with their observations of reduced activity in the same interacting network, supporting the mechanism of "synergistic heterozygosity" [276], where combined mutations contribute to disease development. The results emphasise ORVAL's effectiveness in predicting complex genetic interactions and validating known variant/gene combinations.

Collaboration on a genetic study validated by ORVAL

We also collaborated with a team of geneticists to analyse patient's data in ORVAL, which resulted in a clinical study that we co-authored [264]. Three probands (P1,P2,P3), affected by familial 46,XY partial gonadal dysgenesis with incomplete penetrance, and their family members were analysed in this study. This condition is a rare genetic disorder that affects the development of the gonads in male individuals.

Exome sequencing was performed on all three probands to identify potential disease-causing variants. The NR5A1 c.991 - 1G>C splice-site variant was identified as the primary causative variant in cases P1 and P2, while P3 didn't carry this variant. The authors hypothesised digenic variant combinations in NR5A1-OTX2 in P1 and NR5A1-PROP1 in P2.

ORVAL was used for validation rather than discovery in this study. The authors used ORVAL to estimate the probability of a hypothesised combined pathogenic effect of the variant pairs.

The authors hypothesised that OTX2 and PROP1 may interact with NR5A1 to regulate gonadal development and differentiation and contribute to the pathogenesis of familial 46,XY partial gonadal dysgenesis with incomplete penetrance. However, further research is needed to support these findings and elucidate the underlying molecular mechanisms involved.

Concluding remarks

In conclusion, the ORVAL platform has made a significant impact on genetic research since its public release in January 2019. Its global adoption has enabled researchers to analyse their genomic data, leading to the discovery and validation of pathogenic variant combinations. An evaluation of ORVAL with independent case studies showcased its ability to support the oligogenic hypothesis, identifying trigenic oligogenic networks, and supporting the pathogenicity of specific gene pairs (see Subsection 4.5.1).

The platform's growing number of citations and usage in numerous clinical studies highlight its importance in the field, particularly its effectiveness in predicting complex genetic interactions and providing valuable insights into various diseases and conditions.

A notable example is the collaboration with geneticists in co-authored research [264]. The gene pathogenicity network analysis has been a valuable tool for researchers in understanding gene interactions in multiple patients. This success, combined with the independent evaluation results, demonstrates the platform's significant contribution to our growing understanding of complex genetic interactions and oligogenicity.

4.6 Improvements and evolution of the platform

Following the completion of my personal contributions to the ORVAL platform up until December 2021, presented in the previous sections, the responsibility for its maintenance and development has been handed over to Emma Verkinderen, a bioinformatics engineer. After a comprehensive handover process, she contributed in the evolution of the annotation database and the automatisation of its data integration pipeline as well as the update of predictive models. Additionally, she was entrusted with the migration of ORVAL to a cloud infrastructure, supported by the *101 Genomes Foundation*[10] – a foundation focused on providing solutions for analysing rare diseases.

The purpose of this section is to emphasise that ORVAL is a platform undergoing continuous enhancement and development, reflecting a long-term vision for its growth. This commitment is further supported by Elixir Belgium, which has included ORVAL as part of their node as a *mature service*[11] in 2023. Additionally, the platform is now referred in the Bio.tools registry[12].

4.6.1 VarCoPP 2.0: improved pathogenicity prediction

One notable improvement in the platform is the integration of the new Variant Combinations Pathogenicity Predictor, *VarCoPP 2.0* [277]. This updated version builds upon the success of its predecessor, introducing significant enhancements in both predictive performance and computational efficiency.

[10] 101 Genomes Foundation: https://www.f101g.org/en/

[11] Elixir Belgium - ORVAL node service: https://www.elixir-belgium.org/services/orval

[12] Bio.tools registry - ORVAL: https://bio.tools/Oligogenic_resource_for_variant_analysis

VarCoPP 2.0 incorporates several key improvements. The first of these is the use of a Balanced Random Forest algorithm, which greatly simplifies the model while still effectively addressing the class imbalance issue, therefore reducing the computation time and memory requirements and facilitating the analysis of larger data sets.

Another noteworthy enhancement in VarCoPP 2.0 is the refinement of its training set, which now consists of higher-quality instances from the Oligogenic Diseases Database (OLIDA) [71]. This modification ensures that the model's predictions are based on more confidently linked evidence of pathogenicity, ultimately improving its performance.

Additionally, VarCoPP 2.0 benefits from the inclusion of an updated and diverse set of features. A more careful selection process, involving an original wrapper method, was employed to identify the most relevant features for the model. One such feature was derived from the knowledge graph called BOCK (Biological networks and Oligogenic Combinations as a Knowledge graph) presented in Chapter 5, which effectively captures the relationships between genes and their potential pathogenic interactions.

The distance between two genes in BOCK (KG Distance) is computed using the Dijkstra algorithm and then normalised to account for the heterogeneity of the graph. Notably, gene pair features in VarCoPP 2.0, including KG Distance, emerge as the second most important predictive feature group. This demonstrates that VarCoPP 2.0 places more emphasis on the biological relatedness between genes, contributing to its enhanced performance.

As a result of these enhancements, VarCoPP 2.0 exhibits a significant improvement in both sensitivity (95% in cross-validation and 98% during testing) and specificity (with a 5% False Positive rate). This improved performance, coupled with its faster execution time, enables more accurate analysis of larger data sets associated with oligogenic diseases. With the simplification of the model, the support score (SS) of the original VarCoPP, reflecting the number of trees supporting a prediction, was removed. Additionally, with the improved accuracy, the confidence scores have been adapted into two confidence zones: a 99% and a 99.9% confidence-zone. To reflect these changes, the visualisation of the variant pathogenicity results has been adapted from an S-plot to a beeswarm plot (Figure 4.16).

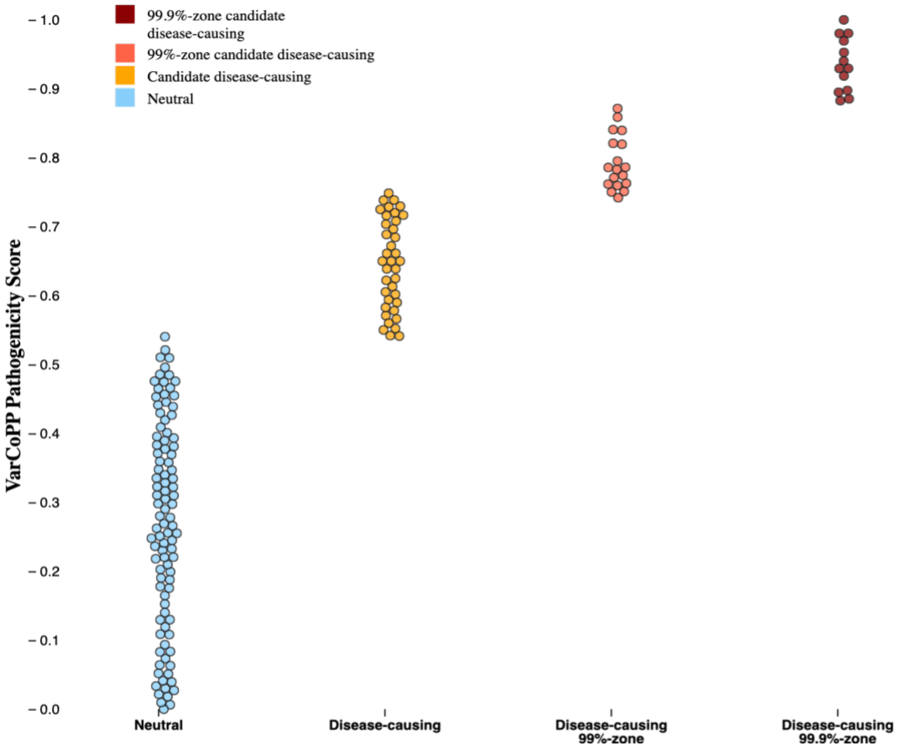


Figure 4.16. Example showcasing the new graphical representation, as a beeswarm plot, of VarCoPP 2.0 results in the ORVAL platform. Each variant combination is represented as a dot, coloured with the class and, for the pathogenic class, according to the two newly defined confidence zones.

4.6.2 Automatising and extending annotation data integration

Significant improvements have been made to the existing annotation data integration pipeline, which consolidates heterogeneous data from various public bioinformatics databases and pre-computed results (Section 4.3). The enhancements focus on automating and streamlining the integration process using Snakemake[13] [278] and Docker[14] technologies, which are crucial for promoting efficiency, reproducibility, and adaptability. *Snakemake*, as a workflow management system, orchestrates the pipeline, allowing for a more streamlined and interconnected workflow, handling dependencies between scripts and managing error recovery. This is essential for maintaining a robust and efficient pipeline compared to using a set of independent scripts. *Docker* ensures a consistent software environment across different platforms, which is vital for easy cloud deployment and replication by

[13] Snakemake workflow management system: https://snakemake.readthedocs.io

[14] Docker containerisation technology: https://www.docker.com

researchers for their specific needs.

The Snakemake pipeline is designed to fetch, parse, and consolidate the sources for the annotation database, accommodating different genome builds and generating TSV files compatible with the database updated schema. Snakemake manages its workflow with rules, encapsulating each of the data integration tasks and their dependencies (see Figure A.1). The inclusion of a Dockerfile simplifies the setup and deployment of the required software environment, facilitating the transfer of the ORVAL platform into a cloud infrastructure. This is crucial as the large annotation database requires regular updates from heterogeneous sources, and providing instructions and code to build instances of the annotation database on the cloud allows for easy transfer to third-parties without the need to run the pipeline on premises.

The new automated annotation data integration pipeline builds upon the previous multi-stage design, while also adding support for the *grch38/hg38* genome assembly. This addition is important for staying up-to-date with the latest genomic data and accommodating diverse research needs. The pipeline still integrates the aforementioned resources using the same scripts initially developed (see Section 4.3). However, updates have been made to re-calculate the biological distance using up-to-date STRING db [258] and to incorporate new data types, such as the BOCK knowledge graph (Chapter 5) distance, the pre-calculated Inheritance Specific Pathogenicity Predictor (ISPP) [181] gene-level feature, and the Biological Process similarity calculated with the SimGIC method [140], which are new features incorporated into the new VarCoPP 2.0 model.

In conclusion, the successful transfer of the ORVAL platform ensures its long-term development and maintenance, extending beyond the personal contributions presented in this thesis. As new models become available in the future, the platform's modular architecture will allow for easy adaptation. The annotation database, a vital yet challenging component of ORVAL, is now more manageable thanks to the full automation of the data integration pipeline. The use of Snakemake and Docker improved the pipeline to ensure the integrity and reliability of the integrated data while promoting ease of use, reproducibility, and adaptability. This improvement not only simplifies future updates and feature additions but also facilitates the migration of ORVAL to a cloud infrastructure.

4.7 Conclusion

In conclusion, ORVAL is a powerful web-based platform designed to assist geneticists and clinicians in identifying and understanding pathogenic variant combinations in the context of oligogenic diseases. It leverages predictive models and heterogenous biological data to provide a comprehensive and user-friendly platform for the analysis of genomic variants.

The technical development of ORVAL has focused on creating an efficient and versatile platform that can handle the complexities of genomic data analysis and the large number of potential combinations to process. The annotation database integrated into ORVAL is a crucial aspect of its functionality, providing a wealth of relevant biological information that can be used for predictive modelling and interpretation.

ORVAL offers a streamlined workflow for the submission, filtering, and processing of genomic data, handled with a dedicated job submission system. The platform accepts manual input as well as VCF files and also provides a range of filtering and sorting options, enabling users to focus on relevant results based on their specific research interests.

The platform's analysis capabilities span multiple levels of interpretation, from a ranking of pathogenic variant combinations to the exploration of gene pathogenicity networks. The platform's integration of annotation data and biological networks further enhances its predictive capabilities, enabling a more comprehensive understanding of the complex relationships between genes and their potential roles in disease development.

ORVAL is a platform in continuous development, with recent improvements focusing on enhancing its predictive performance, computational efficiency, and user experience. The incorporation of VarCoPP 2.0, the updated annotation database, and the streamlined data integration pipeline all contribute to the platform's ongoing growth and success.

Finally, it is essential to emphasise the impact of ORVAL on genetic research. Geneticists worldwide have adopted the platform to explore and validate pathogenic variant combinations, leading to a better understanding of various diseases. Its growing number of citations and usage in numerous clinical studies highlights its importance in the field and demonstrates its significant contribution to our understanding of complex genetic interactions and oligogenicity.

Chapter 5

BOCK: a knowledge graph to contextualise oligogenic disease information

In recent years, knowledge graphs (KGs) have emerged as a robust means to represent and integrate heterogeneous data. A KG is a semantic network that connects diverse entities (nodes) through various relationships (edges), enriched with semantic information and properties. This integration facilitates enhanced decision-making, information retrieval, and knowledge discovery, proving invaluable in applications ranging from search engines and recommendation systems to broader domains [124, 131, 237].

KGs and heterogenous networks have demonstrated several advantages for biomedical applications, such as facilitating a deeper understanding of the intricate mechanisms underlying various diseases, uncovering previously unknown relationships and interactions, potentially leading to novel insights into disease mechanisms and therapeutic targets, and enhancing collaboration between researchers and medical professionals [132, 142, 238, 251].

Given their potential in capturing complex relationships, KGs present a promising approach for the study of oligogenic diseases, where the involvement of multiple genes and the complex interplay between them cannot be easily captured by traditional predictive approaches.

BOCK (Biological networks and Oligogenic Combinations as a Knowledge graph), a novel KG developed as part of this PhD thesis, integrates multiple public biological network and ontology resources together with curated oligogenic disease information from published clinical cases. By incorporating oligogenic disease information within a broader biological context and leveraging the inherent advantages of knowledge graphs, BOCK offers a comprehensive and semantically-rich representation that captures the complexity and diversity of relationships relevant to oligogenic diseases. This knowledge graph opens the way for new applications ranging from advanced querying and question-answering systems to sophisticated predictive modelling.

In this chapter, we present the development process of BOCK and its applications in oligogenic disease research. We start with an overview of BOCK's structure and its various formats (Section 5.1), followed by the selection and integration of source databases (Section 5.2). Next, we showcase the flexible querying and information retrieval capabilities of the knowledge graph model in Section 5.3. We then discuss the selection of relevant oligogenic gene pairs and identification of neutral gene pairs as negative controls (Section 5.4). Then, we provide a comprehensive network analysis of genes and gene pairs involving in oligogenic diseases using commonly used topology metrics, highlighting potential predictive features and biases in the data (Section 5.5). Concluding the chapter, we present some preliminary results using recent KG embedding techniques for accurately identifying potential pathogenic gene pairs (Section 5.6).

The BOCK knowledge graph has been published in *BMC Bioinformatics* (forthcoming) [164] and is available as open data, in multiple exchange formats, on a Zenodo repository at: `https://doi.org/10.5281/zenodo.7185679`.

5.1 BOCK: structure and representations

This section presents the architecture of BOCK (Biological networks and Oligogenic Combination as a Knowledge graph) and outlines how it is made available for use. We first inspect the structure, detailing how biological entities and their relationships are encapsulated by nodes and edges, and present statistics illustrating the scale of the graph. Subsequently, we discuss the different forms in which BOCK is provided, highlighting its flexible adaptability to diverse research environments and technical needs.

5.1.1 Structure and composition of the knowledge graph

Using the oligogenic information from the clinical literature gathered in the Oligogenic Disease Database (OLIDA) [71] (see Material and Methods - Subsection 3.6.1) and multiple public biological network resources and biomedical ontologies, we constructed BOCK, a KG that puts oligogenic combinations into a biological context.

In KGs, nodes and edges can be qualified with types, facilitating the integration of heterogeneous concepts into a single graph without information loss. In BOCK, nodes represent biological entities and biomedical concepts defined by a specific node type, a unique Uniform Resource Identifier (URI) linking the node to its source database entry, as well as optional node properties. Edges represent relationships between these entities, defined by a specific type and an optional confidence score, indicative of the quality or the strength of the relationship (see Material and Methods - Subsection 3.7.1).

BOCK is structured according to the schema presented in Figure 5.1, illustrating all possible node types (or metanodes) and edge types (or metaedges). The current version of BOCK as of April 2023 comprises 83,703 nodes of 10 different types (Figure 5.2), including genes, diseases, biological processes, molecular functions, and cellular components, among others, representing the various levels of biological organisation. The knowledge graph consists of 2,659,346 edges of 17 distinct types, capturing diverse relationships and associations between these entities (Table 5.1).

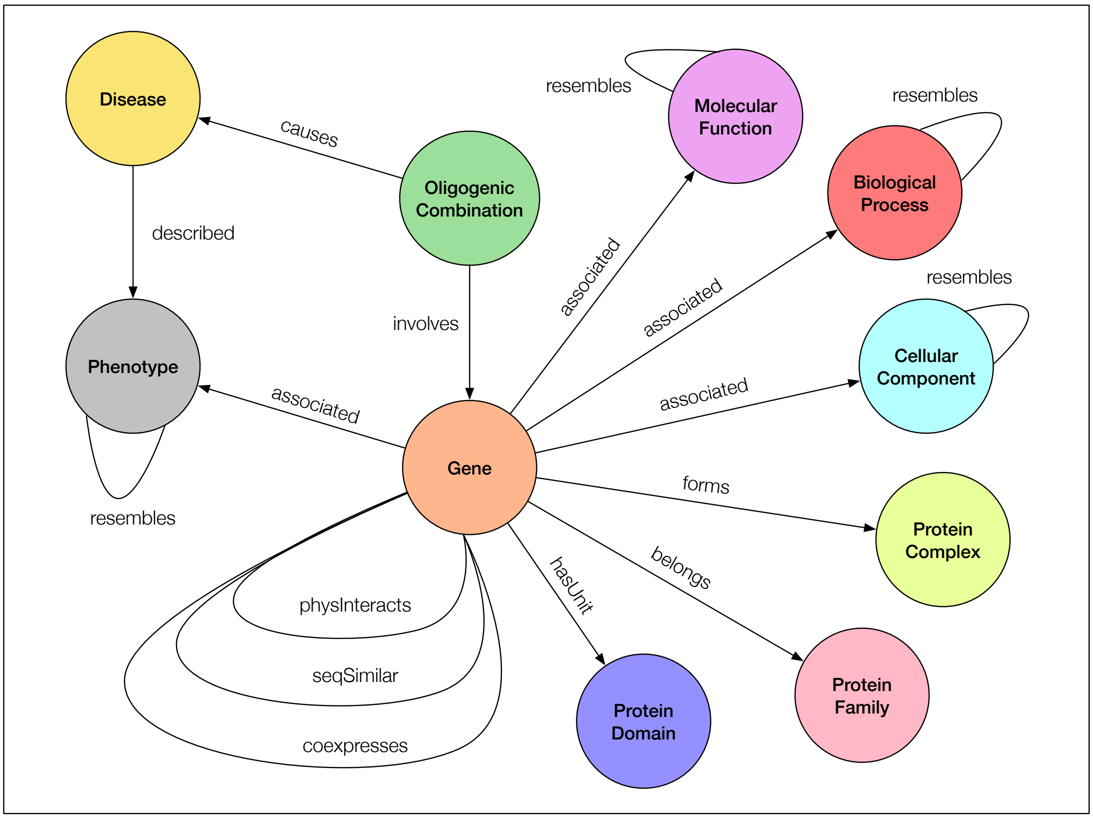


Figure 5.1. Knowledge graph schema of BOCK, representing the different node types (*i.e.* metanodes) and their relationships (*i.e.* metaedges). Undirectional relationships are symbolised by lines (—), denoting connections without a specific implied direction. Directional relationships, represented by arrows (→), indicate associations where the source entity has an influence or describes the target entity.

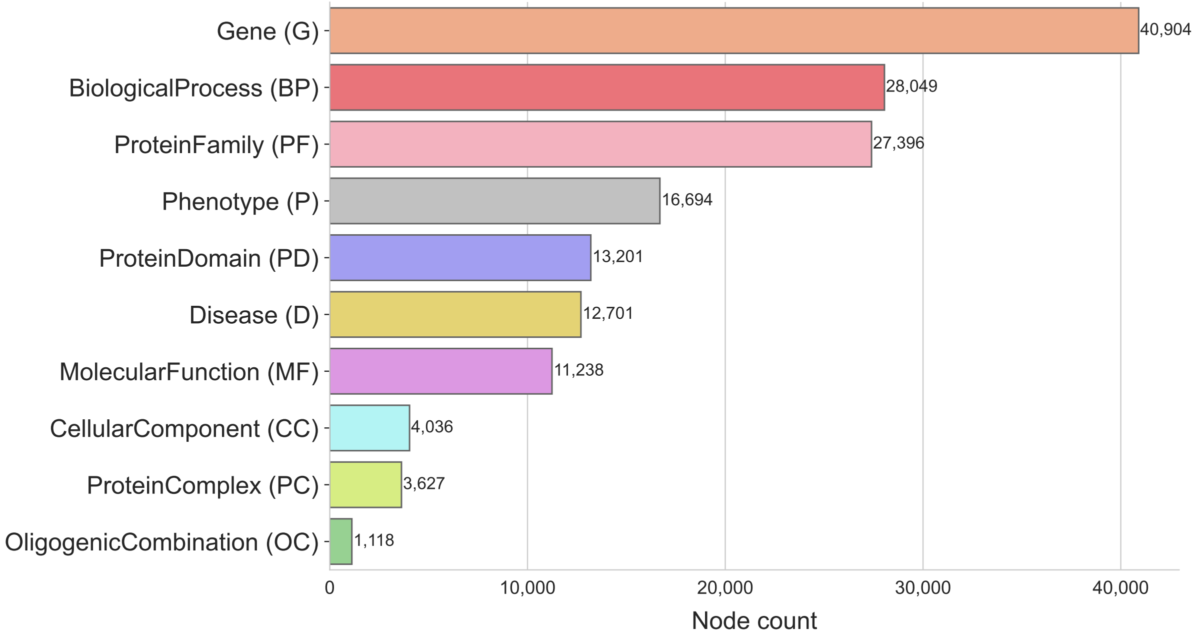


Figure 5.2. Number of nodes in the KG per metanode. We define an abbreviation for each metanode, in parenthesis, to simplify all metapath and rule notations in the following sections (see also Table 5.1).

Metaedge	Abbreviation	# Edges	# Sources	# Targets
Gene–coexpresses–Gene	GeG	1,338,764	14,940	14,940
Gene–physinteracts–Gene	GpG	329,801	17,062	17,062
Disease–described→Phenotype	DdP	233,175	12,676	10,423
Gene–associated→Phenotype	GaP	209,416	4,870	9,151
Gene–seqsimilar–Gene	GsG	186,445	12,226	12,226
Gene–associated→BiologicalProcess	GaBP	93,676	16,323	10,570
Gene–associated→CellularComponent	GaCC	58,432	16,978	691
Gene–belongs→ProteinFamily	GbPF	45,454	19,657	11,187
Gene–associated→MolecularFunction	GaMF	43,331	14,540	4,042
Gene–hasunit→ProteinDomain	GuPD	41,314	15,828	6,636
BiologicalProcess–resembles–BiologicalProcess	BPrBP	33,102	10,811	10,811
Phenotype–resembles–Phenotype	PrP	16,000	7,681	7,681
Gene–forms→ProteinComplex	GfPC	14,531	4,357	3,604
MolecularFunction–resembles–MolecularFunction	MFrMF	11,239	3,710	3,710
OligogenicCombination–involves→Gene	OCiG	2,700	1,118	907
OligogenicCombination–causes→Disease	OCcD	1,173	1,118	175
CellularComponent–resembles–CellularComponent	CCrCC	793	483	483

Table 5.1. Knowledge graph edge types. Each type of edge (*i.e.* metaedge) in the KG is defined uniquely by its source and target node types with the relationship name in between. Directed metaedges are indicated by an arrow on the relationship. We define abbreviations for each metaedge to simplify further notations. The table presents statistics on the number of corresponding edges, source nodes and target nodes for each metaedge, ordered by decreasing number of edges.

5.1.2 Representations and exchange formats

We provide the complete BOCK knowledge graph in various formats that facilitate its exchange and use, ensuring its accessibility, compatibility, and ease of use. The BOCK knowledge graph is available at `https://doi.org/10.5281/zenodo.7185679`. Details about these standards and technologies are provided in Material and Methods, Section 3.7.2.

The primary format for the BOCK knowledge graph is RDF, along with its data model as an OWL file, ensuring that BOCK is compliant with Semantic Web standards and can be easily integrated with other semantic resources. The RDF version of BOCK links entities to a given identifier registry via a Uniform Resource Identifiers (URI), using the Identifiers.org resolution service whenever possible. Additionally, some properties, such as PubMed IDs (PMID), Digital Object Identifiers (DOI), and UniProt accessions, are also semantically linked (see Material and Methods - Subsection 3.7.2).

For users who prefer to work with standard graph libraries and tools, we also provide BOCK in the Graph Markup Language (GraphML) format. This format enables the exchange and processing of graph data across a wide range of software tools and libraries, ensuring that BOCK is accessible to researchers with various technical needs.

Additionally, we have implemented an instance of the BOCK knowledge graph using the Neo4j graph database system. Its primary focus here is to provide an accessible and efficient platform for querying and retrieving information from the BOCK knowledge graph. Neo4j's user-friendly interface facilitates data integration, visualisation, and querying, while its Cypher query language allows for expressive and flexible querying of the graph, enabling users to extract complex patterns and relationships specific to their research interests.

By providing the BOCK knowledge graph in these different formats, we ensure that researchers can easily access, manipulate, and integrate the data with various tools and platforms, facilitating a wide range of applications for this resource.

5.2 Systematic integration of biological resources

The construction of a comprehensive knowledge graph for oligogenic disease research depends on the effective integration of diverse biological resources. Data integration in knowledge graphs involves the consolidation of information from multiple sources, transforming and harmonising them into a unified, structured representation. Integrating data from various biological databases is a complex task that demands meticulous selection of high-quality resources, maintaining compatibility across data sources, while preserving the integrity of the original information (see Material and Methods - Subsection 3.7.2).

In this section, we discuss the approach taken for the systematic integration of biological resources in the BOCK knowledge graph. We detail the criteria for choosing databases, the mappings and harmonisation of identifiers, the data filtering and transformation operated, and the innovative methods for inferring functional ontology-based relationships.

5.2.1 Selection of high-quality biological databases

Sources for creating the KG were selected based on domain knowledge and according to multiple criteria:

1. **Relevance in human disease aetiology**: by considering multiple biological levels of organisation often affected in pathologies and by selecting strictly for human-derived data;
2. **Quality control**: by favouring resources based on clear curation policies and substantial accuracy in the case of electronically inferred annotations;
3. **Gene coverage**: by only considering resources linking at least 20% of all human genes;
4. **Accessibility and interoperability**: by choosing resources from public and free-to-use databases, attributing each entity with a unique and retrievable identifier.

Compared to more generic KGs, we selected specifically networks relevant to understanding the molecular mechanisms of epistasis, placing genes as the central entities, and focusing on trusted resources describing a large set of human genes and their interactions. Table 5.2 describes all the selected source databases along with their versions.

We covered different types of bioinformatics resources: gene and protein interaction networks (Mentha [192], STRING db [258] and TCSBN [195]), functional annotations (InterPro [196] and CORUM [187]), and biomedical ontologies along with their annotations (Gene Ontology [189] and Human Phenotype Ontology [129]). In addition, we included gene features (from dbNSFP [185]) and database unique identifiers for mapping these resources (HGNC [177], Ensembl [174] and UniProt [178] (more details about these resources can be found in Material and Methods - Subsection 3.2.2, 3.3.2, 3.4.2)

Each database contributes to a certain *KG component* (or layer), representing a biologically meaningful subset of metanodes and metaedges in BOCK (see Table 5.4). Note that some of the databases used for integration, such as Ensembl or UniProt, are used across multiple components to handle the mappings of identifiers across resources, as discussed next.

Source database	Version	Extracted information	KG components
OLIDA [71]	04-2022	Oligogenic gene combinations, DOI, timestamp, ethnicity, curation scores	OLIDA
Mentha [192]	10-10-2022	Manually curated protein-protein interactions in human	PPI
STRING db [258]	v11.5	Human protein sequence raw blastp scores	SEQSIM
TCSBN [195]	26-10-2020	Tissue-specific co-expression network from normal human tissues (derived from GTEx)	COEXP
GTEx [196]	V8	Median gene-level RNA-seq transcript per millions (TPM) by tissue ; Tissue names and sample sizes	COEXP
InterPro [196]	87.0	Human protein domains and families	DOMAIN, FAMILY
CORUM [187]	4.0	Human protein complexes	COMPLEX
Gene Ontology [189]	10-07-2022	Gene ontology and annotations on human genes (positive associations with qualifiers “enables”, “involved in”, “is active in” and “located in”), excluding IEA and ND evidence codes	PROCESS, FUNCTION, CELLCMP
Human Phenotype Ontology [129]	06-2022	Human ’phenotypic abnormality’ sub-ontology (non-obsolete), gene annotations and disease associations	PHENO, DISEASE
dbNSFP [185]	v4.3	RVIS [180] and GDI [179] gene scores	All with Gene
HGNC [177]	2022-10-12	Official gene names and Ensembl mappings	All with Gene
Ensembl [174]	Release 107	Ensembl identifiers, gene name and entrez id mappings	All with Gene
UniProt [178]	2022_04	UniProt mappings to Ensembl identifiers	PPI, SEQSIM, COMPLEX, DOMAIN, FAMILY

Table 5.2. Source databases to construct BOCK Bioinformatics databases used with their version used for this study. The specific extracted information from each database is indicated as well as the corresponding KG component it is used for (see Table 5.4).

5.2.2 Identifier harmonisation and gene-level collapsing

Considering that clinical studies on oligogenic cases rarely report the effect of variants on specific encoded proteins, we chose to reduce the model complexity of BOCK by collapsing all protein identifiers at the gene level into entities of type “Gene”. Gene and protein identifiers from all integrated resources were collapsed and mapped into their corresponding Ensembl identifiers. The databases Ensembl [174], UniProt [178] and HGNC [177] were used as a reference to handle potential identifier mapping ambiguities.

Edges linking protein pairs were also collapsed as edges between their associated genes, with an associated score computed as the maximum of all original scores.

Two properties were also added to the "Gene" entity: the Residual Variation Intolerance Score (RVIS) [180] and the human Gene Damage Index (GDI) [179], obtained from dbNSFP [185] (see Material and Methods - Subsection 3.2.2).

The non-redundant contribution of each network source for genes is detailed in Table 5.3.

Network resource	# **Genes**	# **Exclusive genes**
Domains (InterPro)	20302	691
Functions (GO)	18344	46
PPI (Mentha)	17062	89
Coexpression (TSCBN)	14940	33
Seq. similarity (STRING)	12226	67
Phenotypes (HPO)	4870	32
Complexes (CORUM)	4357	0
Oligogenic combinations (OLIDA)	907	0

Table 5.3. BOCK source network gene contributions. We calculated *#Genes* as the total number of genes present in each source network. *#Exclusive genes* was obtained by considering genes that are exclusively present in each of the specific network resource.

5.2.3 Data processing and integration

In order to integrate the source databases information into BOCK, specific pre-processing steps were applied, including data mappings, transformation or calculation of edge scores, and edge filtering. These integration steps are summarised in Table 5.4 and detailed next.

KG component	Node types	Edge types	Edge score	Filtering
OLIDA	OligogenicCombination, Disease	causes	None	None
	OligogenicCombination, Gene	involves	None	None
DISEASE	Disease, Phenotype	described	Disease-Phenotype frequency	None
PPI	Gene	physInteracts	Original Mentha score (interaction confidence)	None
SEQSIM	Gene	seqSimilar	Blast Score Ratio (BSR)	align. coverage $\geq$ 50% ; BSR $\geq$ 0.2
COEXP	Gene	coexpresses	CI_{low}(cor)	tissue-sample-size > 70 ; $Z_{tpm}(G_1)$ & $Z_{tpm}(G_2) \geq$ -3 ; $CI_{low}(cor) \geq$ 0.80 ; p-val-adj < 0.01
DOMAIN	ProteinDomain, Gene	hasUnit	$FI(PD)$	Only "$Active\_site$", "$Binding\_site$", "$Conserved\_site$", "Domain", "PTM", "Repeat"
FAMILY	ProteinFamily, Gene	belongs	$FI(PF)$	Only "Family" and "$Homologous\_superfamily$"
COMPLEX	ProteinComplex, Gene	forms	$FI(PC)$	None
PROCESS	BiologicalProcess, Gene	associated	$FI(BP)$	None
	BiologicalProcess	resembles	$SimGIC(BP_1, BP_2)$	Sim. score $\geq$ 0.5
FUNCTION	MolecularFunction, Gene	associated	$FI(MF)$	None
	MolecularFunction	resembles	$SimGIC(MF_1, MF_2)$	Sim. score $\geq$ 0.5
CELLCMP	CellularComponent, Gene	associated	$FI(CC)$	None
	CellularComponent	resembles	$SimGIC(CC_1, CC_2)$	Sim. score $\geq$ 0.5
PHENO	Phenotype, Gene	associated	$FI(P)$	None
	Phenotype	resembles	$SimGIC(P_1, P_2)$	Sim. score $\geq$ 0.5

Table 5.4. KG components integration into BOCK Data integration of the BOCK knowledge graph by component. Each component describes a different biological level or view and has been integrated into the knowledge graph from a source network or ontology database (see Table 5.2) into KG node types and edge types. For most components, an edge score is computed and attributed as an edge property. For some components, the integration process required the application of a filtering stage to limit the noise and the size of the integrated network. More details about the computation of scores and filtering steps are provided in the Method section.

Oligogenic combinations

OLIDA aggregates curated information about oligogenic diseases gathered from the medical literature [71] (see Material and Methods - Subsection 3.6.1). Each entry consists of a genetic variant combination involving several genes linked with contextual information, such as the associated disease, the source scientific article, the suspected oligogenic effect and its curation confidence scores.

The BOCK KG encodes the relational information from OLIDA by linking the involved “Gene” and “Disease” entities via a dedicated “OligogenicCombination” node pointing to the OLIDA identifier of a given oligogenic variant combination. Additional properties have been added to this node, such as the OLIDA curation confidence scores, the publication DOI and timestamp, the ethnicity of the associated patient and the suspected oligogenic effect [81].

Protein interactions

The protein-protein interactions (PPI) were extracted from Mentha [192], a resource aggregating primary protein-protein interaction databases within the IMEx consortium. Mentha focuses on experimentally determined direct protein interactions and, therefore, does not contain any computationally inferred interactions. All interactions from the Mentha human interactome were integrated as edges of type “physInteract” in the KG, linking two entities of type “Gene” and weighted by the provided Mentha reliability score.

Protein sequence similarity

Protein sequence similarity links were derived from pre-calculated BLAST pairwise protein alignment bit scores [198] obtained through the STRING homology file [279]. To determine the presence of a sequence similarity link between two proteins, we used criteria recommended for homology detection in human [280].

First, we retained only proteins where the aligned region covers at least 50% of the shorter protein. Second, we computed the Blast Score Ratio (BSR) as a ratio of the BLAST bit-score to the smaller of the two self-alignment bit-scores [281]. This transformation bounds all alignment scores in the interval $[0, 1]$ and allowed us to set a universal minimum BSR threshold value of 0.2 to retain significant edges. This threshold translates to a bit-score of approximately 200 for the alignment of average length proteins (472 aa [282]). A minimum BSR cutoff value of 0.2 was determined following an analysis (see Figure 5.3) that demonstrated its capacity to capture functional similarity signal [283].

Retained edges, scored by the BSR, were incorporated into BOCK as “seqSimilar” type, linking “Gene” entities.

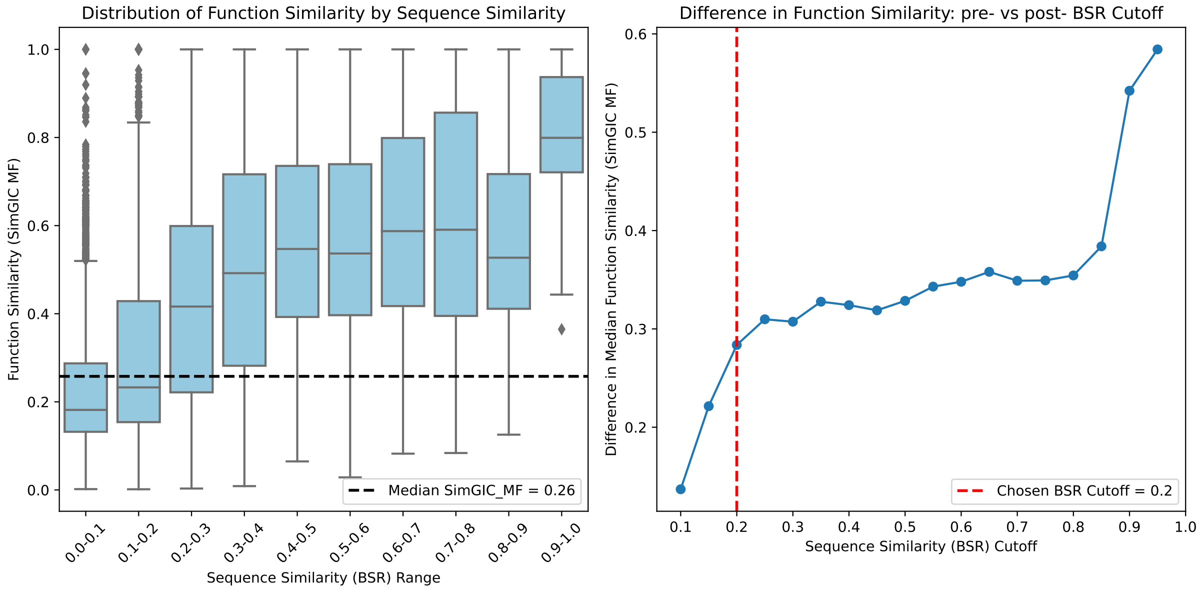


Figure 5.3. Empirical determination of sequence similarity cutoff value. To find a meaningful threshold of sequence similarity for integration into BOCK as “seqSimilar” edges, we studied the relationship between sequence similarity, computed by Blast Score Ratio (BSR) [281] – a standardised representation of sequence similarity – and the functional similarity – calculated with SimGIC [140] based on shared Molecular Function (MF) annotations from the Gene Ontology. **(A)** Functional Similarity (SimGIC MF) distribution across Sequence Similarity (BSR) ranges. The dashed line represents the overall median SimGIC MF score across all gene pairs. **(B)** Differential in median SimGIC MF for gene pairs segregated by each tentative BSR cutoff. The red dashed vertical line indicates the selected BSR cutoff of 0.2. This cutoff corresponds to the initial point of substantial elevation in functional similarity (SimGIC MF), effectively segregating low-similarity, potentially noise-introducing gene pairs, while capturing functional relationship signal.

Tissue co-expression

Tissue co-expression data was extracted from the TCSBN database [195] that provides pre-calculated co-expression correlation statistics generated from 46 human tissues using the tissue specific RNA-seq data. We focused exclusively on the data computed from normal tissues originally from the GTEx database (Genotype Tissue Expression) [196] and ignored cancer-related cell-lines data.

To enhance signal strength, we applied several filters to the original sources: (1) Tissues with fewer than 70 samples were excluded per GTEx recommendations, and redundant subtypes were consolidated. (2) Co-expression relationships involving a gene with a z-score below -3 in any given tissue, as per the standardised GTEx median tissue gene expression levels, were discarded [284]. (3) We retained edges with significant adjusted p-values (< 0.01) and strong correlations ($\rho \geq 0.80$). Correlations were conservatively estimated using the lower bound of the Fisher-transformed confidence interval, taking into account tissue sample size [285]. These were integrated into the KG as type “Gene”, linked by “coexpresses”, and scored by the maximum correlation value across tissues. We recorded the set of tissues associated to the highly correlated and significant co-expression value into the “in” edge property.

Protein domain and families

We extracted protein domain and family information, as well as their annotations on human proteins, from the InterPro database [286]. Genes sharing domains or families are often found to be functionally associated. Entries from InterPro were integrated as “ProteinDomain” and “ProteinFamily” node types, with properties such as the name and the member database, and linked to nodes of type “Gene” via edges of types “hasUnit” and “belongs” respectively.

Protein complexes

Protein complexes were extracted from the CORUM database [187], a resource of manually annotated protein complexes from mammalian organisms. We selected complexes found in human and integrated each complex as a “ProteinComplex” node linked, via an edge “forms”, to its sub-units corresponding “Gene” entities.

Phenotype and disease information

We extracted phenotypic information from the Human Phenotype Ontology (HPO) [129], a resource standardising terminologies for human phenotype abnormalities. We integrated all non-obsolete terms under the sub-ontology “Phenotypic abnormality” as “Phenotype” nodes. The provided phenotype-gene annotations were integrated by linking the associated “Gene” entities with an edge of type “associated”.

Diseases were integrated based on known associations with phenotypes coming from medical literature and reference disease databases such as OMIM and Orphanet [129, 205]. An edge of type "described" was created between "Disease" and "Phenotype" entities, scored with the frequency of the phenotype if available.

Gene ontology annotations

The Gene Ontology (GO) knowledge base provides a standardised, controlled vocabulary for gene functional information [189]. We integrated all three provided sub-ontologies as corresponding entities, namely "BiologicalProcess" (BP), "Molecular Function" (MF) and "Cellular Component" (CC) into the KG, excluding root terms and terms marked as obsolete. Relationships to human genes were retrieved from the provided gene ontology annotation (GOA) file. We filtered the retained associations to keep only positive relationships flagged with the qualifiers "enables", "involved_in", "is_active_in" and "located_in". In order to keep only evidence-based reviewed associations, we also discarded all annotations inferred by electronic annotation (evidence code IEA) and those where no supporting biological data is available (evidence code ND). The retained annotations were integrated by connecting "Gene" entities to the corresponding GO entity via the edge type "associated".

5.2.4 Information-based inference of ontology relationships

Gene functional annotation relationships

In the KG, many entities linked to "Gene" correspond to functional annotation terms, represented as nodes of type: "ProteinDomain", "ProteinFamily", "ProteinComplex", "Phenotype", "BiologicalProcess", "MolecularFunction" and "CellularComponent".

We assigned a score to the edges between one gene and an annotation term, estimating how informative these relationships are. This score was determined by looking at how frequently an annotation term occurs on human genes, with infrequent terms receiving higher scores and more common terms receiving lower scores.

More formally, we defined a metric, the functional information (FI), given an annotation term $t \in T$ of a specific entity type. This metric corresponds to the information content of the term t (see Material and Methods - Subsection 3.4.3), scaled by the maximum information content and is therefore bound in the $[0, 1]$ interval, facilitating its comparison across all functional edges (Equation (5.1)). The information content of a term t is derived from the ratio of the count of genes associated with that term, $g(t)$, and its descendants in the ontology t_s, to the count of genes associated with all terms of the same entity type, $g(T)$.

$$\mathrm{FI}(t) = \frac{-\log\left(|g(t \ \cup \ (\bigcup_{t_s \sqsubseteq t} t_s))| \ / \ |g(T)|\right)}{\log\left(|g(T)|\right)} \tag{5.1}$$

Semantic similarity relationships

The KG incorporates terms from two extensive ontologies, Gene Ontology (GO) and Human Phenotype Ontology (HPO). Both ontologies are structured as hierarchies of terms, interconnected by semantic associations in a directed acyclic graph. By extracting the 'is a' relationships, we obtained the term subclass hierarchies, which allowed us to compute semantic similarity links between each pair of terms.

The semantic similarity was estimated by SimcGIC [211], a pairwise adaptation of the SimGIC measure [140], which has been shown to effectively capture the expected relationship between functional and sequence similarity (see Material and Methods - Subsection 3.4.3). This measure considers both the hierarchical relationship between terms and the information content of their shared ancestors, providing a balance between shared and distinct semantics akin to the Jaccard index. This property ensures that all similarity scores are comprised between 0 and 1, enhancing their interpretability and comparability across all pairs of terms. We also found this method to be more sensitive to the path distance in the ontology hierarchy, compared to methods considering only the IC of the Lowest Common Ancestor (LCA), such as Resnik's measure, making it more suitable to represent a "shortcut" for navigating the underlying ontology hierarchy.

We integrated these semantic relationships with an edge of type "resembles" whenever the semantic similarity between two terms is higher than 0.5.

5.3 Knowledge graph querying and exploration

The BOCK knowledge graph provides researchers with a powerful tool to explore the complex relationships between genes, diseases, and a range of biological entities. One way to interact with BOCK is through the graph database Neo4j, which can be queried via the Cypher query language (see Material and Methods - Subsection 3.7.4). In this section, we present various examples that demonstrate the potential of BOCK in deriving biomedical insights, using the Cypher query language to explore the integrated data. The examples presented in this section are color-coded based on their representation in the BOCK knowledge graph schema (see Figure 5.1).

```
MATCH (oc:OligogenicCombination)-[i:involves]->(g:Gene),
      (oc)-[c:causes]->(d:Disease)
WHERE oc.FAMmanual = 3 OR (oc.STATmeta = 3 AND oc.STATmanual = 3)
AND oc.id IN ['OLI075', 'OLI199', 'OLI179', 'OLI081', 'OLI111']
RETURN oc,d,g,i,c
```

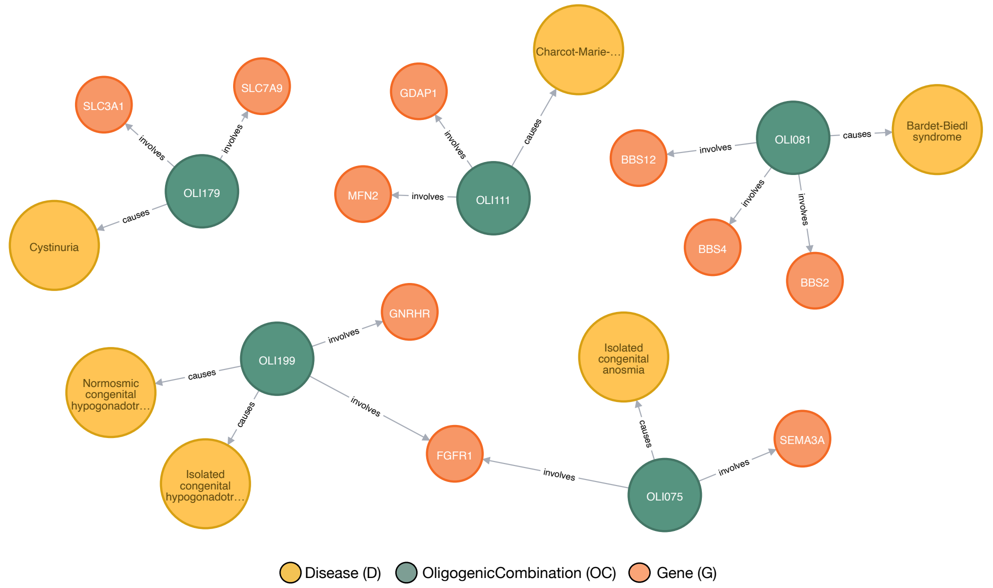


Figure 5.4. Cypher query demonstrating the retrieval of some OligogenicCombinations with strong evidence levels, their involved genes, and the diseases they cause. The query filters OligogenicCombinations based on a FAMmanual score of 3 or a combination of STATmeta and STATmanual scores both equal to 3, indicating strong evidence levels (see Table 5.5). The query returns the OligogenicCombinations (oc), the diseases they cause (d), the involved genes (g), and the corresponding relationships (i and c).

The first example (Figure 5.4) illustrates how the oligogenic cases are represented inside BOCK along with their involved genes and the diseases they cause. Based on the OLIDA curation scores [71], it is possible to select combinations with a strong evidence level (Table 5.5). The query reveals examples of digenic diseases, a trigenic disease, and even instances where distinct digenic diseases share a common causal gene.

The BOCK knowledge graph also facilitates the examination of connections between genes involved in oligogenic diseases. We show, in Figure 5.5, a sample of the relationships between genes *ADGRV1* and *PDZD7*, which are known to be involved in Usher syndrome type 2. This example highlights the complex interconnections between these genes, with paths of varying lengths and comprising different edges and nodes of diverse types, such as protein complexes, protein families, protein domains, cellular locations, and genes interconnected with coexpresses relationships and protein physical interactions.

```
MATCH (oc:OligogenicCombination{id:'OLI021'})-[i1:involves]->(g1:Gene),
      (oc)-[i2:involves]->(g2:Gene),
      (oc)-[c:causes]->(d:Disease)
OPTIONAL MATCH path = (g1)-[*1..3]-(g2)
WHERE NONE(x IN nodes(path) WHERE x:OligogenicCombination OR x:Disease OR
    ↪x:Phenotype) AND length(path) > 1
WITH oc, d, g1, g2, i1, i2, c, collect(path)[0..10] as paths
RETURN oc, d, g1, g2, i1, i2, c, paths
```

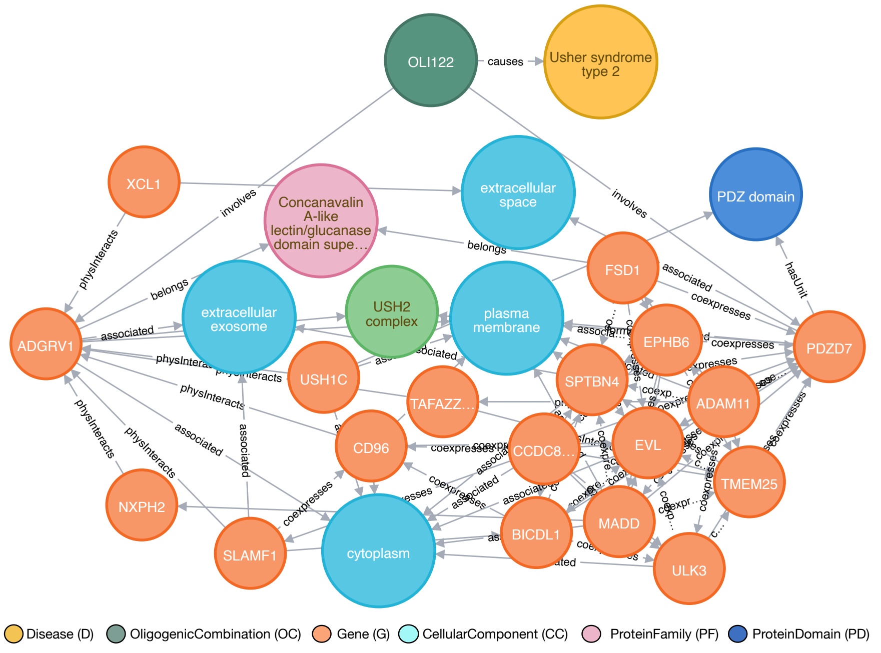


Figure 5.5. Illustration of a Cypher query for retrieving an OligogenicCombination with its involved genes, associated disease, and a subset of paths between the involved genes. The query retrieves OligogenicCombination *OLI021*, its involved genes (g1 and g2), *ADGRV1* and *PDZD7*, the disease it causes (d), *Usher syndrome type 2*, and a subset of 1 to 3-hop connections between the two involved genes, excluding paths through OligogenicCombination, Disease, and Phenotype nodes.

More specifically, it is possible to explore specific patterns between two genes. In

Figure 5.6, we explore a specific pattern of indirect Gene-Gene relationships between two genes ($CDK4$ and MYC) involved in MODY, an oligogenic disease. Users can explore the list of intermediate genes (here 6 genes) with a similar sequence with one gene and a protein-protein physical interaction with the other gene. Using the underlying edge scores can help prioritising which intermediate gene might be the most relevant.

```
MATCH (g1:Gene)-[:involves]-(oc:OligogenicCombination{id:'OLI513'})
      -[:involves]-(g2:Gene)
WITH g1,g2
MATCH p=(g1)-[:seqSimilar]-(:Gene)-[:physInteracts]-(g2)
RETURN p
```

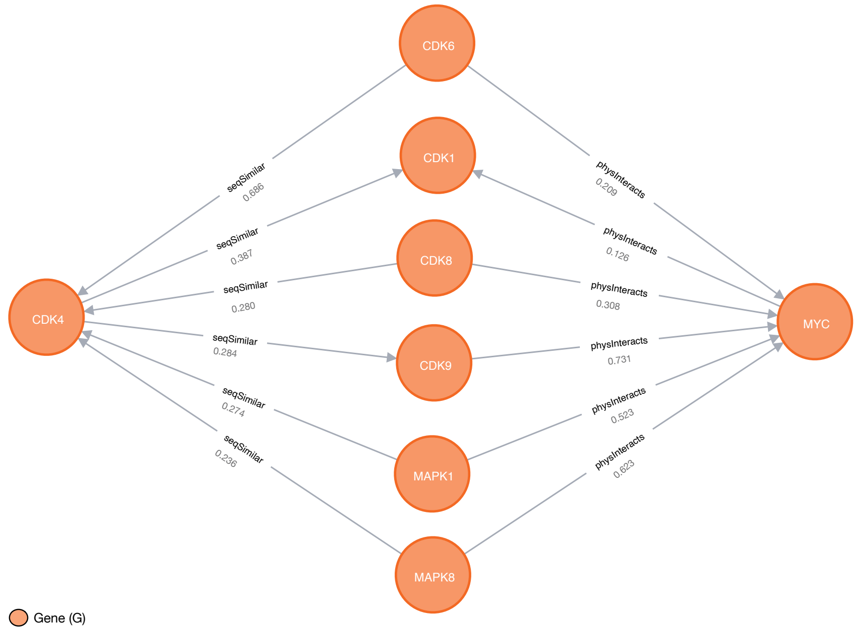


Figure 5.6. Illustration of a Cypher query for retrieving a specific pattern of indirect Gene-Gene relationships between two genes ($g1$, $g2$) involved in a specific oligogenic combination oc (OLI513: causing the MODY disease). The pattern involves an intermediate gene connected via a `seqSimilar` relationship to one gene ($g1$) and a `physInteracts` relationship to the other gene ($g2$). Relationship scores, detailed in Table 5.4 reflect the maximum protein-protein sequence similarity and physical interaction confidence score.

Users can also explore functional annotations of genes, such as their associations with protein families and domains. In the example presented in Figure 5.7, all genes returned by the query are linked to the Small GTPase (IPR001806) protein family and are associated with multiple protein domains. Functional annotations relationships are scored based on term frequency, with highly specific annotations receiving higher scores. For example, the "Sigma-54 interaction domain, ATP-binding site 1" (IPR025662) is linked to 2 genes and has a score of 0.901, while the more ubiquitous "Pleckstrin homology domain" (IPR001849) is linked to 275 genes and has a score of 0.199, as explained in Section 5.2.4.

```
MATCH (g:Gene)-[b:belongs]->(pf:ProteinFamily)
OPTIONAL MATCH (g)-[h:hasUnit]->(pd:ProteinDomain)
RETURN g, b, pf, h, pd
LIMIT 10
```

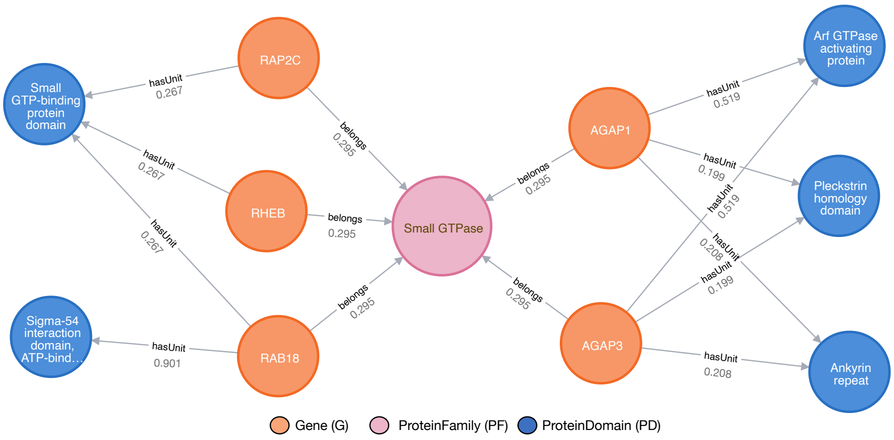


Figure 5.7. Cypher query retrieving some genes, their associated protein families and their associated protein domains. Genes are connected to ProteinFamily nodes through the `belongs` relationship, and the same genes (via the *g* variable) are also linked to ProteinDomain nodes through the `hasUnit` relationship. The edge scores represent the frequency of annotation of the functional annotation terms, with higher scores indicating more specific and informative annotations, as detailed in Section 5.2.4.

The BOCK knowledge graph enables users to investigate shared and related biological processes between genes. For example, the query presented in Figure 5.8 explore these relationships for the genes $SHROOM2$ and $MYO7A$, known to be associated with Atypical hemolytic uremic syndrome (OLIDA ID: OLI1032, PMID: 34391192). While the first part of the query retrieves a single common biological process, the second part, using the `resembles` links, uncovers additional biological processes that are exclusively associated to each gene but closely related to each other. Scores associated to the `resembles` links are computed based on the semantic similarity between terms (see Section 5.2.4).

```
MATCH p1=(g1:Gene{name:'SHROOM2'})-[:associated]
    ↪->(:BiologicalProcess)<-[:associated]-(g2:Gene{name:'MYO7A'})
OPTIONAL MATCH p2=(g1)-[:associated]->(bp1:BiologicalProcess)
    ↪-[:resembles]-(bp2:BiologicalProcess)<-[:associated]-(g2)
RETURN p1, p2
```

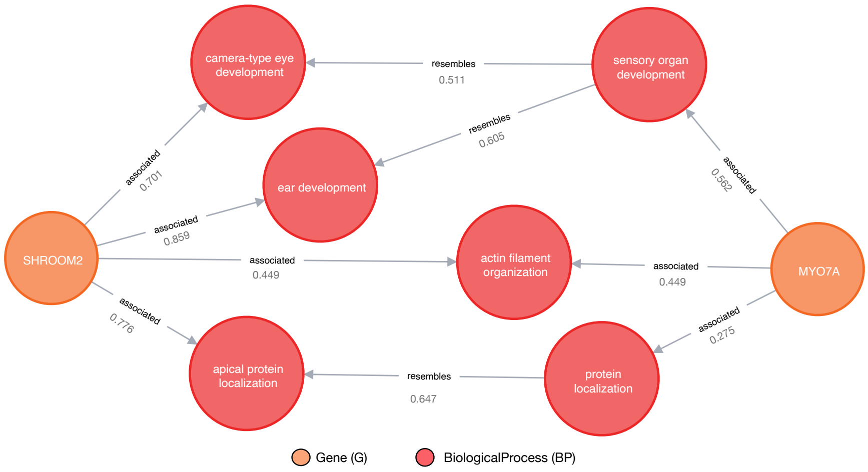


Figure 5.8. Cypher query illustrating the retrieval of common and related biological processes between two genes (g1 and g2) known to be associated with an oligogenic disease. The query returns the shared biological processes (p1) and related biological processes (p2) between the genes. The results emphasise the utility of the `resembles` relationship in uncovering related biological processes in the BOCK knowledge graph. Scores associated to the `resembles` links are computed based on the semantic similarity between terms, as detailed in Section 5.2.4.

Using the BOCK knowledge graph, it is also possible to explore genetic relationships between diseases with similar phenotypes. In Figure 5.9, we query two diseases using their Orphanet IDs: Marfan syndrome (orphanet:558) and Loeys-Dietz syndrome (orphanet:60030), both of which present similar phenotypes. We can retrieve phenotypes describing each disease and associated with a gene. By ranking the gene associations by the number of paths, we can see that the gene $FBN1$ has the most connections to related phenotypes, providing valuable insights into the genetic underpinnings of these two syndromes and their overlapping clinical manifestations.

These examples demonstrate how the BOCK knowledge graph's versatility and depth enable researchers to formulate queries to retrieve genetic relationships, disease mechanisms, and functional annotations. By employing the Cypher query language, users can efficiently access the vast amounts of integrated data to uncover meaningful biomedical insights. This practical approach complements the advanced graph analytics presented in Section 5.5 and sets the stage for machine-learning applications, as discussed in Chapter 6.

```
MATCH paths=(d1:Disease{id:'orphanet:558'})-[:described]
    ↪->(:Phenotype)<-[:associated]-(g:Gene)-[:associated]
    ↪->(:Phenotype)<-[:described]-(d2:Disease{id:'orphanet:60030'})
WITH g, COLLECT(paths) AS agg_paths, COUNT(paths) AS path_count
ORDER BY path_count DESC
RETURN g, agg_paths
LIMIT 1
```

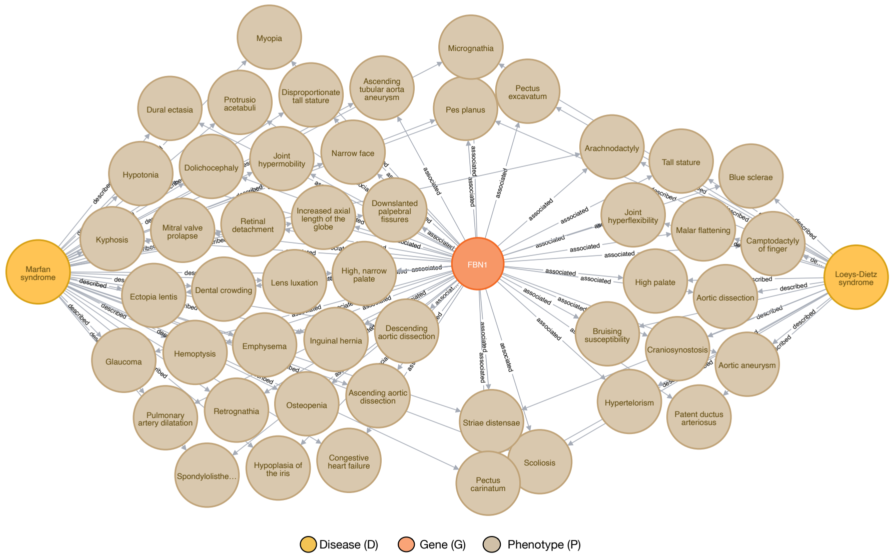


Figure 5.9. Cypher query retrieving genes connecting two diseases (d1 and d2) with similar phenotypes through shared or related phenotypes. The query returns the genes (g) and the paths (agg_paths) connecting the two diseases. The results highlight the potential of BOCK to identify genes that may be involved in multiple diseases and underline the importance of phenotype information in understanding the relationships between diseases.

5.4 Selection and weighting of gene combinations

In this section, we present the methodology for extracting gene pairs from BOCK. The primary focus is the identification of disease-causing gene pairs with evidence of oligogenicity, along with the selection of a set of neutral gene pairs for comparative purposes. These chosen gene pairs are essential for the upcoming in-depth analysis detailed in section Section 5.5. Furthermore, these sets are used for training and evaluation of our machine learning approach, which aims to detect novel pathogenic gene pairs, such as the KG embedding approach discussed in Section 5.6 and the interpretable predictive methodology introduced in Chapter 6.

5.4.1 Disease-causing gene pairs with reliable evidences

In this research, we leverage the extensive oligogenic combinations integrated into BOCK from the OLIDA database, a valuable resource compiling information about oligogenic diseases from medical literature [71] (refer to Section 5.2). Using entity properties associated with each `OligogenicCombination`, such as the OLIDA curation confidence scores, our focus is centered on digenic gene pairs that demonstrate reliable evidence.

We leveraged specifically the familial and statistical evidence scores to attribute a weak, moderate, and strong evidence level for each pathogenic variant combination. A weight was attributed to these instances according to the three defined levels of confidence (see Table 5.5).

Evidence level	Weight	Criteria based on OLIDA curation scores				
		FAMmanual	OR	(STATmeta	AND	STATmanual)
Weak	0.33	1	||	1	&	1
Moderate	0.67	2	||	2	-	-
Strong	1.00	3	||	3	-	-

Table 5.5. Oligogenic evidence levels based on familial and statistical scores. For each variant combination present in OLIDA, different types of curation confidence scores have been assigned according to the strength of evidences provided (*e.g.* relatives genotype information, cohort statistical analyses, etc). See [71] for more information on OLIDA confidence scores. For this method, we created an *Evidence level*, solely based on familial (FAMmanual) and statistical (STATmeta, STATmanual) curation confidence scores, associated with a linear weight. If none of the criteria are matching, the variant combination is discarded.

All variant combinations involving two genes satisfying at least a weak evidence level were considered, amounting to a total of 794 variant combinations. These were subsequently aggregated at the gene pair level, weighted by the maximum confidence level. A total of 441 disease-causing gene pairs (denoted as D) were selected after aggregation (Table 5.6).

Each selected gene pair, denoted as (G_S, G_T) for source/target, was ordered using the 'RVIS' gene property, the Residual Variation Intolerance Score (RVIS) [180], in ascending order, in line with the VarCoPP 2.0 predictor [277]. This ordering suggests that the source gene possesses fewer common functional genetic variations than the target gene, in relation to their number of neutral variants, meaning the source gene is more "intolerant" to variations. By considering their evidence level, we were able to assign the corresponding

Evidence level	**# variant combinations**	**# gene pairs**	**# diseases**
Discarded	324	130	71
Weak	577	280	108
Moderate	131	101	71
Strong	86	60	30
Selected	794	**441**	153

Table 5.6. Pathogenic gene pair selection according to evidence levels. Evidence levels have been attributed for each digenic variant combination in OLIDA (Table 5.5). All variant combinations passing at least the *Weak* evidence level criteria have been selected (a total of 794) and weighted accordingly and the remaining variant combinations have been discarded. Statistics are presented with a gene-level and disease-level aggregation, by attributing the highest evidence level of each set.

weight from Table 5.5 to each gene pair. This weighting provides valuable information for the machine-learning predictor, highlighting which instances should be the most influential during training.

Selection of an hold-out test set from recent literature

To provide an independent testing of the predictive models, 15 pathogenic gene pairs were held-out. This test set was selected based on an automatic procedure designed to favour diverse, confident and recently published cases: first, all disease-causing gene pairs were ranked by their first associated article publication date, then for each gene pair from the most recent to the oldest, gene pairs were chosen if their evidence level was at least Moderate (see Table 5.5) and if none of their genes overlapped with the previously selected ones. Details about the selected gene pairs are provided in Table 5.7.

5.4.2 Neutral gene pair selection using healthy controls

We developed a process to select a subset of gene pairs that, when mutated together, are unlikely to be causal of a disease. This strategy leverages data from healthy individuals. Our hypothesis is that gene pairs recurrently exhibiting mutation patterns of pathogenicity and frequency similar to those seen in digenic diseases across a large set of healthy individuals are less likely to contribute to disease phenotypes. A step-by-step overview of this selection process is presented in Figure 5.11.

We assessed the potential relevance of gene pairs based on the pathogenicity of the

Gene pair	Evidence level	Disease	PMID	Publication date
MYH7-ANKRD1	Moderate	Left ventricular noncompaction	34752814	06/11/2021
LAMA3-LAMB3	Moderate	Severe generalized junctional epidermolysis bullosa	34837689	01/10/2021
TSHR-SLC26A4	Moderate	Congenital hypothyroidism	34374102	10/08/2021
JAG1-DUOXA1	Moderate	Congenital hypothyroidism	34374102	10/08/2021
CDCA8-DUOX2	Moderate	Congenital hypothyroidism	34374102	10/08/2021
HOXB3-TG	Moderate	Congenital hypothyroidism	34374102	10/08/2021
MYO7A-SHROOM2	Strong	Atypical hemolytic uremic syndrome	34391192	02/08/2021
PKHD1-PKD1	Moderate	Autosomal recessive polycystic kidney disease	34032358	25/05/2021
POLG-PPFIA4	Moderate	Isolated focal cortical dysplasia	34095804	07/05/2021
SLC20A2-PDGFRB	Moderate	Bilateral striopallidodentate calcinosis	33793087	01/04/2021
TRAPPC11-TTN	Strong	Limb-girdle muscular dystrophy	33746696	04/03/2021
SPG7-SPAST	Strong	Hereditary spastic paraplegia	33598982	20/02/2021
MITF-C2orf74	Moderate	Waardenburg syndrome	33571247	12/02/2021
CHD7-CDON	Moderate	Kallmann syndrome	33208564	17/11/2020
BMPR2-NOTCH3	Moderate	Idiopathic pulmonary arterial hypertension	33007923	30/09/2020

Table 5.7. Characteristics of pathogenic gene pairs in the held-out test set. Details on the selected held-out test set made of 15 pathogenic gene pairs. This test set was selected based on an automatic procedure designed to favour diverse, confident and recently published cases: first, all disease-causing gene pairs were ranked by their first associated article publication date, then for each gene pair from the most recent to the oldest, gene pairs were chosen if their evidence level was at least Moderate (see Table 5.5) and if none of their genes overlapped with the previously selected ones.

known digenic variant combinations (see "Selected" in Table 5.6), estimated by the Combined Annotation Dependent Depletion (CADD) raw score [86] (see Material and Methods - Subsection 3.1.4).

We considered two key distributions within pathogenic combinations: $min(CADD(G_1), CADD(G_2))$ $(RefMinPatho)$ and $max(CADD(G_1), CADD(G_2))$ $(RefMaxPatho)$. These distributions are depicted in Figure 5.10.A and B, respectively.

From these distributions, we established pathogenicity thresholds: $Q1(RefMinPatho) = 1.89$ and $Q1(RefMaxPatho) = 3.57$. These thresholds represent the cutoffs above which 75% of pathogenic combinations have the pathogenicity scores of their least and most impacted genes, respectively.

Moreover, to mimic the frequency characteristic of digenic diseases, we selected a minimum allele frequency (MAF) cutoff of 0.03. This threshold was determined based on its capability to capture the majority of variants involved in digenic diseases (as established as a variant filter in the ORVAL platform [165]). Our neutral gene pair selection from healthy individuals thus focuses on gene pairs presenting variants as rare and pathogenic as the majority of known pathogenic variant combinations, increasing the likelihood that these gene pairs do not contribute to disease phenotypes.

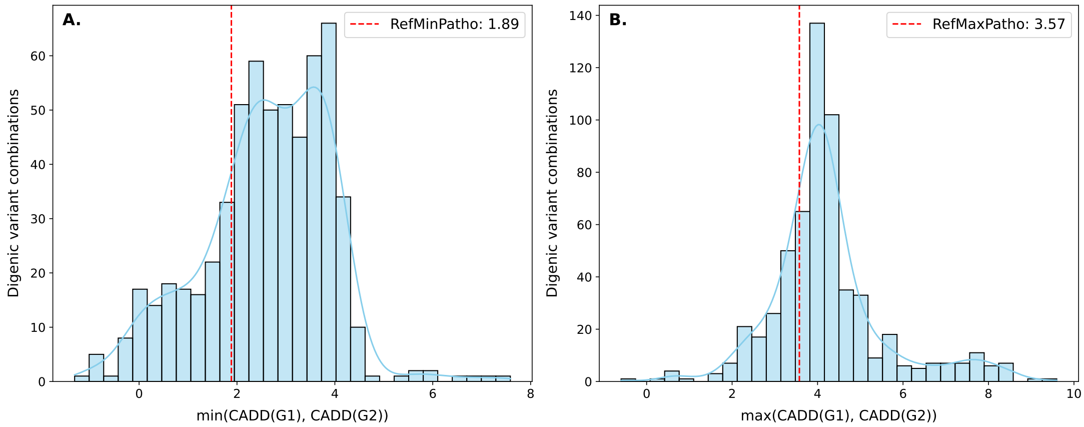


Figure 5.10. CADD score distributions of pathogenic gene pairs. The pathogenicity of digenic variants (detailed in Table 5.6) is quantified using the Combined Annotation Dependent Depletion (CADD) raw scores. For each gene, the maximum CADD score across all its variants (CADD(G)) is computed. **A.** The distribution of the lesser of the two CADD(G) scores from each gene pair (RefMinPatho). **B.** The distribution of the greater of the two CADD(G) scores from each gene pair (RefMaxPatho). The first quartile (Q1) of each distribution, represented by the red dotted line, serves as a pathogenicity threshold.

Building upon the criteria for pathogenicity and frequency, we sought to identify gene pairs from healthy individuals that presented similar variant characteristics. This process involved a multi-step filtering protocol applied to variant data from 2490 healthy individuals from the 1000 Genomes Project (1KGP) [18] (see Material and Methods - Subsection 3.1.2).

Initially, we filtered the variants based on a minimum allele frequency (MAF) cutoff of 0.03, reflecting the rarity characteristic of variants involved in digenic diseases. Variants were then aggregated at the gene level for each individual, preserving only the maximum CADD score for each gene. To ensure pathogenicity resemblance, each gene's CADD score had to be greater than or equal to the established $Q1(RefMinPatho) = 1.89$. Additionally, genes not present in BOCK or without any neighbours (*i.e.* not connected in the KG) are discarded at this stage.

We then generated gene pairs from the candidate genes of each individual. To emulate the characteristics of pathogenic pairs, we required that the maximum CADD score across both genes was at least $Q1(RefMaxPatho)$. This ensured these gene pairs also shared similar pathogenicity characteristics with known disease-causing gene pairs.

Candidate neutral gene pairs were then generated and filtered, based on their fre-

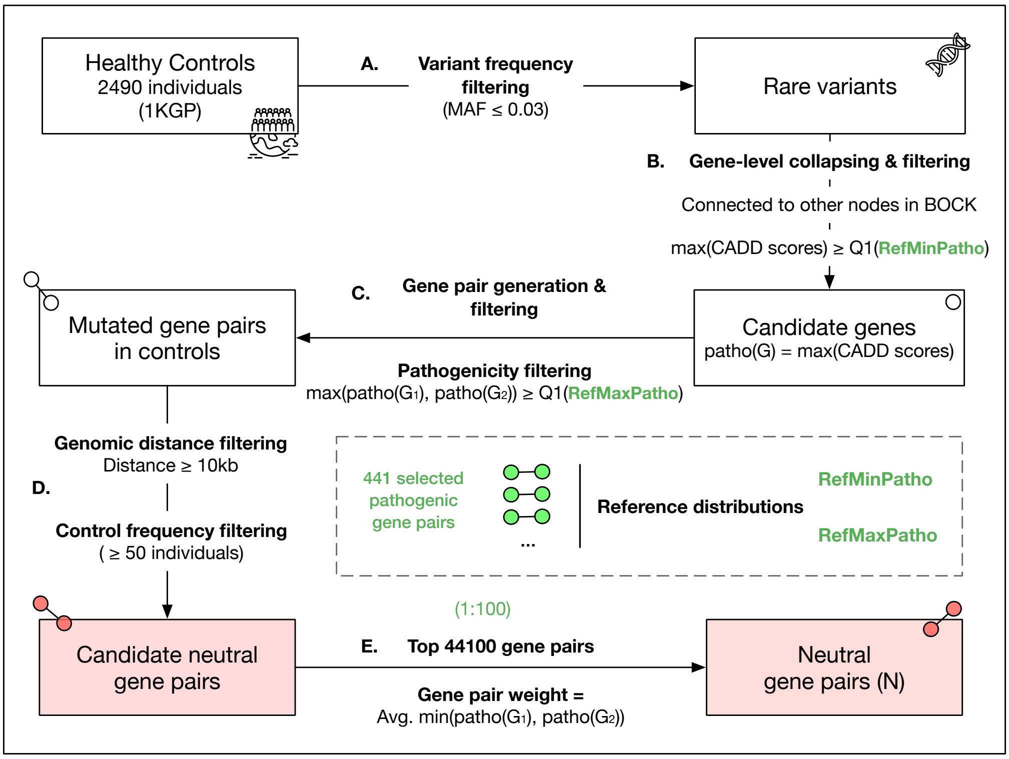

Healthy Controls
2490 individuals
(1KGP)
A.
Variant frequency filtering
(MAF ≤ 0.03)
Rare variants
B.
Gene-level collapsing & filtering
Connected to other nodes in BOCK
max(CADD scores) ≥ Q1(RefMinPatho)
C.
Gene pair generation & filtering
Mutated gene pairs in controls
Candidate genes
patho(G) = max(CADD scores)
Pathogenicity filtering
max(patho(G₁), patho(G₂)) ≥ Q1(RefMaxPatho)
Genomic distance filtering
Distance ≥ 10kb
D.
Control frequency filtering
(≥ 50 individuals)
441 selected pathogenic gene pairs
...
Reference distributions
RefMinPatho
RefMaxPatho
(1:100)
Candidate neutral gene pairs
E.
Top 44100 gene pairs
Gene pair weight =
Avg. min(patho(G₁), patho(G₂))
Neutral gene pairs (N)

Figure 5.11. Generation of neutral gene pair sets. This diagram outlines the process for generating Neutral gene pair sets (N) from healthy control individuals' data. **A.** Rare variants from healthy controls (1KGP) are selected based on allele frequency. **B.** These rare variants are aggregated at the gene level to create candidate genes, with the maximum CADD score across all variants representing each gene's pathogenicity potential ($patho(G)$). Genes not present in BOCK or without any neighbours are discarded. Additionally, genes with pathogenicity score below $Q1(RefMinPatho)$ are filtered out (see Figure 5.10.A). **C.** Candidate gene pairs are formed from these genes, ensuring the maximum pathogenicity in each pair surpasses $Q1(RefMaxPatho)$ (see Figure 5.10.B). **D.** Additional filters are applied to gene pairs based on genomic distance and their frequency among controls to derive neutral gene pairs. **E.** Each neutral gene pair is weighted based on the minimum pathogenicity score across all occurrences, then the top 44100 gene pairs are selected to form the Neutral set. This selection reflects an imbalance ratio of 1:100 relative to disease-causing gene pairs, mirroring the infrequency of pathogenic gene interactions.

quency, to ensure that each selected gene pair occurred in a minimum of 50 healthy individuals. Additionally, in order to minimise the impact of linkage disequilibrium, we incorporated an additional constraint: neutral gene pairs had to be either from different chromosomes or at least 10kb apart in genomic coordinates.

The weight of each gene pair was then calculated by averaging the lesser gene-aggregated CADD score across all associated individuals, and this weight was scaled against the maximum value across all selected gene pairs.

To create the final set of neutral gene pairs (N), we chose the top 44,100 pairs with the highest scores, yielding an imbalance ratio of approximately 1:100 between disease-causing and neutral gene pairs. This ratio acknowledges the fact that pathogenic gene interactions are far rarer than non-pathogenic ones [287]. Despite a considerable number of potential gene interactions, only a fraction contribute to disease phenotypes, justifying the larger imbalance in our data.

We present some statistics of the chosen set of neutral gene pairs in Figure 5.12. It can be noted that the majority of selected gene pairs are associated with fewer than 100 control individuals (median: 68 individuals). The associated weight, taking into account both the number of control individuals and the underlying variant pathogenicity, predominantly ranges between 0.37 and 0.8 (median: 0.61).

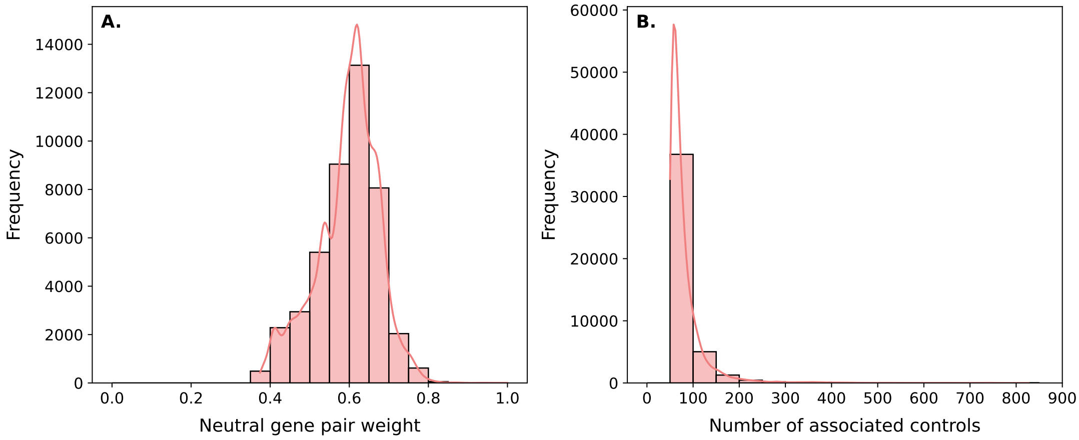


Figure 5.12. Selected gene pairs weight and associated controls distribution The characteristics of the selected set of neutral gene pairs (N) are presented as distributions. **A.** Neutral gene pair weights, calculated by averaging the lesser of the two gene-aggregated CADD scores. **B.** Number of control individuals (from the 1000 Genomes Project cohort) associated with each neutral gene pair.

5.5 Network analysis of oligogenic gene combinations

In this section, we explore the topological characteristics of the knowledge graph we constructed, focusing specifically on pathogenic and likely neutral gene pairs described in the previous section (Section 5.4). By analysing the neighbourhood and relationships of these instances in the graph, we aim to identify distinctive topological features that might aid in their inference and reveal potential biases.

5.5.1 Gene set connectivity and annotation biases

We begin our exploration by first analysing individual genes composing these sets, in particular their connectivity and the diversity of these relationships in BOCK.

Comparison by edge type diversity

We first analysed the proportion of genes connected with at least one edge of a certain type (also referred as gene annotation coverage) in BOCK. We evaluated four gene sets: *Disease genes* gathered from disease-gene associations in Orphanet data [205], *Digenic genes* from the pathogenic gene pairs collected in Subsection 5.4.1, *Neutral genes* from the neutral gene pairs collected in Subsection 5.4.2 and *Connected genes* corresponding to all genes in BOCK with at least one connection (Figure 5.13).

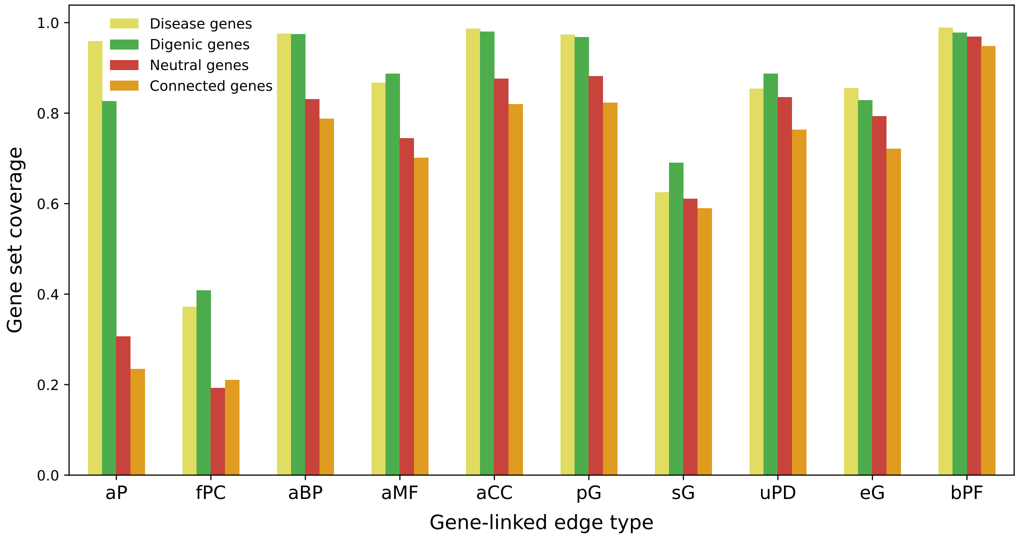


Figure 5.13. Gene set coverages comparison for different edge types. Considering genes from different sets, we determine the fraction of genes connected with a specific edge type (abbreviated according to Table 5.1). We compare different gene sets in BOCK: *Connected genes*: all genes with at least one neighbour in BOCK ; *Disease genes*: all genes involved in a disease, based on Orphanet data ; *Digenic genes*: all genes involved in a digenic gene pair from the selected set defined in Subsection 5.4.1 and *Neutral genes*: all genes involved in a neutral gene pair from the selected set defined in Subsection 5.4.2. Edge types are ordered by decreasing ratio difference between digenic genes and neutral genes. [aP:associated-Phenotype; fPC:forms-ProteinComplex; aBP:associated-BiologicalProcess; aMF:associated-MolecularFunction; aCC:associated-CellularComponent; pG:physInteracts-Gene; sG:seqSimilar-Gene; uPD:hasUnit-ProteinDomain; eG:coexpresses-Gene; bPF:belongs-ProteinFamily].

This analysis, on the one hand, underscores a similar annotation coverage ratio across all edge types for both disease-associated genes and digenic genes. Notably, a divergence is observed in the case of phenotype associations (aP), likely resulting from the distinct roles some genes play in digenic diseases, where they function as modifiers rather than direct causes of specific phenotypes. Neutral genes, on the other hand, largely mirror the trend seen for all BOCK-connected genes, showing a modest average enrichment in coverage across all edge types (1.07-fold).

However, digenic genes, like their disease-related counterparts, significantly deviate from the general trend seen in BOCK's connected genes (and therefore the neutral genes). These genes are generally more annotated, evidenced by a notable 1.55-fold average coverage enrichment across all different edge types.

More specifically, digenic genes show their most significant annotation coverage discrepancy for associated-Phenotype (aP) – a 3.52-fold enrichment compared to all connected genes. This is followed by forms-ProteinComplex (fPC) with a 1.95-fold enrichment. Other edge types, such as associated-BiologicalProcess (aBP), show a far smaller difference.

The significant annotation divergence for phenotype associations is attributable to a double bias: first, phenotype annotations primarily originate from literature, possibly even the studies from which digenic gene pairs were extracted; second, research aimed at identifying causal genes are often limited to gene panels already associated with known diseases or phenotypes. It is therefore important to consider the strong phenotype annotation bias and the limited coverage of human genes, which make up approximately 23% of genes connected in BOCK, when planning future machine learning applications.

Comparison by edge type connectivity

Further insights can be derived from analysing degree centrality, a measure that denotes a gene's level of connectivity within a network (see Material and Methods - Subsection 3.3.1). In KGs, this metric can be further dissected into type-specific degrees, indicating how many neighbouring connections exist for a specific edge type.

Interestingly, digenic genes exhibit greater connectivity than their neutral counterparts for most edge types (Figure 5.14). The edge types associated-Phenotype (aP) (median difference: 19.0), physInteracts-Gene (pG) (median difference: 7.0), and associated-

BiologicalProcess (aBP) (median difference: 4.0) demonstrate the largest connectivity disparity when compared to neutral genes. Conversely, the degree of other Gene-Gene interaction types and forms-ProteinComplex (fPC) show no substantial differences between these sets.

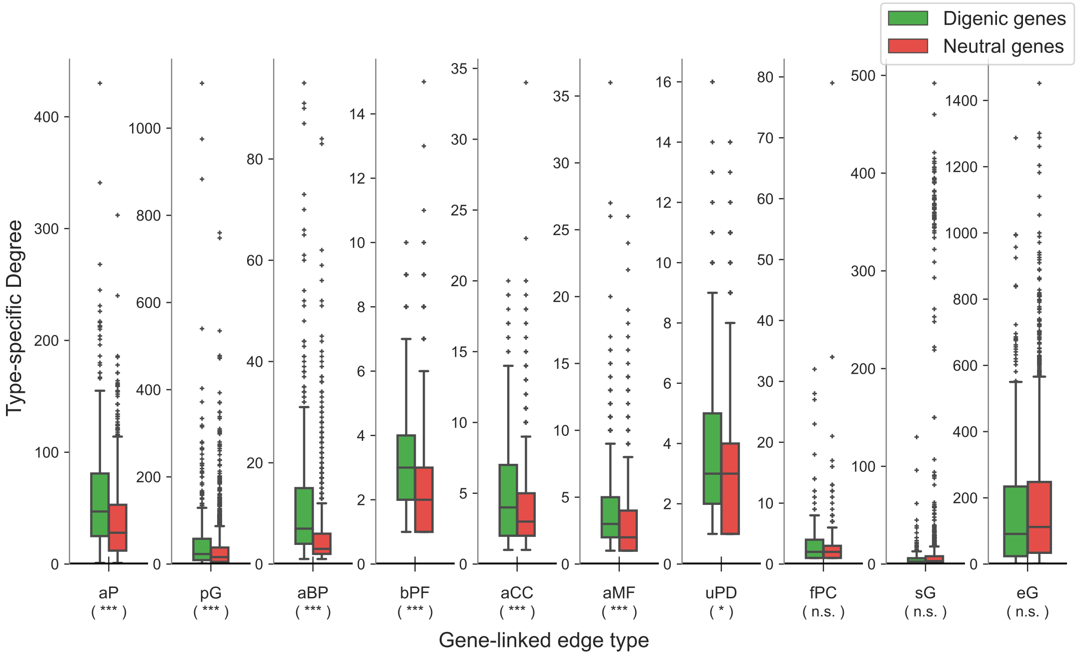


Figure 5.14. Degree distribution comparison between digenic and neutral genes for different edge types. The comparison is made between genes involved in digenic (see Subsection 5.4.1) and neutral (see Subsection 5.4.2) gene pairs. Each subplot represents a specific gene-linked edge type in BOCK (abbreviated as per Table 5.1), with boxplots depicting the type-specific degree centrality of genes within each set. Edge types are ordered by the median difference between the digenic and neutral sets, and significant distinctions between the two gene sets are indicated below each type (p-values obtained via Mann-Whitney U tests with Bonferroni correction: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). [aP:associated-Phenotype; pG:physInteracts-Gene; aBP:associated-BiologicalProcess; bPF:belongs-ProteinFamily; aCC:associated-CellularComponent; aMF:associated-MolecularFunction; uPD:hasUnit-ProteinDomain; fPC:forms-ProteinComplex; sG:seqSimilar-Gene; eG:coexpresses-Gene;].

Comparing the distribution of digenic and disease-related genes (see Table A.2 in Appendix) reveals that digenic genes possess a greater depth of Phenotype annotation. Furthermore, they tend to have more annotations linked to Biological Processes than disease-related genes. In contrast, neutral genes exhibit similar degree distributions across all types compared to all connected genes.

To summarise, these analyses reveal that digenic and disease-related genes share similar annotation diversity and connectivity, while neutral genes resemble the overall gene characteristics within BOCK, suggesting that our neutral gene pair selection process is effectively selecting genes that are representative of human gene average topologies.

When classifying gene pairs based on these genes, we should therefore be mindful of known biases observed in previous work identifying single pathogenic genes with network topology [288]. Indeed, disease genes tend to have more connections in biological networks, due to the well-known *study or literature bias* [289, 290, 291]. The same study bias directly impact phenotype coverage and depth. Therefore, much of the performance of classifiers using degree-based features or phenotype annotations may simply be due to these study biases, rarely accounted for in evaluation and test datasets.

5.5.2 Reachability of disease-causing gene pairs

We then focused our attention to the relationship between disease-causing gene pairs by examining their node pair reachability, that is, the potential to navigate from one gene in a pair to the other within a specified maximum path length. We label two genes as 'reachable' within a given graph if there exists at least one path, irrespective of direction, connecting them within the constraints of the defined path length.

To manage computational complexity and maintain the significance of gene relationships, we restricted our analysis to paths of four steps or fewer. Our results, as depicted in Figure 5.15.a and b, show that only a minority of disease-causing gene pairs have close relationships (1 or 2-hops) in each individual component. Considering longer paths (3 or 4-hops) significantly increases the coverage of pathogenic gene pair, but this comes at the price of additional noise, reduced influence and computational cost.

Integrating these components as a single graph (*GENE-CENTRIC*, *COMPOSITE*) makes the majority of gene pair reachable for 2-hop paths (up to 90% for 3-hop paths) (Figure 5.15.c). We also find that paths of lengths ≤ 3 are sufficient to connect all disease-causing gene pairs in the knowledge graph (*ALL*), even when excluding potentially bias-prone Phenotype information (*ALL_NOPHENO*) (Figure 5.15.c).

These findings underscore the value of integrative approaches in studying gene relationships in oligogenic diseases, given their inherent heterogeneity and the diverse epistatic mechanisms at play.

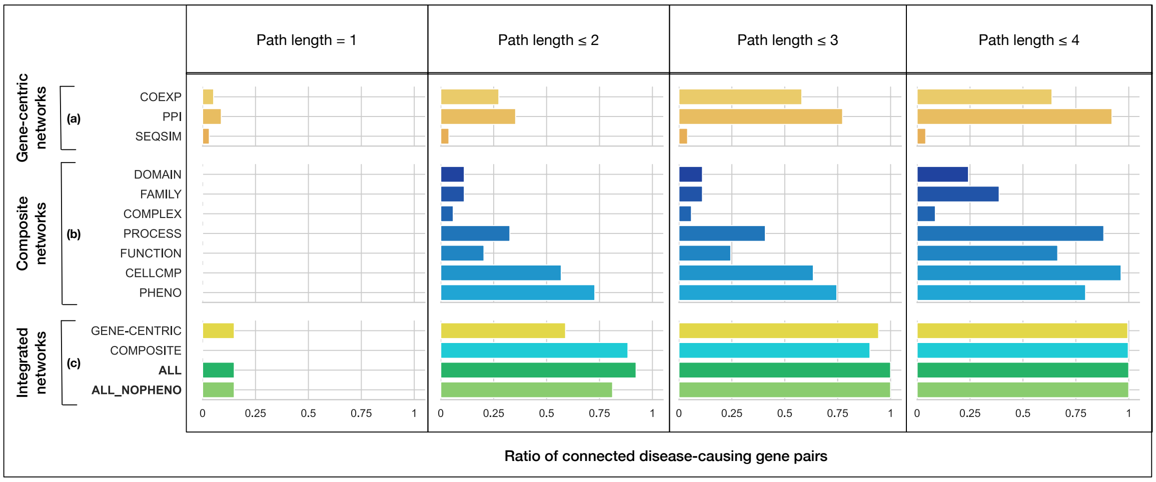


Figure 5.15. Ratio of reachable disease-causing gene pairs in different components of BOCK. A gene pair is considered reachable if there exists a path between the two genes, regardless of directionality, that can be traversed given a path length cutoff. Nodes of types "Disease" and "OligogenicCombination" were excluded and BOCK was decomposed into: a) gene-centric networks (COEXP: Gene-coexpresses , PPI: Gene-physInteracts, SEQSIM: Gene-seqSimilar) shown merged as GENE-CENTRIC ; b) composite networks (DOMAIN: Gene-hasUnit-ProteinDomain, FAMILY: Gene-belongs-ProteinFamily, PROCESS: Gene-associated-BiologicalProcess, FUNCTION: Gene-associated-MolecularFunction, CELLCMP: Gene-associated-CellularComponent, PHENO: Gene-associated-Phenotype) shown merged as COMPOSITE ; c) integrated networks (GENE-CENTRIC: merge of a), COMPOSITE: merge of b), ALL: merge of a) and b), ALL_NOPHENO: a subset of ALL excluding paths traversing "Phenotype" nodes). The ratios of reachable disease-causing gene pairs with at least a weak evidence-level (see Table 5.5) are presented for these components according to path length cut-offs ranging from 1 to 4.

5.5.3 Topology-driven link prediction

In this section, we aim to analyse the predictive capabilities of various graph topology methods and measurements within the BOCK system, specifically for distinguishing pathogenic gene pairs from neutral ones (Section 5.4). For the forthcoming analyses, we chose to remove nodes of type *Disease* and *OligogenicCombination* from BOCK, and we chose to break down the knowledge graph into components as outlined in Table 5.4.

Topological predictive measures

We assessed the following topology metrics, each of which captures a different aspect of topological information: Total Neighbours (TN), Preferential Attachment (PA), Jaccard Index (JI), Adamic Adar (AA), inverse Shortest Path Length (iSPL), and Random Walk with Restart (RWR). These metrics are further detailed in the Material and Methods (Subsection 3.3.3).

To assess the performance of each topological metric for a given knowledge graph component, we trained and evaluated a logistic regression model based on that topological metric within the network component. The model's evaluation was carried out with a 10-fold stratified cross-validation. The heatmap in Figure 5.16 summarises the results, with performance measured by the Area Under the Receiver Operating Characteristic (AUROC) and the Area Under the Precision-Recall Curve (AUPRC).

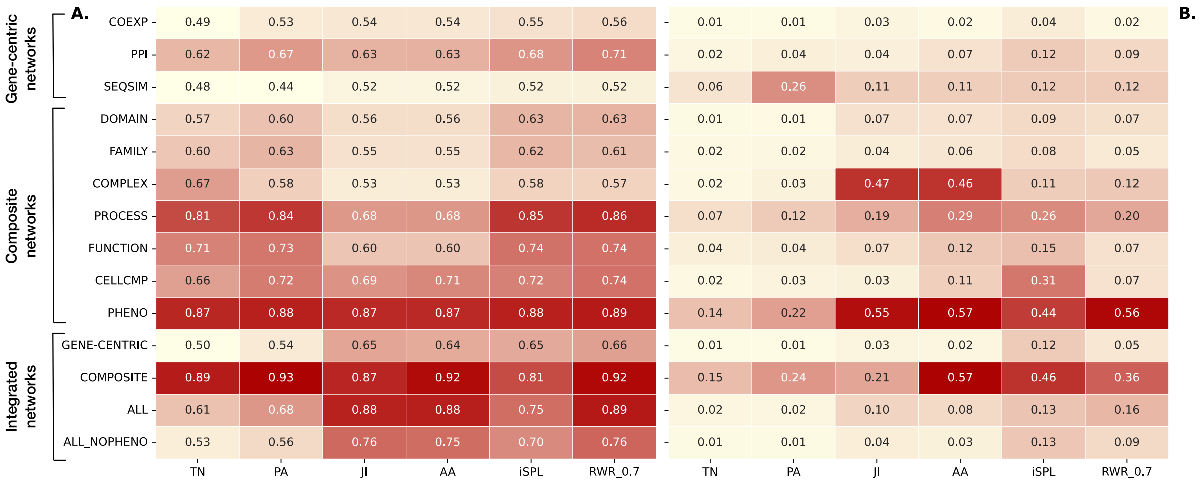


Figure 5.16. AUROC performance of common topological measures in BOCK and its components for pathogenicity prediction. Heatmaps showing the mean Area Under the Receiver Operating Characteristic (AUROC) of a logistic regression model trained with different topological measures (TN: Total Neighbours, PA: Preferential Attachment, JI: Jaccard Index, AA: Adamic Adar, iSPL: inverse Shortest Path Length, RWR: Random Walk with Restart (with restart probability = 0.7 [148])) and evaluated with a 10-fold stratified cross-validation. Results are reported for each KG component (see Table 5.4) and integrated networks (GENE-CENTRIC: Fusion of gene-centric networks, COMPOSITE: Fusion of composite networks, ALL: entire KG without *Disease* and *OligogenicCombination* ; ALL_NOPHENO: ALL without Phenotype).

The results reveal that certain network components offer more predictive power than others across all types of measurements. All components that include the Phenotype layer (PHENO, COMPOSITE, and ALL) perform the best. However, as shown in earlier analyses (see Subsection 5.5.1), this high performance is largely influenced by a study bias affecting all disease-related genes. Coexpression and sequence similarity related components, conversely, do not yield as much predictive power.

The different metrics we studied provide varied insights. TN and PA provide basic measures calculating the overall connectivity of the gene pairs, but do not deliver as much discriminative power compared to other measures, across all types of components.

JI and AA are metrics that account for the direct common neighbours of the entity pair. These measures can offer higher precision when considering individual components but lack the recall offered by other more flexible metrics. For integrated networks, they offer a good predictive power, though less robust than RWR.

iSPL and RWR provide measures of connectivity, taking into account indirect, potentially long-range, relationships. Notably, they perform the best for the individual component PROCESS, suggesting that biological processes are more readily captured via indirect interactions. RWR, however, significantly outperforms the simpler iSPL for integrated networks. More details about RWR results are provided next.

Random-walk analysis

The Random Walk with Restart (RWR) is a measure that offers a unique perspective on connectivity within a network. It accounts for both direct and indirect interactions by incorporating a probabilistic element that enables the random walker to either move to a neighboring node or return to the start. To explore the efficacy of RWR, we applied it to our existing models in the same manner as before, and we present the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves in Figure 5.17, with RWR applied to BOCK, both with and without phenotype data.

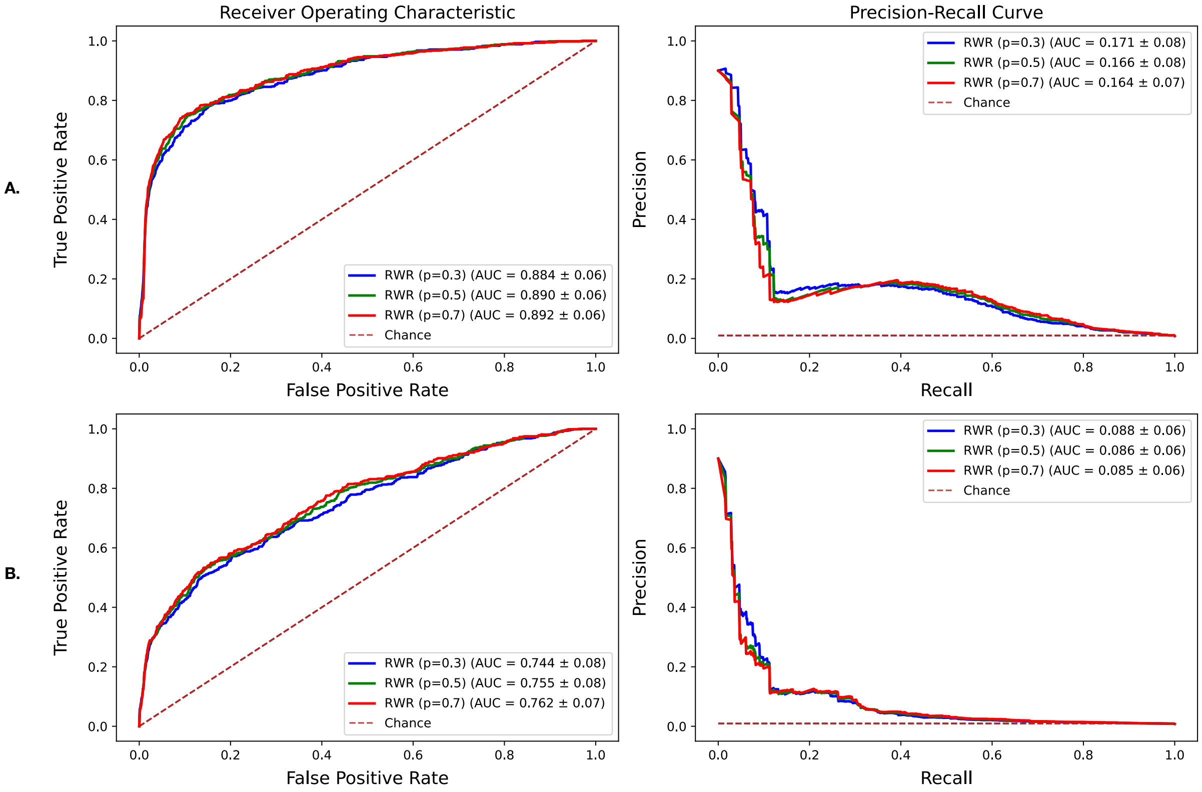


Figure 5.17. Performance of Random Walk with Restart for gene pair classification. We use the Random Walk with Restart (RWR) algorithm combined with a logistic regression model to predict the probability that a gene pair is disease-causing, trained on the selected disease-causing and neutral gene pairs. We evaluate this classifier with a 10-fold stratified cross-validation. The RWR algorithm is evaluated for three restart probabilities (p=0.3, 0.5, 0.7) on the knowledge graph **(A)** including and **(B)** excluding Phenotype information.

Our analysis indicates that the RWR probability alone delivers a respectable performance baseline for assessing the predictive power of the topological proximity within BOCK. This holds true for the knowledge graph that includes Phenotype information (mean AUPRC = 0.159) and also for the one that excludes Phenotype information (mean AUPRC = 0.084). Interestingly, there is no significant performance difference when considering different restart probabilities, consistent with previous findings [148].

Thus, the RWR analysis further strengthens the argument that indirect and potentially long-range relationships, play an important role in predicting pathogenic gene interactions.

However, while the RWR measure is effective, it may not fully capture the complexity and diversity of entities and relationships found in knowledge graphs. Consequently, it's necessary to explore more advanced and tailored techniques, such as the one we present in the following section, to improve prediction accuracy and better harness the richness of knowledge graph data.

5.6 Towards high-accuracy pathogenicity prediction using KG embeddings

In this section, we aim to show some preliminary results on the application of a state-of-the-art KG embedding technique to predict the pathogenicity of gene pairs. These analyses were conducted by *Inas Bosch*, a master student whom I had the opportunity to supervise. These findings, detailed in her master's thesis (see Section 2.5), demonstrate the potential of KG embeddings for accurate predictions and help paving the way for future applications of BOCK in accurately detecting pathogenic gene pairs.

While traditional link prediction approaches directly leverages the connections present in our knowledge graph, BOCK, KG embedding approaches aim to create a latent representation of entities and relationships, which can capture more efficiently the semantic and structural intricacies of such network.

Our aim in this study was to leverage these representations in a downstream link prediction task for discerning pathogenic gene pairs, with the same training dataset (Section 5.4) and classification task as delineated in the previous section (Subsection 5.5.3).

5.6.1 KG embedding model and downstream classification

Choice and evaluation of KG embedding methodology

A key decision was the choice of the Knowledge Graph (KG) embedding model. Different KG embedding models were investigated based on models proposed in the PyKEEN library [250], with DistMult [252] and ComplEx [292] emerging as prominent contenders [250]. Given DistMult's comparable performance to ComplEx on the YAGO3-10 dataset and its inherent simplicity (while ComplEx uses vectors of complex numbers), the DistMult model was chosen (see Material and Methods - Subsection 3.7.5).

To hits@10 metric metric was used to assess the quality of the knowledge graph representation, by measuring how frequently the correct link appears within the top 10 predictions. Our results yielded scores of 0.15 when including phenotype data and 0.14 when omitting it. This relatively low performance, compared to the one on the YAGO3-10 [293] benchmark dataset, may be attributed to the incompleteness and noise usually observed in biological networks. Nevertheless, this metric gauges a general graph competition objective, which may not reflect performance on a more specific downstream task of gene pair pathogenicity prediction.

From KG embeddings to a pathogenicity prediction model

We devised a workflow, summarised in Figure 5.18, following a two step procedures. The initial phase produces latent vectorial representations for all entities and relationships in BOCK. We focused here on the entity embeddings and more specifically the Gene entities. By using a specific vectorial transformation, we could obtain a unique vectorial representation for each gene pair. The next phase uses these gene pair representations as features to train a Balanced Random Forest classifier (BRF) [224], using the labelled gene pairs presented in Section 5.4 for training. The predictor can then infer a probability of belonging to the disease-causing class and be evaluated in cross-validation and on an independent dataset.

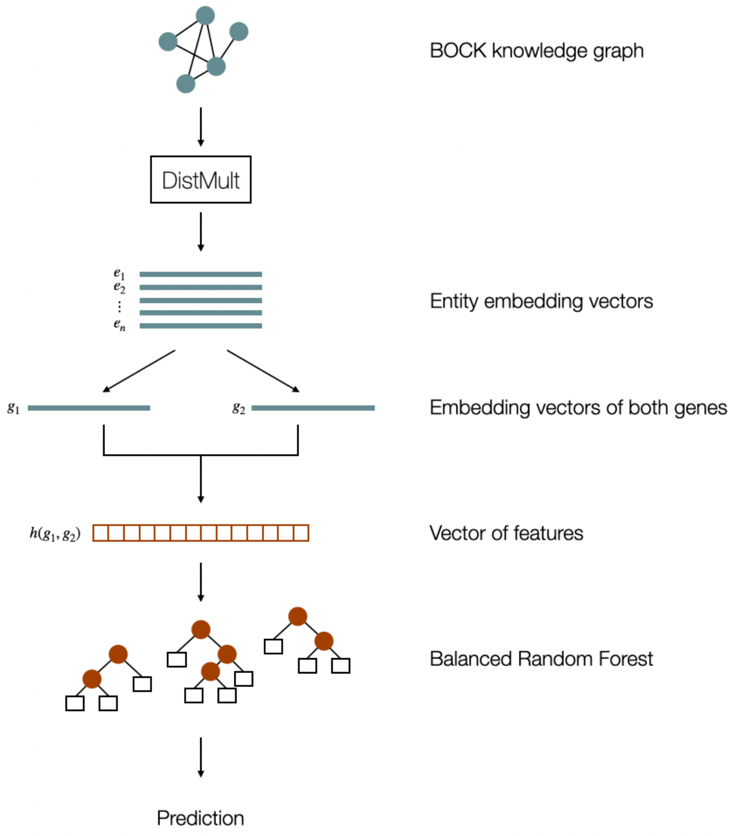


Figure 5.18. Workflow from KG embeddings to pathogenicity prediction. The prediction process is structured in two main phases. The initial phase involves using an embedding model to derive embeddings (*i.e* latent vectorial representations) for all entities in the graph. This representation captures intricate semantics and topology inherent in the knowledge graph. Gene pairs (g_1 ,g_2) can then be represented as single vectors via a transformation h integrating g_1 and g_2 vectors. In the subsequent phase, these embeddings serve as input to a Balanced Random Forest classifier which then predicts the pathogenicity of the gene pair. Credit figure: Inas Bosch.

5.6.2 Prediction performance and key findings

Optimal parameters

The performance of an embedding generation model is significantly influenced by two key parameters: the embedding dimension and the number of epochs. A larger embedding dimension allows the model to encapsulate more nuanced aspects of the knowledge graph, and a greater number of epochs gives the model more opportunities to refine its representation of the graph. However, it's important to note that augmenting these parameters can yield diminishing returns beyond a certain point, due to increased dimensionality and computational demands.

Considering the BRF model used for classification, we investigated the impact of varying the number of trees. The number of trees is a key parameter in a BRF as it can influence the model's robustness and ability to generalise, while potentially affecting computational efficiency.

The optimal parameters, determined based on their performance in pathogenicity prediction tasks with and without phenotype information in the knowledge graph, are summarised in Table 5.8.

Parameter	With phenotype	Without phenotype
Embedding dimension	200	300
Number of epochs	50	50
Number of trees	340	540

Table 5.8. Optimized parameter values for the models incorporating and excluding phenotype information. Credit Table: Inas Bosch.

Prediction performance and influence of vector transformation

We conducted an analysis of the downstream predictive performance influenced by two categories of embedding vector transformations: embedding-aggregating [254] and embedding-preserving transformations. The former modifies the components of embedding vectors , while the latter combines but maintains the integrity of individual embedding vectors. We evaluated various embedding-aggregating transformations including Average, Hadamard, WeightedL1, and WeightedL2 [254].

Two types of embedding-preserving transformations were evaluated: the *concatenate* and *double-concatenate* methods. *Concatenate* is a straightforward approach that joins two gene vectors in an order defined by RVIS. On the other hand, *double-concatenate* incorporates both possible orderings of the gene vectors, generating two feature vectors for every gene pair. Consequently, the Random Forest model produces two probabilities for each pair, which are then averaged to yield a singular prediction.

Our primary results, as depicted in Table 5.9 and evaluated using a BRF model with 100 trees evaluated with stratified 10-fold cross-validation, indicate that the preservation approach outperforms the aggregation approach. This implies that retaining original vector components benefits the classification of gene pair pathogenicity. Moreover, using only the first or second gene vector already offers high performance, suggesting that a large part of the predictive signal comes from single gene information. Interestingly, the *double-concatenate* method yields better results than the simple concatenation, which indicates that the gene order does not impact the prediction.

Vector Transformation	Without phenotype information		With phenotype information	
	AUROC	AUPRC	AUROC	AUPRC
Average	0.72	0.16	0.93	0.47
Hadamard	0.62	0.04	0.92	0.51
WeightedL1	0.62	0.03	0.85	0.23
WeightedL2	0.66	0.03	0.84	0.24
First gene	0.91	0.56	0.95	0.66
Second gene	0.92	0.55	0.95	0.65
Concatenate	0.91	0.54	0.95	0.65
Double-concatenate	**0.96**	**0.72**	**0.97**	**0.67**

Table 5.9. Predictive performance of models for different embedding vector transformation. Considering BOCK with and without the Phenotype information, AUROC and AUPRC performance, obtained by stratified 10-fold cross-validation are shown for each embedding vector transformation.

Our final models, incorporating optimal BRF configurations, produced impressive results. The model without phenotype information achieved an AUROC of 0.98 and an AUPRC of 0.84, successfully recalling 14/15 pathogenic gene pairs in the independent test dataset. Meanwhile, the model including phenotype information reached an AUROC of 0.97 and an AUPRC of 0.70, recalling 13/14 pathogenic gene pairs from the same dataset.

Despite the high predictive accuracies of models using latent representations from KG embedding methods, these models are often viewed as 'black boxes' due to the irreversible loss of information in the initial embedding stage, rendering any post-hoc interpretability of the downstream classifier ineffective. Additionally, the model's strong performance using just single gene information hints that predictions may rely heavily on monogenic signals. It is unclear whether the model takes into account the relationship types and connectivity between gene pairs or simply processes their information independently.

Given the importance of interpretability in the biomedical field, we suggest considering alternative methodologies as a complementary approach, like the one we introduce in the next chapter (Chapter 6), in order to allow for domain validation and help geneticists and researchers derive meaningful insights.

5.7 Conclusion

Knowledge graphs have emerged as essential structures for consolidating complex datasets and providing detailed insights, particularly in biomedical research. In this landscape, we developed BOCK, a new knowledge graph that merges data from different oligogenic diseases with information from several biological networks and biomedical ontologies. This combination aims to capture the complex relationships found in oligogenic diseases, offering researchers a potential framework to systematically study gene relationships backed by solid evidence. Additionally, BOCK opens up new opportunities for feature engineering and machine-learning applications, laying a foundation for future predictive models.

Ensuring data quality and accuracy was a primary focus during BOCK's development. We carefully selected the data sources for BOCK to reduce the typical inconsistencies seen in large knowledge graphs. Looking ahead, expanding and consolidating the data in BOCK will be important. This includes considering additional types of biological relationships and possibly fusing reliable sources to increase its robustness.

BOCK has already found applications to enhance the Variant Combination Pathogenicity Predictor (VarCoPP 2.0 [277]) and as an essential component of the High-throughput Oligogenic Prioritizer (HOP) (future work). Preliminary results on the use of knowledge graph embedding techniques on BOCK are also very promising. Early findings suggest that these methods can accurately detect potential pathogenic gene pairs in BOCK. As we continue to refine these techniques and enrich BOCK's data, we hope to drive significant progress in oligogenic disease research.

Chapter 6

ARBOCK: A knowledge graph approach to predict and interpret pathogenic gene interactions

This chapter builds upon the earlier phases of our research, beginning with the creation of ORVAL, a web platform assisting geneticists in exploring and interpreting patient variant data (Chapter 4). ORVAL incorporates two machine learning models able to identify pathogenic variant combinations and to differentiate their effects. It also supplements these predictions with variant filters, feature contribution analysis and network representations. However, while ORVAL makes efforts to connect predictions to biological background knowledge, it primarily relies on post-hoc analyses for interpretation. This highlights the necessity for an approach that can provide clearer insights into predictive patterns and offer more direct connections to underlying biological mechanisms.

To further explore such complex patterns, our attention turned to the construction of the Biological Oligogenic Knowledge Graph (BOCK), presented in Chapter 5. BOCK serves as a way to contextualise oligogenic disease information, integrating data from curated clinical evidences together with multiple biological networks and ontologies. BOCK's creation opens up possibilities to leverage multi-level information for complex query formulation and inference. Nevertheless, our experience revealed that manually crafting these queries was not only intricate but also resulted in large graph subsets, making the exploration of the results challenging. Initial attempts to infer links between gene pairs showed that while a predictive signal for pathogenicity was discernible, these methods either yielded limited performance or lacked insightful explanations.

In light of these challenges, in Chapter 6, we developed ARBOCK - a novel machine learning approach designed to harness the rich information contained within BOCK. ARBOCK's primary goal is to decipher intricate relational patterns between pathogenic gene pairs, predict their pathogenicity potential, and more importantly, provide meaningful explanations for these predictions, addressing the interpretability gap.

The chapter begins by exploring the concept of metapaths and their role in creating interpretable predictions (Subsection 6.2.1). We then introduce our framework that combine these metapaths as rules for link prediction (Section 6.3). Subsequent sections investigate the discovered patterns associated with pathogenic gene pairs (Section 6.4) and evaluate the predictive performance of our predictive rule-based model (Section 6.5). We round off the chapter by demonstrating the inherent interpretability of this model and the ability to explain predicted pathogenic gene pairs with subgraphs from BOCK (Section 6.7).

We aim to demonstrate, in this chapter, the potential of biological knowledge graph like BOCK to support explainable predictions in the context of biomedical genetics. In doing so, we hope to offer a new approach to validate predictions using background knowledge and for generating mechanistic hypotheses, ultimately enhancing our understanding of oligogenic diseases.

The ARBOCK framework has been published, together with BOCK, in *BMC Bioinformatics* (forthcoming) [164].

It is provided as an open-source library at: `https://github.com/oligogenic/ARBOCK` with all the scripts and models ensuring the reproducibility of the results presented in this thesis (see Subsection 6.3.6).

6.1 Modelling assumptions and objectives of ARBOCK

In this work, we aimed to develop a novel classifier able to identify gene pairs where the presence of certain mutation patterns could lead to a disease or influence its manifestation. We follow similar modelling simplifications as the Variant Combination Pathogenicity Predictor (VarCoPP) [83, 277] in several respects: treating the problem as a binary classification task, centering our attention on pairwise gene interactions, and using the same data sources for training our model (see Material and Methods - Subsection 3.6.2).

However, while VarCoPP consider a range of features at the variant, gene and gene combination level, we focus solely on features reflecting relationships between gene pairs. Though the individual gene and variant characteristics hold significance, the patterns of their relationships alone might provide sufficient predictive information to identify candidate gene pairs. By focusing exclusively on short and long-range relationship features and by designing a fully interpretable model, our approach seeks to provide a deeper understanding of potential molecular mechanisms that underlie disease pathogenesis.

6.2 Capturing complex relationships with metapaths

We propose to model complex relationships between gene pairs based on path information contained in our designed knowledge graph, BOCK. More precisely, we abstract this path information with metapaths. Metapaths, or path types, provide a way to capture and quantify the relationships between entities in a graph and can also be used as KG queries, enhancing interpretability (see Material and Methods - Subsection 3.7.3).

In this section, we explore metapaths extracted from pathogenic gene pairs in BOCK and assess their predictive power. We also describe a simple method to rank and select paths underlying metapaths and demonstrate how this can improve their interpretability when used as KG query.

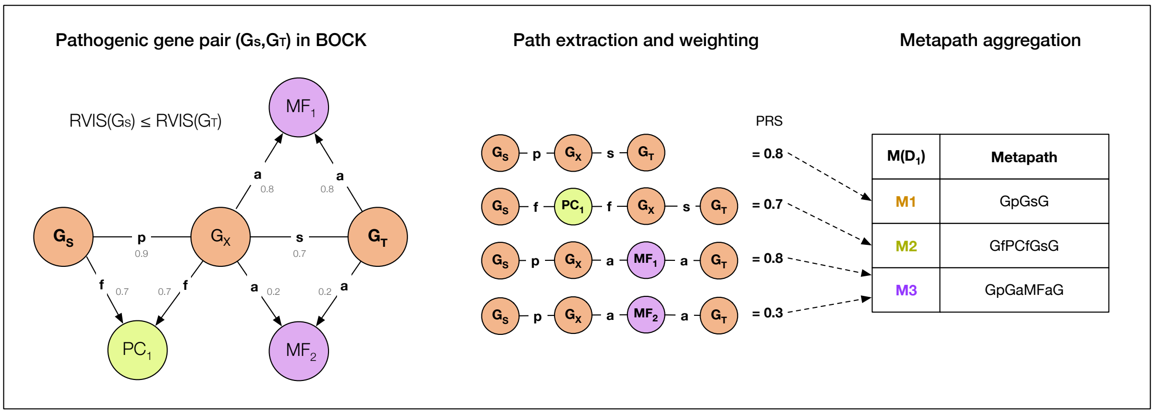
6.2.1 Leveraging metapaths for interpretable predictions

We first explain the methodology employed to extract metapath features describing pathogenic gene pairs from BOCK and then provide an overview of metapath predictive power.

Pathogenic gene pair paths to metapaths

Following the same logic of VarCoPP [277], gene pairs in our training dataset (see Section 5.4) are first ordered into a source gene (G_S) and a target gene (G_T), according to the Residual Variation Intolerance Score (RVIS) [180], indicating that the source gene has less common functional genetic variation than the target gene, in relation with their respective number of neutral variants (*i.e* the source gene is more "intolerant" to variations).

Then, all paths in BOCK up to a defined length cutoff are extracted from G_S to G_T. For this work, we set this length cutoff to 3, which ensures full connectivity among known pathogenic gene pairs (Figure 5.15 in Chapter 5). These paths are then weighted based on the traversed edge weights (see next section) and aggregated as metapaths, a sequence of node types and edge types (Figure 6.1). The traversal process is not constrained by the original edge directionality; however, this directionality is encoded within the resulting metapath.



M(D₁)	Metapath
M1	GpGsG
M2	GfPCfGsG
M3	GpGaMFaG

Figure 6.1. . The diagram outlines how to extract metapaths for a given gene pair in BOCK. **(1)** Given a disease-associated gene pair (G_S,G_T), oriented by Residual Variation Intolerance Score (RVIS) all paths in BOCK starting at the gene node G_S and ending at the gene node G_T are collected, up to a certain predetermined path length cutoff. Although this traversal disregard edge directionality, the original direction of the edges is encoded in the recorded paths; Each path is attributed a reliability score (PRS) based on the original edge weight. Paths are then aggregated into their metapaths (*i.e.* path types), which are here represented in their compact representation using abbreviated metaedges (see Chapter 5 - Table 5.1 and Material and Methods - Subsection 3.7.3).

The extracted metapaths take the form:

$$M = G_S \xrightarrow{R_1} E_2 \xrightarrow{R_2} \ldots \xrightarrow{R_{n-1}} G_T \tag{6.1}$$

where $R_i \in \mathcal{RT}$ is a specific edge type (or metaedge) and $E_j \in \mathcal{ET}$ a node type (or metanode) in BOCK (see Material and Methods - Subsection 3.7.3).

A metapath can be seen as a KG query returning a set of KG paths for a given gene pair. We say that a metapath *matches* a given gene pair if applying it to that gene pair returns a non-null set of KG paths.

To succinctly represent metapaths in our results, we use a compact representation based on the abbreviated metaedges from BOCK (see Chapter 5 - Table 5.1). Further details on this representation can be found in Material and Methods - Subsection 3.7.3.

Metapaths as predictive features

Whether a metapath matches or not a gene pair can be interpreted as a binary feature, enabling its analysis in relation to the class label in the training set (refer to Chapter 5 - Section 5.4). Figure 6.2 illustrates all metapaths derived from mining paths of length up to 3 from the pathogenic gene pairs. Metapaths are represented according to their individual pathogenic gene pair coverage (the ratio of pathogenic gene pairs that present the metapath) and its correlation to the pathogenic class (measured by the Matthews correlation coefficient).

The figure reveals an over-representation of longer metapaths, a consequence of the combinatorial effect. Direct relationships (metapaths of length 1) individually cover only a small fraction of pathogenic gene pairs (up to 9% for GpG). Metapaths composed solely of Gene entities (depicted in orange) cover an increasing number of pathogenic gene pairs as their length increases. However, these metapaths do not individually provide a strong predictive signal.

Metapaths incorporating Phenotype and Biological Process entities demonstrate both high predictive power and substantial coverage of pathogenic gene pairs. This is particularly true for direct associations (e.g., $GaPaG$, $GaBPaG$) and related associations based

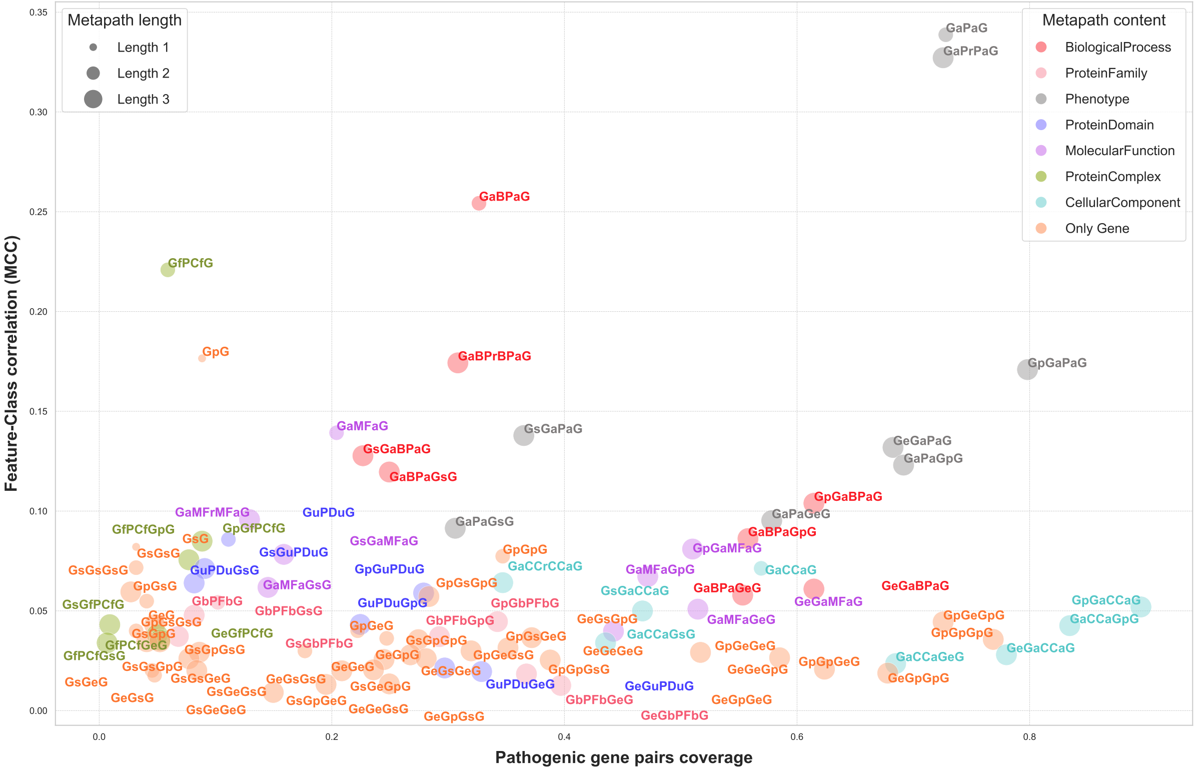


Figure 6.2. Analysis of metapaths based on coverage and predictive power. This visualisation presents an analysis of all metapaths up to a length of 3, based on their presence or absence in selected pathogenic and neutral gene pairs from our training set (refer to Chapter 5 - Section 5.4). Each metapath is treated as a binary feature. The "Pathogenic gene pair coverage" on the x-axis represents the proportion of pathogenic gene pairs that present the metapath. The "Feature-Class correlation (MCC)" on the y-axis measures the Matthews correlation coefficient, indicating the correlation between the presence of the metapath and the pathogenic class. The colour of each point corresponds to the entity types contained within the metapath (excluding the Gene entity), with metapaths containing only the Gene entity coloured in orange. The size of each point is proportional to the length of the metapath it represents.

on the *resembles* edges (e.g., $GaPrPaG$, $GaBPrBPaG$). Heterogeneous metapaths that traverse both a Gene and one of these entities (e.g., $GaGaPaG$, $GaBPaGaG$) offer higher pathogenic gene pair coverage and maintain a relatively good predictive power.

Certain metapaths, such as those involving the $ProteinComplex$ entity, provide limited pathogenic gene pair coverage. Others, like those composed of $CellularComponent$, offer lower predictive power.

While some metapaths may individually exhibit relatively low predictive power, the power can be enhanced when these metapaths are assembled, in a conjunctive manner, into predictive patterns. Moreover, applying these patterns in a complementary or disjunctive manner could lead to a broader coverage of pathogenic gene pairs.

6.2.2 Heterogenous path reliability score

To increase the interpretability and signal strength of our approach, we aim to select the most relevant paths from the mined metapaths. Each mined metapath can be seen as a KG query, which, when applied to a specific entity pair (in our case, a gene pair), yields zero to multiple paths. However, as the metapath length and connectivity between gene pairs in the KG increase, the number of yielded paths can become substantial. Therefore, to aid interpretability, we propose ranking paths based on their reliability, estimated from the quality, strength, or informativeness of the traversed edges. This way, we can filter out less trustworthy paths and focus on the most relevant ones.

Calculating the path reliability score

Scoring and ranking paths in a knowledge graph can be challenging, given their varying lengths and the use of different edge types and node types. To address this issue, we introduce a path reliability score (PRS) to rank paths corresponding to a specific metapath. The PRS (Equation 6.2) is computed as the geometric mean of all the edge weights composing the path.

$$PRS(P) = \left(\prod_{e \in P} w(e) \right)^{1/|P|} \tag{6.2}$$

Since all edge weights have been calibrated to range from 0 to 1, the PRS will also fall within this range, with 1 indicating maximum reliability (all edges have maximum weight) and 0 indicating minimum reliability (at least one edge has zero weight). By using the geometric mean, paths traversing even a single low-weight edge in the graph will be assigned a relatively low PRS, making them less likely to be considered in the analysis.

Filtering paths based on the reliability score

Given a specific metapath, the distribution of PRS can vary across gene pairs, with some gene pairs presenting mainly unreliable paths, some showing mostly reliable paths, and most exhibiting a mix of both (Figure 6.3.A).

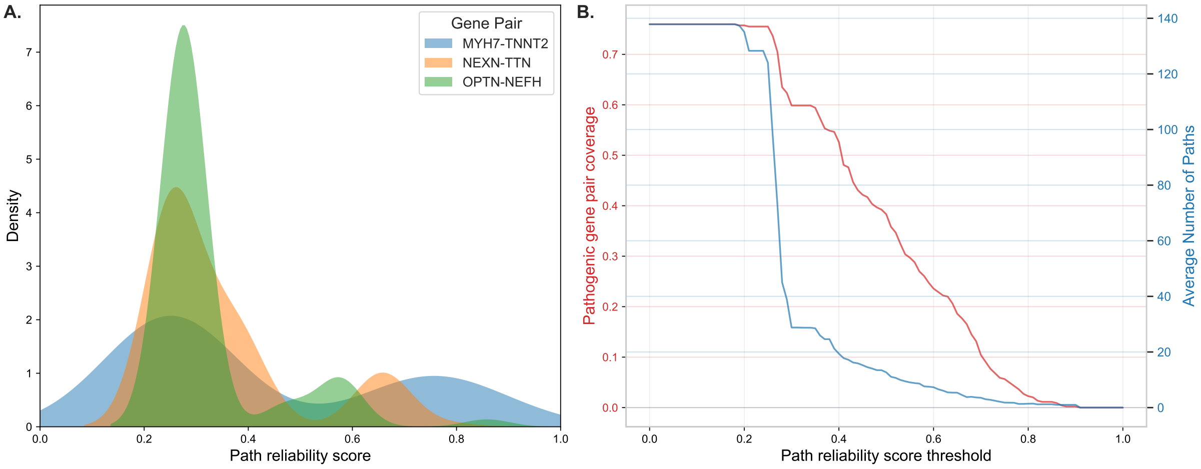
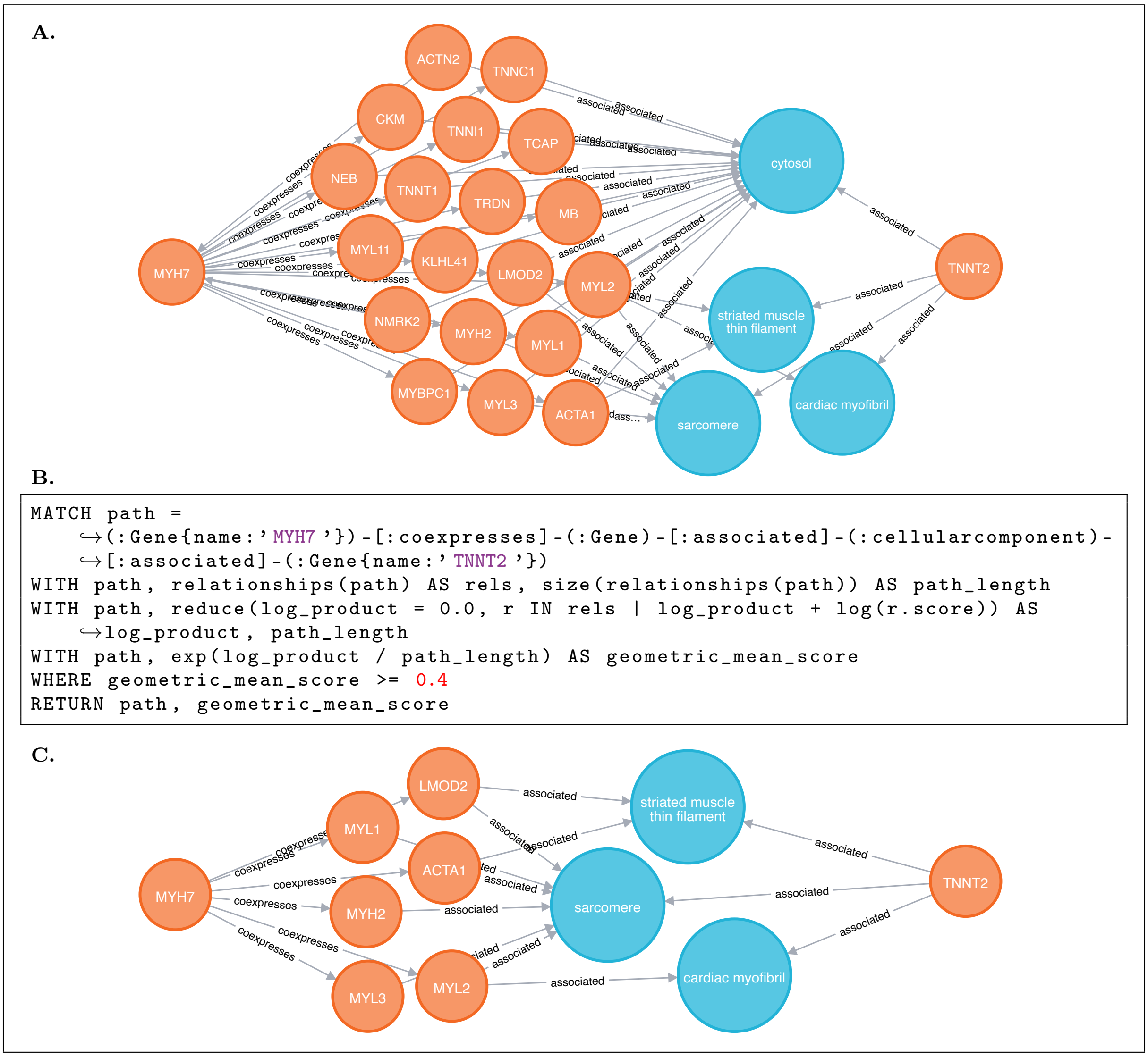


Figure 6.3. Path reliability score distributions and thresholding effect. The figure demonstrates the impact of path reliability score (PRS) thresholding on the assessment of a specific metapath, denoted as *GeGaCCaG*. This metapath represents a type of path where an intermediate gene is coexpressed with the source gene and is associated functionally to the same cellular component as the target gene. To investigate the effect of PRS thresholding, we evaluated the presence of this metapath across all disease-causing gene pairs and retrieved their corresponding path scores. **A.** The graph presents the distribution of path reliability scores for three selected gene pairs, highlighting the variable nature of path scores for the same metapath. **B.** By imposing a minimum PRS threshold, we analyze its impact on two key metrics: *Pathogenic gene pair coverage* (the proportion of pathogenic gene pairs presenting the metapath with at least one path score above the threshold) and the *Average Number of Paths* (the average number of paths passing the threshold for all covered pathogenic gene pairs).

To enhance the interpretability of metapaths (*i.e.* how many KG paths on average are yielded for that metapath) while maintaining sufficient coverage, we can set a minimum path reliability score (PRS) threshold as an additional condition for a metapath to match a gene pair. This conditional threshold selectively retains the most reliable paths, prioritising trustworthy relationships and improving interpretability. However, overly strict PRS thresholding may reduce the coverage of the metapath (see Figure 6.3.B). Striving to strike a balance between interpretability and coverage, we aim to find a minimum PRS threshold that preserves a few highly reliable paths across as many pathogenic gene pairs as possible. For example, as depicted in Figure 6.3.B, a threshold of approximately 0.3 still achieves a coverage of 0.6 (*vs.* 0.76 initially) with an average of 30 paths (*vs.* 138 initially) returned per covered pathogenic gene pair. This approach can be likened to reducing noise while retaining essential signal, ensuring meaningful relationships are effectively captured.

We illustrate, in Figure 6.4, how the metapath previously shown, can be turned into



```
MATCH path =
    ↪(:Gene{name:'MYH7'})-[:coexpresses]-(:Gene)-[:associated]-(:cellularcomponent)-
    ↪[:associated]-(:Gene{name:'TNNT2'})
WITH path, relationships(path) AS rels, size(relationships(path)) AS path_length
WITH path, reduce(log_product = 0.0, r IN rels | log_product + log(r.score)) AS
    ↪log_product, path_length
WITH path, exp(log_product / path_length) AS geometric_mean_score
WHERE geometric_mean_score >= 0.4
RETURN path, geometric_mean_score
```

Figure 6.4. Metapath KG querying with path reliability thresholding. We illustrate the process of applying the path reliability score (PRS) threshold on the metapath *GeGaCCaG* (see Figure 6.3 for more details), applied as a KG query anchored on the gene pair ($MYH7$, $TNNT2$). **A.** Subgraph obtained by querying the metapath without PRS threshold. **B.** KG query, expressed in the Cypher language (Neo4J), applying both the metapath and the PRS threshold (here 0.4). **C.** Subgraph result of the query in B., where paths of lower quality have been filtered (*e.g* all paths going through cytosol, which association edge harbour a relatively small edge score due to its low information content).

a KG query, with a minimum PRS threshold that select a reduced set of reliable paths. For instance, in the example given, all paths traversing the cellular component *Cytosol* are filtered, due to its low information content (*i.e* lack of specificity).

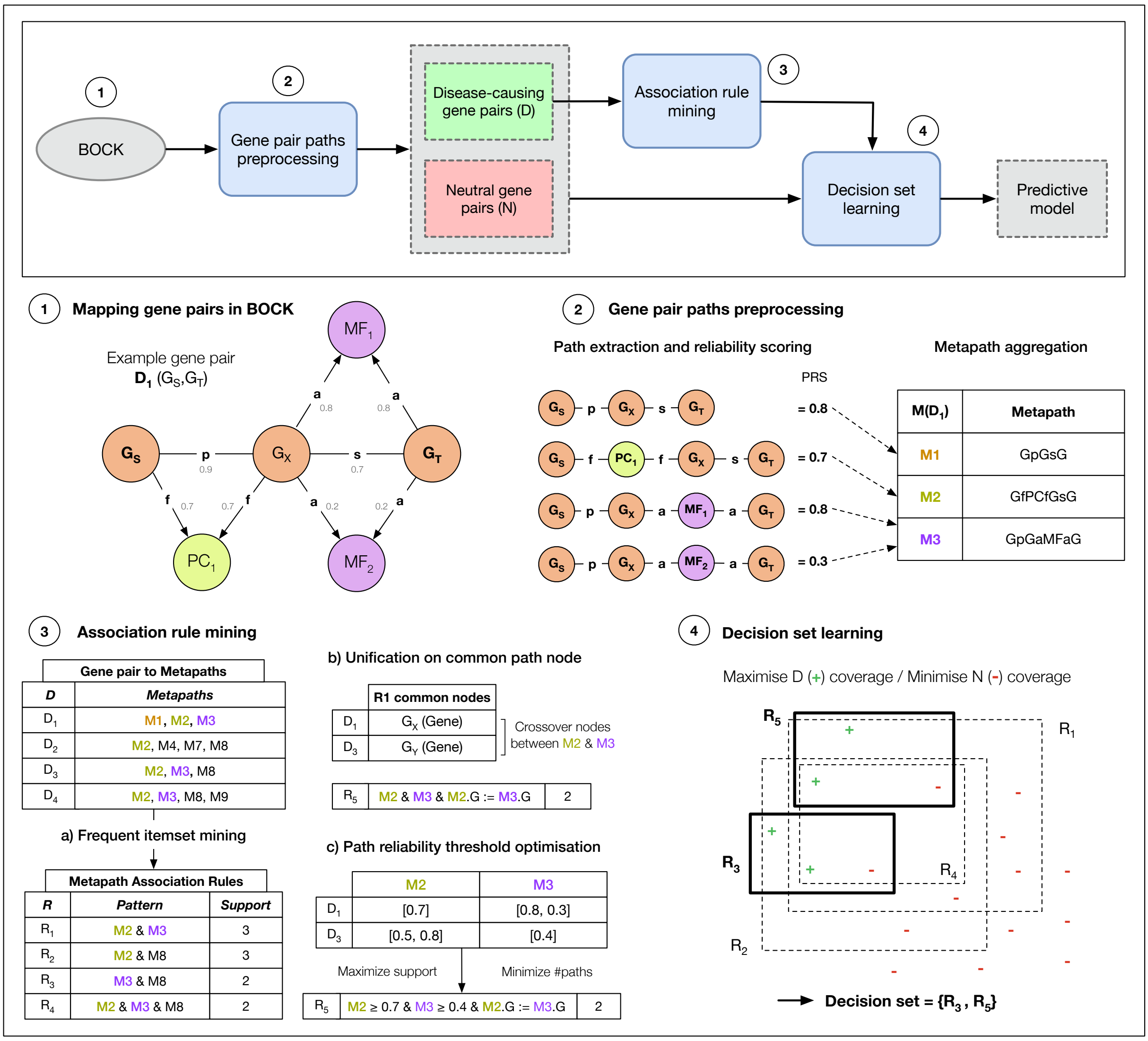
In conclusion, the extraction of metapaths is a powerful approach to automatically generate meaningful features capturing diverse short and long-range interactions between gene pairs within the knowledge graph, BOCK. The introduction of the path reliability score offers a systematic way to filter paths underlying metapaths, facilitating their manual inspection, and thereby enhancing their interpretability.

To achieve our goal of accurately detecting pathogenic gene pairs while ensuring full interpretability, our next section presents a novel methodology that leverages the extracted metapaths as rules and incorporates path score thresholding to build a robust and interpretable classifier.

6.3 A novel approach to leverage metapaths as rules for interpretable link prediction

To address the challenge of accurately detecting pathogenic gene pairs while ensuring interpretability, we introduce a novel methodology called ARBOCK (Associated Rule learning on BOCK). This novel approach harnesses path characteristics (known as metapaths, as discussed in Subsection 6.2.1) that connect possible pathogenic gene pairs within the BOCK framework. Our aim here is to construct a rule-based model that can classify these gene pairs with accuracy while providing explanations in the form of subgraphs in BOCK.

We begin by giving a broad overview of this innovative predictive framework with its key parameters. We then explain the specifics of each stage in the process. Finally, we provide details on the availability of the framework and the reproducibility of the research presented in this chapter.


1
BOCK
2
Gene pair paths preprocessing
Disease-causing gene pairs (D)
Neutral gene pairs (N)
3
Association rule mining
4
Decision set learning
Predictive model
1 Mapping gene pairs in BOCK
Example gene pair
D1 (GS,GT)
GS
p 0.9
GX
s 0.7
GT
a 0.8
a 0.8
MF1
f 0.7
f 0.7
PC1
a 0.2
a 0.2
MF2
2 Gene pair paths preprocessing
Path extraction and reliability scoring
PRS
GS – p – GX – s – GT = 0.8
GS – f – PC1 – f – GX – s – GT = 0.7
GS – p – GX – a – MF1 – a – GT = 0.8
GS – p – GX – a – MF2 – a – GT = 0.3
Metapath aggregation
3 Association rule mining
b) Unification on common path node
Crossover nodes between M2 & M3
a) Frequent itemset mining
c) Path reliability threshold optimisation
Maximize support
Minimize #paths
4 Decision set learning
Maximise D (+) coverage / Minimise N (-) coverage
R5
R1
R3
R4
R2
+
-
Decision set = {R3 , R5}

M(D1)	Metapath
M1	GpGsG
M2	GfPCfGsG
M3	GpGaMFaG

Gene pair to Metapaths

D	*Metapaths*
D_1	M1, M2, M3
D_2	M2, M4, M7, M8
D_3	M2, M3, M8
D_4	M2, M3, M8, M9

Metapath Association Rules

R	*Pattern*	*Support*
R_1	M2 & M3	3
R_2	M2 & M8	3
R_3	M3 & M8	2
R_4	M2 & M3 & M8	2

	R1 common nodes
D_1	G_X (Gene)
D_3	G_Y (Gene)

R_5	M2 & M3 & M2.G := M3.G	2

	M2	M3
D_1	[0.7]	[0.8, 0.3]
D_3	[0.5, 0.8]	[0.4]

R_5	M2 ≥ 0.7 & M3 ≥ 0.4 & M2.G := M3.G	2

Figure 6.5. KG-based associative classifier training workflow. The diagram outlines how our framework uses labelled gene pairs and the path information in BOCK to train a rule-based model predicting pathogenic gene interactions. **(1)** For a given disease-causing gene pair (*e.g.* D_1 (G_S,G_T)), all paths in BOCK, up to a chosen length cutoff, between nodes G_S and G_T are collected, ignoring edge directions. Paths traversing “OligogenicCombination” or “Disease” entities are excluded ; **(2)** Each path is attributed a reliability score based on the original edge weight. Paths are then aggregated into their metapaths (*i.e.* path types) (M) ; **(3.a)** Association rules (R) are mined by finding frequent patterns of metapaths occurring in disease-causing gene pairs (D). Rules are extended with additional metapath conditions as long as their support (*i.e.* the weighted frequency of the pattern) is greater than a defined threshold (*minsup*) ; **(3.b)** Rules can be extended with a unification condition (*e.g.* node G_X common to metapaths $M1$ and $M2$) if such pattern remains frequent ; **(3.c)** The mined rules are refined with path reliability thresholds aiming to filter paths of lower quality while preserving a high rule support ; **(4)** Using all pre-mined rules (R) and training data made of disease-causing gene pairs (D) as positive examples and a set of putative neutral gene pairs (N) as negative examples, a decision set model is trained by selecting a subset of predictive rules.

6.3.1 Framework overview and parameters

Built on associative classifier principles [231, 233, 294], this two-step model starts by generating a rule set from local patterns of the pathogenic gene pairs (Figure 6.5.2 and Figure 6.5.3). The second step combines these rules into a decision set (DS) classifier [235], enabling identification of potential pathogenic gene pairs (Figure 6.5.4). This approach was chosen to balance interpretability and performance.

The model's parameters, summarised in Table 6.1, can be adjusted to manage computational complexity and the volume of discovered patterns. These parameters were empirically determined in this study to optimise predictive performance while minimising explanation complexity (see Subsection 6.6.1). Details on each step of this framework are provided next.

Parameter	Description	Step	Algorithm	Range	Set value
`path_cutoff`	Maximum number of edges traversed for a path	Path extraction	Depth-first search	$[1, \infty[$	3
`minsup_ratio`	Minimum fraction of oligogenic samples to consider a pattern frequent	Rule mining	Weighted apriori	$]0, 1[$	0.2
`max_rule_length`	Maximum number of metapaths in a rule	Rule mining	Weighted apriori	$[2, \infty[$	3
α	Relative importance of positive coverage over negative coverage	Model training	Greedy weighted set cover	$]0, 1[$	0.5

Table 6.1. Framework parameters summary. This table presents the parameters defined in the ARBOCK approach. For each parameter, a description, its relevant step in the process, the algorithm employed in that step, the range of potential values, and the specific value chosen for this study are provided.

6.3.2 Path traversal and confidence scoring

First, the relational information between all selected gene pairs is captured via a path traversal of the oligogenic KG, a path filtering procedure and a subsequent aggregation of paths into path types, also known as metapaths (Figure 6.5.2).

In the initial phase of the method, a traversal is conducted over all potential paths within the oligogenic KG between each identified gene pair. Each path begins from the gene with the lowest Residual Variant Intolerance Score (RVIS) [180] and ends at the gene with the highest score. The traversal process is not constrained by the original edge directionality; however, this directionality is encoded within the resulting metapath.

Finally, the paths are aggregated into metapaths, a sequence of node types and edge types that records the semantic pattern of the relationship. The original path information is recorded for later use in the mining stage.

To be able to compare and rank paths following the same metapath based on the informativeness and quality of relationships, a path reliability score is also attributed to each path by computing the geometric mean of all composing edge scores (see Subsection 6.2.2).

Note that, for the subsequent analyses presented in this research, the `path_cutoff` parameter, limiting the maximum number of edges in a path, was set to 3, which ensures full connectivity among known pathogenic gene pairs (see Chapter 5 - Figure 5.15). We also discarded paths traversing "OligogenicCombination" or "Disease" entities, as these inherently contain the answers to our predictive problem. Additionally, paths traversing inconsistent edge properties were discarded such as paths crossing multiple "coexpresses" edges, with discordant tissue properties.

6.3.3 Association rule mining on paths

After extracting paths and metapaths, frequent associations of metapaths are extracted from the disease-causing gene pairs using a level-wise search based on the Apriori algorithm [218], designed for the efficient exhaustive discovery of frequent patterns over transactional data (Figure 6.5.3.a). These patterns are considered as class association rules (CAR) [231] in the form: *conditional pattern* $\rightarrow$ *class label*, with *disease-causing* (l_D) as the only class label (see Material and Methods - Subsection 3.5.3).

Each mined rule r is associated with a support value based on the covered disease-causing gene pairs $\mathcal{C}_r(D)$. We set the `minsup_ratio` parameter controlling the minimum relative support to consider a rule valid to 0.2 for this study. Both support and `minsup_ratio` were adjusted according to the weight associated with each gene pair in order to give more importance to instances with higher confidence. For this study, the `max_rule_length` parameter was set to 3, limiting the maximum number of metapath conditions in a rule.

We extended the mining of simple metapath associations by searching for unification constraints between metapaths (Figure 6.5.3.b). Unifications are variables expressed in multiple conditions that can be substituted with the same value. This concept has been adapted to metapaths by searching for nodes at the intersection of paths associated with at least two different metapaths. Unified patterns were limited to one unification constraint and these patterns have the same minimum support constraint.

Finally, to minimise the redundancy of mined patterns, we limited the mining to *closed itemsets*, by discarding all patterns where at least one of its superset pattern has the same support count [230] (see Material and Methods - Subsection 3.5.3).

6.3.4 Optimisation of path reliability thresholds

Subsequently, rules are optimised with path reliability thresholds to improve interpretability and limit potentially spurious paths (Figure 6.5.3.c).

The aim of this stage is to refine rules with conditions filtering out paths below a certain path reliability score threshold (Subsection 6.2.2). We implemented this refinement stage by searching, for every metapath composing a rule, a minimum threshold in the interval $[0, 1]$ conditioning the associated paths in the training set to be scored with a value higher or equal to that threshold (see Subsection 6.2.2 for an illustration of this process).

Note that, if high threshold values for a rule are set, fewer paths may be yielded, but the rule support may be decreased or the rule may even be invalidated if the support doesn't meet the minimum support constraint; therefore, this search was implemented to take this tradeoff into account.

A differential evolution algorithm [295, 296] was used for this search, considering the non-linear problem to be solved (see Material and Methods - Subsection 3.5.2). This algorithm works by evolving and combining a population of solutions (here, the optimal threshold values for a rule), retaining the most fit candidate solutions at each generation.

We implemented the fitness of a rule, given a set of thresholds Θ, as defined in Equation 6.4. This fitness function is influenced positively by the rule support ($i.e$ the coverage of disease-causing instances $\mathcal{C}_r(D)$) and negatively by the number of paths returned on matching instances $P_r(d)$ (with $d \in D$) on average (Equation 6.3). A fitness of 0 was returned for thresholds where the rule support was lower than `minsup_ratio`, to enforce the minimum support constraint. Gene pair weights were used to adjust both the rule support and the average number of paths calculation.

$$\bar{\mathrm{P}}_r(D) = \sum_{d \,\in\, D} |P_r(d)| \,/\, |\mathcal{C}_r(D)| \tag{6.3}$$

$$\mathrm{f}\,(r, \Theta) = \frac{1}{2} \cdot \left(\frac{\left|\mathcal{C}_{r|\Theta}(D)\right|}{|\mathcal{C}_r(D)|} + \left(1 - \frac{\bar{P}_{r|\Theta}(D)}{\bar{P}_r(D)}\right)\right) \tag{6.4}$$

This optimisation was performed with the differential evolution DE/best/1/bin scheme [297] shown to be the most accurate and robust strategy, regardless of the characteristics of the problem to be solved. The algorithm was set with the following hyperparameters: a population size of 50, up to 1000 generations, a recombination constant of 0.7 and a mutation constant dithering from 0.5 to 1.

6.3.5 Training of a decision set classifier

After mining all rules from the disease-causing gene pair set, the rules are then applied to the neutral gene pairs (N) to estimate the rule negative coverage $\mathcal{C}_r(N)$ for any rule r.

Using these rules, we trained a decision set (DS) model [235], a type of associative classifier [294] based on a collection of unordered rules interpreted as disjunction (see Material and Methods - Subsection 3.5.4). Training a decision set consists of two phases: first, selecting a representative subset of rules out of an initial rule set and second, estimating the class probabilities associated with each model decision (Figure 6.5.4).

We implemented the first phase with the weighted set cover algorithm, inspired by the RUDIK rule mining method [155]. This greedy heuristic can find a representative subset of rules in reasonable time constraint (see Material and Methods - Subsection 3.5.4). In this approach, a weight is assigned to a candidate rule set (Equation 6.5), with lower weights given to rule sets that have a high coverage of disease-causing gene pairs ($\mathcal{C}_R(D)$) and a low coverage of neutral gene pairs ($\mathcal{C}_R(N)$). We adapted the coverage calculation to take into account instance weights. A parameter $\alpha \in [0,1]$ calibrates the relative importance of the positive coverage or negative coverage.

$$w(R) = \alpha \cdot \left(1 - \frac{|\mathcal{C}_R(D)|}{|D|}\right) + (1-\alpha) \cdot \left(\frac{|\mathcal{C}_R(N)|}{|N|}\right) \tag{6.5}$$

This algorithm also defines a marginal weight w_m (Equation 6.6), that quantifies the weight increase by adding a rule r to the decision set of rules R.

$$w_m(r) = w\left(R \cup \{r\}\right) - w(R) \tag{6.6}$$

The greedy procedure starts with an empty decision set solution R. Then, at each iteration, it picks the rule from the original rule set with the minimum marginal weight and adds it to the solution R. The procedure stops when the marginal weight is greater than or equal to 0.

The selected rules were then used to build a predictive model. We assigned to each rule a probability estimate for the disease-causing class label (l_D), defined in Equation 6.7. This estimate corresponds to the precision of the rule corrected by the class imbalance in the training dataset. Note that probability estimates returned by the model are uncalibrated and mainly serve as the basis for binary classification, according to a classification threshold (see Material and Methods - Subsection 3.5.1).

$$p(l_D|r) = \frac{|\mathcal{C}_r(D)|}{|\mathcal{C}_r(D)| + \frac{|D|}{|N|} \cdot |\mathcal{C}_r(N)|} \tag{6.7}$$

The model decision process was set up according to these criteria, following the same logic proposed by [235]:

1. if a gene pair matches multiple rules in the decision set, the rule with the highest probability estimate is chosen;
2. If a gene pair does not match any of the rules, it is predicted as neutral with a probability estimate based on uncovered training instances.

The trained DS model takes a gene pair and its associated BOCK-paths as input and returns a probability estimate of pathogenicity along with the matched rules, if any.

6.3.6 Availability of the framework and reproducibility

The ARBOCK framework is provided as an open-source software at:

https://github.com/oligogenic/ARBOCK. It includes all scripts necessary to make predictions and generate explanations, as well as scripts to train and evaluate new models. We provide a Python notebook to reproduce the presented plots and tables. The predictive models presented in this study and the predictive result files are available in the same repository, under the `models/` folder.

6.3.7 Runtimes

The current implementation, tested on an Intel Core i7, is capable of retrieving path data from BOCK at an average speed of 1.16 seconds (SD = 2.2) for each gene pair (using `path_cutoff`=3), and produces predictions along with corresponding explanations at an average rate of 0.5 milliseconds (SD = 2.9) for each gene pair. Excluding the path retrieval, the training of a new model (rule mining + decision set learning), using the parameters and the full data presented in this study, took an average of 29.26 mn parallelised on 8 cores.

6.3.8 Preliminaries on the presented analyses

In the following presented analyses, we applied our novel predictive framework, ARBOCK, on two labelled training datasets: a positive set (D) of 441 disease-causing gene pairs with established familial or statistical evidence of pathogenicity, and a negative set (N) of 44,100 putative neutral gene pairs, selected from a cohort of healthy individuals. From the positive set, we also held a test set of 15 pathogenic gene pairs from the recent literature with reliable clinical evidence.

Each gene pair is given a weight signifying the confidence in its label. Detailed process for these sets' creation is described in Chapter 5 - Selection and weighting of gene combinations.

6.4 Frequent graph patterns associated with pathogenic gene pairs

Considering only the rule mining part of our framework and the previously defined parameters (see Table 6.1), we now explore common patterns occurring in the 426 pathogenic gene pairs from the training dataset. These patterns, represented as rules, are integral to the following predictive modelling stage. Each rule is formulated as a set of conditions, involving multiple metapaths, associated to a class label (here, the disease-causing label l_D).

Upon considering all valid paths (*incl. Pheno*), 6917 rules were mined. Conversely, excluding paths traversing Phenotype nodes (*excl. Pheno*) resulted in 4124 rules.

6.4.1 Association rule metrics

We decided to explore rules based on their coverage over the disease-causing (D) and neutral (N) sets. We used two relevance measures commonly used in data mining, the rule support (*i.e.* the ratio of disease-causing gene-pairs associated with the conditional clause) and the rule confidence (*i.e.* the likelihood of a gene-pair being pathogenic when it matches the conditional clause). The confidence distribution of the rules, in relation to their support is presented in Figure 6.6.

The individual rules extracted from the oligogenic cases in BOCK range from a support of 0.2 (*i.e.* the minimal reachable bound) to 0.62, with a distribution skewed towards low support, owing to the high number of complex and more specific patterns mined. The majority of extracted rules individually exhibit a confidence in the range $[0.56, 0.98]$, with only 204/4124 rules with a confidence higher than 0.9. Integrating both distributions shows that the most interesting rules (with a confidence higher than 0.9) individually only cover up to 34% of pathogenic cases. Therefore, a combination of multiple rules is necessary to cover the entire spectrum of known oligogenic interactions.

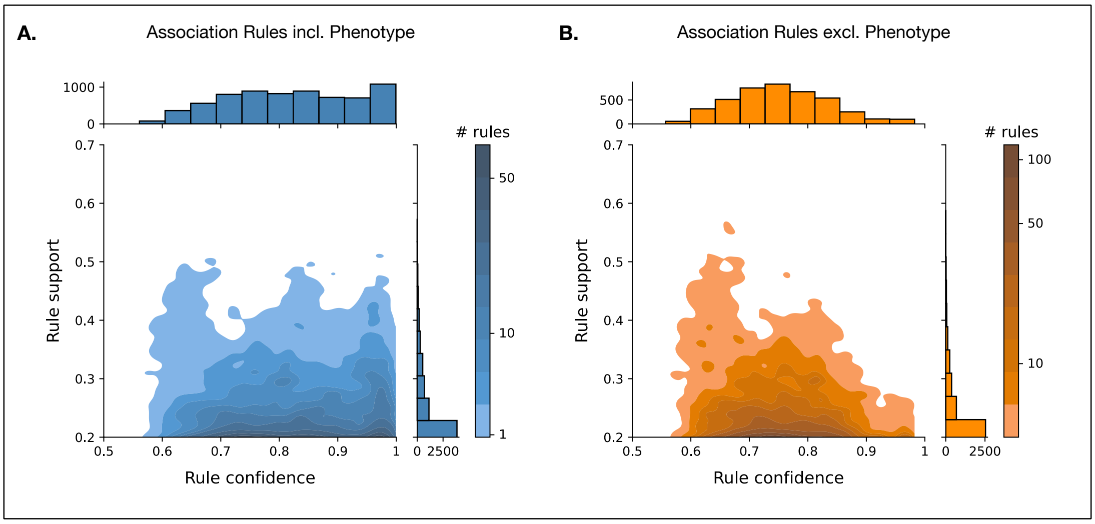


Figure 6.6. Distribution of all candidate association rules obtained by mining all disease-causing gene pairs in the training set, according to two metrics: the rule support (*i.e.* the ratio of pathogenic gene pairs associated with the rule pattern) and the rule confidence (*i.e.* the likelihood of a gene pair being pathogenic when it matches the rule pattern). Note that, due to the high imbalance in the dataset, the rule confidence metric was adapted to rebalance the negative coverage. The darkness of the shade increases logarithmically with the number of rules. **A.** Rules mined from paths including Phenotype (*incl. Pheno*). **B.** Rules mined from paths excluding Phenotype (*excl. Pheno*).

6.4.2 Rule metapath and metaedge content

To better understand the relationship between the rule metapath content and its predictive power, we analysed metapaths significantly associated with higher rule confidence, enabling us to shed light on the most influential types of relationships (Figure 6.7).

Analysing the rules' conditions, 16 metapaths emerged as significantly associated with higher confidence rules (Figure 6.7.A). High-confidence rules often include metapaths related to similar phenotypes (50%), biological processes (31.2%), and molecular functions (12.5%), and involve intermediate genes (75%) linked with a diverse range of relationships. Metapaths containing phenotype information, specifically *GaPaG* and *GaPrPaG* (reflecting common and related phenotypes between gene pairs), hold the most influence among high-confidence rules.

Analysis of the rules' conditions revealed 16 metapaths significantly associated with higher confidence rules, of which 8 were unique to this setting (Figure 6.7.B). High-confidence rules predominantly present heterogenous metapaths associated with a wider range of entities including biological processes (43.8%), molecular functions (18.8%), cellu-

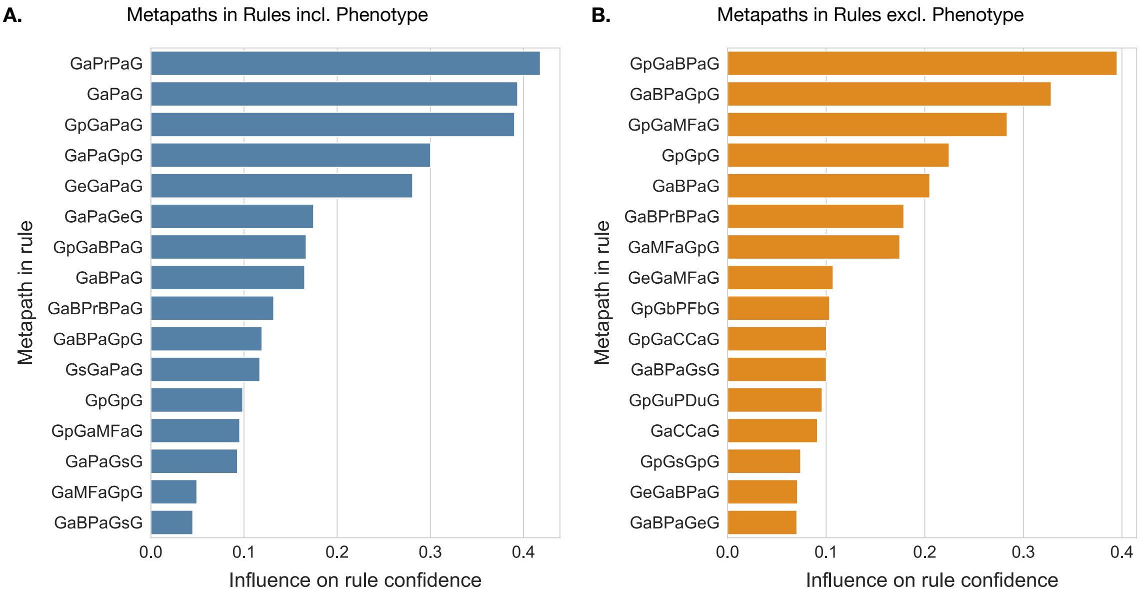


Figure 6.7. Metapath influence on rule confidence Metapaths significantly associated with higher rule confidence are shown. Rule confidence distributions are compared for each metapath, considering both its presence and absence as a rule condition. The significance of the difference is determined using a one-tailed Wilcoxon ranksum test with Bonferroni correction (adjusted p-value ≤ 0.01). Metapaths are ranked based on their effect size, measured by the rank-biserial correlation. **A.** Rules mined from paths including Phenotype (*incl. Pheno*). **B.** Rules mined from paths excluding Phenotype (*excl. Pheno*).

lar components (12.5%), protein families (6.2%), and protein domains (6.2%). Metapaths related to biological processes and molecular functions are most influential. Particularly, $GpGaBPaG$ and its reverse, which capture shared processes between a gene pair and an interacting gene, stand out in the highest confidence rules.

6.5 Decision set model predictive performance

In this section, we evaluate the full decision set models produced by the ARBOCK framework. We first evaluate the predictive performance of these models. We then explore the effect of including Phenotype paths and generalise these observations to the DiGePred model [85].

6.5.1 Predictive performance

Using our new approach with the parameters described previously (see Table 6.1), we trained decision set (DS) models based on 426 pathogenic gene-pairs and 42,600 neutral gene pairs (imbalance ratio of 1:100), holding-out 15 recently published and high-quality pathogenic gene pairs for independent testing (see Chapter 5 - Section 5.4).

We analysed two DS models: *DS incl.Pheno* (including Phenotype-traversing paths) and *DS excl. Pheno* (excluding Phenotype-traversing paths). Both were evaluated using a stratified 10-fold cross-validation strategy, enabling a robust evaluation of predictive performance (see Figure 6.8 for an overview of the model evaluation process).

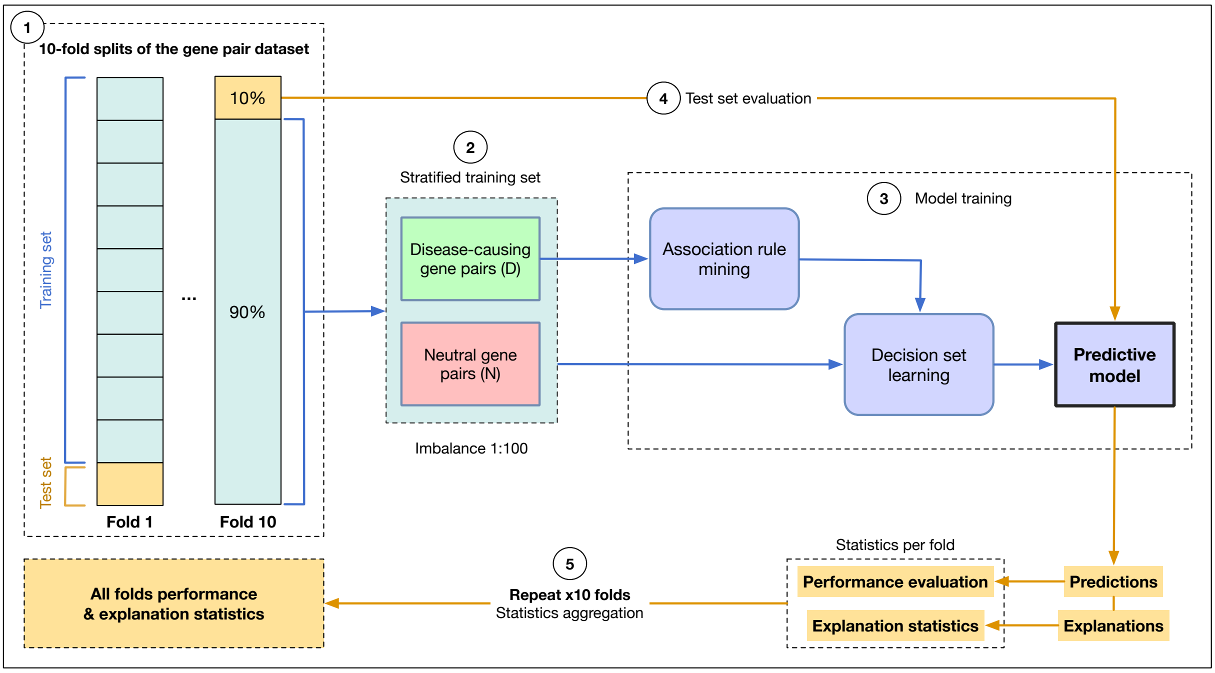


Figure 6.8. ARBOCK evaluation by cross-validation. Diagram summarising the evaluation of the ARBOCK models. After holding-out an independent test set, the remaining gene pairs comprised of 426 disease-causing gene pairs (D) and 42,600 neutral gene pairs (N) (and their associated paths from BOCK) are used to evaluate the predictive performance of the model using a stratified 10-fold cross-validation strategy. **(1)** First, the dataset is splitted into 10 subsamples which are used to create 10 cross-validation folds. Each fold is segmented into a training set (90% of the fold) and a test set (10%). **(2)** The subsamples are sampled in a stratified way, conserving the 1:100 imbalance between the D and the N class. **(3)** The model is trained using the training data of a given fold, according to the procedure described in Section 6.3. **(4)** The test set of the given fold is evaluated against the trained model to obtain predictions and explanations enabling the computation of predictive performance and explanation statistics (*e.g* number of rules, number of paths). **(5)** The procedure is repeated for each fold and all statistics are aggregated for further analyses.

Based on the cross-validation results, we defined optimal classification thresholds for each model based on the ROC curve. The final trained models were tested on the held-out independent test set. We reported predictive performance measured with the Area Under the Receiver Operating Characteristic (AUROC) and the Area Under the Precision-Recall

Curve (AUPRC) (see Material and Methods - Subsection 3.5.1).

We summarised essential information to assess this machine learning approach following the DOME recommendations [298] in Table A.3.

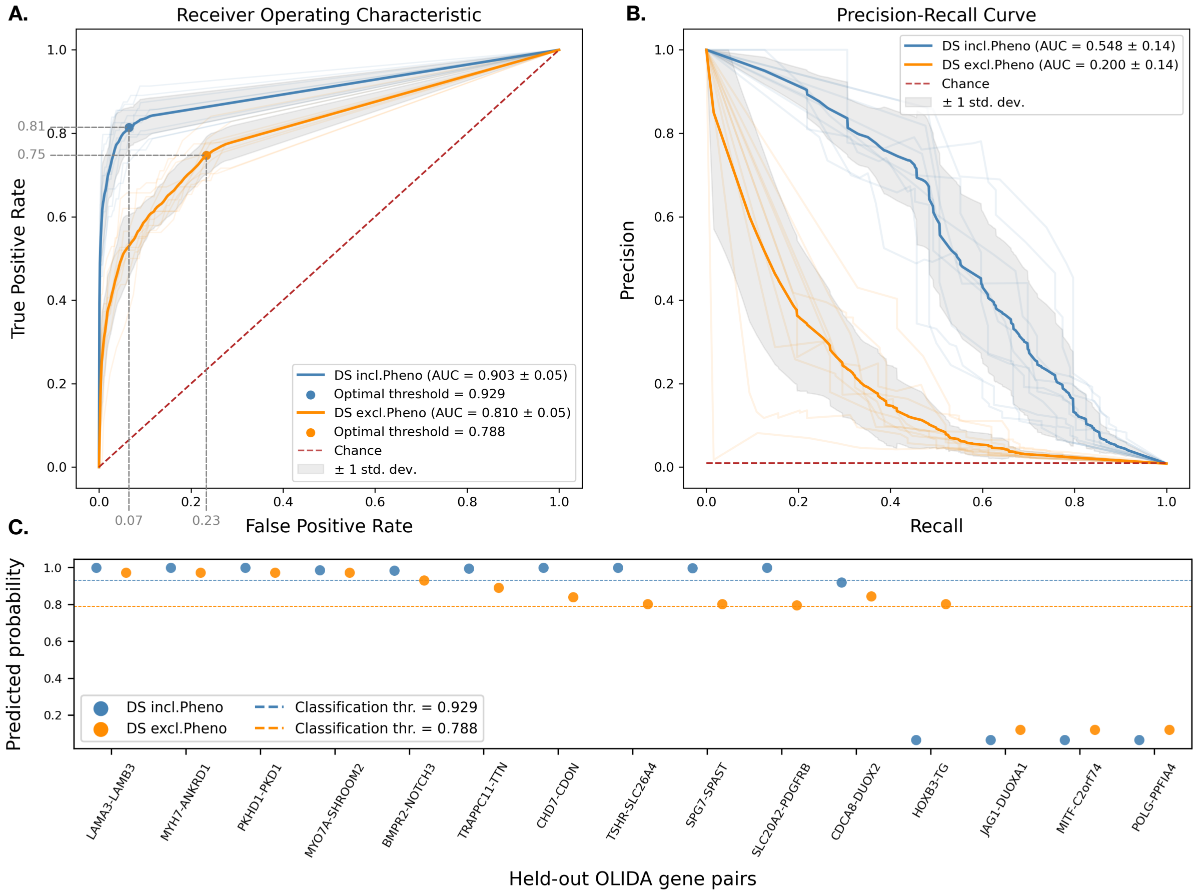


Figure 6.9. Decision set models performance. Two DS models have been evaluated: one trained with all valid paths (*DS incl.Pheno*) and one trained without Phenotype-traversing paths (*DS excl. Pheno*), on a stratified 10-fold cross-validation setting. A test set of 15 pathogenic gene pairs from recent literature has been held out for independent evaluation. (A) Receiver operating characteristic (ROC) curve obtained by averaging all fold curves. The best classification threshold is evaluated using the geometric mean between the sensitivity and the specificity. (B) Precision-Recall curve obtained by averaging all fold curves. (C) Test set predicted probabilities, displayed in decreasing order. The horizontal lines represent the optimal thresholds for binary classification.

The performance of these models is depicted in Figure 6.9. The *DS incl.Pheno* model achieves an AUROC of 0.903 (SD = 0.03) and AUPRC of 0.548 (SD = 0.07), recalling 81.8% of pathogenic gene pairs with a 6.6% false positive rate at the optimal classification threshold. The final model, consisting of 35 rules, successfully predicts 10/15 held-out pathogenic gene pairs.

In comparison, the *DS excl.Pheno* model achieves an AUROC of 0.810 (SD = 0.03) and an AUPRC of 0.200 (SD = 0.07), recalling 75.9% of pathogenic gene pairs with a 24.1% false positive rate at the optimal classification threshold. The fully trained model, consisting of 27 rules, identifies 12/15 held-out pathogenic gene pairs.

6.5.2 Annotation bias and the use of phenotype information

Comparison of decision set models

While the model incorporating phenotype information delivers promising results, its limitations must be considered. Specifically, the disproportionate Gene-Phenotype annotation coverage in disease-associated genes could introduce bias in machine-learning models trained on this dataset (see Chapter 5 - Section 5.5.1). This suggests that models trained with phenotype association features could primarily make decisions based on these features and thus, disproportionately identifying gene pairs from the limited 23.5% pool of phenotype-annotated genes.

The phenotype-inclusive model (*DS incl. Phenotype*) is highly influenced by phenotype annotation. However, this model can nonetheless capture indirect Phenotype relationships due to the metapath-based design of its rules. This property enables the coverage of a wider pool of genes than methods relying on direct Phenotypic associations only (see example of the gene pair *MYO7A-SHROOM2* in Table 6.2).

Comparison with the DiGePred model

We examined the Digenic Gene Predictor (DiGePred) [85], a statistical machine-learning method that reports high-accuracy (reported average AUROC of 0.972 in cross-validation [85]) in predicting pathogenic gene pairs. This model assigns 44% of feature importance to phenotype-based characteristics, indicating a strong reliance on such features.

We evaluated this model on our independent test set comprised of recently published instances (*i.e* potentially less affected by such knowledge bias). From the results presented in Table 6.2, we can note that DiGePred was able to correctly identify 4/15 pathogenic gene pairs. All four correctly predicted instances were fully annotated with phenotype terms. The remaining non-predicted gene pairs often lacked complete phenotype annotation (6/11) or had limited common phenotype terms (Jaccard Index between 0.01 and 0.11), making them harder to identify by this model, which uses features such as the Phenotype Jaccard index and the total number of phenotypes associated for each gene.

Gene pair (A,B)	**#Pheno. A**	**#Pheno. B**	**Phenotype Similarity**	**DiGePred**	**DS incl.Pheno**	**DS excl.Pheno**
HOXB3-TG	0	32	0.00	0.00	0.06	0.80
CDCA8-DUOX2	0	38	0.00	0.01	0.92	0.84
MYO7A-SHROOM2	34	0	0.00	0.02	0.98	0.97
MITF-C2orf74	58	0	0.00	0.02	0.06	0.12
JAG1-DUOXA1	96	0	0.00	0.01	0.06	0.12
POLG-PPFIA4	245	0	0.00	0.03	0.06	0.12
BMPR2-NOTCH3	19	133	0.01	0.05	0.98	0.93
MYH7-ANKRD1	171	8	0.05	0.23	1.00	0.97
PKHD1-PKD1	75	27	0.10	0.03	1.00	0.97
CHD7-CDON	217	116	0.10	0.38	1.00	0.84
TRAPPC11-TTN	73	114	0.11	0.26	0.99	0.89
SPG7-SPAST	53	34	0.21	0.66	1.00	0.80
TSHR-SLC26A4	60	35	0.28	0.88	1.00	0.80
SLC20A2-PDGFRB	35	108	0.32	0.72	1.00	0.80
LAMA3-LAMB3	71	74	0.84	0.90	1.00	0.97

Table 6.2. Independent test set predicted probabilities in relation with phenotype. Table presenting the 15 pathogenic gene pairs held-out as independent test set. For each gene pair, the number of phenotype terms (HPO) linked to to the first and second gene is shown, as well as the Jaccard similarity index between HPO terms of the two genes (Phenotype Similarity). The predicted probabilities are shown for the three benchmarked models: "DiGePred" [85] (a gene pair pathogenicity predictor with features at the gene and relationship level, including strong phenotype-based features), "DS incl.Pheno" (the decision set trained with inclusion of paths traversing nodes of type "Phenotype") and "DS excl.Pheno" (the decision set trained without phenotype-based paths). Green and red colors represent the predicted classification labels, respectively "disease-causing" and "neutral", based on the classification optimal thresholds of these models (DiGePred=0.496 ; DS incl.Pheno=0.929 ; DS excl.Pheno=0.788). Rows are ordered by the Phenotype similarity. The influence of phenotype annotations over gene pairs can be observed for the "DiGePred" and the "DS incl.Phenotype" models, both reliant on phenotype-related features.

6.6 Model assessment in varying settings

In this section, we examine how our model performs under varying conditions. We first explore the impact of ARBOCK framework's parameters on performance and interpretability. Subsequently, we assess the influence of excluding metaedges from the biological knowledge graph, BOCK. Through this investigation, we aim to determine the optimal set of parameters and assess the importance of each knowledge graph component.

6.6.1 Framework parameter tuning

We present here in more details the parameters used in the ARBOCK framework (Table 6.1) and how their choice was empirically determined. In particular, we analyse the effect of the `alpha`, `minsup_ratio` and `max_rule_length` parameters on the model's performance, measured by the Area Under the Receiver Operating Characteristic (AUROC), and explanation complexity, measured mainly by the number of paths on average present in the explanation subgraph, for True Positive (TP) instances. Note that for this benchmark, the choice of AUROC as performance metric stems from its ability to meaningfully evaluate and compare models irrespective of class imbalance and threshold settings, without being artificially inflated by models that merely predict the majority class (*e.g* sets of parameters leading to a model with only the default rule), considering our high class imbalance setting.

Influence of decision set alpha parameter

Figure 6.10 present a comprehensive analysis of the best performing models (considering all other parameters) according to the alpha parameter of the decision set model. At α=0.5, performance is maximised and higher values lead to more explanation complexity, measured as the number of matching rules and explanation path count. This model corresponds to `path_cutoff`=3 ; `minsup_ratio`=0.2 ; `max_rule_length`=3.

Influence of rule mining parameters

Figure 6.11 shows the combined effect of `minsup_ratio` and `max_rule_length`, for fixed α=0.5 and `path_cutoff`=3, on performance and explanation path count. Higher `max_rule_length` leads to a similar performance but at the cost of increasing explanation complexity. Lower `minsup_ratio` leads to better performance and a smaller number of paths in explanations.

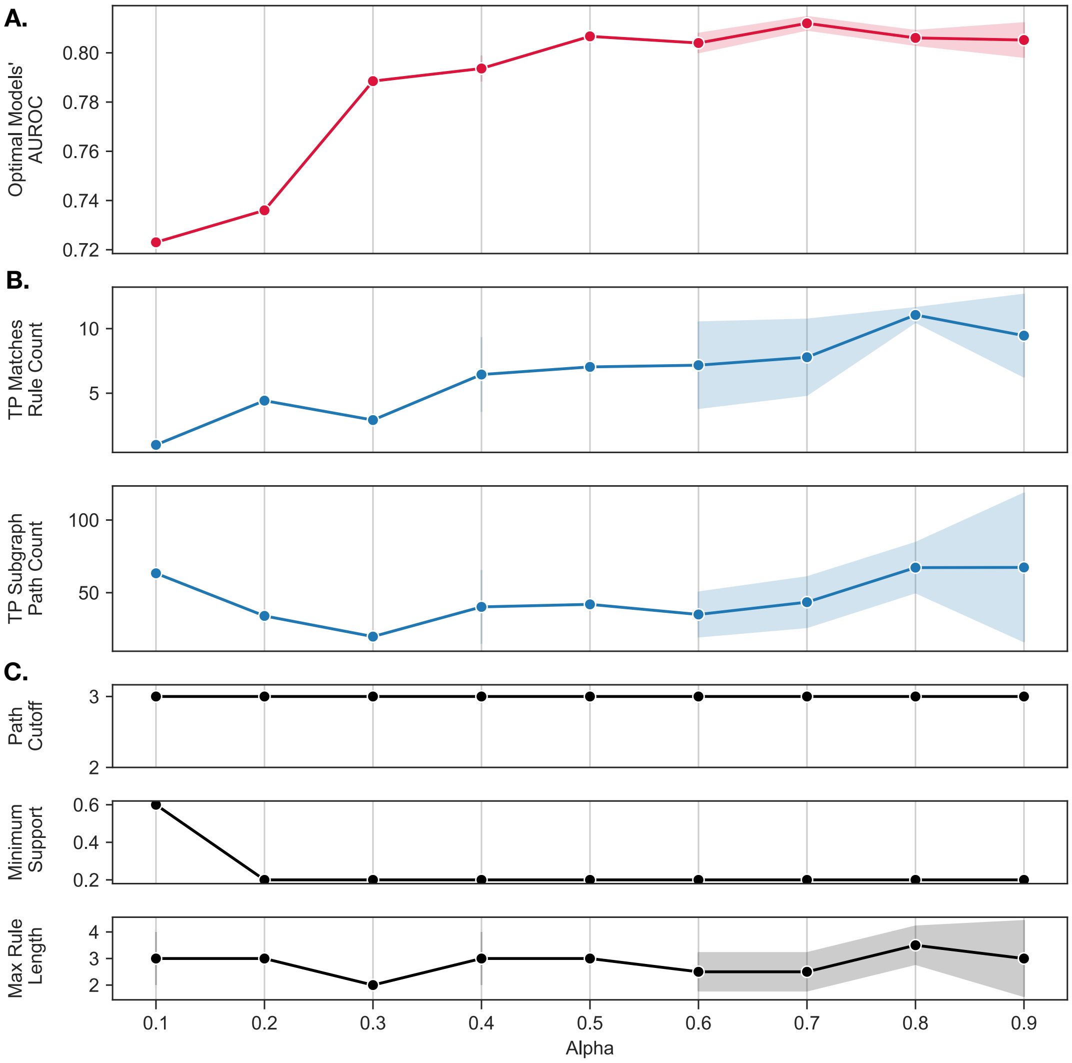


Figure 6.10. Influence of decision model's alpha parameter tuning Evaluating decision set models over a spectrum of the decision model's alpha (α) values (0.1-0.9) and variable parameters - Path Cutoff (`path_cutoff`) {2,3}, Minimum Support (`minsup_ratio`) {0.2, 0.4, 0.6, 0.8}, Max Rule Length (`max_rule_length`) {2,3,4}, we highlight the performance and characteristics of the optimal models for each alpha value. This evaluation is performed on the model excluding Phenotype, prone to study bias. Optimal models for each alpha are selected based on mean AUROC from a stratified cross-validation, keeping all within 0.01 of the highest performer to ensure comparable accuracy levels. **(A)** Mean AUROC of top-performing models evaluated via stratified cross-validation on test sets **(B)** Measures of explanation complexity for the associated models, calculated from predicted true positive (TP) gene pairs on the test sets: Rule Count measures the mean number of rules per match, and Path Count measures the mean number of paths across all rules, on average per match. **(C)** Parameter distribution for the corresponding top-performing models.

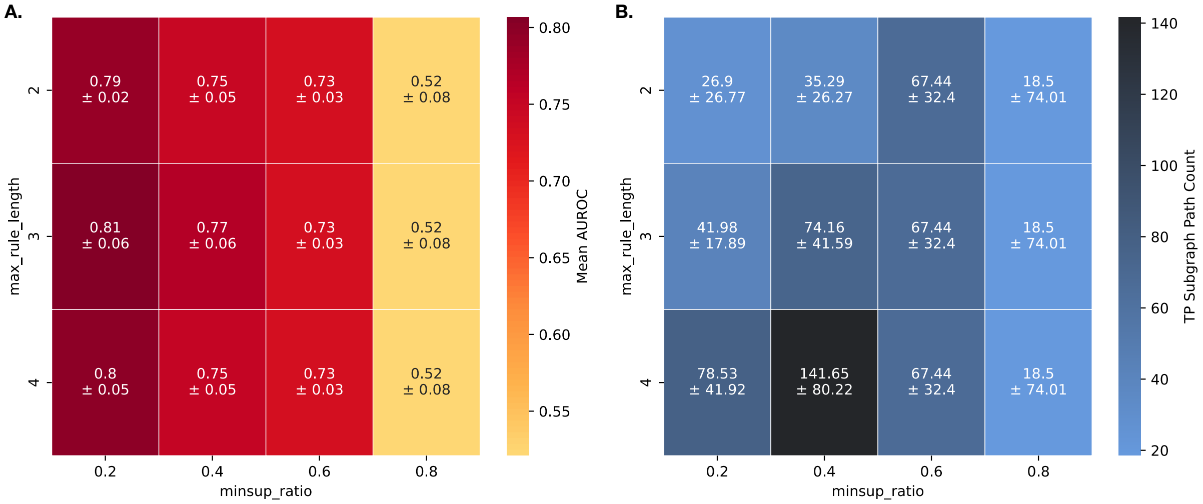


Figure 6.11. Influence of minimum support and max rule length parameters tuning. Heatmaps showing the influence of parameters `minsup_ratio` and `max_rule_length` on the performance and explanation complexity for decision models trained with paths excluding Phenotype and with fixed parameters: $\alpha = 0.5$ and `path_cutoff` $= 3$.
(A) Effect on the model performance, evaluated with mean AUROC on a stratified cross-validation. **(B)** Effect on the model explanation complexity, evaluated with the Path Count measure, measuring the mean number of paths across all rules, on average per True Positive matches.

Fixing `path_cutoff`$=2$, as presented in Appendix Figure A.2 and Figure A.3, leads to a maximum performance of 0.67 for all evaluated sets of parameters but yields simpler explanations in terms of path count.

6.6.2 Knowledge graph metaedge ablation study

The biological knowledge graph, BOCK, contains a multitude of biological relationships represented as different types of edges, also referred to as metaedges (as summarised in Table 5.1). In this experiment, we seek to understand the contribution of each metaedge in both the predictive performance and the complexity of the explanations provided by decision set models, by comparing models trained with different versions of BOCK omitting one metaedge.

We first assessed the baseline performance and complexity on a 10-fold stratified cross-validation, similarly as done when evaluating *DS excl. Pheno*, described in Results (excluding Phenotype-based metaedges, *i.e GaP* and *PrP* metaedges). Using this baseline setting as reference, we evaluate, in a similar manner, decision models based on BOCK excluding an additional metaedge, from the list of remaining metaedges taken into consideration. For each metaedge-excluded model, we report metrics of predictive performance (Figure 6.12) and explanation complexity (Figure 6.13), based on the 10 test folds.

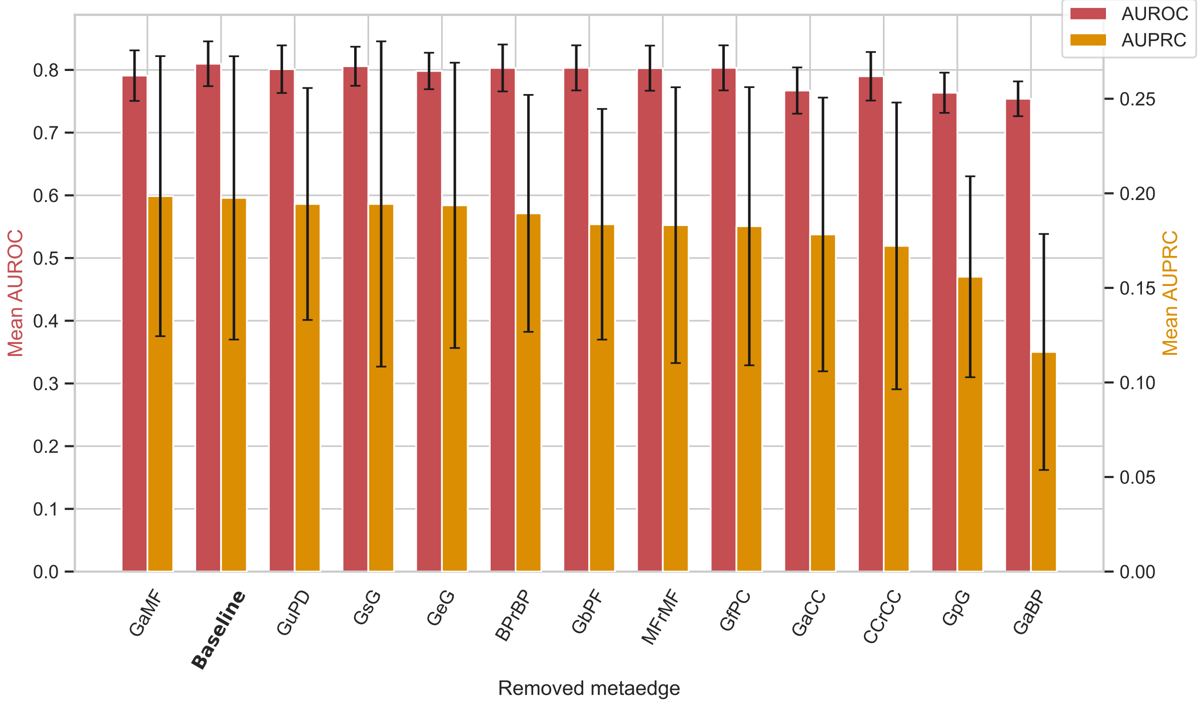


Figure 6.12. Influence of BOCK metaedge ablation on performance. Considering a decision set model, its predictive performance is estimated by the Mean Area Under the Receiver Operating Characteristic Curve (AUROC, red) and the Mean Area Under the Precision-Recall Curve (AUPRC, orange). These metrics are evaluated on a 10-fold stratified cross-validation. The 'Baseline' bar (bold) represents the distribution of this metric for a model excluding Phenotype relationship (*GaP* metaedge). Based on that baseline, all other bars represent models where an additional metaedge (type of edge, as summarised in Table 5.1) has been removed from BOCK. Error bars indicate the standard deviation of this metric for all evaluated test folds.

From this experiment, we can observe that eliminating physical interaction edges (*GpG*) and associations with biological processes (*GaBP*) from BOCK adversely affects classifier performance. Removing other metaedges has a negligible effect on predictive performance.

We measure explanation complexity as the average number of paths obtained for explanations of True Positive gene pairs. A lower number of path leads to explanations that are easier to interpret. Interestingly, omitting gene functional relationships results in a higher average number of paths yielded in explanations, possibly due to an increased prevalence of gene-gene relationships in rules. The removal of the coexpression relationship

GeG reduces the number of paths in the provided explanations, likely owing to the high frequency of *GeG* edges in BOCK.

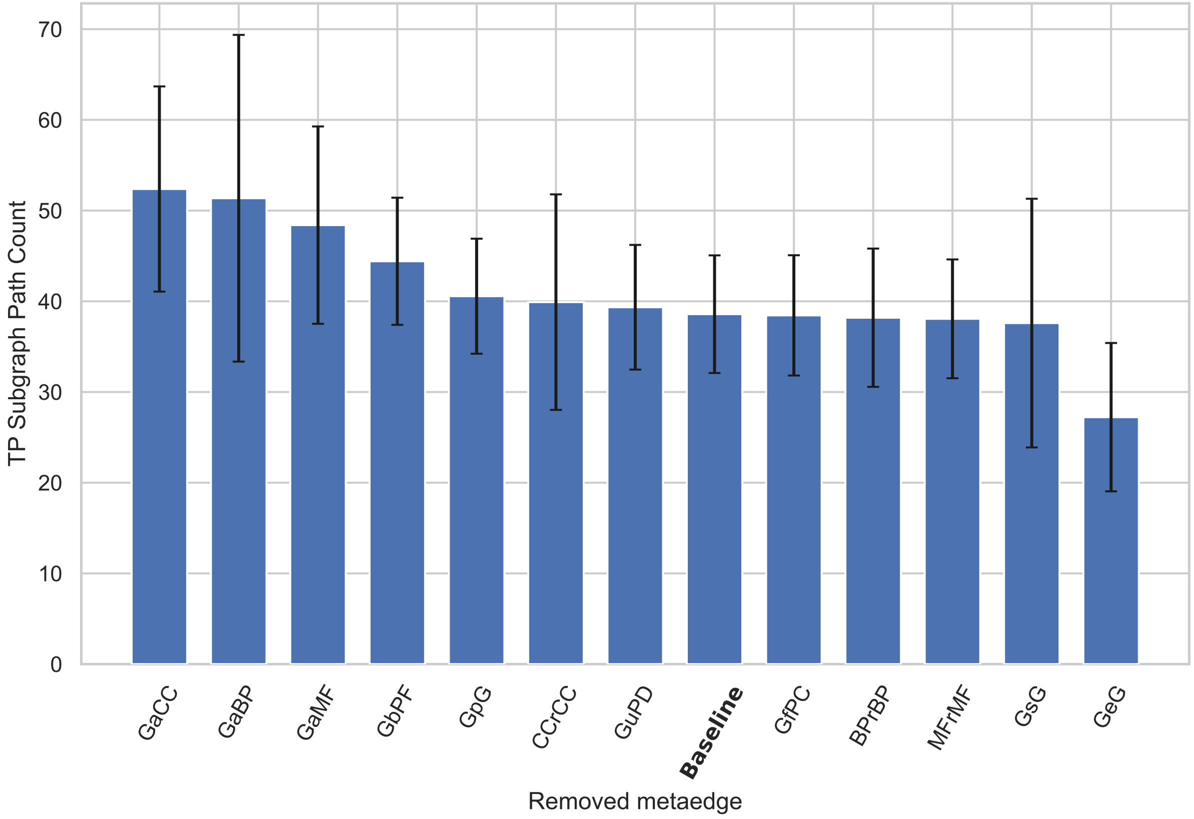


Figure 6.13. Influence of BOCK metaedge ablation on explanation complexity. Considering a decision set model, the explanation complexity is estimated by the mean count of paths of explanation subgraphs returned for the True Positive (TP) gene pairs. This metric is evaluated on a 10-fold stratified cross-validation. The 'Baseline' bar (bold) represents the distribution of this metric for a model excluding Phenotype relationship (*GaP* metaedge). Based on that baseline, all other bars represent models where an additional metaedge (type of edge, as summarised in Table 5.1) has been removed from BOCK. Error bars indicate the standard deviation of this metric for all evaluated test folds.

6.7 Explaining predictions: from rules to paths

The predictive model presented in this work offers both global interpretability and context-specific explanations for pathogenic gene pairs. On the one hand, the simplicity of this model model allows users to examine all the rules that contribute to a pathogenic prediction. On the second hand, it provides transparent predictions by returning the matching rules associated with each predicted pathogenic gene pair.

6.7.1 Inspecting the predictive rules

The rules for our predictive models, *DS incl. Phenotype* and *DS excl. Phenotype*, are listed in Appendices Table A.5 and Table A.4 respectively. These rules can be understood through their metapaths, unification conditions, and path reliability thresholds. Additionally, their impact on the disease-causing and neutral datasets is provided, highlighting the gene pairs each rule covers from the training data.

Interestingly, a significant portion of the rules in our models have unification conditions (30/35 in *DS incl. Phenotype* and 23/27 in *DS excl. Phenotype*), providing more complexity and therefore more specific patterns to detect pathogenic gene pairs. For the *DS excl. Phenotype* model, some rules exhibit high negative coverage. Their inclusion by the model was certainly due to their ability to capture more pathogenic instances that might otherwise be overlooked, indicating potential avenues for refining and uncovering more precise patterns.

6.7.2 Explaining predictions with KG subgraphs

From rules to KG subgraphs

Every rule produced by ARBOCK can be transformed into a KG query on BOCK. Hence, for any gene pair predicted as pathogenic by an ARBOCK model, and based on the corresponding matched rules, we can extract specific subgraphs from BOCK. These subgraphs act as contextual explanations.

These explanatory subgraphs detail specific relationships and entities related to the gene pair in question. Such insights can assist geneticists in evaluating the predictions made based on their domain knowledge and can highlight potential molecular mechanisms involved in the targeted disease.

To produce these contextual explanations, we begin by identifying the matched rules associated with a positively predicted gene pair. These rules are then used to query the KG. The resulting paths are assembled into a coherent set of nodes and edges, forming the desired subgraph.

Our model is able to save these explanations for each prediction, in the Graph Markup

Language (GraphML) format [240], which can be in turn explored via graph visualisation software such as Cytoscape [299]. Additionally, we demonstrated in Subsection 6.2.1 that all metapaths composing a rule along with their path reliability score thresholds could be easily translated into a KG query language such as Cypher (see Figure 6.4 for a query example).

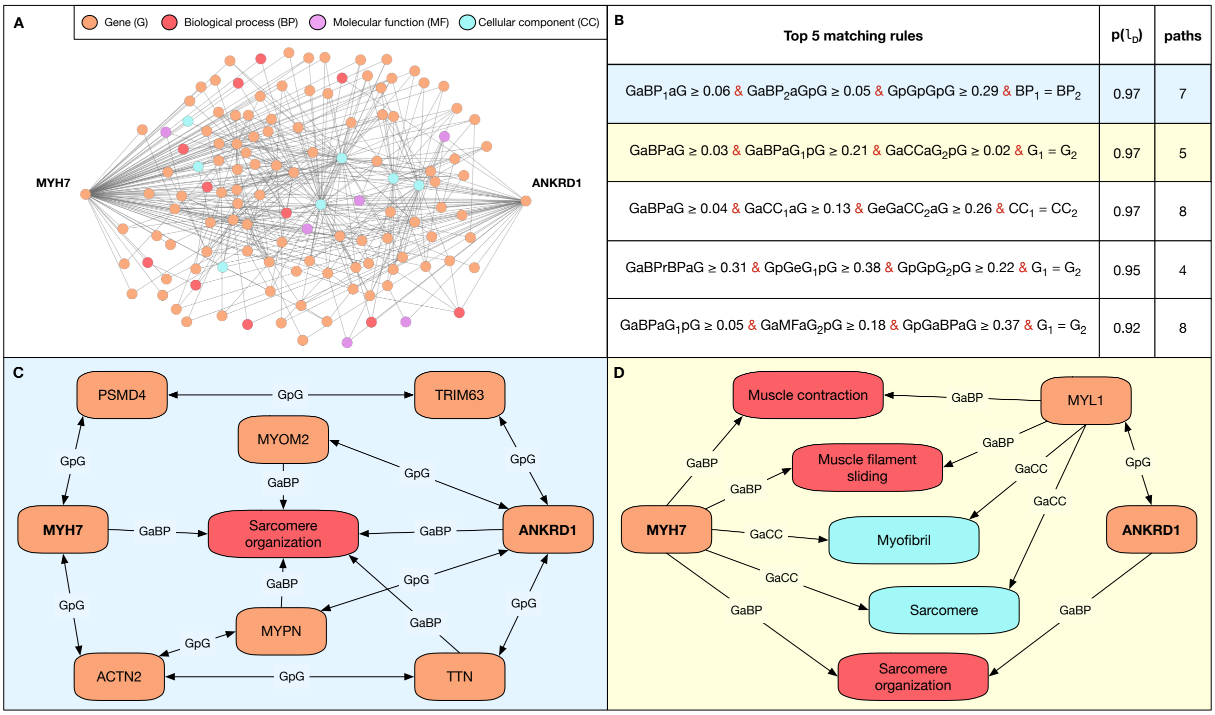
Illustrative examples with test set pathogenic gene pairs

To illustrate the type of explanation produced by ARBOCK, we consider the gene pair *MYH7-ANKRD1* from the independent test set, which was predicted as pathogenic with a high probability estimate by both decision set models. Previous studies have demonstrated the involvement of this gene pair, with a digenic pattern, in Left ventricular noncompaction disease (LVNC) (ORPHA:54260; HP:0030682) associated with Dilated cardiomyopathy (DCM) phenotype (HP:0001644) based on familial evidence [300]. Exploring paths of up to length 3 between these two genes in BOCK (excluding "Phenotype", "Disease" and "OligogenicCombination" entities) reveals a large subgraph comprising 342 paths, 127 nodes, and 447 edges (Figure 6.14.A).

The *DS excl.Pheno* model applied to the *MYH7-ANKRD1* gene pair returns matching rules ranked by their associated probabilities. Figure 6.14.B displays the top 5 rules along with the number of paths retrieved by querying the KG. Due to stringent path thresholds and the use of unification conditions, each of the top rules yields only a few paths. The first two rules are showcased in Figure 6.14.C and D, providing graphical explanations.

The first rule corresponds to a pattern where both genes of the combination share a common biological process (GaBP$_1$aG) and where a third gene, physically interacting, is also involved in the same biological process (GaBP$_2$aGpG ; BP$_1$=BP$_2$). Both genes from the pair are also linked with a long-range physical interaction.

The second rule describes a pattern where a central gene (G$_1$=G$_2$) physically interacts with the second gene while sharing a biological process (GaBPaGpG) and a common cellular component (GaCCaGpG) with the first gene. Both genes of the pair also share a common biological process (GaBPaG).



Top 5 matching rules	$p(l_D)$	paths
$GaBP_1aG \geq 0.06$ & $GaBP_2aGpG \geq 0.05$ & $GpGpGpG \geq 0.29$ & $BP_1 = BP_2$	0.97	7
$GaBPaG \geq 0.03$ & $GaBPaG_1pG \geq 0.21$ & $GaCCaG_2pG \geq 0.02$ & $G_1 = G_2$	0.97	5
$GaBPaG \geq 0.04$ & $GaCC_1aG \geq 0.13$ & $GeGaCC_2aG \geq 0.26$ & $CC_1 = CC_2$	0.97	8
$GaBPrBPaG \geq 0.31$ & $GpGeG_1pG \geq 0.38$ & $GpGpG_2pG \geq 0.22$ & $G_1 = G_2$	0.95	4
$GaBPaG_1pG \geq 0.05$ & $GaMFaG_2pG \geq 0.18$ & $GpGaBPaG \geq 0.37$ & $G_1 = G_2$	0.92	8

Figure 6.14. Predictive explanations generated by querying matching rules on the KG This figure showcases the example of the digenic gene pair *MYH7-ANKRD1*, part of the independent test set and predicted as disease-causing with the highest probability estimate by the *DS excl.Pheno* model. **(A)** Subgraph extracted by traversing all paths (excluding those traversing "Phenotype", "Disease" and "OligogenicCombination" nodes) of a length ≤ 3. A total of 342 paths, 127 nodes and 447 edges exists. **(B)** Top 5 matching rules ranked by their associated probability estimate. Each rule is written in their abbreviated form (see Table 5.1) with its conditions separated by &. Indices for node types (*e.g.* BP_1) are used in unification conditions (*e.g.* $BP_1 = BP_2$) to constrain entities to be the same across different metapaths. The numerical value associated with each metapath (*e.g.* ≥ 0.21) sets the path reliability threshold, which conditions the minimum path reliability score of all underlying paths. We display the number of paths obtained by querying the KG with the rule with that specific gene pair. **(C)** Returned explanation subgraph for the 1^{st} rule based on the 7 matching paths. **(D)** Returned explanation subgraph for the 2^{nd} rule based on the 5 matching paths. Entity types are represented with the same colors as in (A). Explanations subgraphs for the 3^{rd}, 4^{th} and 5^{th} rules are provided in Appendix Figure A.4.

In 4 out of the 5 presented rules, functional entities associated with the sarcomere (GO:0030017, GO:0045214) are shown relevant both via direct and indirect paths (Figure 6.14 and Appendix Figure A.4). Mutations in sarcomere protein genes have been linked to both LVNC and DCM diseases [301, 302]. Among traversed genes, *ACTN2* has been previously associated with LVNC [303], *TTN* to LVNC and DCM [304, 305] and *MYPN* to DCM [306]. The association of other genes postulates novel hypotheses for further exploration. For example, *MYL1*, *MYOM2*, *TRIM63* and *PSMD4* have been broadly associated with myopathies [307, 308, 309, 310] but not directly to LVNC or DCM yet, and could therefore be considered as potential targets to investigate.

To illustrate the limits of our approach, consider the gene pair *HOXB3*-*TG*, part of the independent test set, involved in congenital hypothyroidism (ORPHA:442) with a digenic pattern [311]. This gene pair is predicted as neutral by the *DS incl.Pheno* but as disease-causing by the *DS excl.Pheno* model, though with the lowest predicted probability estimate among the 12/15 positively predicted gene pairs (see 6.9.C). The gene pair is predicted as neutral by the *DS excl.Pheno* model as rules in this model tend to constrain the presence of associated phenotypes while the gene *HOXB3* is not associated with any phenotype (see Appendix Figure A.6.A). We could also observe that for the *DS excl.Pheno* model, a single rule is matched (see Appendix Figure A.5). However, the corresponding contextual explanation does not provide any relevant entities to explain the disease phenotype, as the traversed biological process and cellular location terms lack specificity and the traversed genes do not show any biological relevance or prior associations to the disease. Nonetheless, we could observe that BOCK contains potentially more informative and relevant paths to explain the physiopathology (see Appendix Figure A.6.B). This case serves to highlight a limitation of the current predictive approach: while the models may classify a gene pair as disease-causing, the rationale for such a classification may not be biologically informative or relevant.

Advantages of contextual explanations

These examples demonstrate the strength of the ARBOCK approach compared to traditional statistical machine-learning techniques. ARBOCK models not only provide predictions but also contextualise them with explanations directly related to background knowledge, thereby bridging computational predictions with biological interpretations. Unlike explanations derived from abstract features that can be challenging to decipher, ARBOCK offers insights into concrete entities and relationships, tailored to the specific gene pair of interest. This type of explanation provides geneticists with valuable insights, enabling both the expert validation of predictions and the discovery of novel, yet to be explored, mechanistic hypotheses.

6.8 Conclusion

In this chapter, we provided an extensive overview of the development and functionality of ARBOCK, an interpretable machine learning method that leverages the comprehensive data in the BOCK knowledge graph (KG) for predicting pathogenic gene interactions. We initially emphasised the potential of metapaths as predictive features and demonstrated their utility as queries for retrieving relevant information from the KG. We subsequently introduced an innovative predictive framework, integrating these metapaths to mine complex rules for link prediction. Thereafter, we examined the specific patterns characterising pathogenic gene pairs and assessed the performance of the rule-based model. Finally, we emphasised the inherent interpretability of the model, showcasing its capability to explain predicted pathogenic gene pairs via subgraphs derived from BOCK.

The development of ARBOCK arose from the need for improved interpretability in predicting gene interactions associated with oligogenic diseases. Beyond simply predicting pathogenic gene interactions, we demonstrated that ARBOCK's models, grounded in both KG paths and rules, offer direct insights into the contextual knowledge behind the predictions. This gives geneticists an intuitive means to understand the predictions. These contextual explanations not only enhance the trustworthiness of the models by allowing for the assessment of their validity but also aid researchers in formulating hypotheses about the underlying biological mechanisms related to the genetic disease under consideration.

The ARBOCK approach, while promising, is not without challenges. A predominant concern relates to accuracy, as we showed that our model excluding phenotype paths can still predict many false positives. While ARBOCK enhances the interpretability of genetic predictions, navigating multiple rule-based explanations without a systematic evaluation method can introduce ambiguity for the user. Ensuring a balance between interpretability and accuracy remains crucial for real-world applications and further advancements.

Moving forward, there are opportunities to further refine ARBOCK. Introducing novel techniques to improve the relevance of explanations, and adding more features at the gene level may strengthen its predictive capabilities. This approach could also be combined with novel or existing statistical approaches for improved accuracy. As of now, ARBOCK serves as a significant advancement, highlighting the value of knowledge graph-based models in genetic research and laying the groundwork for future, more refined interpretable tools.

Chapter 7

Discussion and Conclusion

This final chapter reflects upon the main objectives of our research and the corresponding contributions, each of which was directed towards enhancing the understanding of predictive approaches related to oligogenic diseases. Firstly, we improved the interpretability of existing predictive models by providing context to predictions. This was done by designing a novel integrated platform aiding geneticists in the analysis and exploration of patient-level variant data. Secondly, we developed a comprehensive biological knowledge graph, integrating data from known oligogenic diseases with multi-scale biological networks, thus offering a robust tool for exploring the complex interrelationships within oligogenic diseases. Lastly, we implemented an interpretable machine learning model that not only predicts the pathogenic potential of gene pairs but also provides insightful explanations into the underlying mechanisms of these diseases.

Beyond these key contributions, this chapter will also highlight some of the limitations and challenges encountered during our research. Recognising these allows us to outline potential areas for improvement and refinement in our methodologies and results. This will be followed by a discussion of potential future directions for research.

7.1 Scientific contributions

The field of medical genetics has experienced a transformative shift in recent years, with the traditional Mendelian model of genetic inheritance being challenged by an expanding understanding of human genetic disorders. These disorders often exhibit incomplete penetrance, high phenotypic variability or locus heterogeneity, leading to the consideration of alternative genetic models (Section 1.2).

The oligogenic model of inheritance has gained increasing attention over the last years. Oligogenic diseases, where a combination of causative variants is distributed among two or a few genes, offer a unique perspective for understanding genetic disorders that do not fit the traditional 'one gene - one phenotype' model. This model is particularly relevant for studying rare diseases, providing a bridge between the established monogenic and the poorly understood complex or polygenic disorders (Section 1.3).

Despite the promising potential of the oligogenic model, significant challenges have remained, particularly with regards to the interpretability of predictive methods that have been developed to detect pathogenic variant or gene combinations. Existing tools, while demonstrating accurate predictions, often lack clear explanations to enhance our understanding of causal characteristics behind these gene interactions. This presents a substantial hurdle for geneticists trying to validate these predictions or use them as a basis for generating new hypotheses about disease mechanisms.

To address these challenges, our research has made three major scientific contributions.

Firstly, we developed ORVAL an interactive web platform designed for the prediction, contextualisation and analysis of disease-causing variant combinations. This platform integrates established machine-learning models identifying pathogenic variant combinations and leverages post-hoc interpretability strategies to explain them, while also placing predictions in the context of existing biological knowledge. It equips geneticists with a user-friendly integrated tool, enabling the submission of patient-specific variant data and the exploration of multiple predictive results at different biological level. In particular, it provides insights into feature contributions, gene pathogenicity networks, and gene module mapping to biological networks (see Chapter 4). This resource bridges the gap between complex genetic predictions from statistical machine-learning methods and practical analysis aimed to support their interpretation for the end-users, giving geneticists an accessible tool to aid their research and diagnoses.

Secondly, we constructed a comprehensive biological knowledge graph (KG), BOCK, that integrates data from known oligogenic diseases with information from multiple biological networks. This integration captures the complex relationships and intricate interactions of oligogenic diseases, offering a rich, multi-dimensional perspective on these disorders (see Chapter 5). By using this KG, researchers can explore oligogenic disease information on a more systematic level and confidently examine gene relationships with traceable evidences. This new resource also expands potential feature engineering and machine-learning techniques that could be applied to solve existing or new predictive tasks.

Lastly, we proposed a novel interpretable model that leverages BOCK to provide meaningful explanations for pathogenic gene interactions. This model is built upon the semantics of paths between gene pairs in BOCK, effectively learning and applying rules for predicting potential pathogenic interactions (see Chapter 6). Unlike many existing machine-learning models, this approach has been designed to be fully transparent, facilitating the interpretation and assessment of predictions and providing geneticists with a much-needed tool to investigate the causal mechanisms behind oligogenic diseases.

Our research contributes significantly to the understanding of disease genetics, providing novel insights into interpretable machine learning in the biomedical field, the use of knowledge graphs, and furthering our understanding of oligogenic disease underlying patterns. Through our various developments and contributions, we explain how these areas are interconnected and vital for medical genetic research.

7.1.1 Explainable machine-learning in medical genetics

The demand for explainability in machine learning is strong within the biomedical field, especially in genetics. Experts often seek to understand the rationale behind a particular prediction, allowing them to apply their domain knowledge when validating the output. In this thesis, we explored two strategies to explain predictions: the use of post-hoc explanation techniques for gaining insights into existing statistical models (considered as *black-box models*) and the use of inherently interpretable models (or *white-box models*) (Section 1.4).

In the ORVAL platform (Chapter 4), we implemented a post-hoc explanation strategy, where black-box predictions are individually explained by their feature contributions or contextualised within domain knowledge. For example, we provided users with a detailed breakdown of feature contributions for each prediction made by the *Variant Combination Pathogenicity Predictor (VarCoPP)* [83], allowing the most influential features driving a specific prediction to be scrutinised. This approach focuses on the relative importance of individual features in a specific decision, but does not explicitly reveal how these features interact or offer a general decision-making rationale for the model [95, 104]. Therefore, to supplement these explanations, ORVAL leverages prior knowledge to provide contextual information for the predictions (as discussed in the next section), thereby enabling to assess the biological significance of biological relationships between predicted candidate genes.

When developing ARBOCK (Chapter 6), we chose to develop an inherently interpretable model that can provide, at the same time an interpretation of the model's inner working and decisions while also providing insightful contextual explanations. ARBOCK employs a rule-based model that provides both global and local interpretability, revealing feature interaction and complex patterns to the users while also directly revealing the model's inner working. This model also uses highly interpretable features based on metapaths extracted from BOCK. This provides our model with the ability to generate predictive explanations in the form of KG subgraphs. Compared to models using features based on abstract and complex scores, such as in VarCoPP, this approach can provide concrete domain knowledge and make explanations more intuitive for geneticists and researchers.

Our research underlines that the interpretability of a model does not depend solely on the model's characteristics or post-hoc explanation techniques; the choice of features used also significantly impacts the explainability of models. Moreover, the use of contextual explanations based on background knowledge in both ORVAL and ARBOCK can improve trust in predictions by allowing domain experts to assess them based their domain expertise [133].

As part of our exploration into rule-based models (Chapter 6), we considered using an associative classification approach [231, 294], guided by the 'Local Patterns to Global Models' (LeGo) principle [233]. This two-step approach begins by mining promising individual patterns before integrating them into a coherent model for a broader task. We performed the mining stage in an unsupervised manner on the minority class instances – the disease-causing gene pairs – using a frequent itemset mining approach. Our findings suggested that this approach not only permits a more extensive exploration of the feature space compared to traditional greedy methods but also helps discover useful rules in an imbalanced dataset [312]. This method controls the number of pathogenic gene pairs supporting a given pattern, limiting the risk of model overfitting. Finally, the final model was designed as a decision set classifier [235] due to its inherent ability to enhance interpretability by enabling each rule to be examined and interpreted independently of other rules.

These findings can be particularly valuable in the biomedical field, where imbalanced datasets with rare positive cases are common. Furthermore, our findings emphasise the importance of model structure when interpretability is a primary concern.

7.1.2 Leveraging background knowledge for biological insights

Another central theme of this thesis was to harness structured background knowledge – comprising biological networks, biomedical ontologies, and our custom-built KG, BOCK – to uncover biological insights.

In the case of ORVAL (Chapter 4), our approach transform predictions into a predicted pathogenic gene network that can be further explored in relation with background knowledge to assess their biological significance. We developed an analysis of protein-protein interactions with cellular compartments and provided pathway information from the *Reactome* ontology. The latter relates to *Gene Set Enrichment Analysis (GSEA)* [139], a method used to detect significant overrepresentation of certain gene functional annotations within a gene set. Additionally, by examining protein-protein interaction subnetworks that stem from a pathogenic module, we present evidence of both direct and indirect protein interactions to the user. This is further reinforced by indications of their localisation in particular cellular compartments. Collectively, these analyses can highlight biological interaction patterns, such as the simultaneous involvement of genes in a protein complex or a particular biological process, providing further support for their potential coordinated effect in causing the disease phenotype (Section 1.3).

In our exploration of BOCK and ARBOCK (Chapter 5 and Chapter 6), we considered three distinct strategies to leverage heterogeneous network data for predicting potential pathogenic gene pairs: traditional link prediction metrics, KG embedding methods, and a fully interpretable technique based on metapath patterns. Each of these methods has its unique strengths and weaknesses concerning model accuracy and interpretability. One of these approaches, investigated with a master student, involved the use of a KG embedding model. This technique provided exceptional accuracy on identifying potential pathogenic gene pairs and would require further investigations. Nevertheless, it presented interpretability challenges due to its complex, opaque nature. Traditional link prediction metrics, on the other hand, offered more interpretability but are not ideally suited to handle the inherent complexity of a KG. Lastly, our metapath-based interpretable approach, developed as part of our core research on ARBOCK, presents a good balance of predictive power and high interpretability. This last method represents an exciting opportunity to explore the rich patterns contained within knowledge graphs effectively.

When dealing with our knowledge graph, BOCK, extracting meaningful patterns in a knowledge graph, aiming to solve a particular link prediction tasks, can be a challenging endeavour considering the volume and complexity of these graphs. Our ARBOCK framework addresses this challenge by mining association rules from a combinations of different path types or metapaths. To the best of our knowledge, this is an original approach that extends over other KG rule mining approaches. Compared to existing KG rule mining approaches such as AMIE, AnyBURL, and RUDIK, our strategy differs significantly. While these systems attempt to exhaustively mine rules for all types of relationships in a knowledge graph, our approach focuses on mining rules predictive of a specific label. We do not incorporate rules with constant attributes in our search, and our scope is restricted to rules based on paths connecting two entities. Additionally, by mining frequent metapath associations, our method seeks rules that go beyond single "path-rules", instead uncovering more specific and complex subgraph patterns. These patterns also include unification conditions, further enriching their specificity as well as path score thresholds, reducing explanation complexity.

Additionally, in link prediction problems in networks and knowledge graphs, incompleteness and noise are often concerns. We demonstrated that using metapaths of various lengths can help capturing indirect associations between gene pairs, increasing the coverage of common associations. Furthermore, to handle noise within the knowledge graphs, we focused on reliable, high-quality sources for constructing BOCK, and introduced operators capable of filtering out low-quality or less informative paths in ARBOCK.

Previous methods have explored the potential of KG paths for scientific hypothesis generation. The *RPath* method [154] leverages paths within a knowledge graph to prioritise drugs for specific diseases, guided by transcriptomic data. It further unveils target proteins along these paths. Similarly, the *PoLo* method [151] uses policy-guided walks, informed by logical rules, to predict biomedical links for drug repurposing tasks. Contrastingly, our approach uses these KG paths to instruct a machine-learning model that in turn identifies meaningful paths as explanations. These paths can then also be used to generate scientific hypotheses regarding the mechanisms and interactors involved in oligogenic diseases. Therefore, our work contributes to the growing trend of harnessing the path information in KGs to generate scientifically relevant insights.

7.1.3 Contribution to oligogenic disease research

Our research offers distinct contributions to oligogenic disease research through the use of the ORVAL variant analysis platform and the exploration of prior knowledge interaction patterns in BOCK via our novel method ARBOCK.

ORVAL is a concrete research contribution, directly addressed to geneticists and researchers via a user-friendly analysis web platform (Chapter 4). It promotes the use of predictive approaches on variant combinations and enhance these results with a wealth of background knowledge and annotations, aiding in biological interpretation. Notably, the platform creates patient-specific oligogenic networks, potentially elucidating higher-order gene interactions, such as those involving multiple gene synergies or multiple modifier genes and enable the investigation of digenic effects in variant combinations [82].

More practically, ORVAL distinguishes itself not only with these features but also with its speed and intuitive design. By using indexed pre-computed variant-level features, it allows geneticists to efficiently process variant data and obtain predictions in a short time.

Furthermore, ORVAL's modular structure accommodates future updates and integration of additional predictive methods such as VarCoPP 2.0, which has been implemented already, and the High-Throughput Oligogenic Prioritiser (HOP) developed as part of *Barbara Gravel* PhD thesis. ORVAL has also proven its practical utility, being employed in clinical research for both the discovery and validation of variant combinations.

In the context of BOCK and ARBOCK, we have shown their potential to highlight possible mechanisms underlying oligogenic diseases. Indeed, some reported oligogenic cases exhibit functional evidence of negative epistasis [77, 78, 79]. The ability to detect or predict involved molecular mechanisms could therefore significantly advance the investigation of digenic disorders, particularly in cases where the effects of genes are not fully understood [78]. Gene interactions can be complex, with diseases often resulting from indirect effects mediated through other genes [313]. Our approach in ARBOCK aims to reveal such cases, helping to reveal, via the generated explanations, the common involvement of biological processes, protein complexes or functions which could indicate synergistic mechanisms. It could also help identify other target genes that might be indirectly responsible for disease causation. We believe that continuing in this research direction could contribute substantially to our understanding of complex genetic diseases.

In conclusion, our research serves as a significant step forward in the understanding and prediction of oligogenic diseases. Through the integration of biological knowledge into predictive models, and by providing novel tools designed for interpretability and transparency, we have moved closer to revealing the hidden complexity of oligogenic diseases. This work not only empowers geneticists with the means to more effectively interpret and validate predictions but also opens up new avenues for future research on these disorders.

7.2 General issues and limitations

Our thesis work made significant contributions through the development of BOCK, ARBOCK, and ORVAL. However, there remain certain limitations that need to be addressed in future research.

7.2.1 Limitations in prediction accuracy

In ORVAL, the effective application of the machine learning predictors, despite their high accuracy, presents a challenge due to the nature of genomic data. The predictors, VarCoPP and its subsequent version, have exhibited impressive accuracy levels, achieving a False Positive (FP) rate of only 5% and sensitivity of 95% in VarCoPP 2.0. However, when applying these models to full exomes, the sheer volume of variant combinations to evaluate results in a large absolute number of incorrect predictions. Even with the low FP rate, the combinatorial explosion in whole exome analysis could overwhelm geneticists with false leads. Therefore, for whole exome submissions, users are strongly advised to use all available filtering options and to submit a gene panel to manage the number of predictions. Nevertheless, these filters limit the exploration of combinations to well-known genes, which can hinder discoveries.

Regarding ARBOCK, when considering all prior knowledge contained in BOCK, such model could reach a sensitivity of 81% and yield a FP rate of 7%. However, as discussed in Section 6.5, this model's performance evaluation may be overly optimistic, overestimating the True Positive (TP) rate due to the coverage discrepancy between known pathogenic genes and the total gene pool, subsequently limiting its broad applicability. Conversely, the model evaluated without phenotype information, while free from this bias, offers a sensitivity of 75% but exhibit a high FP rate of 23%. The lack of specificity in this model underscores the necessity for integrating novel biological information in BOCK, refining the rule mining with additional features, or coupling this model with more accurate machine-learning methods for enhanced prediction accuracy.

When evaluated on the identical classification task and using the same training data as ARBOCK, the Distmult-based model, as described in Section 5.6, exhibits superior predictive accuracy. This performance gap can be attributed to two primary factors. First, the Distmult model works in a larger feature space, with embeddings having dimensions between 200 and 300. Second, it employs a balanced random forest with a tree count ranging from 340 to 540 and no constraint on tree depth, in contrast to ARBOCK's simpler rule-based approach with at most four conditions per rule. This added complexity enables the Distmult-based model to identify more nuanced patterns that distinguish pathogenic gene pairs. However, it raises questions about the nature of the features it leverages for prediction. Specifically, it is not clear if the model takes into account the types of relationships and connectivity between gene pairs, or if it is primarily influenced by single-gene information. Moreover, the high accuracy of the Distmult-based model comes at the expense of interpretability. Models that employ knowledge graph embeddings are often labeled as "black boxes", a significant concern in medical genetics where interpretability is crucial. The initial stage of embedding leads to a loss of information, limiting the effectiveness of any subsequent efforts to interpret the model.

7.2.2 Challenges with big data and computational efficiency

In our work on BOCK and ARBOCK, the greatest computational burden arise from the inherent characteristics of pattern mining and path traversal in large-scale knowledge graphs. The size and complexity of knowledge graphs present significant challenges. The generation of rules via association mining in ARBOCK can also generate a large number of candidate patterns, requiring stringent constraints on the search space to limit the computational burden. While our methodologies have shown promising results, their efficiency is inevitably impacted by the intensive computations required.

Additionally, the aspiration to examine higher-order genetic combinations (*e.g.* trigenic, tetragenic, ...) and to integrate additional machine learning methods into ORVAL introduces further computational complexity, challenged by the combinatorial nature of these analyses. These computational constraints also limit the current use of ORVAL to single patient analyses. This restriction can limit its broader clinical applicability, as analysing cohorts of patients is a recurrent aspect of clinical research.

7.2.3 Limits of model interpretability and explanations

ORVAL, while providing post-hoc interpretability analyses and contextualisation of predictions, is limited by the scope of these methods. The feature contribution analysis, which explains the relative importance of features in VarCoPP predictions, can only show their individual contribution and assume an additive relationship [104]. But this method cannot display intricate feature interactions or to identify the underlying decision thresholds considered by the model. Furthermore, these features, extracted from sophisticated bioinformatics scores like the Inheritance-mode specific pathogenicity prioritisation (ISPP) score [181], can be challenging to understand due to their abstract nature. These features are even more complex to interpret when considering their potential interactions.

Moreover, ORVAL employs background knowledge to provide an interpretation of gene modules from the pathogenic gene network. While this approach can assist in interpreting the network, it doesn't directly illustrate the underlying rationale for the prediction. Therefore, it serves more as an aid to interpretation rather than as a concrete predictive explanation.

ARBOCK, despite delivering improved interpretability through a rule-based model, also faces limitations. Its decision set model allows for multiple explanations due to the independent nature (`OR` relationship) and potential overlap of rules. When the associated probabilities of the rules are similar, it can be difficult to identify the most relevant explanation.

Furthermore, graphical explanations, obtained by translating rules into KG queries, can result in extensive subgraphs, particularly in dense regions of the KG. Even with the application of path score thresholds and unification conditions, explanation complexity can potentially remain high, leading to a need for more manual effort on the user side.

Finally, the task of validating the explanations provided by ORVAL and ARBOCK presents a challenge that has not been addressed in this thesis. While explanations are designed to be informative, their utility is subject to evaluation by domain experts on a case-by-case basis. In ARBOCK, efforts were made to enhance the quality of explanations through objective metrics, such as the number of rules or paths and the introduction of a path reliability score to filter out low-quality or non-informative paths. However, the biological relevance of these explanations is often tied to the specific disease aetiology under investigation. This aspect of biomedical relevance was not considered in the current research, leaving room for further work to assess the utility of the generated explanations.

7.2.4 Issues with the quality of background knowledge

Working with background knowledge and data in our research presents several unique challenges. Aspects like noise, incompleteness, and bias in our data sources can impact the quality and reliability of our predictions.

Our predictions, generated using tools such as ARBOCK and ORVAL, are closely tied to the quality of data in the OLIDA database. Despite deriving from peer-reviewed publications, this database can contain spurious associations, leading to noise. While a substantial number of reported clinical cases provide limited familial and statistical evidences [80], the curated evidence scores can mitigate this problem to some extent [71] but require time-consuming curation efforts.

Given that oligogenic disease research is a relatively new field, the data in OLIDA is understandably incomplete. Novel evidence of oligogenicity for various genetic diseases is yet to be discovered. Furthermore, a bias is present due to a disproportionate focus on certain digenic diseases. For instance, the discovery of digenic causes for some ciliopathies (*e.g* Bardet-Biedl syndrome) has led to increased research into oligogenic causes for other ciliopathies (*e.g* Retinitis Pigmentosa). Such investigative bias results in an overrepresentation of certain diseases, challenging the machine-learning assumption of data point independence.

Biological networks also present their challenges. Despite strict quality criteria, these networks can be noisy, impacting inference accuracy. Weighted biological networks, such as co-expression networks or sequence similarity networks, can provide a guide for filtering potentially spurious associations. However, finding a suitable minimum cutoff to remove noise is a non-trivial task often left to the researcher's discretion.

Background knowledge can also be incomplete and biased. Sequence analysis-derived networks (like protein domains, families, and sequence similarity) are often considered exhaustive and unbiased, while high-throughput experiments-derived networks (like co-expression and protein-protein network data) while dense present some biases [289, 290]. In contrast, curated experimental or clinical evidence-derived networks, such as Gene Ontology or Human Phenotype Ontology associations, are sparse and can be very biased towards over-studied, especially disease-related genes. Although automatic annotation techniques can somewhat improve coverage for these networks, they may introduce additional bias [127].

Finally, bioinformatics databases often contain redundant information, a situation exacerbated by inference methods commonly applied to increase the coverage of these databases. For example, protein functions can be inferred from protein domains through methods like InterPro2GO [186], or from sequence similarity via Ensembl Gene Trees [184]. To mitigate this issue in BOCK, we selectively filtered subsets of inferred data from the source databases. For instance, in Gene Ontology, we focused on specific annotation evidence codes excluding electronically inferred ones. However, despite these precautions, some inference bias may still exist as inferred annotations also guide biologists in their formal investigations and data curation. Consequently, even annotations considered empirically supported can be influenced by these underlying inference methods. This could manifest as artifactual patterns of co-occurrence in the knowledge graph, capturing these redundancies rather than true biological relationships.

7.2.5 Knowledge graph data integration simplifications

In the creation of the BOCK knowledge graph, we opted for certain simplifications during data integration. Specifically, we chose to connect biomedical ontology terms based on high semantic similarity, thereby losing the original hierarchical structure of the ontologies. This decision was partly motivated by the ARBOCK framework's limitation on path length during training. While shorter paths make the computational process more efficient, they may not traverse complex ontology structures, potentially missing out on semantically connected terms.

We also applied pre-filtering to co-expression and sequence similarity links to manage the graph's complexity. While this reduces computational demands, it could omit biologically meaningful but weaker interactions, which may affect the accuracy of models derived from the graph.

These simplifications, although practical for computational reasons, introduce limitations that could affect the quality of inferences and the granularity of queries made on the graph. Further research is needed to assess these limitations.

7.2.6 Assumptions in modelling pathogenic gene interactions

The methodologies and tools developed in this research show promise but are based on several assumptions and modelling simplifications that warrant further examination.

The first simplification in our approach involves abstracting genetic information to the gene level, effectively ignoring the identity and position of specific variants within the gene. This is applicable both for generating explanations in ORVAL and for constructing our predictive model in ARBOCK. This level of abstraction allows us to focus on the biological interaction context, operating under the assumption that a damaging variant exists within the gene, sufficiently altering its function, without specifying which variant it is or detailing its specific molecular consequences. However, the impact of a specific variant on protein folding and function is important for understanding its system-level interactions and potential pathogenicity. We investigated the idea of mapping variant information to protein domains to understand their impact on specific functions or interactions. However, this approach is constrained not only by the limited data available for mapping variants to domains in our training dataset but also by the lack of comprehensive networks capturing the impact of variants on interactions and on domain-domain interactions.

Another assumption lies in the construction of pathogenic gene networks, such as done in ORVAL, where we assume that oligogenic disease patterns can be inferred by aggregating multiple pairwise predictions at the gene level. This strategy compensates for the scarcity of training data for oligogenic cases involving more than two genes. However, the validity of this assumption remains to be confirmed.

Regarding the selection of neutral gene pairs used to train our predictive approaches, assumptions include considering the maximum CADD variant as an effective indicator of a gene's pathogenicity potential or setting up the arbitrary threshold of 50 healthy individuals exhibiting gene pair with pathogenic potential as a condition for considering the gene pair as neutral. These assumptions introduce an inherent uncertainty in the neutral gene pair selection process and further research should ensure of the robustness of this selection strategy.

Lastly, our predictive models frame the problem as a binary classification: disease-causing (digenic) or neutral. This approach does not account for gene pairs that may independently contribute to disease. While our current methodology focuses on interaction patterns and is less likely to capture monogenic signals, future methods, such as those based on KG embeddings, may not be as robust against this issue, as indicated by preliminary work (Section 5.6). Therefore, additional studies are needed to validate these assumptions and assess their influence on both the predictive accuracy and interpretability of our models.

7.3 Future perspectives

Building on this research, there are several directions future work could take, primarily addressing current limitations and further improving the approaches and tools we have developed.

7.3.1 Improvements on the training data

To further advance our understanding of oligogenic diseases, refining and expanding our training data would lead to improved accuracy and open the way to novel predictive tasks.

Expansion and improvements of the OLIDA database

The Oligogenic Disease Database (OLIDA) will gain significant enhancements through a community curation effort, the expansion of the scientific literature and the establishment of standards for reporting these cases [71, 80], which will potentially improve both the quantity and quality of our training dataset. To streamline the curation process, a semi-automated curation approach via an active learning approach is being developed by *Charlotte Nachtegael*, which could improve the task of curating these articles.

Improved selection of non-oligogenic variant data

The utility and precision of pathogenicity predictors such as VarCoPP and ARBOCK could be further improved by enriching our selection of neutral variant and gene combinations, tapping into larger cohort data such as provided by the UK10K and 100,000 genomes projects [23, 314].

In addition, the identification of monogenic cases versus oligogenic ones could be more accurately achieved by making use of known pathogenic variant associations, such as provided in ClinVar [33]. The use of such databases could serve as a powerful tool to distinguish cases with independent gene contributions from those involving the synergy of multiple genes, thereby refining the results of these predictive tools.

Incorporating experimentally validated epistasis data

Given the essential role of BOCK and ARBOCK in understanding disease mechanisms, we find it promising to consider experimental datasets that examine negative epistasis, which could enlighten mechanisms behind oligogenic diseases [57, 77, 78, 79]. Synthetic Sick or Lethal (SSL) data, derived from model organisms or cell line screenings and available in databases such as BioGRID [315], present valuable resources, that could be transferred to human through homology approaches [316]. This integration could allow a comparative study between these experimentally validated interactions and those described in OLIDA, with the aid of patterns discovered in BOCK, potentially yielding significant advancements in our understanding of biological mechanisms driving these diseases.

7.3.2 Expansion of the ORVAL platform

While the ORVAL platform serves as a powerful tool in disease variant prediction, certain enhancements could improve its clinical utility, user interface, and the quality of underlying data.

Incorporating trio variant data

Rare disease research often involves comparing the genotypic and phenotypic data of a patient with those of their parents to study inheritance patterns, which is known as *trio analysis* [47]. A logical next step could be to enhance ORVAL's ability to accept trios as input, by enabling the submission of additional variant data from both parents into the patient's analysis. This feature could help in filtering out variant combinations already present in both parental datasets, particularly when the parents exhibit no signs of the disease. Furthermore, the platform could help differentiate between de novo variants and inherited ones. In terms of results in the presented gene pathogenicity network, it could highlight the distinction between newly emerging genes and those that also feature in the parents' genetic networks.

Refining the oligogenic network

Several enhancements could be made to the gene pathogenicity network provided in ORVAL. For instance, the platform could be updated to visually represent the frequency of disease-causing gene-pairs. Further work could also be done to devise a less biased aggregation method, for example by normalising scores by the length of the considered genes. Nevertheless, this type of aggregations falls into complex considerations related to the burden of variants and collapsing strategies [47].

Furthermore, the integration of the BOCK knowledge graph and the ARBOCK method into ORVAL could supplement the gene pathogenicity network analaysis.

Scaling from single-patient to multi-exome analyses

ORVAL's nature as a public platform might limit its use for analysing sensitive medical data due to strict data exchange policies. The current configuration also restricts the number of parallel jobs that can be submitted to five, thus limiting large-scale analyses. To overcome these limitations, a private cloud service version of ORVAL is being explored. This adaptation, supported by the Foundation 101 Genome [1] and led by *Emma Verkinderen* and *Nassim Versbraegen*, would ensure compliance with data exchange policies while also facilitating concurrent analyses on multiple patients.

Furthermore, the development of new types of methods and visualisations would be needed to handle patient heterogeneity and highlight statistically significant variant or gene combinations. These developments would first require higher exome-level accuracy, a challenge currently being addressed with the development, by *Barbara Gravel*, of HOP: a variant combination prioritiser that leverages BOCK to rank variant combinations by

[1] https://www.f101g.org/

integrating the pathogenicity predictions from VarCoPP with phenotypic information that has been diffused onto the knowledge graph.

Enhancements to the Annotation Database

Our internal annotation database enables rapid feature and annotation retrieval in ORVAL. Adding new types of annotations could enhance the interpretability of predicted results. In addition, it will be necessary to keep these resources updated with the latest database versions and genome assemblies. For example, the recent construction of a human pan-genome [317] could help reduce ethnicity bias in the current reference genome assembly.

Moreover, our integration of pre-computed CADD results and annotations has its limitations, as some InDel annotations can be missing. A potential solution could be to run the CADD predictor for these missing variants in real-time and store the results in our database. Although this could slow down the annotation process, it would enhance both the coverage and accuracy of the results provided in ORVAL.

7.3.3 Future directions on knowledge graph applications

In considering the future perspectives of the developed KG, BOCK, and its applied methods, such as ARBOCK, we identified important directions to consider, including data consolidation, interpretability improvements, feature incorporation, and the extension to wider applications.

Expansion of the BOCK knowledge graph

Our first point of focus revolves around augmenting the datasets integrated into BOCK. Currently, we rely on one resource per component type (*e.g.* co-expression data are derived from TCSBN/GTEx [195, 196]). However, a future strategy could consider consolidating diverse resources to enrich data quality. Our selection of thresholds, currently mainly informed by literature recommendations, should also be reassessed for their optimisation within our specific tasks. Further, refining the balance between our strategy to simplify graph structure and the preservation of ontology hierarchy information could yield more nuanced and accurate results. Additionally, the inclusion of more specialised networks such as gene regulatory networks could provide deeper insights gene-gene relationships.

Improvement of ARBOCK path mining

For ARBOCK, the development of systematic assessment methods could enhance the quality of generated explanations, potentially guided by phenotype relevance. Additional enhancements could include incorporating gene-level or path-level features to improve the accuracy of the rule mining process. For example, paths could be scored based on the essentiality of the traversed genes. We could also explore alternate path scores such as the Degree-Weighted Path Count (DWPC) [162, 163].

In depth analysis of global topological properties in BOCK

Future research could also examine the topological properties of our knowledge graph, BOCK. Traditional network analyses often focus on homogeneous networks, investigating properties like diameter, clustering coefficient, or whether the network exhibits small-world or scale-free characteristics [191, 318]. However, there is limited research on how the integration of heterogeneous resources impacts these properties in a biomedical knowledge graph. Understanding these global topological features could offer valuable insights into the reachability of node pairs via paths and the efficacy of data mining and inference methods applied to the graph. For instance, the small-world property, if present, could imply efficient information propagation but might also introduce challenges in distinguishing biologically meaningful paths from random connections. Conversely, a scale-free topology could indicate the presence of highly connected nodes, which could be of particular interest in the study of disease mechanisms. Investigating these properties could provide a better understanding that informs the design and interpretation of machine learning and data mining approaches applied to BOCK.

Exploration of KG latent representations

Looking towards knowledge graph (KG) applications, the exploration of KG latent representations via KG embedding methods could help in identifying potential pathogenic gene interactions with a high accuracy, as shown by preliminary results and presented in Section 5.6. However, we have shown that this approach comes at the cost of interpretability and can lead to unexpected behaviours, such as predictions made solely based on the monogenic signal of one of the genes. Future research should focus on reconciling accuracy and interpretability. This could be done via techniques like the grey-box approach [319], which combines a black-box classifier (here the KG embedding approach) and a white-box surrogate classifier (such as ARBOCK) in a semi-supervised framework. Additionally, this could also employ post-hoc interpretation techniques like CRIAGE and XKE [320, 321]. Another potential strategy could involve guiding the KG embedding approach by focusing on specific path-queries or metapaths, combining the benefits of latent approaches and the interpretability of these query results [322, 323].

Widening applications and availability

As BOCK and its predictive methods evolve, ensuring broad availability and user-friendliness of these tools is imperative. Integration into platforms like ORVAL could provide additional insights.

Beyond identifying pathogenic variant combinations, the scope of BOCK and ARBOCK could also be widened to identify modifier genes [56, 61] or to suggest drugs suitable for repurposing [163]. While these new applications would require additional data, they could build upon the foundational principles established in this thesis research.

7.3.4 Assessment of predictive explanations by domain experts

A notable aspect of this research is the focus on generating explanations grounded in domain knowledge, as opposed to relying solely on abstract features. This approach aims to make the explanations more interpretable for domain experts, such as researchers and clinicians. One avenue for future work could involve conducting surveys among these experts to measure the conditions under which an explanation is deemed satisfactory or unsatisfactory [122].

To facilitate this, the ORVAL platform could incorporate a grading system that collects user feedback on the quality of each proposed explanation. This real-world input could then be used to refine the explanation-generating algorithms to better meet the needs of clinicians and researchers specialising in oligogenic diseases.

Additionally, automated methods for assessing the quality of these explanations could be explored. For instance, external knowledge sources like PanelApp [324] could be employed to evaluate the relevance of genes involved in the paths generated by the ARBOCK framework. This automated assessment could serve as a complementary approach to human expert evaluation, offering a more comprehensive understanding of the explanation's quality and relevance.

7.4 Conclusion

The field of medical genetics is evolving, moving away from the traditional 'one gene - one phenotype' model to acknowledge the complexity of oligogenic inheritance, where a few genes collectively influence disease phenotypes. Our research responds to this evolution and addresses the critical need for predictive tools that are not only accurate and efficient, but also interpretable.

To respond to this need, our research focused on leveraging prior biomedical knowledge through two distinct interpretation strategies. In the initial approach, we aimed to improve the interpretability of opaque predictive models, using post-hoc techniques to explain and contextualise their predictions. This led to the development of an integrated web platform that facilitates patient-level variant analyses for geneticists and places variant combination pathogenicity predictions in the context of biological knowledge. In our subsequent approach, we emphasised full transparency in modelling. We first constructed a comprehensive knowledge graph to better understand oligogenic diseases. We then introduced a transparent machine learning model that offers direct insights into the mechanisms behind pathogenic gene interactions.

By integrating structured biological knowledge with predictive models and improving their interpretability, we have provided geneticists with tools that can increase confidence in predictions, allow expert validation, and drive the generation of new hypotheses about disease mechanisms. This research has highlighted the value of knowledge graph for improved inference and insights and the importance of interpretable machine learning, especially in the biomedical field.

However, our research also revealed multiple challenges, including limitations in prediction accuracy, computational efficiency, and model interpretability that may restrict their broader clinical applicability. Additionally, the quality of data from current databases presents another obstacle, as noisy or incomplete data can negatively impact the performance of our methods.

Looking forward, we see a clear path for improvement. Enhancements in data collection, expanding and refining our current approaches, and addressing computational and interpretative issues present significant opportunities. By addressing these limitations, we can further advance our understanding of oligogenic diseases.

In conclusion, our work underscores the impactful role of integrating knowledge graphs and interpretable machine learning in genetic disease research. It provides the scientific community with practical tools to further this area of study. As we advance on this path, we are assured these approaches will equip geneticists with the knowledge and tools they need to uncover the hidden complexity of genetic diseases.

Appendices

A.1 Appendices: Tables

Ref.	Year	Publication type	Disease
[325]	2020	Clinical study	Hereditary angioedema
[326]	2020	Clinical study	Bethlem myopathy
[327]	2020	Meta-study	Various
[328]	2020	Clinical study	Inherited bleeding and platelet disorders
[85]	2020	Method article	Various
[329]	2021	Meta-study	Various
[330]	2021	Literature review	Various
[331]	2022	Opinion article	Hearing loss
[332]	2022	Literature review	Early onset neuronal diseases
[333]	2022	Meta-study	Rare diseases
[334]	2022	Method article	Cardial channelopathies
[335]	2022	Method article	Congenital heart defects
[336]	2022	Meta-study	Hypoplastic left heart syndrome
[337]	2022	Clinical study	Familial arrhythmogenic right ventricular cardiomyopathy
[338]	2022	Literature review	Neuromuscular disorders
[71]	2022	Database article	Various
[80]	2023	Opinion article	Various
[339]	2023	Clinical study	Long QT syndrome, Brugada Syndrome
[340]	2023	Clinical study	Severe adult obesity

Table A.1. Research papers mentioning the ORVAL platform. All scientific publications citing ORVAL as inspiration, comparison or potential perspective are listed up to April 2023. Other citations using ORVAL in genetics studies are listed in Table 4.2

Gene-linked edge type	**Digenic genes vs. Disease genes**	**Neutral genes vs. Connected genes**
associated-Phenotype (aP)	11.00 (***)	-3.00 (n.s.)
physInteracts-Gene (pG)	-2.00 (n.s.)	-3.00 (***)
associated-BiologicalProcess (aBP)	2.00 (***)	0.00 (n.s.)
belongs-ProteinFamily (bPF)	0.00 (*)	0.00 (***)
associated-CellularComponent (aCC)	0.00 (*)	0.00 (***)
associated-MolecularFunction (aMF)	0.00 (n.s.)	0.00 (n.s.)
hasUnit-ProteinDomain (uPD)	0.00 (**)	1.00 (***)
forms-ProteinComplex (fPC)	0.00 (n.s.)	0.00 (*)
seqSimilar-Gene (sG)	-0.50 (n.s.)	0.00 (**)
coexpresses-Gene (eG)	-10.00 (n.s.)	10.00 (n.s.)

Table A.2. Median differences in edge-type specific degree for related gene sets. The degree distribution for each gene-linked edge types in BOCK is analysed. The median difference is compared for pairs of related distributions: digenic genes *vs.* disease genes (first column), and neutral genes *vs.* connected genes (second column), respectively. The p-values indicating the significance of the differences are shown in parentheses (obtained via Mann-Whitney U tests with Bonferroni correction: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

DOME	Version	1.0
Data	Provenance	OLIDA [71], Mentha [192], STRING db [258], TCSBN [195], GTEx [196], InterPro [196], CORUM [187], Gene Ontology [189], Human Phenotype Ontology [129], dbNSFP [185], HGNC [177], Ensembl [174], UniProt [178]. All these sources are merged as the knowledge graph BOCK. Negative gene pair data are generated from the 1000 genomes Project [18].
	Dataset splits	426 positive instances, 42.600 negative instances for training data. 15 positive instances as validation set.
	Redundancy between data splits	No overlap
	Availability of data	Yes: BOCK knowledge graph at: `doi.org/10.5281/zenodo.7185679` and 1000 genome project at: `www.internationalgenome.org`
Optimization	Algorithm	Associative classification with a weighted set cover approach.
	Meta-predictions	No.
	Data encoding	Transformation of gene pair paths into metapath-based rules.
	Parameters	path cutoff = 3 ; minsup ratio = 0.2 ; max rule length = 3 ; α = 0.5
	Features	Metapath features obtained by aggregating the path information for all gene pairs of the positive set.
	Fitting	To avoid overfitting, rules can only associate 3 different metapaths and needs to be supported by at least 20% of positive training instances.
	Regularization	No
	Availability of configuration	Yes: `github.com/oligogenic/bock_rule_mining`
Model	Interpretability	White box model (rule-based) with knowledge-based explanations for gene pairs predicted as positive.
	Output	Classification probability (with predicted class) and explanation subgraphs for positively predicted instances
	Execution time	1000 samples in 0.56 seconds on a single Intel i5 core.
	Availability of software	Yes, Github: `github.com/oligogenic/bock_rule_mining`
Evaluation	Evaluation method	Both stratified 10-fold cross validation and evaluation on 15 independent validation positive instances
	Performance measures	Precision, Recall, ROC AUC, PR AUC
	Comparison	Model with and without phenotype information and the DiGePred model on the independent validation set.
	Confidence	Stability of the ROC AUC measure over different folds of cross-validation (std. 0.03).
	Availability of evaluation	Yes: Github: `github.com/oligogenic/bock_rule_mining`

Table A.3. DOME recommendation table consisting of essential information to assess the machine learning approach [298]. These criteria correspond to the decision set models presented in this study.

Rule	D coverage	N coverage	$p(l_D \mid r)$
$GaBPrBP_1aG \geq 0.33$ & $GpGaBP_2aG \geq 0.20$ & $GpGaCCaG \geq 0.28$ & $BP_1 = BP_2$	82	201	0.976
$GaBP_1aG \geq 0.06$ & $GaBP_2aGpG \geq 0.05$ & $GpGpGpG \geq 0.29$ & $BP_1 = BP_2$	81	210	0.975
$GaBPaG \geq 0.12$ & $GpGaCCaG \geq 0.26$	104	288	0.973
$GaBPaG \geq 0.03$ & $GaBPaG_1pG \geq 0.22$ & $GaCCaG_2pG \geq 0.03$ & $G_1 = G_2$	99	302	0.970
$GaBPaG \geq 0.04$ & $GaCC_1aG \geq 0.13$ & $GeGaCC_2aG \geq 0.26$ & $CC_1 = CC_2$	74	235	0.969
$GaBPrBPaG \geq 0.31$ & $GpGeG_1pG \geq 0.38$ & $GpGpG_2pG \geq 0.22$ & $G_1 = G_2$	81	396	0.953
$GaBPaG_1pG \geq 0.06$ & $GaMFaG_2pG \geq 0.19$ & $GpGaBPaG \geq 0.38$ & $G_1 = G_2$	103	691	0.937
$GpG_1aBPaG \geq 0.42$ & $GpG_2aCCaG \geq 0.26$ & $G_1 = G_2$	122	831	0.936
$GaCCaG_1pG \geq 0.02$ & $GaBPaGsG \geq 0.02$ & $GpGpG_2pG \geq 0.19$ & $G_1 = G_2$	74	618	0.923
$GaBPaG_1pG \geq 0.19$ & $GaMFaG_2pG \geq 0.18$ & $GpGaMFaG \geq 0.36$ & $G_1 = G_2$	81	683	0.922
$GaBPaG_1eG \geq 0.12$ & $GaMFaG_2eG \geq 0.29$ & $GpGaBPaG \geq 0.38$ & $G_1 = G_2$	86	728	0.922
$GsGaBPaG \geq 0.10$ & $GsGaCCaG \geq 0.17$	80	725	0.917
$GaBPaGsG \geq 0.13$ & $GpGaCCaG \geq 0.19$	90	846	0.914
$GaBPaGpG \geq 0.36$ & $GpG_1aCCaG \geq 0.13$ & $GpG_2bPFbG \geq 0.01$ & $G_1 = G_2$	83	786	0.913
$GaBPaG_1sG \geq 0.19$ & $GaCCaG_2sG \geq 0.16$ & $G_1 = G_2$	83	858	0.906
$GeG_1aBPaG \geq 0.25$ & $GeG_2aMFaG \geq 0.37$ & $GpGaBPaG \geq 0.36$ & $G_1 = G_2$	105	1169	0.900
$GaBPaGpG \geq 0.29$ & $GpG_1bPFbG \geq 0.05$ & $GpG_2uPDuG \geq 0.05$ & $G_1 = G_2$	79	887	0.899
$GaBPaG_1pG \geq 0.21$ & $GaCCaG_2pG \geq 0.03$ & $GpGuPDuG \geq 0.10$ & $G_1 = G_2$	84	986	0.895
$GaCC_1aG \geq 0.12$ & $GpGaCC_2aG \geq 0.23$ & $GpGsGpG \geq 0.19$ & $CC_1 = CC_2$	81	1001	0.890
$GeG_1aCCaG \geq 0.02$ & $GeG_2sGpG \geq 0.34$ & $GpGaMFaG \geq 0.33$ & $G_1 = G_2$	84	1097	0.884
$GeG_1aCCaG \geq 0.19$ & $GeG_2aMFaG \geq 0.09$ & $GpGaBPaG \geq 0.38$ & $G_1 = G_2$	119	1604	0.881
$GaBPaGeG \geq 0.64$ & $GeG_1aBPaG \geq 0.05$ & $GeG_2pGpG \geq 0.24$ & $G_1 = G_2$	90	1403	0.865
$GpG_1aMFaG \geq 0.18$ & $GpG_2eGpG \geq 0.28$ & $GpGsGpG \geq 0.18$ & $G_1 = G_2$	81	1358	0.856
$GpG_1aCCaG \geq 0.12$ & $GpG_2aMFaG \geq 0.24$ & $GpGbPFbG \geq 0.08$ & $G_1 = G_2$	97	1657	0.854
$GaBPaGeG \geq 0.58$ & $GeGaCC_1aG \geq 0.31$ & $GpGaCC_2aG \geq 0.22$ & $CC_1 = CC_2$	115	2511	0.821
$GpGaBPaG \geq 0.14$ & $GpGaCCaG \geq 0.25$	218	4769	0.821
$GaCCrCC_1aG \geq 0.12$ & $GeGaBPaG \geq 0.29$ & $GeGaCC_2aG \geq 0.06$ & $CC_1 = CC_2$	106	2481	0.810

Table A.4. Rules from a decision set trained without phenotype information. Table presenting the rules of a decision set trained with the parameters chosen in Table 6.1, excluding phenotype information (DS excl. Pheno). Rules are depicted with multiple conditions separated by &. Each condition can be either a metapath shown in its compact representation together with a path reliability score minimum threshold (see Subsection 6.2.1 or Subsection 6.3.4) an optional unification condition (see Subsection 6.3.3) targeting entity types from the metapaths. We also provide the number of training instances covered during training (D: Disease-causing, N: Neutral) and the associated probability of the rule $p(l_D|r)$ (see Subsection 6.3.5).

Rule	D coverage	N coverage	$p(l_D \mid r)$
$GaBPaG \geq 0.08$ & $GaP_1aG \geq 0.37$ & $GpGaP_2aG \geq 0.35$ & $P_1 = P_2$	81	12	0.999
$GaBPaG \geq 0.13$ & $GaPaG \geq 0.55$ & $GaPrPaG \geq 0.48$	94	6	0.999
$GaPaG \geq 0.55$ & $GaMFaG_1pG \geq 0.22$ & $GpGpG_2pG \geq 0.28$ & $G_1 = G_2$	81	12	0.999
$GaBP_1aG \geq 0.02$ & $GaPaG \geq 0.42$ & $GpGaBP_2aG \geq 0.09$ & $BP_1 = BP_2$	74	12	0.998
$GaCC_1aG \geq 0.08$ & $GaPaG \geq 0.50$ & $GaCC_2aGsG \geq 0.15$ & $CC_1 = CC_2$	88	17	0.998
$GaPaG \geq 0.43$ & $GaBPrBP_1aG \geq 0.32$ & $GpGaBP_2aG \geq 0.18$ & $BP_1 = BP_2$	72	13	0.998
$GaPaG \geq 0.53$ & $GaMFaG_1pG \geq 0.21$ & $GaPaG_2pG \geq 0.27$ & $G_1 = G_2$	91	17	0.998
$GaPaG \geq 0.53$ & $GaBPaG_1pG \geq 0.11$ & $GaPaG_2pG \geq 0.30$ & $G_1 = G_2$	110	18	0.998
$GaPaG \geq 0.54$ & $GpG_1aBPaG \geq 0.07$ & $GpG_2aMFaG \geq 0.19$ & $G_1 = G_2$	89	14	0.998
$GaBPaG \geq 0.15$ & $GaP_1aG \geq 0.06$ & $GaPrP_2aG \geq 0.26$ & $P_1 = P_2$	105	19	0.998
$GaPaG \geq 0.54$ & $GaMFaG_1pG \geq 0.22$ & $GpGeG_2pG \geq 0.32$ & $G_1 = G_2$	80	14	0.998
$GaPaG \geq 0.54$ & $GpG_1aMFaG \geq 0.22$ & $GpG_2eGpG \geq 0.33$ & $G_1 = G_2$	82	14	0.998
$GaPaG \geq 0.52$ & $GaP_1aGpG \geq 0.11$ & $GaPrP_2aG \geq 0.20$ & $P_1 = P_2$	143	44	0.997
$GaPaG \geq 0.50$ & $GaPrPaG \geq 0.45$	225	105	0.995
$GaBPaG \geq 0.00$ & $GpG_1aCCaG \geq 0.22$ & $GpG_2aPaG \geq 0.38$ & $G_1 = G_2$	78	58	0.993
$GaBPaG \geq 0.13$ & $GaPaG_1pG \geq 0.18$ & $GpGpG_2pG \geq 0.15$ & $G_1 = G_2$	87	126	0.986
$GaPaG \geq 0.24$ & $GaCC_1aGpG \geq 0.21$ & $GpGaCC_2aG \geq 0.21$ & $CC_1 = CC_2$	106	160	0.985
$GaPaG \geq 0.41$ & $GpGeG_1eG \geq 0.24$ & $GpGpG_2eG \geq 0.24$ & $G_1 = G_2$	117	180	0.985
$GaBPaG \geq 0.05$ & $GeG_1aCCaG \geq 0.20$ & $GeG_2aPaG \geq 0.44$ & $G_1 = G_2$	79	130	0.984
$GaCC_1aG \geq 0.07$ & $GaPaG \geq 0.20$ & $GsGaCC_2aG \geq 0.04$ & $CC_1 = CC_2$	100	189	0.981
$GaP_1aGpG \geq 0.14$ & $GaP_2rPaG \geq 0.07$ & $GpGaMFaG \geq 0.28$ & $P_1 = P_2$	128	258	0.980
$GaBPaG_1pG \geq 0.22$ & $GpGaPaG \geq 0.38$ & $GpG_2pG \geq 0.12$ & $G_1 = G_2$	83	166	0.980
$GaPaG \geq 0.18$ & $GaCCaG_1pG \geq 0.12$ & $GpG_2pGpG \geq 0.12$ & $G_1 = G_2$	82	183	0.978
$GaMFaG_1pG \geq 0.26$ & $GaPaG_2pG \geq 0.31$ & $GpGaMFaG \geq 0.34$ & $G_1 = G_2$	78	186	0.977
$GaBPaGsG \geq 0.14$ & $GpGaCCaG \geq 0.17$ & $GpGaPaG \geq 0.25$	78	182	0.977
$GaBPaG \geq 0.14$ & $GaBPaG_1eG \geq 0.34$ & $GpGpG_2eG \geq 0.04$ & $G_1 = G_2$	76	235	0.970
$GaPaG \geq 0.22$ & $GpGaCCaG \geq 0.20$	239	750	0.970
$GaPrPaG \geq 0.24$ & $GpGaPaG \geq 0.38$	235	750	0.969
$GaBPaGpG \geq 0.35$ & $GeG_1aBPaG \geq 0.14$ & $GeG_2aPaG \geq 0.34$ & $G_1 = G_2$	115	457	0.962
$GaBPaGpG \geq 0.44$ & $GpG_1aBPaG \geq 0.05$ & $GpG_2aMFaG \geq 0.19$ & $G_1 = G_2$	100	422	0.960
$GaCCrCCaG \geq 0.03$ & $GpG_1aCCaG \geq 0.20$ & $GpG_2aPaG \geq 0.34$ & $G_1 = G_2$	85	432	0.952
$GaCCrCCaG \geq 0.13$ & $GeGaP_1aG \geq 0.48$ & $GpGaP_2aG \geq 0.28$ & $P_1 = P_2$	96	692	0.933
$GeGaPaG \geq 0.59$ & $GpG_1aBPaG \geq 0.01$ & $GpG_2aCCaG \geq 0.11$ & $G_1 = G_2$	153	1203	0.927
$GaBPaGpG \geq 0.37$ & $GeGaBP_1aG \geq 0.10$ & $GpGaBP_2aG \geq 0.04$ & $BP_1 = BP_2$	116	984	0.922
$GaBPaGsG \geq 0.03$ & $GeGaCC_1aG \geq 0.27$ & $GpGaCC_2aG \geq 0.17$ & $CC_1 = CC_2$	72	636	0.919

Table A.5. Rules from a decision set trained with phenotype. Table presenting the rules of a decision set trained with the parameters chosen in Table 6.1, including phenotype information (DS incl. Pheno). Rules are depicted with multiple conditions separated by &. Each condition can be either a metapath shown in its compact representation together with a path reliability score minimum threshold (see Subsection 6.2.1 or Subsection 6.3.4) an optional unification condition (see Subsection 6.3.3) targeting entity types from the metapaths. We also provide the number of training instances covered during training (D: Disease-causing, N: Neutral) and the associated probability of the rule $p(l_D|r)$ (see Subsection 6.3.5).

A.2 Appendices: Figures

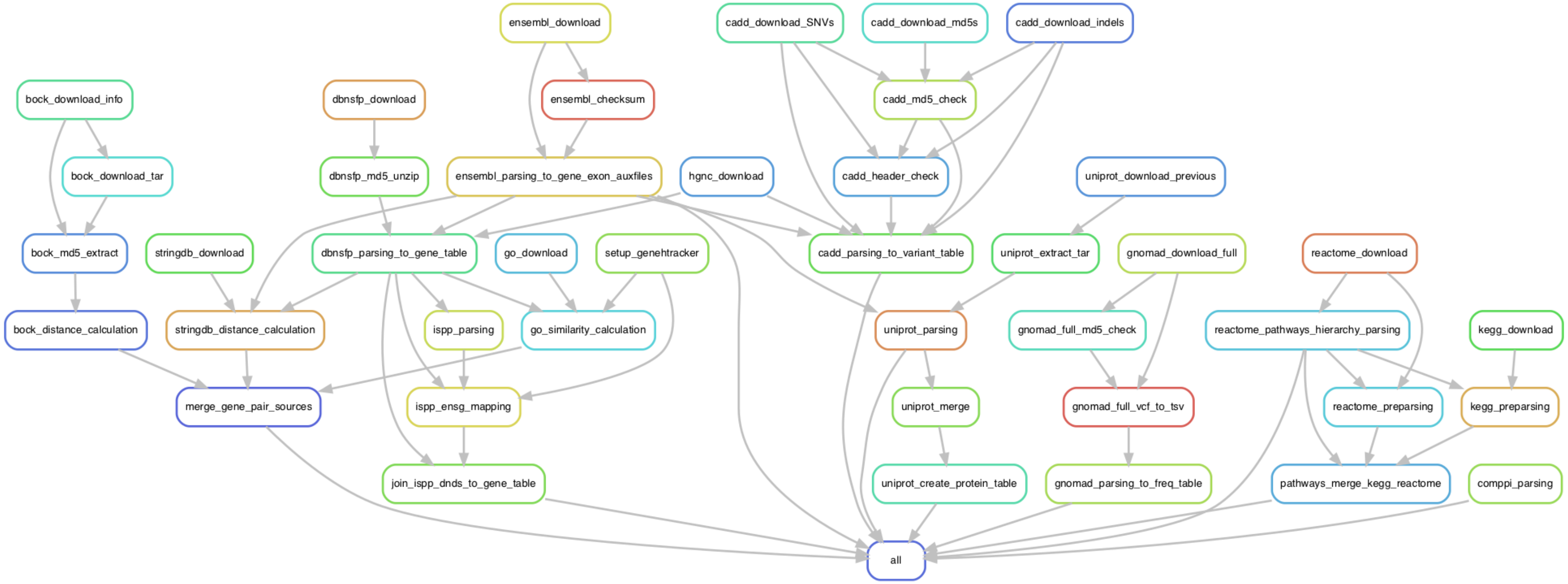


Figure A.1. Overview of the new automatised annotation database data integration pipeline, managed by the Snakemake workflow management system. Workflows are described via a set of rules with dependencies represented as arrows. The 'all' rules encompasses the whole workflow process (Credit: Emma Verkinderen).

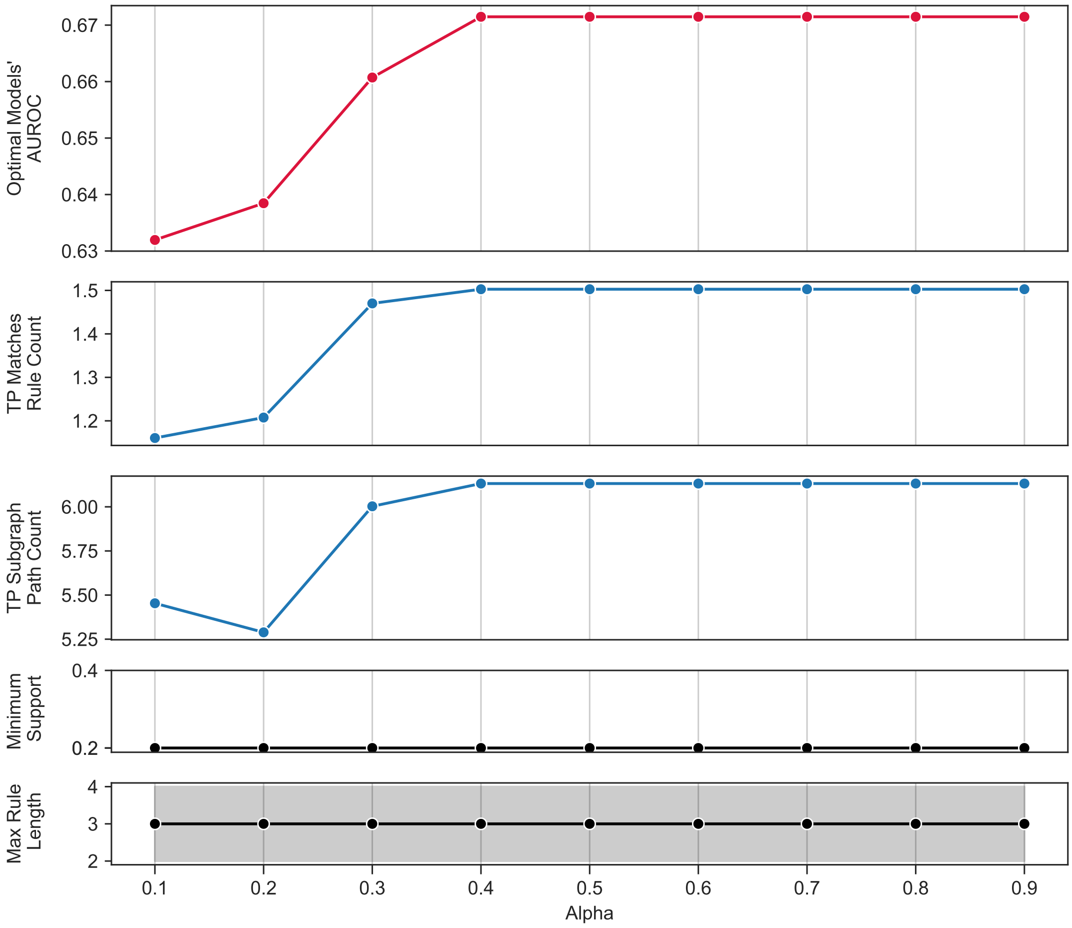


Figure A.2. Influence of decision model's alpha parameter tuning with path_cutoff=2 Evaluating decision set models with a fixed Path Cutoff = 2, over a spectrum of the decision model's alpha (α) values (0.1-0.9) and variable parameters - Minimum Support (`minsup_ratio`) {0.2, 0.4, 0.6, 0.8}, Max Rule Length (`max_rule_length`) {2,3,4}, we highlight the performance and characteristics of the optimal models for each alpha value. This evaluation is performed on the model excluding Phenotype, prone to study bias. Optimal models for each alpha are selected based on mean AUROC from a stratified cross-validation, keeping all within 0.01 of the highest performer to ensure comparable accuracy levels. **(A)** Mean AUROC of top-performing models evaluated via stratified cross-validation on test sets **(B)** Measures of explanation complexity for the associated models, calculated from predicted true positive (TP) gene pairs on the test sets: Rule Count measures the mean number of rules per match, and Path Count measures the mean number of paths across all rules, on average per match. **(C)** Parameter distribution for the corresponding top-performing models.

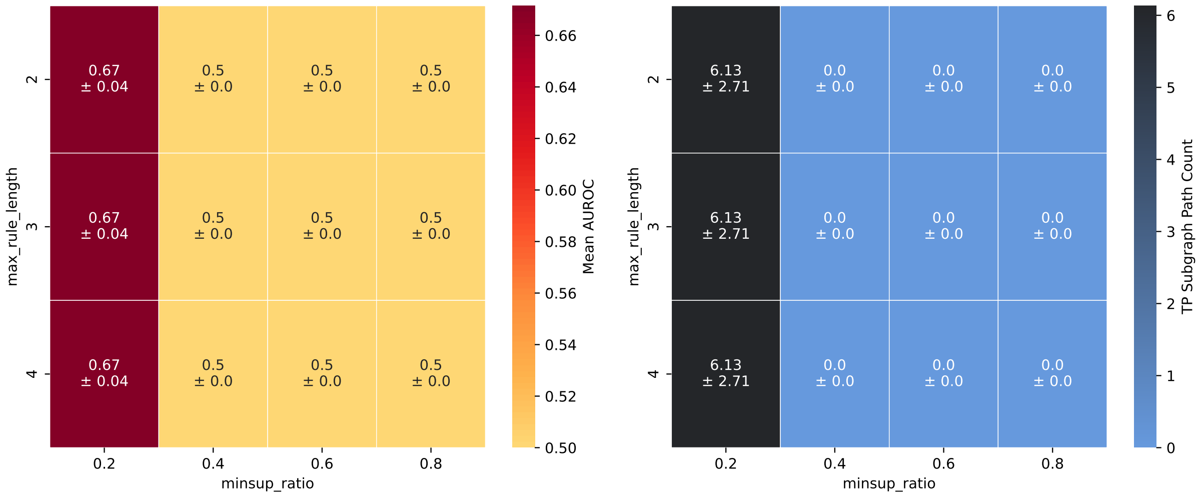


Figure A.3. Influence of minimum support and max rule length parameters tuning, with path_cutoff = 2. Heatmaps showing the influence of parameters `minsup_ratio` and `max_rule_length` on the performance and explanation complexity for decision models trained with paths excluding Phenotype and with fixed parameters: $\alpha = 0.5$ and `path_cutoff` $= 2$.

(A) Effect on the model performance, evaluated with mean AUROC on a stratified cross-validation. **(B)** Effect on the model explanation complexity, evaluated with the Path Count measure, measuring the mean number of paths across all rules, on average per True Positive matches.

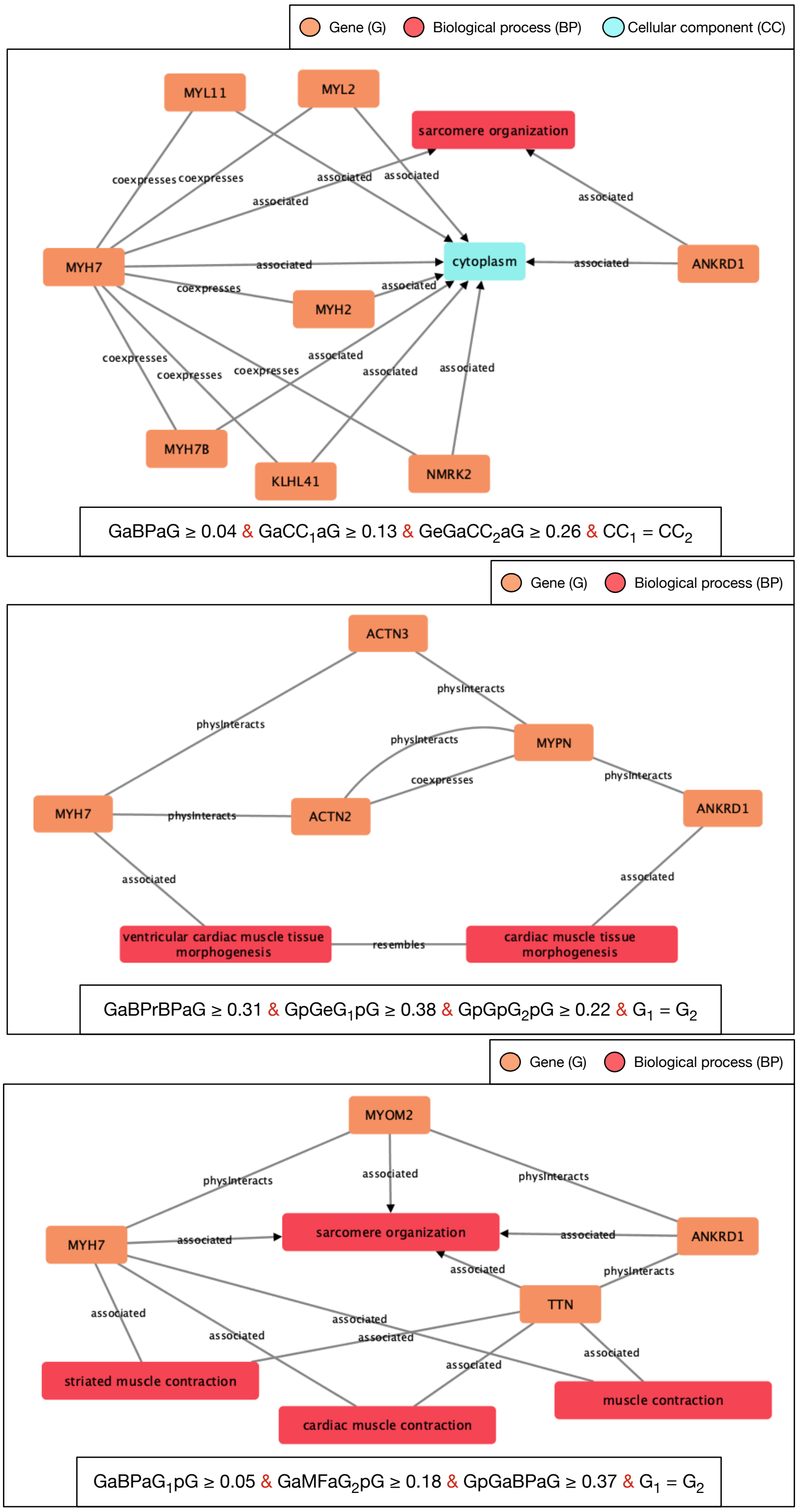
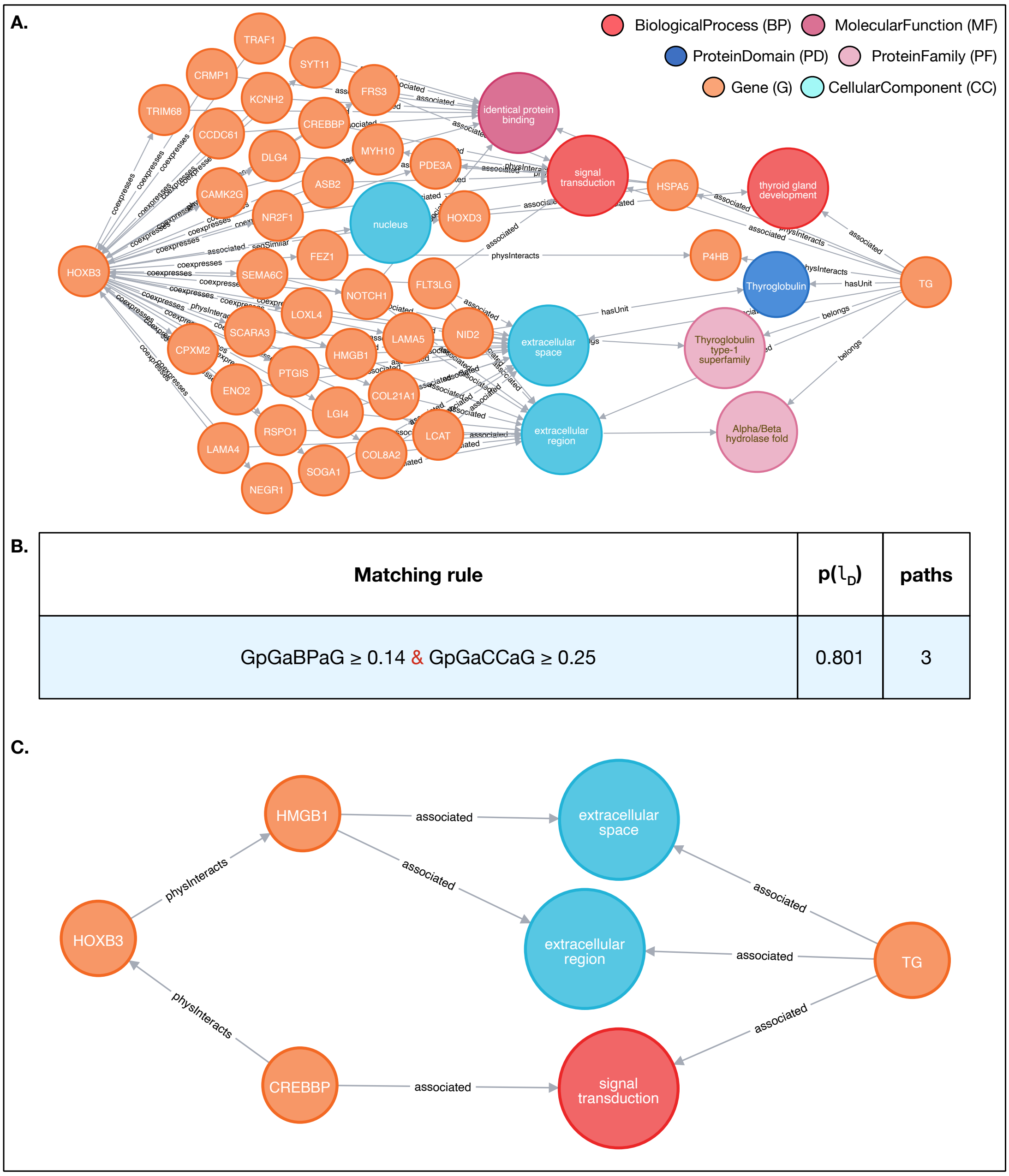


Figure A.4. The three figures below showcase the remaining graphical explanations for the 3^{rd}, 4^{th} and 5^{th} rules (in order of predicted probability), matching the digenic gene pair *MYH7-ANKRD1* in decreasing order of predicted probability (see Figure 6.14). Each rule is written in their abbreviated form (see Table 5.1) with its conditions separated by &. Indices for node types (*e.g.* G_1) are used in unification conditions (*e.g.* $G_1 = G_2$) to constrain entities to be the same across different metapaths. The numerical value associated to each metapath (*e.g.* ≥ 0.31) sets the path reliability threshold, which conditions the minimum path reliability score of all underlying paths.



Matching rule	$p(\ell_D)$	paths
GpGaBPaG ≥ 0.14 & GpGaCCaG ≥ 0.25	0.801	3

Figure A.5. Predictive explanation for the gene pair *HOXB3-TG*. This figure showcases the example of the digenic gene pair *HOXB3-TG*, causing congenital hypothyroidism, part of the independent test set and predicted as disease-causing by the *DS excl.Pheno* model with the lowest probability among the 12/15 predicted gene pairs and predicted as neutral by the *DS incl.Pheno* model. **(A)** Subgraph extracted by traversing all paths (excluding those traversing "Phenotype", "Disease" and "OligogenicCombination" nodes) of a length ≤ 3. A total of 102 paths, 48 nodes and 97 edges exists. **(B)** Matching rule from the *DS excl.Pheno* model, written in its abbreviated form (see Table 5.1) with its conditions separated by &. The numerical value associated with each metapath (*e.g.* ≥ 0.14) sets the path reliability threshold, which conditions the minimum path reliability score of all underlying paths. We display the number of paths obtained by querying the KG with the rule with that specific gene pair. **(C)** Returned explanation subgraph for the rule in (B). Note that, for this gene pair, this explanation provides limited understanding as both *CellularComponent* and *BiologicalProcess* are general terms and the traversed genes do not have any known associations with the disease.

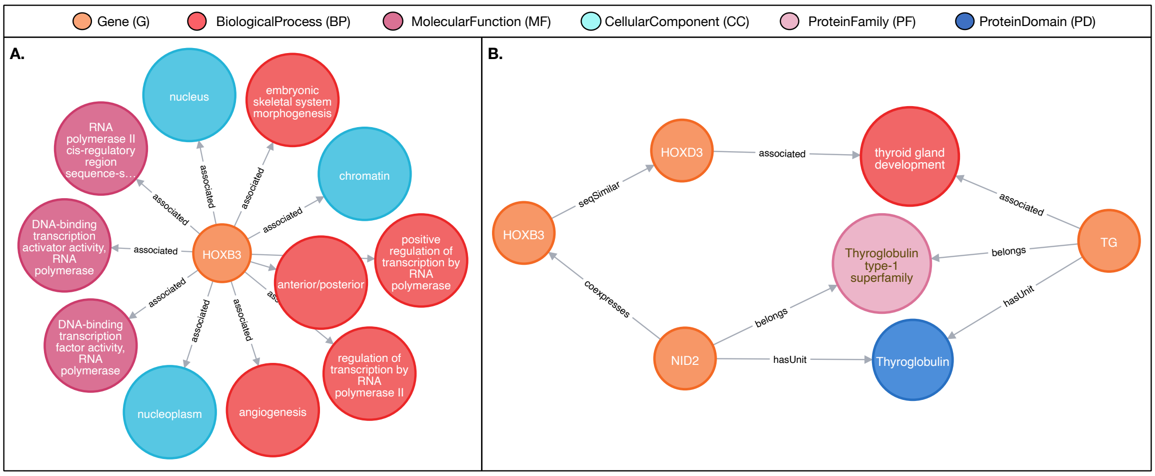

Gene (G)
BiologicalProcess (BP)
MolecularFunction (MF)
CellularComponent (CC)
ProteinFamily (PF)
ProteinDomain (PD)
A.
nucleus
embryonic skeletal system morphogenesis
RNA polymerase II cis-regulatory region sequence-s…
chromatin
DNA-binding transcription activator activity, RNA polymerase
HOXB3
positive regulation of transcription by RNA polymerase
anterior/posterior
DNA-binding transcription factor activity, RNA polymerase
regulation of transcription by RNA polymerase II
nucleoplasm
angiogenesis
associated
B.
HOXD3
associated
thyroid gland development
HOXB3
seqSimilar
associated
Thyroglobulin type-1 superfamily
belongs
TG
coexpresses
belongs
hasUnit
NID2
hasUnit
Thyroglobulin

Figure A.6. Additional details regarding the gene pair *HOXB3-TG* in BOCK. This figure showcases the example of the digenic gene pair *HOXB3-TG*, causing congenital hypothyroidism, part of the independent test set and predicted as disease-causing by the *DS excl.Pheno* model with the lowest probability among the 12/15 predicted gene pairs and predicted as neutral by the *DS incl.Pheno* model. **A.** All *HOXB3* neighbouring nodes connected with the "associated" edge type in BOCK. *HOXB3* is not associated with any known Phenotype. **B.** Selection of paths between *HOXB3-TG* traversing more relevant entities for understanding the physiopathology of hypothyroidism. This example illustrates that BOCK contains potentially more informative paths for this gene pair than those associated with the matching predictive rule as shown in Figure A.5.

Bibliography

[1] Dahm, R. Friedrich Miescher and the discovery of DNA. (2005).

[2] Watson, J. D. and Crick, F. H. (1953) Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature,* **171**(4356), 737–738.

[3] Meselson, M. and Stahl, F. W. (1958) The replication of DNA in Escherichia coli. *Proceedings of the National Academy of Sciences,* **44**(7), 671–682.

[4] NIRENBERG, M. W. and MATTHAEI, J. H. (1961) The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides.. *Proceedings of the National Academy of Sciences of the United States of America,* **47**(10), 1588–1602.

[5] Jacob, F. and Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. (1961).

[6] Crick, F. (1970) Central dogma of molecular biology. *Nature,* **227**(5258), 561–563.

[7] Sanger, F. and Coulson, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology,* **94**(3), 441–448.

[8] Shendure, J. and Ji, H. Next-generation DNA sequencing. (2008).

[9] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**(6822), 860–921.

[10] Brittain, H. K., Scott, R., and Thomas, E. (2017) The rise of the genome and personalised medicine. *Clinical Medicine, Journal of the Royal College of Physicians of London,* **17**(6), 545–551.

[11] Speicher, M. R., Antonarakis, S. E., and Motulsky, A. G. (2010) Vogel and Motulsky's human genetics: Problems and approaches (fourth edition), , .

[12] Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., and Mattick, J. S. Non-coding RNAs: Regulators of disease. (2010).

[13] Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America,* **104**(49), 19428–19433.

[14] Nilsen, T. W. and Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. (2010).

[15] Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., et al. (2012) An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature,* **489**(7414), 57.

[16] Vallender, E. J. and Lahn, B. T. Positive selection on the human genome. (2004).

[17] Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. Human genetic variation and its contribution to complex traits. (2009).

[18] Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., et al. A global reference for human genetic variation. (2015).

[19] MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science,* **335**(6070), 823–828.

[20] Hunt, S. E., Moore, B., Amode, R. M., Armean, I. M., Lemos, D., Mushtaq, A., Parton, A., Schuilenburg, H., Szpak, M., Thormann, A., et al. (2022) Annotating and prioritizing genomic variants using the Ensembl Variant Effect PredictorA tutorial. *Human Mutation,* **43**(8), 986–997.

[21] Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature,* **581**(7809), 434–443.

[22] Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'Ang, L. Y., Huang, W., Liu, B., Shen, Y., Tam, P. K. H., et al. (2003) The international HapMap project. *Nature,* **426**(6968), 789–796.

[23] Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R., Xu, C., Futema, M., Lawson, D., et al. (2015) The UK10K project identifies rare variants in health and disease. *Nature,* **526**(7571), 82–89.

[24] Montserrat Moliner, A. and Waligora, J. (2017) The European Union policy in the field of rare diseases. In *Advances in Experimental Medicine and Biology* Vol. 1031, pp. 561–587.

[25] Boycott, K. M., Rath, A., Chong, J. X., Hartley, T., Alkuraya, F. S., Baynam, G., Brookes, A. J., Brudno, M., Carracedo, A., den Dunnen, J. T., et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. (2017).

[26] Maroilley, T. and Tarailo-Graovac, M. (2019) Uncovering Missing Heritability in Rare Diseases. *Genes,* **10**(4), 1–18.

[27] Ferreira, C. R. The burden of rare diseases. (2019).

[28] Greene, D., Pirri, D., Frudd, K., Sackey, E., Al-Owain, M., Giese, A. P., Ramzan, K., Riaz, S., Yamanaka, I., Boeckx, N., et al. (2023) Genetic association analysis of 77,539 genomes reveals rare disease etiologies. *Nature Medicine,* **29**(3), 679–688.

[29] Seaby, E. G. and Ennis, S. (2020) Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. *Briefings in Functional Genomics,* **19**(4), 243–258.

[30] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009) Finding the missing heritability of complex diseases. *Nature,* **461**(7265), 747–753.

[31] Zlotogora, J. Penetrance and expressivity in the molecular age. (2003).

[32] Myers, R. H. (2004) Huntington's Disease Genetics. *NeuroRx,* **1**(2), 255–262.

[33] Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018) ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research,* **46**(D1), D1062–D1067.

[34] Amberger, J. S. and Hamosh, A. (2017) Searching online mendelian inheritance in man (OMIM): A knowledgebase of human genes and genetic phenotypes. *Current Protocols in Bioinformatics,* **2017**(1), 1.2.1–1.2.12.

[35] Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., and Rath, A. (2020) Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics,* **28**(2), 165–173.

[36] Martin, H. C., Jones, W. D., McIntyre, R., Sanchez-Andrade, G., Sanderson, M., Stephenson, J. D., Jones, C. P., Handsaker, J., Gallone, G., Bruntraeger, M., et al. (2018) Quantifying the contribution of recessive coding variation to developmental disorders. *Science,* **362**(6419), 1161–1164.

[37] Ramoni, R. B., Mulvihill, J. J., Adams, D. R., Allard, P., Ashley, E. A., Bernstein, J. A., Gahl, W. A., Hamid, R., Loscalzo, J., McCray, A. T., et al. The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. (2017).

[38] Garcia, F. A. d. O., de Andrade, E. S., and Palmero, E. I. Insights on variant analysis in silico tools for pathogenicity prediction. (2022).

[39] Kumar, P., Henikoff, S., and Ng, P. C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols,* **4**(7), 1073–1082.

[40] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. A method and server for predicting damaging missense mutations. (2010).

[41] Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics,* **46**(3), 310–315.

[42] Moreau, Y. and Tranchevent, L. C. (2012) Computational tools for prioritizing candidate genes: Boosting disease gene discovery. *Nature Reviews Genetics,* **13**(8), 523–536.

[43] Tranchevent, L. C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., De Moor, B., Aerts, S., and Moreau, Y. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species.. *Nucleic acids research,* **36**(Web Server issue).

[44] Sifrim, A., Popovic, D., Tranchevent, L. C., Ardeshirdavani, A., Sakai, R., Konings, P., Vermeesch, J. R., Aerts, J., De Moor, B., and Moreau, Y. (2013) EXtasy: Variant prioritization by genomic data fusion. *Nature Methods,* **10**(11), 1083–1086.

[45] Ghiassian, S. D., Menche, J., and Barabási, A. L. (2015) A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLoS Computational Biology,* **11**(4).

[46] Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. Rare-variant association analysis: Study designs and statistical tests. (2014).

[47] Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A. S., and Goldstein,

D. B. (2019) Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics,* **20**(12), 747–759.

[48] Hasegawa, T., Kojima, K., Kawai, Y., Misawa, K., Mimori, T., and Nagasaki, M. (2016) AP-SKAT: Highly-efficient genome-wide rare variant association test. *BMC Genomics,* **17**(1), 745.

[49] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. (2008).

[50] Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., Zhou, H., Tian, L., Prakash, O., Lemire, M., et al. (2016) Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature Biotechnology,* **34**(5), 531–538.

[51] Posey, J. E., Harel, T., Liu, P., Rosenfeld, J. A., James, R. A., Coban Akdemir, Z. H., Walkiewicz, M., Bi, W., Xiao, R., Ding, Y., et al. (2017) Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *New England Journal of Medicine,* **376**(1), 21–31.

[52] Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C., and Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease. (2013).

[53] Badano, J. L. and Katsanis, N. Beyond mendel: An evolving view of human genetic disease transmission. (2002).

[54] Katsanis, N. (2016) The continuum of causality in human genetic disorders. *Genome Biology,* **17**(1).

[55] Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell,* **169**(7), 1177–1186.

[56] Kousi, M. and Katsanis, N. (2015) Genetic modifiers and oligogenic inheritance. *Cold Spring Harbor Perspectives in Medicine,* **5**(6), 1–22.

[57] Robinson, J. F. and Katsanis, N. (2010) Oligogenic disease. In Speicher, M. R., Motulsky, A. G., and Antonarakis, S. E., (eds.), *Vogel and Motulsky's Human Genetics: Problems and Approaches (Fourth Edition)*, pp. 243–262 Springer Berlin Heidelberg Berlin, Heidelberg.

[58] Lupski, J. R. (2012) Digenic inheritance and Mendelian disease. *Nature Genetics,* **44**(12), 1291–1292.

[59] Deltas, C. (2018) Digenic inheritance and genetic modifiers. *Clinical Genetics,* **93**(3), 429–438.

[60] Weiler, C. A. and Drumm, M. L. Genetic influences on cystic fibrosis lung disease severity. (2013).

[61] Rahit, K. M. H. and Tarailo-Graovac, M. (2020) Genetic Modifiers and Rare Mendelian Disease. *Genes,* **11**(3).

[62] Kajiwara, K., Berson, E. L., and Dryja, T. P. (1994) Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science,* **264**(5165), 1604–1608.

[63] Zhu, X., Need, A. C., Petrovski, S., and Goldstein, D. B. One gene, many neuropsychiatric disorders: Lessons from Mendelian diseases. (2014).

[64] Schäffer, A. A. (2013) Digenic inheritance in medical genetics. *Journal of Medical Genetics,* **50**(10), 641–652.

[65] Katsanis, N. (2004) The oligogenic properties of Bardet-Biedl syndrome. *Human Molecular Genetics,* **13**(REV. ISS. 1), R65—-R71.

[66] M'Hamdi, O., Ouertani, I., and Chaabouni-Bouhamed, H. Update on the genetics of bardet-biedl syndrome. (2014).

[67] Zheng, Q. Y., Yan, D., Ouyang, X. M., Du, L. L., Yu, H., Chang, B., Johnson, K. R., and Liu, X. Z. (2005) Digenic inheritance of deafness caused by mutations in genes encoding cadherin 23 and protocadherin 15 in mice and humans. *Human molecular genetics,* **14**(1), 103.

[68] Gazzo, A. M., Daneels, D., Cilia, E., Bonduelle, M., Abramowicz, M., Van Dooren, S., Smits, G., and Lenaerts, T. (2016) DIDA: A curated and annotated digenic diseases database. *Nucleic Acids Research,* **44**(D1), D900–D907.

[69] Gifford, C. A., Ranade, S. S., Samarakoon, R., Salunga, H. T., Yvanka De Soysa, T., Huang, Y., Zhou, P., Elfenbein, A., Wyman, S. K., Bui, Y. K., et al. (2019) Oligogenic inheritance of a human heart disease involving a genetic modifier. *Science,* **364**(6443), 865–870.

[70] Wang, H., Kong, X., Pei, Y., Cui, X., Zhu, Y., He, Z., Wang, Y., Zhang, L., Zhuo, L., Chen, C., et al. (2020) Mutation spectrum analysis of 29 causative genes in 43 Chinese patients with congenital hypothyroidism. *Molecular Medicine Reports,* **22**(1), 297–309.

[71] Nachtegael, C., Gravel, B., Dillen, A., Smits, G., Nowé, A., Papadimitriou, S., and Lenaerts, T. (2022) Scaling up oligogenic diseases research with OLIDA: The Oligogenic Diseases Database. *Database,* **2022**(March), 1–15.

[72] Moore, J. H. A global view of epistasis. (2005).

[73] Domingo, J., Baeza-Centurion, P., and Lehner, B. The Causes and Consequences of Genetic Interactions (Epistasis). (2019).

[74] Lehner, B. (2011) Molecular mechanisms of epistasis within and between genes. *Trends in Genetics,* **27**(8), 323–331.

[75] Draper, N., Walker, E. A., Bujalska, I. J., Tomlinson, J. W., Chalder, S. M., Arlt, W., Lavery, G. G., Bedendo, O., Ray, D. W., Laing, I., et al. (2003) Mutations in the genes encoding 11β-hydroxysteroid dehydrogenase type 1 and hexose-6-phosphate

dehydrogenase interact to cause cortisone reductase deficiency. *Nature Genetics,* **34**(4), 434–439.

[76] Steingrímsson, E., Tessarollo, L., Pathak, B., Hou, L., Arnheiter, H., Copeland, N. G., and Jenkins, N. A. (2002) Mitf and Tfe3, two members of the Mitf-Tfe family of bHLH-Zip transcription factors, have important but functionally redundant roles in osteoclast development. *Proceedings of the National Academy of Sciences of the United States of America,* **99**(7), 4477–4482.

[77] Cooper-Knock, J., Robins, H., Niedermoser, I., Wyles, M., Heath, P. R., Higginbottom, A., Walsh, T., Kazoka, M., Al Kheifat, A., Al-Chalabi, A., et al. (2017) Targeted genetic screen in amyotrophic lateral sclerosis reveals novel genetic variants with synergistic effect on clinical phenotype. *Frontiers in Molecular Neuroscience,* **10**(November), 1–11.

[78] Ameratunga, R., Woon, S. T., Bryant, V. L., Steele, R., Slade, C., Leung, E. Y., and Lehnert, K. (2018) Clinical implications of digenic inheritance and epistasis in primary immunodeficiency disorders. *Frontiers in Immunology,* **8**(JAN), 1–8.

[79] Badano, J. L., Leitch, C. C., Ansley, S. J., May-Simera, H., Lawson, S., Lewis, R. A., Beales, P. L., Dietz, H. C., Fisher, S., and Katsanis, N. (2006) Dissection of epistasis in oligogenic Bardet-Biedl syndrome. *Nature,* **439**(7074), 326–330.

[80] Papadimitriou, S., Gravel, B., Nachtegael, C., De Baere, E., Loeys, B., Vikkula, M., Smits, G., and Lenaerts, T. (2023) Toward reporting standards for the pathogenicity of variant combinations involved in multilocus/oligogenic diseases. *Human Genetics and Genomics Advances,* **4**(1), 100165.

[81] Gazzo, A., Raimondi, D., Daneels, D., Moreau, Y., Smits, G., Van Dooren, S., and Lenaerts, T. (2017) Understanding mutational effects in digenic diseases. *Nucleic Acids Research,* **45**(15), e140.

[82] Versbraegen, N., Fouché, A., Nachtegael, C., Papadimitriou, S., Gazzo, A., Smits, G., and Lenaerts, T. (2019) Using game theory and decision decomposition to effec-

tively discern and characterise bi-locus diseases. *Artificial Intelligence in Medicine,* **99**.

[83] Papadimitriou, S., Gazzo, A., Versbraegen, N., Nachtegael, C., Aerts, J., Moreau, Y., Van Dooren, S., Nowé, A., Smits, G., and Lenaerts, T. (2019) Predicting disease-causing variant combinations. *Proceedings of the National Academy of Sciences of the United States of America,* **116**(24), 11878–11887.

[84] Boudellioua, I., Kulmanov, M., Schofield, P. N., Gkoutos, G. V., and Hoehndorf, R. (2018) OligoPVP: Phenotype-driven analysis of individual genomic information to prioritize oligogenic disease variants. *Scientific Reports,* **8**(1).

[85] Mukherjee, S., Cogan, J. D., Newman, J. H., Phillips, J. A., Hamid, R., Meiler, J., and Capra, J. A. (2021) Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network. *American Journal of Human Genetics,* **108**(10), 1946–1963.

[86] Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019) CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research,* **47**(D1), D886–D894.

[87] Shihab, H. A., Rogers, M. F., Campbell, C., and Gaunt, T. R. (2017) HIPred: an integrative approach to predicting haploinsufficient genes. *Bioinformatics (Oxford, England),* **33**(12), 1751–1757.

[88] Cassereau, J., Casasnovas, C., Gueguen, N., Malinge, M. C., Guillet, V., Reynier, P., Bonneau, D., Amati-Bonneau, P., Banchs, I., Volpini, V., et al. (2011) Simultaneous MFN2 and GDAP1 mutations cause major mitochondrial defects in a patient with CMT. *Neurology,* **76**(17), 1524–1526.

[89] Tan, V. H., Duff, H., Kuriachan, V., and Gerull, B. (2014) Congenital long QT syndrome: Severe Torsades de pointes provoked by epinephrine in a digenic mutation carrier. *Heart and Lung: Journal of Acute and Critical Care,* **43**(6), 541–545.

[90] Kelberman, D., Islam, L., Holder, S. E., Jacques, T. S., Calvas, P., Hennekam, R. C., Nischal, K. K., and Sowden, J. C. (2011) Digenic inheritance of mutations in FOXC1 and PITX2: Correlating transcription factor function and axenfeld-rieger disease severity. *Human Mutation,* **32**(10), 1144–1152.

[91] Ito, T., Young, M. J., Li, R., Jain, S., Wernitznig, A., Krill-Burger, J. M., Lemke, C. T., Monducci, D., Rodriguez, D. J., Chang, L., et al. (2021) Paralog knockout profiling identifies DUSP4 and DUSP6 as a digenic dependence in MAPK pathway-driven cancers. *Nature Genetics,* **53**(12), 1664–1672.

[92] Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., et al. (2018) A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology,* **36**(10), 983.

[93] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature,* **596**(7873), 583–589.

[94] Lipton, Z. C. (2016) The Mythos of Model Interpretability. *Communications of the ACM,* **61**(10), 35–43.

[95] Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 2019 1:5,* **1**(5), 206–215.

[96] Doshi-Velez, F. and Kim, B. (2017) Towards A Rigorous Science of Interpretable Machine Learning.

[97] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., et al. (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion,* **58**, 82–115.

[98] Mittelstadt, B., Russell, C., and Wachter, S. (2019) Explaining explanations in AI. *FAT*2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency,* pp. 279–288.

[99] Lim, B. Y., Dey, A. K., and Avrahami, D. (2009) Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Conference on Human Factors in Computing Systems - Proceedings,* pp. 2119–2128.

[100] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984) Classification and regression trees. *Classification and Regression Trees,* pp. 1–358.

[101] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016) "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* **13-17-Augu**, 1135–1144.

[102] Lundberg, S. M. A Unified Approach to Interpreting Model Predictions. (2017).

[103] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR),* **51**(5).

[104] Gosiewska, A. and Biecek, P. (2019) Do Not Trust Additive Explanations.

[105] Cox, D. R., Hinkley, D. V., Reid, N., Rubin, D. B., and Silverman, B. W. (1989) Generalized Linear Models. *Regression Analysis with Application G.B. Wetherill,* (2), 28.

[106] Cohen, W. W. (1995) Fast Effective Rule Induction. In *Machine Learning Proceedings 1995* pp. 115–123.

[107] Cover, T. M. and Hart, P. E. (1967) Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory,* **13**(1), 21–27.

[108] Thompson, M., Duda, R. O., and Hart, P. E. (1974) Pattern Classification and Scene Analysis. *Leonardo,* **7**(4), 370.

[109] Wachter, S., Mittelstadt, B., and Russell, C. (2018) COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR. *Harvard Journal of Law & Technology,* **31**(2).

[110] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018) Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence,* **32**(1), 1527–1535.

[111] Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015) Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *http://dx.doi.org/10.1080/10618600.2014.907095,* **24**(1), 44–65.

[112] Watson, D. S. (2022) Interpretable machine learning for genomics. *Human Genetics,* **141**(9), 1499–1513.

[113] Vayena, E., Blasimme, A., and Cohen, I. G. (2018) Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine,* **15**(11), e1002689.

[114] Yu, M. K., Ma, J., Fisher, J., Kreisberg, J. F., Raphael, B. J., and Ideker, T. Visible Machine Learning for Biomedicine. (2018).

[115] Azodi, C. B., Tang, J., and Shiu, S. H. Opening the Black Box: Interpretable Machine Learning for Geneticists. (2020).

[116] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P. M., Zietz, M., Hoffman, M. M., et al. (2018) Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface,* **15**(141).

[117] Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., and Sölkner, J. (2013) Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics,* **4**(DEC), 58395.

[118] Levy, J. J., Titus, A. J., Petersen, C. L., Chen, Y., Salas, L. A., and Christensen, B. C. (2020) MethylNet: An automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinformatics,* **21**(1), 1–15.

[119] Drouin, A., Letarte, G., Raymond, F., Marchand, M., Corbeil, J., and Laviolette, F. (2019) Interpretable genotype-to-phenotype classifiers with performance guarantees. *Scientific Reports,* **9**(1), 1–13.

[120] Anguita-Ruiz, A., Segura-Delgado, A., Alcalá, R., Aguilera, C. M., and Alcalá-Fdez, J. (2020) EXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. *PLoS Computational Biology,* **16**(4), e1007792.

[121] Garbulowski, M., Smolinska, K., Diamanti, K., Pan, G., Maqbool, K., Feuk, L., and Komorowski, J. (2021) Interpretable Machine Learning Reveals Dissimilarities Between Subtypes of Autism Spectrum Disorder. *Frontiers in Genetics,* **12**, 618277.

[122] Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. (2019) What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Proceedings of Machine Learning Research,*.

[123] Yang, J., Song, H., Cao, K., Song, J., and Zhou, J. (2018) Comprehensive analysis of Helicobacter pylori infection-associated diseases based on miRNA-mRNA interaction network. *Briefings in Bioinformatics,* **20**(4), 1492–1501.

[124] Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., et al. (2021) Knowledge graphs. *ACM Computing Surveys,* **54**(4).

[125] Krötzsch, M., Stepanova, D., and Goos, G. (2019) Reasoning Web. Explainable Artificial Intelligence, Vol. 11810 of Lecture Notes in Computer Science, Springer International Publishing, Cham.

[126] Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011) Network medicine: A network-based approach to human disease. *Nature Reviews Genetics,* **12**(1), 56–68.

[127] Dessimoz, C. and Walker, J. M. (2017) The Gene Ontology Handbook, Vol. 1446, , .

[128] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. Gene ontology: Tool for the unification of biology. (2000).

[129] Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., et al. (2021) The human phenotype ontology in 2021. *Nucleic Acids Research,* **49**(D1), D1207–D1217.

[130] Schriml, L. M., Munro, J. B., Schor, M., Olley, D., McCracken, C., Felix, V., Baron, J. A., Jackson, R., Bello, S. M., Bearer, C., et al. (2022) The Human Disease Ontology 2022 update. *Nucleic Acids Research,* **50**(D1), D1255–D1261.

[131] Fensel, D., imek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., and Wahler, A. (2020) Knowledge Graphs, , .

[132] Nicholson, D. N. and Greene, C. S. Constructing knowledge graphs and their biomedical applications. (2020).

[133] Tiddi, I. and Schlobach, S. (2022) Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence,* **302**, 103627.

[134] Charitou, T., Bryan, K., and Lynn, D. J. (2016) Using biological networks to integrate, visualize and analyze genomics data. *Genetics Selection Evolution,* **48**(1), 1–12.

[135] Lee, B., Zhang, S., Poleksic, A., and Xie, L. Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis. (2020).

[136] Gligorijević, V. and Pržulj, N. (2015) Methods for biological data integration: perspectives and challenges. *Journal of The Royal Society Interface,* **12**(112).

[137] Fortunato, S. and Hric, D. (2016) Community detection in networks: A user guide. *Physics Reports,* **659**, 1–44.

[138] Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., et al. Visualization of omics data for systems biology. (2010).

[139] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America,* **102**(43), 15545–15550.

[140] Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falcão, A. O., and Couto, F. M. (2008) Metrics for GO based protein semantic similarity: A systematic evaluation. *BMC Bioinformatics,* **9**(SUPPL. 5).

[141] Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009) Semantic Similarity in Biomedical Ontologies. *PLOS Computational Biology,* **5**(7), e1000443.

[142] Tran, V. D., Sperduti, A., Backofen, R., Backofen, R., and Costa, F. (2020) Heterogeneous networks integration for disease-gene prioritization with node kernels. *Bioinformatics,* **36**(9), 2649–2656.

[143] Rao, A., Vg, S., Joseph, T., Kotte, S., Sivadasan, N., and Srinivasan, R. (2018) Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. *BMC Medical Genomics,* **11**(1), 1–12.

[144] Valentini, G., Paccanaro, A., Caniza, H., Romero, A. E., and Re, M. (2014) An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artificial Intelligence in Medicine,* **61**(2), 63–78.

[145] Banerjee, J., Taroni, J. N., Allaway, R. J., Prasad, D. V., Guinney, J., and Greene, C. (2023) Machine learning in rare disease. *Nature Methods 2023,* pp. 1–12.

[146] Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G. D., and

Morris, Q. (2018) GeneMANIA update 2018. *Nucleic Acids Research,* **46**(W1), W60–W64.

[147] Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., and Sharan, R. (2019) To embed or not: Network embedding as a paradigm in computational biology. *Frontiers in Genetics,* **10**(MAY), 381.

[148] Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E., and Baudot, A. (2019) Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics,* **35**(3), 497–505.

[149] Rajabi, E. and Etminani, K. (2022) Knowledge-graph-based explainable AI: A systematic review. *Journal of Information Science,*.

[150] Lv, X., Cao, Y., Hou, L., Li, J., Liu, Z., Zhang, Y., and Dai, Z. (2021) Is Multi-Hop Reasoning Really Explainable? Towards Benchmarking Reasoning Interpretability. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings,* pp. 8899–8911.

[151] Liu, Y., Hildebrandt, M., Joblin, M., Ringsquandl, M., Raissouni, R., and Tresp, V. (2021) Neural Multi-hop Reasoning with Logical Rules on Biomedical Knowledge Graphs. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* **12731 LNCS**, 375–391.

[152] Babur, Ö., Luna, A., Korkut, A., Durupinar, F., Siper, M. C., Dogrusoz, U., Vaca Jacome, A. S., Peckner, R., Christianson, K. E., Jaffe, J. D., et al. (2021) Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *Patterns,* **2**(6).

[153] Chindelevitch, L., Ziemek, D., Enayetallah, A., Randhawa, R., Sidders, B., Brockel, C., and Huang, E. S. (2012) Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics,* **28**(8), 1114–1121.

[154] Domingo-Fernandez, D., Gadiya, Y., Patel, A., Mubeen, S., Rivas-Barragan, D., Diana, C. W., Misra, B. B., Healey, D., Rokicki, J., and Colluru, V. (2022) Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. *PLoS computational biology,* **18**(2).

[155] Ortona, S., Meduri, V. V., and Papotti, P. (2018) RuDiK: Rule discovery in knowledge bases. *Proceedings of the VLDB Endowment,* **11**(12), 1946–1949.

[156] Meilicke, C., Chekol, M. W., Ruffinelli, D., and Stuckenschmidt, H., Anytime bottom-up rule learning for knowledge graph completion. Technical report, IJCAI Macau (2019).

[157] Galárraga, L., Teflioudi, C., Hose, K., and Suchanek, F. M. (2015) Fast rule mining in ontological knowledge bases with AMIE+. *VLDB Journal,* **24**(6), 707–730.

[158] Rossi, A., Barbosa, D., Firmani, D., Matinata, A., and Merialdo, P. (2021) Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data,* **15**(2).

[159] Meilicke, C., Fink, M., Wang, Y., Ruffinelli, D., Gemulla, R., and Stuckenschmidt, H. (2018) Fine-grained evaluation of rule- and embedding-based systems for knowledge graph completion. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* Vol. 11136 LNCS, .

[160] Sun, Y., Han, J., Yan, X., Yu, P. S., and Wu, T., Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. Technical Report 11, VLDB Seattle (2011).

[161] Meng, C., Cheng, R., Maniu, S., Senellart, P., and Zhang, W. (2015) Discovering meta-paths in large heterogeneous information networks. *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web,* pp. 754–764.

[162] Himmelstein, D. S. and Baranzini, S. E. (2015) Heterogeneous Network Edge Pre-

diction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS Computational Biology,* **11**(7).

[163] Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S. E. (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife,* **6**, 1–35.

[164] Renaux, A., Terwagne, C., Cochez, M., Tiddi, I., Nowé, A., and Lenaerts, T. (2023) A Knowledge Graph approach to predict and interpret Disease-causing Gene Interactions. *BMC Bioinformatics,*.

[165] Renaux, A., Papadimitriou, S., Versbraegen, N., Nachtegael, C., Boutry, S., Nowé, A., Smits, G., and Lenaerts, T. (2019) ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Research,* **47**(W1), W93–W98.

[166] Bohannan, Z. S. and Mitrofanova, A. (2019) Calling Variants in the Clinic: Informed Variant Calling Decisions Based on Biological, Clinical, and Laboratory Variables. *Computational and Structural Biotechnology Journal,* **17**, 561–569.

[167] Caspar, S. M., Dubacher, N., Kopps, A. M., Meienberg, J., Henggeler, C., and Matyas, G. Clinical sequencing: From raw data to diagnosis with lifetime value. (2018).

[168] Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* **25**(14), 1754–1760.

[169] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* **20**(9), 1297–1303.

[170] Seaby, E. G., Pengelly, R. J., and Ennis, S. (2016) Exome sequencing explained: A practical guide to its clinical application. *Briefings in Functional Genomics,* **15**(5), 374–384.

[171] den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., Mcgowan-Jordan, J., Roux, A. F., Smith, T., Antonarakis, S. E., and Taschner, P. E. (2016) HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation,* **37**(6), 564–569.

[172] The SAM/BAM Format Specification Working Group (2021) The Variant Call Format Specification. *The SAM/BAM Format Specification Working Group,* (May), 1–36.

[173] Manolio, T. A., Brooks, L. D., and Collins, F. S. A HapMap harvest of insights into the genetics of common disease. (2008).

[174] Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Ridwan Amode, M., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., et al. (2021) Ensembl 2021. *Nucleic Acids Research,* **49**(D1), D884–D891.

[175] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001) DbSNP: The NCBI database of genetic variation. *Nucleic Acids Research,* **29**(1), 308–311.

[176] Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Research,* **43**(D1), D789–D798.

[177] Tweedie, S., Braschi, B., Gray, K., Jones, T. E., Seal, R. L., Yates, B., and Bruford, E. A. (2021) Genenames.org: The HGNC and VGNC resources in 2021. *Nucleic Acids Research,* **49**(D1), D939–D946.

[178] Bateman, A. (2019) UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research,* **47**(D1), D506–D515.

[179] Itan, Y., Shang, L., Boisson, B., Patin, E., Bolze, A., Moncada-Vélez, M., Scott, E., Ciancanelli, M. J., Lafaille, F. G., Markle, J. G., et al. (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings*

of the National Academy of Sciences of the United States of America, **112**(44), 13615–13620.

[180] Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., and Goldstein, D. B. (2013) Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics,* **9**(8), e1003709.

[181] Hsu, J. S., Kwan, J. S., Pan, Z., Garcia-Barcelo, M. M., Sham, P. C., and Li, M. (2016) Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein coding genes. *Bioinformatics,* **32**(20), 3065–3071.

[182] Huang, N., Lee, I., Marcotte, E. M., and Hurles, M. E. (2010) Characterising and predicting haploinsufficiency in the human genome. *PLoS Genetics,* **6**(10), 1–11.

[183] Georgi, B., Voight, B. F., and Bućan, M. (2013) From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLOS Genetics,* **9**(5), e1003484.

[184] Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M., Amode, R., Brent, S., et al. (2016) Ensembl comparative genomics resources. *Database : the journal of biological databases and curation,* **2016**.

[185] Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine,* **12**(1), 1–8.

[186] Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H. Y., Dosztanyi, Z., El-Gebali, S., Fraser, M., et al. (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research,* **45**(D1), D190–D199.

[187] Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. (2019) CORUM: The comprehensive resource of mammalian protein complexes - 2019. *Nucleic Acids Research,* **47**(D1), D559–D563.

[188] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Research,* **46**(D1), D649–D655.

[189] The Gene Ontology, C., That, I., Acencio, M., Lægreid, A., Kuiper, M., and Among, O. (2019) The Gene Ontology Resource: 20 years and still GOing strong.. *Nucleic Acids Research,* **8**(47), D330—-D338.

[190] Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O., Danis, D., Gourdine, J. P., Gargano, M., Harris, N. L., Matentzoglu, N., McMurry, J. A., et al. (2019) Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research,* **47**(D1), D1018–D1027.

[191] Barabási, A.-l. and Oltvai, Z. N. (2004) Network biology: understanding the cell's functional organization. **5**(February).

[192] Calderone, A., Castagnoli, L., and Cesareni, G. Mentha: A resource for browsing integrated protein-interaction networks. (2013).

[193] Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., Kim, D. K., Kishore, N., Hao, T., et al. (2019) Network-based prediction of protein interactions. *Nature Communications,* **10**(1), 1–8.

[194] Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., et al. (2017) The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research,* **45**(D1), D362–D368.

[195] Lee, S., Zhang, C., Arif, M., Liu, Z., Benfeitas, R., Bidkhori, G., Deshmukh, S., Al Shobky, M., Lovric, A., Boren, J., et al. (2018) TCSBN: A database of tissue and cancer specific biological networks. *Nucleic Acids Research,* **46**(D1), D595–D600.

[196] Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., et al. (2020) The GTEx Consortium

atlas of genetic regulatory effects across human tissues. *Science,* **369**(6509), 1318–1330.

[197] Obayashi, T., Kodate, S., Hibara, H., Kagaya, Y., and Kinoshita, K. (2023) COXPRESdb v8: an animal gene coexpression database navigating from a global view to detailed investigations. *Nucleic Acids Research,* **51**(D1), D80–D87.

[198] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology,* **215**(3), 403–410.

[199] Assenov, Y., Ramírez, F., Schelhorn, S. E. E., Lengauer, T., and Albrecht, M. (2008) Computing topological parameters of biological networks. *Bioinformatics,* **24**(2), 282–284.

[200] Xia, F., Liu, J., Nie, H., Fu, Y., Wan, L., and Kong, X. (2020) Random Walks: A Review of Algorithms and Applications. *IEEE Transactions on Emerging Topics in Computational Intelligence,* **4**(2), 95–107.

[201] Ata, S. K., Wu, M., Fang, Y., Ou-Yang, L., Kwoh, C. K., and Li, X. L. Recent advances in network-based methods for disease gene prediction. (2021).

[202] Gruber, T. R. (1993) A translation approach to portable ontology specifications. *Knowledge Acquisition,* **5**(2), 199–220.

[203] Splendiani, A., Donato, M., and Drghici, S. (2014) Ontologies for bioinformatics. In *Springer Handbook of Bio-/Neuroinformatics* Vol. 2, pp. 441–461 SAGE Publications.

[204] Hendler, J. and van Harmelen, F. The Semantic Web: Webizing Knowledge Representation. (2008).

[205] Rath, A., Olry, A., Dhombres, F., Brandt, M. M., Urbero, B., and Ayme, S. (2012) Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation,* **33**(5), 803–808.

[206] Xiang, J., Zhang, J., Zhao, Y., Wu, F. X., and Li, M. (2022) Biomedical data, computational methods and tools for evaluating diseasedisease associations. *Briefings in Bioinformatics,* **23**(2).

[207] Papatheodorou, I., Oellrich, A., and Smedley, D. (2015) Linking gene expression to phenotypes via pathway information. *Journal of Biomedical Semantics,* **6**(1), 1–7.

[208] Shannon, C. E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal,* **27**(3), 379–423.

[209] Resnik, P. (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy.

[210] Wu, Z. and Palmer, M. (1994) Verb semantics and lexical selection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* Association for Computational Linguistics (ACL) Vol. 1994-June, pp. 133–138.

[211] Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., and Montmain, J. (2014) A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of Biomedical Informatics,* **48**, 38–53.

[212] Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., et al. Machine learning in bioinformatics. (2006).

[213] Dhal, P. and Azad, C. (2022) A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence,* **52**(4), 4543–4581.

[214] Sohail, A. and Arif, F. (2020) Supervised and unsupervised algorithms for bioinformatics and data science. *Progress in Biophysics and Molecular Biology,* **151**, 14–22.

[215] Nick, T. G. and Campbell, K. M. Logistic regression.. (2007).

[216] Breiman, L. (2001) Random forests. *Machine Learning,* **45**(1), 5–32.

[217] Mannor, S., Jin, X., Han, J., Jin, X., Han, J., Jin, X., Han, J., and Zhang, X. (2011) K-Means Clustering. In *Encyclopedia of Machine Learning* pp. 563–564 Springer, Boston, MA.

[218] Agrawal, R. and Srikant, R. (1994) Fast Algorithms for Mining Association Rules. In *Proc. of 20th International Conference on Very Large Data Bases, {VLDB'94}* pp. 487–499.

[219] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019) Reconciling modern machine-learning practice and the classical biasvariance trade-off. *Proceedings of the National Academy of Sciences of the United States of America,* **116**(32), 15849–15854.

[220] Kohavi, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference of Artificial Intelligence,*.

[221] Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters,* **27**(8), 861–874.

[222] Davis, J. and Goadrich, M. (2006) The relationship between precision-recall and ROC curves. In *ACM International Conference Proceeding Series* Vol. 148, pp. 233–240.

[223] Krawczyk, B. Learning from imbalanced data: open challenges and future directions. (2016).

[224] Chen, C., Liaw, A., and Breiman, L. (2004) Using random forest to learn imbalanced data. Statistics Department of University of California at Berkeley. *University of California, Berkeley,* **110**(1-12).

[225] Liu, X. Y. and Zhou, Z. H. (2006) The influence of class imbalance on cost-sensitive learning: An empirical study. In *Proceedings - IEEE International Conference on Data Mining, ICDM* pp. 970–974.

[226] Deng, Q., Gao, J., Ge, D., He, S., Jiang, B., Li, X., Wang, Z., Yang, C., and Ye, Y. (2020) Modern optimization theory and applications. *Scientia Sinica Mathematica,* **50**(7), 899–968.

[227] Eltaeib, T. and Mahmood, A. (2018) Differential evolution: A survey and analysis. *Applied Sciences (Switzerland),* **8**(10).

[228] Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T. N., Berghe, W. V., Goethals, B., and Laukens, K. (2015) A primer to frequent itemset mining for bioinformatics. *Briefings in Bioinformatics,* **16**(2), 216–231.

[229] Al-Maolegi, M. and Arkok, B. (2014) An Improved Apriori Algorithm For Association Rules. *International Journal on Natural Language Computing,* **3**(1), 21–29.

[230] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999) Efficient mining of association rules using closed itemset lattices. *Information Systems,* **24**(1), 25–46.

[231] Liu, B., Hsu, W., Ma, Y., and Ma, B. (1998) Integrating Classification and Association Rule Mining. *Knowledge Discovery and Data Mining,* pp. 80–86.

[232] Palczewska, A., Palczewski, J., Robinson, R. M., and Neagu, D. (2013) Interpreting random forest classification models using a feature contribution method. *arXiv:1312.1121 [cs],*.

[233] Knobbe, A., Cr, B., and Scholz, M. (2008) From Local Patterns to Global Models : The LeGo Approach to Data Mining. *From Local Patterns to Global Models: Pro-ceedings of the ECML/PKDD-08 Workshop (LeGo-08), Antwerp, Belgium,* pp. 1–16.

[234] Fürnkranz, J., Gamberger, D., and Lavrač, N. (2012) Foundations of Rule Learning, Vol. 6, , .

[235] Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016) Interpretable decision sets: A joint framework for description and prediction. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* **13-17-Augu**, 1675–1684.

[236] Fuernkranz, J. (1999) Separate-and-conquer rule learning. *Artificial Intelligence Review,* **13**(1), 3–54.

[237] Wilcke, X., Bloem, P., and de Boer, V. (2017) The knowledge graph as the default data model for learning on heterogeneous knowledge. *Data Science,* **1**(1-2), 39–57.

[238] Chandak, P., Huang, K., and Zitnik, M. (2023) Building a knowledge graph to enable precision medicine. *Scientific Data 2023 10:1,* **10**(1), 1–16.

[239] Bernal-Llinares, M., Ferrer-Gómez, J., Juty, N., Goble, C., Wimalaratne, S. M., and Hermjakob, H. (2021) Identifiers.org: Compact Identifier services in the cloud. *Bioinformatics,* **37**(12), 1781–1782.

[240] Eiglsperger, M. and Pich, C. (2020) Graph Markup Language (GraphML). *Handbook of Graph Drawing and Visualization,* pp. 532–557.

[241] Sun, Y., Norick, B., Han, J., Yan, X., Yu, P. S., and Yu, X. (2013) PathSelClus: Integrating meta-path selection with user-guided Object clustering in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data,* **7**(3), 1348–1356.

[242] Cao, X., Zheng, Y., Shi, C., Li, J., and Wu, B. (2017) Meta-path-based link prediction in schema-rich heterogeneous information network. *International Journal of Data Science and Analytics,* **3**(4), 285–296.

[243] Dong, Y., Chawla, N. V., and Swami, A. (2017) Metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Vol. Part F1296, pp. 135–144.

[244] Shi, B. and Weninger, T. (2016) Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems,* **104**, 123–133.

[245] Ma, W., Jin, W., Zhang, M., Wang, C., Cao, Y., Liu, Y., Ma, S., and Ren, X. (2019) Jointly learning explainable rules for recommendation with knowledge graph. In *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019* pp. 1210–1221.

[246] Lao, N. and Cohen, W. W. (2010) Relational retrieval using a combination of path-constrained random walks. In *Machine Learning* Springer Vol. 81, pp. 53–67.

[247] Pérez, J., Arenas, M., and Gutierrez, C. (2009) Semantics and complexity of SPARQL. *ACM Transactions on Database Systems,* **34**(3).

[248] Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., Plantikow, S., Rydberg, M., Selmer, P., and Taylor, A. (2018) Cypher: An evolving query language for property graphs. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* Association for Computing Machinery pp. 1433–1445.

[249] Wang, Q., Mao, Z., Wang, B., and Guo, L. Knowledge graph embedding: A survey of approaches and applications. (2017).

[250] Ali, M., Berrendorf, M., Hoyt, C. T., Vermue, L., Galkin, M., Sharifzadeh, S., Fischer, A., Tresp, V., and Lehmann, J. (2022) Bringing Light into the Dark: A Large-Scale Evaluation of Knowledge Graph Embedding Models under a Unified Framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **44**(12), 8825–8845.

[251] Su, C., Tong, J., Zhu, Y., Cui, P., and Wang, F. (2018) Network embedding in biomedical data science. *Briefings in Bioinformatics,* **21**(1), 182–197.

[252] Yang, B., tau Yih, W., He, X., Gao, J., and Deng, L. (2015) Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.*

[253] Nickel, M., Tresp, V., and Kriegel, H. P. (2011) A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011* pp. 809–816.

[254] Grover, A. and Leskovec, J. (2016) Node2vec: Scalable feature learning for networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Vol. 13-17-Augu, pp. 855–864.

[255] Veres, D. V., Gyurkó, D. M., Thaler, B., Szalay, K. Z., Fazekas, D., Korcsmáros, T., and Csermely, P. (2015) ComPPI: A cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic Acids Research,* **43**(D1), D485–D493.

[256] Itan, Y., Mazel, M., Mazel, B., Abhyankar, A., Nitschke, P., Quintana-Murci, L., Boisson-Dupuis, S., Boisson, B., Abel, L., Zhang, S. Y., et al. (2014) HGCS: An online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics,* **15**(1).

[257] Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research,* **28**(1), 27–30.

[258] Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., et al. (2021) The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research,* **49**(D1), D605–D612.

[259] Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016) dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation,* **37**(3), 235–241.

[260] Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., et al. (2020) Ensembl 2020. *Nucleic Acids Research,* **48**(D1), D682–D688.

[261] Castellana, S., Mastroianno, S., Palumbo, P., Palumbo, O., Biagini, T., Leone, M. P., De Luca, G., Potenza, D. R., Amico, C. M., Mazza, T., Russo, A., Di Stolfo, G., and Carella, M. (2019) Sudden death in mild hypertrophic cardiomyopathy with compound DSG2/DSC2/MYH6 mutations: Revisiting phenotype after genetic assessment in a master runner athlete. *Journal of Electrocardiology,* **53**, 95–99.

[262] Zullo, A., Frisso, G., Detta, N., Sarubbi, B., Romeo, E., Cordella, A., Vanoye, C., Calabrò, R., George, A., and Salvatore, F. (2017) Allelic Complexity in Long

QT Syndrome: A Family-Case Study. *International Journal of Molecular Sciences,* **18**(8), 1633.

[263] Byrjalsen, A., Hansen, T. V., Stoltze, U. K., Mehrjouy, M. M., Barnkob, N. M., Hjalgrim, L. L., Mathiasen, R., Lautrup, C. K., Gregersen, P. A., Hasle, H., et al. (2020) Nationwide germline whole genome sequencing of 198 consecutive pediatric cancer patients reveals a high frequency of cancer prone syndromes. *PLoS Genetics,* **16**(12), 1–24.

[264] Laan, M., Kasak, L., Timinskas, K., Grigorova, M., Venclovas, Č., Renaux, A., Lenaerts, T., and Punab, M. (2021) NR5A1 c.991-1G > C splice-site variant causes familial 46,XY partial gonadal dysgenesis with incomplete penetrance. *Clinical Endocrinology,* **94**(4), 656–666.

[265] Cerminara, M., Spirito, G., Pisciotta, L., Squillario, M., Servetti, M., Divizia, M. T., Lerone, M., Berloco, B., Boeri, S., Nobili, L., et al. (2021) Case Report: Whole Exome Sequencing Revealed Disease-Causing Variants in Two Genes in a Patient With Autism Spectrum Disorder, Intellectual Disability, Hyperactivity, Sleep and Gastrointestinal Disturbances. *Frontiers in Genetics,* **12**(February), 1–8.

[266] Zhao, T., Ma, Y., Zhang, Z., Xian, J., Geng, X., Wang, F., Huang, J., Yang, Z., Luo, Y., and Lin, Y. (2021) Young and early-onset dilated cardiomyopathy with malignant ventricular arrhythmia and sudden cardiac death induced by the heterozygous LDB3, MYH6, and SYNE1 missense mutations. *Annals of Noninvasive Electrocardiology,* **26**(4), 1–14.

[267] Costantini, A., Valta, H., Suomi, A. M., Mäkitie, O., and Taylan, F. (2021) Oligogenic Inheritance of Monoallelic TRIP11, FKBP10, NEK1, TBX5, and NBAS Variants Leading to a Phenotype Similar to Odontochondrodysplasia. *Frontiers in Genetics,* **12**, 680838.

[268] Dallali, H., Kheriji, N., Kammoun, W., Mrad, M., Soltani, M., Trabelsi, H., Hamdi, W., Bahlous, A., Ben Ahmed, M., Mahjoub, F., et al. (2021) Multiallelic Rare Vari-

ants in BBS Genes Support an Oligogenic Ciliopathy in a Non-obese Juvenile-Onset Syndromic Diabetic Patient: A Case Report. *Frontiers in Genetics,* **12**, 664963.

[269] Brancaccio, M., Mennitti, C., Cesaro, A., Monda, E., D'Argenio, V., Casaburi, G., Mazzaccara, C., Ranieri, A., Fimiani, F., Barretta, F., et al. (2021) Multidisciplinary In-Depth Investigation in a Young Athlete Suffering from Syncope Caused by Myocardial Bridge. *Diagnostics,* **11**(11), 2144.

[270] Tack, L. J., Spinoit, A. F., Hoebeke, P., Riedl, S., Springer, A., Tonnhofer, U., Hiess, M., Weninger, J., Mahmoud, A., Tilleman, K., et al. (2022) Endocrine outcome and seminal parameters in young adult men born with hypospadias: A cross-sectional cohort study. *eBioMedicine,* **81**.

[271] Perea-Romero, I., Solarat, C., Blanco-Kelly, F., Sanchez-Navarro, I., Bea-Mascato, B., Martin-Salazar, E., Lorda-Sanchez, I., Swafiri, S. T., Avila-Fernandez, A., Martin-Merida, I., et al. (2022) Allelic overload and its clinical modifier effect in Bardet-Biedl syndrome. *npj Genomic Medicine,* **7**(1), 1–7.

[272] Najjar, D., Chikhaoui, A., Zarrouk, S., Azouz, S., Kamoun, W., Nassib, N., Bouchoucha, S., and Yacoub-Youssef, H. (2022) Combining Gene Mutation with Expression of Candidate Genes to Improve Diagnosis of Escobar Syndrome. *Genes,* **13**(10).

[273] Varzari, A., Deyneko, I. V., Bruun, G. H., Dembic, M., Hofmann, W., Cebotari, V. M., Ginda, S. S., Andresen, B. S., and Illig, T. (2022) Candidate genes and sequence variants for susceptibility to mycobacterial infection identified by whole-exome sequencing. *Frontiers in Genetics,* **13**.

[274] Martinez de Lapiscina, I., Kouri, C., Aurrekoetxea, J., Sanchez, M., Naamneh Elzenaty, R., Sauter, K.-S., Camats, N., Grau, G., Rica, I., Rodriguez, A., et al. (2023) The NR5A1/SF-1 variant p.Gly146Ala cannot explain the phenotype of individuals with a difference of sex development. *preprint,*.

[275] Jacquemin, V., Versbraegen, N., Duerinckx, S., Massart, A., Soblet, J., Perazzolo, C., Deconinck, N., Brischoux-Boucher, E., De Leener, A., Revencu, N., et al. (2023)

Congenital hydrocephalus: new Mendelian mutations and evidence for oligogenic inheritance. *Human Genomics,* **17**(1), 1–14.

[276] Vockley, J., Rinaldo, P., Bennett, M. J., Matern, D., and Vladutiu, G. D. (2000) Synergistic heterozygosity: disease resulting from multiple partial defects in one or more metabolic pathways. *Molecular genetics and metabolism,* **71**(1-2), 10–18.

[277] Versbraegen, N., Gravel, B., Nachtegael, C., Renaux, A., Verkinderen, E., Nowé, A., Lenaerts, T., and Papadimitriou, S. (2023) Faster and more accurate pathogenic combination predictions with VarCoPP2.0. *BMC bioinformatics,* **24**(1), 179.

[278] Köster, J. and Rahmann, S. (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics,* **28**(19), 2520–2522.

[279] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., et al. (2019) STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research,* **47**(D1), D607–D613.

[280] Chen, W. H., Zhao, X. M., van Noort, V., and Bork, P. (2013) Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs. *PLoS Computational Biology,* **9**(5), e1003073.

[281] Rasko, D. A., Myers, G. S., and Ravel, J. (2005) Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics,* **6**.

[282] Tiessen, A., Pérez-Rodríguez, P., and Delaye-Arredondo, L. (2012) Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Research Notes,* **5**(1), 1–23.

[283] Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein engineering,* **12**(2), 85–94.

[284] Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., Quinlan, A. R., Nickerson, D. A., and Eichler, E. E. (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Research,* **22**(8), 1525–1532.

[285] Malatras, A., Michalopoulos, I., Duguez, S., Butler-Browne, G., Spuler, S., and Duddy, W. J. (2020) MyoMiner: Explore gene co-expression in normal and pathological muscle. *BMC Medical Genomics,* **13**(1).

[286] Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., et al. (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research,* **49**(D1), D344–D354.

[287] Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., et al. (2016) A global genetic interaction network maps a wiring diagram of cellular function. *Science (New York, N.Y.),* **353**(6306).

[288] Xu, J. and Li, Y. (2006) Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics,* **22**(22), 2800–2805.

[289] Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A. L. (2015) Uncovering disease-disease relationships through the incomplete interactome. *Science,* **347**(6224), 841.

[290] Gillis, J., Ballouz, S., and Pavlidis, P. (2014) Bias tradeoffs in the creation and analysis of protein-protein interaction networks. *Journal of Proteomics,* **100**, 44–54.

[291] Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A. L. (2007) The human disease network. *Proceedings of the National Academy of Sciences of the United States of America,* **104**(21), 8685–8690.

[292] Trouillon, T., Welbl, J., Riedel, S., Ciaussier, E., and Bouchard, G. (2016) Complex

embeddings for simple link prediction. In *33rd International Conference on Machine Learning, ICML 2016* Vol. 5, pp. 3021–3032.

[293] Mahdisoltani, F., Biega, J., and Suchanek, F. M. (2015) YAGO3: A knowledge base from multilingual wikipedias. In *CIDR 2015 - 7th Biennial Conference on Innovative Data Systems Research.*

[294] Abdelhamid, N. and Thabtah, F. (2014) Associative Classification Approaches: Review and Comparison. *Journal of Information and Knowledge Management,* **13**(3).

[295] Storn, R. and Price, K. (1997) Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization,* **11**(4), 341–359.

[296] Das, S. and Suganthan, P. N. (2011) Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation,* **15**(1), 4–31.

[297] Mezura-Montes, E., Velázquez-Reyes, J., and Coello Coello, C. A. (2006) A comparative study of differential evolution variants for global optimization. *GECCO 2006 - Genetic and Evolutionary Computation Conference,* **1**, 485–492.

[298] Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Capriotti, E., Casadio, R., Capella-Gutierrez, S., Cirillo, D., Del Conte, A., et al. (2021) DOME: recommendations for supervised machine learning validation in biology. *Nature Methods,* **18**(10), 1122–1127.

[299] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research,* **13**(11), 2498–2504.

[300] Collyer, J., Xu, F., Munkhsaikhan, U., Alberson, N. F., Orgil, B. O., Zhang, W., Czosek, R. J., Lu, L., Jefferies, J. L., Towbin, J. A., et al. (2022) Combining whole exome sequencing with in silico analysis and clinical data to identify candidate

variants in pediatric left ventricular noncompaction. *International Journal of Cardiology,* **347**, 29–37.

[301] Klaassen, S., Probst, S., Oechslin, E., Gerull, B., Krings, G., Schuler, P., Greutmann, M., Hürlimann, D., Yegitbasi, M., Pons, L., et al. (2008) Mutations in sarcomere protein genes in left ventricular noncompaction. *Circulation,* **117**(22), 2893–2901.

[302] Dellefave, L. and McNally, E. M. (2010) The genetics of dilated cardiomyopathy. *Current Opinion in Cardiology,* **25**(3), 198–204.

[303] Bagnall, R. D., Molloy, L. K., Kalman, J. M., and Semsarian, C. (2014) Exome sequencing identifies a mutation in the ACTN2 gene in a family with idiopathic ventricular fibrillation, left ventricular noncompaction, and sudden death. *BMC Medical Genetics,* **15**(1), 1–9.

[304] Richard, P., Ader, F., Roux, M., Donal, E., Eicher, J. C., Aoutil, N., Huttin, O., Selton-Suty, C., Coisne, D., Jondeau, G., et al. (2019) Targeted panel sequencing in adult patients with left ventricular non-compaction reveals a large genetic heterogeneity. *Clinical Genetics,* **95**(3), 356–367.

[305] Gerull, B., Gramlich, M., Atherton, J., McNabb, M., Trombitás, K., Sasse-Klaassen, S., Seidman, J. G., Seidman, C., Granzier, H., Labeit, S., et al. (2002) Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. *Nature Genetics,* **30**(2), 201–204.

[306] Duboscq-Bidot, L., Xu, P., Charron, P., Neyroud, N., Dilanian, G., Millaire, A., Bors, V., Komajda, M., and Villard, E. (2008) Mutations in the Z-band protein myopalladin gene and idiopathic dilated cardiomyopathy. *Cardiovascular Research,* **77**(1), 118–125.

[307] Ravenscroft, G., Zaharieva, I. T., Bortolotti, C. A., Lambrughi, M., Pignataro, M., Borsari, M., Sewry, C. A., Phadke, R., Haliloglu, G., Ong, R., et al. (2018) Bi-allelic mutations in MYL1 cause a severe congenital myopathy. *Human molecular genetics,* **27**(24), 4263–4272.

[308] Lamber, E. P., Guicheney, P., and Pinotsis, N. (2022) The role of the M-band myomesin proteins in muscle integrity and cardiac disease. *Journal of Biomedical Science,* **29**(1), 1–15.

[309] Salazar-Mendiguchiá, J., Ochoa, J. P., Palomino-Doza, J., Domínguez, F., Diéz-López, C., Akhtar, M., Ramiro-León, S., Clemente, M. M., Pérez-Cejas, A., Robledo, M., et al. (2020) Mutations in TRIM63 cause an autosomal-recessive form of hypertrophic cardiomyopathy. *Heart,* **106**(17), 1342–1348.

[310] Predmore, J. M., Wang, P., Davis, F., Bartolone, S., Westfall, M. V., Dyke, D. B., Pagani, F., Powell, S. R., and Day, S. M. (2010) Ubiquitin proteasome dysfunction in human hypertrophic and dilated cardiomyopathies. *Circulation,* **121**(8), 997–1004.

[311] Zhang, R. J., lin Yang, G., Cheng, F., Sun, F., Fang, Y., Zhang, C. X., Wang, Z., Wu, F. Y., Zhang, J. X., Zhao, S. X., et al. (2022) The mutation screening in candidate genes related to thyroid dysgenesis by targeted next-generation sequencing panel in the Chinese congenital hypothyroidism. *Clinical Endocrinology,* **96**(4), 617–626.

[312] Zhang, J., Bloedorn, E., Rosen, L., and Venese, D. (2004) Learning rules from highly unbalanced data sets. *Proceedings - Fourth IEEE International Conference on Data Mining, ICDM 2004,* pp. 571–574.

[313] Nguyen, T. P. and Jordán, F. (2010) A quantitative approach to study indirect effects among disease proteins in the human protein interaction network. *BMC Systems Biology,* **4**.

[314] Peplow, M. (2016) The 100000 Genomes Project. *BMJ,* **353**.

[315] Oughtred, R., Rust, J., Chang, C., Breitkreutz, B. J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., et al. (2021) The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science,* **30**(1), 187–200.

[316] Topatana, W., Juengpanich, S., Li, S., Cao, J., Hu, J., Lee, J., Suliyanto, K., Ma, D., Zhang, B., Chen, M., et al. Advances in synthetic lethality for cancer therapy: Cellular mechanism and clinical translation. (2020).

[317] Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., et al. (2023) A draft human pangenome reference. *Nature 2023 617:7960,* **617**(7960), 312–324.

[318] Barabási, A. L. (2009) Scale-free networks: A decade and beyond. *Science,* **325**(5939), 412–413.

[319] Grau, I., Sengupta, D., Matilde, M., Lorenzo, G., and Nowe, A. (2016) Grey-Box Model: An ensemble approach for addressing semi-supervised classification problems. In *Benelearn 2016: Belgian-Dutch Conference on Machine Learning.*

[320] Pezeshkpour, P., Tian, Y., and Singh, S. (2019) Investigating robustness and interpretability of link prediction via adversarial modifications. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* Vol. 1, pp. 3336–3347.

[321] Gusmão, A. C., Correia, A. H. C., De Bona, G., and Cozman, F. G. (2018) Interpreting Embedding Models of Knowledge Bases: A Pedagogical Approach.

[322] Hamilton, W. L., Bajaj, P., Zitnik, M., Jurafsky, D., and Leskovec, J. (2018) Embedding logical queries on knowledge graphs. In *Advances in Neural Information Processing Systems* Vol. 2018-Decem, pp. 2026–2037.

[323] Alivanistos, D., Berrendorf, M., Cochez, M., and Galkin, M. (2022) Query Embedding On Hyper-Relational Knowledge Graphs. In *ICLR 2022 - 10th International Conference on Learning Representations.*

[324] Martin, A. R., Williams, E., Foulger, R. E., Leigh, S., Daugherty, L. C., Niblock, O., Leong, I. U., Smith, K. R., Gerasimenko, O., Haraldsdottir, E., et al. Pan-

elApp crowdsources expert knowledge to establish consensus diagnostic gene panels. (2019).

[325] Bork, K., Zibat, A., Ferrari, D. M., Wollnik, B., Schön, M. P., Wulff, K., and Lippert, U. (2020) Hereditary angioedema in a single family with specific mutations in both plasminogen and SERPING1 genes. *JDDG - Journal of the German Society of Dermatology,* **18**(3), 215–223.

[326] Choi, E., Shin, S., Lee, S., Lee, S. J., and Park, J. (2020) Coexistence of digenic mutations in the collagen VI genes (COL6A1 and COL6A3) leads to Bethlem myopathy. *Clinica Chimica Acta,* **508**(March), 28–32.

[327] Gamba, A., Salmona, M., and Bazzoni, G. (2020) Quantitative analysis of proteins which are members of the same protein complex but cause locus heterogeneity in disease. *Scientific Reports,* **10**(1), 1–10.

[328] Ver Donck, F., Downes, K., and Freson, K. (2020) Strengths and limitations of high-throughput sequencing for the diagnosis of inherited bleeding and platelet disorders. *Journal of Thrombosis and Haemostasis,* **18**(8), 1839–1845.

[329] Quaio, C. R. D. C., Obando, M. J. R., Perazzio, S. F., Dutra, A. P., Chung, C. H., Moreira, C. M., Filho, G. M. N., Sacramento-Bobotis, P. R., Penna, M. G., de Souza, R. R. F., et al. (2021) Exome sequencing and targeted gene panels: A simulated comparison of diagnostic yield using data from 158 patients with rare diseases. *Genetics and Molecular Biology,* **44**(4).

[330] Zhou, Y. and Lauschke, V. M. (2021) Computational Tools to Assess the Functional Consequences of Rare and Noncoding Pharmacogenetic Variability. *Clinical Pharmacology and Therapeutics,* **110**(3), 626–636.

[331] Kremer, H. (2022) Novel gene discovery for hearing loss and other routes to increased diagnostic rates. *Human Genetics,* **141**(3-4), 383–386.

[332] Kadlubowska, M. K. and Schrauwen, I. (2022) Methods to Improve Molecular Diagnosis in Genomic Cold Cases in Pediatric Neurology. *Genes,* **13**(2).

[333] Frederiksen, S. D., Avramović, V., Maroilley, T., Lehman, A., Arbour, L., and Tarailo-Graovac, M. (2022) Rare disorders have many faces: in silico characterization of rare disorder spectrum. *Orphanet Journal of Rare Diseases,* **17**(1), 1–18.

[334] Draelos, R. L., Ezekian, J. E., Zhuang, F., Moya-Mendez, M. E., Zhang, Z., Rosamilia, M. B., Manivannan, P. K., Henao, R., and Landstrom, A. P. (2022) GENESIS: Gene-Specific Machine Learning Models for Variants of Uncertain Significance Found in Catecholaminergic Polymorphic Ventricular Tachycardia and Long QT Syndrome-Associated Genes. *Circulation: Arrhythmia and Electrophysiology,* **15**(4), E010326.

[335] Pittman, M., Lee, K., Srivastava, D., and Pollard, K. S. (2022) An oligogenic inheritance test detects risk genes and their interactions in congenital heart defects and developmental comorbidities. *bioRxiv,* p. 2022.04.08.487704.

[336] Anfinson, M., Fitts, R. H., Lough, J. W., James, J. M., Simpson, P. M., Handler, S. S., Mitchell, M. E., and Tomita-Mitchell, A. (2022) Significance of α-Myosin Heavy Chain (MYH6) Variants in Hypoplastic Left Heart Syndrome and Related Cardiovascular Diseases. *Journal of Cardiovascular Development and Disease,* **9**(5), 144.

[337] Chen, J., Ma, Y., Li, H., Lin, Z., Yang, Z., Zhang, Q., Wang, F., Lin, Y., Ye, Z., and Lin, Y. (2022) Rare and potential pathogenic mutations of LMNA and LAMA4 associated with familial arrhythmogenic right ventricular cardiomyopathy/dysplasia with right ventricular heart failure, cerebral thromboembolism and hereditary electrocardiogram abnormality. *Orphanet Journal of Rare Diseases,* **17**(1), 1–17.

[338] Koczwara, K. E., Lake, N. J., DeSimone, A. M., and Lek, M. (2022) Neuromuscular disorders: finding the missing genetic diagnoses. *Trends in genetics : TIG,* **38**(9), 956–971.

[339] O'Neill, M. J., Sala, L., Denjoy, I., Wada, Y., Kozek, K., Crotti, L., Dagradi, F., Kotta, M. C., Spazzolini, C., Leenhardt, A., et al. (2023) Continuous Bayesian

variant interpretation accounts for incomplete penetrance among Mendelian cardiac channelopathies. *Genetics in Medicine,* **25**(3).

[340] Almansoori, S., Alsters, S., Walters, R., and Blakemore, A. (2023) Oligogenic inheritance in severe adult obesity. *preprint,*.