LipidCRED: Lipid Computational Reaction and Enzyme Database

Qassim Alkassir

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the M.Sc. degree in
Biochemistry, with specialization in Bioinformatics

Department of Biochemistry, Microbiology, and Immunology
Faculty of Medicine
University of Ottawa

Acknowledgements

First and foremost, I would like to express my profound gratitude to my supervisors, Dr. Steffany Bennett and Dr. Miroslava Cuperlovic-Culf. Dr. Bennett, thank you for welcoming me into your lab, introducing me to the fascinating world of lipidomics, and providing invaluable mentorship that pushed me to grow as a scientist and researcher. Dr. Cuperlovic-Culf, I am equally indebted to you for agreeing to co-supervise this project and for offering dedicated guidance, critical insights, and mentorship that were essential throughout my project. Your combined expertise, support, and thoughtful advice were instrumental in the success of this work. I feel incredibly fortunate to have benefited from the mentorship of two exceptional researchers.

My sincere appreciation extends to the members of the Bennett lab and my colleagues at the National Research Council (NRC). The stimulating environment, insightful discussions, and camaraderie made the challenges of research more manageable and enjoyable. A special thank you is owed to Dr. Irina Alecu and Thao Nguyen for generously sharing their expertise in lipid biochemistry, which provided crucial context and understanding for my work.

Finally, and most profoundly, this journey would not have been possible without the unwavering love, support, and boundless encouragement of my incredible parents, Ali A. and Sahar A., and my beloved siblings Batoul, Ghina, and Aasem. Though 10,000 km separated us, your belief in me never faltered. Your faith was a constant source of strength, your cheers my greatest motivation, and your sacrifices deeply appreciated. This achievement is a testament to your endless love and support, and it is as much yours as it is mine. Thank you for everything, from the bottom of my heart.

Abstract

Mass spectrometry-based lipidomic approaches generate vast datasets of individual lipid species, but translating these compendiums of abundances into meaningful biological insights, particularly regarding enzymatic pathways, remains a significant challenge. Existing bioinformatics tools often struggle with nomenclature heterogeneity, limited reaction coverage, the need for extensive *post-hoc* validation of enzyme predictions, and/or cumbersome user workflows. To address these limitations, I developed Lipid Computational Reaction and Enzyme Database (LipidCRED), a novel bioinformatics platform designed for comprehensive and validated exploration of lipid metabolism. LipidCRED integrates a robust relational database, amalgamating curated information from SwissLipids, Rhea, UniProtKB, and the HUGO Gene Nomenclature Committee (HGNC), with a sophisticated software engine. Key innovations include advanced nomenclature parsing, a "LipidMatcher" module that intelligently infers specific molecular products from class-level reactions using carbon-chain matching rules, and an emphasis on pre-validated enzyme data. A user-friendly R/Shiny web interface ensures broad accessibility. Benchmarking against leading tools (LipidOne, LINEX, BioPAN) demonstrated LipidCRED's competitive performance in validated lipid-enzyme associations, uniquely providing highly accurate enzyme lists requiring minimal filtering. By streamlining the path from complex lipid lists to reliable enzymatic insights, LipidCRED empowers researchers to more effectively unravel the roles of lipid metabolism in health and disease.

Table of Contents

List of Abbreviations

3KDS	3-ketodihydrosphingosine
3NF	Third Normal Form
AC(s)	UniProtKB accession numbers
AD	Alzheimer's disease
AGPAT	Alkylglycerolphosphate 2-*O*-acyltransferase
AGPR	Acylglycerone-phosphate reductase
AGPS	Alkylglycerone-phosphate synthase
B4GALT5	β-1,4-galactosyltransferase 5
B4GALT6	β-1,4-galactosyltransferase 6
C1P	Ceramide-1-phosphate
CDases	Ceramidases
CDP	Cytidine 5'-diphosphate
CDS	Cytidylyltransferase
Cer	Ceramide
CERK	Ceramide kinase
CERS	Ceramide synthase
CGCDases	Ceramide glycocerebrosidases
CGT	Ceramide glycotransferase
CID(s)	PubChem Compound Identifier(s)
CL	Cardiolipins
CLS	Cardiolipin synthase
CoA	Coenzyme-A
CPT	Choline phosphotransferase
CSV	Comma-Separated Values
DAG	Diacylglycerol
DAGK	Diacylglycerol kinase

DEGS1	Sphingolipid Δ4-desaturase DES1
DEGS2	Sphingolipid Δ4-desaturase DES2
DG(O)	1-alkyl,2-acylglycerol
DG(P)	1-*O*-alk-1′-enyl-2-acyl-*sn*-glycerol
DGATs	Diacylglycerol O-acyltransferases
DOI(s)	Digital Object Identifier(s)
EC number	Enzyme Commission (EC) number
EPT	Ethanolamine phosphotransferase
FA	Fatty acyl
FADS3	Fatty acid desaturase 3
GalCer	Galactosylceramide
GCase	Glucosylceramidase
GL	Glycerolipids
GLB1	LacCer β-galactosidase
GlcCer	Glucosylceramide
GP	Glycerophospholipids
GPAT	Glycerone-phosphate *O*-acyltransferase
GSL	Glycosphingolipids
HCOP	HGNC Comparison of Orthology Predictions
Hex2Cer	Dihexosylceramide
HexCer	Hexosylceramide
HGNC	HUGO (Human Genome Organisation) Gene Nomenclature Committee
KDSR	3-ketodihydrosphingosine reductase
LacCer	Lactosylceramide
LCB(s)	Long-chain base(s)
LipidCRED	Lipid Computational Reaction and Enzyme Database
LMISSD	LIPID MAPS In-Silico Structure Database

LMSD	LIPID MAPS Structure Database
LPA	Monoacylglycerophosphate
LPA(O)	1-*O*-alkyl-*sn*-glycero-3-phosphate
LPAATs	LPA acyltransferases
LPC	Monoacylglycerophosphocholine
LPC(O)	Monoalkylglycerophosphate
LPE	Monoacylglycerophosphoethanolamine
LPE(O)	Monoalkylglycerophosphoethanolamine
LPI	Monoacylglycerophosphoinositol
LPLATs	Lyso-phospholipid acyltransferases
LRU	Least Recently Used
LSI	Lipidomics Standards Initiative
MS	Mass spectrometry
NAE(s)	*N*-acylethanolamines
NAPE(s)	*N*-acyl phosphatidylethanolamines
OEA	Ontology enrichment analysis
PA	Glycerophosphate
PA(O)	1-*O*-alkyl-2-acyl-*sn*-glycero-3-phosphate
PAP	Phosphatidic acid phosphatase
PC	Glycerophosphocholine
PC(O)	1-alkyl,2-acylglycerophosphocholines
PC(P)	1-*O*-alk-1′-enyl-2-acyl-*sn*-glycero-3-phosphocholine
PD	Parkinson’s disease
PE	Glycerophosphoethanolamine
PE(O)	1-alkyl,2-acylglycerophosphoethanolamine
PE(P)	1-*O*-alk-1′-enyl-2-acyl-*sn*-glycero-3-phosphoethanolamine
PED	Plasmanylethanolamine desaturase

PEMT	Phosphatidylethanolamine *N*-methyltransferase
PG	Glycerophosphoglycerol
PGP	Glycerophosphoglycerophosphate
PGPP	Phosphatidylglycerolphosphate phosphatase
PGPS	PGP synthase
PH	Phosphohydrolase
PI	Glycerophosphoinositol
PIS	PI synthase
PK	Polyketide
PLA	Phospholipases
PLA1/2	Phospholipases A1/2
PLC	Phospholipase C
PLD	Phospholipase D
PLPPs	Phospholipid phosphatases
PNPLA1	Patatin like phospholipase domain containing 1
PR	Prenol lipid
PS	Glycerophosphoserine
PSD1	Phosphatidylserine decarboxylase
PSS1	Phosphatidylserine synthases-1
PSS2	phosphatidylserine synthases-2
S1P	Sphingosine-1-phosphate
SGPL1	S1P lyase 1
SGPPs	Sphingosine-1-phosphate phosphatases
SL	Saccharolipid
SM	Sphingomyelin
SMases	Sphingomyelinases
SMS1	Sphingomyelin synthase 1
SMS2	Sphingomyelin synthase 2

SMSs	Sphingomyelin synthases
SP	Sphingolipids
SPHKs	Sphingosine kinases
SPT	Serine palmitoyltransferase
ST	Sterol lipid
TG	Triacylglycerol
UGCG	UDP-glucose ceramide glycosyltransferase
UI	User interface

List of Tables

List of Figures

List of Algorithms

Chapter 1: General Introduction

1.1 Computational lipidomics and lipid metabolism

Computational lipidomics, a subdiscipline of bioinformatics, aims to derive knowledge from high throughput mass spectrometry-based lipidomics by leveraging advances in computer science to provide methods for analysis, simulation and interpretation [1]. Lipidomics refers to the high-throughput acquisition and quantification of individual lipid species [2]. Lipids (named after the Greek "lipos", meaning fat) are essential biological molecules that play vital roles across all biological systems. Formally, lipids are defined as “hydrophobic or amphipathic small molecules that originate entirely, or in part, by carbanion-based condensations of thioesters and/or by carbocation-based condensations of isoprene units” [3]. The entire collection of chemically distinct lipid species in a cell, organ, or a biological system is termed as a “lipidome” [4]. Lipids are divided into eight categories based on their chemical properties: fatty acyls (FA), glycerolipids (GL), glycerophospholipids (GP), sphingolipids (SP), sterol lipids (ST), prenol lipids (PR), saccharolipids (SL) and polyketides (PK). Each category is subdivided into classes and subclasses depending on the headgroup and the linkage type between the structural backbone and the hydrocarbon chains.

In this thesis, the focus is on SPs and GPs, initially viewed as simple but key structural components of membranes, each class having unique physiochemical properties. However, their roles as bioactive lipids have been cemented as mounting evidence points to the involvement of individual molecular species in various essential functions, including – SPs in intra- and inter-cellular signalling, cell migration and inflammation [5], and GPs as signalling mediators or precursors of many signalling molecules, in addition to playing key roles in bioenergetics, cell

proliferation and apoptosis [6,7]. Studies by Vonkova [8] and Koberlin [9] have demonstrated close bidirectional interactions between SPs and GPs [10] with a growing number of studies demonstrating the association between dysregulation of SP and GP metabolism and various pathologies [5]. As a result, this thesis will focus on providing bioinformatic tools to interrogate their metabolism in lipidomic datasets.

Lipid homeostasis in healthy conditions is maintained by a multitude of nutritionally and metabolically regulated processes. Thus, it comes as no surprise that aberrant metabolism of SPs and GPs has been implicated in numerous human diseases, such as cancer [11], insulin-resistant diabetes [12], Alzheimer's disease (AD) [13,14] and Parkinson's disease (PD) [15]. As more lipidomic experiments are performed, coupled with a growing interest in the network analysis of lipidome and with the mounting data repositories [16,17], it is imperative that more computational tools are developed to ensure analysis of these increasingly larger datasets and to continue pushing research efforts studying how lipid metabolism promotes healthy or disease conditions.

1.1.1 Brief overview of sphingolipid and glycerophospholipid metabolism

Lipid metabolism combines numerous enzymatic reactions within and between classes and subclasses of lipids. The majority of enzymes involved in lipid metabolism are multispecific, i.e., they catalyze changes in structural features of many different molecular lipid species belonging to the same lipid class, albeit with different rate constants and substrate specificities, most of which have yet to be empirically determined [18]. As such, lipid metabolic pathways appear deceivingly simple, abstracting the fact that each class reaction is performed on hundreds of different molecular lipid species within, and sometimes between classes, leading to varying concentrations of individual lipid molecules, each with potentially unique functions [19]. Thus, the capacity of lipidomics to quantify individual molecular species requires a better understanding of the function

of individual lipids, the complexity of their metabolic patterns, and the important subtle differences in enzymatic activities on cell function.

Both SP and GP metabolism involve a complex interplay between *de novo* synthesis and remodelling pathways. SP metabolism begins with structurally simple molecules, an amino acid and a fatty acyl coenzyme-A (CoA), eventually producing thousands of distinct lipid species, and ends with the release of small metabolites which are either recycled into glycerophospholipids or back into sphingolipids (Figure 1.1). This process involves a variety of enzymes with oxidoreductase, transferase, or hydrolase activity working in a concerted manner [20,21].

The enzymes involved in the catalysis of individual reactions of sphingolipid metabolism are broadly conserved across fungi, plant, and animal kingdoms [22]. Ceramides are at the center of sphingolipid metabolism and are the precursor for all complex sphingolipids. There are four major pathways for the production of ceramides within the cell: *de novo* synthesis, sphingomyelin hydrolysis, the modification pathway, and the salvage pathway [23]. Complex sphingolipids can subsequently be created by remodelling ceramide via various enzymes catalyzing reactions such as phosphorylation and head groups modification (Figure 1.1).

The first step in the *de novo* synthetic pathway (Figure 1.1a) is the conversion of serine and fatty acyl-CoA into 3-ketodihydrosphingosine (3KDS) via the actions of the serine palmitoyltransferase (SPT) complex. The subunit identities of the SPT complex dictate which acyl-CoA, in the range of C_{14}-C_{18}, can be used as a substrate, forming a variety of long chain bases (LCBs) that differ by structure and function and form the sphingoid backbone [24,25]. The most commonly used fatty acyl-CoA is palmitoyl-CoA in mammals [26,27]. The product formed by the SPT complex is further reduced into dihydrosphinganine by 3-ketodihydrosphingosine reductase

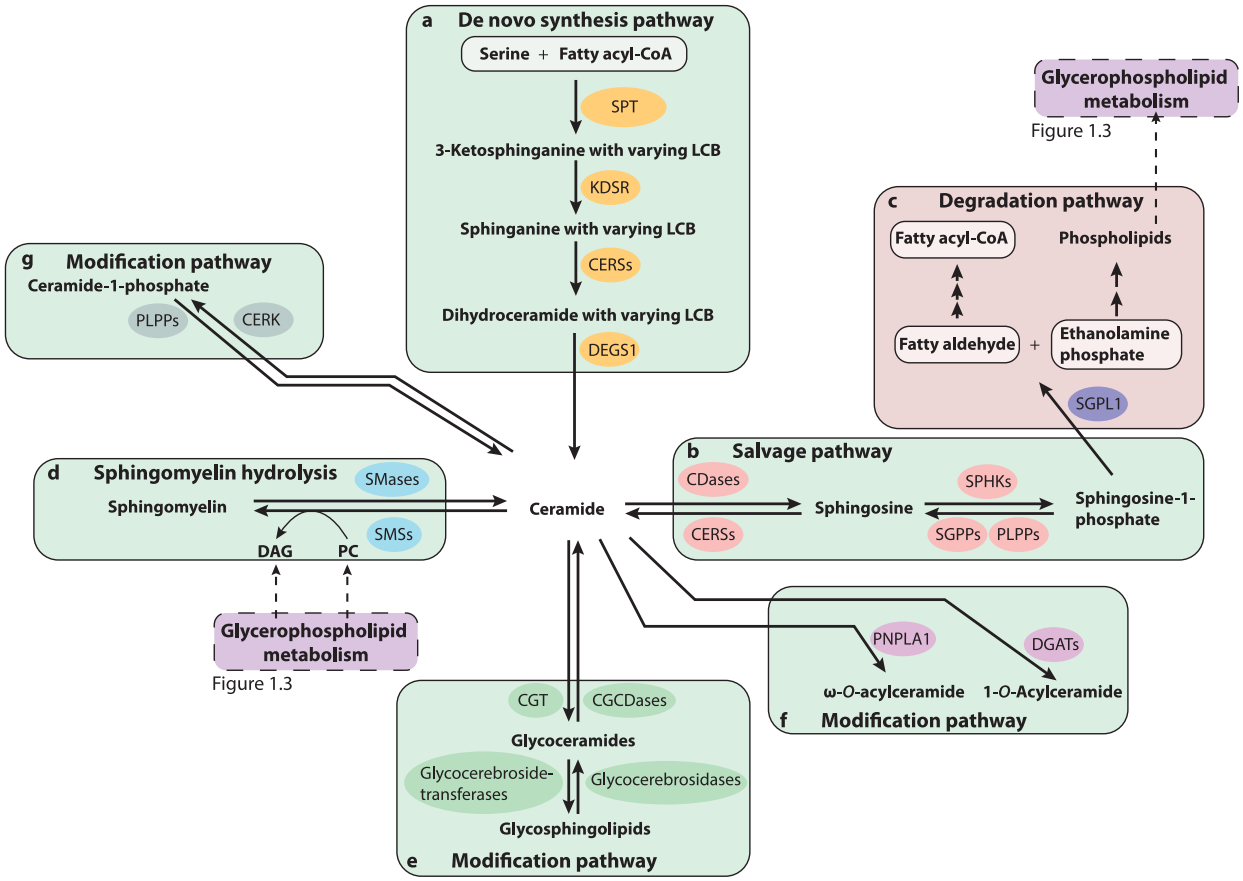


Figure 1.1 Overview of sphingolipid metabolism in mammals. a) The *de novo* synthesis pathway is initiated in the endoplasmic reticulum via the action of serine palmitoyltransferase (SPT) and culminates in the generation of ceramides. **b)** In the salvage pathway, sphingosine can be recycled back to ceramide or be phosphorylated to form sphingosine-1-phosphate (S1P). **c)** The degradation pathway is the only exit point of sphingolipid metabolism, wherein S1P is hydrolyzed to a fatty aldehyde and ethanolamine phosphate. **d)** The sphingomyelin hydrolysis pathway maintains complex sphingolipids such as sphingomyelin. **e-g)** Ceramides can be converted to other complex sphingolipids through modifications of the headgroup at the 1-hydroxyl position. In this schematic, enzymes are indicated within ovals, either as their gene name or enzyme family name. The dashed purple boxes indicate bridge points between glycerophospholipid and sphingolipid metabolism. Metabolites in white boxes are not included in the database. Abbreviations: CDases, ceramidases; CERK, Ceramide kinase; CERSs, ceramide synthases; CGCDases, ceramide glycocerebrosidases; CGT, ceramide glycotransferase; DEGS1, sphingolipid Δ4-desaturase DES1; DGATs, diacylglycerol *O*-acyltransferases; KDSR, 3-ketodihydrosphingosine reductase; LCB, long-chain base, PLPPs, phospholipid phosphatases; PNPLA1, patatin like phospholipase domain containing 1; SGPL1, sphingosine-1-phosphate lyase 1; SGPPs, sphingosine-1-phosphate phosphatases; SMases, sphingomyelinases; SMSs, sphingomyelin synthases; SPHKs, sphingosine kinases; SPT, Serine palmitoyltransferase; UGCG, UDP-glucose ceramide glycosyltransferase.

(KDSR) [28,29]. Dihydrosphinganine is then further processed by ceramide synthase (CERS) enzymes to form dihydroceramides by the addition of an acyl group from a fatty acyl-CoA to the free amino group of dihydrosphinganine via amide linkage [30]. The formation of ceramides, the centre of sphingolipid metabolism, requires the introduction of a 4,5-*trans* (Δ4*E*) double bond, catalyzed by the sphingolipid Δ4-desaturase DES1 (DEGS1) [31].

Sphingoid bases can also be salvaged from ceramides via the salvage pathway (Figure 1.1b) or targeted for degradation via the degradation pathway (Figure 1.1c). In the salvage pathway, hydrolysis of the acyl chain by ceramidases (CDases) salvages free sphingoid bases that can be reacylated cyclically to reform ceramides by CERSs. Alternatively, in the degradation pathway, the sphingoid bases salvaged from sphingosine can be phosphorylated, generating sphingosine-1-phosphates (S1Ps), where sphingosine refers, in this case, to any sphingoid base also known as long chain bases (LCBs). S1Ps undergo irreversible degradation by S1P lyase 1 (SGPL1), yielding phosphatidylethanolamine and hexadecenal [32].

In the sphingomyelin hydrolysis pathway (Figure 1.1d), ceramides can be further processed to form complex sphingolipids such as sphingomyelin (SM), each defined by the molecular identity of their *N*-acyl chain and their sphingoid base. To form SM, sphingomyelin synthase 1 (SMS1) and SMS2 transfer the phosphocholine group of glycerophosphocholine (PC) to ceramides, forming a sphingomyelin and remodelling the PC to diacylglycerol (DAG) [33]. Ceramides can be generated by reverse reactions where the phosphocholine headgroup is hydrolyzed from SM by sphingomyelinases (SMases).

The modification pathway can both expand and regenerate ceramides through multiple headgroup modifications (Figure 1.1e-g). Glycosphingolipids (GSLs) are formed through the sequential addition of sugar groups to the 1-hydroxyl group of ceramide (Figure 1.1e). For

example, glucosylceramides (GlcCers) are formed through the action of UDP-glucose ceramide glycosyltransferase (UGCG) [34]. Further expansions can be made by adding a galactose group to a GlcCer, forming lactosylceramide (LacCer), via glycosyltransferases β-1,4-galactosyltransferase 5 (B4GALT5) and B4GALT6 [35]. Subsequently, LacCer can be further modified for the biosynthesis of various classes of GSLs such as globosides and gangliosides [36]. The breakdown of these GSLs is equally important to prevent their accumulation and serve the purpose of regenerating ceramides. GlcCer, a crucial molecule in SP metabolism due to its role as a precursor of the numerous GSLs that originate from the LacCer motif, can be hydrolyzed back into ceramide (Cer) via glucosylceramidase (GCase). Additionally, LacCer can be hydrolyzed back into GlcCer through the action of LacCer β-galactosidase (GLB1) [37].

Two other modifications constitute the modification pathway. Ceramides can be converted to 1-*O*-acylceramide by diacylglycerol O-acyltransferases (DGATs) [38] or omega-acylceramides by PNPLA1 (Figure 1.1f) [39]. The enzymatic processes responsible for the degradation of 1-O-acylceramides and omega-O-acylceramides have not been identified. Alternatively, ceramides can be phosphorylated by ceramide kinase (CERK), forming ceramide 1-phosphate (C1P). Ceramide 1-phosphates can be converted back to ceramides by phospholipid phosphatases (PLPPs) (Figure 1.1g) [40].

Beyond the canonical SP pathway initiated with serine, alternative metabolic routes arise when the SPT complex, which catalyzes the first committed step in *de novo* synthesis, utilizes different amino acid substrates. Instead of serine, the SPT complex can condense palmitoyl-CoA with other amino acids, notably L-alanine or glycine. This use of alternative substrates leads to the formation of atypical sphingoid bases lacking the crucial C1-hydroxyl group: 1-deoxysphinganine (from L-alanine) and 1-deoxymethylsphinganine (from glycine), respectively (Figure 1.2) [41,42].

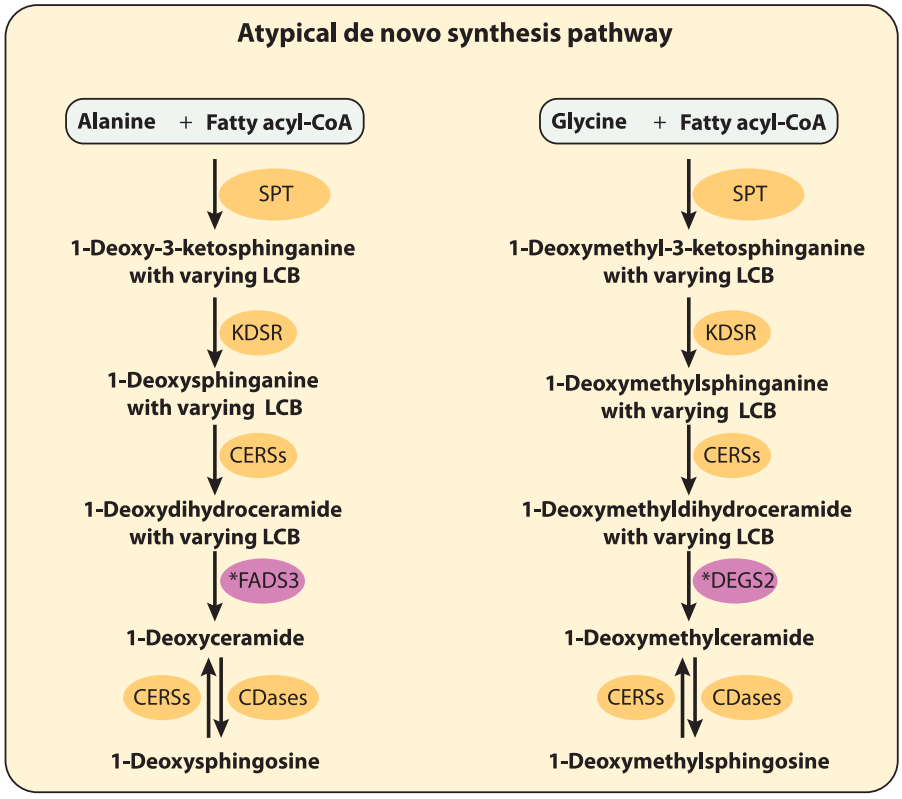


Figure 1.2 Atypical sphingolipid synthesis in mammals. Alternate activity of serine palmitoyltransferase (SPT) towards amino acids other than serine, namely alanine and glycine results in the formation of atypical categories of sphingolipids, 1-deoxysphingolipids and 1-deoxymethylsphingolipids, respectively. Enzymes are indicated within ovals, either as their gene name or enzyme family name. *FADS3: this enzyme introduces a Δ14,15-*cis* double bond onto 1-deoxydihydroceramide to form 1-deoxyceramide. *DEGS2: this enzyme introduces a Δ3,4-*cis* double bond onto 1-deoxymethyldihydroceramide to form 1-deoxymethylceramide. Metabolites in white boxes are not included in the database. Abbreviations: CDases, ceramidases; CERSs, ceramide synthases; DEGS2, Sphingolipid Δ4-desaturase DES2; FADS3, fatty acid desaturase 3; KDSR, 3-ketodihydrosphingosine reductase; SPT, serine palmitoyltransferase.

The first three metabolic steps remain the same, to conjugate alanine/glycine with a fatty acyl-CoA via SPT, forming 1-deoxyketosphinganine/1-deoxymethylsphinganine, which are then reduced via KDSR to form the respective atypical sphingoid bases lacking the crucial C1-hydroxyl group, 1-deoxysphinganine/1-deoxymethylsphinganine. Following their formation, these atypical bases enter steps analogous to the canonical pathway but yielding different products. The CERS enzymes acylate 1-deoxysphinganine to form 1-deoxydihydroceramide, and similarly acylate 1-deoxymethylsphinganine to form 1-deoxymethyldihydroceramide. A subsequent desaturation step occurs, differing significantly from the canonical pathway where DEGS1 introduces a (4*E*) double bond to form ceramide. In the atypical pathway, desaturation introduces a double bond at a different position, typically (14*Z*) in 1-deoxydihydroceramide to form 1-deoxyceramide, which is catalyzed by FADS3 [43], and (3*E*) in 1-deoxymethyldihydroceramide to form 1-deoxymethylceramide, catalyzed by DEGS2 [44,45]. CDases can hydrolyze 1-deoxyceramide to form 1-deoxysphingosine and, likewise hydrolyze 1-deoxymethylceramide to release 1-deoxymethylsphingosine. Crucially, this reaction is reversible: CERSs can also utilize 1-deoxysphingosine and 1-deoxymethylsphingosine as substrates for acylation, regenerating 1-deoxyceramide and 1-deoxymethylceramide, respectively. Despite this interconversion between the ceramide and sphingoid base forms, the fundamental absence of the C1-hydroxyl group in all these atypical lipids prevents their further modification into complex sphingolipids (like SM or glycosphingolipids). It also blocks their entry into the canonical degradation pathway initiated by sphingosine kinases (SPHKs), as the 1-phosphate derivatives cannot be formed [44]. Consequently, while not completely static due to the CERS/CDase equilibrium, these atypical sphingolipids represent metabolic cul-de-sacs within the major sphingolipids pathways, potentially leading to their accumulation and associated lipotoxicity [44].

With respect to GP metabolism, the primary route for *de novo* synthesis of PC and glycerophosphoethanolamine (PE) is through the Kennedy pathway [46,47] (Figure 1.3A). The PC and PE branches of the Kennedy pathway are based upon cytidine 5'-diphosphate (CDP)-choline, for the synthesis of PC, and CDP-ethanolamine for the synthesis of PE. Consequently, the two branches of the Kennedy pathway are often referred to as the CDP-choline and CDP-ethanolamine pathways, respectively.

In the CDP-choline pathway, DAG is converted to PC via CDP-choline:1,2-diacylglycerol choline phosphotransferase (CPT). Alternatively, PC is formed from the sequential methylation of the choline headgroup of PE via phosphatidylethanolamine *N*-methyltransferase (PEMT), referred to as the PEMT pathway [48,49]. PE is synthesized by the CDP-ethanolamine pathway, which couples DAG to CDP-ethanolamine via CDP-ethanolamine:1,2-diacylglycerol ethanolamine phosphotransferase (EPT). PE can also be synthesized by decarboxylation of phosphatidylserine via phosphatidylserine decarboxylase [50]. Glycerophosphoserine (PS) is synthesized from PC or PE through a base-exchange reaction via phosphatidylserine synthases-1 (PSS1) and -2 (PSS2). PSS1 catalyzes the exchange reactions where serine replaces the choline or ethanolamine group, whereas PSS2 exchanges the ethanolamine group for a serine [51]. Glycerophosphoinositol (PI) is synthesized from *myo*-inositol and CDP-DAG via PI synthase (PIS) [52,53]. Similarly, CDP-DAG is combined with glycerol-3-phosphate to form glycerophosphoglycerophosphate (PGP) in a reaction catalyzed by PGP synthase (PGPS). PGP is then dephosphorylated by PTPMT1 to form glycerophosphoglycerol (PG). Cardiolipin is formed by cardiolipin synthase which transfers the phosphatidic acid group from CDP-DAG to the 3-hydroxyl group on the glycerol moiety of PG [54].

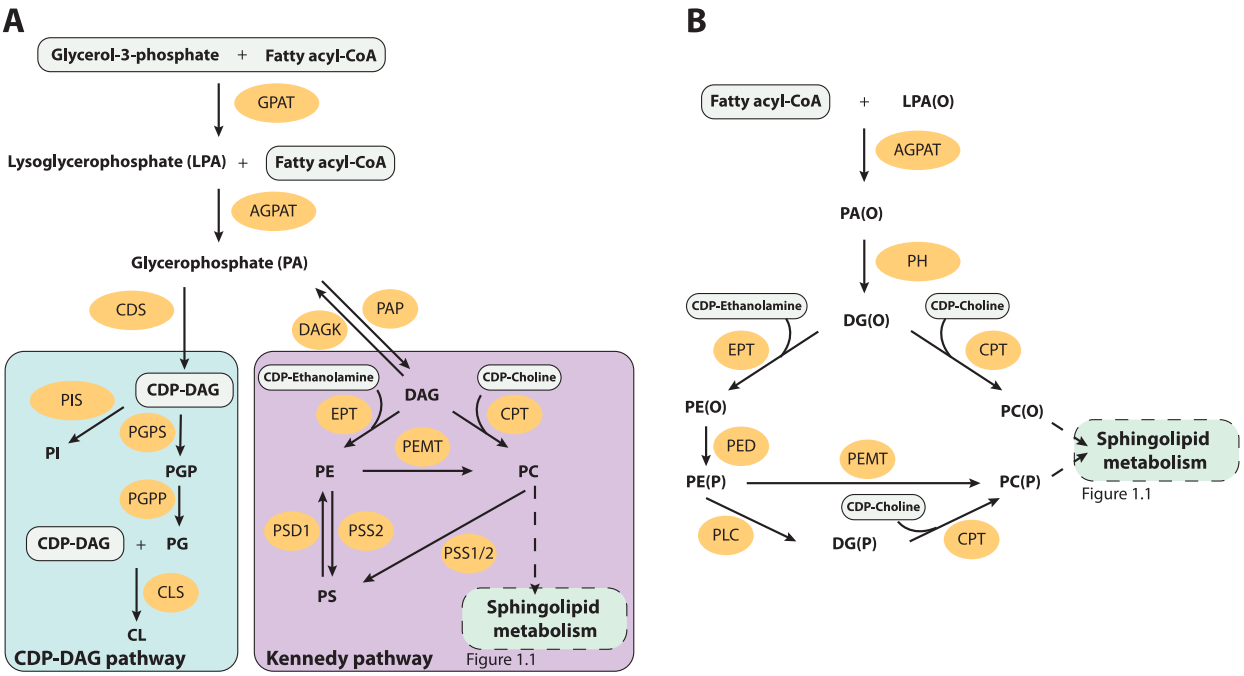


Figure 1.3 Overview of glycerophospholipid synthesis pathways in mammals. A) Synthesis of ester glycerophospholipids branches from glycerophosphate (PA) into either the CDP-DAG or Kennedy pathway to generate various glycerophospholipid classes. **B)** Ether glycerophospholipid synthesis progresses through an analogous Kennedy pathway to produce ether and vinyl-ether lipids by utilizing an alkyl-PA. The dashed green boxes indicate bridge points between glycerophospholipid and sphingolipid metabolism. Enzymes are indicated within orange ovals. Metabolites in white boxes are not included in the database. Some metabolites are omitted for brevity; more detailed description is provided in the text. Abbreviations: AGPAT, 1-acylglycerol-3-phosphate acyltransferases; CDS, cytidylyltransferase; CLS, cardiolipin synthase; CPT, choline phosphotransferase; DAGK, diacylglycerol kinase; EPT, ethanolamine phosphotransferase; GPAT, glycerol-3-phosphate acyltransferases; PAP, phosphatidic acid phosphatase; PED, plasmanylethanolamine desaturase; PEMT, phosphatidylethanolamine N-methyltransferase; PGPP, phosphatidylglycerolphosphate phosphatase; PGPS, phosphatidylglycerolphosphate synthase; PH, phosphohydrolase; PIS, phosphatidylinositol synthase; PLC, phospholipase C; PSD1, phosphatidylserine decarboxylase; PSS1/2, phosphatidylserine synthase 1/2.

The lipids described thus far can be grouped as acyl lipids, owing to the ester bond that links their fatty acyl chains to the glycerol backbone. However, fatty acyl chains can also be linked via an ether or vinyl ether bonds, creating ether lipids (alkyl and alkenyl glycerols). Ether lipids have an alkyl chain linked to the sn-1 position of the glycerol backbone via an ether bond. Compared to the ester bond in acyl lipids, this ether bond imparts lipids with higher metabolic stability. Ether lipids are found in many organisms: bacteria, protozoa, plants, and mammals including humans [55].

Ether lipid biosynthesis starts with a peroxisomal enzyme, glycerone-phosphate *O*-acyltransferase (GPAT), that acylates glycerone phosphate (dihydroacetone phosphate, DHAP) with a fatty acyl-CoA. The resulting 1-acyl-glycerone 3-phosphate exchanges its acyl group for a fatty alcohol, yielding 1-*O*-alkyl-glycerone 3-phosphate in a complex reaction catalyzed by alkylglycerone-phosphate synthase (AGPS). The next step generates the first glycerol-based intermediate, 1-*O*-alkyl-*sn*-glycero-3-phosphate, abbreviated as LPA(O), by acylglycerone-phosphate reductase (AGPR). After LPA(O) is transported to the endoplasmic reticulum, it is esterified with acyl-CoA at the *sn*-2 position of glycerol by alkylglycerolphosphate 2-*O*-acyltransferase (AGPAT), yielding 1-*O*-alkyl-2-acyl-*sn*-glycero-3-phosphate (PA(O)) (Figure 1.3B). The phosphate is removed via phosphatide phosphatase, yielding 1-*O*-alkyl-2-acyl-*sn*-glycerol (DG(O)), which can then either be converted into 1-*O*-alkyl-2-acyl-*sn*-glycero-3-phosphoethanolamine (PE(O)) using CDP-ethanolamine via EPT or into 1-*O*-alkyl-2-acyl-*sn*-glycero-3-phosphocholine (PC(O)) using CDP-choline via CPT. PE(O) can then be oxidized by plasmanylethanolamine desaturase (PED) to yield the vinyl double bond in 1-*O*-alk-1′-enyl-2-acyl-*sn*-glycero-3-phosphoethanolamine (PE(P)). Interestingly, PC(O) is not a substrate for PED [54]. 1-*O*-alk-1′-enyl-2-acyl-*sn*-glycero-3-phosphocholine (PC(P)) is instead formed by either

methylation of PE(P) via PEMT, or by a salvage pathway involving phospholipase C (PLC) hydrolysis of PE(P) to 1-*O*-alk-1′-enyl-2-acyl-*sn*-glycerol (DG(P)), which is then converted to PC(P) via CPT [54].Following the interconversion of polar headgroups during *de novo* synthesis, the fatty acid moieties of glycerophospholipids can undergo remodelling, a process referred to as Land's remodelling cycle (Figure 1.4) [56]. This process of fatty acid hydrolysis and re-esterification creates new glycerophospholipid species, effectively bypassing the *de novo* synthesis pathway. The cycle is orchestrated by the concerted action of phospholipases A1/2 (PLA1/2) and LPA acyltransferases (LPAATs) [57,58].

1.2 Lipid structure and classification and their impact on metabolic pathways

The first challenge facing bioinformatic interrogation of lipid diversity is the ability to map changes in lipid metabolism at the molecular level derived from measurements of lipid abundances. For this, a detailed understanding of lipid identities and classification is required.

1.2.1 Lipid class categories

GPs and SPs are the two most abundant lipid categories in mammals, making up the majority of structural lipids of the plasma membrane [59]. Both GPs and SPs consist of three building blocks, the modification of which is the source of their molecular diversity (Figure 1.5). GPs consist of a glycerol backbone, a polar headgroup attached via a phosphodiester linkage to the *sn*-3 hydroxyl group of glycerol, and up to two hydrocarbon chains (Figure 1.5A). SPs consist of a sphingoid LCB, a headgroup (or lack thereof), and a fatty acyl chain (Figure 1.5B) [54,60] .

The head group attached to the glycerol backbone defines the lipid class of GPs. The GP lipid classes include PC, PE, PI, PS, and PG which are defined by choline, ethanolamine, inositol, serine and glycerol head group, respectively. Additionally, the molecular species in each of these

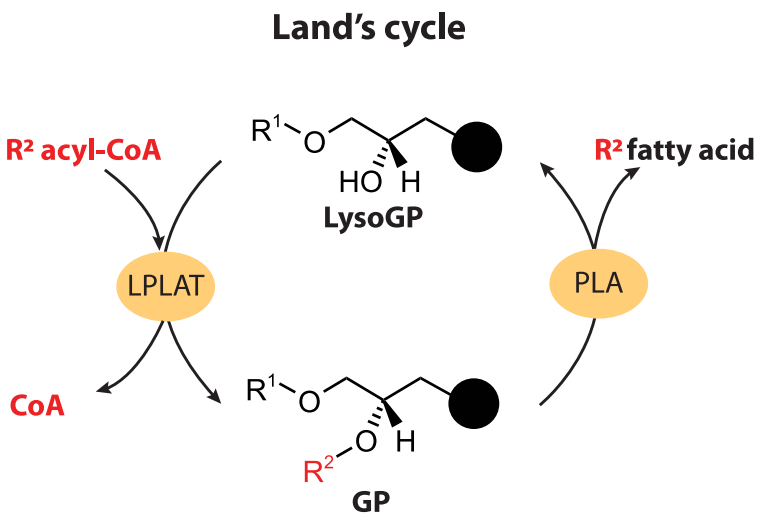
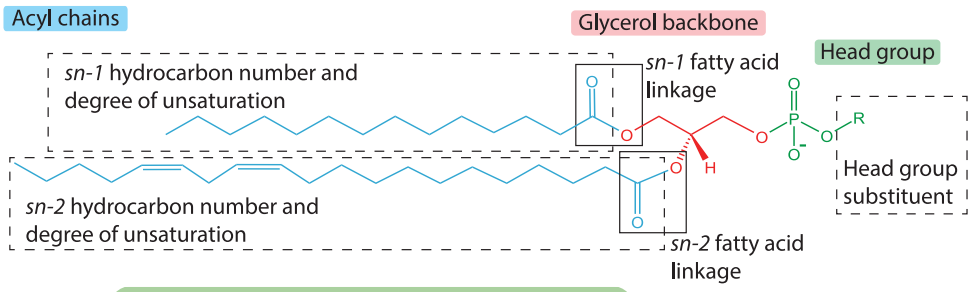
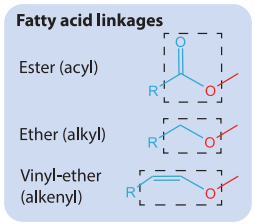


Figure 1.4 Remodelling of glycerophospholipids. Following *de novo* synthesis, glycerophospholipids can undergo remodelling, also called Land's cycle. This process involves fatty acid hydrolysis and re-esterification to create new glycerophospholipid species and is orchestrated by phospholipases (PLA) and lyso-phospholipid acyltransferases (LPLATs), seen in orange ovals. The glycerophospholipid head group is indicated by a black circle.

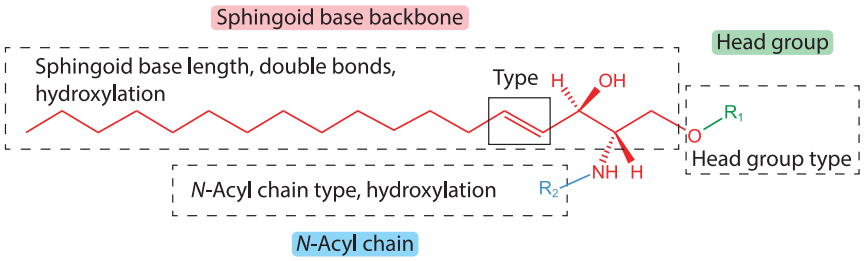
A Glycerophospholipid diversity



Glycerophospholipid	R (head) group
PA	-
PC	Choline
PE	Ethanolamine
PS	Serine
PI	Inositol
PG	Glycerol



B Sphingolipid diversity



Sphingolipid	R_1 (head) group	R_2 group
Sph	Hydroxyl	$-H_2$
S1P	Phosphate	$-H_2$
Cer	Hydroxyl	*N*-Acyl
C1P	Phosphate	*N*-Acyl
SM	Phosphocholine	*N*-Acyl
GlcCer	Glucose	*N*-Acyl
GalCer	Galactose	*N*-Acyl
Complex Glycosphingolipids	Oligosaccharides	*N*-Acyl

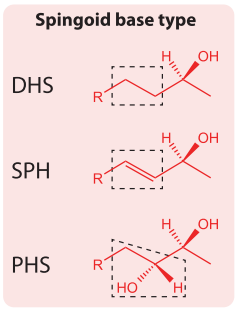


Figure 1.5 Chemical diversity of sphingolipids and glycerophospholipids. A) Glycerophospholipids have a glycerol backbone with fatty acids at the *sn*-1 and *sn*-2 positions. The headgroup consists of a phosphate and a substituent that defines the glycerophospholipid class and name. Ether glycerophospholipids have an ether or vinyl-ether linkage at the *sn*-1 position. **B)** Sphingolipids consist of a sphingoid base (backbone), an *N*-acyl chain and a headgroup. The sphingoid base type is defined by its hydroxylation and unsaturation, whereas the headgroup defines the sphingolipid class and name. The boxed parts of the structures represent the building blocks that confer diversity. Abbreviations: C1P, ceramide-1-phosphate; Cer, ceramide; DHS, sphinganine; GalCer, galactosylceramide; GlcCer, glucosylceramide; PA, glycerophosphate; PC, glycerophosphocholine; PE, glycerophosphoethanolamine; PG, glycerophosphoglycerol; PHS, 4-hydroxy-sphinganine; PI, glycerophosphoinositol; PS, Glycerophosphoserine; S1P, sphingosine-1-phosphate; SM, sphingomyelin; Sph, sphingosine.

classes can be further classified into subclasses based on the type of linkage of the aliphatic chains attached to the glycerol backbone. The hydrocarbons can be attached at the *sn*-1 and *sn*-2 positions of glycerol via an ester bond, an ether bond, or a vinyl-ether bond, known as acyl, alkyl, or alkenyl bonds, respectively. If a fatty acyl chain is absent at either the *sn*-1 or *sn*-2 positions, replaced by a hydroxyl group, the lipid is known as a *lyso* GP [54].

SPs are classified by their sphingoid LCB, amide-linked hydrocarbon chains (or lack thereof), and head groups. Sphingoid LCBs, which form the backbone of SPs, consist of a hydrocarbon chain and an amide group. The various combinations of hydrocarbon chain length, degree of unsaturation, hydroxylation, and structure of the head group moiety determine the specific subclass of a particular sphingoid LCB. For example, Sphingosine is the most common sphingoid LCB, known by the shorthand nomenclature as Sph(d18:1), as its carbon chain includes 18 carbon atoms, and a Δ4*E* double bond (18:1) and hydroxyl groups at position 1 and 3 of the LCB (d) [60]. This nomenclature indicates the chain length and number of double bonds, with the prefix 'd' indicating a dihydroxy base. The saturated analogue of sphingosine is called sphinganine, or Sph(d18:0). The attachment of a fatty acyl chain to a sphingoid base via an amide-linkage, creates the Cer class of sphingolipids. The specific subclass of Cer is determined by its precursor sphingoid LCB. For example, sphingosine generates Cers, whereas sphinganine generates dihydroceramides. The length and degree of unsaturation of the fatty acyl chain attached to the sphingoid LCB can vary significantly, contributing to their diversity and consequently granting them different structures and levels of hydrophobicity. Mammalian SPs contains a wide range of fatty acyls, ranging from C_2 to C_{32} [60]. The last region that contributes to the heterogeneity of SPs is the head group. Complex SPs are formed via modifications of the C1 of ceramide as explained in section 1.1.1.

1.2.2 Lipid nomenclature

Advances in mass spectrometry (MS)-based lipidomics have revolutionized the study of lipids with its power for the identification and quantification of lipid molecules in cells, tissues, and biofluids. However, an ongoing challenge in lipidomic analysis is the structural identification of lipids, where the depth of structural detail obtained depends on the analytical workflow, and is reflected in the annotation nomenclature as proposed by Liebisch [61]. The shorthand nomenclature used in this thesis to describe GPs is as follows: first the class is denoted by the head group, followed by the carbon chain length, degree of unsaturation, and hydroxylation (if present) of the *sn*-1 fatty acyl and *sn*-2 fatty acyl chain. The most basic annotation of lipids is by lipid class and the sum composition of carbons and double bonds in the fatty acyl chains. For GPs, PC(34:1) is an example of a glycerophosphocholine molecule identified and annotated to the lipid species level, with phosphocholine as the lipid class, and a total of 34 carbons with one degree of unsaturation summed across the *sn*-1 and *sn*-2 positions. Advanced mass spectrometry methodologies yield molecular lipid species where the positional isomeric level of the fatty acyl chains is known and is indicated by a forward slash "/". However, when there's certainty in the composition of the fatty acyl constituents, but not in their *sn*-position, an underscore "_" is used instead. For example, PC(16:0_18:1) correctly identifies 16:0 and 18:1 as the two acyl chains but not their placement on the glycerol backbone, whereas PC(16:0/18:1) indicates that 16:0 and 18:1 chains are found at the *sn*-1 and *sn*-2 positions, respectively. *Lyso* GPs are indicated by adding "L" as prefix to the headgroup and placing '0:0' in the place of the absent fatty acyl chain. O- indicates an ether linkage, P- indicates a vinyl-ether linkage, and the absence of a designation indicates an ester linkage to the glycerol backbone.

The shorthand nomenclature used in this thesis to describe SPs is as follows: first the subclass is denoted by the head group, followed by the backbone specification which includes the carbon chain length, number of hydroxy groups, and degree of unsaturation, and finally the carbon chain length and degree of unsaturation of the *N*-acyl chain, if present. Cer(34:1) is an example of a ceramide identified and annotated to the lipid species level, where the sphingoid LCB is unknown and thus the sum of sphingoid LCB and fatty acyl chain is shown as 34 total number of carbons:1 degree of unsaturation. Utilizing advanced analytical methods can elucidate SPs to the hydroxyl group level, which annotates the number of hydroxyl groups in the sphingoid LCB. When experimentation resolves SPs to the fatty acyl level, the *N*-linked fatty acyl chain is annotated by the number of carbons:number of double bonds and placed after a forward slash "/", separating it from the sphingoid LCB specification. For example, Cer(d34:1) is a ceramide molecule annotated to the hydroxyl group level, with "d" indicating that its backbone is a (d)ihydroxy sphingoid LCB, with 34 total carbons and 1 double bond within the backbone and *N*-acyl chain combined. Whereas Cer(d18:0/16:0) identifies the hydroxyl groups and the number of carbons and double bonds in the sphingoid LCB and *N*-acyl chain, with 18 and 16 carbons, respectively, and no double bonds.

1.2.3 A revised understanding of lipid metabolism

As extensively reviewed elsewhere, the rapid development of MS technologies over the past 30 years has propelled lipid analysis [62]. Prior to the 1990s, a major limitation was that lipid coverage was limited and sporadic, covering only a few species from each lipid class and distinguishing only one or two lipid classes of interest [63]. Since then, the burgeoning research interest in lipid biology has led to the expansion of the known lipidome. For example, *N*-acyl phosphatidylethanolamines (NAPEs) were first isolated and characterized from pea seeds in 1968 [64]. However, it wasn't until decades after their discovery that researchers began to take interest

in their biological functions [65,66]. Beginning in the 1990s and onwards, the association of NAPEs and *N*-acylethanolamines (NAEs) with neurotoxicity was elucidated. At the same time, the enzymes involved in their metabolism were characterized, namely NAPE-hydrolyzing phospholipase D (PLD), *N*-acyltransferase, and NAE-hydrolyzing amidase, thereby shedding more light on their biological roles [65,66]. In addition to the discovery of previously unknown lipid classes, lipids belonging to the same class were found to elicit different, at times opposite, biological roles, based on their specific fatty acyl chains [67–70]. Such discoveries fueled the rapid advancement of analytical approaches that enable the identification and quantification of molecular lipid species.

Pathway databases such as KEGG [71], WikiPathways [72], BioCyc [73], Reactome [74], and Rhea [75] serve as knowledgebases storing and organizing these reactions and biological processes curated from published literature. They provide visualizations of the myriad of interactions underlying biological processes. For example, complex metabolic reactions can be represented with a minimal set of symbols, lines, and arrows. In the context of lipid metabolism, reactions are generalized to the lipid class level, for example, PE → PC is a head group modifying reaction catalyzed by phosphatidylethanolamine N-methyltransferase (PEMT). This reaction can be seen in Figure 1.3, which displays glycerophospholipid metabolism as a network of class reactions, and this representation mirrors the approach pathway databases use to display lipid metabolism. However, this approach is an oversimplification of the inherent diversity and complexity of lipid metabolism. Since experimental data from lipidomics at the molecular species level is available, a more accurate representation of the PE → PC reaction would require hundreds of arrows originating from each individual PE species to its corresponding PC species, e.g.,

PE(16:0/18:0) → PC(16:0/18:0), PE(20:1/24:0) → PC(20:1/24:0), PE(20:0/22:0) → PC(20:0/22:0), etc.

Given this diversity and the strengths and limitations of current mass spectrometry assessments of lipidomic abundances, we require a revision to the "class" model of lipid metabolism. By embracing the individuality of molecular lipid species, we will be able to uncover their specific biological functions and mechanisms within biological systems, which paves the road to more accurate interpretations. To accomplish this task, computational tools are crucial to fully embrace the power of data generated from lipidomics studies and contribute to our understanding of lipid metabolism.

1.3 Bioinformatics for lipidomics

1.3.1 Overview of lipidome bioinformatics

Improvements in MS technologies and developments in bioinformatics have synergistically contributed to advancements in the measurement and interpretation of lipidomics data in recent years. Lipidomics studies analyze large numbers of samples, and the amount of experimental data generated is increasing rapidly. Recognizing the immense potential within these data to yield significant biological insights has been a primary catalyst for bioinformatic development, aiming to overcome the inherent challenges in data processing and enable more impactful analysis and interpretations. These challenges are continually being addressed as new algorithms and tools are developed to enable effective computational processing of data. Lipidomic computational tools can be broadly divided into three categories: lipid identification and quantification, data analysis and visualization, and biological interpretation.

Advances in experimental workflows, such as advanced applications of liquid chromatography tandem mass spectrometry (LC-MS/MS), have resulted in better structural information, and, consequently, the generation of more data. The development of software solutions for the automated identification and quantification of lipids from large-scale high-throughput MS data forms the first frontier of current computational challenges in lipidomics. This task involves a number of analytical steps, and a review of software tools dedicated to this processing is available elsewhere [17] and on the LIPID MAPS website as part of their lipidomics tool guide (https://www.lipidmaps.org/resources/tools/ms).

Once the identification and quantification of lipids from biological samples is complete, various computational approaches are employed to analyze the resulting high-dimensional data and begin uncovering biological meaning. These strategies range from univariate methods, examining the connection between single lipid species and experimental groups, to multivariate statistical methods that explore relationships among multiple lipids and associated metadata (e.g., age, sex) simultaneously [76,77]. Furthermore, broader data mining approaches, often facilitated by general-purpose platforms (e.g., Orange [78], WEKA [79], KNIME [80], Cytoscape [81]), are utilized. These include network and graph-based methods, which allow for the systemic analysis of relationships between lipids, often inferred from abundance data using correlation- or classification-based approaches [82]. Several general purpose tools, such as SiDCO [83] and WGCNA [84], along with foundational algorithms reviewed in [85], are available for correlation and network analysis. Additionally, specialized platforms like MetaboAnalyst [86], LipoStar2 [87], liputils [88], and LipidSuite [89] focus more on metabolite and lipid analysis and provide different types of analysis, including network derivation. These methods are discussed in detail by Hoffmann et al. [17]. Beyond these, machine learning approaches are increasingly being applied

directly to experimental lipidomic data for predictive modeling, for instance, in disease biomarker discovery [90] and within integrated analysis tools [91].

While statistical and network analyses are crucial for identifying patterns and potential lipid changes, they are data driven, making it difficult to link observations and observed perturbations. Achieving this deeper biological interpretation often requires integrating the analytical results with existing biochemical knowledge. Given the complexity of lipid metabolism, with its vast number of species and reactions, computational systems biology approaches are invaluable. This interpretation phase typically involves leveraging curated pathways and knowledge-based network strategies. Key steps often include mapping the identified lipids to standardized identifiers compatible with knowledge bases and ontology databases, performing lipid ontology or pathway enrichment analyses to identify over-represented biological pathways, and conducting pathway or network analysis that overlays experimental data onto known metabolic maps [92]. Analysis of individual lipid molecules rather than class investigations is essential, as distinct molecular species possess unique structural properties that dictate their specific functions [93]. These bioinformatic solutions are essential for disentangling the complexity of lipid metabolism and translating large datasets into meaningful biological insights.

1.3.2 Lipid-oriented databases

Curated lipid-oriented databases, which organize historical and newly published individual lipid structures, are essential for researchers who aim to identify the specific molecules in their biological samples. Over the last 5-10 years, the size of lipidomics experimental datasets generated from MS and MS/MS has greatly increased, which has instigated the need to create searchable databases to help with analysis. The most widely used lipid-oriented databases are provided by

LIPID MAPS [94] and SwissLipids [95], with general databases also including lipid information such as PubChem [96], HMDB [97], KEGG [71], and CHEBI [98].

LIPID MAPS provides several databases where lipids are catalogued according to the LIPID MAPS nomenclature and classification scheme [3,99,100]. The LIPID MAPS Structure Database (LMSD) contains 48,691 unique lipid structures (April 2025) obtained and manually curated from LIPID MAPS Consortium's core laboratories and partners, other lipid databases, scientific literature, and computationally generated structures based on commonly occurring mammalian fatty acid chains. LMSD provides both bulk (lipid species) annotations for MS data based on shorthand notation and fully annotated names of structurally-defined lipids [101]. Since 2022, LMSD began including lipid reaction data to link lipids and their biochemical reactions catalyzed by enzymes [94]. These reactions have been incorporated from Rhea [75], Reactome [74], and WikiPathways [72].

Rhea [75] is an expert-curated knowledgebase of chemical and transport reactions and is the standard for enzyme and transporter annotation in UniProtKB [102]. Currently, Rhea covers 17,614 unique reactions involving 14,498 unique reaction participants (April 2025). The Reactome Knowledgebase [74] is a manually curated pathway database, focused on manual annotation of signaling and metabolic molecules and their relations organized into biological pathways and processes. The database contains 22,732 pathways spanning 15 organisms, including 2,769 *Homo sapiens* pathways (April 2025).

WikiPathways [72] is a biological pathway database and open science platform dedicated to community-driven curation of biological pathways. To date, WikiPathways stores 3,218 pathways (April 2025). In addition to LMSD, LIPID MAPS added the In-Silico Structure Database (LMISSD), which is populated via computational expansion of head groups and chains of common

lipid classes. LMISSD aims to enumerate all possible structures available using a large set of acyl/alkyl chains, and contains 1,132,490 molecules, ~25,000 of which are in common with LMSD [94].

The SwissLipids database is an expert-curated resource of lipids and their metabolism [95]. Experimentally characterized lipids are curated from primary literature and mapped to the chemical ontology database ChEBI [98]. Lipid metabolism is described using the Rhea knowledgebase, which itself is based on ChEBI. Enzymes, transporters, and interacting proteins involved in lipid metabolism are described using the UniProt Knowledgebase UniProtKB [102], which uses Rhea as its reference. The SwissLipids knowledgebase includes a total of 779,689 known and theoretical lipid structures belonging to over 550 lipid classes, which are organized into two hierarchical lipid classifications, one that follows the structural classification system of LIPID MAPS [99] and another based on MS shorthand notation [103], linking MS-based lipid identification to structures.

To study lipid metabolism, it is important to relate lipids to the genes and proteins which are involved in their metabolism. BRENDA is a comprehensive database storing functional and molecular information of enzymes obtained mainly from literature references [104]. BRENDA also utilizes text mining algorithms to provide more complete coverage of the literature, culminating in 8476 enzymes in its database (April 2025). UniProtKB provides another comprehensive resource for functional information on proteins, and consists of two sections: a section of manually curated records extracted from literature, and a section of computationally analyzed records awaiting manual annotation, named UniProtKB/Swiss-Prot (reviewed, manually annotated) and UniProtKB/TrEMBL (unreviewed, automatically annotated), respectively [102]. BRENDA is cross-referenced within UniProtKB and, despite not being lipid- or enzyme-specific,

UniProtKB is a valuable resource for lipid-related genes and proteins, providing a total of 391,866 of proteins (reviewed + unreviewed) which include 118,427 reviewed proteins spanning 14,778 different taxonomic species (April 2025). Other sources link lipids and enzymes at a lipid category level, including MetabolicAtlas [105] and BioCyc [73].

1.3.3 Ontology enrichment analysis

Ontology enrichment analysis (OEA) provides a different approach to investigating lipidomics data without needing enzyme information. OEA groups lipids based on shared physicochemical and biological properties such as membrane fluidity, intrinsic curvature, subcellular compartment, degree of unsaturation, and class/subclass, after which statistical methods are applied to determine whether certain ontology terms are enriched, or, in other words, overrepresented compared to a target list or higher ranked in a list of lipids ordered by a statistic (e.g., fold change, *p*-value) than expected by chance [106]. The goal of this analysis is to identify significant associations and relationships within the lipidomic data, and, in turn, elucidate underlying biological mechanisms. For example, ontology enrichment is commonly used to identify lipid classes associated with cellular signalling pathways [92].

There are numerous lipidomic ontology enrichment tools, which can be divided into two categories based on whether they utilize a database or not. Lipid Mini-On [107], LipidSig [91], and LipidSuite [89] are notable examples of database-independent tools. These tools utilize text mining to obtain ontology terms by parsing individual lipid species names to extract their structural characteristics such as lipid class, hydrocarbon chain length, and degree of unsaturation, followed by statistical analysis that produces a variety of visualizations of lipid enrichment by structural characteristics. Conversely, instead of focusing on structural characteristics, database-dependent enrichment tools can generate more biologically contextualized results. Both LION/web [108] and

LipiDisease [109] are freely available tools that utilize a database to perform ontology enrichment analysis. Currently, LION/web contains over 50,000 lipid species, which are linked with four major branches: lipid classification, chemical and physical properties, function, and sub-cellular component, resulting in over 250,000 connections. LipiDisease aims to perform disease enrichment analysis given a set of lipids. This analysis is built around extracting lipid and disease associations using the biomedical literature from the PubMed database, covering 4270 diseases and 4798 lipids. However, given the dependence of both LION/web and LipiDisease on a database, a lipid must exist in their database to be included in the enrichment analysis. Looking ahead, curated knowledgebases such as those within LION/web and LipiDisease offer significant potential as functional resources for developing and training next-generation machine learning algorithms focused specifically on predicting lipid-mediated biological outcomes, as was shown for general enzymatic reactions [110].

1.3.4 Network-based lipidomic tools

Network analysis-based lipidomics is a computational approach intended to give meaning to high-throughput lipidomics data by inferring connections between lipids, which can subsequently be linked to specific biological pathways or processes. Unlike enrichment analysis, which provides a statistical representation of selected biomolecules within a predefined group (e.g., a pathway), network analysis is centred around studying the interactions and relationships among lipids. This approach allows a more comprehensive understanding of the systems that drive homeostasis in lipid metabolism [111].

Network inference can be performed directly from data using various algorithms, such as correlation-based methods, Bayesian networks, and mutual information approaches [82,112]. However, despite significant efforts in this area, purely data-derived networks can lead to

misleading information due to challenges like inferring indirect interactions, sensitivity to threshold errors, and dataset-specific biases [82]. To mitigate these issues, knowledge-based networks leverage existing biological knowledge about molecular interactions by combining enzymatic reaction information from databases or known pathways to create a map of direct, known interactions rooted in current understanding of possible reactions. In the context of lipidomics, integrating lipid-specific databases with protein knowledgebases such as UniProtKB, enables the derivation of knowledge-based lipid-protein networks of lipid metabolism.

Relating lipidomics data and associating it with known metabolic pathways is an important step towards building a better understanding of the lipidome in different systems. Existing pathway databases such as WikiPathways [72], Reactome [74], and KEGG [71] organize current knowledge of molecular interactions, reactions, and relation networks of biological processes. Computational methods can utilize these tools by connecting pathways to databases of biological annotations and experimental data to create effective systems biology approaches. However, in lipidomics, given the enormous diversity of complex lipids (e.g., GPs and SPs), most pathway tools map individual lipid species to their lipid class, truncating their complexity and subsequently obscuring their individual functional contributions [16].

To perform lipid pathway analysis, several tools are available, each with slight differences in their implementation. BioPAN, which is hosted by LIPID MAPS, combines lipidomic profiles with current knowledge of lipid metabolism using a statistical workflow [113]. The statistical model is based on the model described by [114,115], which was developed for gene expression data and later adopted by Nguyen et al. [116] for lipidomics data. BioPAN uses quantitative lipidomics data to compare two experimental conditions (e.g. control vs. treated) by calculating a statistical score for all possible lipid pathways to predict active/suppressed pathways between the

two conditions. This analysis can be performed at the lipid class, subclass, and molecular species level, where input lipids are mapped as reactants and products based on BioPAN's manually collated database of mammalian lipid metabolic pathways. The results are visualized using an interactive network along with a table listing genes involved in pathways that are significantly activated/suppressed. The workflow employed by BioPAN can be described as a hybrid approach, combining biological knowledge (collated database) with data-driven methods calculating pathway score.

Another notable tool is LINEX^2 [18], which is a lipid network analysis framework aimed at generating dataset-specific lipid interaction networks. The lipid reactions driving their analysis are curated from the Rhea and Reactome databases and subsequently extended from the lipid class level to the lipid species level. The generated data-specific lipid networks are analyzed using an in-house network enrichment algorithm to propose candidate enzymes driving enzymatic dysregulation. An important aspect of this method is the usage of hypernetworks, which acts as a proxy of the multispecificity of lipid enzymes, allowing the enrichment algorithm to connect solutions from the same class reactions. The resulting set of candidate enzymes can be used to drive further investigation into potential disease mechanisms. Similarly to BioPAN, the combination of data- and knowledge-driven analysis, classifies LINEX^2 as a hybrid approach to network analysis.

A recent addition to the available network-based lipid analysis tools is LipidOne 2.0, which offers a range of statistical analyses in addition to its lipid pathway analysis [117]. Its lipid pathway investigation is applied to three structural levels of lipids: classes, molecular species, and lipid building blocks (acyl, alkyl, or alkenyl chains). Similarly to BioPAN, the analysis approach utilizes the statistical model developed by [114,115] and later adopted for lipids by Nguyen et al. [116],

where lipid metabolic pathways are explored by comparing two experimental conditions to produce a list of genes/enzymes associated with the lipid transformations. An advantage of their approach is that users have the flexibility to examine lipid transformations across the three levels of analysis.

While the aforementioned network-based lipid analysis tools represent impressive contributions, a limitation remains in the limited coverage of reactions in terms of both enzymatic and lipid species coverage, restriction to mammalian lipid metabolism as observed in BioPAN, curation, as well as highly specific input nomenclature and presentation of results and information. LINEX^2 was found to not differentiate between forward and reverse reactions and their respective enzymes. LipidOne 2.0 utilizes a specific, limited nomenclature model based on the standardized shorthand annotation nomenclature proposed by Liebisch et al. [100], which includes sum composition and molecular species level, and therefore does not accept nomenclature at the *sn*-position or DB-position level.

1.4 Hypothesis and Objectives

The overall goal of my thesis is to address the lack of easily accessible lipidomic software solutions that can return all possible lipid reactions represented in any given lipidomic dataset. To achieve this goal, I hypothesized that by developing a relational database where lipid reactions are generalized to the class level, chain length specifications are maintained at the molecular level, and lipid nomenclature input is standardized, this integrated software solution would be able to provide a comprehensive reaction list known to date, including enzymatic drivers, for any lipidomic dataset within the categories of glycerophospholipids and sphingolipids. Thus, I outlined three objectives to address this hypothesis:

1. Develop lipid input standardization solutions and lipid reaction processing and inference software linked to a relational database of all known lipid metabolic reactions for GPs and SPs. (**Chapter 2**).

2. Develop a graphical user interface. (**Chapter 3**).

3. Benchmark the reaction list against currently available tools to compare reaction, lipid, and enzyme coverage. (**Chapter 4**).

Chapter 2: Lipid Computational Reaction and Enzyme Database (LipidCRED)

2.1 Objective

This thesis began with the assumption that lipidomic research is hindered by the lack of comprehensive and queryable databases that provide researchers with all the metabolic pathways represented in their datasets. My goal was to bridge this gap by developing LipidCRED, a system encompassing a relational database and backend software. LipidCRED is designed to integrate and generalize lipid reaction data from public sources, include orthologous enzymatic drivers, and address some of the difficulties in lipid syntax with the multiple lipid nomenclatures used in lipidomics and lipid biochemistry that often confound database querying. To achieve this, I aimed to create algorithms for parsing lipid nomenclature, establishing rules for reactant-product carbon-chain matching in lipid reactions, and handling both first and second-order reactions. This was accomplished by building a knowledgebase integrating publicly available data from SwissLipids (for lipid information and classification), UniProtKB (for enzyme annotation), HUGO Gene Nomenclature Committee's (HGNC) (for orthology), and, crucially, Rhea (as the primary source for lipid reactions, cross-referenced by both SwissLipids and UniProtKB). Through this integration and the developed algorithms, LipidCRED was designed to generate comprehensive outputs, including reaction lists and adjacency matrices, representing potential metabolic transformations and their associated enzymes relevant to a user-input lipid list across numerous species.

2.2 Statement of Author Contributions

To address the objectives of this chapter, Q. Alkassir executed the development of the LipidCRED system. This encompassed the creation and subsequent extension of the lipid reaction and enzyme database and development of the Python backend (including the core lipid

reaction processing algorithms). The lipid nomenclature translation system is powered by another database built in collaboration with A. Surendra and E. Thompson.

2.3 Introduction

As reviewed in Chapter 1, while significant progress has been made in lipid identification and quantification, the biological interpretation of large-scale lipidomic datasets remains a bottleneck. Computational tools designed for pathway and network analysis are essential for translating lists of identified lipids into meaningful biological context. However, existing tools often face limitations when analyzing lipidomics data. As discussed in Section 1.3.4, existing tools may have restricted reaction coverage, fail to differentiate reaction directionality, or are constrained by the specific lipid nomenclature formats they can accept. Furthermore, a persistent challenge lies in systematically connecting experimentally identified lipids, often measured at the molecular species level, to relevant enzymatic reactions which may be defined only at the broader lipid class level in knowledgebases. The inherent complexity and variability in lipid nomenclature further complicates the querying and integration of data from diverse sources.

Addressing these gaps requires a dedicated bioinformatics solution capable of integrating reaction knowledge with enzyme information across species, robustly handling diverse nomenclature inputs, and intelligently mapping reactions across different levels of lipid structural detail.

2.4 Process and Results

2.4.1 Overall architecture

The LipidCRED software design can be divided into five main classes: *LipidTranslator*, *DataAccess*, *LipidProcessor*, *ReactionProcessor*, and *OutputGenerator*; where each class handles

a set of operations in the overall workflow of lipid analysis. The overarching driver class for LipidCRED is *LipidAnalysisWorkflow*, which is the program's entry point. This class initializes the five main classes, loads the configuration file, and establishes a connection with the database. The configuration file, *lipid_config.yaml,* is utilized by multiple classes as it stores information on lipid classes and subclasses, carbon-chain matching strategies, second-order reactions, output file name, and database access credentials.

Once the initialization is complete, LipidCRED's analysis workflow can be divided into four steps (Figure 2.1):

1. **Input preprocessing and standardization:** Handled collaboratively by *LipidTranslator* (for translating various lipid nomenclature) and *LipidPreprocessor* (for parsing lipids into their components and categorizing them by subclass, to optimize subsequent matching).
2. **Lipid processing:** Orchestrated by the *LipidProcessor* class, which retrieves lipid reaction information from the database via *DataAccess*.
3. **Reaction processing:** Performed by the *ReactionProcessor* class (utilized by *LipidProcessor*) to determine valid reactions based on defined rules and matching strategies, leveraging the subclass categorization from the preprocessing step.
4. **Output generation:** Managed by the *OutputGenerator* class to produce the final analysis files.

The *DataAccess* class serves as a crucial interface to the underlying database throughout the lipid and reaction processing steps.

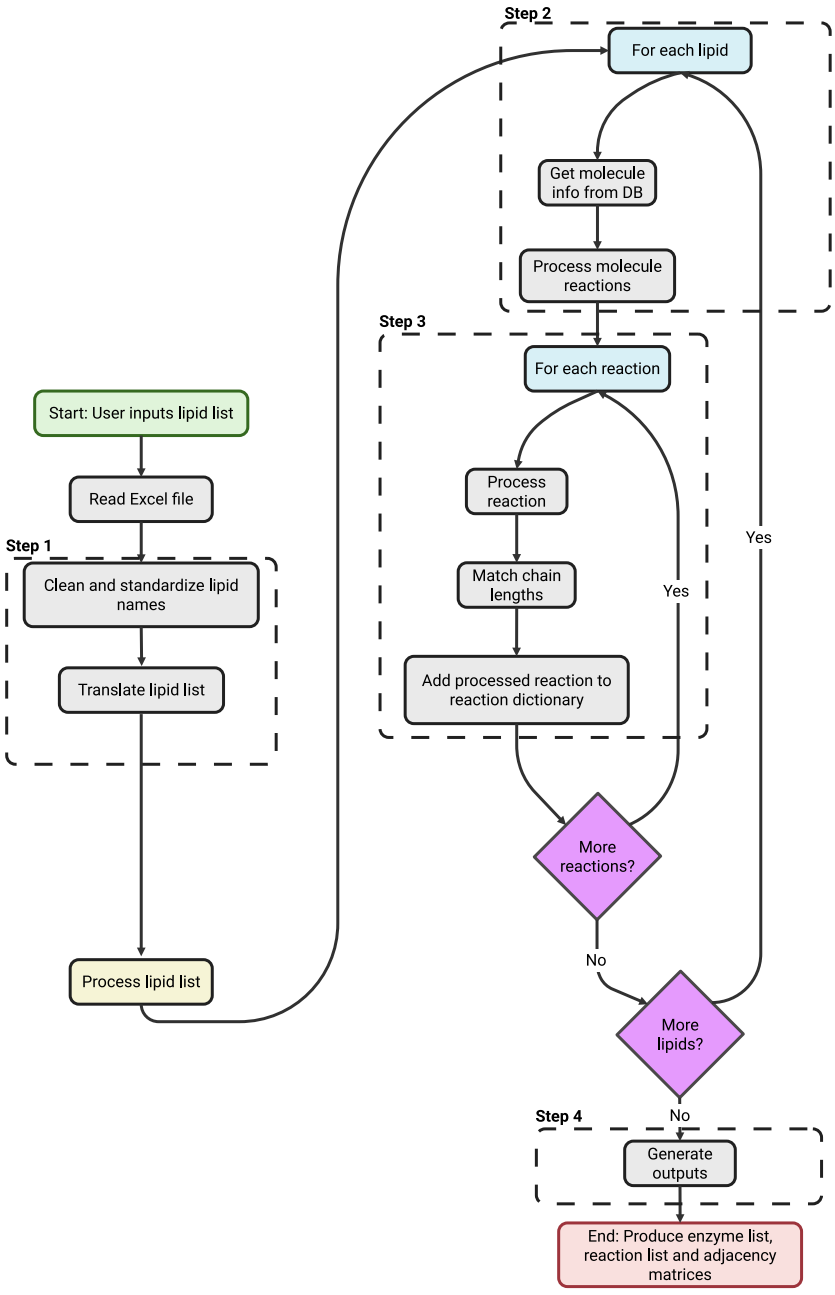


Figure 2.1 LipidCRED analysis pipeline overview. This flowchart outlines the primary stages of the LipidCRED workflow. Input lipids are first standardized and translated. Each standardized lipid then undergoes processing where associated reactions are retrieved from the database. Each reaction is processed ensuring correct chain length matching between the reactant and product. Once all lipids and their reactions have been processed, the results are aggregated in the form of an enzyme list, adjacency matrix and reaction list ready for download.

2.4.2 Input Preprocessing and Standardization

To address the complexity of lipid nomenclature, I developed a multi-component system for robust standardization and parsing. This process utilizes the *LipidTranslator*, *LipidParser*, and *LipidPreprocessor* classes, which I designed along with the *LipidComponents* data structures.

The *LipidTranslator* class is the initial component I implemented for standardizing input lipid names, which utilizes a PostgreSQL database. This translation database integrates primary identifiers and known synonyms from LIPID MAPS® Structure Database (LMSD) [101] and SwissLipids [118]. It contains 42,495 lipid species, encompassing various synonym types such as common names, different abbreviation formats (e.g., PC(16:0/18:0), and PC(34:0)) and systematic names. Additionally, it includes cross-references (unique identifiers) to major chemical and biological databases like PubChem (PubChem CIDs) [96], the Human Metabolome Database (HMDB IDs) [97], and ChEBI (ChEBI IDs) [98]. The purpose of including these diverse identifiers and synonyms was to maximize the probability of mapping user-provided lipid names, regardless of the specific nomenclature used, to the internal standard used throughout LipidCRED. *LipidTranslator* also performs essential cleaning operations, removing positional isomers or stereochemical details (e.g., converting PC(16:0/18:1(9Z)) to PC(16:0/18:1)) to focus on the core structure relevant for carbon-chain matching. Furthermore, it handles specific ambiguous lipid species like 'HexCer' by generating both possible specific isomers (glucosylceramide (GlcCer) and galactosylceramide (GalCer)) for comprehensive analysis.

Following translation and cleaning by *LipidTranslator*, the *LipidPreprocessor* class orchestrates the structural analysis. It utilizes the *LipidParser* framework, which selects the appropriate parsing strategy based on the lipid class/subclass. The parser breaks down the standardized lipid name into its constituent parts (e.g., headgroup, backbone, and fatty acyl chain),

storing this information in structured *LipidComponents* objects. A key function of *LipidPreprocessor* is then to categorize these parsed lipids based on their class/subclass. This categorization creates an internal map (e.g., all 'PC' lipids grouped together, all 'PC' lipids with a '16:0' acyl chain in the *sn*-1 position grouped together, etc.) that significantly optimizes the subsequent reaction processing steps by reducing the search space for potential reaction partners.

The representation of these structural components of lipids relies on a hierarchical lipid component model designed, implemented primarily through the *LipidComponents* class hierarchy. This model is hierarchical in two senses: firstly, specific lipid classes (e.g., sphingolipids, glycerophospholipids) are represented by dedicated subclasses like *SphingoLipidComponents*, *GlycerophosphoLipidComponents* that inherit from a common abstract *LipidComponents* base class. Each subclass encapsulates the distinct structural attributes relevant to its lipid class/subclass (Figure 2.2). For example, *SphingoLipidComponents* tracks the headgroup, backbone, *N*-acyl chain, and optional hydroxyl modification. Secondly, this component-based structure enables a conceptual hierarchy for grouping lipids based on shared features extracted during parsing.

The hierarchical lipid component model (developed as a core part of LipidCRED's design) relies on accurate parsing of lipids and their components. To achieve robust and extensible parsing, I implemented a system leveraging established software design patterns. The *LipidParserFactory* utilizes the Factory pattern to dynamically instantiate the appropriate parser strategy based on the lipid's class/subclass [119]. The Factory maps the lipid class/subclass to a parsing strategy, and determines the appropriate parser for each lipid at runtime. The abstract *LipidParserStrategy* class defines a common interface for all parser implementations, following the Strategy pattern [119]. Each concrete strategy (e.g., *SphingoLipidParserStrategy*, *GlycerophosphoLipidParserStrategy*) implements specialized parsing logic for its respective lipid class, isolating parsing rules, and

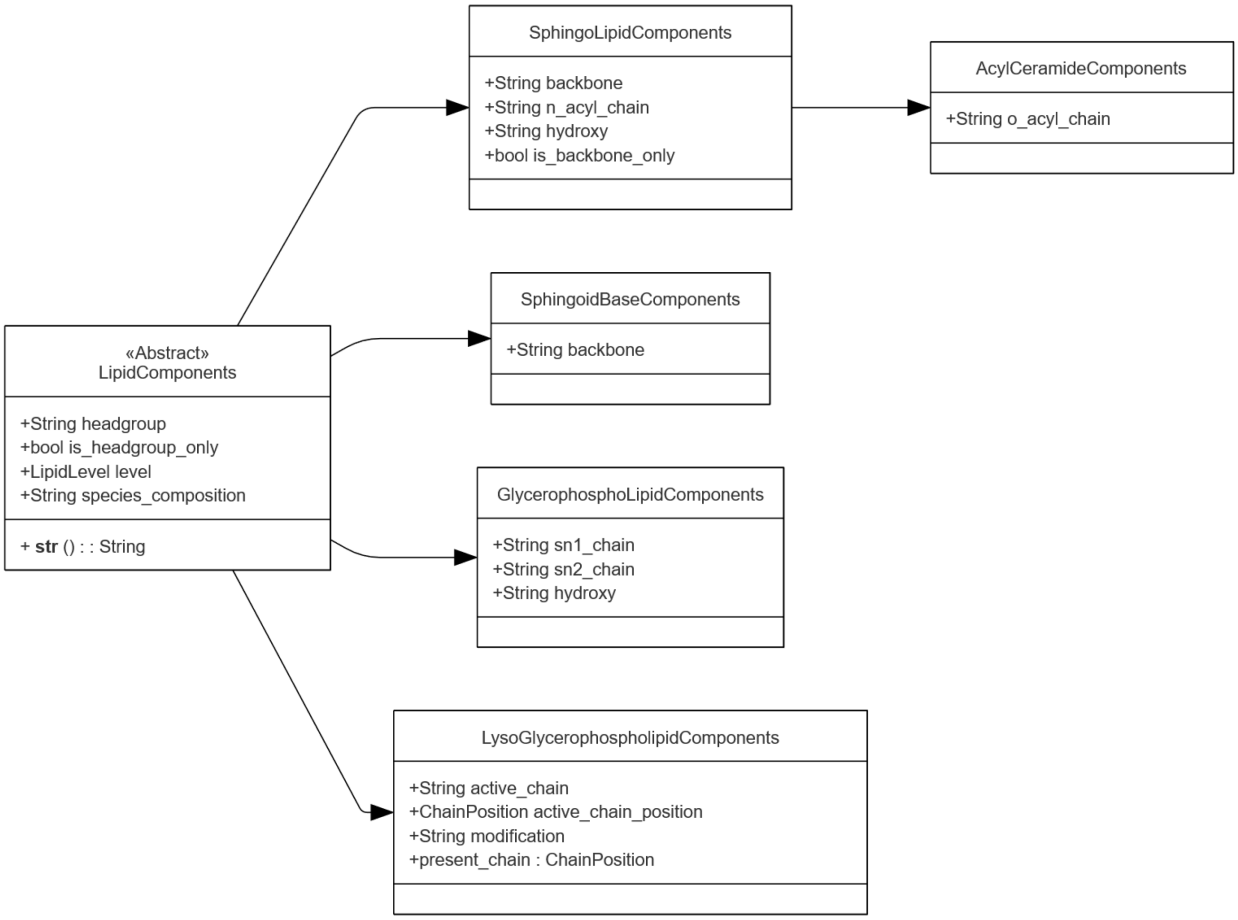


Figure 2.2 The LipidComponents class hierarchy for structured lipid representation. This UML class diagram illustrates the inheritance structure of the *LipidComponents* model designed for LipidCRED. The abstract base class, *LipidComponents*, defines common attributes and methods for all lipid structural representations. Specific lipid types are represented by dedicated subclasses (e.g., *SphingoLipidComponents*, *GlycerophosphoLipidComponents*, *LysoGlycerophospholipidComponents*) which inherit from *LipidComponents* and encapsulate the distinct structural attributes relevant to their respective lipid classes (e.g., backbone, *sn*-1/*sn*-2 chains). Note that *AcylCeramideComponents* inherits from *SphingoLipidComponents*, further specializing the representation for this particular subclass.

allowing for easier addition of new lipid classes. To handle the specialized parsing logic, I implemented regular expression patterns to handle optional modifications (e.g., hydroxylation, deuteration) and class-specific nomenclature rules.

The outcome of this parsing process is a populated *LipidComponents* object specific to each lipid's structure. This structured information directly enables the hierarchical categorization mentioned earlier. For instance, parsing Cer(d18:1/16:0(OH)) using *SphingoLipidParserStrategy* populates a *SphingoLipidComponents* object with: class='Cer', backbone='d18:1', *N*-acyl chain='16:0', hydroxy='OH' (Figure 2.3A). This object can then be categorized conceptually: it belongs to the 'Ceramides' group (based on class), and also to the more specific 'Ceramides with d18:1 backbone' subgroup (based on the backbone component), illustrating the conceptual hierarchy facilitated by the component model (Figure 2.3B).

2.4.3 Lipid Processing

Once the input list of lipids has been standardized and translated as necessary (as described below in section 2.4.2), the *LipidProcessor* class takes over. Its primary role is to iterate through each processed input lipid, retrieve all associated reactions using the *DataAccess* class, and then coordinate the detailed analysis of these reactions via the *ReactionProcessor*. The results – validated reactions connecting lipids within the input set – are progressively stored before final output generation.

An important step bridges the initial translation and the final recording of reactions. Following the *LipidTranslator*'s work (which includes generating potential isomers for ambiguous inputs like 'HexCer' into 'GlcCer' and 'GalCer' to capture all possible reactions associated with either epimer), I initiate the *TranslationMapper* class. This class holds the mapping between the

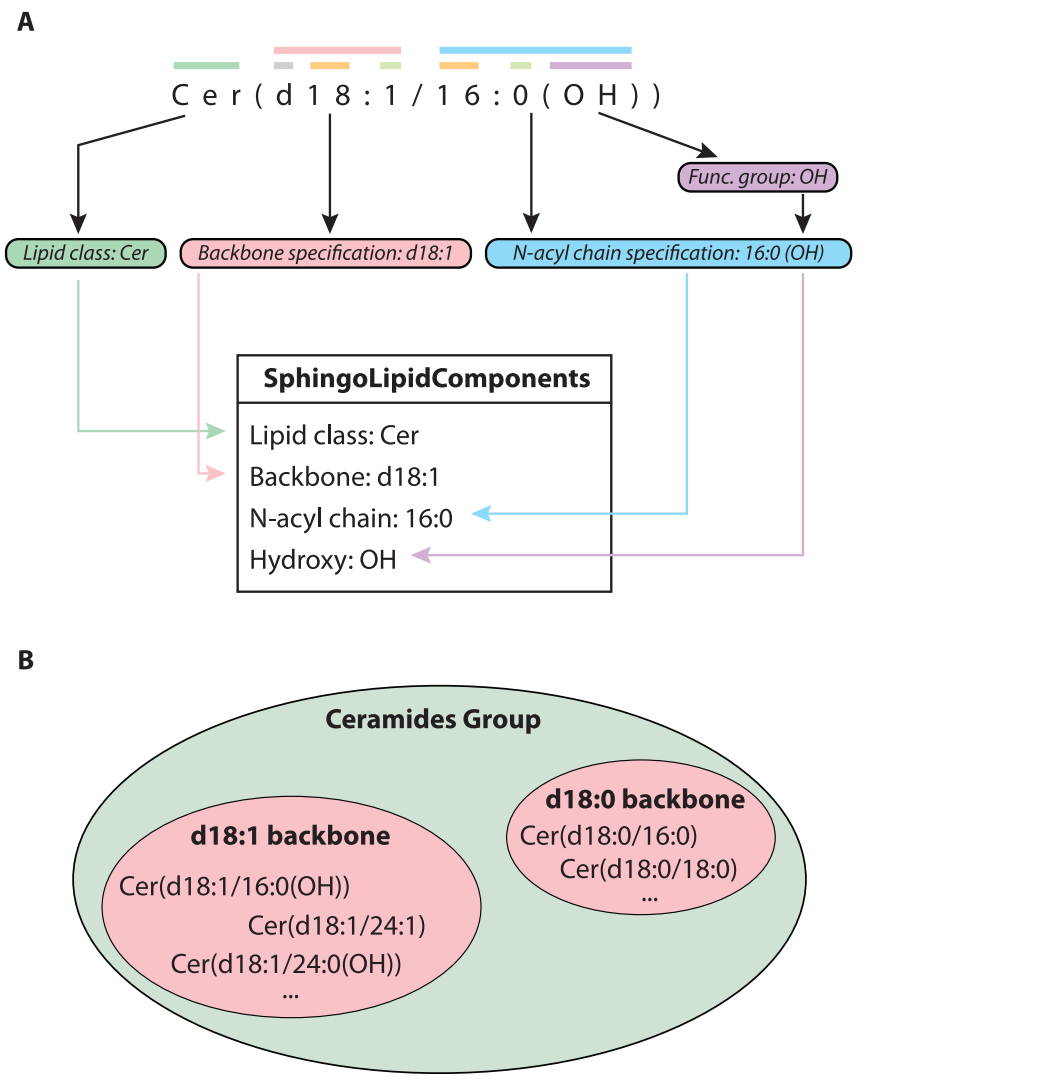


Figure 2.3 Conceptual hierarchical categorization from parsed lipid components. A) Schematic showing how a Cer(d18:1/16:0(OH)) lipid is parsed into a SphingoLipidComponents objects. **B)** The distinct attributes (lipid class 'Cer', backbone 'd18:1') then allow for conceptual grouping at different levels of specificity (e.g., 'Ceramides' groups, 'Ceramides with d18:1 backbone' subgroup).

original user-provided names and all their translated/generated internal representations. Crucially, *TranslationMapper* provides a *reverse_translate_reaction* method. This method is invoked just before a processed reaction is finalized (within *LipidProcessor._add_single_reaction_to_dict*). Its purpose is to convert any internally used translated lipid names (like 'GlcCer') back into the format originally provided by the user (e.g., 'HexCer'). This ensures the final output accurately reflects the user's input data, while still allowing the internal processing to leverage specific isomer information and parsing strategies where necessary.

2.4.4 Hierarchical Reaction Search

To ensure comprehensive retrieval of relevant metabolic pathways, I developed a hierarchical reaction search algorithm. This algorithm leverages the structured lipid classification inherent in SwissLipids to connect specific molecular lipid species provided by the user to their broader class or subclass reactions. The primary source for curated reaction information is the Rhea database [75], which provides standardized biochemical reactions. This information is accessed largely through the cross-references provided within SwissLipids, supplementing directly from Rhea where necessary. Within LipidCRED's database, these reactions are associated with lipids at different levels of structural resolution: the molecular species level; the lipid class level (e.g., 'PC' reactions); or, particularly for sphingolipids, the backbone family level (e.g., reactions specific to 'Cer(d18:1)' derivatives).

A hierarchical search algorithm systematically queries for reactions in the following order for each input lipid (e.g., Cer(d18:1/16:0)) (Figure 2.4):

1. **Molecular species level:** It first searches for reactions directly linked to the specific molecular species identifier in the database (e.g., Cer(d18:1/16:0)).

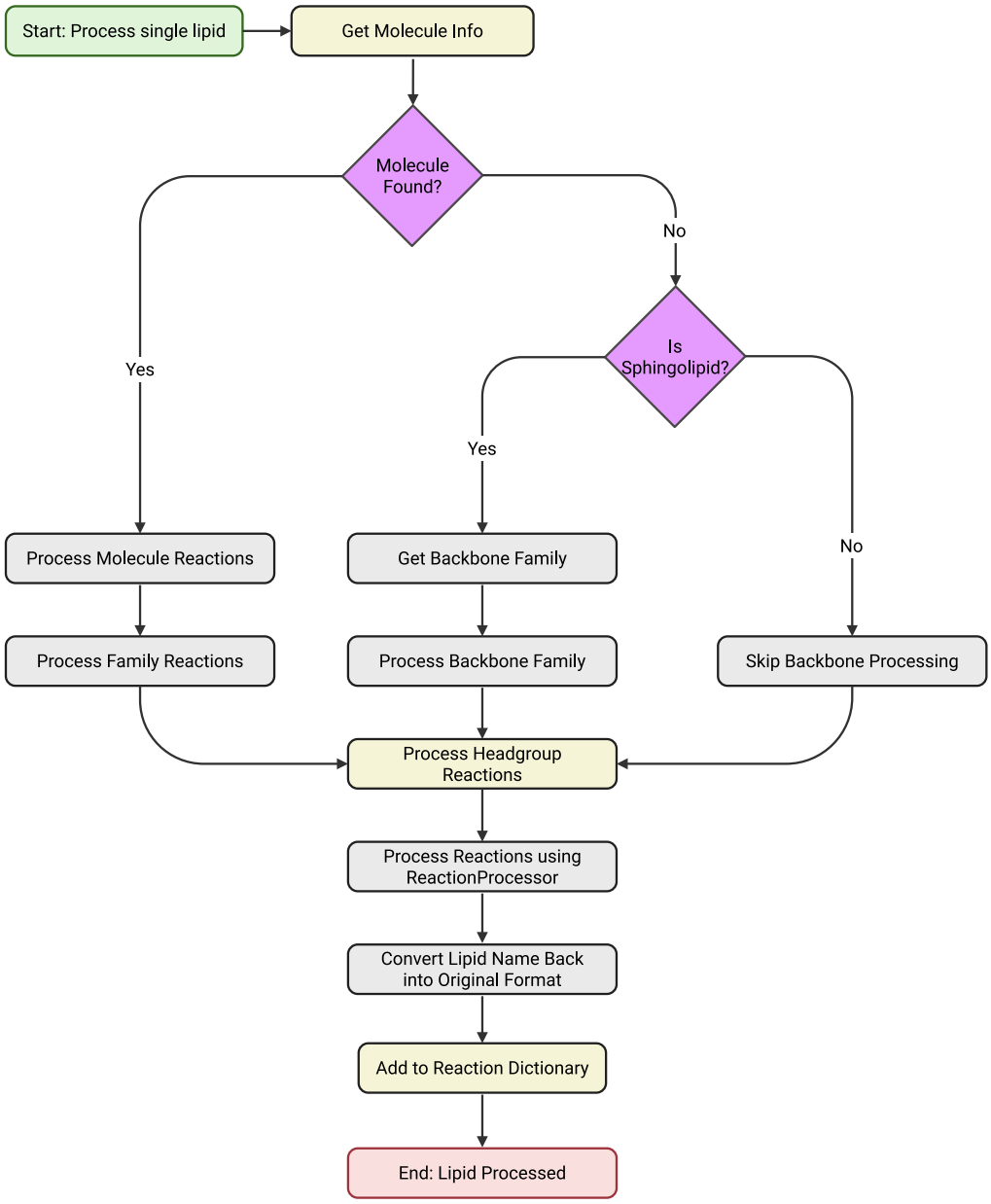


Figure 2.4 Hierarchical reaction search and lipid processing loop. This flowchart illustrates the process for retrieving reactions associated with a single input lipid. Initially, the system searches the database for reactions linked directly to the specific molecular lipid species. Additionally, if the lipid is a sphingolipid, it searches for reactions associated with its backbone family. Finally, reactions linked to the lipid's general class are retrieved. All identified reactions are then passed to the *ReactionProcessor*. Lipid names are converted back to their original name before adding the processed reaction to the growing reaction list.

2. **Backbone level (sphingolipids only):** If the lipid is a sphingolipid, it utilizes the *swisslipids_lipid_class* attribute stored in the database for that molecule, which points to its backbone family (e.g., 'Cer(d18:1)'). It then searches for reactions associated with this backbone family identifier.
3. **Class level:** Finally, it searches for reactions associated with the lipid class. This is achieved both by using the *swisslipids_lipid_class* attribute to find the parent class (e.g., 'Cer') and by using the lipid class parsed directly from the lipid name by *LipidParser*.

When a reaction is found at the backbone or class level (e.g., a generic 'Cer' → 'GlcCer' reaction), the product ('GlcCer') lacks specific carbon-chain information. In these cases, the system uses the *LipidMatcher* component (detailed in 2.4.5) to *synthesize* the predicted molecular product (e.g., GlcCer(d18:1/16:0)) based on the structure of the *original reactant* (Cer(d18:1/16:0)) and the class of the product ('GlcCer'). This predicted product is then evaluated against the user's input list in the subsequent processing step.

2.4.5 Reaction Processing and Product Carbon-Chain Matching

Lipid reactions within LipidCRED are processed based on the number of lipid reactants involved, categorized as first-order or second-order. It's important to note that 'first-order' here specifically refers to reactions involving only *one* lipid species as a reactant recognized by LipidCRED. Non-lipid co-factors or reactants (like ATP, water, or fatty acyl-CoA) present in the underlying biochemical reaction (e.g., Rhea) are not processed or required as input lipids. Second-order reactions involve two distinct lipid molecules present in LipidCRED's database as reactants (Figure 2.5). Distinct processing logic was implemented for each type within the *ReactionProcessor* class (Figure 2.6).

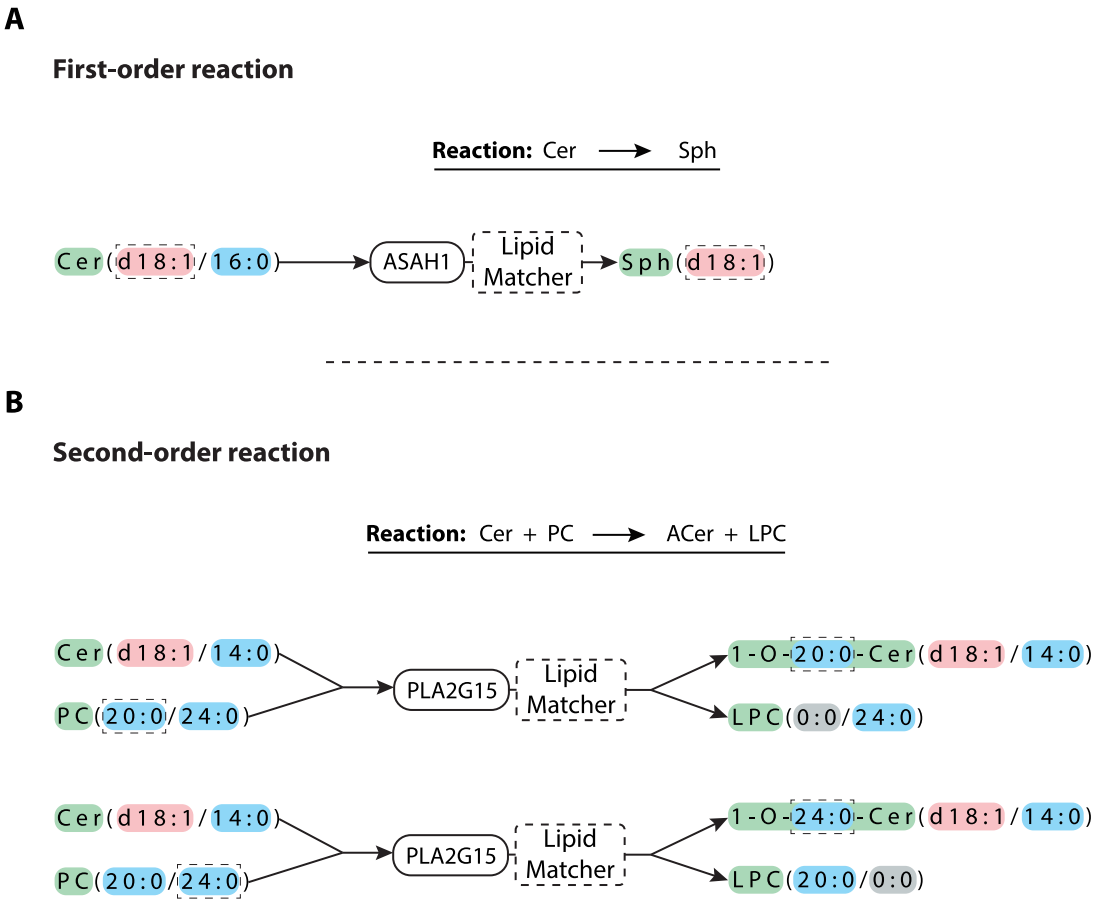


Figure 2.5 Representation and processing of first- versus second-order reactions. A) The conversion of a Cer to an Sph, through the loss of the *N*-acyl chain illustrates a first-order reaction, involving one reactant. This reaction represents a pairwise (one-to-one) relationship, which in a network is seen as one edge connecting two nodes. The enzyme responsible for the catalysis (ASAH1) is placed in between the reactant and product. The components of each lipid are color-coded, green: lipid class, red: backbone, blue: *N*-acyl chain. During the reaction processing, the LipidMatcher module performs the necessary carbon chain length matching strategy, to produce the correctly matched product, Sph(d18:1), which matches the backbone of the reactant, Cer(d18:1/16:0). **B)** Cer can react with PC, where the PC can donate either its *sn*-1 or *sn*-2 hydrocarbon to form an AcylCeramide (ACer), with the respective LPC forming as the byproduct. The Ceramide reactant has the same color-coding as in **A.** The PC is color coded as follows: green: lipid class, blue: *sn*-1 and *sn*-2 chains, and grey represents the absence of a chain (0:0). Cer and PC are combined when representing second-order reactions and are linked to the enzyme (PLA2G15) involved in their reactions, which is subsequently linked to each possible product. During the reaction processing, the LipidMatcher module extract the necessary components from each reactant to form the correctly matched products. The dotted boxes represent the *sn*-1 or *sn*-2 that was used to form the ACer.

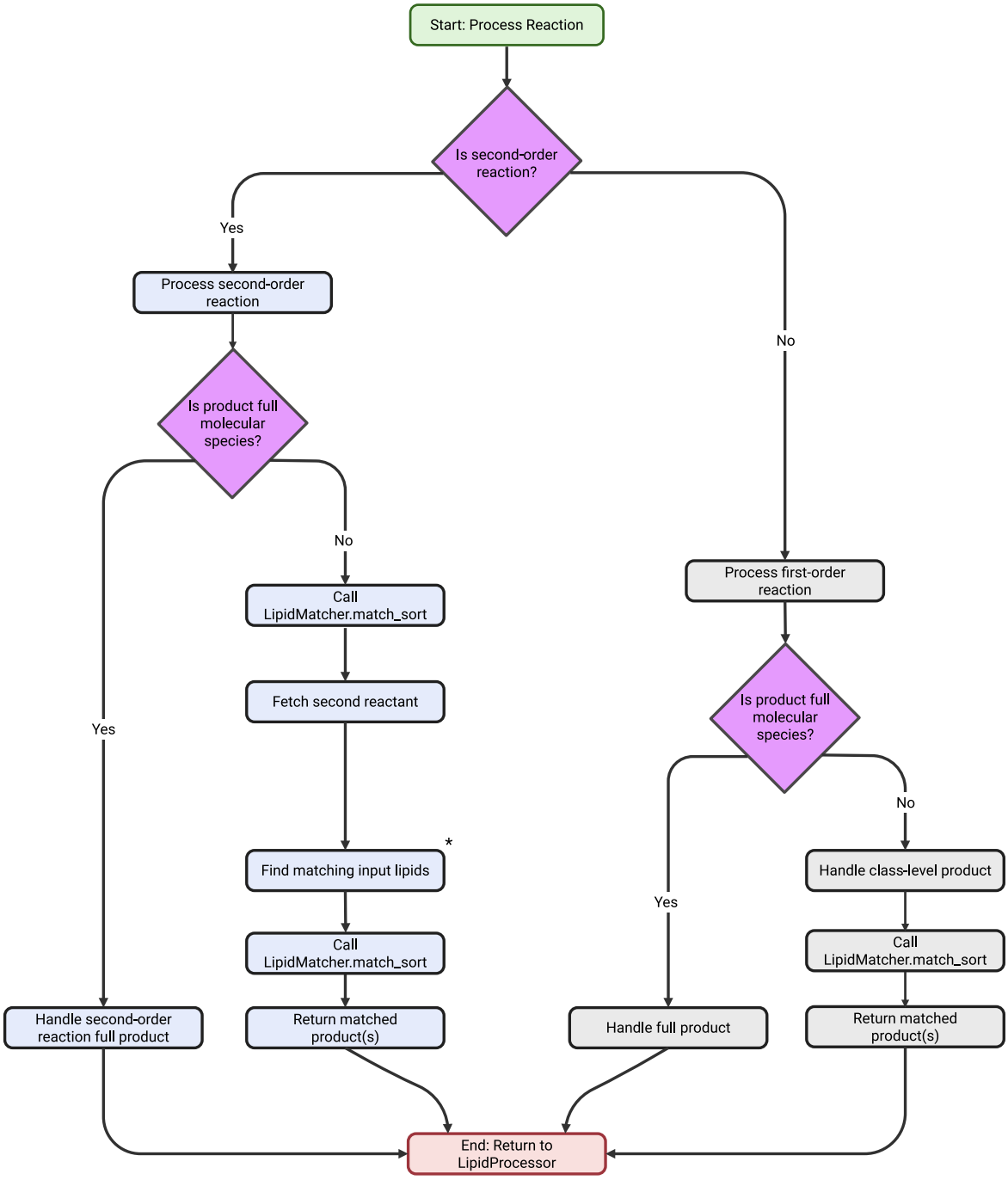


Figure 2.6 Reaction processing logic. This flowchart depicts the distinct paths taken by the ReactionProcessor based on reaction order. First-order reactions proceed directly to product carbon-chain matching using LipidMatcher based on the single reactant. For second-order reactions, if matching is necessary, the input list of lipids is used to fetch possible second reactants to produce a correctly matched product. * refers to the fact that if there are no lipids belonging to the class of the second reactant in the input list, the abbreviated lipid class name is used as a placeholder.

The processing workflow for both reaction orders shares the same initial steps: retrieving the reaction details (reactant(s) class/species, product class/species, enzyme, and Rhea ID) identified by the hierarchical search (2.4.4). The process then diverges:

- **First-order reaction processing:** The *ReactionProcessor* uses the *LipidMatcher.match_sort* method. This method takes the single input lipid reactant and the product definition from the reaction details.
    1. If the reaction was found at the *molecular species level*, *match_sort* confirms the product directly, as no carbon-chain matching is necessary.
    2. If the reaction was found at the *class or backbone level*, *match_sort* synthesizes the specific molecular product by applying the reaction transformation rule (matching rule) and inheriting the necessary carbon-chain information from the single lipid reactant (as described in 2.4.4).

  The resulting specific product is then checked against the user's input list (using the *TranslationMapper* to handle any necessary name reversions) before being added to the results.
- **Second-order reaction processing:** Handling these reactions is more complex as it requires identifying the *second* lipid reactant partner from the user's input list.
    1. The system first identifies the required *class* of the second lipid reactant from the reaction definition stored in the database.
    2. It calls *LipidMatcher.match_sort* using the *first* input lipid reactant and the product definition. The outcome determines the next steps:
        - **Case 1: Product synthesized directly (match_sort returns a specific product):** This indicates the product's structure is

dependent only on the *first* reactant, but the second reactant is still biochemically necessary. The system searches the user's input list for all lipids belonging to the required second reactant class. For each valid pair found, a reaction instance is created using the already synthesized product. It is important to note that if *no* lipids of the required second reactant class are present in the input list, the system uses the *class name itself* (e.g., 'PC') as a placeholder for the second reactant. This ensures the potential reaction pathway and its associated enzyme are still recorded, indicating the biochemical step without a specific partner from the dataset.

- **Case 2: Second-order matching required (*match_sort* returns MatchResult.SUPER_MATCH):** This signifies that the product structure requires carbon-chain information from *both* reactants. The system searches the input list for suitable second reactant partners of the required class. For each potential partner found, it calls *LipidMatcher.super_match_sort* to synthesize the product using components from both lipids.
- **Case 3: Insufficient Information (match_sort returns MatchResult.NO_MATCH):** The first reactant alone does not provide the necessary structural information to determine the product. The system searches the input list for suitable second reactant partners. For each potential partner, it re-attempts product

synthesis using *LipidMatcher.match_sort*, this time leveraging the second reactant's structure.

3. In all second-order cases, any successfully synthesized product is then checked against the user's input list (via *TranslationMapper*) before the corresponding reaction instance (including both identified reactants) is added to the results.

A crucial component enabling this logic is the carbon-chain matching system within the *LipidMatcher* class. The selection of the appropriate matching logic is guided by the lipid_config.yaml configuration file, where mappings are defined between specific reactant-product class transformations (e.g., 'Cer_GlcCer', or 'LPC_PC') and designated matching strategies. Based on the reaction type, *match_sort* (or *super_match_sort*) applies different internal matching strategies. I developed three main categories of strategies:

1. **Complete chain inheritance:** Handles cases where the product fully retains the carbon chains of the reactant. For sphingolipids, this creates a matched product that shares the same backbone and *N*-acyl chain as the reactant, or the same *sn*-1/*sn*-2 chains for glycerophospholipids (Figure 2.7).
2. **Input-dependent:** Used when the product requires adding components not present in the primary reactant (e.g., LPC → PC requires finding suitable PC in the input that matches the LPC's existing chain). This often involves searching the input list for potential products matching the partial structure derived from the reactant (Figure 2.8).

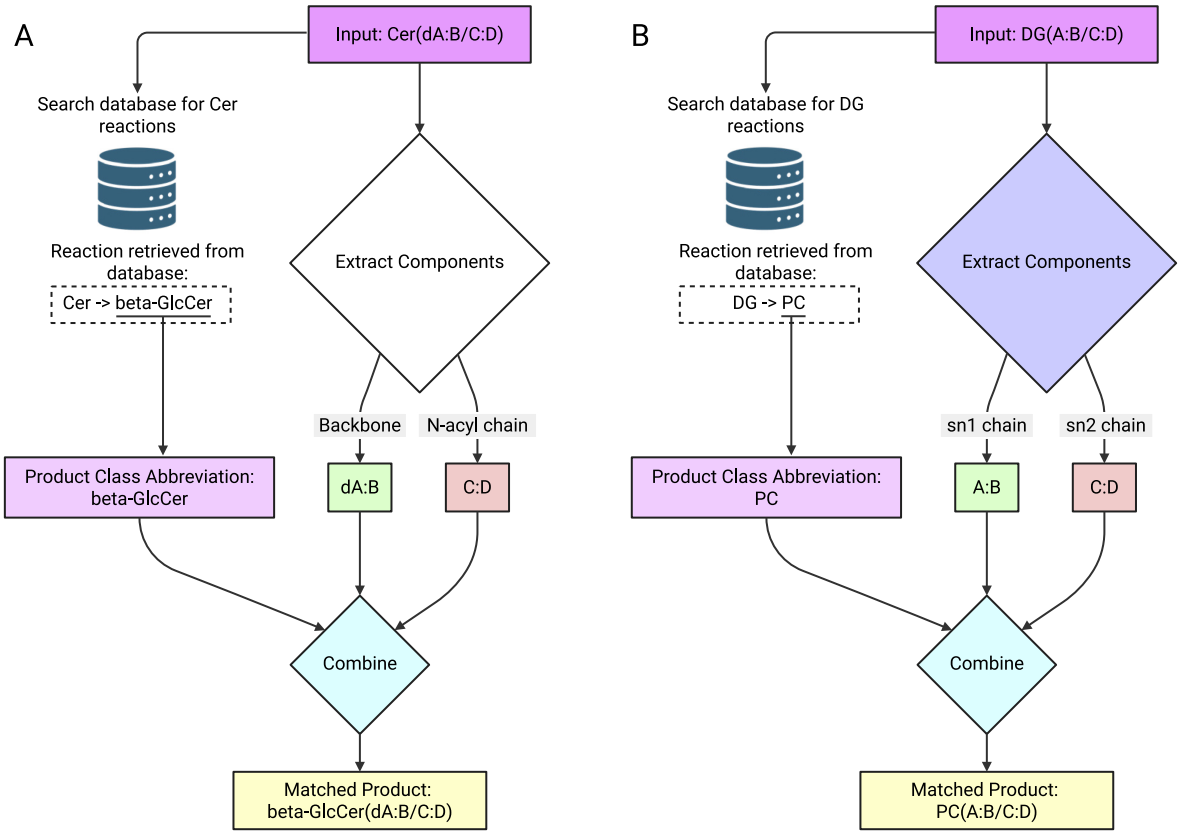


Figure 2.7 "Complete chain inheritance" matching strategy for sphingolipids and phospholipids. This strategy applies when the reaction product completely retains the carbon-chains of the reactant. **A) Sphingolipid example:** For the reaction Cer → beta-GlcCer, the system retrieves the product class ('beta-GlcCer'). It extracts the backbone (xA:B) and *N*-acyl chain (C:D) from the input reactant (Cer(xA:B/C:D)) and combines them with the product class to synthesize the matched product: beta-GlcCer(xA:B/C:D). **B) Phospholipid example:** For the reaction DG → PC, the system retrieves the product class ('PC'). It extracts the *sn*-1 (A:B) and *sn*-2 (C:D) chains from the input reactant (DG(A:B/C:D)) and combines them to synthesize the matched product: PC(A:B/C:D). Database lookup retrieves the reactant-product pair (e.g., Cer → beta-GlcCer).

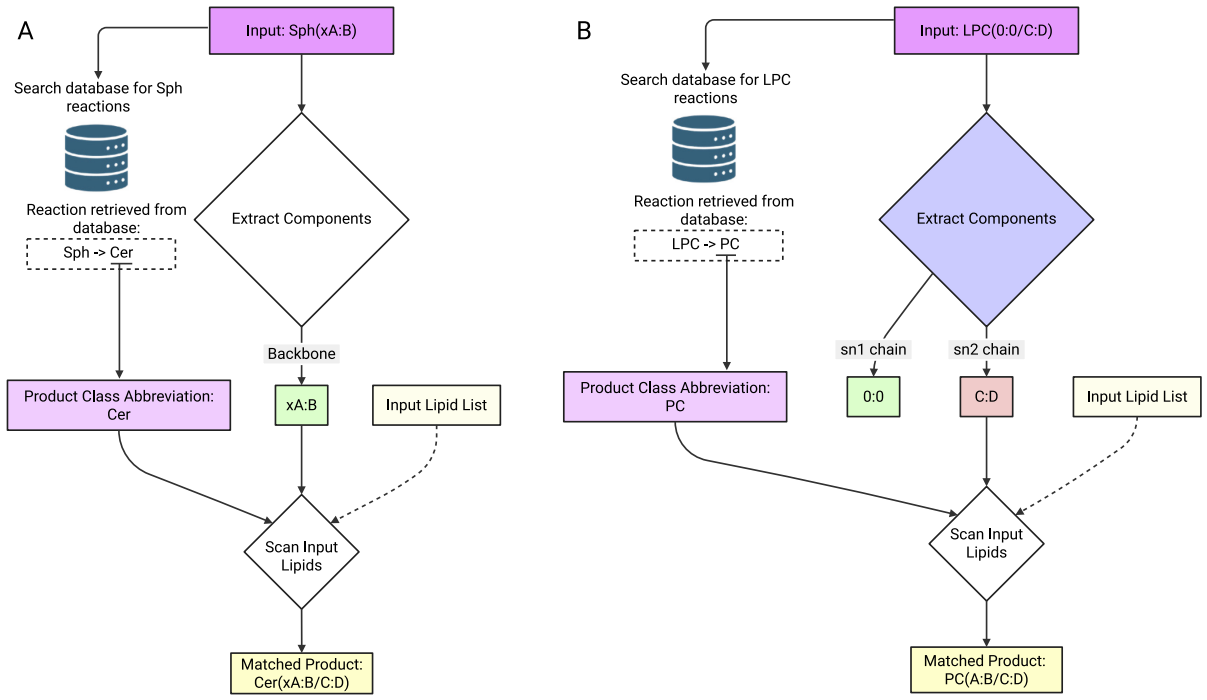


Figure 2.8 "Input-dependent" matching strategies for sphingolipids and phospholipids. A) *N*-acyl matching strategy for sphingolipids: for the reaction Sph → Cer, the system retrieves the product class ('Cer'). It retains the backbone (xA:B) from the input reactant (Sph(xA:B)). It then searches the input list for all Ceramide species that share this backbone (xA:B), identifying potential N-acyl chains (C:D) present in the dataset to form specific matched products like Cer(xA:B/C:D). **B)** Acyl chain addition: for the reaction LPC → PC, the system retrieves the product class ('PC'). It retains the existing acyl chain (e.g., *sn*-2 chain C:D) from the input reactant (LPC(0:0/C:D)). It then searches the input list for PC species that contain this specific chain (C:D at *sn*-2), identifying potential partner chains (A:B at *sn*-1) present in the dataset to form specific matched products like PC(A:B/C:D). Database lookup retrieves the reactant-product pair (e.g., Sph → Cer).

3. **Dual contribution (second-order):** Specifically for reactions where both lipid reactants contribute structural elements to the final product (e.g., Cer + PC → 1-*O*-acyl-Cer + LPC) (Figure 2.9).

Beyond these broad types, I implemented more specific strategies defined in the configuration file, such as N_ACYL_MATCH_DESATURASE for sphingolipid desaturation reactions (e.g., Cer(d18:0/16:0) → Cer(d18:1/16:0)) and BACKBONE_MATCH for reactions involving the hydrolysis of the *N*-acyl chain of sphingolipids, and thus only the backbone is retained (e.g., Cer(d18:1/16:0) → Sph(d18:1) via acid ceramidase activity). The *match_sort* method orchestrates the selection and application of the appropriate underlying logic based on the specific reaction pattern (Figure 2.10)

Finally, before adding any validated reaction (first- or second-order) to the output list, two checks occur:

1. The *TranslationMapper.reverse_translate_reaction* method ensures all lipid names in the reaction correspond to the user's original input nomenclature.
2. The system verifies that the reverse-translated *product* lipid exists within the user's original input list. Only reactions producing lipids found in the input dataset are retained.

This final filtering step ensures the output directly reflects the potential metabolic transformations *within the provided lipidomic dataset*. Duplication of reactions (e.g., from GlcCer and GalCer reactions mapping back to HexCer) is handled inherently by using a dictionary structure for the results, where the unique key combines the reactant(s), product, and specific enzyme, ensuring distinct enzymatic pathways are preserved.

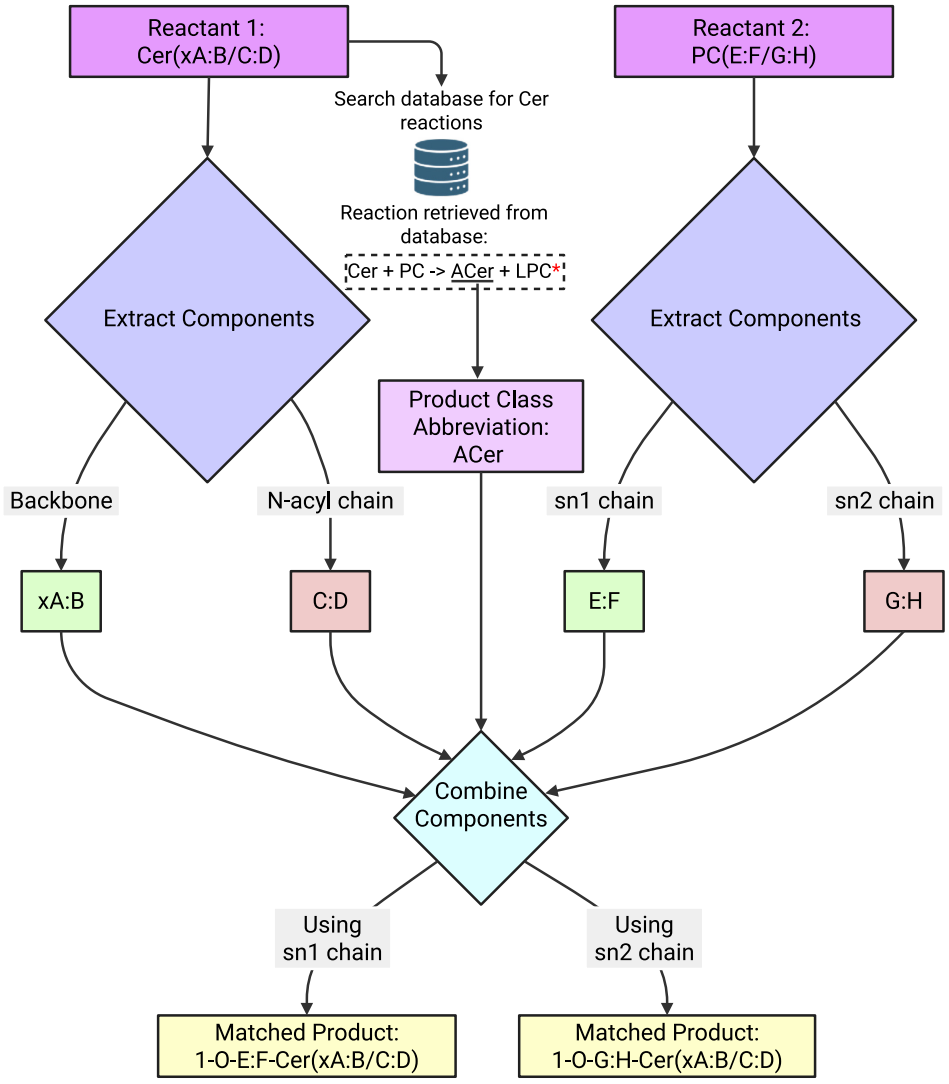


Figure 2.9 "Dual contribution" second-order matching strategy. This illustrates matching for reactions where both reactants contribute structural components. For the reaction Cer + PC → ACer + LPC (*LPC pathway omitted), the system retrieves the product class ('ACer'). It extracts the backbone (xA:B) and N-acyl chain (C:D) from the Cer reactant and combines them with *either* the *sn*-1 (E:F) *or* the *sn*-2 (G:H) chain from the PC reactant, synthesizing two potential matched products: 1-O-E:F-Cer(xA:B/C:D) and 1-O-G:H-Cer(xA:B/C:D).

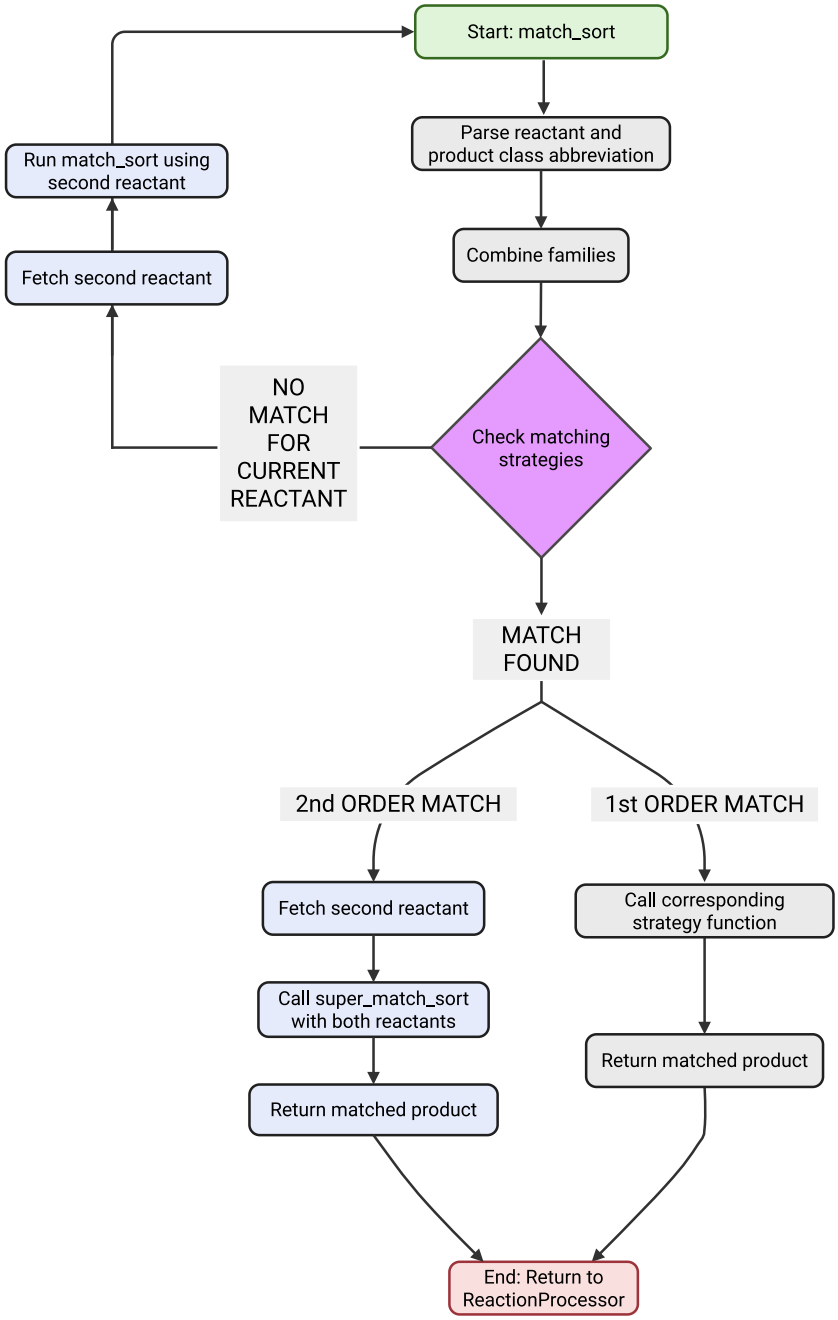


Figure 2.10 match_sort workflow for strategy selection and execution. Each This flowchart details the internal logic of the match_sort function within LipidMatcher. It begins by parsing reactant and product classes and combining them to identify potential matching rules. It checks configured strategies: (1) If a specific **1st Order Match** is found, the corresponding strategy function is called, and the matched product is returned. (2) If a **2nd Order Match** (requiring both reactants for synthesis, i.e., SUPER_MATCH) is identified, the system fetches the second reactant partner from the input list (if available) and calls super_match_sort before returning the product. (3) If **No Match for the current reactant** is found (often in second-order reactions where the first reactant is insufficient, i.e.,NO_MATCH), the system fetches the second reactant and recursively calls match_sort using the second reactant before returning the product.

2.4.6 Output File Generation

Finally, after all input lipids and their reactions have been processed, I developed a modular output generation system to present the results of the lipid analysis workflow in multiple standardized formats. The system's architecture utilizes an *OutputGenerator* protocol, implemented through specialized generator classes: *CombinedOutputGenerator* for network/reaction files, *EnzymeOutputGenerator* for the enzyme summary table, and *TranslateOutputGenerator* for the translation record. This output generation system handles the creation of five key files representing the reaction network:

1. **Enzyme-annotated adjacency matrix:** This matrix displays the pairwise relationships between lipids identified in the input list. Cells indicate reactions, listing the specific enzyme(s) that catalyze the transformation from the row lipid (reactant) to the column lipid (product) (Table 2.1).
2. **Binary adjacency matrix:** A simplified version where cells contain only 1 (reaction present) or 0 (reaction absent) (Table 2.2).
3. **Reaction list:** Provides a detailed tabular summary of all identified reactions, listing the specific reactant(s), product, enzyme(s), and available literature/database references (e.g., Rhea IDs, Digital Object Identifier (DOIs)) (Table 2.3).
4. **Enzyme list:** A concise table summarizing all unique enzymes involved in the identified reactions, listing their gene names alongside their corresponding UniProtKB identifiers (Table 2.4).
5. **Translation record:** Lists each original user-provided lipid name alongside all the standardized or isomer-specific names generated and used internally by LipidCRED during processing

Table 2.1 Enzyme-annotated adjacency matrix output example.

Reactants \ Products	Cer(d18:1/16:0)	S1P(d18:1)	SM(d18:1/16:0)	Sph(d18:1)	beta-GlcCer(d18:1/16:0)	...
Cer(d18:1/16:0)			SGMS1$PC;SGMS2$PC	ACER1;ACER2;ACER3;ASAH1;ASAH2	GBA1$beta-GlcChol;GBA2$beta-GlcChol;UGCG	...
S1P(d18:1)				PLPP1;PLPP2;PLPP3;SGPP1;SGPP2		...
SM(d18:1/16:0)	ENPP7;SGMS1$DG;SGMS2$DG;SMPD1;SMPD2;SMPD3					...
Sph(d18:1)	ASAH1;ASAH2;CERS1;CERS2;CERS3;CERS4;CERS5;CERS6	SPHK1;SPHK2				...
beta-GlcCer(d18:1/16:0)	GBA1;GBA1$Chol;GBA2;GBA2$Chol;GBA3					...
...	...	...	...	...	...	...

Table 2.2 Binary adjacency matrix output example.

Products / Reactants	Cer(d18:1/16:0)	S1P(d18:1)	SM(d18:1/16:0)	Sph(d18:1)	beta-GlcCer(d18:1/16:0)	...
Cer(d18:1/16:0)	0	0	1	1	1	...
S1P(d18:1)	0	0	0	1	0	...
SM(d18:1/16:0)	1	0	0	0	0	...
Sph(d18:1)	1	1	0	0	0	...
beta-GlcCer(d18:1/16:0)	1	0	0	0	0	...
...	...	...	...	...	...	...

Table 2.3 Reaction list output example.

Reactant	Product	Enzyme	Rhea ID	DOI
Cer(d18:1/16:0)	Sph(d18:1)	ACER2	20857;38892	
Cer(d18:1/16:0)	Sph(d18:1)	ASAH2	38892	
Cer(d18:1/16:0)	Sph(d18:1)	ASAH1	38892	
Cer(d18:1/16:0)	beta-GlcCer(d18:1/16:0)	UGCG	12089	
Cer(d18:1/16:0)	Sph(d18:1)	ACER1	20857	
Cer(d18:1/16:0)	Sph(d18:1)	ACER3	20857	
S1P(d18:1)	Sph(d18:1)	SGPP1	27519	
S1P(d18:1)	Sph(d18:1)	PLPP3	27519	
S1P(d18:1)	Sph(d18:1)	SGPP2	27519	
SM(d18:1/16:0)	Cer(d18:1/16:0)	ENPP7	19254;45301	
SM(d18:1/16:0)	Cer(d18:1/16:0)	SMPD1	19254;45301	
Sph(d18:1)	S1P(d18:1)	SPHK2	35848;51496	
Sph(d18:1)	S1P(d18:1)	SPHK1	35848;51496	
beta-GlcCer(d18:1/16:0)	Cer(d18:1/16:0)	GBA2	13270	
beta-GlcCer(d18:1/16:0)	Cer(d18:1/16:0)	GBA1	13270	
Cer(d18:1/16:0)	SM(d18:1/16:0)	SGMS1$PC	18766	
Cer(d18:1/16:0)	SM(d18:1/16:0)	SGMS2$PC	18766;43325	
		GBA1$beta-		10.1074/jbc.RA
Cer(d18:1/16:0)	beta-GlcCer(d18:1/16:0)	GlcChol		119.012502
		GBA2$beta-		10.1074/jbc.RA
Cer(d18:1/16:0)	beta-GlcCer(d18:1/16:0)	GlcChol		119.012502
SM(d18:1/16:0)	Cer(d18:1/16:0)	SGMS1$DG	43321	
SM(d18:1/16:0)	Cer(d18:1/16:0)	SGMS2$DG	43321	

Table 2.4 Enzyme list output example.

Gene Name	UniProt AC
ACER2	Q5QJU3
ASAH2	Q9NR71
ASAH1	Q13510
UGCG	Q16739
ACER1	Q8TDN7
ACER3	Q9NUN7
SGPP1	Q9BX95
PLPP3	O14495
SGPP2	Q8IWX5
ENPP7	Q6UWV6
SMPD1	P17405
SPHK2	Q9NRA0
SPHK1	Q9NYA1
GBA2	Q9HCG7
GBA1	P04062
SGMS1	Q86VZ5
SGMS2	Q8NHU3
CERS1	P27544
…	…

To accurately represent second-order reactions within these formats, I implemented a notation combining the enzyme name and the necessary second lipid reactant, displayed as ENZYME$LIPID. This LIPID component corresponds to the specific second reactant partner identified from the input list during reaction processing, or, importantly, the lipid class name itself if no specific partner from that class was found within the user's dataset (detailed in Section 2.4.5). This notation indicates the requirement of the second lipid species for the reaction to proceed.

2.4.7 LipidCRED Database Overview

The knowledge backbone of LipidCRED lies within its database. I designed a relational database to store lipid reaction information, obtained from various public sources, in a queryable and easily extensible format. Lipid reactions and their associated enzymes are linked with their respective organism and literature references across six tables (Figure 2.11)

2.4.8 Database Design

The goal of this database was to collect, organize, and integrate publicly available lipid reaction data, alongside associated enzymes and literature references, into a searchable and computationally useful format. The SwissLipids knowledgebase served as an ideal initial data source due to its expert curation of lipid species information and metabolism across numerous organisms. As of April 2025, SwissLipids contained annotations for 779,689 lipids, 1,396 enzymes and 7,170 distinct pieces of curated evidence from 1,581 peer-reviewed publications. These annotations cover multiple lipid categories including GP, GL, SP, SL, and FA [118].

A key feature leveraged from SwissLipids is its hierarchical classification system. This system organizes lipid entities based on structural features, progressing from broad categories down to specific molecular species, following established lipid nomenclature standards [100]. This

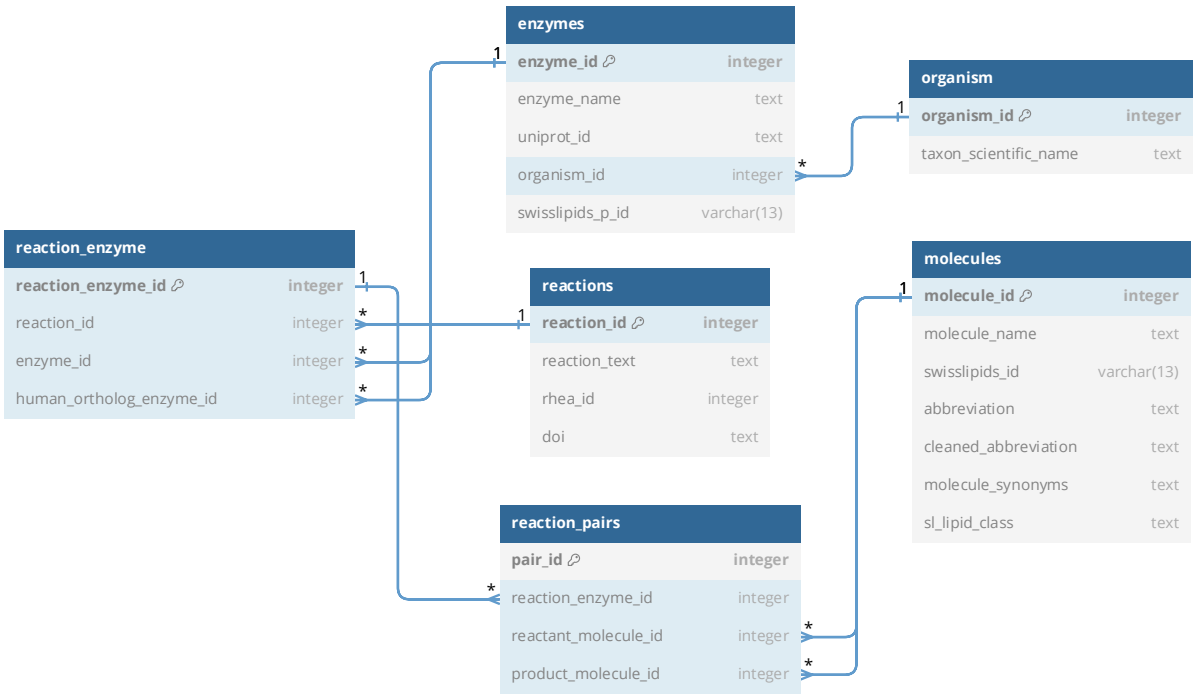


Figure 2.11 Entity-Relationship Diagram of LipidCRED's database. The database consists of six primary tables: organism, enzymes, molecules, reactions, reaction_enzyme, reaction_pairs. The 'enzymes' table is linked to 'organism' through the organism_id foreign key, establishing a connection between enzymes and their source organisms. The 'reaction_enzyme' table serves as a junction between 'reactions' and 'enzymes', allowing for many-to-many relationships between these entities. The 'reaction_pairs' table connects specific reactions (via reaction_enzyme) to the molecules involved, distinguishing between reactants and products. This database diagram was created using the online tool dbdiagram.io.

classification is essential for LipidCRED's ability to generalize reactions and perform the hierarchical reaction search (Section 2.4.4). Furthermore, SwissLipids provides valuable cross-references linking lipid entries to other databases (like HMDB, LMSD), metabolic reactions (via Rhea IDs), and enzymes (via UniProtKB accession numbers (ACs)), facilitating data integration within LipidCRED.

2.4.9 Data Acquisition and Processing

I obtained the primary lipid and reaction data from SwissLipids' publicly available *lipids.tsv* and *enzymes.tsv* files. The *lipids.tsv* file contains numerous lipid descriptors including names, abbreviations, synonyms, lipid classes (crucial for the hierarchical approach), and external database identifiers (LIPID MAPS, HMDB, ChEBI). The *enzymes.tsv* file contained lipid reaction information linked to UniProtKB ACs, gene names, organism taxonomy (Taxon ID), reaction text, and Rhea IDs. The *lipids.tsv* file contained 779,250 lipid species, and *enzymes.tsv* contained 5,713 unique reaction and enzyme combinations across 70 organisms. Both files required preprocessing steps to clean and standardize their contents. I implemented custom parsers to handle formatting inconsistencies, such as shifted columns and unrecognized characters, ensuring data cleanliness before database insertion.

To structure this information, I designed the database schema comprising core tables for organisms, reactions, enzymes, and molecules (Algorithm 2.1). The schema adhered to Third Normal Form (3NF) principles to eliminate data redundancy while maintaining data integrity. The *organisms* table uses the taxonomic identifier (Taxon ID) as its primary key and a biologically relevant unique identifier. The *reactions* table stores unique reaction information, including literature references as Rhea ID and DOI. The *enzymes* table implements referential integrity

Algorithm 2.1 LipidCRED schema definition.

```
CREATE TABLE lipograph.organism (
    organism_id INTEGER NOT NULL PRIMARY KEY,
    taxon_scientific_name TEXT
);

CREATE TABLE lipograph.reactions (
    reaction_id SERIAL PRIMARY KEY,
    reaction_text TEXT,
    rhea_id INTEGER,
    doi TEXT
);

CREATE TABLE lipograph.enzymes (
    enzyme_id SERIAL PRIMARY KEY,
    enzyme_name TEXT,
    uniprot_id TEXT,
    organism_id INTEGER REFERENCES lipograph.organism,
    swisslipids_p_id VARCHAR(13)
);

CREATE TABLE lipograph.molecules (
    molecule_id SERIAL PRIMARY KEY,
    molecule_name TEXT,
    swisslipids_id VARCHAR(13),
    abbreviation TEXT,
    cleaned_abbreviation TEXT,
    molecule_synonyms TEXT,
    sl_lipid_class TEXT
);

CREATE TABLE lipograph.reaction_enzyme (
    reaction_enzyme_id SERIAL PRIMARY KEY,
    reaction_id INTEGER REFERENCES lipograph.reactions,
    enzyme_id INTEGER REFERENCES lipograph.enzymes,
    human_ortholog_enzyme_id INTEGER REFERENCES lipograph.enzymes
);

CREATE TABLE lipograph.reaction_pairs (
    pair_id SERIAL PRIMARY KEY,
    reaction_enzyme_id INTEGER NOT NULL REFERENCES lipograph.reaction_enzyme,
    reactant_molecule_id INTEGER NOT NULL REFERENCES lipograph.molecules,
    product_molecule_id INTEGER NOT NULL REFERENCES lipograph.molecules
);
```

through a foreign key relationship with the *organisms* table. The *molecules* table stores information about each lipid entity, including its SwissLipids ID, name, common abbreviation, synonyms, and its assigned lipid class based on the SwissLipids hierarchy. To ensure consistent nomenclature for internal processing and matching, an additional 'cleaned abbreviation' column was added that uses the original lipid abbreviation column and removes any double bond positional information and standardizes hydroxylation nomenclature.

Following the core tables, I implemented two junction tables to represent the complex relationships in lipid metabolism, *reaction_enzyme* and *reaction_pairs.* The *reaction_enzyme* table represents the many-to-many relationships between reactions and enzymes, where one reaction can be catalyzed by multiple enzymes, or one enzyme catalyzing multiple reactions. The *reaction_pairs* table links specific reaction-enzyme instances to the specific lipid molecules involved as reactants and products, using foreign keys to the *molecules* table.

The process of populating these tables was carefully ordered to maintain referential integrity (Algorithm 2.2). First, the *organisms* table was populated by extracting unique protein taxa from the *enzymes.tsv* file. This was followed by populating the *reactions* table with unique reaction texts and their associated literature references, Rhea ID, and DOI. The *enzymes* table population included establishing a foreign key relationship to the *organisms* table to ensure that each enzyme was correctly associated with its organism. This was particularly important, as some enzymes share gene names but have different UniProtKB ACs due to species differences. The *reaction_enzyme* junction table population involved creating associations between reactions and enzymes based on matching reaction texts and UniProtKB ACs from the *enzymes.tsv* file.

The most complex aspect of the database population was the *reaction_pairs* table, which required sophisticated parsing of reaction texts. I developed an algorithm (Algorithm 2.3) to

Algorithm 2.2 LipidCRED schema population.

```
-- Populate organisms table
INSERT INTO lipograph.organism (organism_id, taxon_scientific_name)
SELECT DISTINCT CAST(protein_taxon as INTEGER), taxon_scientific_name
FROM enzymes_tsv
WHERE protein_taxon IS NOT NULL;

-- Populate reactions table
INSERT INTO lipograph.reactions (reaction_text, rhea_id)
SELECT DISTINCT reaction_text, rhea_id
FROM enzymes_tsv
WHERE reaction_text IS NOT NULL;

-- Populate enzymes table
INSERT INTO lipograph.enzymes (enzyme_name, uniprot_id, organism_id,
swisslipids_p_id)
SELECT DISTINCT t.gene_name, t.uniprotkb, o.organism_id, t.swisslipids_id
FROM enzymes_tsv t
JOIN organism o ON CAST(t.protein_taxon as INT) = o.organism_id;

-- Populate reaction_enzyme table
INSERT INTO lipograph.reaction_enzyme (reaction_id, enzyme_id)
SELECT r.reaction_id, e.enzyme_id
FROM enzymes_tsv t
JOIN enzymes e ON t.uniprotkb = e.uniprot_id
JOIN reactions r ON t.reaction_text = r.reaction_text;
```

Algorithm 2.3 Reaction pairs population process.

```
Function AddReactionPairs(reaction_enzyme_id, reaction_text):
    /* Example input:
       reaction_enzyme_id: 151
       reaction_text: "Ceramide + 1,2-diacyl-sn-glycero-3-phosphocholine => 1,2-
       diacyl-sn-glycerol + Sphingomyelin"
    */

    // Split reaction text into components
    reaction_components = SplitReactionText(reaction_text)

    reactants = reaction_components.left_side   // ["Ceramide", "1,2-diacyl-sn-
    glycero-3-phosphocholine"]
    products = reaction_components.right_side   // ["1,2-diacyl-sn-glycerol",
    "Sphingomyelin"]

    // Generate all possible reactant-product combinations
    combinations = GenerateReactionCombinations(reactants, products)

    For each (reactant, product) in combinations:
        // Look up molecule_id for reactant
        reactant_id = QueryMoleculeId(reactant)
        // Look up molecule_id for product
        product_id = QueryMoleculeId(product)
        // Check if pair already exists
        if not PairExists(reaction_enzyme_id, reactant_id, product_id):
            // Add new reaction pair to database
            Insert into reaction_pairs (
                reaction_enzyme_id,
                reactant_molecule_id,
                product_molecule_id
            ) values (
                reaction_enzyme_id,
                reactant_id,
                product_id
            )

    Commit transaction
```

process reaction texts and create appropriate reactant-product pairs. This process handles both reversible and irreversible reactions, maintains proper references to the *reaction_enzyme* table, and includes validation steps to ensure data integrity across the database tables.

The resulting database structure efficiently represents the complex web of relationships in lipid metabolism while maintaining data integrity and allowing for flexible querying of reactions, enzymes and lipids. The modular and normalized design of the schema allows for extensibility as new lipid reaction data becomes available.

2.4.10 Expanding Enzyme and Organism Coverage

A key objective for LipidCRED was to extend its utility beyond the organisms and reactions directly annotated in the initial SwissLipids data. I employed a multi-step strategy to maximize enzyme and organism coverage:

1. **Initial enzyme data:** The *enzymes.tsv* file from SwissLipids provided an initial set of enzymes and reactions, covering 70 unique organisms.
2. **Orthology-based inference:** I incorporated orthologous enzymes to the database. Orthologs are homologous genes in different species that have originated through speciation from a single gene in a common ancestor [120]. I obtained orthologous enzyme information from the HGNC Comparison of Orthology Predictions (HCOP) database [120]. HGNC's HCOP database incorporates multiple sources of gene and ortholog information in order to generate orthology assertions and construct the final HCOP orthologs table (Figure 2.12). I downloaded all human orthologs data from HGNC, covering orthology assertions between human genes and those in 19 other species (including rat, mouse, chimp, and cattle). To ensure high confidence, I cross-

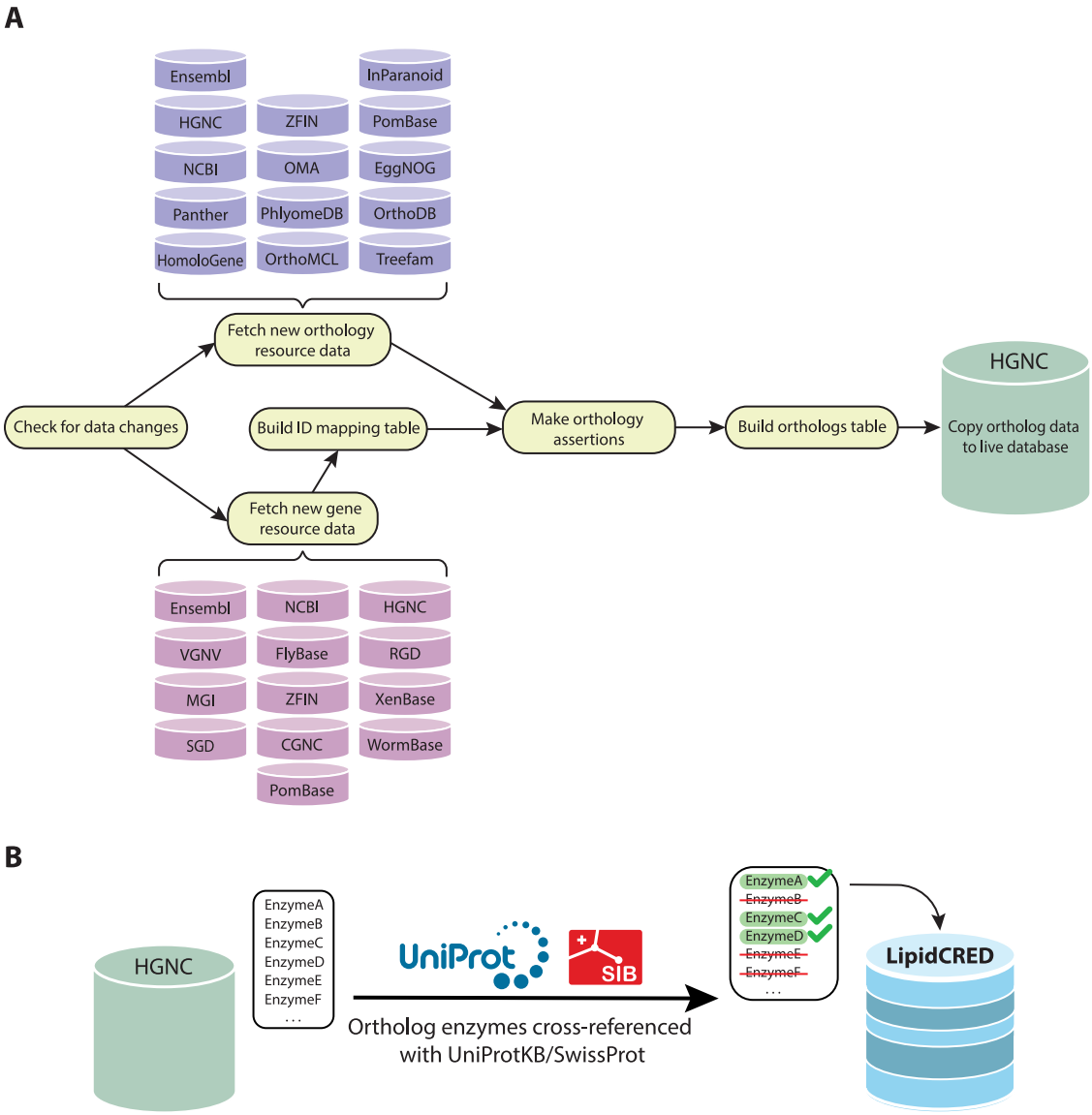


Figure 2.12 HGNC orthology prediction pipeline and incorporation of orthologous enzymes into LipidCRED. A) This flowchart illustrates the process HGNC follows to integrate orthology data from multiple sources, including Ensembl, NCBI, EggNOG, PhylomeDB, OrthoMCL and many others seen above in purple and pink. The pipeline begins by fetching and comparing new data from various orthology and gene resources. It then builds an ID mapping table, generates orthology assertions, and constructs the final HCOP orthologs table. The resulting data is used to update the live database. **B)** Orthologous enzymes for human genes that are involved in lipid metabolism were obtained from HGNC and subsequently cross-referenced with the manually curated knowledgebase of UniProtKB/Swiss-Prot. Only those enzymes that passed this cross-reference were added into LipidCRED's database.

referenced these predicted orthologs with the UniProtKB/Swiss-Prot database [121], retaining only those orthologous enzyme entries that have undergone manual review and annotation within Swiss-Prot.

3. **Targeted expansion via Rhea IDs and UniProtKB:** During the development and testing of LipidCRED using various sample lipid lists, I encountered situations where a known reaction (present in the *reactions* table via its Rhea ID from SwissLipids) lacked an associated enzyme entry for the specific organism being queried. To address these gaps incrementally, I developed a process: when an enzyme was missing for a known reaction in a target organism, a script queried the UniProtKB/Swiss-Prot database using the reaction's Rhea ID. This query retrieved reviewed enzyme entries linked to that Rhea ID across any organism available in UniProtKB/Swiss-Prot. These newly found enzymes and their associated organism information were then added to the LipidCRED enzymes table.

This multi-step strategy led to the current scope of LipidCRED, which currently encompasses 3,075 enzyme entries spanning 519 unique organisms (Figure 2.13).

2.4.11 Reaction Generalization

Many of the reactions obtained from SwissLipids were found to be at the molecular species level. As a result, a user would need to input the specific lipid molecular species involved in that reaction to see its reaction information. However, following the general principle of lipid enzyme specificity, where lipid enzymes catalyze lipid class reactions [18], I implemented an algorithm to generalize reactions to the lipid class level. To generalize a reaction, the hierarchical classification system of SwissLipids (and stored in LipidCRED's *molecules* table via the *swisslipids_lipid_class*

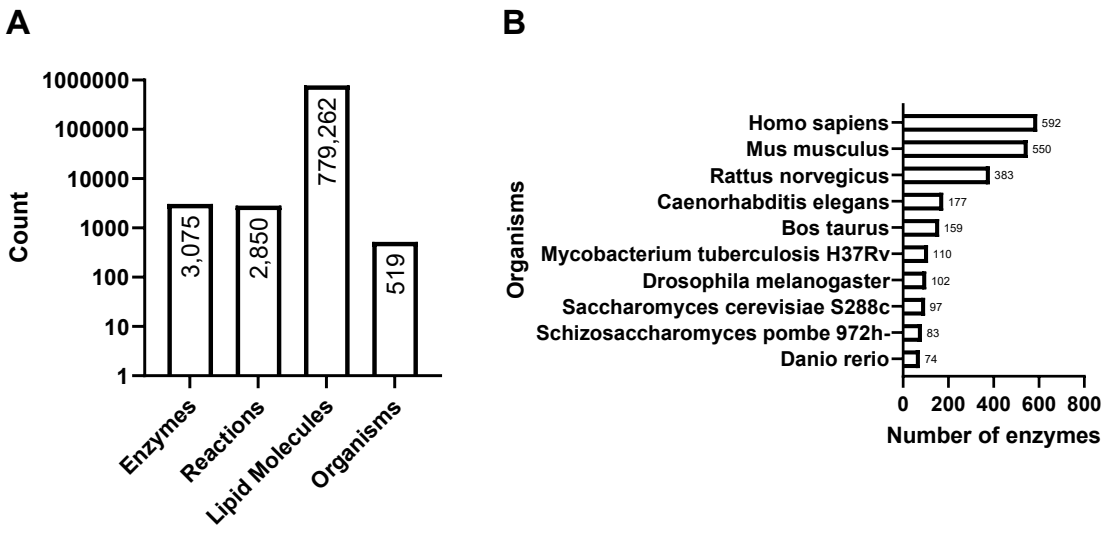


Figure 2.13 Figure 2.13 Overall database statistics. A) Overall database statistics showing the total number of unique entries for enzymes, reactions, lipid molecules, and organisms in the LipidCRED database. **B)** Distribution of enzymes across the top 10 organisms represented in the database. Homo sapiens leads with 592 enzymes, followed by Mus musculus (550) and Rattus norvegicus (383).

attribute) was utilized. For reactions originally defined at the molecular species level (e.g., the conversion of Cer(d18:1/16:0) to Sph(d18:1) catalyzed by ASAH1), the algorithm identifies the parent lipid *class* for both the reactant ('Cer') and the product ('Sph') using the hierarchy. It then stores this generalized class-level reaction (Cer → Sph, catalyzed by ASAH1) as an additional entry in the database, linked appropriately in the *reaction_pairs* table using the class-level identifiers from the *molecules* table.

The key benefit of this generalization is that any input lipid can now be associated with the reactions of the lipid class it belongs to during the hierarchical reaction search (Section 2.4.4). This allows LipidCRED to infer potential metabolic steps even for lipid species that were not the exact ones used in the original curated reaction evidence, including newly discovered lipids or those simply not yet curated at the species level for that specific reaction. When such a generalized reaction is matched to a specific input lipid, the product remains initially defined at the class level (e.g., 'Sph'). As described in Section 2.4.5, the *LipidMatcher* then synthesizes the specific molecular product (e.g., 'Sph(d18:1)' derived from 'Cer(d18:1/16:0)') based on the structure of the actual input reactant.

2.4.12 Query Optimization and Performance

The hierarchical reaction search algorithm, while comprehensive, requires multiple database queries for each input lipid as it traverses from the molecular species level, through backbone (for sphingolipids), up to the class level. To mitigate the potential performance bottleneck associated with repeated database access, especially with large input lipid lists, I implemented caching for database query results using the Least Recently Used (LRU) strategy [122].

LRU caching is a standard optimization technique where the results of recent, computationally expensive operations (in this case, database queries) are temporarily stored in memory. When the same query is requested again, the result is retrieved instantly from the cache. The 'Least Recently Used' aspect refers to the policy for managing the cache size: when the cache is full and a new result needs to be stored, the oldest (least recently accessed) cached item is discarded.

This strategy proved particularly effective within LipidCRED for several reasons. Queries for molecule information (*get_molecule_info*), reaction data associated with a specific molecule or class (*get_reactions*), and lipid family details (*get_family_molecule*) are frequently repeated during a single analysis run. For instance, multiple input lipids belonging to the same class (e.g., various PCs) will all trigger the same class-level reaction queries. Similarly, lipids sharing backbone structures (like sphingolipids) might trigger repeated family-level queries. By caching the results of these queries after their first execution, subsequent requests for the same data are served almost instantaneously from memory. As demonstrated in Table 2.5, this caching implementation yielded substantial performance improvements for key methods within the reaction processing pipeline, significantly reducing the overall analysis time, especially for larger datasets.

2.4.13 Database and Software Maintenance

The LipidCRED database schema and software architecture were designed with modularity and extensibility in mind. The normalized database schema (Section 2.4.9), adhering to 3NF principles, isolates different types of information (lipids, enzymes, reactions, and organisms) into distinct tables. This structure minimizes data redundancy and makes it conceptually simpler to add

Table 2.5 Performance improvement of specific functions and overall workflow after implementation of database query result caching.

Query Type	Without Cache (s)	With Cache (s)	Performance Improvement
get_molecule_info	0.049308	0.004642	90.59%
get_reactions	0.011981	0.001155	90.36%
get_family_molecule	0.055469	0.005575	89.95%
get_second_reactant	0.010538	0.000951	90.97%
Full run *	618.553757	112.015926	81.89%

*Note: Total processing times measured for the complete LipidCRED workflow using an input dataset (193 lipids) for Human (organism ID 9606). Benchmarks were run on the Linux server hosting LipidCRED.

new lipid entries, enzyme annotations, or reaction data from future sources without requiring extensive modifications to existing tables.

Similarly, the software architecture, utilizing design patterns like Factory and Strategy for parsing (Section 2.4.2) and distinct classes for different workflow stages (Section 2.4.1), promotes modularity. This separation of concerns makes the codebase easier to maintain and facilitates the future integration of new functionalities. For instance, adding support for a new lipid class would primarily involve creating a new *LipidParserStrategy* and potentially a new *LipidComponents* subclass rather than altering large portions of the existing code. While a fully automated pipeline for pulling updates from online databases is not currently implemented, the modular design principles employed throughout LipidCRED provide a strong foundation for such future enhancements and ensure the system's long-term maintainability and adaptability.

2.5 Discussion

The development of LipidCRED, including the backend software and relational database described in this chapter, was motivated by the need for more integrated and comprehensive tools to interpret lipidomic datasets within the context of metabolic pathways. While numerous resources address individual aspects of lipid biology, LipidCRED was specifically designed to bridge critical gaps related to nomenclature heterogeneity, the connection between molecular lipid species and lipid class reactions, and the integration of enzyme information across diverse organisms, thereby addressing limitations identified in Chapter 1.

A central achievement of LipidCRED’s implementation and design is the creation of a unified relational database in which I integrated curated information from key resources like SwissLipids, Rhea, UniProtKB, and HGNC. This database currently provides a foundation

encompassing data on approximately 780,000 lipid species, 3,075 enzymes, and 519 unique organisms. To interact with this database and process input lipid lists, I developed a modular software built upon key functionalities:

- A multi-stage process for robust nomenclature handling, using *LipidTranslator* for synonym mapping/cleaning, *LipidParser* (employing Factory/Strategy patterns selected for extensibility) for accurate structural parsing, and *LipidComponents* for hierarchical representation of lipids.
- A novel hierarchical reaction search algorithm designed to systematically query the database at multiple lipid structural levels.
- A reaction generalization strategy implemented to infer class-level reactions from species-level data, thereby enhancing lipid-enzyme associations.
- Dedicated logic for processing first- and second-order reactions, featuring a sophisticated and configurable *LipidMatcher* system built to predict specific molecular products based on defined carbon-chain matching rules.
- A multi-faceted strategy employed to expand enzyme and organism coverage, combining initial data, orthology inference, and targeted Rhea-ID-driven expansion.
- Performance optimizations, such as LRU caching to manage the computational demands.
- Standardized outputs to facilitate downstream analysis.

This approach to building LipidCRED offers several advantages. By consolidating diverse data sources into an optimized structure, it provides a more comprehensive knowledgebase than available elsewhere. The nomenclature translation system, combining synonym mapping with

lipid-class-aware parsing, allows LipidCRED to handle a wide variety of input formats robustly. A key feature of LipidCRED stems from the synergistic relationship between the hierarchical reaction processing logic and the carbon-chain matching algorithm. This framework links specific input molecular lipid species to relevant class-level reactions and then predicts the specific molecular product via *LipidMatcher*'s algorithm, thereby significantly enhancing LipidCRED's ability to find lipid-enzyme associations from any input lipid dataset. Furthermore, the broad organism coverage, achieved through multi-source integration, increases the tool's applicability, while also providing accurately filtered results focused on reactions relevant within the user's specific dataset and organism of choice.

It is important to note that LipidCRED's accuracy is fundamentally tied to the veracity of the underlying public databases. The enzyme expansion, while effective, was performed incrementally, and thus a systematic expansion querying UniProtKB for all Rhea IDs remains a potential future task. The current lipid category scope (GP and SP) needs to be broadened, expanding the parsing and matching rules to include fatty acid transformations for example. Finally, while I designed the system for modularity (Section 2.4.13), database updates currently require manual intervention. Developing the automated update pipeline I initially envisioned remains an important future goal to ensure long-term accuracy and relevance.

In conclusion, the LipidCRED backend and database presented here provides a powerful, integrated engine for exploring potential lipid metabolic reactions and their associated enzymes. Its nomenclature handling, hierarchical reaction mapping, product carbon-chain prediction, and multi-organism enzyme integration address key challenges in the field. User interaction with this engine is facilitated through a graphical user interface developed using R/Shiny, the details of which are presented in Chapter 3. While I foresee future work to further expand its scope and

automation, LipidCRED currently offers a valuable resource for translating lipidomic data into biochemical context.

Chapter 3: LipidCRED User Interface / Web application

3.1 Objective

My objective in this chapter was to design and implement a user-friendly web interface for LipidCRED using RShiny, providing researchers with an efficient platform to upload, process, and visualize reactions for their lipidomic datasets. The overarching objective was to deliver a comprehensive output package, containing a reaction list (inclusive of literature evidence when available), reaction adjacency matrix (inclusive of enzymes), binary adjacency matrix, list of all enzymes involved with user's lipid set, and a translation of lipid names into standardized nomenclature, thereby empowering users to explore and analyze their data with ease and flexibility. The result of this endeavor is a robust and scalable software solution for the investigation of lipid enzymes underlying lipidomic abundances.

3.2 Statement of Author Contributions

Q. Alkassir designed and implemented the LipidCRED web interface and underlying data processing functionalities described in this chapter. A. Surendra provided essential support in deploying the application to the server. E. Hashimoto-Roth kindly provided a template for the R/Shiny user interface layout.

3.3 Introduction

To bridge the gap between complex backend computations and the end-user researcher, a graphical user interface is essential. Command-line tools, while powerful, can present a barrier to entry for scientists without extensive computational expertise. Therefore, a web-based application for LipidCRED was designed. This approach offers platform independence (accessible via web browser), simplifies the input/output process, and allows for a guided, step-by-step workflow.

My primary design philosophy for the LipidCRED interface was centered around clarity, efficiency, and ease of use. I aimed for an intuitive layout where users could quickly understand how to upload their data, set parameters, run the analysis, and retrieve the results. Providing clear feedback during processing and delivering results in multiple, standardized formats were also key considerations.

For the implementation, I chose the R/Shiny framework [123,124]. R provides a powerful environment for data handling, and the Shiny package allows for the rapid development of interactive web applications directly from R code, significantly accelerating the user interface (UI) development process. The *reticulate* package [125] enables seamless integration between the R/Shiny frontend and the Python-based analytical backend (detailed in Chapter 2), allowing the UI to trigger computations and retrieve results effectively.

This chapter details the implementation specifics of the LipidCRED web application, walks through its features and usage from a user perspective, and discusses its availability.

3.4 Implementation

The LipidCRED Shiny application was structured using best practices for organization and maintainability, leveraging the capabilities of the *shinydashboard* layout [126], which provides a familiar and organized interface with a sidebar for navigation and a main body for content display.

Architectural Choices: To ensure maintainability and clarity, a modular design strategy was adopted. The application's UI is divided into distinct sections: Home, Analyze, Sample Data, Troubleshoot, and Citing), each managed by separate R scripts within a dedicated *modules/* directory (Figure 3.1). This separation allows for easier development, testing, and modification of individual application components. Similarly, server-side logic for distinct tasks, such as handling

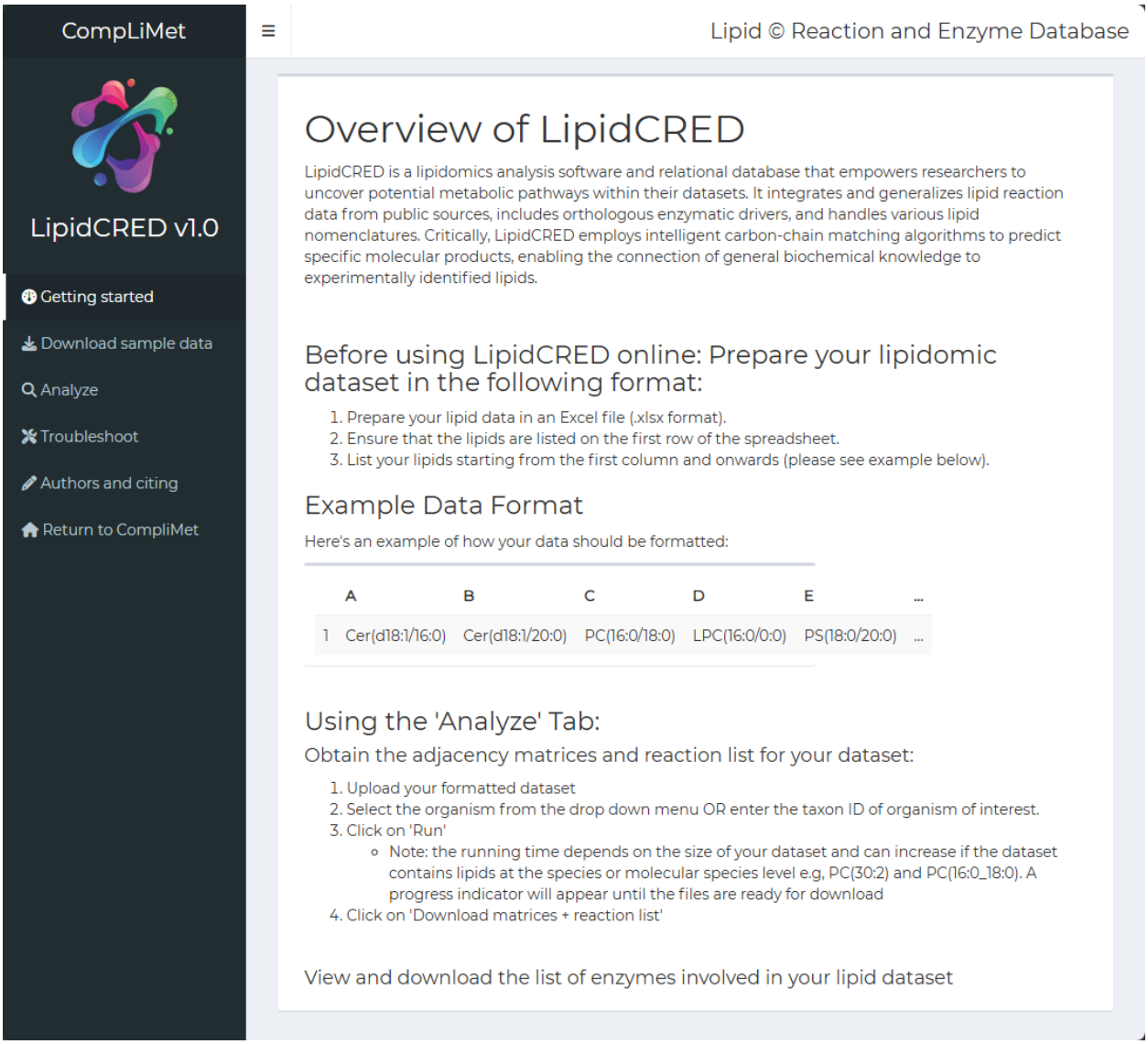


Figure 3.1 LipidCRED Web Application Interface Overview. Screenshot showing the main dashboard layout of the LipidCRED application, built using R/Shiny and the *shinydashboard* package. The image displays the consistent structure featuring the sidebar navigation menu (left) with links to different sections ("Getting Started", "Analyze", "Sample Data", etc.) and the main content panel (right).

user file uploads, initiating the backend analysis, preparing results for display (e.g. the enzyme table using the *DT* R package [127]), and packaging files for download (using the zip package), is encapsulated within dedicated functions organized in a *functions/* directory. This functional separation promotes code reusability and readability.

Backend Integration: A critical aspect of the implementation was connecting the R/Shiny frontend to the core LipidCRED analysis backend, developed in Python (Chapter 2). The *reticulate* R package was employed [125], which allows R to call Python functions and exchange data. When the user initiates the analysis, specific R functions within the Shiny server logic invoke the main Python *LipidAnalysisWorkflow* class, passing necessary parameters like the input file path and selected organism ID. The Python backend performs the computations and writes output files (matrices, reaction lists, enzyme lists, and translation records) to a temporary location. The Shiny application then accesses these files to display the results (like the interactive enzyme table) or package them for download.

User Experience and Reactivity: Shiny's reactive programming model was used to create a dynamic and responsive interface. User actions, such as uploading a file or clicking the "Run" button, trigger specific server-side operations. Reactive values track the state of the application (e.g., whether processing is complete), allowing UI elements (like download buttons or results tables) to update automatically when ready. Visual feedback during processing, including messages and loading spinners (using *shinybusy* [128]), keeps the user informed about the application's status.

This implementation approach, combining R/Shiny's interactive web capabilities with *reticulate* for Python integration and adhering to modular design principles, resulted in a

functional, maintainable, and user-friendly application that effectively serves as the frontend for the analyses performed by the LipidCRED backend.

3.5 Features and Usage

The LipidCRED web application is designed to guide researchers through the analysis workflow in a logical sequence.

Guided Workflow and Initial Setup: Upon accessing the application, users are directed to a "Getting Started" section that provides an overview of LipidCRED and clearly outlines the required input data format (a simple Excel spreadsheet listing lipid names). To facilitate a demonstration run, a 'Download sample data" section provides a correctly formatted example file. This instructional section ensures users can quickly prepare their data and understand the application's requirements.

Input and Analysis Configuration: The main analysis is performed within the "Analyze" section (Figure 3.2). Here, users upload their prepared lipid list file. They are then prompted to specify essential parameters for the analysis, most importantly the target organism (either by selecting from common options like *Homo sapiens, Mus musculus, Rattus norvegicus*, or by providing a specific NCBI Taxon ID). Once the input file and organism are set, the user initiates the core LipidCRED analysis by clicking the "Run" button.

Processing Feedback and Results Access: During the backend computation, the interface provides visual feedback, including status messages and loading indicators, informing the user that the analysis is in progress. Upon completion, the results become accessible within the "Analyze" tab:

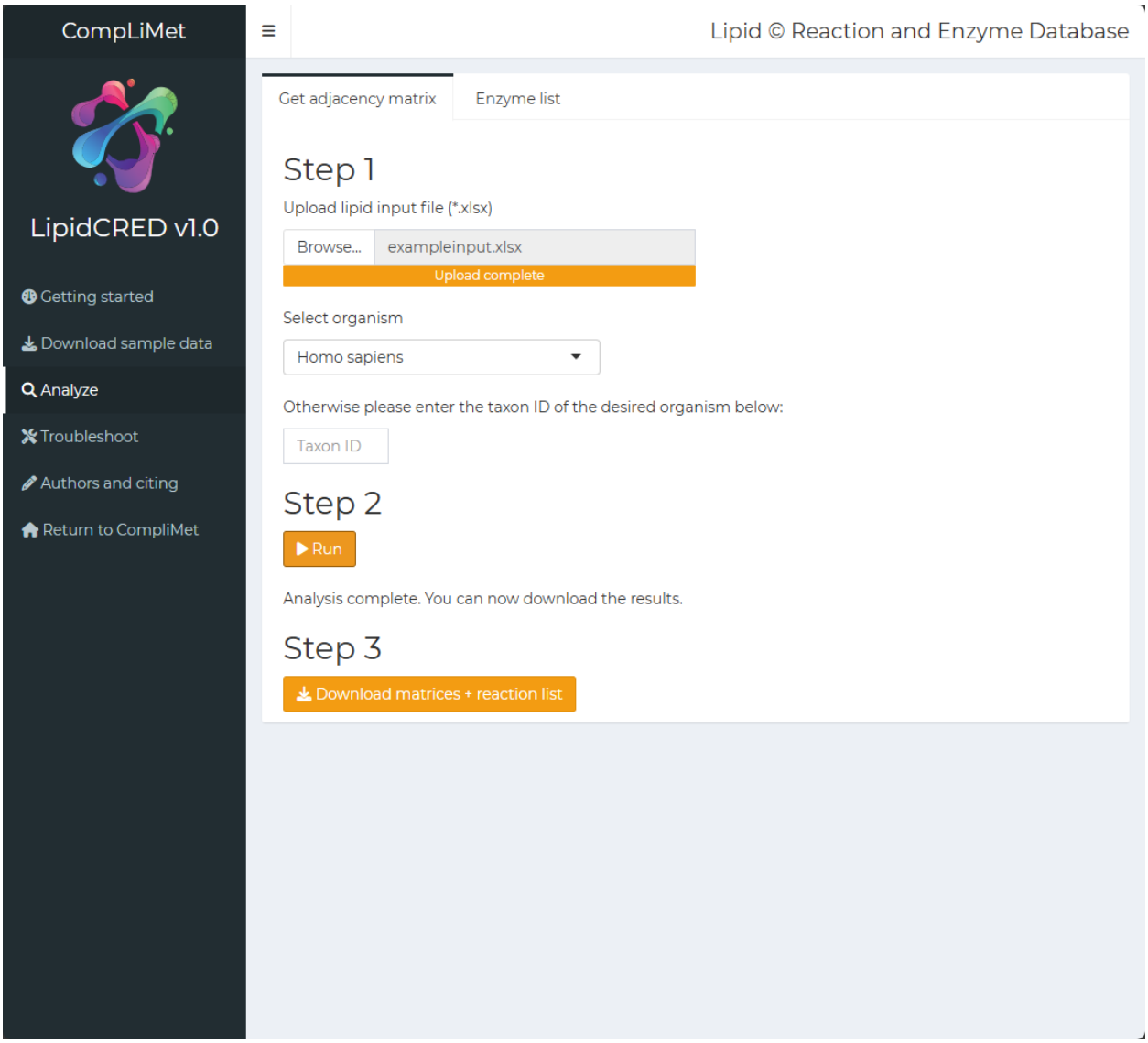


Figure 3.2 Data Input and Analysis in LipidCRED. Screenshot showing the primary analysis interface within the "Analyze" tab *after* the backend processing has successfully completed. Key elements shown include: (1) confirmation that processing is finished (e.g., via status text "Analysis complete. You can now download the results."), and (2) the now-active "Download matrices + reaction list" button, enabling the user to retrieve the output in a zipped package.

- **Interactive Enzyme Summary:** A dedicated sub-tab presents a searchable and sortable table listing all enzymes identified as being associated with the input lipids, with UniProtKB IDs and gene names as identifiers (Figure 3.3). This allows for quick inspection of the key enzymatic players directly within the browser. This table can also be downloaded separately as a comma-separated valued (CSV) file.
- **Comprehensive Download Package:** The primary results are delivered as a consolidated zip archive. Clicking the main download button provides the user with multiple standardized output files suitable for offline analysis, or use in other software (e.g., loading binary adjacency matrix into Cytoscape [129]). This package includes the detailed reaction list (including literature evidence links where available), the enzyme-annotated adjacency matrix, a simplified binary adjacency matrix, the full enzyme list (same as the interactive table), and a translation record of all nomenclature translations. Providing these diverse formats caters to different downstream analysis needs.

Supporting Information: Beyond the core analysis workflow, the application includes dedicated sections for troubleshooting, offering guidance on common issues (like matrix size limits), and for Authors and Citing, providing contact information and instructions for citing LipidCRED in publications.

3.6 Conclusion

Through the integration of R/Shiny and Python via *reticulate*, I successfully developed a functional and user-friendly web application for LipidCRED. The application provides an intuitive interface for researchers to upload their lipid list, specify analysis parameters, execute the backend

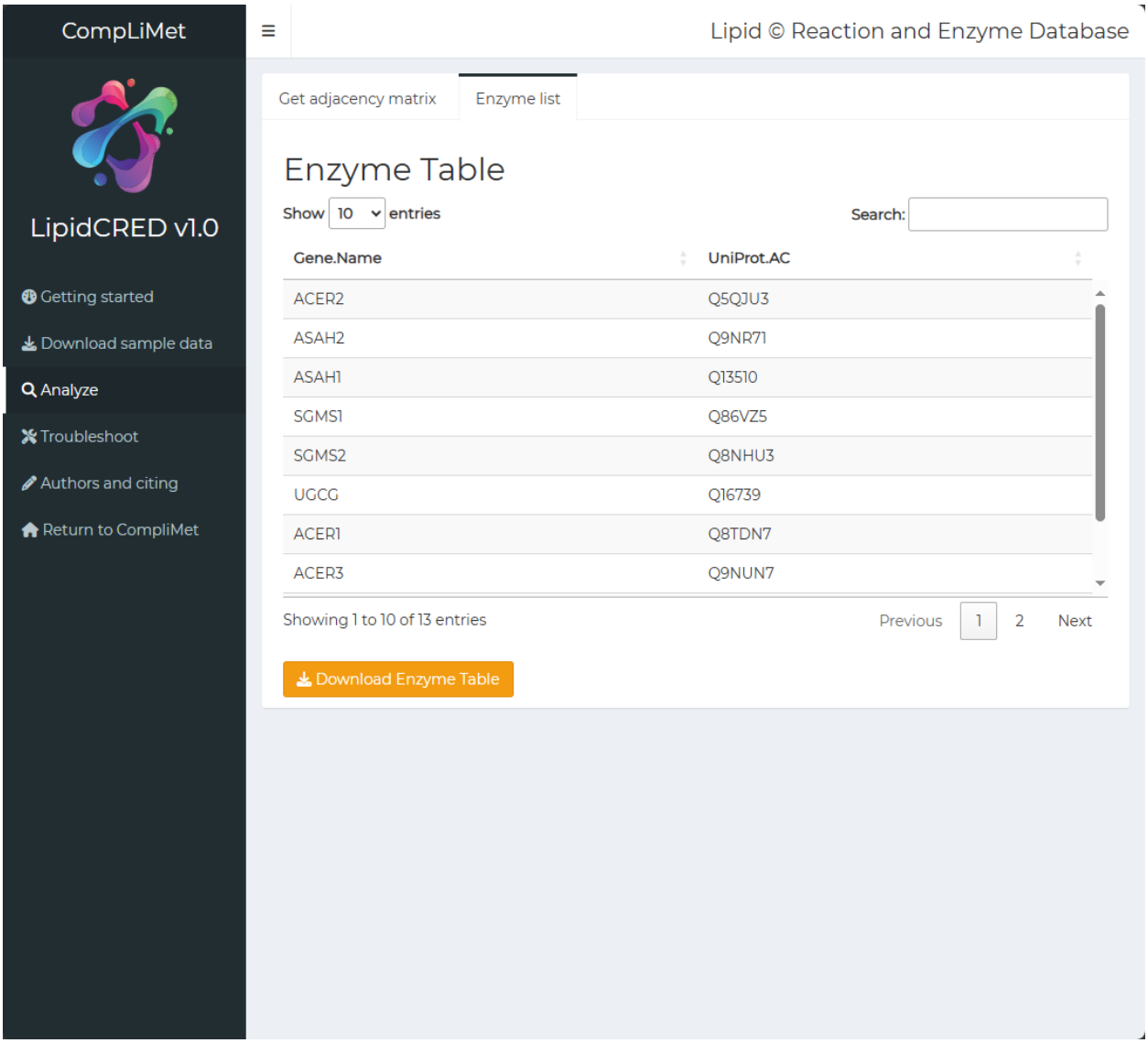


Figure 3.3 Interactive Enzyme Summary Table. Screenshot illustrating the enzyme results table within a sub-tab of the "Analyze" section, displayed after successful backend processing. The figure shows the interactive table, rendered using the *DT* R package, which lists identified enzymes (Gene Name, UniProtKB ID). Features such as sorting, filtering/searching, and pagination provided by *DT* are visible. The "Download Enzyme Table" button, allowing users to export this specific data, is also shown below the table.

processing pipeline described in Chapter 2, and retrieve comprehensive results in multiple, accessible formats. Key features include clear workflow guidance, dynamic user feedback, interactive results display (enzyme table) and organized downloadable outputs. This interface effectively lowers the barriers to entry for utilizing LipidCRED's analytical capabilities, making the complex task of lipid-enzyme association analysis more accessible to the broader lipidomics research community. It serves as a critical component in translating the computational methods developed in this thesis into a practical and impactful research tool.

3.7 Availability

LipidCRED is currently accessible for thesis review via _.

Chapter 4: Benchmarking LipidCRED: A Comparative Analysis

4.1 Objective

My goal in this chapter was to evaluate and benchmark LipidCRED's performance against leading bioinformatics tools for lipid metabolism investigation: BioPAN, LINEX, and LipidOne.

4.2 Introduction

The rapid advancement of lipidomics technologies necessitates the development of robust bioinformatics tools capable of interpreting complex lipid datasets. Identifying the enzymes and reactions associated with measured lipids is crucial for understanding the underlying biological processes. To address this need, I developed LipidCRED, a database and software application that provides comprehensive and accurate lipid-enzyme associations. To evaluate its performance and position it within the existing landscape of bioinformatics resources, I benchmarked it against existing tools.

The current landscape of databases and bioinformatic tools that support lipid metabolism research is growing to keep up with the growth of lipidomics. These tools typically address one or more aspects in the lipidomics workflow as summarized in Chapter 1. In addition to dedicated, lipid-specific databases and tools, there are numerous resources that have overlap in functionality or include lipids in their databases. Table 4.1 provides an overview of some of the major general-purpose and lipid-specific databases and tools. Methodologies presented in Table 4.1 are categorized based on their purpose and functionality, with examples of databases and enrichment or statistical analyses methods. Certain tools such as Lipid Mini-ON and LION/web focus on enrichment analysis, while applications like lipidr and LipidSuite specialize in visualization of

Table 4.1 Overview of databases and bioinformatics tools relevant to lipidomics.

Category	Tool / Database	Batch input	Pathway Information	Manual Curation	Lipid analysis specific?	Enzymes?	Ref
Database	KEGG	Yes	Yes	Yes	No	Yes	[71]
Database	PubChem	Yes	Yes	Yes	No	No	[96]
Database	UniProtKB	Yes	No	Yes	No	Yes	[102]
Database	HMDB	No	Yes	Yes	No	Yes	[97]
Database	SwissLipids	No	No	Yes	Yes	Yes	[118]
Database	LIPID MAPS	Yes	No	Yes	Yes	Yes	[94]
Database	WikiPathways	No	Yes	Yes	No	Yes	[72]
Ontology, Enrichment	Lipid Mini-On	Yes	No	Yes	Yes	No	[107]
Ontology, Enrichment	LION/web	Yes	No	Yes	Yes	No	[108]
Ontology, Enrichment	LipiDisease	Yes	No	Yes	Yes	No	[109]
Statistical Analysis, Visualization	MetaboAnalyst	Yes	Yes	Yes	No	No	[86]
Statistical Analysis, Visualization	lipidr	Yes	No	Yes	Yes	No	[130]
Statistical Analysis, Visualization	LipidSuite	Yes	No	No	Yes	No	[89]
Ontology, Enrichment, Pathway Analysis	BioPAN	Yes	Yes	Yes	Yes	Yes	[113]
Statistical Analysis, Visualization, Pathway Analysis	Lipostar 2.0	Yes	Yes	Yes	Yes	No	[87]
Statistical Analysis, Visualization, Pathway Analysis	LipidOne 2.2	Yes	Yes	Yes	Yes	Yes	[117]
Statistical Analysis, Enrichment, Pathway Analysis	LINEX^2	Yes	Yes	Yes	Yes	Yes	[18]

lipid reactions. LIPID MAPS, Rhea and UniProtKB provide information about the lipid molecules and their reactions by organizing information about lipid structures, molecular characteristics, enzymes, and reactions. Direct comparison of these tools becomes challenging as they differ in their primary purposes, acceptable input types as well as provided output information. For example, WikiPathways is a general repository of biological pathways including lipid metabolism with number of reactions organized within the database, providing only single input option as a pathway search and not batch lipid molecule input. SwissLipids acts as a knowledgebase of lipids, their structures and biology, particularly enzymatic reactions allowing users to batch input lipids and map them to other molecular identifiers or obtain the corresponding reactions and enzymes. For my benchmarking tests, I focused on tools that share LipidCRED's core functionality of processing bulk, batch lipid data and providing insights on the reactions and enzymes involved with provided list of lipid molecules.

In this chapter, I detail the benchmarking process undertaken to compare LipidCRED against three leading tools with similar core functionalities: LipidOne (v2.2) [117], LINEX (v2) [18], and BioPAN [113]. I selected these specific tools because, like LipidCRED, they accept lists of lipids as input and return information about associated enzymes and/or reactions, making them suitable for direct comparison.

My primary objectives for this benchmarking study were to compare: 1) the lipid coverage achieved by each tool, which was defined as the percentage of input lipids associated with at least one enzyme; 2) the total number of enzymes reported by each tool for a given dataset; and to introduce 3) an enzyme validation step, assessing the biological relevance and accuracy of the reported enzymes.

4.3 Methods

4.3.1 Benchmarking Datasets

Several different publicly available datasets were utilized for the benchmarking test including data originally provided by Alabed et al. [131], Hahn et al. [132], and Wang et al. [133]. These three datasets correspond to the demonstration datasets presented by LipidOne, BioPAN, and LINEX, respectively. By choosing these datasets, I am ensuring favourable conditions for each of the three tools, at least when running them with their corresponding demonstration dataset. Additionally, I obtained lipidomics data from Hornburg et al. [134] to serve the role of an independent test case with recently published data. I generated a summary of the lipid composition of each dataset (Table 4.2), detailing the number of lipids per category and class.

4.3.2 Data Preprocessing

Each dataset had differences in lipid nomenclature, and at the same time, input for each tool required a specific lipid nomenclature and file format. Thus, to ensure a fair comparison, I preprocessed each dataset to fit the input requirements of each of the four tools. To run datasets on LipidOne, the specifics on data preparation as detailed on the tool's website were followed, which necessitated the use of the molecular species level of the Lipidomics Standards Initiative (LSI) guidelines [100]. Similarly, for BioPAN and LINEX, the specific instructions on their respective websites were followed, with appropriate lipid nomenclature, to prepare the datasets.

4.3.3 Tool Execution Workflow

I ran each preprocessed dataset through LipidCRED and the three benchmarking tools (Figure 4.1). Where possible, consistent parameters were used, specifically selecting *Homo sapiens* as the target organism for all analyses.

Table 4.2 Lipid categories and classes within the benchmarking datasets.

Lipid Category / Class	Hornburg et al.	Hahn et al.	Wang et al.	Alabed et al.
Fatty Acyls (FA)				
CAR	0	13	0	13
CoA	0	5	0	0
FA	26	25	0	0
Total FA	26	43	0	13
Glycerolipids (GL)				
DG	47	22	13	13
DG(O-)	0	21	0	0
MG	0	7	0	0
TG	97	80	22	66
TG(O-)	0	25	0	0
Total GL	144	155	35	79
Glycerophospholipids (GP)				
CL	0	24	0	0
LPA	0	4	6	0
LPC	16	18	24	6
LPC(O-)	0	5	3	0
LPE	9	14	24	0
LPE(O-)	0	6	0	0
LPG	0	3	0	0
LPI	0	8	5	0
LPS	0	6	0	0
PA	0	20	7	4
PC	79	40	23	24
PC(O-)	0	20	26	0
PE	38	37	27	48
PE(O-)	17	29	24	0
PE(P-)	31	0	0	0
PG	0	20	13	0
PI	26	28	19	7
PS	0	29	27	12
Total GP	216	311	228	101
Sphingolipids (SP)				
Cer	14	21	36	4
Hex2Cer	7	0	0	0
HexCer	7	0	2	0
SM	12	21	19	7
SPBP	0	1	0	0
Total SP	40	43	57	11
Sterol Lipids (ST)				
CE	26	7	14	4
ST	0	0	0	2
Total ST	26	7	14	6
Total all lipids	452	559	334	210

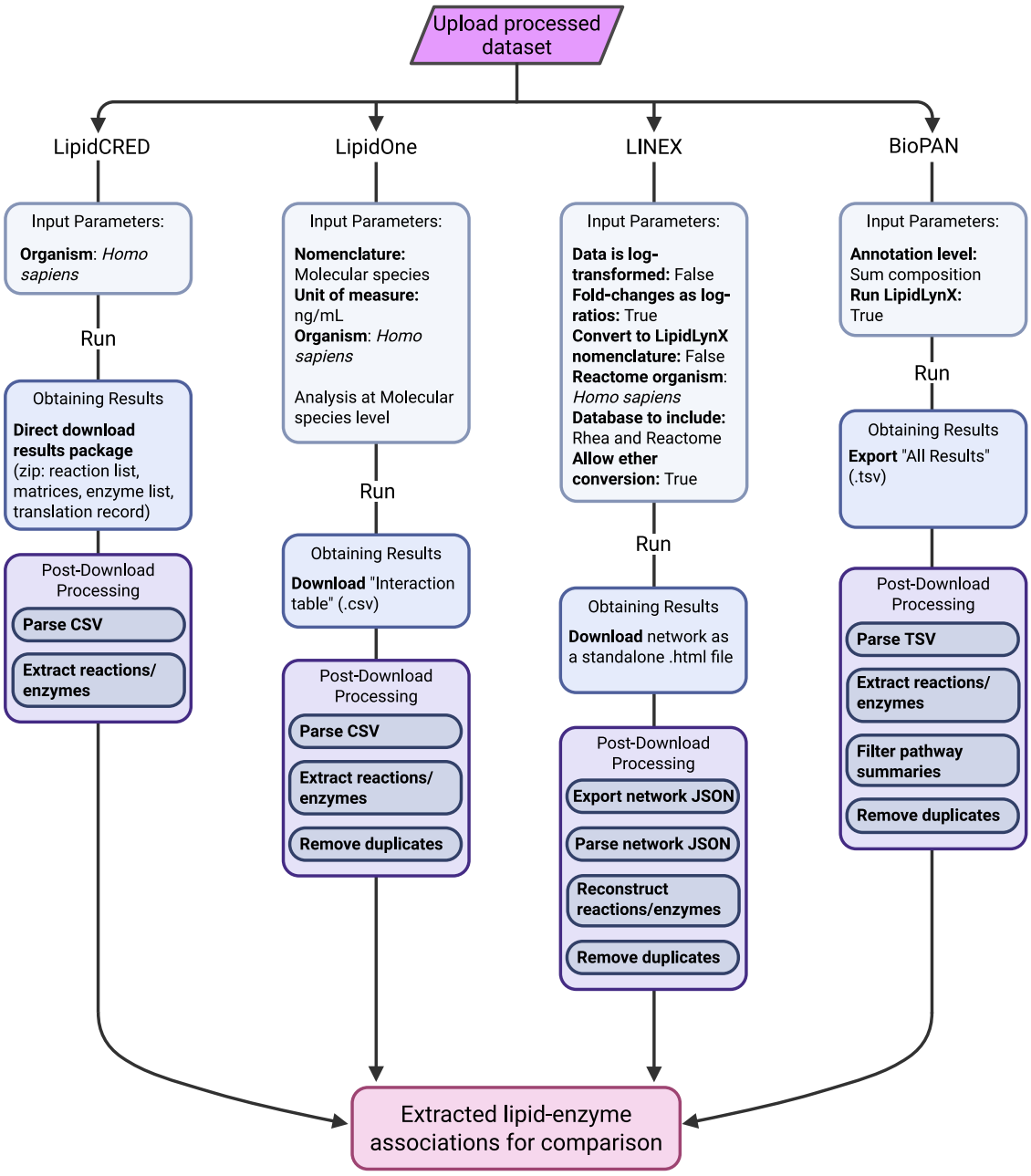


Figure 4.1 Comparative workflow for analysis execution and data extraction. Flowchart illustrating the steps involved in processing an input dataset (preprocessed for tool-specific nomenclature and format requirements) through each benchmarked tool: LipidCRED, LipidOne, LINEX, and BioPAN. The diagram details the input parameters configured for each tool to ensure consistency (e.g., specifying *Homo sapiens* as the target organism where applicable) and outlines the distinct procedures required to obtain the primary results output. Additionally, it highlights the varying levels of post-download processing necessary to extract the final lipid-enzyme association data used for the comparative metrics in this study.

LipidCRED: I processed the datasets using the standard LipidCRED workflow selecting *Homo sapiens*. I then used the reaction list output file containing lipid-enzyme associations for my analysis.

LipidOne (v2.2): Following the tool's online documentation, I prepared the datasets using the molecular species level nomenclature and ran the analyses specifying *Homo sapiens*. I performed both 'Lipid Class' and 'Lipid Molecular Species' level analyses sequentially, as required by the tool. I downloaded the resulting 'Interaction Table' for my evaluation.

LINEX (v2): I uploaded the datasets and ran the analysis using default settings, which specify *Homo sapiens*. LINEX does not provide tabular output listing reactions and enzymes, instead it provides an offline network visualization (as .html file) which then allows downloading the network data as a JSON file. I parsed this JSON output to extract the lipid-enzyme associations needed for my comparison.

BioPAN: I uploaded the datasets, selected "sum composition" as the annotation level, and enabled the "Run LipidLynxX" option. The platform did not provide the option to select the target organism. I exported the results table and, to avoid redundant counting from pathway summaries, I only considered individual reaction steps when associating enzymes with lipids in my analysis pipeline.

4.3.4 Output Processing and Metric Calculation

To handle the different output formats, a Python script was developed to parse the files from each tool (LipidCRED's reaction list, LipidOne's Interaction Table, LINEX's JSON data, and BioPAN's results table). Using this script, the key metrics, lipid coverage and total enzyme count (reported) were calculated. For lipid coverage, this script identified the number of unique input

lipids associated with at least one enzyme for each tool and dataset. For the total enzyme count, this script extracted all unique gene names reported by each tool for each dataset. Lipid coverage was then calculated as:

$$\frac{Number\ of\ unique\ input\ lipids\ with\ >0\ associated\ enzyme(s)}{Total\ number\ of\ unique\ input\ lipids} \times 100\%$$

4.3.5 Enzyme Validation Protocol

A crucial step in my benchmarking strategy was to validate the reported enzymes. I implemented a protocol where all unique gene names reported by each tool were cross-referenced against the UniProtKB/Swiss-Prot database [102] (release 2025_01). An enzyme was defined as "Validated" only if it met all the following criteria:

1. Correct organism: the gene entry belonged to *Homo sapiens* (TaxID: 9606).
2. Catalytic activity: the gene entry possessed at least one annotated Enzyme Commission (EC) number, indicating known catalytic function.
3. Valid entry: the gene entry corresponded to a valid, reviewed entry in UniProtKB/Swiss-Prot.

Based on this, the total enzyme count (validated) was calculated, which represented the count of unique enzymes from the "Reported" set that passed my validation protocol. Additionally, coverage was recalculated using only associations involving validated enzymes. An input lipid was considered covered in the validated analysis only if it was associated with at least one validated enzyme, according to the criteria listed above.

4.4 Results

4.4.1 Enzyme validation reveals differences in lipid coverage across tools

As shown in Figure 4.2, LipidCRED achieved validated lipid coverage ranging from 39.5% (Alabed et al.) to 77.2% (Wang et al.). LipidCRED yielded the highest validated lipid coverage percentage among the four tools in two of the four datasets (Hornburg et al. and Wang et al.), with LINEX was predominant in the two other datasets (Hahn et al. and Alabed et al.). The application of the enzyme validation protocol resulted in varied reduction in coverage across the tools and datasets. For instance, BioPAN's reported coverage in the Hahn et al. dataset decreased from 41.5% to 28.8% after validation. This reduction was due to the exclusion of reported genes that did not meet all validation criteria; examples include CERT1 (a lipid transfer protein, not classified by an EC number) and SCD3 (a gene studied in mouse, not human, though BioPAN lacked an organism selection option).

LipidOne's coverage decreased minimally, within 0.4-2.0% across the four datasets, suggesting most of its reported enzymes met the criteria, although examples like NSMAF (related to TNF receptor signaling, lacking EC number) were excluded upon validation. Similarly, LINEX's coverage post validation decreased by 0.7-2.4% across three datasets (Hornburg et al., Wang et al., and Alabed et al.) due to non-human and non-catalytic genes (e.g., GPC1, MBO). For LipidCRED, the reported and validated genes were identical, and thus, its coverage showed no difference upon applying the validation protocol.

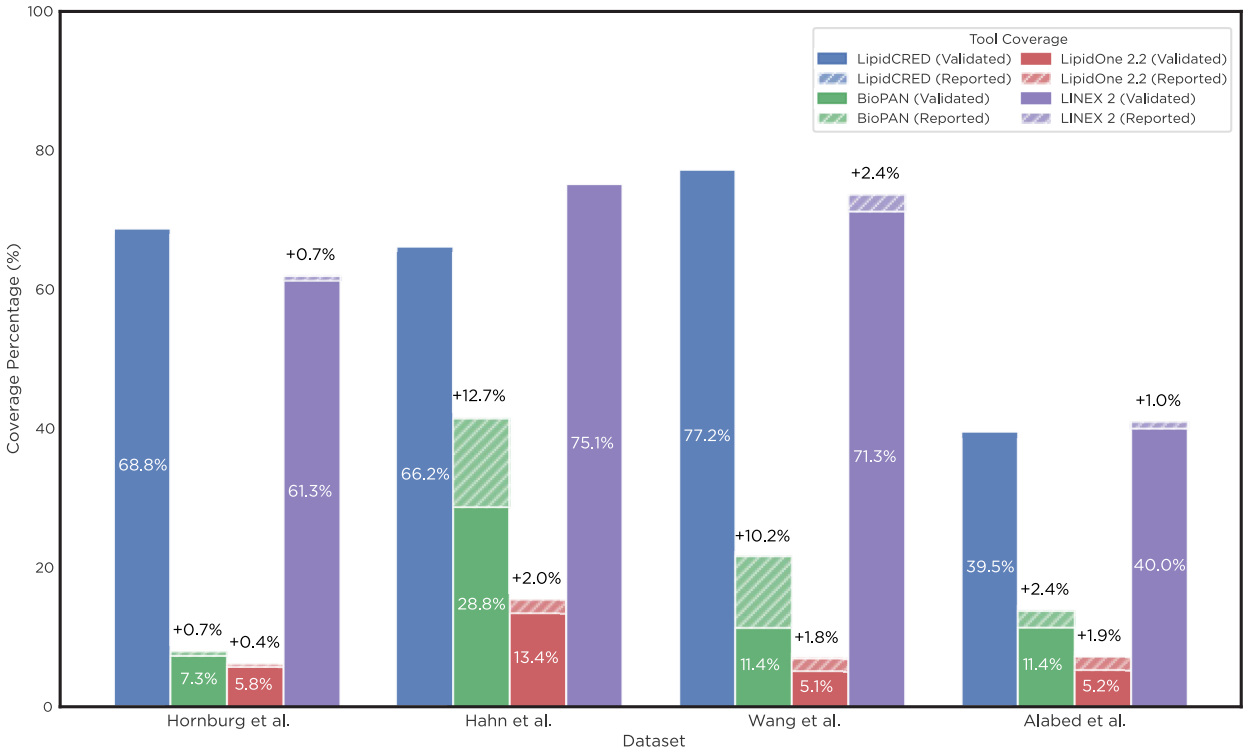


Figure 4.2 Comparison of reported and validated lipid coverage percentages. Lipid coverage (%) calculated for LipidCRED, BioPAN, LipidOne, and LINEX across four datasets. Analyses were performed targeting *Homo sapiens* enzymes. Solid bars represent “Validated” coverage, calculated considering only associations with *Homo sapiens* enzymes, possessing an EC number, and present as a reviewed entry in UniProtKB/Swiss-Prot (release 2025_01). Hatched overlays represent the additional coverage percentage points derived from associations that did not meet all validation criteria (“Reported” coverage minus “Validated” coverage). Total bar height (solid + hatch) indicated the “Reported” coverage before validation. Input lipid list for each dataset were preprocessed for compatibility with each tool.

4.4.2 Validation refines enzyme sets, revealing distinct validated enzyme contributions from each tool

Comparing Figure 4.3 (Reported) and Figure 4.4 (Validated) reveals the quantitative impact of the enzyme validation process on the enzyme sets identified by each tool. The total number of unique enzymes provided by each tool either decreased (BioPAN, LINEX, and LipidOne) or stayed the same (LipidCRED) after validation. For example, in the Hornburg et al. dataset, the total number of enzymes reported uniquely by LINEX decreased from 38 (Figure 4.3) to 22 (Figure 4.4), a reduction of 42.1%. This reduction reflects the removal of a number of LINEX-specific genes flagged during validation, such as 'DGAT2L7P' (a pseudogene), non-human genes like 'LRO1', and invalid entries like 'NAN' which could not be linked to any catalytic enzymes (human or otherwise). For the same dataset, the unique counts for BioPAN and LipidOne decreased from 3 to 2 (33.3% reduction). The validation step only affected the enzymes uniquely reported by each tool. The overlap patterns between tools (i.e., enzymes reported by multiple tools) remained consistent before and after validation, suggesting that enzymes identified concordantly across the tools are more likely to be well-characterized and meet the validation criteria.

4.4.3 Validated enzyme associations show class-specific distribution and vary across tools

The heatmap in Figure 4.5 details the distribution of enzyme associations across lipid classes. When comparing the validated number of identified enzymes per lipid class, LipidCRED consistently leads the competition by dominating the number of lipid class for which it has the highest number of enzymes. For example, in the Wang et al. dataset, LipidCRED returns the highest number of enzymes for 13 of the 18 lipid classes (PC, DG, LPC, PA, PE, LPE, Cer, LPA, SM, PC(O-), LPC(O-), HexCer, LPI), followed by LINEX, which leads the race for 2 lipid classes

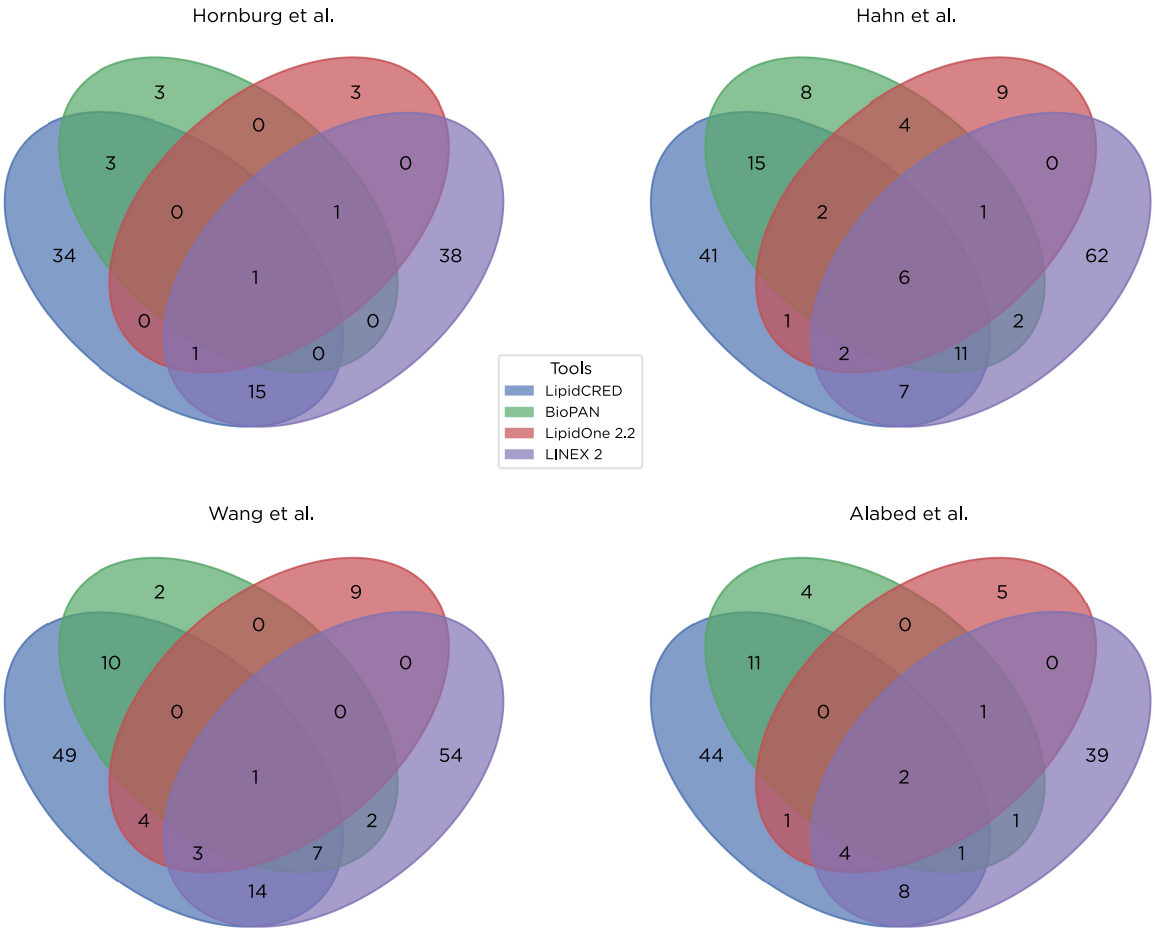


Figure 4.3 Overlap of reported gene names. Four-set Venn diagrams displaying the number of unique and shared gene names reported by LipidCRED, BioPAN, LipidOne, and LINEX for each of the four benchmarking datasets. These counts represent the enzymes reported before the validation protocol was applied. Numbers in overlapping regions indicate enzymes reported by multiple tools; numbers in non-overlapping regions indicate enzymes unique to a single tool for that dataset. Analyses were performed targeting *Homo sapiens*.

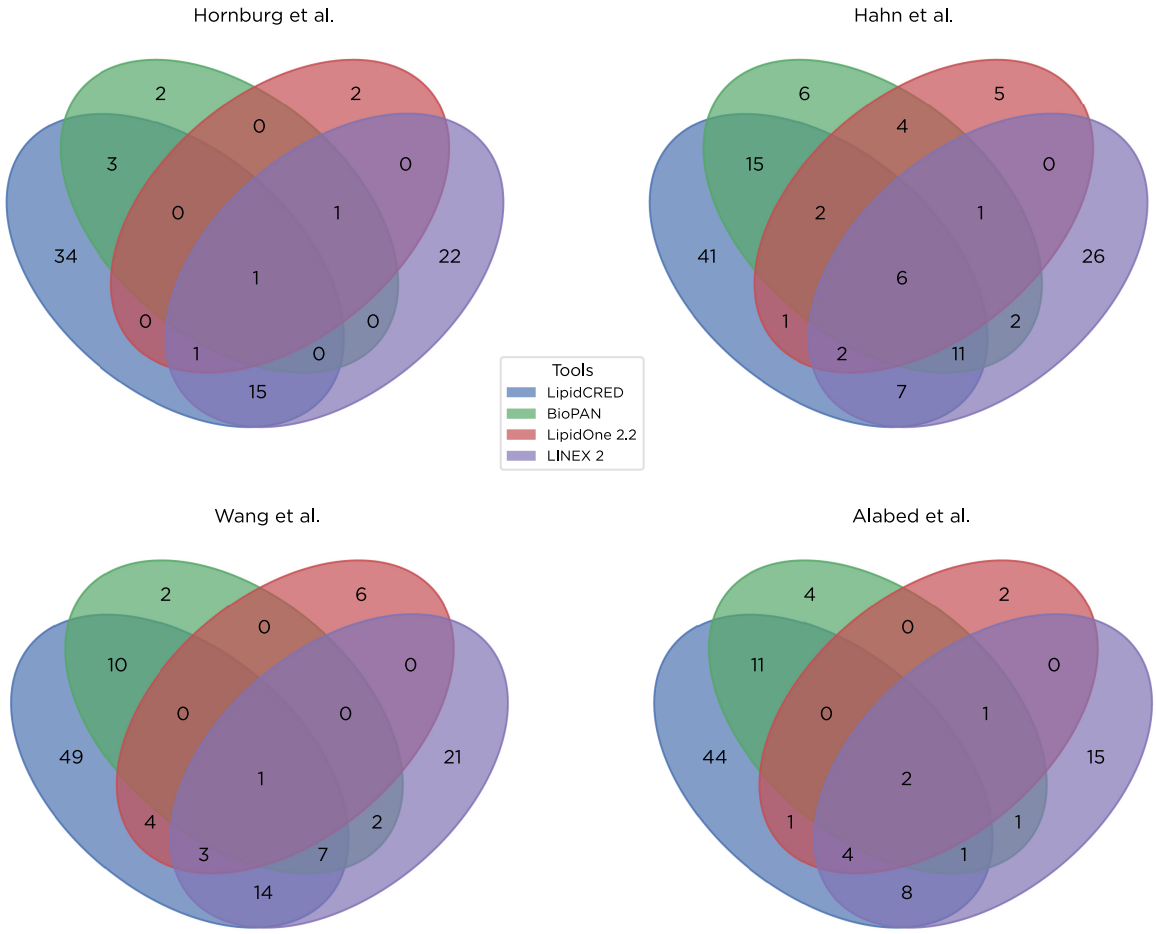


Figure 4.4 Overlap of validated gene names. Four-set Venn diagrams displaying the number of unique and shared gene names confirmed as Validated (*Homo sapiens*, possessing an EC number, reviewed entry in UniProtKB/Swiss-Prot (release 2025_01)) for each tool across the four benchmarking datasets. Numbers indicate the count of validated enzymes in each segment of the diagrams, illustrating overlap and unique contributions after applying the validation criteria described in Methods (Section 5.3.5).

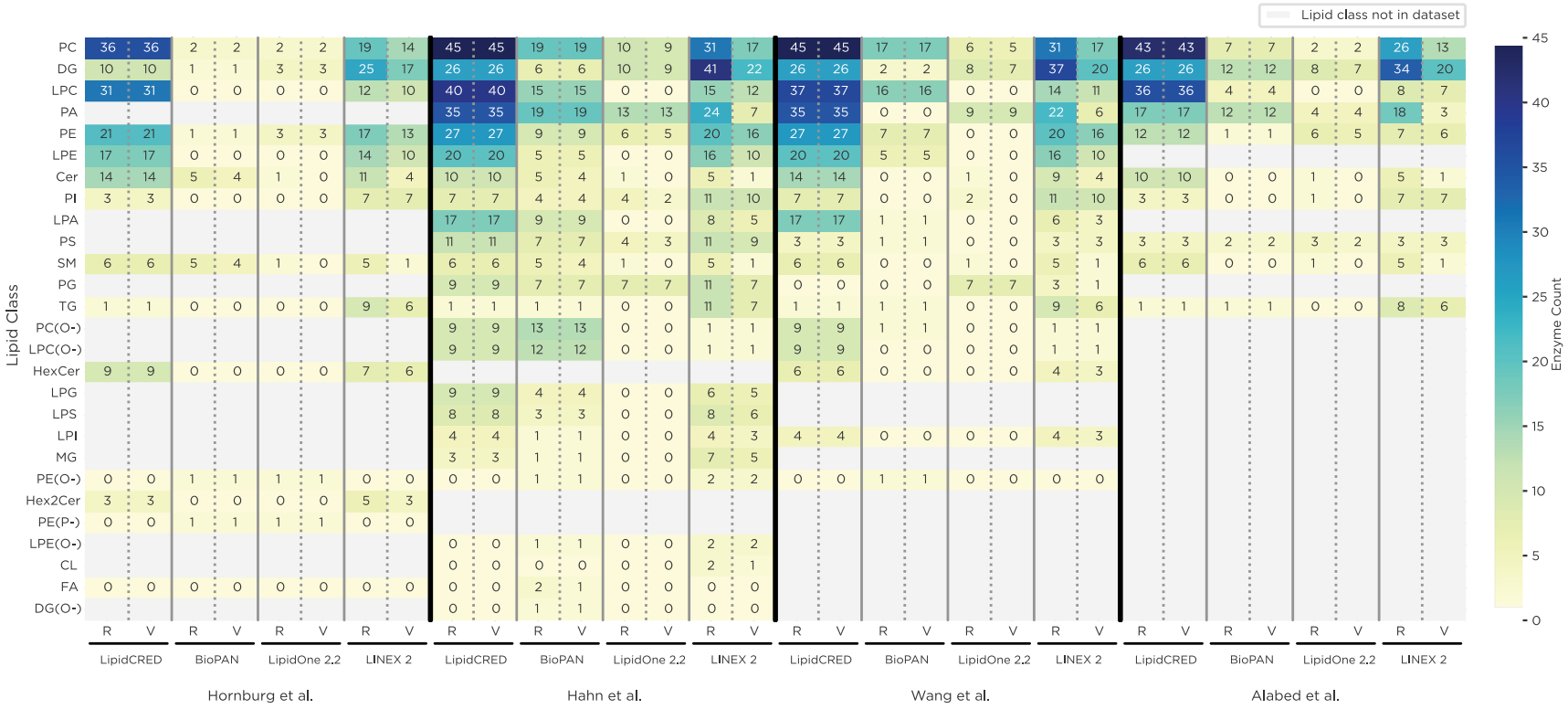

Lipid class not in dataset
Lipid Class
PC 36 36 2 2 2 2 19 14 45 45 19 19 10 9 31 17 45 45 17 17 6 5 31 17 43 43 7 7 2 2 26 13
DG 10 10 1 1 3 3 25 17 26 26 6 6 10 9 41 22 26 26 2 2 8 7 37 20 26 26 12 12 8 7 34 20
LPC 31 31 0 0 0 0 12 10 40 40 15 15 0 0 15 12 37 37 16 16 0 0 14 11 36 36 4 4 0 0 8 7
PA 35 35 19 19 13 13 24 7 35 35 0 0 9 9 22 6 17 17 12 12 4 4 18 3
PE 21 21 1 1 3 3 17 13 27 27 9 9 6 5 20 16 27 27 7 7 0 0 20 16 12 12 1 1 6 5 7 6
LPE 17 17 0 0 0 0 14 10 20 20 5 5 0 0 16 10 20 20 5 5 0 0 16 10
Cer 14 14 5 4 1 0 11 4 10 10 5 4 1 0 5 1 14 14 0 0 1 0 9 4 10 10 0 0 1 0 5 1
PI 3 3 0 0 0 0 7 7 7 7 4 4 4 2 11 10 7 7 0 0 2 0 11 10 3 3 0 0 1 0 7 7
LPA 17 17 9 9 0 0 8 5 17 17 1 1 0 0 6 3
PS 11 11 7 7 4 3 11 9 3 3 1 1 0 0 3 3 3 3 2 2 3 2 3 3
SM 6 6 5 4 1 0 5 1 6 6 5 4 1 0 5 1 6 6 0 0 1 0 5 1 6 6 0 0 1 0 5 1
PG 9 9 7 7 7 7 11 7 0 0 0 0 7 7 3 1
TG 1 1 0 0 0 0 9 6 1 1 1 1 0 0 11 7 1 1 1 1 0 0 9 6 1 1 1 1 0 0 8 6
PC(O-) 9 9 13 13 0 0 1 1 9 9 1 1 0 0 1 1
LPC(O-) 9 9 12 12 0 0 1 1 9 9 0 0 0 0 1 1
HexCer 9 9 0 0 0 0 7 6 6 6 0 0 0 0 4 3
LPG 9 9 4 4 0 0 6 5
LPS 8 8 3 3 0 0 8 6
LPI 4 4 1 1 0 0 4 3 4 4 0 0 0 0 4 3
MG 3 3 1 1 0 0 7 5
PE(O-) 0 0 1 1 1 1 0 0 0 0 1 1 0 0 2 2 0 0 1 1 0 0 0 0
Hex2Cer 3 3 0 0 0 0 5 3
PE(P-) 0 0 1 1 1 1 0 0
LPE(O-) 0 0 1 1 0 0 2 2
CL 0 0 0 0 0 0 2 1
FA 0 0 0 0 0 0 0 0 0 0 2 1 0 0 0 0 0 0 0 0
DG(O-) 0 0 1 1 0 0 0 0
R V R V R V R V R V R V R V R V R V R V R V R V R V R V R V R V R V R V R V R V
LipidCRED BioPAN LipidOne 2.2 LINEX 2 LipidCRED BioPAN LipidOne 2.2 LINEX 2 LipidCRED BioPAN LipidOne 2.2 LINEX 2 LipidCRED BioPAN LipidOne 2.2 LINEX 2
Hornburg et al.
Hahn et al.
Wang et al.
Alabed et al.
Enzyme Count
45
40
35
30
25
20
15
10
5
0

Figure 4.5 Reported (R) versus Validated (V) enzyme counts per lipid class. Heatmap illustrating the number of unique enzymes associated with individual lipid classes by LipidCRED, BioPAN, LipidOne, and LINEX across the four datasets. Each tool is represented by two adjacent columns: 'R' for Reported enzyme counts (before validation) and 'V' for Validated enzyme counts (meeting criteria outlined in Section 5.3.5: Homo sapiens, EC number present, reviewed UniProtKB/Swiss-Prot entry (release 2025_01)). Color intensity corresponds to the enzyme count. Lipid classes are ordered based on the total enzyme counts across all tools and datasets. A cell with 0 indicates that no enzymes were reported. Grey cells indicate that the corresponding lipid class was not present in the input lipid list for that specific dataset.

(PI and TG), and finally BioPAN and LipidOne taking the lead for PE(O-) and PG, respectively. The remaining lipid class PS is tied at three enzymes for both LipidCRED and LINEX.

The validation step appears to have tipped the scales for LipidCRED for a few lipid classes across the four datasets. For instance, the reported (R) number of enzymes associated with the DG class is dominated by LINEX, however, after validation (V), LINEX falls back behind LipidCRED for three of the four datasets (Hahn et al., Wang et al. and Alabed et al.). It is worth mentioning that, despite outperforming the three other tools in the majority of lipid classes, LipidCRED returns no enzymes for a few lipid classes present in certain datasets. For example, in the Hahn et al. dataset, LipidCRED reports zero enzymes for PE(O-), LPE(O-), FA, CL, and DG(O-), which otherwise have 1 or more validated enzymes reported by BioPAN and/or LINEX. This absence of reported enzymes by LipidCRED for these specific classes stems directly from the current scope of its underlying reaction knowledge base; reactions involving free fatty acids (FA) or cardiolipins (CL), along with certain transformations specific to ether lipids (PE(O-), LPE(O-), DG(O-)), are not included in the current version of the database.

4.5 Discussion

This benchmarking study aimed to evaluate the performance of LipidCRED relative to three established tools (LipidOne (v2.2), LINEX (v2), BioPAN) using metrics focused on lipid coverage and enzyme identification, with a critical focus on enzyme validation. A key finding is LipidCRED's strong performance in terms of validated lipid coverage, achieving up to 77.2% coverage and slightly superior to LINEX (which reached up to 75.1% validated coverage). Strikingly, LipidCRED's initially reported enzyme set is identical to its validated enzyme set, underlining the effectiveness of the data sourcing and processing pipeline implemented. This contrasts sharply with the other tools, particularly LINEX, where validation of its reported enzyme

sets significantly reduced the number of unique enzymes reported given false positives in reaction reports. BioPAN and LipidOne, which generally reported fewer enzymes overall, showed much smaller reductions upon validation (<2.4%), suggesting higher level of curation of the databases, but they still contained a small proportion of non-validated entries. The validation criteria employed here aimed to identify potentially erroneous or irrelevant enzyme entries reported by the three tools, either due to inclusion of non-enzymes, errors in naming, duplications or non-Human enzymes presented in Human reaction outputs. For example, LINEX reported REQ_38260 as an enzyme catalyzing glycosphingolipid hydrolysis (Hex2Cer → Cer) in the Hornburg et al. dataset, which upon cross-referencing with UniProtKB/Swiss-Prot was revealed to be a bacterial enzyme belonging to *Rhodococcus hoagii* (strain 103S, TaxID 685727). Some gene names reported did belong to *Homo sapiens* but upon validation appeared to have no catalytic activity. For instance, NSMAF was reported by LipidOne as the enzyme involved in converting SM to Cer, but validation confirmed it is a protein factor associated with sphingomyelinase activation but lacks catalytic activity [135]. Similarly, BioPAN erroneously reported CERT1, which is a shuttling enzyme with no catalytic activity, as being associated with the synthesis of SM from Cer [136].

In LipidCRED's workflow, the reported enzyme set is strictly constrained to the specified organism by leveraging UniProtKB/Swiss-Prot reviewed entries and HCOP orthology data cross-referenced against Swiss-Prot (Chapter 2, Section 2.4.10), while also focusing on catalytic enzymes by primarily sourcing reactions linked to enzymes with known activity in SwissLipids/Rhea. LipidCRED's database is carefully constructed based on these curated sources. Also, a multitude of design decisions that went into LipidCRED were aimed at ensuring high level of validation of information. The normalized relational database (Chapter 2) was intentionally designed to allow for discriminant queries focused on specific reactions, molecular or class lipids,

enzymes and organisms. This design choice, coupled with the in-house *LipidMatcher* algorithm (Chapter 2, Section 2.4.5), ensures consistency and accuracy when reporting lipid reactions and the associated enzymes. On the other hand, the other tools, while allowing *Homo sapiens* selection, seem to draw from broader or differently curated knowledge bases. Specifically, LINEX utilizes Rhea and Reactome knowledgebases as sources for lipid reactions and their associated enzymes. Although Reactome maps reactions to organisms, Rhea itself neither provides this mapping nor denotes organism-specific enzymes, which explains the presence of non-human enzymes in the set reported by LINEX during benchmarking. This inherent rigor in LipidCRED potentially offers users a more reliable starting point, minimizing the need for extensive post-hoc filtering of enzyme lists.

Analysis of enzymes uniquely provided by each tool as well as shared enzymes (Section 4.4.2) revealed that the validation step primarily impacted the enzymes uniquely reported by BioPAN, LINEX, and LipidOne, while leaving LipidCRED's unique enzyme count unchanged. This suggests that enzymes reported by multiple tools are generally more likely to meet the validation criteria. However, the presence of unique validated enzymes reported by the other tools but not by LipidCRED, provides an avenue for future improvement of LipidCRED by incorporating the relevant reactions and these enzymes into its knowledge base. Based on my analysis of the unique validated enzymes reported by other tools, their absence in LipidCRED often reflects specific differences in nomenclature, database scope, or currently implemented matching strategies:

- **Nomenclature/synonym issues:** Some apparent absences are due to synonym differences. For example, PEDS1 (reported uniquely by BioPAN in the Hornburg et al. dataset) catalyzes the PE(O-) → PE(P-) reaction, which is present in

LipidCRED but associated with TMEM189, a synonym of PEDS1. Similarly, AAPT1 (reported by LINEX) catalyzes DG → PC, and is an alternative name for CHPT1 (the primary gene name in UniProtKB), which is present in LipidCRED.

- **Missing reactions/enzymes in source data:** Other enzymes are missing because the specific reaction they catalyze is not currently captured in LipidCRED's source data integration pipeline (primarily derived from SwissLipids/Rhea associations). For instance, SELENOI (reported by LipidOne), associated with the DG(O-) → PE(O-) reaction, is absent from LipidCRED's database. Likewise, SMPD4 (reported by BioPAN for SM → Cer) is absent, although LipidCRED includes other sphingomyelinases (SMPD1, SMPD2, SMPD3) for this reaction.
- **Reaction/matching strategy scope:** Some absences occur because the required reaction type or matching strategy is not yet implemented. For example, the reaction synthesizing cardiolipin (CL) from PG (PG → CL) associated with CRLS1 (reported by LINEX and BioPAN), is currently outside the scope of LipidCRED's *LipidMatcher* strategies.

Addressing these specific cases through targeted expansion of the reaction and enzyme database, and implementing corresponding *LipidMatcher* strategies (Chapter 2, Section 2.4.5) are key priorities for enhancing LipidCRED's comprehensiveness.

When examining enzyme associations at the lipid class level (Section 4.4.3), LipidCRED associated the highest number of validated enzymes with the majority of lipid classes across the datasets. This finding confirms LipidCRED's utility for analyzing typical lipidomics datasets encompassing glycerophospholipid and sphingolipid species. However, this detailed analysis also pinpointed specific areas where LipidCRED's current version shows gaps. The absence of

validated enzyme associations for certain classes like ether-linked lipids (PE(O-)) or cardiolipins (CL) in some datasets, where other tools did find at least one validated hit, clearly identifies targets for future development and expansion of LipidCRED's knowledge base. These omissions reflect the current scope of the reaction and enzyme databases integrated within LipidCRED, which currently lacks comprehensive coverage for reactions specifically involving ether-linked lipids, cardiolipins, and free fatty acids (FA). The database storing LipidCRED's reactions and enzymes is designed with expansion as a core design choice (Chapter 2, Section 2.4.9 & 2.4.13), thus when missing reactions or enzymes are identified, they can be readily incorporated into LipidCRED.

4.6 Conclusion

The insights gained from this benchmarking informed LipidCRED's future development trajectory. Mainly, it will be a priority to address the identified gaps by expanding the reaction and enzyme database within LipidCRED. To do this, I plan to conduct future benchmarking using more datasets to help identify any other gaps, while also collecting feedback from users to evaluate user experience aspects compared to other tools (input formats, output clarity, visualization options). Furthermore, the benchmarking process itself highlighted significant differences in usability and data accessibility between tools. As illustrated in the workflow comparison (Figure 4.1), obtaining comparable results from tools like LINEX required multiple non-trivial steps by the user as well as network file parsing, whereas LipidCRED provides a direct download of clearly structured output files.

Chapter 5: General Discussion

In this thesis, I present the development, implementation, and evaluation of LipidCRED, a novel bioinformatics platform designed to navigate the complexities of lipid metabolism. Chapter 2 detailed the construction of its core database and computational engine; Chapter 3 described the creation of its user-friendly web interface; and Chapter 4 presented a comprehensive benchmarking study comparing LipidCRED against contemporary tools (LipidOne, LINEX, BioPAN). This benchmarking revealed key differences in performance and usability among the evaluated tools. Notably, LipidCRED demonstrated competitive validated lipid coverage while ensuring high accuracy of the reported enzymes, whereas other tools required varying degrees of *post hoc* filtering or complex data extraction workflows. Having established LipidCRED's technical foundation and comparative performance, in this chapter, I place the motivations behind LipidCRED's creation in context. I contextualize its unique contributions to the field of lipid bioinformatics when compared with similar available applications; I also reflect on the challenges encountered and its inherent limitations and discuss promising avenues for future development and application.

5.1 Addressing the Challenge: The Genesis of LipidCRED

The field of lipidomics has undergone remarkable growth, yet extracting meaningful biological insights from complex lipid datasets remains a significant hurdle. My exploration of the existing landscape identified a need for a tool that could reliably associate lipids with enzymatic reactions, handle diverse nomenclatures, and provide accessible, validated outputs. Existing tools, while valuable, often presented limitations as confirmed by the benchmarking analysis (Chapter 4). For instance, while LINEX often reported a high number of initial enzyme associations and

achieved good, validated coverage, its results required significant subsequent curation by the user of non-validated entries (e.g., non-human enzymes) and involved a cumbersome multi-step process to extract usable data. BioPAN and LipidOne, though providing structured output, both required validation filtering (to exclude non-catalytic proteins) and reported fewer associations overall in tested datasets. Furthermore, the accessibility of results varied greatly, contrasting the direct, comprehensive package from LipidCRED with the more complex retrieval needed for others. These observed variations among existing tools – in terms of output reliability, scope and user workflow – highlighted the opportunity for a new platform like LipidCRED. It was this confluence of challenges that motivated the development of LipidCRED, the central contribution of this thesis. My objective was therefore: to create a platform that streamlines the path from lipid list to functional enzymatic insight.

5.2 LipidCRED's Approach

To meet this objective, I designed LipidCRED with several core innovations aimed at overcoming the limitations of previous approaches. Its foundation lies in a structured relational database (Chapter 2), carefully designed not just to store, but also to logically relate information about lipids, reactions, enzymes, and organisms. This structure incorporates curated data emphasizing data quality from the outset, for instance, by cross-referencing orthologous enzymes against the reviewed UniProtKB/Swiss-Prot database [121]. This design choice means the enzymatic associations LipidCRED provides have a curated level of biological validation, thereby minimizing the noise of non-reviewed or non-catalytic entries often encountered elsewhere, as substantiated by my benchmarking results (Chapter 4).

Perhaps the most significant conceptual advance is the *LipidMatcher* module (Chapter 2). Instead of relying solely on pre-defined, molecule-specific reactions, LipidCRED employs novel

reaction generalizations and intelligent carbon-chain matching algorithms specifically designed for each lipid class and reaction type. This allows the system to computationally infer specific reaction products based on class-level knowledge and the user's input lipid pool. This moves beyond a simple database lookup towards predictive biochemistry based on the experimental evidence for other members of the lipid class, drastically expanding the potential reaction space. LipidCRED can, thus, be used for newly identified (or "to be identified') lipid species currently not present in reaction databases. In addition to providing information about the thus-far unknown lipid molecules, this computational inference tackles the sheer combinatorial complexity of the lipidome in a way static databases cannot [137]. This approach directly addresses a key limitation highlighted by Lee et al. [138], who noted that simply observing metabolite level changes is often insufficient to infer pathway activity. By providing a biochemical blueprint of known and inferred reactions linked directly to specific gene products (enzymes), LipidCRED offers the necessary context required for more robust pathway interpretation, moving beyond simple enrichment statistics towards mechanism-based understanding.

Furthermore, to address practical challenges, significant effort was invested into developing a robust nomenclature parsing and translation system (Chapter 2) to handle the diverse and often inconsistent ways lipids are named, a common frustration point in lipid bioinformatics [139]. The overall backend architecture was implemented using Object-Oriented Programming (OOP) principles and a modular design (Chapter 2), separating concerns into distinct classes (e.g., *LipidTranslator*, *LipidProcessor*, *ReactionProcessor*, *LipidMatcher*). This software engineering approach not only facilitated the initial development but also makes LipidCRED extensible and maintainable, allowing for the easier addition of new functionalities, parsers for new lipid classes, or refined matching strategies in the future without requiring extensive rewrites of the core system.

Finally, recognizing that the results provided by even the best analysis is inadequate if inaccessible, I focused on providing clear, structured, and comprehensive outputs (Chapter 2), including detailed reaction lists with evidence, enzymes summaries, and adjacency matrices. This facilitates immediate downstream use – addressing issues with more opaque or difficult-to-parse outputs encountered with some tools benchmarked against LipidCRED (Chapter 4).

5.3 Addressing the Challenges and Limitations

LipidCRED's accuracy and comprehensiveness are fundamentally tied to the quality and scope of its source databases (SwissLipids, UniProtKB, and HGNC). Gaps or errors in these sources can be reflected in LipidCRED. The benchmarking identified specific classes (e.g., ether lipids, cardiolipin) needing better representation. Moreover, while powerful, *LipidMatcher*'s algorithms rely on defined rules for handling hydrocarbon chains. Unusual enzymatic mechanisms or complex rearrangements might not be captured by the current matching strategies, thus requiring manual addition. Similarly, despite sophisticated parsing of nomenclature, certain ambiguous cases can still pose challenges. Therefore, continuous refinement of the parsing rules and potentially including user feedback mechanisms are necessary. Understanding these limitations is key to interpreting LipidCRED's results appropriately and identifying priorities for its continued improvement.

5.4 Future Directions and Perspectives

The development and evaluation of LipidCRED presented in this thesis highlights strengths of this new application but also identifies exciting avenues for its future enhancement. Building upon the current framework, my immediate priority is to expand the scope and depth of LipidCRED's knowledge base. This future work will initially focus on expanding the representation of currently underrepresented lipid classes and subclasses within the Sphingolipid

and Glycerophospholipid categories, where gaps were identified during benchmarking (Chapter 4). Successfully incorporating more reactions will necessitate the concurrent development of new matching strategies within the *LipidMatcher* module (Section 2.4.5) to accurately predict reaction products. Following this consolidation, I plan to broaden LipidCRED's coverage significantly by systematically integrating reaction data and developing appropriate parsing and matching logic for other major lipid categories, including Glycerolipids, Sterol lipids, and Fatty Acyls, thereby moving toward a comprehensive representation of the entire lipidome.

Beyond expanding lipid coverage, I intend to implement an automated procedure for regularly updating the underlying knowledge base by pulling and processing new data from source databases like SwissLipids and UniProtKB to ensure LipidCRED remains current and maintainable in the rapidly evolving field of lipidomics. This automation, considered during the initial database design (Chapter 2), is vital for the long-term relevance and reliability of LipidCRED. Through these planned enhancements, I envision LipidCRED evolving into an even more powerful and comprehensive resource for researchers studying lipid metabolism.

5.5 Concluding Remarks: Significance of LipidCRED

This thesis presented LipidCRED, a novel platform conceived and developed to address critical issues in functional lipidomic analyses. Through its unique combination of a structured, high-quality knowledge base, intelligent reaction inference via the *LipidMatcher*, and clear, accessible outputs, LipidCRED empowers researchers to explore the enzymatic landscape underlying their lipid data more effectively. The rigorous benchmarking validated its performance and highlighted the significance of its design choices, particularly its emphasis on inherent data quality.

While future developments will undoubtedly enhance its capabilities, LipidCRED stands as a significant contribution. It represents a step towards more sophisticated, reliable, and user-friendly computational tools desperately needed in the complex field of lipidomics. By facilitating the transition from lipid lists to functional hypotheses about enzymatic activity, LipidCRED aids researchers in unraveling the intricate roles that lipids play in health and disease. The work presented here not only delivers a functional tool but also offers a perspective on how computational approaches, thoughtfully designed and rigorously validated, can push the boundaries of biological discovery in the post-genomic era.

References

[1]. Pauling, J. & Klipp, E. Computational Lipidomics and Lipid Bioinformatics: Filling In the Blanks. *J Integr Bioinform* **13**, 299 (2016).

[2]. Han, X. & Gross, R. W. Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. *Journal of Lipid Research* **44**, 1071–1079 (2003).

[3]. Fahy, E. *et al.* Update of the LIPID MAPS comprehensive classification system for lipids. *J Lipid Res* **50 Suppl**, S9-14 (2009).

[4]. Kishimoto, K., Urade, R., Ogawa, T. & Moriyama, T. Nondestructive Quantification of Neutral Lipids by Thin-Layer Chromatography and Laser-Fluorescent Scanning: Suitable Methods for "Lipidome" Analysis. *Biochemical and Biophysical Research Communications* **281**, 657–662 (2001).

[5]. Hannun, Y. A. & Obeid, L. M. Sphingolipids and their metabolism in physiology and disease. *Nat Rev Mol Cell Biol* **19**, 175–191 (2018).

[6]. Han, X. Lipidomics for studying metabolism. *Nat Rev Endocrinol* **12**, 668–679 (2016).

[7]. Xu, S. *et al.* Spatially and temporally probing distinctive glycerophospholipid alterations in Alzheimer's disease mouse brain via high-resolution ion mobility-enabled sn-position resolved lipidomics. *Nat Commun* **15**, 6252 (2024).

[8]. Vonkova, I. *et al.* Lipid Cooperativity as a General Membrane-Recruitment Principle for PH Domains. *Cell Reports* **12**, 1519–1530 (2015).

[9]. Köberlin, M. S. *et al.* A Conserved Circular Network of Coregulated Lipids Modulates Innate Immune Responses. *Cell* **162**, 170–183 (2015).

[10]. Rodriguez-Cuenca, S., Pellegrinelli, V., Campbell, M., Oresic, M. & Vidal-Puig, A. Sphingolipids and glycerophospholipids – The “ying and yang” of lipotoxicity in metabolic diseases. *Progress in Lipid Research* **66**, 14–29 (2017).

[11]. Ackerman, D. & Simon, M. C. Hypoxia, lipids, and cancer: surviving the harsh tumor microenvironment. *Trends in Cell Biology* **24**, 472–478 (2014).

[12]. Kim, E. J., Ramachandran, R. & Wierzbicki, A. S. Lipidomics in diabetes. *Current Opinion in Endocrinology, Diabetes and Obesity* **29**, 124 (2022).

[13]. Bennett, S. A. *et al.* Using neurolipidomics to identify phospholipid mediators of synaptic (dys)function in Alzheimer’s Disease. *Front. Physiol.* **4**, (2013).

[14]. Touboul, D. & Gaudin, M. Lipidomics of Alzheimer’s Disease. *Bioanalysis* **6**, 541–561 (2014).

[15]. Alecu, I. & Bennett, S. A. L. Dysregulated Lipid Metabolism and Its Role in α-Synucleinopathy in Parkinson’s Disease. *Front. Neurosci.* **13**, (2019).

[16]. Züllig, T., Trötzmüller, M. & Köfeler, H. C. Lipidomics from sample preparation to data analysis: a primer. *Anal Bioanal Chem* **412**, 2191–2209 (2020).

[17]. Hoffmann, N. *et al.* A Current Encyclopedia of Bioinformatics Tools, Data Formats and Resources for Mass Spectrometry Lipidomics. *Metabolites* **12**, 584 (2022).

[18]. Rose, T. D. *et al.* Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data. *Briefings in Bioinformatics* **24**, bbac572 (2023).

[19]. Hyötyläinen, T. & Orešič, M. Systems biology strategies to study lipidomes in health and disease. *Progress in Lipid Research* **55**, 43–60 (2014).

[20]. Dubot, P., Sabourdy, F. & Levade, T. Human genetic defects of sphingolipid synthesis. *Journal of Inherited Metabolic Disease* **48**, e12745 (2025).

[21]. Vial, H. J. & Wengelnik, K. Phospholipid Metabolism. in *Encyclopedia of Malaria* (eds. Hommel, M. & Kremsner, P. G.) 1–23 (Springer, New York, NY, 2021).

[22]. Marquês, J. T., Marinho, H. S. & de Almeida, R. F. M. Sphingolipid hydroxylation in mammals, yeast and plants – An integrated view. *Progress in Lipid Research* **71**, 18–42 (2018).

[23]. Bikman, B. T. & Summers, S. A. Ceramides as modulators of cellular and whole-body metabolism. *J Clin Invest* **121**, 4222–4230 (2011).

[24]. Lone, M. A. *et al.* Subunit composition of the mammalian serine-palmitoyltransferase defines the spectrum of straight and methyl-branched long-chain bases. *Proceedings of the National Academy of Sciences* **117**, 15591–15598 (2020).

[25]. Lone, M. A., Bourquin, F. & Hornemann, T. Serine palmitoyltransferase subunit 3 and metabolic diseases. *Sphingolipid Metabolism and Metabolic Disease* 47–56 (2022).

[26]. Weiss, B. & Stoffel, W. Human and murine serine-palmitoyl-CoA transferase--cloning, expression and characterization of the key enzyme in sphingolipid synthesis. *Eur J Biochem* **249**, 239–247 (1997).

[27]. Hornemann, T., Richard, S., Rütti, M. F., Wei, Y. & von Eckardstein, A. Cloning and initial characterization of a new subunit for mammalian serine-palmitoyltransferase. *J Biol Chem* **281**, 37275–37281 (2006).

[28]. Beeler, T. *et al.* The Saccharomyces cerevisiae TSC10/YBR265w gene encoding 3-ketosphinganine reductase is identified in a screen for temperature-sensitive suppressors of the Ca2+-sensitive csg2Delta mutant. *J Biol Chem* **273**, 30688–30694 (1998).

[29]. Kihara, A. & Igarashi, Y. FVT-1 is a mammalian 3-ketodihydrosphingosine reductase with an active site that faces the cytosolic side of the endoplasmic reticulum membrane. *J Biol Chem* **279**, 49243–49250 (2004).

[30]. Levy, M. & Futerman, A. H. Mammalian ceramide synthases. *IUBMB Life* **62**, 347–356 (2010).

[31]. Kuo, A. & Hla, T. Regulation of cellular and systemic sphingolipid homeostasis. *Nat Rev Mol Cell Biol* **25**, 802–821 (2024).

[32]. Saba, J. D. Fifty years of lyase and a moment of truth: sphingosine phosphate lyase from discovery to disease. *J Lipid Res* **60**, 456–463 (2019).

[33]. Huitema, K., van den Dikkenberg, J., Brouwers, J. F. H. M. & Holthuis, J. C. M. Identification of a family of animal sphingomyelin synthases. *EMBO J* **23**, 33–44 (2004).

[34]. Zhang, T., de Waard, A. A., Wuhrer, M. & Spaapen, R. M. The Role of Glycosphingolipids in Immune Cell Functions. *Front. Immunol.* **10**, (2019).

[35]. Allende, M. L. & Proia, R. L. Simplifying complexity: genetically resculpting glycosphingolipid synthesis pathways in mice to reveal function. *Glycoconj J* **31**, 613–622 (2014).

[36]. D'Angelo, G., Capasso, S., Sticco, L. & Russo, D. Glycosphingolipids: synthesis and functions. *The FEBS Journal* **280**, 6338–6353 (2013).

[37]. Sabourdy, F. *et al.* Functions of sphingolipid metabolism in mammals — Lessons from genetic defects. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1781**, 145–183 (2008).

[38]. Senkal, C. E. *et al.* Ceramide Is Metabolized to Acylceramide and Stored in Lipid Droplets. *Cell Metab* **25**, 686–697 (2017).

[39]. Hirabayashi, T., Murakami, M. & Kihara, A. The role of PNPLA1 in ω-*O*-acylceramide synthesis and skin barrier function. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1864**, 869–879 (2019).

[40]. Hait, N. C. & Maiti, A. The Role of Sphingosine-1-Phosphate and Ceramide-1-Phosphate in Inflammation and Cancer. *Mediators of Inflammation* **2017**, 4806541 (2017).

[41]. Penno, A. *et al.* Hereditary Sensory Neuropathy Type 1 Is Caused by the Accumulation of Two Neurotoxic Sphingolipids*♦. *Journal of Biological Chemistry* **285**, 11178–11187 (2010).

[42]. Zitomer, N. C. *et al.* Ceramide Synthase Inhibition by Fumonisin B1 Causes Accumulation of 1-Deoxysphinganine. *Journal of Biological Chemistry* **284**, 4786–4795 (2009).

[43]. Karsai, G. *et al.* FADS3 is a Δ14Z sphingoid base desaturase that contributes to gender differences in the human plasma sphingolipidome. *Journal of Biological Chemistry* **295**, 1889–1897 (2020).

[44]. Lone, M. A., Santos, T., Alecu, I., Silva, L. C. & Hornemann, T. 1-Deoxysphingolipids. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1864**, 512–521 (2019).

[45]. Santos, T. C. B. *et al.* The long chain base unsaturation has a stronger impact on 1-deoxy(methyl)-sphingolipids biophysical properties than the structure of its C1 functional group. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1863**, 183628 (2021).

[46]. Gibellini, F. & Smith, T. K. The Kennedy pathway—De novo synthesis of phosphatidylethanolamine and phosphatidylcholine. *IUBMB Life* **62**, 414–428 (2010).

[47]. Dowhan, W. & Bogdanov, M. Eugene P. Kennedy's Legacy: Defining Bacterial Phospholipid Pathways and Function. *Front. Mol. Biosci.* **8**, (2021).

[48]. Cole, L. K., Vance, J. E. & Vance, D. E. Phosphatidylcholine biosynthesis and lipoprotein metabolism. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1821**, 754–761 (2012).

[49]. Vance, J. E. & Vance, D. E. Phospholipid biosynthesis in mammalian cells. *Biochem. Cell Biol.* **82**, 113–128 (2004).

[50]. Steenbergen, R. *et al.* Disruption of the Phosphatidylserine Decarboxylase Gene in Mice Causes Embryonic Lethality and Mitochondrial Defects*. *Journal of Biological Chemistry* **280**, 40032–40040 (2005).

[51]. Vance, J. E. & Tasseva, G. Formation and function of phosphatidylserine and phosphatidylethanolamine in mammalian cells. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1831**, 543–554 (2013).

[52]. Holub, B. J. Metabolism and Function of myo-Inositol and Inositol Phospholipids. *Annual Review of Nutrition* **6**, 563–597 (1986).

[53]. Blunsom, N. J. & Cockcroft, S. Phosphatidylinositol synthesis at the endoplasmic reticulum. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1865**, 158471 (2020).

[54]. Ridgway, N. D. Phospholipid synthesis in mammalian cells. in *Biochemistry of lipids, lipoproteins and membranes* 227–258 (Elsevier, 2021).

[55]. Watschinger, K. & Werner, E. R. Orphan enzymes in ether lipid metabolism. *Biochimie* **95**, 59–65 (2013).

[56]. Lands, W. E. M. Metabolism of Glycerolipids: II. THE ENZYMATIC ACYLATION OF LYSOLECITHIN. *Journal of Biological Chemistry* **235**, 2233–2237 (1960).

[57]. Shindou, H. & Shimizu, T. Acyl-CoA:Lysophospholipid Acyltransferases *. *Journal of Biological Chemistry* **284**, 1–5 (2009).

[58]. Kita, Y., Shindou, H. & Shimizu, T. Cytosolic phospholipase A2 and lysophospholipid acyltransferases. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1864**, 838–845 (2019).

[59]. Casares, D., Escribá, P. V. & Rosselló, C. A. Membrane Lipid Composition: Effect on Membrane and Organelle Structure, Function and Compartmentalization and Therapeutic Avenues. *International Journal of Molecular Sciences* **20**, 2167 (2019).

[60]. Futerman, A. H. Chapter 9 - Sphingolipids. in *Biochemistry of Lipids, Lipoproteins and Membranes (Seventh Edition)* (eds. Ridgway, N. D. & McLeod, R. S.) 281–316 (Elsevier, 2021).

[61]. Liebisch, G., Ekroos, K., Hermansson, M. & Ejsing, C. S. Reporting of lipidomics data should be standardized. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1862**, 747–751 (2017).

[62]. Lam, S. M., Tian, H. & Shui, G. Lipidomics, en route to accurate quantitation. *Biochim Biophys Acta Mol Cell Biol Lipids* **1862**, 752–761 (2017).

[63]. Lam, S. M., Wang, Z., Li, B. & Shui, G. High-coverage lipidomics for functional lipid and pathway analyses. *Analytica Chimica Acta* **1147**, 199–210 (2021).

[64]. Quarles, R. H., Clarke, N. & Dawson, R. M. C. Isolation of N-acyl phosphatidylethanolamine from pea seeds. *Biochemical and Biophysical Research Communications* **33**, 964–968 (1968).

[65]. Hansen, H. S., Lauritzen, L., Moesgaard, B., Strand, A. M. & Hansen, H. H. Formation of N-acyl-phosphatidylethanolamines and N-acetylethanolamines: proposed role in neurotoxicity. *Biochem Pharmacol* **55**, 719–725 (1998).

[66]. Hansen, H. S., Moesgaard, B., Hansen, H. H., Schousboe, A. & Petersen, G. Formation ofN-acyl-phosphatidylethanolamine andN-acylethanolamine (including anandamide) during glutamate-induced neurotoxicity. *Lipids* **34**, S327–S330 (1999).

[67]. Guan, X. L. *et al.* Non-targeted profiling of lipids during kainate-induced neuronal injury. *FASEB J* **20**, 1152–1161 (2006).

[68]. Schmid, P. C. *et al.* Occurrence and postmortem generation of anandamide and other long-chain N-acylethanolamines in mammalian brain. *FEBS Lett* **375**, 117–120 (1995).

[69]. Astarita, G. & Piomelli, D. Lipidomic analysis of endocannabinoid metabolism in biological samples. *J Chromatogr B Analyt Technol Biomed Life Sci* **877**, 2755–2767 (2009).

[70]. Hishikawa, D., Valentine, W. J., Iizuka-Hishikawa, Y., Shindou, H. & Shimizu, T. Metabolism and functions of docosahexaenoic acid-containing membrane glycerophospholipids. *FEBS Letters* **591**, 2730–2744 (2017).

[71]. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research* **49**, D545–D551 (2021).

[72]. Agrawal, A. *et al.* WikiPathways 2024: next generation pathway database. *Nucleic Acids Research* **52**, D679–D689 (2024).

[73]. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics* **20**, 1085 (2019).

[74]. Milacic, M. *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research* **52**, D672–D678 (2024).

[75]. Bansal, P. *et al.* Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Research* **50**, D693–D700 (2022).

[76]. Saccenti, E., Hoefsloot, H. C. J., Smilde, A. K., Westerhuis, J. A. & Hendriks, M. M. W. B. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* **10**, 361–374 (2014).

[77]. Chappel, J. R., Kirkwood-Donelson, K. I., Reif, D. M. & Baker, E. S. From big data to big insights: statistical and bioinformatic approaches for exploring the lipidome. *Analytical and bioanalytical chemistry* **416**, 2189 (2023).

[78]. Demšar, J. *et al.* Orange: Data mining toolbox in python. *Journal of Machine Learning Research* **14**, 2349–2353 (2013).

[79]. Frank, E., Hall, M. A. & Witten, I. H. The WEKA workbench. Online appendix for 'data mining: Practical machine learning tools and techniques'. (2016).

[80]. Berthold, M. R. *et al.* KNIME - the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explor. Newsl.* **11**, 26–31 (2009).

[81]. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

[82]. Nguyen-Tran, T., Alkassir, Q., Bennett, S. A. & Cuperlovic-Culf, M. Network development and comparison in lipidomics and metabolomics. in *Metabolomics: Recent advances and future applications* 39–57 (Springer, 2023).

[83]. Monti, F. *et al.* Signed Distance Correlation (SiDCo): an online implementation of distance correlation and partial distance correlation for data-driven network analysis. *Bioinformatics* **39**, btad210 (2023).

[84]. DiLeo, M. V., Strahan, G. D., Bakker, M. den & Hoekenga, O. A. Weighted Correlation Network Analysis (WGCNA) Applied to the Tomato Fruit Metabolome. *PLOS ONE* **6**, e26683 (2011).

[85]. Lei, J., Cai, Z., He, X., Zheng, W. & Liu, J. An approach of gene regulatory network construction using mixed entropy optimizing context-related likelihood mutual information. *Bioinformatics* **39**, btac717 (2023).

[86]. Pang, Z. *et al.* MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Research* **52**, W398–W406 (2024).

[87]. Goracci, L. *et al.* Lipostar, a Comprehensive Platform-Neutral Cheminformatics Tool for Lipidomics. *Anal. Chem.* **89**, 6257–6264 (2017).

[88]. Manzini, S., Busnelli, M., Colombo, A., Kiamehr, M. & Chiesa, G. liputils: a Python module to manage individual fatty acid moieties from complex lipids. *Sci Rep* **10**, 13368 (2020).

[89]. Mohamed, A. & Hill, M. M. LipidSuite: interactive web server for lipidomics differential and enrichment analysis. *Nucleic Acids Research* **49**, W346–W351 (2021).

[90]. Castañé, H. *et al.* Coupling Machine Learning and Lipidomics as a Tool to Investigate Metabolic Dysfunction-Associated Fatty Liver Disease. A General Overview. *Biomolecules* **11**, 473 (2021).

[91]. Lin, W.-J. *et al.* LipidSig: a web-based tool for lipidomic data analysis. *Nucleic Acids Research* **49**, W336–W345 (2021).

[92]. Ni, Z. *et al.* Guiding the choice of informatics software and tools for lipidomics research applications. *Nat Methods* **20**, 193–204 (2023).

[93]. Kimura, T., Jennings, W. & Epand, R. M. Roles of specific lipid species in the cell and their molecular mechanism. *Progress in Lipid Research* **62**, 75–92 (2016).

[94]. Conroy, M. J. *et al.* LIPID MAPS: update to databases and tools for the lipidomics community. *Nucleic Acids Research* **52**, D1677–D1682 (2024).

[95]. Aimo, L. *et al.* The SwissLipids knowledgebase for lipid biology. *Bioinformatics* **31**, 2860–2866 (2015).

[96]. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res* **49**, D1388–D1395 (2020).

[97]. Wishart, D. S. *et al.* HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res* **50**, D622–D631 (2022).

[98]. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research* **44**, D1214–D1219 (2016).

[99]. Fahy, E. *et al.* A comprehensive classification system for lipids1. *Journal of Lipid Research* **46**, 839–861 (2005).

[100]. Liebisch, G. *et al.* Update on LIPID MAPS classification, nomenclature, and shorthand notation for MS-derived lipid structures. *J Lipid Res* **61**, 1539–1555 (2020).

[101]. Sud, M. *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Research* **35**, D527–D532 (2007).

[102]. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2023).

[103]. Liebisch, G. *et al.* Shorthand notation for lipid structures derived from mass spectrometry. *Journal of Lipid Research* **54**, 1523–1530 (2013).

[104]. Schomburg, I., Chang, A. & Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* **30**, 47–49 (2002).

[105]. Li, F., Chen, Y., Anton, M. & Nielsen, J. GotEnzymes: an extensive database of enzyme parameter predictions. *Nucleic Acids Research* **51**, D583–D586 (2023).

[106]. Stevens, R. Ontology based document enrichment in bioinformatics. *Comparative and Functional Genomics* **3**, 42–46 (2002).

[107]. Clair, G. *et al.* Lipid Mini-On: mining and ontology tool for enrichment analysis of lipidomic data. *Bioinformatics* **35**, 4507–4508 (2019).

[108]. Molenaar, M. R. *et al.* LION/web: a web-based ontology enrichment tool for lipidomic data analysis. *GigaScience* **8**, giz061 (2019).

[109]. More, P., Bindila, L., Wild, P., Andrade-Navarro, M. & Fontaine, J.-F. LipiDisease: associate lipids to diseases using literature mining. *Bioinformatics* **37**, 3981–3982 (2021).

[110]. Habibpour, M., Razaghi-Moghadam, Z. & Nikoloski, Z. Machine learning of metabolite–protein interactions from model-derived metabolic phenotypes. *NAR Genomics and Bioinformatics* **6**, lqae114 (2024).

[111]. García-Campos, M. A., Espinal-Enríquez, J. & Hernández-Lemus, E. Pathway Analysis: State of the Art. *Front. Physiol.* **6**, (2015).

[112]. Felippe, H., Battiston, F. & Kirkley, A. Network mutual information measures for graph similarity. *Commun Phys* **7**, 1–12 (2024).

[113]. Gaud, C. *et al.* BioPAN: a web-based tool to explore mammalian lipidome metabolic pathways on LIPID MAPS. *F1000Research* **10**, 4 (2021).

[114]. Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS computational biology* **4**, e1000217 (2008).

[115]. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18**, S233–S240 (2002).

[116]. Nguyen, A., Rudge, S. A., Zhang, Q. & Wakelam, M. J. Using lipidomics analysis to determine signalling and metabolic changes in cells. *Current Opinion in Biotechnology* **43**, 96–103 (2017).

[117]. Alabed, H. B. R. *et al.* LipidOne 2.0: A Web Tool for Discovering Biological Meanings Hidden in Lipidomic Data. *Current Protocols* **4**, e70009 (2024).

[118]. Aimo, L. *et al.* The SwissLipids knowledgebase for lipid biology. *Bioinformatics* **31**, 2860–2866 (2015).

[119]. Gamma, E., Helm, R., Johnson, R. & Vlissides, J. *Design Patterns: Elements of Reusable Object-Oriented Software*. (Addison-Wesley Longman Publishing Co., Inc., USA, 1995).

[120]. Yates, B., Gray, K. A., Jones, T. E. M. & Bruford, E. A. Updates to HCOP: the HGNC comparison of orthology predictions tool. *Briefings in Bioinformatics* **22**, bbab155 (2021).

[121]. Famiglietti, M. L. *et al.* An enhanced workflow for variant interpretation in UniProtKB/Swiss-Prot improves consistency and reuse in ClinVar. *Database* **2019**, baz040 (2019).

[122]. Johnson, T. & Shasha, D. 2Q: A Low Overhead High Performance Buffer Management Replacement Algorithm. in (1994).

[123]. R Core Team. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/ (2023).

[124]. Chang, W. *et al. Shiny: Web Application Framework for R*. https://shiny.posit.co/ (2025).

[125]. Ushey, K., Allaire, J. & Tang, Y. *Reticulate: Interface to 'Python'*. https://github.com/rstudio/reticulate (2025).

[126]. Chang, W. & Borges Ribeiro, B. *Shinydashboard: Create Dashboards with 'Shiny'*. https://github.com/rstudio/shinydashboard (2025).

[127]. Xie, Y., Cheng, J. & Tan, X. *DT: A Wrapper of the JavaScript Library 'DataTables'*. https://github.com/rstudio/DT (2022).

[128]. Meyer, F. & Perrier, V. *Shinybusy: Busy Indicators and Notifications for 'shiny' Applications*. https://github.com/dreamRs/shinybusy (2024).

[129]. Otasek, D., Morris, J. H., Bouças, J., Pico, A. R. & Demchak, B. Cytoscape Automation: empowering workflow-based network analysis. *Genome Biol* **20**, 185 (2019).

[130]. Mohamed, A., Molendijk, J. & Hill, M. M. lipidr: A Software Tool for Data Mining and Analysis of Lipidomics Datasets. *J. Proteome Res.* **19**, 2890–2897 (2020).

[131]. Alabed, H. B. R. *et al.* Comparison between Sickle Cell Disease Patients and Healthy Donors: Untargeted Lipidomic Study of Erythrocytes. *Int J Mol Sci* **24**, 2529 (2023).

[132]. Hahn, O. *et al.* A nutritional memory effect counteracts the benefits of dietary restriction in old mice. *Nat Metab* **1**, 1059–1073 (2019).

[133]. Wang, Y. *et al.* Shotgun lipidomics-based characterization of the landscape of lipid metabolism in colorectal cancer. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1865**, 158579 (2020).

[134]. Hornburg, D. *et al.* Dynamic lipidome alterations associated with human health, disease and ageing. *Nat Metab* **5**, 1578–1594 (2023).

[135]. Adam-Klages, S. *et al.* FAN, a novel WD-repeat protein, couples the p55 TNF-receptor to neutral sphingomyelinase. *Cell* **86**, 937–947 (1996).

[136]. Kudo, N. *et al.* Crystal structures of the CERT START domain with inhibitors provide insights into the mechanism of ceramide transfer. *J Mol Biol* **396**, 245–251 (2010).

[137]. Dingjan, T. & Futerman, A. H. The fine-tuning of cell membrane lipid bilayers accentuates their compositional complexity. *BioEssays* **43**, 2100021 (2021).

[138]. Lee, K. S., Su, X. & Huan, T. Metabolites are not genes—avoiding the misuse of pathway analysis in metabolomics. *Nature Metabolism* 1–4 (2025).

[139]. Witting, M. *et al.* Challenges and perspectives for naming lipids in the context of lipidomics. *Metabolomics* **20**, 1–9 (2024).