Advancing Lipidomic Bioinformatics: Visualization and phosphoLipid IDentification (VaLID)

Graeme S.V. McDowell

This thesis is submitted to the Faculty of Graduate and Postdoctoral Studies as a partial fulfillment of the requirements for the degree of Masters in Neuroscience

Department of Cellular & Molecular Medicine
Faculty of Medicine
University of Ottawa

Abstract

Lipidomics is a relatively new field under the heading of systems biology. Due to its infancy, the field suffers from significant 'growing pains', one of which is the lack of bioinformatic analytic resources that other "-*omics*" fields enjoy. Here, I describe the creation and validation of the glycerophospholipid identification program VaLID. Using an *in silico* approach, we generated a comprehensive database containing all of the glycerophospholipids within multiple sub-classes: those containing chains of 0 to 30 carbons with up to 6 unsaturations and various linkages. Using Java, I created a web-based computer interface with a search engine and a visualization tool to access this database. In comparing results to current programs, I found that VaLID consistently contained more identity predictions than did the current gold standard LipidMAPS. Results from several tests with real datasets confirm that VaLID is more than capable as a phospholipid identification tool for use in lipidomics.

Contents

List of Tables

List of Figures

List of Abbreviations

ACh	acetylcholine
AD	Alzheimer's disease
CIMS	Carleton Immersive Media Studio
CTPNL	CIHR Training Program in Neurodegenerative Lipidomics
ESI	electrospray ionization
FA	formic acid
GUI	graphical user interface
HB	homogenization buffer
HMDB	Human Metabolome Database
HPLC	high performance liquid chromatography
HRP	horseradish peroxidase
IDE	integrated development environment
IWBBIO	International Work Conference on Bioinformatics and Biomedical Engineering
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC	liquid chromatography
LC-ESI-MS	high performance liquid chromatography electrospray ionization tandem mass spectrometry
MALDI	matrix-assisted laser desorption ionization
MS	mass spectrometry
MS/MS	tandem mass spectrometry
m/z	mass-to-charge ratio
NeuN	neuronal nuclei
PA	glycerophosphate
PAF	platelet activating factor
PBS	phosphate buffered saline
PC	glycerophosphocholine
PE	glycerophosphoethanolamine
PH	pleckstrin homology
PI	glycerophosphoinositol
PIP	glycerophosphoinositol mono-phosphates
PIP_2	glycerophosphoinositol bis-phosphates
PIP_3	glycerophosphoinositol tris –phosphate
PS	glycerophosphoserine
PSD	post-synaptic density
PVDF	polyvinylidene fluoride
RT	retention time
SDF	MDL SDfile
SMILES	simplified molecular-input line entry system
TLC	thin layer chromatography
VaLID	Visualization and phosphoLipid IDentification
XIC	extracted ion chromatography

Acknowledgments

My words cannot adequately express the gratitude I feel at this moment. So forgive me while I try. I would like to give my deepest and sincerest thanks to my supervisor, Dr. Steffany Bennett. Her guidance, advice and assistance were instrumental to the completion of this work; Without her help, VaLID would be unrecognizable. Thank you, Stef, for your constant support, and for trusting me when even I did not. I would also like to thank the Bennett lab members, past and present, for your support, companionship and help throughout the countless hours we've had together. A special thanks to Alexandre P Blanchard, Graeme Taylor, Stephanie Fowler, Mark Akins, Dr. Hongbin Xu, Dr. Yun Wang and Matthew Granger, for your assistance in training me, running samples for me, and helping to deal with all of my experimental animals. Without your help, I would surely have failed, many times over. To Fred Elisma and Daniel Jedrysiak, your help with Java was instrumental to making this program work. I would like to thank my advisory committee members, Dr. Daniel Figeys and Dr. Theodore Perkins for their guidance and insight throughout my Master's project.

It would be unjust of me not to thank my family for their constant and undying love and support. Mom and Dad, thank you for putting up with me in my 'lair', through my culinary experimentations, and for always pushing me to do the best I can do. Dad, thank you for pushing me to complete this monster of a document, and for bouncing ideas back-and-forth. Your linguistic mastery was indispensable to help forge this document into something coherent. Thank you Karin, Marcus, Dani, Althea, Emma and Adam for making me smile, and thank you Wes and Kris for challenging me intellectually, and for inviting me to watch all those UFC matches. To my friends: you all have a very special

place in my heart; I love you all. I have to give special thanks to Paolo for your constant friendship throughout all these years, Ben, Aggy, and the rest of the Gumdo crew, for still putting up with my visits. To the Knights – Nicholas, Larry and Drew – (y)our shenanigans and jokes never cease to brighten my day. To the NotGumdo server – C-team, Parthiv *et al* – thanks for all the great times. To Laura, Mishelle, Cassie, Shayna, Dan, Kim and Guinevere, thanks for all the chats and hangouts; there were never enough of either. To Cynthia and Christina, thanks for the concerts and altogether good times. Thank you to my friends and family at East Wind, everyone at RGNfit and to Natasha, for reminding me that an active body helps an active mind. And thank you to Alicia for being there when I decided to journey down this rabbit-hole. And of course, to anyone I haven't named: I haven't forgotten you, and I never will.

And lastly, I would like to thank you, dear reader. For without you, this work has no meaning. I dedicate this to you[1].

[1] *Bantis zōbrie issa se ossÿngnoti lēdys.*

Chapter 1: Introduction

1.1 The Field of Lipidomics

1.1.1 The Re-emergence of Lipidomics

Fats, or lipids, are one of the major macromolecules in biology, and they play many critical roles in biological systems. Lipids are defined as substances of biological origin that are soluble in organic solvents such as chloroform and methanol, but are only sparingly soluble, if at all, in water [1]. While lipids are created from relatively few chemical building blocks, an enormous number of chemically and structurally diverse molecules can be created from combinations of these hydrophobic fatty acids and backbone structures including glycerol or sphingosines, generating different pigments, vitamins, fatty acids, cholesterols, phospholipids, sphingolipids and many others [2-4]. A significant portion of our genome is devoted to the synthesis, metabolism and regulation of this lipid diversity [4]. Despite a long-established awareness of the basic chemistry of lipids, it was only recently that researchers began to grasp the significant role that lipids play in human health and diseases [5], and researchers are still finding new avenues to explore within the rapidly expanding world of lipids and lipidomics. For example, lipids are being studied as biomarkers for diseases for which biomarkers were previously difficult to obtain [6].

Systems biology, or "-*omics*", seeks to understand not only the individual components of a biological system such as a single gene, protein, or lipid species, but also the system's structure and dynamics [7]. This includes understanding the network of interactions and biochemical pathways, including the mechanisms by which these

interactions are modulated, and then how the system behaves dynamically over time under varying conditions [7]. A systems biology approach therefore focuses not only on identifying the individual components of a biological system, but also on the nature of the interactions between those components, and how they are modulated and modified over time and under different conditions. The popularity of a systems approach is due, in large part, to technical advances in applying this approach to molecular biology, notably the development of high-throughput technologies that enable researchers to collect comprehensive datasets that allow fuller understanding of a biological systems structure and functioning [7]. Systems approaches have been used to profile sets of genes of organisms (genomics), proteins (proteomics), gene expression (transcriptomics), and more recently, the metabolites present in a given cell (metabolomics) under different conditions. One of the subfields within metabolomics concerns itself with lipids; this is the field of lipidomics [8]. Lipidomics is arguably the most recent, and fastest expanding field of the systems-level analysis, focusing on the study of lipids, their biological role, and their interacting partners [2, 8, 9].

In the 1960's and 70's, lipids became an intense focus of research [10], but the field was limited by a lack of analytical technology capable of distinguishing individual lipid species [2]. Only larger family-level analyses were possible at the time. For example, in the 1980s, lipid samples were mostly analyzed by chromatography techniques such as thin-layer chromatography (TLC) or gas chromatography [11], which could only separate general lipid classes [4]. TLC could distinguish sphingomyelins and glycerophosphocholines (PCs), but could not separate the individual lipid species within these classes. Perhaps this is why lipids were assumed not to have information-carrying

functions, so that the plasma membrane, for example, was considered to be a relatively homogenous sea of lipids, peppered with interesting proteins but devoid of significant biological function in and of itself [4]. Hence this field was overshadowed by advances in other fields such as genomics and proteomics [10]. That is, until recently; application of technologies from other "-omics" fields to the study of lipids has initiated a lipid renaissance and this rapidly expanding technological frontier forms the topic of this thesis [2, 9, 10, 12].

1.1.2 Advances in Mass Spectrometry that Led to the Re-emergence of Lipidomics

This re-emergence of lipid biochemistry in the form of lipidomics was driven by advances in both biochemistry and analytical techniques: specifically, the development of electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) [11] mass spectrometry (MS). These 'soft' ionization techniques enabled researchers to generate molecular ions of high-mass, non-volatile compounds, such as lipids, without fragmenting them [13]. This enabled researchers to characterize lipids, via MS, at the molecular level for the first time. MS is one of the most widely used analytical methods in lipidomics; it is able to separate and characterize charged analytes in the gas phase, according to their mass-to-charge ratios (m/z). MS can also provide further information by fragmenting these charged analytes via various methods, such as collision-induced dissociation [13], or by a hard-ionization technique. The soft ionization techniques enabled researchers to produce full-sized, gas-phase lipid ions. In MALDI, introduced in 1988 by Karas and Hillenkamp [14], this is achieved in two steps. In the first, the analyte is dissolved in a solvent containing small organic molecules, called the matrix. This

solution is dried on a metallic substrate, resulting in a 'solid solution' of analyte-doped crystalline matrix. In the second step, the matrix is pulsed with laser photons, inducing a rapid heating of the crystals, causing localized sublimation of these matrix crystals and causing a portion of the crystal surface to move into the gas phase, taking the analyte with it. The analyte molecules, now in the gas phase, ionize during this reaction by a mechanism that is not fully understood [13-16]. The analyte is then read by a mass analyzer. In ESI, first proposed by Malcolm Dole in 1966, and demonstrated by Fenn *et al* [15, 16], analyte in volatile solution is passed through a small-bore, thin-walled tube (often a short length of hypodermic needle tubing), which is maintained at a high potential relative to an opposing counter-electrode. An electric field at the tube tip disperses the liquid into a fine spray of charged droplets. These droplets continue to break up into yet smaller droplets, until each one contains a single, charged, analyte molecule [12, 13, 15]. The charge of the needle will determine whether the ions are positively or negatively charged [15]. The analyte is then read into a mass analyzer. However, due to the soft ionization needed, identification and characterization of lipids depend on tandem MS analyses (MS/MS), which requires multiple mass analyzers [17]. The most commonly employed MS/MS approach uses a triple quadrupole mass spectrometer, which consists of three sets of quadrupoles [17]. Each quadrupole is a device which uses oscillating electric fields to separate ions according to their m/z. It is made up of four perfectly parallel rods arranged in a diamond pattern. Opposing rods carry the same charge, and these rods oscillate charge. An ion entering into the quadrupole will travel straight along the z axis, yet is attracted to the rods of opposite charge on the x or y axes, *i.e.*, a positively charged lipid will be attracted to one of the

negatively charged rods. If the ion reaches the rod before that rod's potential changes, the ion will discharge. Otherwise, the rod switches its potential charge and the ion changes direction along the x- or y- axis, while continuing forward in the z direction [16]. The ion's m/z, and the strength of the electric field along the rods, determines how far the ion is deflected towards the respective rod along the x- or y- axis. By altering the strength of the rods' electric fields, the analyzer effectively selects for different m/z ions, while scanning through the possible ranges. If the ion has the right m/z for the strength of the field, and the frequency with which the rods switch their potential is correct, the ion will stay close to the centre of the rods, allowing it to travel through into the detector [16]. With this new ability to get the analyte charged and into the mass spectrometer, different MS techniques were developed in order to determine what lipids were in the samples.

The MS techniques used with lipidomics fall into two main approaches: direct infusion, or shotgun MS, and chromatography-based MS [8, 11, 13]; they can use either ESI or MALDI as their ion sources. The shotgun lipidomics approach minimizes the use of chromatography, and thus uses no pre-separation of the lipid extract [8, 11]. This approach is accurate, reproducible, highly sensitive and takes less time than other approaches [8]. The shotgun method enables a comprehensive coverage of multiple lipid families simultaneously. Shotgun lipidomics typically utilizes MS/MS to be able to identify lipids based upon their subclass. Four main MS/MS modes are used in lipidomics, including application in shotgun lipidomics. These are: product-ion scan, precursor-ion scan, neutral-loss scan and selected reaction monitoring [17] (Figure 1.1).

Figure 1.1: Product-ion scan, precursor-ion scan, neutral-loss scan and selected reaction monitoring. The schematic depicts MS/MS modes used in both shotgun and chromatography based lipidomic MS. (A) Product Ion Scan mode. The first quadrupole (Q1) selects a target m/z, defined by the experimenter (m_x), and will only allow this mass to pass through the analyzer. The second quadrupole (Q2) acts as a collision chamber, where the ions fragment after colliding with inert gasses, such as nitrogen, and the third quadrupole (Q3) scans the resulting fragment ion peaks (p_1, p_2 and p_3), recording their individual masses with the detector. (B) Precursor Ion Scan mode. Q1 is set to scan the m/z of all the ions sent into the mass spectrometer, Q2 acts as a fragmentation chamber and Q3 is set to monitor for a specific fragment m/z (p_x). If the fragment m/z is detected in Q3, the precursor mass from Q1 that created that fragment mass is recorded. (C) Neutral Loss Scan mode. In neutral loss scan mode, Q1 scans through all m/z of the ions passed to it, similar to precursor ion scan. Q2 again acts as a fragmentation chamber. Q3 will scan the fragment ions, similar to product ion scan. The mass spectrometer then takes the m/z from Q1 and checks all of the m/z that have passed through Q3. If there is a pair of m/z where the mass from Q1 minus the mass from Q3 is equal to the m/z as set by the experimenter (NL), the precursor ion's m/z is recorded (*i.e.*, $m_x - p_x = NL$). (D) Selected Reaction Monitoring. Both Q1 and Q3 are set to only permit an ion of defined m/z to pass through from Q1 into the collision chamber Q2 (m_x), and from Q3 into the detector (p_x). This method can also be programmed to detect multiple discrete m/z in either or both Q1 and Q3, and is then referred to as Multiple Reaction Monitoring.

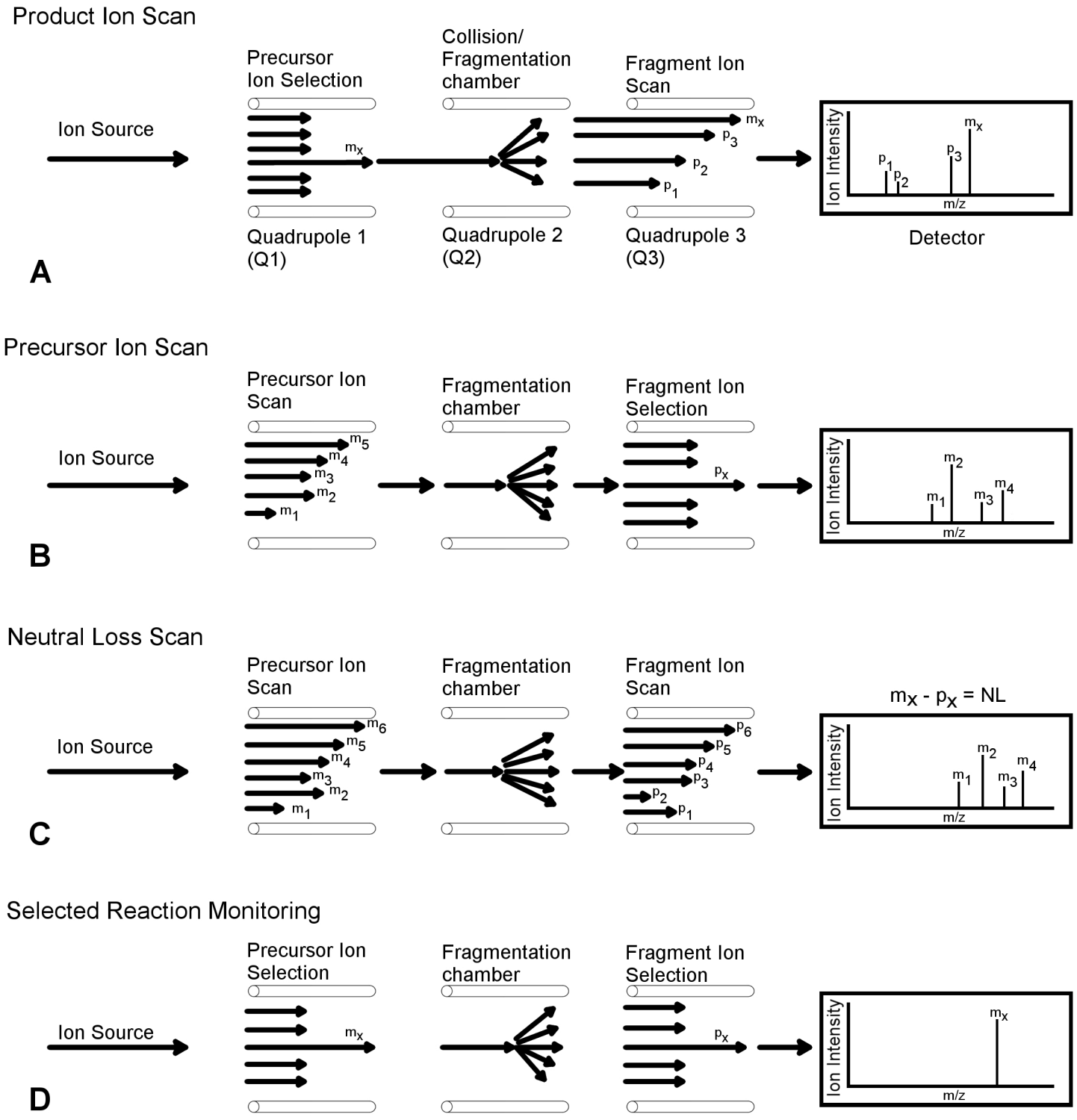

Product Ion Scan
Ion Source
Precursor Ion Selection
Collision/ Fragmentation chamber
Fragment Ion Scan
Quadrupole 1 (Q1)
Quadrupole 2 (Q2)
Quadrupole 3 (Q3)
Detector
Ion Intensity
m/z
A
Precursor Ion Scan
Ion Source
Precursor Ion Scan
Fragmentation chamber
Fragment Ion Selection
Ion Intensity
m/z
B
Neutral Loss Scan
Ion Source
Precursor Ion Scan
Fragmentation chamber
Fragment Ion Scan
mx - px = NL
Ion Intensity
m/z
C
Selected Reaction Monitoring
Ion Source
Precursor Ion Selection
Fragmentation chamber
Fragment Ion Selection
Ion Intensity
m/z
D

Figure 1.1

In product-ion scan analysis (Figure 1.1A), the first analyzer, or quadrupole Q1, selects a particular mass of interest, and any ions with this m/z are selected and then fragmented in the collision cell, quadrupole Q2, and the resultant product ions are analyzed with the second mass analyzer, quadrupole Q3 [17]. The structure of the lipid can be deduced by reconstructing the fragment ions, based on the masses reported from Q3. In precursor-ion scan analysis (Figure 1.1B), Q1 scans the m/z of the ions that enter into the spectrometer, Q2 fragments these ions, and Q3 selects a particular ion of interest [17]. The precursor ions in Q1 that produce this selected ion in Q3 are detected and identified. This is one method of identifying multiple species within a subclass, for example by detecting a peak at 184 m/z representing the PC head group. In neutral-loss scan analysis (Figure 1.1C), both Q1 and Q3 are scanned simultaneously, while Q2 fragments the ions. If a precursor of mass M in Q1, when fragmented produces a product of mass P in Q3, such that M – P = NL, where NL is the mass of the diagnostic peak which the mass spectrometer is set to record, then the precursor ion's mass is recorded [17]. Similar to precursor-ion scan, a class of lipids that have a given neutral-loss fragment (mass NL) can be identified with this mode, such as glycerophosphoethanolamines (PE) with a neutral loss mass of 141 m/z, or the glycerophosphoserines (PS) with a neutral loss mass of 185 m/z. Finally, for selected reaction-monitoring mode, transitions between molecule ion and product ion must be known. In this method, Q1 is locked at a specific m/z, Q2 fragments this ion, and Q3 is also locked at a specific m/z. If both precursor and product ions have the correct mass, then the precursor mass is saved. This technique yields high specificity and sensitivity for one specific ion of interest. If Q1 and/or Q3 are set to monitor multiple m/z's, *i.e.*,

more than one ion for multiple reactions, this is called "multiple reaction monitoring" [17]. While these four methods can help to identify many lipids, either by general class, or potentially via their individual fragmentation pattern, discrimination of isomers of lipids is a challenge not easily overcome using shotgun lipidomics [8, 11, 13], which primarily detects the most abundant species. So in the case of isobaric lipids, or stereoisomers, where two or more lipids have the same elemental composition but different connectivity (this is described in further detail in Figure 2.6 in Chapter 2), and regio-isomers, where the location of the double bonds is different between the species, an alternate method of identification is needed, such as chromatography-based MS.

Chromatography-based MS, by contrast with shotgun MS, uses the power of chromatography, such as liquid chromatography, to pre-separate a complex mixture [8, 11, 13]. This method takes advantage of chromatography's ability to separate molecules with high resolving power and good reproducibility. In fact, to study low abundance lipids, this pre-separation is a necessity. Of all the chromatography methods, perhaps the most widely used for this design are liquid chromatography (LC), or high-performance liquid chromatography (HPLC) [8, 13]. There are two methods by which LC and HPLC separate lipids: by normal phase or reverse phase. In normal phase LC, the stationary phase is hydrophilic, meaning that hydrophilic or polar molecules stick to the column. By increasing the polarity of the solvent, analytes are eluted from the column based on their polarity, with more polar substances staying on the column longer. This method is regularly used to separate different classes of lipids due to their polar head groups [8]. On the other hand, reverse phase LC has a hydrophobic stationary phase, where hydrophobic molecules stick to the column. In an opposite fashion to normal phase LC,

solutes are eluted from the column by making the mobile phase less polar. This separates lipids based on how hydrophobic they are, with the less hydrophobic ones being eluted earlier. Thus, in this method lipids are separated based on their fatty-acyl chains [8], in that lipids with longer hydrocarbon chains, and with fewer unsaturations, are more hydrophobic. With both of these methods, the time a certain lipid (or lipid class) stays bound to the column is termed that lipid's retention time (RT). By using the RT as a measure, along with m/z, it is possible to separate and distinguish between isobaric lipids, as this approach utilizes properties other than mass to separate the lipids. Similar to shotgun MS, chromatography-based MS tends to use MS/MS, and the same four methods (product-ion scan, precursor-ion scan, neutral-loss scan and selected-reaction monitoring or multiple-reaction monitoring modes) are used for further separation and identification.

The final output of both shotgun and chromatography-based MS is a spectrum containing peaks representing the amount of material present at each m/z detected in the sample. In the case of chromatography-based MS, the spectrum also tends to display the RT for each peak, showing the separation by both m/z and RT. This spectrum is then analyzed to determine which lipids are represented in the sample.

These advances in MS technology opened the field of lipidomics to more specific research. Using the general approach of extracting lipids from a sample of interest, often by a method based upon the Bligh and Dyer method [8, 13, 18], and then separating this extract using either shotgun or chromatography-based MS to generate spectra that can be analyzed to determine what the lipid composition is, lipidomic researchers are now able to answer many lipidomic questions. Primarily, the field of lipidomics concerns itself with two major, but seemingly simple questions: How many lipid species are there, and

how do they interact with other molecules? [11] These questions, and others that derive from them, have led researchers to discover more about lipids in the context of biological systems, and to discoveries pointing to the roles of lipids in various diseases [9, 10].

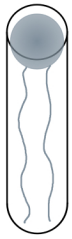
1.2 Why Do We Study Lipids?

Technological advances in MS offered a means by which researchers could gain a detailed understanding of lipids, but it was the discovery that lipids have significant biological functions that formed the driving force bringing researchers back to study lipids, and thence to create the field of lipidomics. Lipids, one of the major macromolecules of the body, comprise between 50% and 60% of the brain's dry weight [12, 19], and make up the plasma and intracellular organelle membranes of every cell in the body. For years, lipids were thought to have only two functions: to serve as energy storage, and to act as the major structural component of membranes [20]. It was not until the discovery by Bergströem, Danielsson and Samuelsson in 1964 that arachidonic acid was the precursor for prostaglandins [20, 21], and that the first biologically active lipids were discovered as platelet activating factors (PAFs) by Benveniste in 1972 [22], later identified as phospholipids by Demopoulos in 1979 [23], that researchers realized lipids were more than just energy storage molecules or structural building blocks, and that they performed biological functions themselves. These discoveries helped re-frame lipids in the minds of researchers, from molecules with limited function to a class of molecules with diverse structures and functions [20].

1.2.1 Lipid Function: Lipids and Membrane Form, Function and Dynamics

We now know that membrane lipids are highly dynamic [12], requiring continuous adjustment to their constituents, remodelled by lipases at the membrane and extracellularly [24], to allow for changes in shape and heightened fluidity, for example allowing for greater membrane curving during endocytosis or exocytosis. The new technical capacity to profile lipid composition at the sub-cellular level, coupled with empirical testing, enabled a new lipid-centric view and understanding of membrane dynamics. Previously, membrane proteins were thought to be the main determinants of membrane structure, but it is now understood that lipids also play a critical role. The differing geometrical structures of the lipid components of a membrane (mostly phospholipids) determine many of the properties of their respective membranes. For example, the shape of a phospholipid within the membrane is determined by the size of the head group and of its hydrophobic tail. In a PC, the size and orientation of the choline head group matches that of the space of the two tails, creating an overall cylindrical shape [2, 12]. PS similarly have a head group that roughly matches the space of the two tails, also creating a cylindrical shape [25]. Because of their cylindrical shapes, when PC molecules aggregate in an aqueous environment, they produce a stable planar monolayer, which will form a bilayer when two planar sheets meet (Figure 1.2). If one of the fatty chains on this PC is hydrolysed, the space the lipid would fill would correspondingly change; if the head-group occupies a larger space than the hydrocarbon tail, as is generally the case with lyso-PCs or with a glycerophosphoinositol tris-phosphate (PIP_3) [25], an inverted conical shape would be created. Inverted conical shapes tend to form positively curved, or convex, monolayers [12, 26]. Conversely, if the

Figure 1.2: Phospholipid space-filling models and the membrane curvature they create. (A) Schematic of the spaces occupied by membrane phospholipids. If the head group and the two chains occupy the same space, the lipid is cylindrical. If the head group is larger than the chains, or one chain is hydrolysed and thus a hydroxyl group, the space occupied is an inverted conical shape. If the fatty acid moiety is larger than the head group, the space occupied is conical. (B) Different membrane shapes generated by diverse phospholipid compositions. If cylindrical lipids aggregate in an aqueous environment, they form a stable planar monolayer with zero membrane curvature, where the head groups solvate with the aqueous environment. If inverse conical lipids aggregate, they form a positively curved, or convex, monolayer. Conversely, when conical lipids aggregate, they tend to form negatively curved, or concave, monolayers.

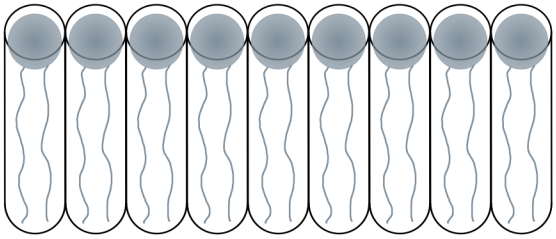


Cylindrical Phospholipid

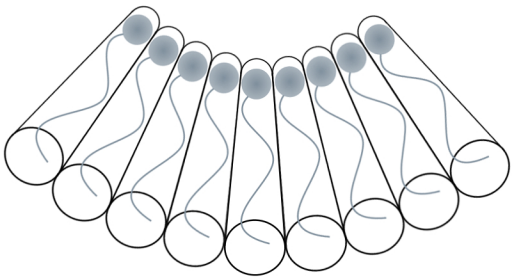
Inverted Conical Phospholipid

Conical Phospholipid

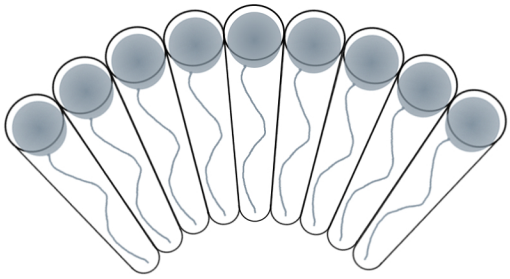
A



Zero membrane curvature
Stable planar



Negative membrane curvature
Conclave



Positive membrane curvature
Convex

B

Figure 1.2

tail's space is larger than that of the head group, for example with PEs due to their reduced head group size and hydration [2], or a glycerophosphate (PA) phospholipid [25], then the lipid would take a conical shape favouring negatively curved, or concave, membranes [2, 12, 26]. These differing structural characteristics of lipids fundamentally alter the plasma membrane's properties and shape, allowing for a multitude of functions. For example, altering the phospholipid composition to generate both positive and negative membrane curvature and the associations between membrane lipids and proteins are both critical steps in clathrin-assisted endocytosis [26], and in exocytosis [12]. The interaction between the actin cytoskeleton and the membrane can also induce curvature [25, 26], often through membrane tension, and this is crucial in the cell's ability to form the structures involved in cell motility, such as filipodia or lamellipodia. Understanding cellular function requires an analysis of the interactions between the membrane, its lipid composition, the proteins that modify membrane lipids, and other structures such as the cell's cytoskeleton, all of which induce both positive and negative curvature. With all these discoveries, lipids not only form the membrane, but their individual properties influence the membrane's shape, enabling, or potentially interfering with, many different cellular processes. To fully explicate these processes requires a lipidomic systems level analysis.

1.2.2 Lipid Function: Lipid-Protein Interactions

While proteins have been shown to modify and influence lipids, as illustrated by the phospholipase family of phospholipid-remodelling proteins [2, 24, 26], or by clathrin in clathrin mediated endocytosis [25, 26], the reverse has also been found: lipids can

influence the activity of proteins. The classic example is with the glycerophosphoinositol phosphates, specifically glycerophosphoinositol 4,5 bis-phosphate ($PI[4,5]P_2$). The pleckstrin homology (PH) domain, or the Phox domain, and others in proteins recognizes and binds to these lipids, anchoring them to the plasma membrane [2, 25] and allowing them to perform their function. This includes a wide range of functions, but which are often involved in a signalling cascade. Localized to the internal leaflet, clusters of PS also form one of the protein docking sites necessary for several key signaling pathways. At least three cascades have been shown to be PS-dependent: (1) the phosphatidylinositol 3-kinase/Akt, (2) the Ras/Raf and (3) the protein kinase C signaling pathways, involved in neuronal survival, neurite growth and synaptogenesis [27].

Lipid-protein interactions go beyond lipid binding domains, however. For example, besides being important in the binding of signalling cascades, PS lipids also modulate properties and functions of membrane receptors. The binding affinity of AMPA receptors is increased by PS in rat telencephalic membranes [27]. PS lipids also affect the function of a variety of other proteins, including tau. PS alters the conformation and antigenic properties of tau, impairing the ability of mitogen-activated protein kinase to phosphorylate tau, and thus tau association with microtubules is decreased [27]. $PI[4,5]P_2$ has also been seen to play a complex regulatory role with the transient receptor potential family of channels; there are a large number of TRP channels where $PI[4,5]P_2$ acts as a necessary co-factor, although there are several cases where $PI[4,5]P_2$ acts as an inhibitor, and can act as both concurrently. These opposing functions are reviewed in detail by Rohacs [28]. For integral membrane proteins, such as γ-secretase, changing any of the properties of the surrounding lipids can alter the activity of

the complex. For example, increasing the fatty-chain length increases γ-activity; changing the position of a double bond (from the 9^{th} position in 18:1 to the 6^{th} position) reduces activity; changing the surrounding lipids from PCs to PSs decreases γ-activity, and changing the surrounding lipids to PI inhibits γ-activity [29]. By direct interaction with proteins, either as anchor points to membranes, or by influencing the protein directly, lipids exert influence on many cellular functions.

1.2.3 Lipid Function: Lipids as Signaling Molecules

Besides influencing membrane form and function, by acting as the scaffolding for signalling cascades, or altering protein function directly, lipids, membrane and otherwise, are also used for the generation of signalling molecules [2]. One class that was previously mentioned, although worth re-stating, is that of PAF lipids. Starting with a PC lipid with an ether-linked, or canonically alkyl linked, *sn*-1 chain and an ester linked, or acyl linked, *sn*-2 chain, when acted on by a phospholipase A_2, such as the calcium-independent phospholipase A_2, results in a *lyso*-PAF. If this lipid molecule is acted upon by a lysophosphatidylcholine acyltransferase, with an acetyl-CoA as a substrate, the acetyl group is then added to the hydroxylated *sn*-2 position, forming the bioactive PAF. This cycle is known as the Lands cycle [30]. PAF is a powerful signaling molecule and is known to stimulate platelets, via the G-protein-coupled PAF receptor, inducing aggregation and secretion of granular constituents [22, 23]. Arachidonic acid is another potentially powerful signaling molecule. It can be produced during the generation of PAF, or in the similar modification of other classes, such as $PI[4,5]P_2$, as long as the fatty acid hydrolysed from the *sn*-2 chain was arachidonic acid. Arachidonic acid forms the

precursor for prostaglandins, which can bind to various G-protein coupled receptors, performing multiple functions throughout the body including regulating inflammation, ensuring embryo implantation and blastocyst growth and development [31].

1.2.4 Lipids and Disease

Lipids are in a state of constant flux. At the membrane, they are in a disordered liquid state [26], able to move almost freely in the plane of their bilayer. Due to the actions of certain proteins, appropriately named flippases, some lipids are able to switch between bilayer leaflets, and they are constantly being remodelled via processes such as the Lands' cycle. Since they perform a variety of functions throughout the body and in vastly different tissue types, lipid levels are tightly regulated both spatially and temporally [2]. In fact, disruption of this regulation is associated with a variety of diseases, such as obesity, atherosclerosis, stroke, hypertension and diabetes, collectively known as the "metabolic syndrome" diseases [2, 9]. Another major class of diseases connected to lipid dysregulation are neurological disorders, such as schizophrenia, Parkinson's disease and Alzheimer's disease [2, 9].

Our current understanding of the vast diversity of structures and functions of lipids was made possible with advances in technology. The large number of cellular and bodily functions that involve lipids is a testament to their diverse potential, yet we are still only scratching the surface. With regards to studying lipids, much to learn we still have, and lipidomics is going to be the framework required to answer these burning questions.

1.3 What Differentiates Lipidomics From Other *–omics* Fields?

While lipidomics is garnering increased attention and lipids are being seen for the structurally and functionally diverse group of molecules that they are, lipidomics is still an emerging field, and with that it is not without its challenges. While the field has similarities to other *–omics* fields, especially sharing a systems approach and in the technology used, there are some key differences that bring about unique challenges for lipidomics, and the implications of these differences are far reaching.

As mentioned previously, the advances in MS technology enabled researchers to identify lipids at the individual species level. While this allowed for further understanding of complex lipid dynamics, it also vastly increased the amount of data generated in lipidomic experiments. Previously, lipid experiments could only distinguish relative amounts of each broad lipid class, yielding tens of data points, researchers are now faced with potentially hundreds of data points per lipid family per experiment. This is further compounded by isobaric lipids: two or more lipid species with the same elemental composition but different structures and thus potentially different properties and effects. Differentiating between isobaric species requires multiple separation techniques, such as the chromatography-based MS methods. Ironically, the vast structural diversity of lipids, which provides lipids with their different properties and thus functions, is the root for the major challenge to identifying, and studying, lipids.

While lipidomics and other fields such as proteomics utilize MS, there are some differences in the techniques. Protein MS techniques to identify proteins digest the protein mixture using proteases and then compare the peptide masses to a large database to identify the proteins within the sample. Alternatively, they may fractionate the

proteins via collision spectra, using both the original peptide mass but also the fragmentation pattern, and then the spectra are compared to comprehensive sequence databases [32]. These strategies rely on using characteristic peaks, representing defining sequence-based signatures, to generate a list of potential protein identities. The fragmentation patterns of lipids, on the other hand, do not produce defining molecular signatures that can be queried in a central database, defining fatty acyl and alkyl chains, for example. Instead they are repeating units of carbon, oxygen, and hydrogen distinguished only by length and degree of unsaturation, making it more challenging to identify each individual lipid. Instead, carbon chain length, degree of unsaturation, linkage, and polar head group identity must be inferred from the structural information inherent in these fragmentation patterns. This complexity is unique to the field of lipidomics, creating a new bioinformatic problem separate from the issues associated with sequence pattern recognition in genomics and proteomics.

Along with the complexities of identifying lipids and distinguishing them from isobaric species, it is not enough to study lipid metabolism in an isolated system [9]. A systems biology approach is necessary to fully understand the interactions between lipids, proteins, and the cellular environment for the biological system as a whole. In order to handle the "medium-scale" lipidomic datasets generated with the more advanced analytical machines, as discussed previously, new techniques are needed [10].

In 2009, Niemelä *et al*. [10] described four analytical challenges facing lipidomics. The first challenge concerns data processing and lipid identification. There is a need for methods and databases that can aid researchers in turning raw instrumental data from experiments into a complete lipidomics dataset that can be analyzed and

interpreted. For example, methods and databases able to convert lists of all the m/z peaks from a MS spectrum into a list of potential lipid identities are required. As mentioned previously, due to the large structural diversity of lipids and their lack of identifying characteristic peaks, this identification is not trivial. Determining which lipids are present is an indispensable first step, without which further experimentation cannot occur. The second challenge lies in statistical data analysis. As described previously, other -*omics* fields, such as proteomics, can use characteristic sequence-based signatures to identify their samples. They can be normalized with respect to other proteins prior to their identification. This is not the case with lipids: lipid extraction efficiency is determined by multiple factors and extracting samples simultaneously does not guarantee that the extraction efficiency will be identical. Another difference is that fragmentation efficiency is different between different individual species, and between subclasses. This means that researchers cannot compare different lipids within the same sample, let alone across samples, and different subclasses cannot be compared; *i.e.,* PSs cannot be directly compared to PCs, and PAF species cannot be normalized against lyso-PAF species, or diacyl species, etc. Individual species can only be compared mathematically to themselves across samples. This is further complicated by the fact that lipidomics datasets tend to generate tens to hundreds of data points per family, per sample, *i.e.*, lipidomics datasets tend to be medium-scale, demanding careful statistical analysis. Here the similarities with other *–omics* datasets, particularly other metabolomics branches, mean that methods and statistical procedures applied in these fields could be used in lipidomics. One such technique is cluster analysis. In this, the main goal is to group samples, or variables, into homogenous groups, allowing researchers to see which lipids

are co-varying, or which lipids separate different sample groups within a study. Once the data have been analyzed, other techniques are required to statistically compare two lipid profiles from different sources, such as different tissues, or species. Regression analysis is one method to compare these profiles. The third challenge is in pathway analyses. While statistical analysis can give insight into how lipids are changing under different conditions, or identifying key lipid species across samples, pathway analysis takes this information and identifies pathways that could be affecting, or affected by, these lipid changes. This pathway analysis can occur by combining data across different fields, such as lipids and proteomics, transcriptomics or genomics. Because of the vast structural diversity of lipids, the number of theoretically possible lipid pathways is extremely large, making representation of all possible pathways impractical. Strategies to isolate and identify which lipid pathways are involved at any given step, and thus the biochemical applications of these lipid changes, need to be developed. And finally, the fourth challenge described by Niemelä *et al* is lipid modeling, specifically in a biophysical context. Lipids are in a tightly controlled homeostasis, due to their essential physiological role. Lipidomics may help to unravel the complexity of the forces that allow maintenance of normal cellular function and phenotypes, but this requires analyzing beyond pathways to include the spatial and temporal context for modeling lipid systems, such as membranes. Simulations of lipid membranes using the composition as measured in lipidomic experiments can lead to new insights, and open new avenues for investigation that may not be readily apparent in more traditional approaches. Modeling approaches which include the spatial and dynamic aspects of lipidomic systems will need to be developed to facilitate these modeling experiments.

The challenges that are facing the lipidomic field, such as those identified by Niemelä *et al* [10], are significant, but not insurmountable. As with other *–omics* fields, lipidomics has the need for new and comprehensive bioinformatic tools to fully mine the data available in this new breed of lipidomic datasets. The pressing need for lipidomic bioinformatics forms the core objective of my thesis.

1.4 Bioinformatics and Lipidomics: Addressing the Lack of Lipidomics-Related Bioinformatics Tools

With the '*–omics* revolution' came an immense surge in the amount of biological data being produced. To handle this increase in data production, computers have become indispensable in biological research. Due to the ease with which computers can handle large quantities of data and model the complex dynamics observed in nature, this is an ideal partnership. This led to the creation and rise of bioinformatics, defined as the application of computational techniques to organize and understand the information associated with biological macromolecules [33]. Bioinformatics centers around the creation and refinement of databases, algorithms, computational and statistical techniques and theory to solve problems arising from the management and analysis of this large amount of data [10]. At its simplest, bioinformatics performs three functions: first, to give a platform for researchers to organize, add to, and access existing data for their experiments; second, to develop tools and resources to aid in the analysis of the experimental data, and third, to use these tools to analyze the data, and then interpret them in a biologically relevant manner [33].

Developments in bioinformatics generally follow technological advances in the *–omics* fields, and thus it is not surprising that the area with the most bioinformatic support is genomics, the first *–omics* discipline addressed by bioinformaticians [10]. Consequently, being a relatively young field, lipidomics does not have nearly as much bioinformatics support as many other fields. This is not to say that lipidomics is entirely devoid of bioinformatic tools to aid researchers, however. Various tools are available, many of which begin to address the challenges laid out by Niemelä *et al.* [10]. One of these tools, dealing with the third challenge, *i.e.*, pathway analysis, is the Kyoto Encyclopedia of Genes and Genomes (KEGG) [10, 34, 35]. This database, initiated in 1995 and continually updated[2], maintains a collection of manually drawn pathway maps from 19 highly integrated databases. It has been used as a reference knowledge base for understanding functions of cellular processes and organism behaviours from large-scale molecular data sets, several of which pertain to lipids [34, 35].

As crucial as pathway analysis is, it is dependent on first identifying the lipids present, without which pathway analysis cannot occur. There is therefore a bottleneck at this step. Naturally, tools have been developed to predict the identities of lipids within a sample. The current strategies with these tools is to use currently available datasets, or to create new reference datasets based on a particular tissue type, then profile those lipids and create a database to which researchers can compare their results. While this strategy can work well, the diversity of lipids means that lipids from one tissue are not necessarily the same as lipids from another. Indeed, lipids from one part of a cell may differ from lipids in other parts of the cell (see Chapter 3). Because of this, lipids may be misclassified, due to differences between the tissue used in the database, and the tissue, or

[2] http://www.genome.jp/kegg/

organism, used in the experiment (see Chapter 2). To address this lack of comprehensive coverage, in this thesis, I created an *in silico* generated lipidomic bioinformatic tool, 'Visualization and phosphoLipid IDentification' (VaLID)[3]. Because of their key roles in the structure of membranes, their numerous biological functions, including metabolic modifications for second-messenger signaling molecules, VaLID focuses on glycerophospholipids. This tool combines a search-engine with a comprehensive glycerophospholipid database, and various visualization features, in a user-friendly package. In creating this tool, I aimed to address the first challenge as identified by Niemelä *et al* [10], lipid identification, and one of the updates to VaLID has opened the way for future development towards comprehensive pathway analyses, addressing their third challenge.

1.5 Objective and Hypothesis

The overall goal of this thesis project was to create and validate a searchable bioinformatic database capable of identifying glycerophospholipids from a given m/z. I hypothesized that by using an *in silico* approach to developing the database, it would be comprehensive, that is it would contain every glycerophospholipid that has been discovered, or that will be discovered. This would aid in the identification of species found in extracted lipid samples, for use with either shotgun or chromatography-based MS. Once this database had been created, the goal was to add features that would make it useful to the field of lipidomic research. I further hypothesized that due to the database being generated *in silico*, it would be comprehensive enough to include all lipids in the

[3] http://www.neurolipidomics.ca

neural datasets used to challenge it, in that it will have predictions for every mass within these datasets. To test these hypotheses, three aims were devised:

1. Create a comprehensive *in silico* database containing glycero-phospholipids, and a computer-based search engine capable of identifying relevant lipids from their m/z, including a visualization feature capable of drawing structural representations of every lipid within the database.
2. Connect this database to existing curated literature sources, by including a feature within VaLID to search these existing literature sources.
3. Validate the computer program by challenging it against the complexities of a real neural dataset, comparing the predicted results identified from the database with those from an existing canonical tool.

Aims 1 and 2, *i.e.*, creating the tool and integrating it with existing literature sources, are presented in Chapter 2, which also addresses the challenge of lipid identification, as described previously by Niemelä *et al* [10]. Chapter 3 describes research addressing Aim 3, the validation of the program, in which a neural dataset was used to see if the program contained predicted identities for all the phospholipids within the complex neural lipidome. The chapter also presents a comparison of the predictive power of the program with LipidMAPS, the current gold-standard canonical database.

Chapter 2: The Development of VaLID – Visualization and phosphoLipid IDentification

2.1 Objectives of This Study

In this chapter, I summarize the research published in two papers Blanchard *et al*. 2013 [36] and McDowell *et al*. 2013 [37], as well as the International Work Conference on Bioinformatics and Biomedical Engineering (IWBBIO) conference proceedings McDowell *et al*. 2014 [38]. The IWBBIO conference proceedings have been further selected to be extended and submitted to a special issue in Genomics and Computational Biology. The overall objective addressed in these papers was to address the need for a comprehensive online search engine capable of facilitating phospholipid identification from mass spectrometry spectra and subsequent visualization of every theoretically possible structure. Here, I combine the two published papers and the published conference proceedings as Chapter 2 to describe creation of VaLID focusing on only my contributions to the program, including the original glycerophospholipid search engine, database, drawing program, and updates made over a 2.5 yr. period. The three papers can be found in Appendix 1.

2.2 Statement of Author Contributions

In VaLID v1.0, described in Blanchard *et al*. 2013 [36], Alexandre P. Blanchard (APB) created the underlying phospholipid database, Nico Valenzuela (NV) created the 3D structural representations for the 58 lipids that were curated by the CIHR Training Program in Neurodegenerative Lipidomics (CTPNL) group, and NV and Dr. Martin

Bertrand created the webpages for the 3D structural models for VaLID v1.0. Sarah Gelbard was the webmaster for VaLID v1.0. She designed the webpage and gave design input on VaLID's GUI. Dr. S Fai's (SF) servers host VaLID and his research team ensure uninterrupted online access. I developed the search and standard 2D visualization algorithms, helped to design the GUI, and coded all of the software, including the search engine, visualization engine, and the code creating the GUI. Dr. SAL Bennett (SALB) conceived the project and wrote the papers with APB and myself. Drs. Hongbin Xu (HX), D Figeys, GW Slater, SF and SALB consulted as scientific experts on the project. In VaLID 2.0[37] [37], APB updated the database by creating and adding all of the glycerophosphoinositols (PI)s to it. I (GSVM) linked VaLID's search engine to the new entries within the database and developed the drawing algorithm for the 2D structures. I also updated the website. For VaLID 3.0 [38], I added the total carbon number nomenclature to the program, and linked VaLID's output to the both PubMed and the Human Metabolome Database (HMDB). I also maintain the program online, respond to all user queries, and update the program in response to user feedback. I wish also to thank our beta-testers who evaluated VaLID v1.0 prior to live launch: David Myers (uVanderbilt), Dattatreya Mellacheruvu (uMichigan), Avinash Shanmugam (uMichigan), Laura Hamilton (Université de Montréal), Matthew Granger (uOttawa), Graham Mazereeuw (Sunnybrook) and Deborah Swartz (uToronto), as well as the expert advisors who gave critical comments to improve the program: Dr. Alex Brown (uVanderbilt), Dr. Theodore Perkins (uOttawa), Dr. Leigh-Anne Swayne (uVictoria), Dr. Shawn Whitehead (uWestern), Dr. Jeff Smith (Carleton), Matthew Cooke (uOttawa) and Marc Léonard (CIMS).

2.3 Introduction

As discussed in Chapter 1, being a relatively new *–omics* field [2], lipidomics lacks the bioinformatic tools necessary to mine the medium to large-scale lipid datasets typical of the field. This bottleneck in data analysis must be addressed to answer key lipidomic questions [10]. Because previously established tools did not completely cover the phospholipidome, we created a new tool, VaLID for online mass spectrometry spectral peak identification.

2.3.1 VaLID in Context: Previous Databases and Neural Datasets

Prior to VaLID's creation, two main tools were available to lipid researchers to assist in identifying the lipids present in their samples based on m/z ratios and parameters of specific mass spectrometry methodologies. For example, database programs created by the LipidMAPS consortium (http://www.lipidmaps.org), or LipidBank (http://lipidbank.jp) are powerful tools that have led the field in terms of standardizing ontologies, notations, and protocols [4, 39]. In the case of LipidMAPS, their databases were created using lipids from literature sources, as well as identified by their team members from specific cell lines [40]. While these sources allowed for a large coverage (~37,000 lipid species, as of December 2014 [40]), I found that many of the lipids (i.e., peaks with unique m/z) I was detecting in my neural datasets (see Chapter 3) were not represented within these databases. Thus, I initiated development of VaLID [36-38].

2.4 Program Description

2.4.1 Programming Language, External Libraries and Programs Used to Create VaLID

From conception of the program, I decided that VaLID should be freely accessible to all researchers, thus available online and cross-platform. Because of these specifications, I choose Oracle's Java™ programming language, as it is a cross-platform language and can be easily integrated into web pages as a web applet, making it ideal for this purpose. When development began on VaLID, Java was in its 6th version. Java 7 was released during development, and the coding was migrated to the newer version of Java. I used the Integrated Development Environment (IDE) Eclipse Kepler as the environment to do the coding. One of the benefits of using Java as the coding language is the wide variety of third party libraries available for use. To give VaLID the functionality it required, I used the JExcelApi libraries, enabling the program to access the glycerophospholipid database, and I used ChemAxon's Marvin libraries, specifically MarvinView, to visualize 2D structural diagrams of the lipids.

2.4.2 VaLID: Structure and Layout

I created VaLID in four parts (Figure 2.1): (1) the graphical user interface (GUI), (2) the underlying glycerophospholipid database, (3) the search function, and (4) the visualization features. The methods responsible for the GUI, search and visualization features work together to produce VaLID's functionality (Figure 2.2). In creating the GUI, I wanted to have a simple design, wherein all of the prompts for the user are clearly

Figure 2.1: A graphical representation of VaLID's modular composition. VaLID comprises of four parts. Three parts, the Phospholipid Database, represented in green, the Search Features, represented in purple and the Drawing features, represented in orange, all make up the functionality of the program, while the fourth part, the Graphical User Interface (GUI), represented in blue, makes up what the user sees while using the program. The code for VaLID was written to ensure seamless connections among these four modules.

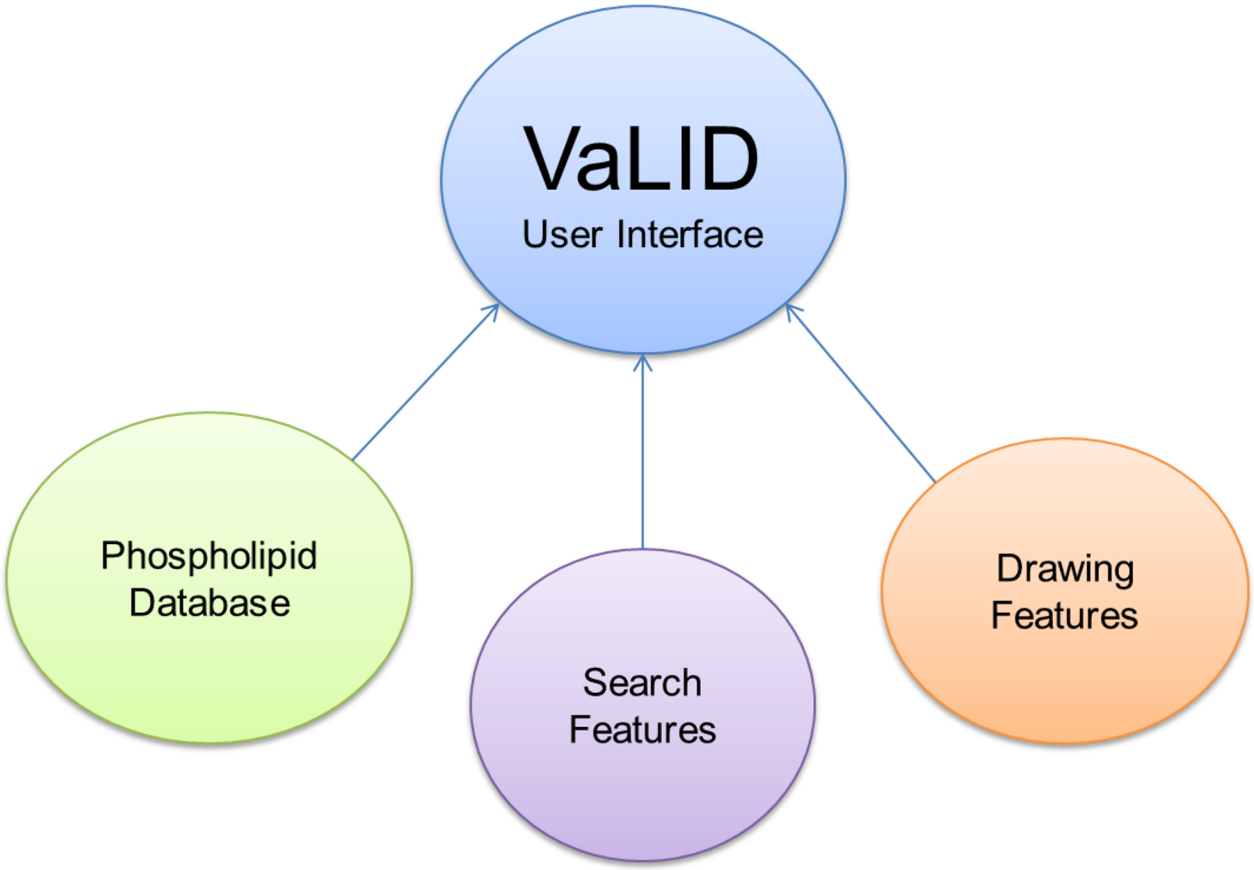

VaLID
User Interface
Phospholipid Database
Search Features
Drawing Features

Figure 2.1

visible and all on one screen. Similarly, I also wanted to have the search results appear in an obvious place, where the user can see as many results as possible, in addition to the search parameters they used. These parameters directed the interface design when VaLID 1.0 was developed (Figure 2.3A) [36].

I initially designed the interface to have two parts: (1) the left side of the screen, where the search parameters and their prompts are located, and (2) the right side of the screen, where the search results are located (Figure 2.3A). The search parameters available to users included an option for either searching by exact or average mass, the ionic mass as dictated by the mass spectrometer; whether to restrict the search to even chains, odd chains or both; a tolerance range around the mass being searched; what subclass to search for within the database; what fatty-chain linkages should be searched for, and what ion mode the mass spectrometer was in for their particular sample. The search results field consists of two panels, labelled "Possible Lipids Include…" and "Possible Isomeric Lipids Include…". When the program finished searching the database, it populated these two panes with the search results. Lipids that appear in the right-hand screen – the *Possible Isomeric Lipids Include* pane – are lipids where the *sn*-1 and *sn*-2 chains of the corresponding lipid in the left hand pane were reversed, where theoretically possible due to the lipid linkage.

As more features were added to VaLID, the GUI was updated to reflect the changes made. However, with these changes, the basic design of the interface stayed the same. By the third, and most current, version, the overall layout of the interface is comparable to v1.0 (Figure 2.3B) but reflects the increased options. More lipid subclasses were

Figure 2.2: The interface, search, and visualization features of VaLID v1.0. A schematic overview of VaLID's functionality at time of launch is depicted. The GUI enables users to search for lipids via their m/z. Once a lipid is highlighted, all of the possible structural conformations can be drawn using the "Display All" button on the interface. A lipid whose name appears in red in the results fields is a lipid that has been curated by the CTPNL affiliated laboratories, which has 3D high resolution images available for download.

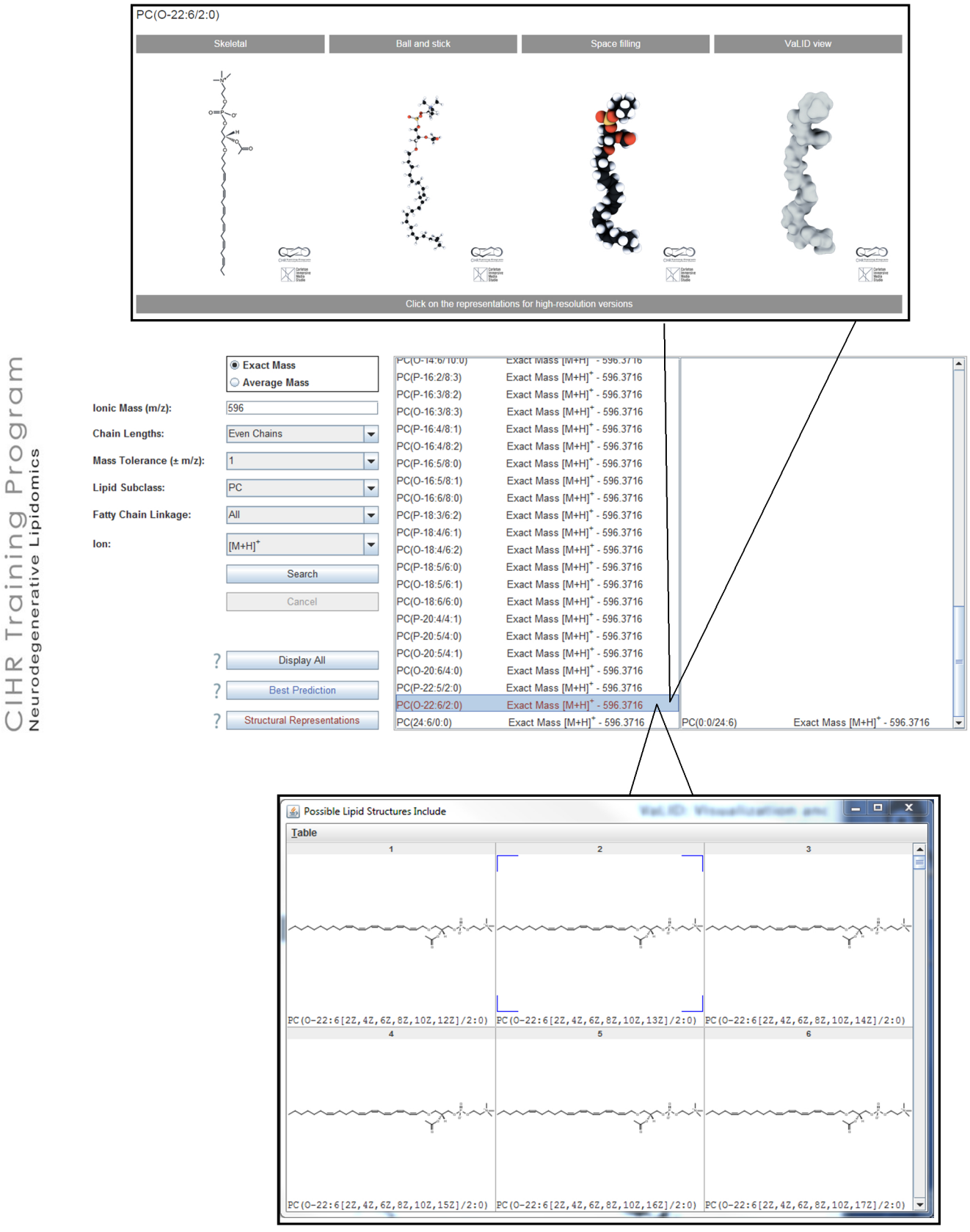

CIHR Training Program
Neurodegenerative Lipidomics
PC(O-22:6/2:0)
Skeletal
Ball and stick
Space filling
VaLID view
Click on the representations for high-resolution versions
Exact Mass
Average Mass
Ionic Mass (m/z):
596
Chain Lengths:
Even Chains
Mass Tolerance (± m/z):
1
Lipid Subclass:
PC
Fatty Chain Linkage:
All
Ion:
[M+H]+
Search
Cancel
Display All
Best Prediction
Structural Representations
PC(P-16:2/8:3) Exact Mass [M+H]+ - 596.3716
PC(P-16:3/8:2) Exact Mass [M+H]+ - 596.3716
PC(O-16:3/8:3) Exact Mass [M+H]+ - 596.3716
PC(P-16:4/8:1) Exact Mass [M+H]+ - 596.3716
PC(O-16:4/8:2) Exact Mass [M+H]+ - 596.3716
PC(P-16:5/8:0) Exact Mass [M+H]+ - 596.3716
PC(O-16:5/8:1) Exact Mass [M+H]+ - 596.3716
PC(O-16:6/8:0) Exact Mass [M+H]+ - 596.3716
PC(P-18:3/6:2) Exact Mass [M+H]+ - 596.3716
PC(P-18:4/6:1) Exact Mass [M+H]+ - 596.3716
PC(O-18:4/6:2) Exact Mass [M+H]+ - 596.3716
PC(P-18:5/6:0) Exact Mass [M+H]+ - 596.3716
PC(O-18:5/6:1) Exact Mass [M+H]+ - 596.3716
PC(O-18:6/6:0) Exact Mass [M+H]+ - 596.3716
PC(P-20:4/4:1) Exact Mass [M+H]+ - 596.3716
PC(P-20:5/4:0) Exact Mass [M+H]+ - 596.3716
PC(O-20:5/4:1) Exact Mass [M+H]+ - 596.3716
PC(O-20:6/4:0) Exact Mass [M+H]+ - 596.3716
PC(P-22:5/2:0) Exact Mass [M+H]+ - 596.3716
PC(O-22:6/2:0) Exact Mass [M+H]+ - 596.3716
PC(24:6/0:0) Exact Mass [M+H]+ - 596.3716
PC(0:0/24:6) Exact Mass [M+H]+ - 596.3716
Possible Lipid Structures Include
Table
1
2
3
PC(O-22:6[2Z,4Z,6Z,8Z,10Z,12Z]/2:0)
PC(O-22:6[2Z,4Z,6Z,8Z,10Z,13Z]/2:0)
PC(O-22:6[2Z,4Z,6Z,8Z,10Z,14Z]/2:0)
4
5
6
PC(O-22:6[2Z,4Z,6Z,8Z,10Z,15Z]/2:0)
PC(O-22:6[2Z,4Z,6Z,8Z,10Z,16Z]/2:0)
PC(O-22:6[2Z,4Z,6Z,8Z,10Z,17Z]/2:0)

Figure 2.2

added to the Lipid Subclass dropdown menu in v2.0 (described in more detail in section 2.4.3, Predicting phosphoinositols: adding the PI subfamily to VaLID's database), and three buttons were added to reflect additional database linkages and data download options in v3.0, the "PubMed Search" button, the "HMDB Search" button, and the "Save Search Selection" button. To increase user friendliness, buttons with functions that require a certain context are disabled until their specific criteria are met. Once met, these buttons are enabled and functional. For example, the "Save Search Selection" button is only enabled, *i.e.*, clickable, when a search has been completed and results have been returned to the screen. Along with the changes on the search parameter side of the screen in v3.0, I added a third panel, underneath the other two, which contains the search results for the "total carbon" results. Predicted carbon chain identities are now returned with all possible isobaric species listed as well as total carbon and total degrees of unsaturation per molecule taking into consideration possible linkages to the glycerol backbone. Selecting an individual species automatically highlights the respective total carbon prediction. Most shotgun lipidomic researchers report only diacyl-PC total carbon lipid identities, lacking a tool to easily predict identities of other linkages. To my knowledge, VaLID v3.2 is the first tool that returns total carbon identity and degree of unsaturation for all possible backbone linkages at a given m/z.

2.4.3 Creating VaLID's Backbone

The Phospholipid Database

The backbone of VaLID is its glycerophospholipid database. This database contains all of the glycerophospholipids searchable within the program. The complete database

Figure 2.3: VaLID's evolving graphical user interface over time. GUI development is depicted. (A) The interface of VaLID 1.0 at launch. The left side of the screen contains all of the search features and options: exact and average mass, ionic mass, chain lengths, mass tolerance, lipid subclass, fatty chain linkage, and ion mode. The search and cancel buttons, space for error messages and a progress bar, help buttons, and the visualization options are also evident. The right hand side of the screen contains the two results panes where lipids and any possible isomeric species were displayed. Displayed are the results of the search for PC with m/z 596, a tolerance of 1, in positive ion mode. (B) The current VaLID interface, VaLID 3.2 (current as of December 2014). The overall layout was kept consistent; however there are two major additions. Firstly, highlighted by the red circle, are the three new buttons. The first, PubMed Search, triggers VaLID to search for citations to the highlighted lipid through the PubMed database. Similarly, the HMDB Search button causes VaLID to search for the highlighted lipid within the HMDB, returning physicochemical properties of the lipid. The third button, the Save Search Selection button, causes VaLID to save the list of lipids to the user's hard-drive as either a text file, or as a comma-separated value file, which can be opened in programs such as Excel. Secondly, highlighted in blue, is the new search-result field. This contains the "total carbon" number for all of the lipids returned in the two previous result fields. This value contains the total number of carbon atoms, and the total number of unsaturations, not separated into individual species as is returned in the right and left panels above. This "total carbon" number represents nomenclature more relevant for shotgun lipidomics profiles.

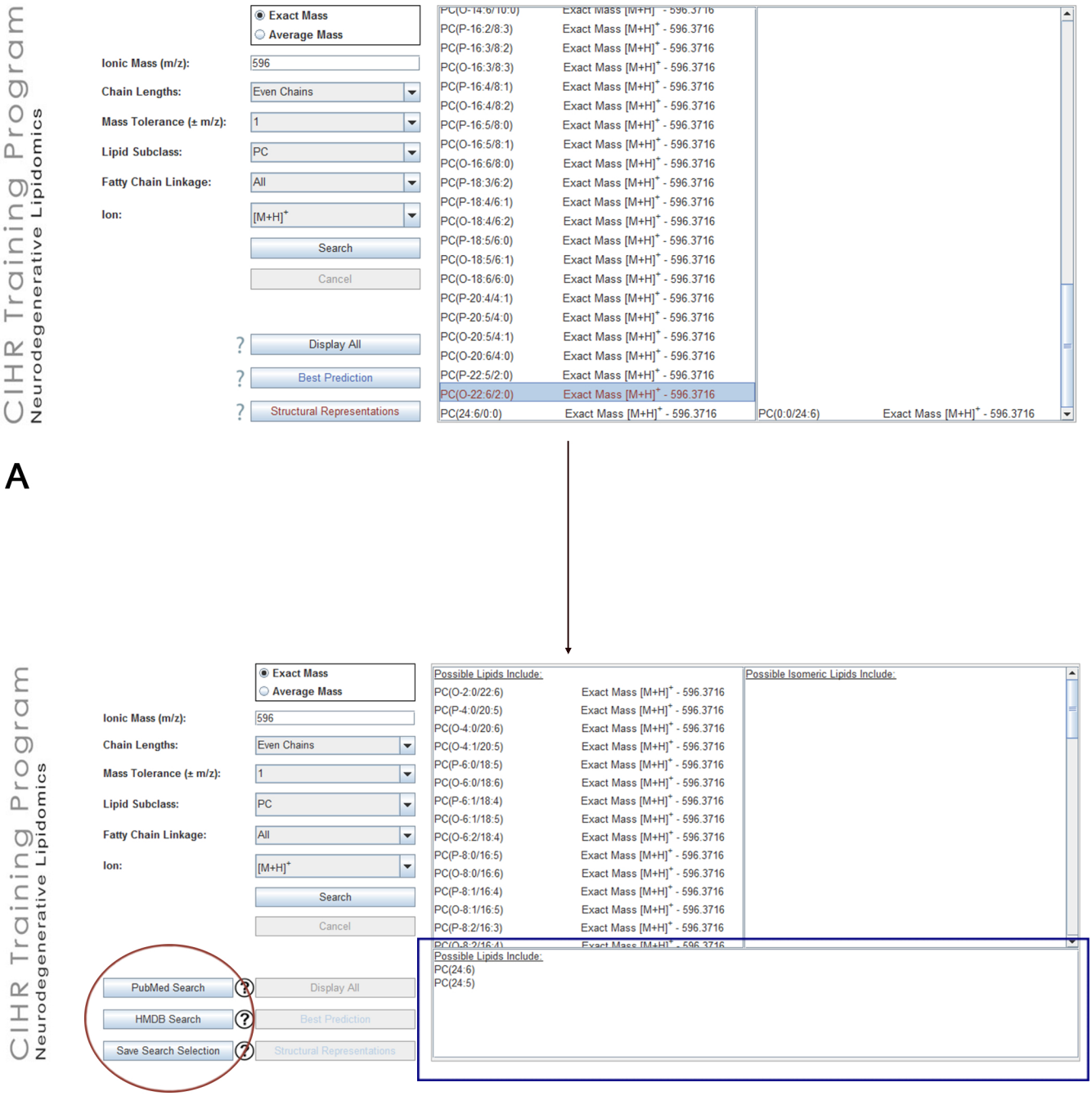

CIHR Training Program
Neurodegenerative Lipidomics
Exact Mass
Average Mass
Ionic Mass (m/z):
596
Chain Lengths:
Even Chains
Mass Tolerance (± m/z):
1
Lipid Subclass:
PC
Fatty Chain Linkage:
All
Ion:
[M+H]+
Search
Cancel
Display All
Best Prediction
Structural Representations
PC(P-16:2/8:3) Exact Mass [M+H]+ - 596.3716
PC(P-16:3/8:2) Exact Mass [M+H]+ - 596.3716
PC(O-16:3/8:3) Exact Mass [M+H]+ - 596.3716
PC(P-16:4/8:1) Exact Mass [M+H]+ - 596.3716
PC(O-16:4/8:2) Exact Mass [M+H]+ - 596.3716
PC(P-16:5/8:0) Exact Mass [M+H]+ - 596.3716
PC(O-16:5/8:1) Exact Mass [M+H]+ - 596.3716
PC(O-16:6/8:0) Exact Mass [M+H]+ - 596.3716
PC(P-18:3/6:2) Exact Mass [M+H]+ - 596.3716
PC(P-18:4/6:1) Exact Mass [M+H]+ - 596.3716
PC(O-18:4/6:2) Exact Mass [M+H]+ - 596.3716
PC(P-18:5/6:0) Exact Mass [M+H]+ - 596.3716
PC(O-18:5/6:1) Exact Mass [M+H]+ - 596.3716
PC(O-18:6/6:0) Exact Mass [M+H]+ - 596.3716
PC(P-20:4/4:1) Exact Mass [M+H]+ - 596.3716
PC(P-20:5/4:0) Exact Mass [M+H]+ - 596.3716
PC(O-20:5/4:1) Exact Mass [M+H]+ - 596.3716
PC(O-20:6/4:0) Exact Mass [M+H]+ - 596.3716
PC(P-22:5/2:0) Exact Mass [M+H]+ - 596.3716
PC(O-22:6/2:0) Exact Mass [M+H]+ - 596.3716
PC(24:6/0:0) Exact Mass [M+H]+ - 596.3716
PC(0:0/24:6) Exact Mass [M+H]+ - 596.3716
A
CIHR Training Program
Neurodegenerative Lipidomics
Exact Mass
Average Mass
Ionic Mass (m/z):
596
Chain Lengths:
Even Chains
Mass Tolerance (± m/z):
1
Lipid Subclass:
PC
Fatty Chain Linkage:
All
Ion:
[M+H]+
Search
Cancel
PubMed Search
HMDB Search
Save Search Selection
Display All
Best Prediction
Structural Representations
Possible Lipids Include:
PC(O-2:0/22:6) Exact Mass [M+H]+ - 596.3716
PC(P-4:0/20:5) Exact Mass [M+H]+ - 596.3716
PC(O-4:0/20:6) Exact Mass [M+H]+ - 596.3716
PC(O-4:1/20:5) Exact Mass [M+H]+ - 596.3716
PC(P-6:0/18:5) Exact Mass [M+H]+ - 596.3716
PC(O-6:0/18:6) Exact Mass [M+H]+ - 596.3716
PC(P-6:1/18:4) Exact Mass [M+H]+ - 596.3716
PC(O-6:1/18:5) Exact Mass [M+H]+ - 596.3716
PC(O-6:2/18:4) Exact Mass [M+H]+ - 596.3716
PC(P-8:0/16:5) Exact Mass [M+H]+ - 596.3716
PC(O-8:0/16:6) Exact Mass [M+H]+ - 596.3716
PC(P-8:1/16:4) Exact Mass [M+H]+ - 596.3716
PC(O-8:1/16:5) Exact Mass [M+H]+ - 596.3716
PC(P-8:2/16:3) Exact Mass [M+H]+ - 596.3716
Possible Isomeric Lipids Include:
Possible Lipids Include:
PC(24:6)
PC(24:5)
B

Figure 2.3

was initially created by my colleague Alexandre P Blanchard, following the method used described in [36]. In summary, the database contains the m/z of all glycerophospholipids in Microsoft Excel files. I then modified this file to have a modular structure (Figure 2.4). The m/z values are divided into smaller databases, one for each lipid subclass. Each one of these sub-databases contains entries for each of the different chain linkages as spreadsheets. Each of these spreadsheets contains all the lipid entries applicable to this chain linkage type. In splitting the database up into these modules, especially at the chain linkage level, the complexity was greatly reduced and thus time spent by the searching algorithm was similarly reduced. Briefly, the mass for each lipid is calculated by virtually fragmenting each species into the glycerophospholipid backbone, containing the phospho-headgroup, and the glycerol backbone with the two hydrocarbon chains. The hydrocarbon chains are permutated from 0 to 30 carbons and from zero to six unsaturations, and the mass calculated. The masses of the chains are then modified to address the effect of linkages; all ester bonds contain an extra oxygen atom, so this mass was added. Each permutation is saved in the database within the corresponding linkage spreadsheet.

Predicting Phosphoinositols: Adding the PI Subfamily to VaLID's Database

We frequently update VaLID, and user feedback is taken into consideration. The best example of this was the effort spent to include the PI subfamily. At time of first release, we did not include PIs given the complexity associated with calculating and visualizing carbon-specific phosphorylation of the PI head group with unique fatty acyl, alkyl, and/or alkenyl *sn*-1 and *sn*-2 chains. Due to user feedback, we increased the

Figure 2.4: The modular nature of the VaLID v1.0 databases. The VaLID database was modular: split into the individual phospholipid subclasses (head groups). Each sub-database was split into separate worksheets, one for each of the different chain linkage types, and all of the possible species with these linkages populate each of those worksheets. The "All Values" sheet contains all the possible species, ranging from 0-30 carbon chains in the *sn*-1 and *sn*-2 position, with up to 6 unsaturations per chain. In the figure, an arrow represents a "populated by" relationship. *I.e.,* a subclass database was 'populated by' worksheets representing each of the different linkage types, which are in turn 'populated by' the individual species appropriate for each linkage.

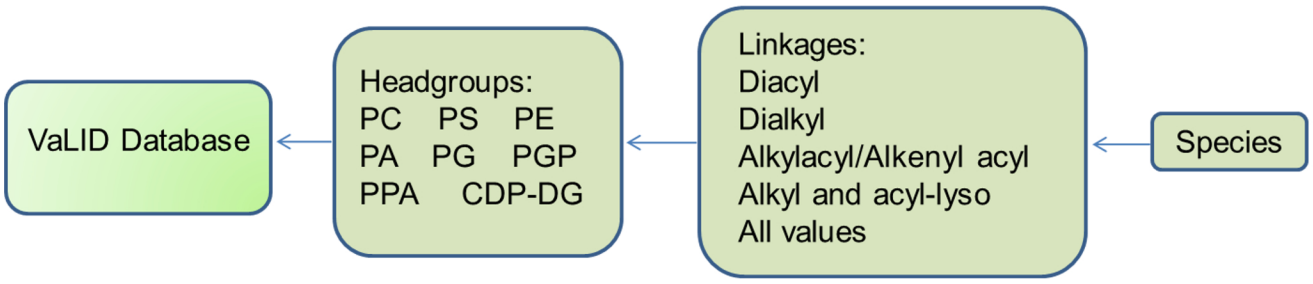

VaLID Database
Headgroups:
PC PS PE
PA PG PGP
PPA CDP-DG
Linkages:
Diacyl
Dialkyl
Alkylacyl/Alkenyl acyl
Alkyl and acyl-lyso
All values
Species

Figure 2.4

database to include the PI subfamily. This effectively doubled the number of species available for users to search [37] (Table 2.1). The same algorithm was used to populate the PI sub-database as was used to populate the rest of the database; I summed the exact mass of the glycerol backbone, the phospho-headgroup, and the masses of the hydrocarbon chains and their linkages, permeating the hydrocarbon chain lengths between 0 and 30 carbons and up to six unsaturations. I repeated this for the average masses and again for the PIs, glycerophosphoinositol mono-phosphates (PIPs), glycerophosphoinositol bis-phosphates (PIP_2s) and PIP_3s.

2.4.4 Searching in VaLID

The search function is the main feature of VaLID as it allows the user to access the database in a user-friendly, helpful, and meaningful way.

Developing a Search Engine: Searching Within the VaLID Databases

The search engine connects the GUI to the database, and comprises the bulk of the methods within the program (Figure 2.5). The algorithm for this segment of the program is multi-faceted. When the user clicks the Search button, the program first collects the specifications provided by the user and ships these specifications into the next method, SearchTask (Figure 2.5). In SearchTask, the program identifies the relevant sub-database and loads it into memory. Once the appropriate database is loaded, the program uses a binary search algorithm to find an instance of the user-requested mass (*Ionic Mass*) in the selected database (Figure 2.5). The ion mode of the mass spectrometer is taken into consideration when locating these masses. Once the program finds the first relevant

Table 2.1: Number of species available for each phospholipid subclass within VaLID.

Phospholipid Subclass	LipidMAPS classification	Abbreviation	Number of species
Glycerophosphates	GP10	PA	92073
Glyceropyrophosphates	GP11	PPA	92073
Glycerophosphocholines	GP01	PC	92073
Glycerophosphoethanolamines	GP02	PE	92073
Glycerophosphoglycerols	GP04	PG	92073
Glycerophosphoglycerolphosphates	GP05	PGP	92073
Glycerophosphoinositols	GP06-GP09	PI, PIP_x	736584
Glycerophosphoserines	GP03	PS	92073
Cytidine 5'-diphosphate glycerols	GP13	CDP-DG	92073
		Total	1473168

value, the program then searches for the first instance of the *Ionic Mass* minus a pre-determined range (i.e., buffer) associated with *Mass Tolerance* using a simple sequential search that includes the first instance of the *Ionic mass* minus the buffer and the *Mass Tolerance* (Figure 2.5). I added the buffer for two reasons: first, if the user has many masses to search around the same inputted mass, this reduces the number of times the program has to re-load the database, and, second, if the user mistyped the mass they were searching for, it is far more forgiving and they can quickly correct their search terms. Once the start and end positions (positions of the first lipid within the buffer and the last lipid within the buffer, respectively) are determined, the program loads all the lipids between these points into an ArrayList (Figure 2.5). In the most recent iteration (VaLID v3.2), when the database is loaded into the array, the program then selectively prints the lipids, according to the specifications of the user – *i.e.*, even or odd chains, or the exact or average mass – to the screen. After the search is complete, the *Save Search Selection* button is enabled. If pressed, a popup appears, prompting the user to save the list of lipids on the search boxes to their desktop (Figure 2.5).

Advancing Search Functions: Integrating PubMed and the HMDB Search Capacity in the VaLID Database

In its third iteration, I sought to connect VaLID to external curated databases. Output can now be searched in both PubMed and the HMDB. This achieves two goals: first, it allows users to easily search external databases for peer-reviewed papers mentioning these lipids and for physicochemical information, and secondly, it lays the foundation so that in the future we can perform pathway analysis, linking specific lipids

Figure 2.5: Method map of VaLID. A diagram of VaLID's classes, including the important methods, and how these methods and classes interact. Classes are outlined by coloured boxes. Green boxes indicate classes and methods that are involved in creating or maintaining the phospholipid database. Purple boxes indicate classes and methods involved in searching through the database for specific results. Orange boxes indicate classes and methods involved in the visualization features of VaLID. The blue box indicates the class and methods responsible for the GUI. The flow-through starts in the LipidAppletV32 class, specifically in the doWork() method for searches of the database, and the method display() for the visualization features.

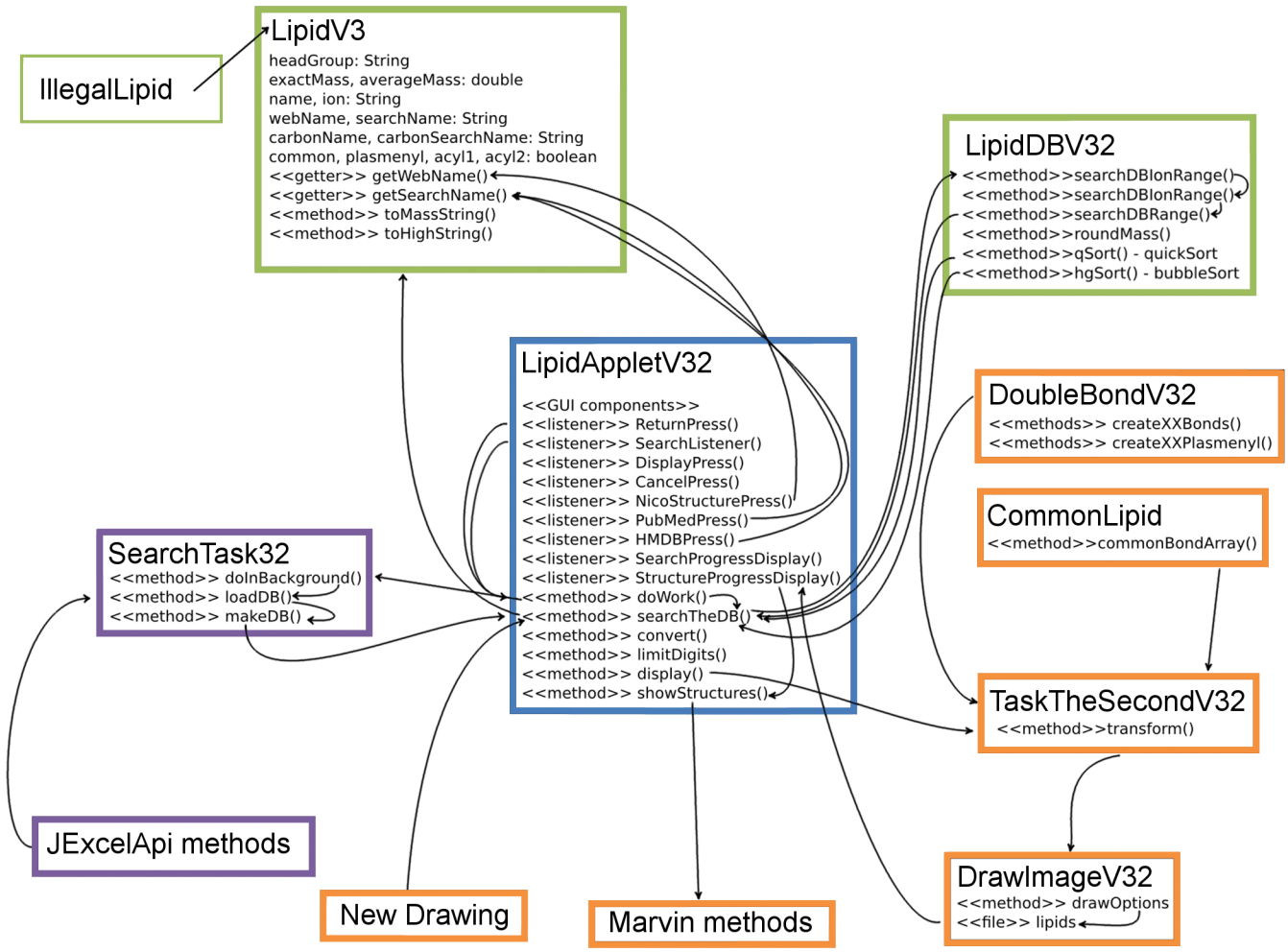

IllegalLipid
LipidV3
headGroup: String
exactMass, averageMass: double
name, ion: String
webName, searchName: String
carbonName, carbonSearchName: String
common, plasmenyl, acyl1, acyl2: boolean
<<getter>> getWebName()
<<getter>> getSearchName()
<<method>> toMassString()
<<method>> toHighString()
LipidDBV32
<<method>>searchDBIonRange()
<<method>>searchDBIonRange()
<<method>>searchDBRange()
<<method>>roundMass()
<<method>>qSort() - quickSort
<<method>>hgSort() - bubbleSort
LipidAppletV32
<<GUI components>>
<<listener>> ReturnPress()
<<listener>> SearchListener()
<<listener>> DisplayPress()
<<listener>> CancelPress()
<<listener>> NicoStructurePress()
<<listener>> PubMedPress()
<<listener>> HMDBPress()
<<listener>> SearchProgressDisplay()
<<listener>> StructureProgressDisplay()
<<method>> doWork()
<<method>> searchTheDB()
<<method>> convert()
<<method>> limitDigits()
<<method>> display()
<<method>> showStructures()
DoubleBondV32
<<methods>> createXXBonds()
<<methods>> createXXPlasmenyl()
CommonLipid
<<method>>commonBondArray()
SearchTask32
<<method>> doInBackground()
<<method>> loadDB()
<<method>> makeDB()
TaskTheSecondV32
<<method>>transform()
JExcelApi methods
New Drawing
Marvin methods
DrawImageV32
<<method>> drawOptions
<<file>> lipids

Figure 2.5

through literature to pathways where they may have a role. To achieve this, I capitalized on the improved search-bar capacity recently added to both of these databases. PubMed's search bar can now accept lipids as search terms – either the common name, such as "dipalmitoylphosphatidylcholine" or the abbreviated name, such as "PC(16:0/16:0)" – and will return instances of these lipids within texts and abstracts of papers. Prior to 2013, PubMed did not recognize "/", "(", ")", or ":" as text names and thus could not search for lipid species using the standardized LipidMAPS nomenclature. HMDB's search feature behaves similarly: by entering in the common or abbreviated name of the lipid it returns data on physical properties, as well as binding partners. Harmonizing search ontology was initially problematic. However, with both of these databases, the search terms proposed by VaLID appear in the URL of the web-page. For example, searching the HMDB for PC(16:0/16:0) returns the URL:

"http://www.hmdb.ca/unearth/q?utf8=✓&query=PC%2816%3A0%2F16%3A0%29&search_type=metabolites&button=".

Changing the "16%3A0%2F16%3A0" to "18%3A1%2F18%3A0" yielded search results for PC(18:1/18:0). This proved that the URL dictates the search results, meaning that if the user modifies the lipid name to be in the same format required by the search function of the external database and injects this text into the URL of these websites, in real time, then the user can query the database for any lipid desired. Because there were some characters (e.g., the check mark) that could not be implemented easily in Java, a streamlined URL was required. Through iterative testing, it was determined that all the

HMDB URL needed was "http://www.hmdb.ca/unearth/q?utf8=&query=" followed by the lipid name and "&searcher=metabolites". Thus, in the *Lipid* class – which defines the Lipid object – there is a property (searchName) which is simply the lipid name, modified to fit the search terms: specifically, replacing special characters with the web-friendly version of their names. For example, "(" becomes "%28", ")" becomes "%29", ":" becomes "%3A" and "/" becomes "%2F" etc. Finally, because *sn*-1 and *sn*-2 carbon chain identities dictate biological function, this input is list-specific. That is, if the user selects their lipid from the left results box, VaLID would search HMDB for the lipid with the *sn* positions as written. If the user selects the right results box, VaLID would search with the alternate positioning. Finally, if the user selects the equivalent lipid using the total carbons option panel, VaLID would search for the total carbon name. Since the lipids in the lipid lists are stored as separate objects, being able to select one list versus the other was relatively simple to achieve. Once the user selects one of the lipids from one of the lipid lists, the program transforms the name into the appropriate 'search friendly name', adds that name to the end of the URL, and open a new web-page with the search results.

Both PubMed and HMDB use very similar patterns for their searches, so the same algorithm is compatible with both external databases. The main difference, of course, was the base URL. For the PubMed search, the URL that VaLID adds the lipid search-name to is "http://www.ncbi.nlm.nih.gov/pubmed/?term=". The rest of the algorithm, including the method for changing the names, remained consistent.

2.4.5 Visualizing the Lipidome: Drawing Features Available Within VaLID

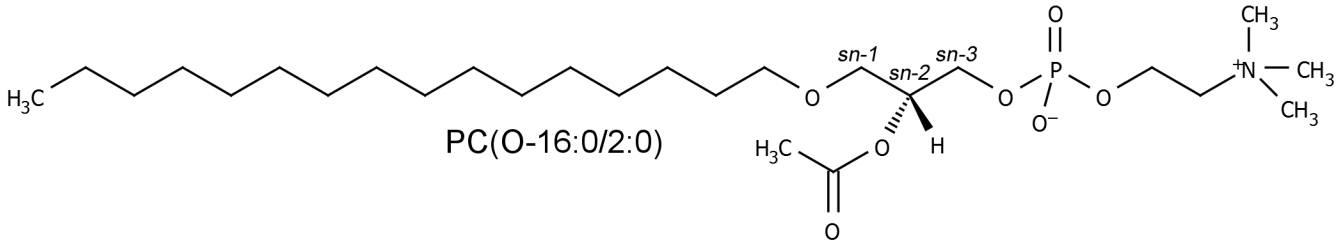
The final part of VaLID, the visualization or drawing feature of VaLID, rounds out the program, and makes it unique in its ability to visualize all lipids within the database.

Original Drawing Techniques of VaLID 1.0

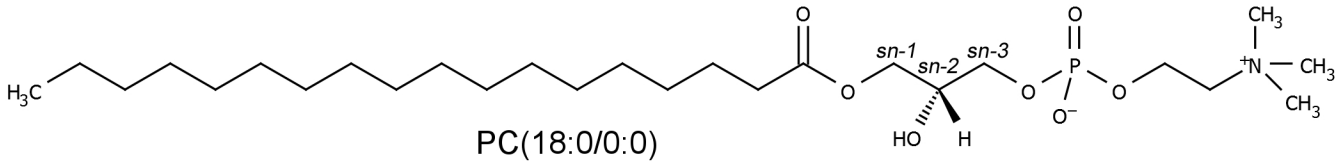
I realized that, since the spirit of the program is to provide the user with all the possibilities, it would be required to visualize every lipid species within the database. Not wishing to submit a posthumous thesis, I opted for automated drawing, as opposed to creating a database of 736,584 lipid image files (as of version 1.0). An algorithm was required to recapitulate every possible structural combination. For example, if there is a carbon chain of six carbons, with one unsaturation, the algorithm must be capable of computing and graphically displaying all of the possible locations for this unsaturation: between carbon two and three (position 2), between carbon three and four (position 3), etc.

Before the coding could begin, the drawing standards were established. To help standardize lipid naming and drawing conventions, I conformed to the LipidMAPS drawing standards [39]. These standards suggest that the hydrophobic hydrocarbon chains be drawn horizontally towards the left of the page, with the glycerol group depicted horizontally with stereochemistry at the *sn* carbons defined (if known), and the phospho-headgroup to the right (Figure 2.6). With the basic structure decided, the next step was coding these structures into an automated drawing method that could be re-iterated in a relatively simple manner – at least in terms that an iterative algorithm could be created – and that consistently produced structures conforming to the LipidMAPS

Figure 2.6: The standardized structural drawing specifications for phospholipids in VaLID. The drawing specifications, as laid out by LipidMAPS [40] dictate that the glycerol backbone be drawn horizontally, with the *sn*-1 and *sn*-2 positions on the left side of the molecule. The phospholipid head group is drawn at the *sn*-3 position, extending to the right, and the hydrocarbon chains drawn extending to the left from the *sn*-1 and *sn*-2 positions, with additional stereochemistry at the *sn*-2 position, if known [40]. The naming convention is also demonstrated. VaLID uses the short-hand naming system, which consists of three parts: the head group, the fatty chain and linkage information for the *sn*-1 chain, and the fatty chain and linkage information for the *sn*-2 chain. The head group is represented in its acronym form (e.g., PC for phosphocholine), the linkage of the fatty chain is represented as follows: "O-" represents an ether bond, "P-" represents a vinyl-ether bond (not shown), and nothing before the number of atoms in the fatty chain represents an ester bond. Following the linkage information, for both the *sn*-1 and *sn*-2 chains is the number of carbons in the chain, a colon, and finally the number of unsaturations. The *sn*-1 and *sn*-2 chains are separated by a forward slash ('/'), giving the general form of: "HG([O-/P-] A:B/[O-] X:Y)", where A and X represent the number of carbons in the fatty chains in the *sn*-1 and *sn*-2 positions, respectively, and B and Y represent the number of unsaturations in their respective chains. As defined in Chapter 1, the lipids depicted in this figure are isobaric; *i.e.*, they both contain the same elemental composition, but have different structures, in this case a PAF molecule and a *lyso*-PC. Although these lipids have the same composition, and thus the same mass, their different structure confers different function.



PC(O-16:0/2:0)



PC(18:0/0:0)

Figure 2.6

conventions. Multiple methods were tested including one promising method, Simplified Molecular-Input Line Entry System (SMILES) wherein 'human-readable text' can be transformed into potentially complex chemical structures[41]. I created an algorithm using SMILES that could procedurally generate a phospholipid of any chain length, with unsaturations at any position along the structure, thus satisfying the first tenant of our drawing feature. I discovered, however, that this method could not force a consistent structure. Sadly, if I created a structure in ChemDraw, saved it as a SMILES file, then re-opened the saved file in ChemDraw, it produced a different structural representation not consistent with our defined LipidMAPS standardization (Figure 2.7). Depending on the lipid being drawn, the structures could be very different from each other – thus, this SMILES-dependent algorithm lacked the required consistency (Figure 2.7).

A method was needed that consistently reproduced bond connectivity and atom location information. Here, exploiting MDL MOL files proved successful [42]. These files contained everything needed to operate the drawing feature; they were human readable files with a well-defined standard explaining each constituent part [42]. They further contain the atomic information – the type of atom, as well as the location in a 3D Cartesian plane – and bond connectivity information. Moreover, many MOL files could be linked into one MDL SDfile (SDF), allowing for multiple structures to be opened in one file. Therefore, I decided to use SDF files as the file type to save each structure. Next, since the atomic and bond information needed to be hard-coded manually, I next calculated bond distances, and angles on the 2D Cartesian plane to serve as a template to create the automated drawing algorithm. The placement of the head groups required standardization. In many sources, the bond from the phosphate group's oxygen to the

Figure 2.7: SMILES-dependent algorithm lacked the required consistency for automated lipid representation conforming to our LipidMAPS drawing standard guidelines. SMILES define a simple nomenclature for representing complex chemical structures but could not be used to reproducibly generate the same representation. (A) The structure was drawn in ChemDraw, following the LipidMAPS drawing criteria [40], and saved as a SMILES document (B). (C) The result when the SMILES document was opened in ChemDraw is technically correct but representationally different from the initial input.

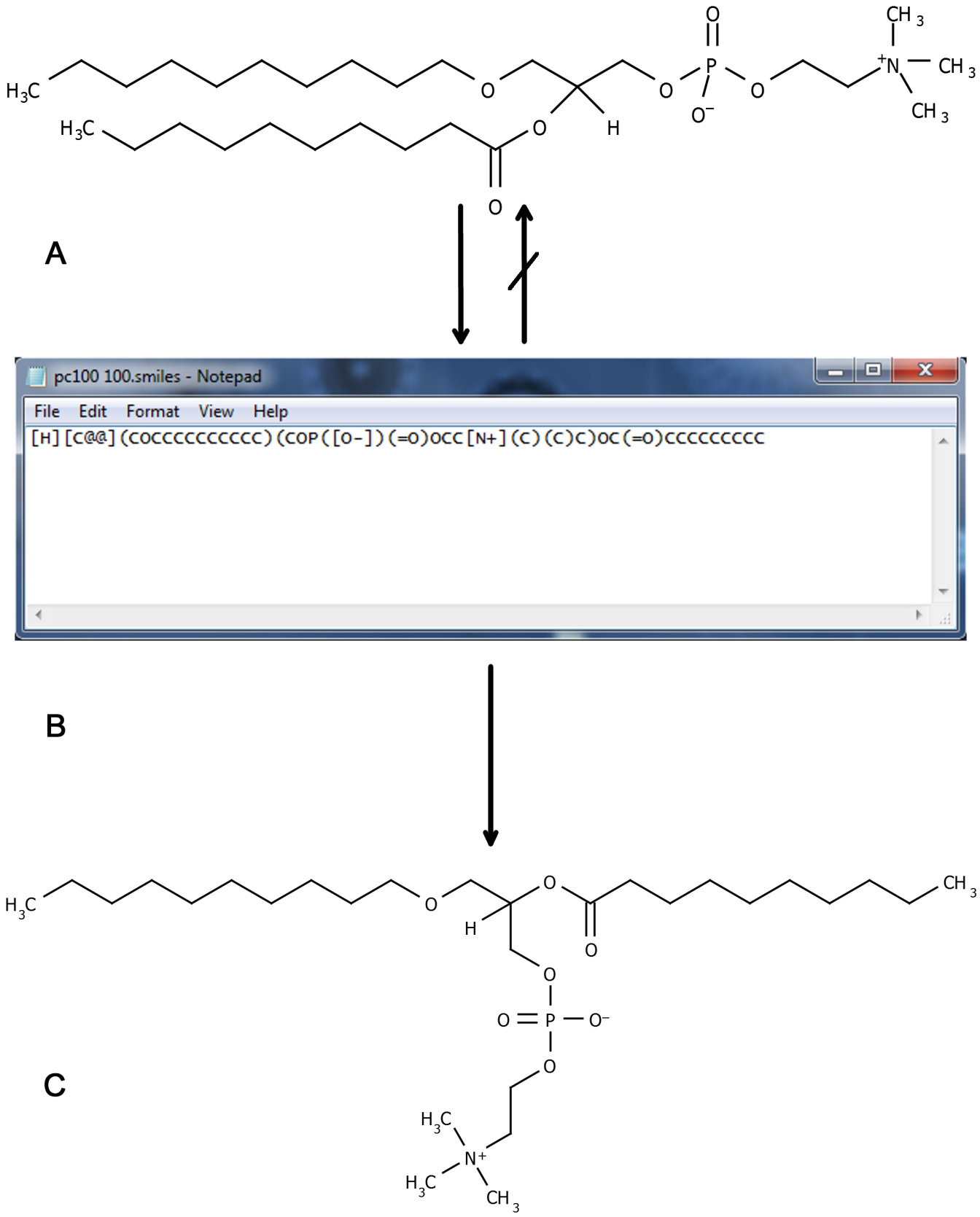

A
pc100 100.smiles - Notepad
File Edit Format View Help
[H][C@@](COCCCCCCCCCC)(COP([O-])(=O)OCC[N+](C)(C)C)OC(=O)CCCCCCCCC
B
C

Figure 2.7

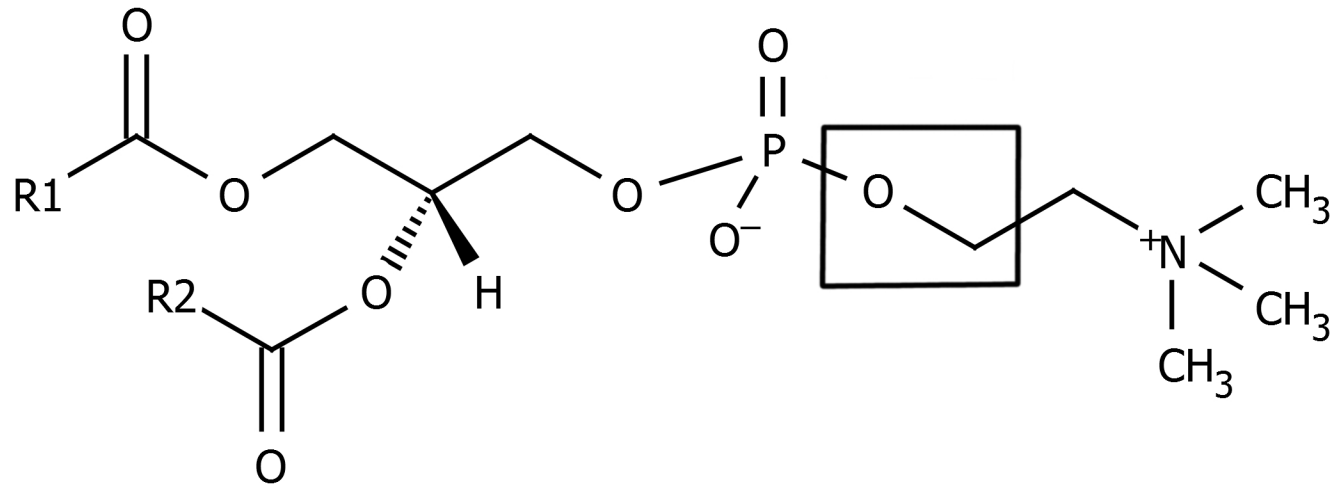
first carbon in the head group was pointing down (Figure 2.8). Due to the nature of the chain leading up to this bond, for VaLID, we flipped this bond to be pointing up (Figure 2.8). Either configuration is possible. Thus, the final algorithm draws the backbones for the *sn*-1 position, the glycerol backbone and the phospho-head group as a single chain, and then the *sn*-2 position, before populating the rest of the atoms in the structure.

To code capacity to draw multiple structures, either simultaneously or sequentially, with the unsaturations in incrementing positions, allowing for all possible combinations to be visualized, a second method was created – "TaskTheSecond" (Figure 2.5). This method operates on a background thread so as to prevent the program from freezing and enables a progress bar to keep track of the time remaining to completion. TaskTheSecond feeds the information into the drawing method, then calls a method designed to permute through the different bond possibilities, and finally saves the results to an array (Figure 2.5). Once all the other specifications (the carbon chain length, head group, etc.) are determined, the drawing method is called for every element in the double bond array, effectively repeating the drawing operation for each bond location. The drawing method also saves the structures to a file, saved in a location specified by the user.

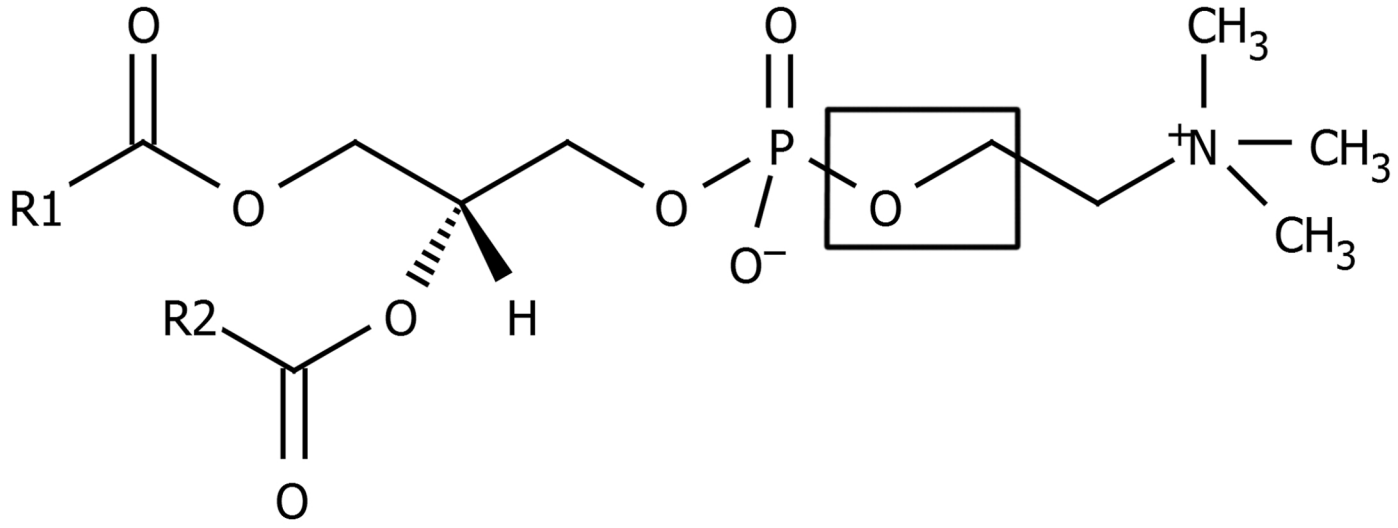
VaLID View for Lipid 3D Structural Representations

To enable future simulations, certain lipids, curated by experimenters within the CTPNL group, have a second type of hard-coded visualization in a static image database. These lipids are returned with a red colour in the program. By selecting one of these curated lipids, and clicking on the "Structural Representations" button, a popup window appears containing four high definition images (Figure 2.9). The four high definition

Figure 2.8: Drawing differences between previous sources and VaLID. Many of the structures found in canonical databases, and in papers, are drawn as appears in panel A, with the bond after the phosphate's oxygen as pointing down (highlighted region). In VaLID, (B) this bond is flipped to point upwards (highlighted region), so as to keep the straight chain continual from the end of the fatty chain to as far along the head group as possible.



A



B

Figure 2.8

images represent different modelling techniques: a high resolution 2D skeletal diagram, a ball and stick model, a space-filling model, and our unique VaLID View model. The ball and stick model, as well as the space-filling model are standard 3D chemical modelling techniques, but the VaLID view model was created by my co-author Nico Valenzuela, and is described in more detail in the first VaLID paper [36]. Briefly, rigid and dynamic models were derived using Maya® nParticles, converted into smooth polygonal meshes. These meshes were directed to the original x, y, and z coordinates and imported as points in space to recapitulate the original molecular structure in an abstracted, organic, form. Resulting VaLID view models are available for download as rigid polygons. They are also available on request fitted with a rig of movable joints between atoms, a process typically used by graphic artists to animate human or animal characters, to facilitate membrane reconstruction and modeling [36].

Expanding VaLID 2.0 to Include Drawing PIs: Unique Challenges with Structures

Adding the PI subfamily posed unique challenges. Similar to the rest of the structures, I needed to first create a template to be able to create the algorithm to draw the structures. Also similar to the other structures, the bond after the phosphate group needed to be rotated from being drawn pointing down to pointing up. This change, while trivial for the previous structures, was non-trivial for the PIs. To change this bond in an inositol, because it is a cyclohexane ring, we would have to perform a "ring/chair flip" [43]. Flipping the ring in this way affects the positioning of substituents on the ring. In published work [40, 44] , the substituents have been drawn in the equatorial position for cyclohexane rings (Figure 2.10A). When rings are flipped (Figure 2.10B), the

Figure 2.9: High definition structural representations available for specific lipids in the database. Lipids that have been curated by the CTPNL group can be visualized in high-definition images in four different styles: (1) 2D skeletal models are available in higher definition than the original drawing function, (2) Ball-and-stick diagrams, and (3) Space-filling models, and (4) a new style, generated using Maya's nParticles for subsequent animation, a method we have termed "VaLID view" [36].

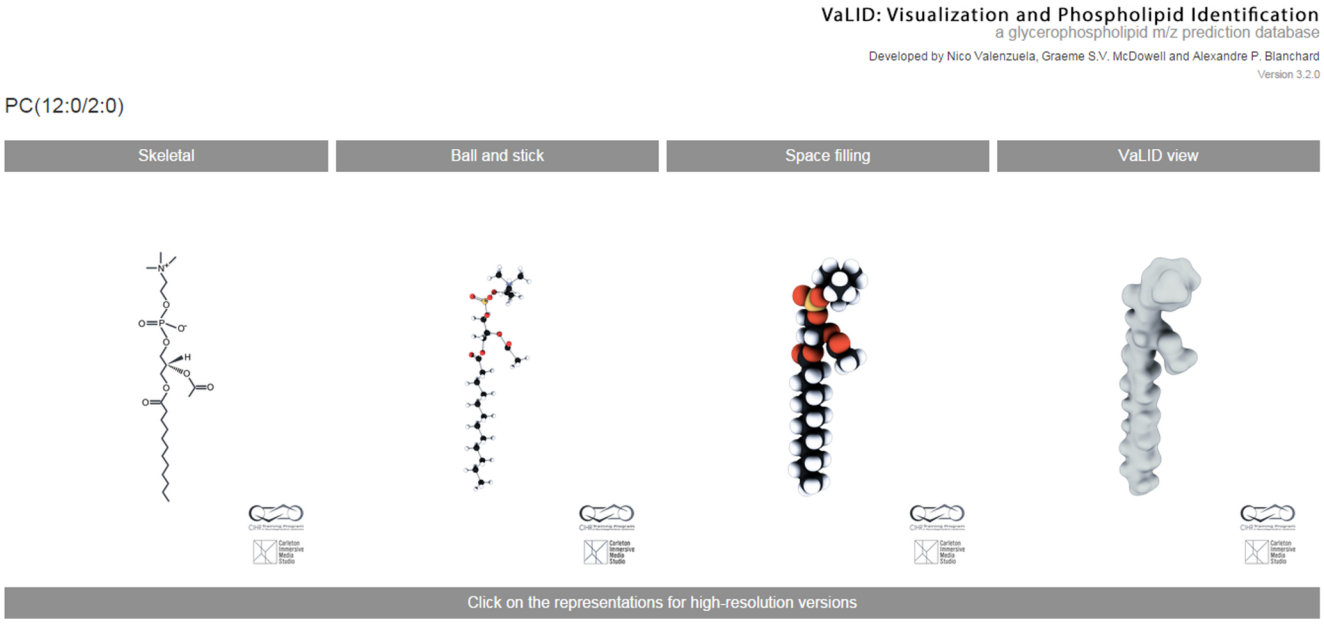
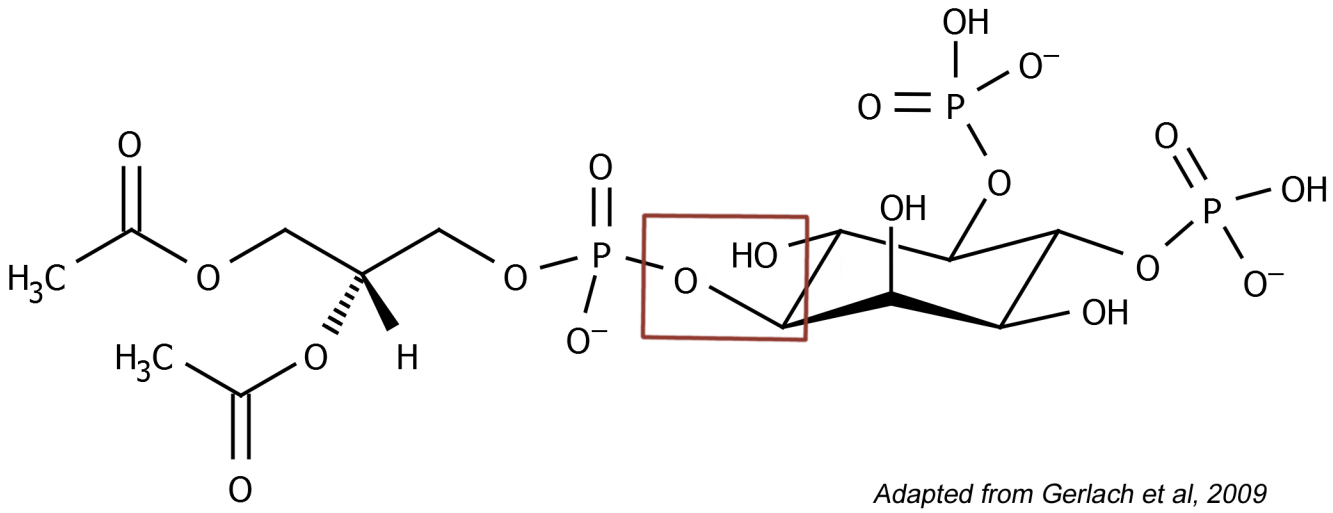

VaLID: Visualization and Phospholipid Identification
a glycerophospholipid m/z prediction database
Developed by Nico Valenzuela, Graeme S.V. McDowell and Alexandre P. Blanchard
Version 3.2.0
PC(12:0/2:0)
Skeletal
Ball and stick
Space filling
VaLID view
Click on the representations for high-resolution versions

Figure 2.9

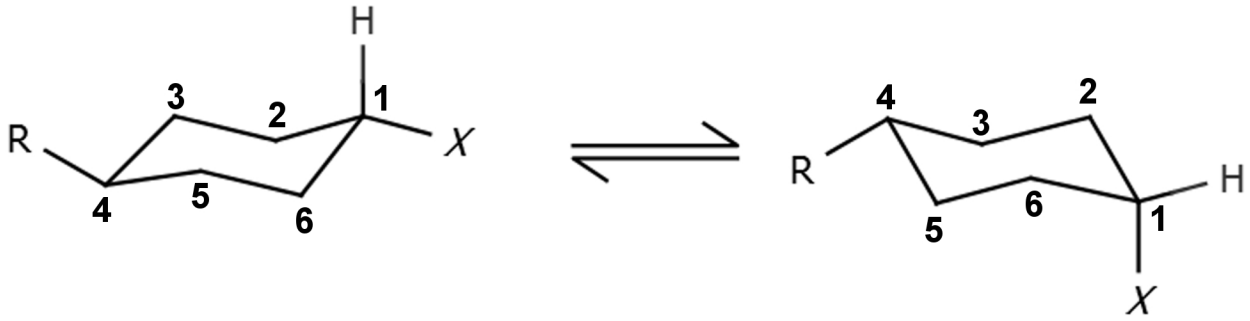
substituents also change orientation, meaning the substituents are drawn in the axial conformation [43] (Figure 2.10C). I took all of these factors into consideration when creating the PI template. I determined the angles and positions for the phosphate groups in the PIP, PIP_2 and PIP_3 variants. Once all templates were created, it was fairly simple to add these structures to the drawing function. In order to reduce file size, and thus make the program faster to load when accessed online, I amalgamated the drawing functions for the PIPs, and PIP_2s. *i.e.*, instead of having an algorithm to draw PI[3]P, and a different one for PI[4]P, or PI[3,4]P_2 and PI[3,5]P_2, I combined all of the algorithms into a single method (drawPIP, or drawPIP2). With this amalgamation, I had to add code that determines which positions have the phosphorylated oxygens and replaces the hydroxyl groups with phosphate groups. To ensure success, I needed the algorithm to recognize the PIP or PIP_2 required by the user. To solve this problem, I added the individual PIP/PIP_2 combinations into the *Lipid Subclass* drop-down menu on the main screen. The drawing algorithm checks this field to determine which positions on the inositol ring to phosphorylate. In the case of the "All", or "All PIP_x", however, no head group information would be returned to the algorithm. To address this challenge, during the database creation stage, another counter is created which keeps track of which phosphorylation position is being used and saves the position to memory. To determine which phosphorylation status to display, this counter is consulted and the proper phosphorylation position is inputted into the drawing function ensuring the proper structure can be drawn.

Figure 2.10: Visualizing phosphatidylinositols in VaLID. (A) A generalized phosphoinositol [4,5] bisphosphate, adapted from previous sources and articles [44]. The phosphate molecules are in equatorial position and the bond following the phosphate oxygen points downwards (A, highlighted region). (B) An overview of the ring-flip for cyclohexane rings. The substituents on carbon 1 are labelled as H and X. In the left panel, H is in the axial position, with X in the equatorial position. After the ring flip, on the right hand side, X on carbon 1 is in the axial position, and H is in the equatorial position. The numbered carbons remain the same. (C) A generalized PI[4,5]P_2 as displayed with VaLID. The phosphate groups are in axial position due to the ring flip. The oxygen bond after the phosphate group points upwards, in accordance with Figure 2.8.



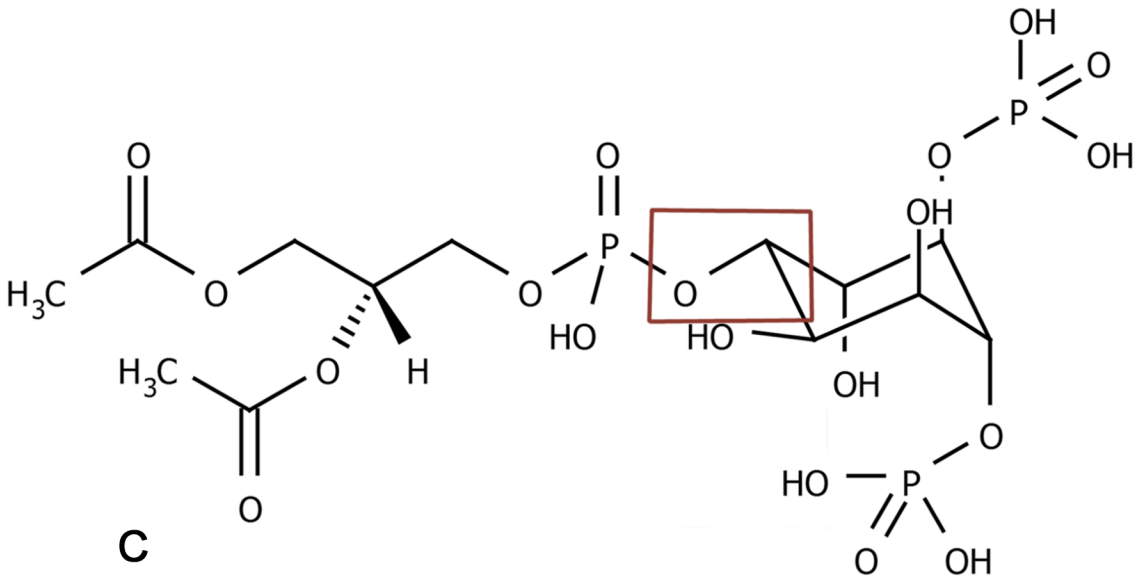
Adapted from Gerlach et al, 2009

A



Adapted from Clayden et al, 2005

B



C

Figure 2.10

2.5 Discussion

The lack of complete glycerophospholipid coverage by previous mass-spectrometry related lipidomic tools sparked the creation of VaLID 1.0, and it was first published in Bioinformatics in 2013 [36]. After receiving user feedback, I further updated the database and visualization features to include PIs, with the update published in 2014 [37] [37]. For IWBBIO 2014 [38], I improved code and linked VaLID to PubMed and the HMDB. When selected for the IWBBIO special issue in Genomics and Computational Biology, I expanded VaLID 3.2 to added new lipid nomenclatures to the search engine useful for researchers using shotgun lipidomic approaches.

2.5.1 Summarizing VaLID 3.2: the Present and Future of the Program

In summary, VaLID 3.2 contains all of the features determined to be necessary for the program to be both useful, and user friendly. The user-friendly user interface clearly lays out all potential search options, with contextual buttons and options greyed out until their respective context occurs. The program uses m/z and mass spectrometer-specific methodologies to search a comprehensive glycerophospholipid database, containing lipids with hydrocarbon chains ranging from 0 carbons to 30 and up to six unsaturations, representing 1,473,168 lipids over nine phospholipid subclasses for the inputted mass, and returns all theoretically possible lipid species and the 'total carbon' number. VaLID is also able to draw all the possible structural combinations of these lipids as 2D representations, and certain specific lipids, curated by the CTPNL group, can be displayed in high resolution 2D and 3D structural representations. Search results and drawings can be saved to desktop in multiple formats. VaLID now further connects to

two external databases, PubMed and the HMDB, providing links to papers or physical properties specifically referencing the lipid in question. To our knowledge, this is the first, and so far only, freely available, web-based lipidomic MS prediction tool combining a comprehensive database, a drawing feature, and linked to curated literature sources.

While this chapter summarizes the current state of the program, this is by no means the end. We intend to update the VaLID database to contain more lipid subclasses, such as glycosylated glycerophospholipids (for example, the glycosylglycerophospholipids [GP14]), and add potential oxidated species. We also plan to add the possibility for trans bonds, include another lipid class, and perhaps further expand to sphingolipids. Currently, identities are suggested by the "best prediction" option based on most prevalent fatty acids in biological tissue [45]. Another potential addition to VaLID could be additional reference markers to help narrow down which lipid amongst the results list is predicted to be most likely the 'correct' lipid in the sample based on spectral properties. For example, for users using a chromatography-MS method, there is pre-separation of the lipids in the sample, adding an extra parameter, RT in this case. If using a reverse-phase liquid chromatography method, lipids are separated by their hydrophobicity, and in some cases, isobaric species can have vastly different RTs, allowing for discrimination between these peaks, and thus, identification. While RT has been difficult to replicate across different chromatographic methods, by using a sufficient number and variety of standards, researchers can mathematically normalize RTs across datasets. Exploiting this could provide a platform for a high-throughput RT based identification strategy. Integrating such a strategy into VaLID, where RT information is taken to rank the predicted identities would provide more precise identifications beyond

expert user knowledge. This is a potential next step for VaLID, and would be beneficial to aid researchers to identifying their species. Also, with the updates to external databases, such as PubMed, creating a feature allowing for pathway analysis – that is, searching for a lipid and determining what metabolites, or pathways, this lipid is involved in, similar to pathway analysis for proteomics will be essential.

2.5.2 Availabilty and Accessibility of VaLID

From the beginning of the project, the overarching goal has been to contribute to advancing the lipidomics field. In that regard, dissemination to all lipid researchers is crucial. Therefore, since its first release [36], VaLID has been web-based, freely available, and open to the public. The current version is available at http://www.neurolipidomics.ca. However, due to our use of external libraries and their corresponding licenses, VaLID is not an open source program, and thus the source code will not be released to the public. We accept and welcome user feedback, and contact information can be found at the program website.

Chapter 3: Validating VaLID

3.1 Objectives of This Phase of the Study

We created VaLID to address the paucity of neural lipid species represented in existing spectral databases. To verify that VaLID filled this gap, I (1) profiled the PC, PS and PE composition of hippocampal synaptic membranes, (2) predicted the identities of these lipids using VaLID, and (3) compared the results to LipidMAPS, the current gold standard. The overarching objective of this research was to establish whether the identity of neural lipids, identified by LC-ESI-MS methodologies but absent from existing databases could be predicted using VaLID.

3.2 Statement of Author Contributions

I am grateful for the assistance I received from my colleagues in Dr. Bennett's laboratory with both methodological guidance and technical assistance. I was mentored by APB on synaptosomal preparations and by Stephanie Fowler (SF) and Mark Akins (MA) with the Western blotting protocol. MA was instrumental in obtaining mouse tissue for the synaptosomal preparations. Graeme Taylor (GT)'s assistance in dissecting the mouse brains was invaluable. HX and Dr. Yun Wang (YW) performed the LC-ESI-MS of my lipid extracts. I analyzed all data.

3.3 Introduction

With the development of any new software, validation is always necessary and VaLID was no exception. We extensively tested the program prior to its original publication, with repeated in-house verification of code for each subsequent version during development and prior to release. We used spectral results to challenge the program; these were theoretical in that they were randomly generated. Moreover, some of the test values were targeted, simply to check for errors in the code. To 'validate VaLID', I required a neural dataset of sufficient biological complexity. Since one of the defining justifications for VaLID was the lack of comprehensive coverage in existing databases for neural datasets, I profiled synaptic membranes, representative of the neuronal chemical synapse, to assess whether VaLID was a more effective tool compared to the gold-standard database, LipidMAPS [40].

3.3.1 The Diversity of Neural Tissue and Their Lipids

Neurons are unique in their function and physiology. Unlike other cell types, their membranes can transverse considerable distances. The axon of a single neuron can reach over a metre in length extending, for example, from cell soma resident in the motor cortex of cerebrum to lower motor neurons in the distal reaches of the spinal cord [46]. They communicate via electrical and chemical signals, are highly compartmentalized, and have unique cell-cell structures at the junction between axons and dendrites called synapses [46]. Synapses are the sites of chemical communication between neurons and other cells; they contain a broad range of chemical neurotransmitters, in addition to localized machinery required for intracellular signalling, including receptors for the

neurotransmitters, as well as exocytosis and endocytosis machinery for neurotransmitter release and reuptake [47]. Synapses are composed of three main components: the presynaptic component – the axon terminal; the postsynaptic component – often a dendritic spine, but could also be neuronal soma, muscle cells, or other potential targets [46]; and a synaptic cleft [47]. Synaptic membranes are also asymmetrical. The pre-synaptic density is what contains neurotransmitter-filled synaptic vesicles and all the material required for both exocytosis and endocytosis; the presynaptic axon terminal (bouton) also contains membrane-bounded organelles such as mitochondria. The neuronal membrane at the bouton is decorated by SNARE protein complexes, required for vesicle docking, ion channels and other protein constituents [47]. When the bouton's membrane depolarizes due to an action potential (or influx of calcium), synaptic vesicles exocytose their contents (neurotransmitters) into the synaptic cleft [47]. Apposing membranes, the post synaptic membrane, have very different compositions. They are enriched in receptors specific to the different neurotransmitters found in the presynaptic vesicles, such as NMDA or AMPA receptors, as two examples, for glutamate, or GABA receptors for GABA, and, at the dendritic spine, can be recognized by a collection of dense lipid bilayers which form complexes with receptors, cytoskeletal scaffolding and other granular/filamentous material (the post-synaptic density (PSD)) [24, 47] visible by electron microscopy on the cytoplasmic surface [47, 48]. The post-synaptic cell receives input from the presynaptic cell, or more accurately multiple pre-synaptic cells at multiple synapses, and integrates excitatory and inhibitory transmissions across all the synapses to dictate membrane depolarization/hyperpolarization. If there are more excitatory signals, for example more excitatory glutamine receptors active than inhibitory GABA receptors

across synapses, the membrane begins to depolarize. A neuron's resting potential is approximately -70mV. With each influx of excitatory signals, the membrane potential becomes slightly more positive, until a threshold at approximately -55mV is reached, at which point voltage-gated sodium channels open, resulting in an influx of positively charged sodium to flood into the cell, causing a depolarization, producing an action potential [46], which propagates along the axon and results in fusion of synaptic vesicles at all presynaptic terminals and chemical transmission to the next cell [46].

Neuronal synaptic membranes are the site of neurotransmitter release and uptake during synaptic transmission [47], and as such their plasma membranes are highly dynamic. During transmission, vesicles constantly fuse to, and bud from, the synaptic membrane. This stresses the importance of the constant flux of the synaptic membrane and its lipid composition is tightly regulated [24]. Phospholipids, the major form of lipids at the membrane [24], are the major component of plasma membranes. Structural phospholipids, which contain hydrocarbon chains at both of their *sn*-1 and *sn*-2 positions, tend to occupy a cylindrical geometry [24], and when many aggregate together, they form a flat or planar membrane and are unsuited for forming small vesicular-like membranes [24, 46], as the extreme curvature cannot be accommodated. In order for synaptic vesicles to bud from the membrane, or for a vesicle to fuse to the membrane, phospholipids at the site must be modified so as to be able to conform to the membrane curvature required. Phospholipases are enzymes which hydrolyse substituents at one of the *sn* positions in phospholipids. Should a phospholipase A_1 or A_2 act on a phospholipid, the hydrocarbon chain at the *sn*-1 or *sn*-2 position, respectively, is hydrolysed, resulting in a more wedge-shaped, single-tailed *lyso*-phospholipid, capable of

adopting a greater curvature [46]. This state of flux is extremely important to synaptic function, and requires careful regulation; disruption of this regulation has been linked to different neurological diseases, such as Alzheimer's disease (AD) [24]. Using advances in lipidomic technologies, researchers have now shown that individual lipid species have different properties and different potential functions, and loss of specific species has been linked to different disease cases. For example, the loss of docosahexaenoic acid during AD has been linked to deformation of dendritic spines [24]. This highly dynamic system, coupled with the discovery that specific lipid species play important roles, point to a unique, and complex, neural lipidomic dataset. Bioinformatic tools were created to help researchers identify the lipids present in datasets, but neural lipidomic datasets, such as ones previously described by our group [49, 50], have identified species not represented in the current canonical databases. After creating VaLID to address this gap in the lipidomic bioinformatic toolkit, to validate the program, I decided to use a complicated neural lipidomic dataset to fully challenge the system. Therefore, I chose the synaptic lipidome, in the form of synaptosomes, for this task.

3.3.2 Using Synaptosomes to Verify That VaLID Can Predict Identity of Novel Neural Phospholipids

In order to verify that VaLID can predict the identities of complex neural phospholipids, I isolated synaptosomes. The term "synaptosome" was first used by Whittaker in 1964 [47, 51, 52]. Originally created to study receptor-ligand interactions involving acetylcholine (ACh), synaptosomes were separated using differential centrifugation followed by a sucrose density gradient [51, 52]. Here, the application of

mild mechanical shearing in a homogenizer releases membranes at the synaptic cleft [47, 52]. The process is not cerebrum-specific. Synaptosomes have been successfully collected from spinal cord, retina, and at the neuromuscular junction of muscle cells [52]. Synaptosomes, themselves, are artificial structures that contain all of the unique synaptic membrane domains described above [47]. As such, they have been referred to as "pinched-off nerve endings" as they are formed when the lipid bilayers of multiple synaptic domains, including the membranes of astrocytic processes infiltrating the synaptic cleft, reseal together after the axon terminals are torn from the cell soma and proximal axonal membranes by homogenization [47]. Thus, synaptosomes contain the complete presynaptic terminal, including mitochondria, synaptic vesicles and synaptic transmission machinery, as well as the postsynaptic membrane and PSD [47, 52]. Gradient fractionation, using either sucrose, Percoll, or sucrose/Ficoll, removes contaminating microsomal and nuclear membranes, allowing synaptic-specific lipid compositions to be harvested [47, 52].

Synaptosomes also represent simple model synapses; they can respire; they take up oxygen; they can synthesize ATP, and many amino acids, neurotransmitters and transmitter precursors are also taken up by carrier-mediated mechanisms, so synaptosomes have been used to study the uptake of metabolites and drugs into synapses and synaptic vesicles [52, 53]. Synaptosomes also release neurotransmitters when depolarized, which requires exocytosis of the transmitter-storing synaptic vesicles, which can be stimulated by electrical stimulation [52]. Taken together, all of this makes synaptosomes an ideal choice for studying the complex nature of synapses and the processes occurring at them. Since we can infer that the membrane altering properties of

synapses exists in synaptosomes, as exocytosis of neurotransmitters from within synaptic vesicles occurs, this would be an ideal model to create a lipidomic dataset to challenge VaLID.

3.4 Materials and Methods

3.4.1 Animals

All experimental protocols were approved by the Animal Care Committee of the University of Ottawa according to guidelines set forth by the Canadian Council on Animal Care. Mouse hippocampal tissue for lipidomic analysis was dissected either from mice obtained on a mixed C57BL/6J/129/FVB background, back-bread for five generations (N5) onto a C57BL/6J lineage in the Bennett laboratory, or mice obtained on a mixed C57BL/6JxC3H background, similarly back-bred to N5 on a C57BL/6J lineage. Mouse left cerebral tissue for protein analysis was dissected from pure C57BL/6J mice back-bred to N15. Mice were kept on a 12 h light-dark cycle with food and water available *ad libitum*. Mice were euthanized by injection with lethal injection of euthanyl (65 mg/mL sodium pentobarbital, Bimeda-MTC Animal Health Ins, Cambridge ON). The appropriate brain region was dissected in ice-cold homogenization buffer (HB, 320 mM sucrose (BioShop, Burlington ON, cat# SUC507), 5mM sodium HEPES (Sigma-Aldrich, St-Louis, MO cat# H3784) and taken for synaptosome preparation. Tissue not used for synaptosome preparation was flash-frozen in liquid nitrogen for future analysis.

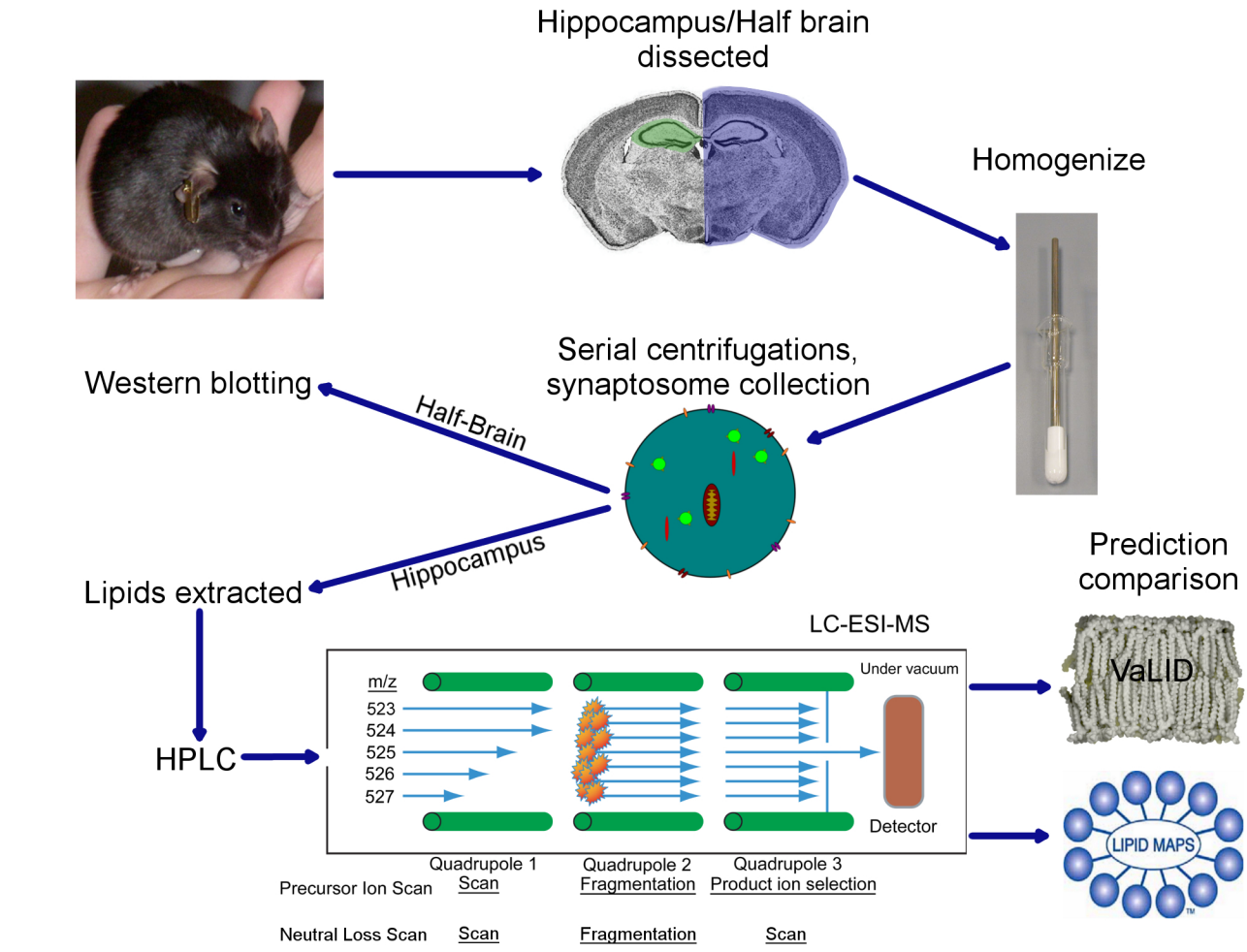
3.4.2 Synaptosome Preparation

Synaptosomes were prepared and collected as we have described in [54] and presented schematically in Figure 3.1A. In brief, mouse brain tissue was dissected and homogenized with a Teflon homogenizer (20 consecutive up and down strokes) in ice-cold HB with protein inhibitors (1µL 1M NaF (Sigma-Aldrich, cat# S-1504), 10 µL 10mg/mL PMSF (EMD Millipore, Billerica, MA, cat# 7110-OP), 10 µL 100mM Na orthovanadate (Sigma-Aldrich, cat# S-6508) and 30 µL stock aprotinin (Sigma-Aldrich, A6279). The resulting homogenate was separated via serial centrifugations, and the synaptosome containing fraction was layered onto a discontinuous Ficoll gradient (13%, 9%, 5% in HB, Sigma-Aldrich, cat# F9378), and separated via ultracentrifugation (Optima XPN-100 Ultracentrifuge) at 22,500 rpm for 35 min at 4^{o}C using a SW41Ti swing-bucket rotor (Beckman Coulter, Mississauga, ON, cat# 331362). Intact synaptosomes were collected from the 13%-9% interface and stored at -80^{o}C in one of three ways: (1) frozen in the 13%-9% Ficoll layer, (2) as a pellet, once the Ficoll layer was removed, or (3) as the pellet in (2), resuspended in HB. The synaptosome pellet was formed by adding HB to the synaptosome-13%-9% Ficoll mixture to a final volume of 5 mL, and centrifuged at 13250 g for 15 minutes at 4^{o}C. The pelleted fraction comprised the synaptosomes; the Ficoll-HB mixture was removed.

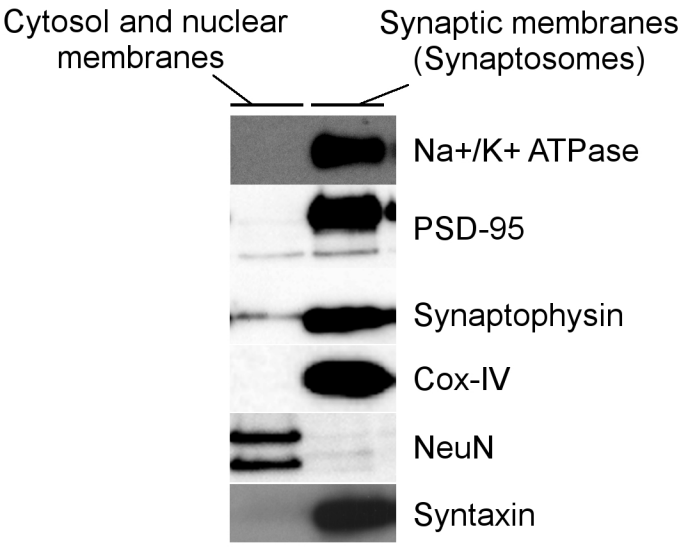
3.4.3 Western Blots

Left cerebral synaptosome samples, stored as previously described in 13%-9% Ficoll, were re-suspended in 250 mL HB, and diluted with HB and 8 M Urea (Fisher Scientific, New Jersey US, cat# U-15) to a 1:2 dilution, with final concentration of

Figure 3.1: Synaptosome fractionation, validation, and lipidomic analysis. (A) Methodological schema. Mouse hippocampal and half-cerebrum tissue were dissected, homogenized, and fractionated via serial centrifugation and density gradient separation. Synaptosomes were collected and fractionations validated by Western blotting (half-cerebrums) or extracted for lipid analysis (hippocampus). Lipids were analyzed by LC-ESI-MS in either precursor ion scan mode or neutral loss scan mode. Identities of lipids were predicted using both VaLID v3.2 and LipidMAPS. (B) Validation of synaptosome methodology by immunoblot analysis of various organelle markers. Na^+/K^+ ATPase was used as a neuronal plasma membrane marker, postsynaptic density protein 95 (PSD-95) was used as a post-synaptic density marker. Synaptophysin, was used as a synaptic vesicle marker, cytochrome c oxidase complex IV (Cox-IV) was used as a mitochondrial marker, feminizing locus on X-3 (Fox-3, also known as NeuN), was used as a neuronal nuclei marker, and syntaxin was used to identify pre-synaptic SNARE protein enriched at pre-synaptic active zones.


Hippocampus/Half brain dissected
Homogenize
Serial centrifugations, synaptosome collection
Western blotting
Half-Brain
Hippocampus
Lipids extracted
HPLC
LC-ESI-MS
m/z
523
524
525
526
527
Under vacuum
Detector
Quadrupole 1
Quadrupole 2
Quadrupole 3
Precursor Ion Scan
Scan
Fragmentation
Product ion selection
Neutral Loss Scan
Scan
Fragmentation
Scan
Prediction comparison
VaLID
LIPID MAPS

A


Cytosol and nuclear membranes
Synaptic membranes (Synaptosomes)
Na+/K+ ATPase
PSD-95
Synaptophysin
Cox-IV
NeuN
Syntaxin

B

Figure 3.1

4 M Urea, and assayed for protein concentration using the BioRad DC protein assay kit (Bio-Rad, Hercules, CA, cat# 500-0111). Due to the low concentration of protein, the protein content was assayed using the low concentration assay conditions. The cytosol and nuclear membrane samples were also diluted, and assayed similarly. Samples were diluted with 10% β-mercaptoethanol (Sigma-Aldrich, cat# M6250), HB with 4 M Urea, and 2x loading buffer (4x LDS Sample Buffer (NuPAGE, Life Technologies, Carlsbad, CA, cat# NP0007), Sample Reducing Agent (NuPAGE, cat# NP0009) and water, 5:2:3 ratio), and boiled for five minutes. Ten µg of protein was the amount used. Proteins were resolved under reducing conditions on 4%-12% Bis-Tris polyacrylamide gels (Life Technologies, Carlsbad, CA, cat# NP0335BOX) and transferred onto Immobilon-P polyvinylidene fluoride (PVDF) membrane (Millipore, cat# IPVH00010) at 220mA for 1 hour. Membranes were blocked in 5% skim milk (Carnation, Smucker Foods, Markham ON) in 10 mM phosphate buffered saline (PBS, 200 mM phosphate buffer [0.5 mM monobasic, Bioshop, Burlington ON, cat# SPM306, 0.5 mM dibasic, Bioshop, cat# SPD307, 6:14:30, monobasic:dibasic:water], 154 mM NaCl [Bioshop, cat# SOD001.10]) with 0.1% Tween-20 (Sigma-Aldrich, cat# P1379) (PBS-T) solution for one hour and incubated in primary antibody diluted in 0.1% PBS-T with 5% skim milk overnight at 4^{o}C (See Table 3.1 for working concentrations). Membranes were rinsed twice in 0.1% PBS-T and thrice in 0.1% PBS-T with 5% skim milk for 10 minutes prior to a 1-3 hr incubation in horseradish peroxidase (HRP) –conjugated anti-mouse (Jackson ImmunoResearch Laboratories, PA, cat# 115-035-146, 1:2000) secondary antibody diluted in 0.1% PBS-T with 5% skim milk. Signal was detected using Immobilon Western Chemiluminescent HRP Substrate (Millipore, cat# WBKLS0500), either on X-

Table 3.1: Antibodies used for verification of synaptosomes

Target	Company	Catalogue number	Concentration	Species	Type
Na^+/K^+ ATPase	BD Transduction	610915	1:10,000	Mouse	Monoclonal
PSD-95	Thermo	MA1-046	1:500	Mouse	Monoclonal
Synaptophysin	Abcam	Ab8049	1:4,000	Mouse	Monoclonal
Cox-IV	Molecular Probes	A21348	1:1,000	Mouse	Monoclonal
NeuN	Chemicon	MAB377	1:250	Mouse	Monoclonal
Syntaxin	Sigma	S0664	1:5,000	Mouse	Monoclonal

ray film, or using a GE Healthcare Life Sciences ImageQuant LAS 4010 (GE Healthcare, #28-9558-10).

3.4.4 Lipid Extractions

Lipids were extracted from murine hippocampal synaptosomes, using a modified acidified Bligh and Dyer extraction [18] as described in [49]. Briefly, synaptosomes were re-suspended in 250 µL 0.1M filtered sodium acetate (Bioshop, cat# SAA304), and of this 50 µL was saved for protein analysis. 0.1 M sodium acetate was added to a final volume of 3.2 mL, and to this 4 mL of 2% (v/v) glacial acetic acid (Fisher Scientific, cat# 351271-212) in methanol (Fisher Scientific, cat# A412P-4) was added, along with 41.3 µL of the internal standard, 10 µM PC(13:0/0:0) (Avanti Polar Lipids, Alabaster, Alabama, cat# 855476). Lipids were extracted by adding 3.8 mL chloroform (Fisher Scientific, cat# C298-500), and the organic phase was collected following a two minute centrifugation at 2000 rpm at 4^{o}C. The aqueous phase was back-extracted three times by sequential addition of 2mL chloroform, centrifugation and organic phase recovery. Chloroform was evaporated under N_2 gas, and lipids were re-suspended in 300 µL of 100% ethanol, and stored in 50 µL aliquots in 2 mL amber glass vials (Chromatographic Specialties Inc, Brockville ON, cat# C779100AW), under a N_2 atmosphere at -80^{o}C.

3.4.5 LC-ESI-MS

LC-ESI-MS/MS analysis was performed on an Agilent 1100 micro-HPLC system (Agilent Technologies, Santa Clara, CA, USA) coupled to an Applied Biosystems API 5500 Q/TRAP mass spectrometer (Applied Biosystems, Foster City, CA) equipped with a

nano-electrospray interface operating in triple quadrupole mode. The mobile phases consisted of 0.1% (v/v) formic acid (FA) (Sigma-Aldrich, US, cat# 56302) and 10 mM ammonium acetate (EMD Milipore, 2145-OP) in water as buffer A, and 0.1% (v/v) FA, 10 mM NH_4Ac in acetonitrile (J.T. Baker, Avantor, Centre Valley, PA, cat# 9829-03)/isopropyl alcohol (Fisher Chemical, cat# A416-4) (5:2 v/v) as buffer B. Sample preparations included 5 µL of lipid extract, 2.5 µL of deuterated lipid standard mixture [1 ng/µL d_4-PC(O-16:0/0:0) (Cayman Chemicals, Ann Arbor, Michigan, cat# 3609060), 1 ng/µL d_4-PC(O-18:0/0:0) (Cayman Chemicals, cat# 10010228), 0.5 ng/µL d_4-PC(O-16:0/2:0) (Cayman Chemicals, cat# 360900) and 0.5 ng/µL d_4-PC(O-18:0/2:0) (Cayman Chemicals, cat# 10010229)] and 15.75 µL of 0.1% FA in water. Lipid separation was achieved on a 75 µm (I.D.) x 100 mm analytical column packed in house with reverse phase Magic C_4 AQ resins (5 µm; 120-Å pore size; Dr. Maisch GmbH, Ammerbuch, Germany). Sample was loaded onto the column by an auto sampler (Agilent Technologies) in 95% buffer A at a flow rate of 10 µL/min for 2.5 minutes. The flow rate was increased to 20 µL/min, and lipid separation was achieved by a linear gradient of buffer B from 5% to 35% in 12.5 min, then the gradient changed to 100% B in 15 min, followed by 100% B for 14 minutes. The column was then equilibrated at 95% A for 15 min. A flow rate of ~300 nL/min was achieved by splitting a 20 µL/min flow through a restriction capillary. Detection of PC was performed in positive ion mode by precursor ion scanning for a diagnostic PC product ion of 184.2 m/z (PC head group). Detection of PE was performed in positive ion mode neutral loss, with the neutral loss mass set to 141.0 m/z, and detection of PS was also performed in positive ion mode neutral loss, with the neutral loss mass set to 185.0 m/z. It was discovered that high molecular mass lipids

were in such high abundance that they were over-saturated on the spectra, potentially confounding analysis. To correct for this, samples were further diluted 10x and re-loaded onto the column.

3.4.6 Data Analysis

Lipid data were analyzed using Absciex' Analyst® 1.5 software. All visible peaks displayed on the 2-dimensional LC elution plot (RT vs. *m/z*) were selected. If the curve of the Extracted Ion Chromatograph (XIC) was unambiguously a peak, the *m/z* was saved. Otherwise, integrated peak area data for the remaining peaks was collected, and if the peak area was greater than $1.0x10^4$, the m/z was saved.

3.4.7 Identifying Potential Species with Both VaLID and LipidMAPS

Once peak m/z ratios were identified, these peak m/z ratios were inputted into VaLID 3.2's search feature[4], using the search parameters: Exact Mass, restricted to Even Chain lengths, ±1 Mass Tolerance, subclass of PC, PE or PS depending on the set of m/z used, all fatty chain linkages, with $[M+H]^+$ ion mode. For each m/z inputted, the search results were saved to a comma separated value file, for analysis.

For the comparison, the same m/z ratios were inputted into the LipidMAPS structural database (LMSD), using the "Search the LMSD with a mass (m/z) value" search feature[5], with the search parameters: Mass +/- 1, category Glycerophospholipids [GP], main class glycerophosphocholines [01], glycerophosphoethanolamines [GP02] or

[4] http://www.neurolipidomics.ca
[5] http://www.lipidmaps.org/tools/ms/LMSD_search_mass_options.php

glycerophosphoserines [GP03] depending on the set of m/z used. All other search parameters were kept constant, and the m/z was inputted in the Mass field. For each m/z inputted, the numbers of results were tabulated, and saved to a file for analysis.

3.5 Results

3.5.1 Isolation and Characterization of Synaptosomes from Mouse Hippocampi

Before performing the lipid extraction and the verification of VaLID, the integrity of synaptosomal fractions prepared from mouse half-cerebrum were established. Proteins from the synaptosome containing 13%-9% Ficoll in HB interface and the cytosol and nuclear membrane-containing fraction were prepared as described above for Western blotting. The results of the Western blot can be seen in Figure 3.1B. The cytosolic and nuclear membrane fraction was deficient in the plasma, mitochondrial, and pre- and post-synaptic membrane and vesicle markers, Na^+/K^+ ATPase, Cox-IV, PSD-95, and syntaxin, and greatly reduced in synaptophysin. This fraction contained markers of neuronal nuclei (NeuN). Conversely, the synaptosome fraction was enriched for synaptophysin, Na^+/K^+ ATPase, PSD-95, Cox-IV and syntaxin but was deficient in NeuN. This characterization clearly demonstrates successful isolation of synaptic membranes.

3.5.2 Collection and Analysis of Lipids from Mouse Hippocampal Synaptosomes

Mouse hippocampi were fractionated into synaptosomes and stored under in 13%-9% Ficoll in HB, as described above. Using the modified Bligh and Dyer method for lipid extraction, as described above, lipids were extracted from the three samples. Lipid samples were loaded onto the LC-ESI-MS. PC, PE and PS spectra were collected and

analyzed. As alluded to above, the high molecular mass lipids on the spectra were saturated, i.e., beyond the detection range of the mass spectrometer. To be able to analyze results for the high molecular peaks, the samples were diluted 10x, and then re-loaded onto the LC-ESI-MS. New spectra were collected and analyzed. The m/z lists from the samples were consolidated and can be seen in Table 3.2, split into the three sub-classes. From Table 3.2, 20 PC m/z were identified in the samples, not including the exogenous standards, with masses ranging from 496.3 m/z, to 834.5 m/z, 58 PE m/z were identified in the samples, with masses ranging from 313.1 m/z to 840.3 m/z, and 29 PS m/z were identified in the samples, with masses ranging from 307.8 m/z to 934.5 m/z.

3.5.3 VaLID Contains Entries for Neural-Specific Samples Not Detected in Other Tissue-Specific Databases

All of the masses from Table 3.2 were inputted into VaLID, and for each m/z, the search results were saved and the number of results tallied for each m/z under their respective subclasses. I then searched for these same m/z ratios on LipidMAPS, and tallied the number of results returned from the LipidMAPS database. The results are summarized in Table 3.3. From Table 3.3, both VaLID and LipidMAPS returned results for the m/z inputted, with VaLID returning between 10.5 and 400 times more results than did LipidMAPS. There were also some search-terms where VaLID returned results and LipidMAPS did not, for example, PE 566.3 m/z, or PS 674.2 m/z. Finally, there were 15 m/z that did not produce any results in either VaLID or LipidMAPS. Every straight-chain, non-modified headgroup lipid identity prediction found in LipidMAPS was represented within VaLID's predictions.

Table 3.2: List of the peak m/z found in the synaptosome samples, not including exogenous standards, PC, PE and PS.

Mass/Charge				
		PC		
496.3	675.7	703.6	706.6	729.5
731.6	732.6	734.5	758.5	760.6
762.6	776.6	782.5	786.6	788.6
806.6	808.6	810.5	832.5	834.5
		PE		
313.1	314.1	316.0	318.0	328.1
330.1	332.1	333.9	342.1	344.1
345.9	356.1	358.1	360.1	371.9
373.9	386.1	454.1	480.3	482.3
502.2	508.3	526.2	350.2	554.2
558.1	566.3	580.4	679.5	690.4
692.3	695.6	704.5	716.3	718.4
724.5	726.5	735.4	738.5	740.4
744.4	748.4	750.4	764.4	766.4
768.5	774.5	776.4	778.3	788.6
790.5	792.3	812.4	814.6	816.6
836.4	837.2	840.3		
		PS		
307.8	309.7	315.7	318.3	319.8
321.7	332.3	376.9	394.0	428.2
468.3	610.5	674.2	680.3	762.5
784.4	788.6	791.6	808.6	810.4
812.6	834.6	836.5	856.6	880.6
882.6	884.5	912.6	934.5	

3.6 Discussion

The present chapter describes the validation of the program VaLID using a lipid dataset generated from murine hippocampal synaptosomes, by comparing the identities predicted by VaLID to that by the current gold-standard, LipidMAPS. Extracting lipids from hippocampal synaptosomes proved successful, as 20 peaks were identified for PCs, 58 peaks were identified for PEs and 29 peaks were identified for PSs (Table 3.2). For the peaks identified in Table 3.2, VaLID identified more potential lipid identities than did LipidMAPS (Table 3.3), and VaLID had predicted identities where LipidMAPS did not. However, there were, 15 m/z identified across the PE and PS spectra that did not correspond to identities found in either VaLID or LipidMAPS. This could represent some other class of lipids, with a fragmentation pattern resulting in a neutral loss of 141.0 or 185.0 m/z, respectively. Taken together, this demonstrates that VaLID is able to cope with the complexity of a real neural dataset, as well as, and even surpassing the current gold-standard LipidMAPS in certain species, and can provide users with more possible identities than do current, more canonical tools.

In this chapter I used murine synaptosomes to generate a neural dataset to challenge VaLID in order to validate its effectiveness with a dataset that had in the past been under-represented. To confirm the protocol, I first isolated synaptosomes from murine half-cerebrum, and through Western blotting, identified multiple proteins present (Figure 3.1B). The presence of the detected proteins have all been reported in synaptosomal preparations previously: Na^+/K^+ ATPase has been shown to localize to the synaptosomal membrane [52], PSD-95, a marker for the post-synaptic density, has also been detected in synaptosomes [47]. Synaptophysin, marker of synaptic vesicles and

Table 3.3: Number of entries in both the VaLID and LipidMAPS databases for each m/z from the PC, PE and PS hippocampal synaptosome lipid data. Accessed February 19th, 2015

m/z	VaLID number	LMSD Number	m/z	VaLID Number	LMSD Number	m/z	VaLID Number	LMSD Number
				PC				
496.3	53	6	**675.7**	311	8	**703.6**	347	13
706.6	348	15	**729.5**	424	11	**731.6**	390	21
732.6	377	23	**734.5**	377	18	**758.5**	398	36
760.6	396	29	**762.6**	394	16	**776.6**	397	23
782.5	405	28	**786.6**	405	43	**788.6**	401	31
806.6	400	25	**808.6**	400	24	**810.5**	402	22
832.5	383	20	**834.5**	385	13			
				PE				
313.1	15	0	**314.1**	12	0	**316.0**	0	0
318.0	0	0	**328.1**	15	0	**330.1**	0	0
332.1	0	0	**333.9**	2	0	**342.1**	16	0
344.1	0	0	**345.9**	0	0	**356.1**	19	0
358.1	0	0	**360.1**	2	0	**371.9**	0	0
373.9	5	0	**386.1**	2	0	**454.1**	53	1
480.3	94	1	**482.3**	90	1	**502.2**	146	1
508.3	132	1	**526.2**	166	1	**530.2**	175	1
554.2	206	0	**558.1**	209	0	**566.3**	210	0
580.4	231	1	**679.5**	361	11	**690.4**	377	19
692.3	377	12	**695.6**	386	5	**704.5**	389	20
716.3	398	19	**718.4**	396	24	**724.5**	401	13
726.5	401	19	**735.4**	397	19	**738.5**	403	12
740.4	405	14	**744.4**	405	20	**748.4**	399	17
750.4	402	8	**764.4**	400	12	**766.4**	400	14
768.5	402	16	**774.5**	394	22	**776.4**	392	12
778.3	393	8	**788.6**	387	21	**790.5**	383	18
792.3	385	11	**812.4**	368	8	**814.6**	364	12
816.6	362	12	**836.4**	338	1	**837.2**	338	1
840.3	332	2						
				PS				
307.8	0	0	**309.7**	0	0	**315.7**	7	0
318.3	0	0	**319.8**	0	0	**321.7**	0	0
332.3	0	0	**376.9**	0	0	**394.0**	18	0
428.2	22	0	**468.3**	56	1	**610.5**	210	0
674.2	309	0	**680.3**	311	6	**762.5**	396	23
784.4	405	13	**788.6**	405	18	**791.6**	386	14
808.6	400	12	**810.4**	400	14	**812.6**	402	14
834.6	383	18	**836.5**	385	11	**856.6**	386	7
880.6	338	0	**882.6**	334	2	**884.5**	332	2
912.6	292	0	**934.5**	260	0			

syntaxin, the SNARE protein and Cox-IV, a mitochondrial marker are all also found in synaptosomes [47, 52]. The lack of neuronal membranes shown by the lack of NeuN, demonstrates that synaptosomes were created and isolated successfully from the protocol.

Lipid extraction also proved successful, in that multiple peaks were detected, beyond the internal and external standards (Table 3.2). While this was expected, since previous studies had successfully extracted phospholipids from mammalian synaptosomes [54-56], unexpectedly, we detected significantly fewer low molecular weight species than previously reported [54], at least with respect to PCs. This discrepancy is likely because Wislet-Gendebien *et al.* detected lipids from lipid-protein complexes, whereas our data is specific to the synaptosomal profile. The lack of low-molecular weight species in our samples suggests that at time of lipid extraction, the synaptosomes were static; i.e., they were not undergoing synaptic transmission. Since both the exocytosis and endocytosis involved in synaptic transmission depend on severe and rapid membrane curvature, a process for which single-chain phospholipids (i.e., low-molecular weight species) are required due to the geometry they can form [24, 46]. It is also possible that during the creation of the synaptosomes low molecular weight species are released or lost, or perhaps are cleared from the synaptic membrane as they are made. However, since the high molecular species were highly abundant, to the point of needing dilution to not have the spectra over-saturated, it is reasonable to expect that if species are lost in the process of creating synaptosomes, the high-molecular weight structural lipids would also be lost. This would suggest that the low molecular weight species might be shipped away from the membrane quickly and are remodelled quickly into high-molecular weight structural membrane lipids, or the membrane is not undergoing any

remodelling. Further exploration into potential mechanisms is required to be able to definitively say why the lower molecular weight species were not as highly represented as the high-molecular weight species.

In comparing the search results from both VaLID and LipidMAPS (Table 3.3), the immediate finding is that for every m/z found there are many more results in VaLID than in LipidMAPS. This validates the program, and the *in silico* approach used, as it contains far more predictions than does the current gold standard, including lipid predictions for lipids not found in LipidMAPS. The overall aim of this chapter was to establish whether or not VaLID could identify neural lipids as identified from LC-ESI-MS spectra. Seeing that VaLID has multiple predictions for the identities of them/z in the list demonstrates that the program is capable of handling a real neural dataset, and the vastly greater number of lipid predictions in VaLID demonstrate that there are potentially hundreds of species absent from current databases. Since lipidomic technologies are still in their infancy [10], and until complete lipidomic profiling of all tissues in biological systems, including across multiple species, occurs, lipidomic tools based solely upon existing literature may erroneously attribute novel lipid findings to inappropriate lipid candidates. Therefore, only with *in silico* created databases, such as the phospholipid database on which VaLID is based, is it possible to truly achieve a comprehensive tool that will contain all possible biologically relevant lipid species.

Chapter 4: Discussion

4.1 Summary of Work Completed

A challenge that faces lipidomic researchers is that of distinguishing among the tens of thousands of lipid species, in multiple classes. I addressed this challenge in the work described in this thesis: I designed, created, and tested the lipidomic bioinformatic tool VaLID. VaLID is a web-based computer program that was coded in Oracle's Java and was released to be freely accessible on the web[6]. To validate the program's effectiveness in identifying lipids, I challenged it with a neural dataset: PCs, PSs and PEs from murine hippocampal synaptosomes – confirmed to be enriched in synaptic membranes by Western blotting. I compared the results to those produced by LipidMAPS[7], recognized at the time as the gold standard lipidomic bioinformatics tool. VaLID consistently characterized more potential phospholipid identities than did LipidMAPS across both of the phospholipid subclasses used, and for some m/z's present in the neural dataset, provided results where LipidMAPS did not. These results suggest that the program can be a useful tool for identifying lipids for lipidomic researchers, especially those studying a variety of phospholipid subclasses from tissues not previously well characterized, such as various neural tissues. VaLID is being increasingly used in laboratories across North America. It is now pertinent to step back and view this work within its broader academic context.

[6] http://www.neurolipidomics.ca
[7] http://www.lipidmaps.org/tools/ms/LMSD_search_mass_options.php

4.2 Context: Why is This Work Important?

4.2.1 What Analytic Tools Existed Before VaLID?

In the approximately 10 years since lipidomics began [2, 8, 11, 35, 57], many tools and databases have been developed to address various challenges associated with studying lipids. Several resources for lipid researchers became available, one of the earliest of which was the Lipid Library, operated by the American Oil Chemists' Society AOCS[8]. While not directly a tool for the identification of lipids or novel biochemical pathways, it is a great resource for providing access to basic information on lipids and lipidomics in general. It contains background information on the biochemistry and physics of lipids, on lipid mass spectrometry, NMR libraries, and many other general lipid topics. Similarly, the CyberLipid[9] website forms an "online, non-profit scientific organization whose purpose is to collect, study and diffuse information on all aspects of lipidology." This site contains resources covering all aspects of lipids and lipidomics, encompassing discussions of practical approaches to lipid analysis including extraction, handling and fractionation of lipid samples. It also contains information on the various types of lipids, a search function to various literature resources on specific sub-classes, and a calendar of lipid-related international events – symposia, exhibitions, conferences, etc. The site is continually updated[10]. While the Lipid Library and CyberLipid are both valuable resources that highlight different areas of the growing research into lipidomics, they do not assist researchers in identifying and classifying their individual lipid samples.

[8] http://lipidlibrary.aocs.org
[9] http://www.cyberlipid.org
[10] Accessed December 7th, 2014

One database that does allow users to classify samples is LIPID BANK[11] (LipidBank). Mentioned briefly in Chapter 2, it was developed in Japan as the first comprehensive lipid database. It provides biochemical data on a variety of lipid classes and specific species. LipidBank is a freely accessible database consisting of fatty acids, glycerolipids, sphingolipids, steroids and various vitamins. As of December, 2014, it contained more than 7000 unique molecular structures, spectral information, names and literature information. Each lipid within the database was manually curated [58]. LipidBank contains a search feature whereby users can sort through the lipids reported in the database, including searching by weight. This bioinformatic feature of the website could help researchers identify species within their samples. LipidBank also contains other biochemical and spectral information, including UV, IR, NMR, or MS spectra, physical and chemical properties, and other information extracted from a range of publications, such as synthesis, metabolism and any other relevant information, along with the reference sources used for this information. Unfortunately, but understandably, since all of this information is manually curated, the LipidBank database contains approximately 7000 entries, 310 of which are glycerophospholipids[12]. At the time LipidBank was created, there was no universal nomenclature, nor were there standards for structural representations; the site was created in Japanese first and later translated into English. These factors make it difficult to find lipids and compare results between LipidBank and other resources. It was not until the LipidMAPS Consortium that these standardization issues were addressed. Nonetheless, if these potential limitations are borne in mind, LipidBank is still a good resource for lipid researchers.

[11] http://lipidbank.jp/
[12] Accessed December 7th, 2014.

While each of these resources and databases helped advance the field, it was the Lipid MAPS Consortium's group of tools that made the strongest contribution. As mentioned in Chapters 2 and 3, the main tool prior to the development of VaLID was the LipidMAPS structural database[13], which has become the major gold-standard tool of lipidomics. Developed by the LipidMAPS Consortium in the mid-2000s, the Lipidomics Gateway became a leader in the field in terms of standardization of lipid ontology, nomenclature, classification and structural representation [35, 40]. In addition, it contained a large array of structural and biochemical information on various fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, and others [35], drawn from a variety of sources. These sources included data generated from the LipidMAPS Consortium's core laboratories and partners, lipids identified in experimental work by these groups using specific cell lines, and biologically relevant lipids curated from other public sources, such as LIPID BANK and LIPIDAT [40]. LipidMAPS has contributed substantially to the field by standardizing nomenclature and structural representations. Indeed, VaLID, like other groups such as KEGG, adopted nomenclature derived from LipidMAPS [34, 59]. Thus, while we found some problems and limitations with LipidMAPS as demonstrated in Chapter 3, it remains a powerful tool, and the consortium behind it has made significant contributions to the field. However, the issues described in Chapters 2 and 3 led to the creation of VaLID.

4.2.2 How Has VaLID Moved the Field Forward?

Chapter 1 presented the challenges facing lipidomics, using a structure proposed by Niemelä *et al* [10]. These challenges included 1) lipid identification, 2) statistical

[13] http://www.lipidmaps.org/data/structure/LMSDSearch.php?Mode=SetupTextOntologySearch or http://www.lipidmaps.org/tools/ms/LMSD_search_mass_options.php

analysis, 3) pathway analysis, and 4) lipid modeling, specifically in biophysical contexts. VaLID was created to address the challenge of lipid identification in a way that had not been used previously. Prior to VaLID, the lipids included in databases were selected from the experimental work undertaken by the creators of the programs. For example, the LipidMAPS Structural Database was populated with lipids curated from their member laboratories and their cell-line data [40] and this experimentally-based approach was the method for most tools. In creating VaLID, we took a completely *in silico* approach, based on computer simulation which, to our knowledge, had not been done previously. By generating a database with every possible lipid, we have created a system that is theoretically capable of identifying any straight-chained lipid (i.e., non-branched). This addresses Niemelä's requirements for a bioinformatic tool or database to "facilitate turning raw instrumental data from experiments into a final lipidomics dataset" [10], as it can theoretically handle any theoretically possible lipid, including those yet to be discovered.

This comprehensiveness forms a significant advance, but brings with it an obvious challenge. VaLID generates what can be a long list of potential phospholipid identities, potentially generating a large number of false positives. While this comprehensiveness will be able to find the species the researcher has, more needs to be done to narrow down the possibilities and definitively identify a species. VaLID lists all of the possible options and does not give any rankings about how likely the prediction is to be the accurate beyond a "best guess" or "best prediction" option based on the frequency of fatty acyl, alkyl, or alkenyl-linked carbon chains being represented in plants and animals, and thus we leave it up to the researcher to identify the lipids within their

datasets. Achieving the next step of complete identification will require other techniques of separation and identification working together to unambiguously distinguish between lipids. Upgrading VaLID to incorporate one or more alternative identification technique is a goal for VaLID's future development.

As described in Chapter 2, VaLID has also begun to address Niemelä's third challenge of pathway analysis, "identifying affected lipid (biochemical, signaling, regulatory) pathways" [10]. While VaLID cannot map pathways of lipids and their interacting partners, we have begun to lay the groundwork for an *in silico* approach to identifying and describing these possible pathways. By linking VaLID to the external literature sources PubMed and the HMDB, users can search for literature referencing any lipid within the VaLID database. This allows researchers to find deeper understanding of various lipid biological properties, and potential binding partners/pathways specific to their lipid of interest. There is still a long way to go towards creating a program as detailed and similar to the Ingenuity pathway analysis software[14] available for genomics or proteomics, for lipidomics. In the meantime, using resources like the KEGG pathway [34] and linking programs such as VaLID to literature sources makes it possible to analyse biochemical pathways in lipidomics.

The main advantage to VaLID lies in the *in silico* methodology used to generate the underlying database. By computationally generating all of the theoretically possible lipid species, it inherently increases VaLID's ability to identify lipids; giving researchers all of the possible identities prevents some lipids from not being identified, as was the case in previous databases. While in some cases, the large number of possible species proposed as identities may be overwhelming, giving researchers all the possibilities can

[14] http://www.ingenuity.com/products/ipa

help propose further separation or experimentation, so they can further identify their species of interest. The identity predictions, coupled with the connection to literature sources, beginning the process of creating biological pathway maps, pushes the field forward. However, VaLID was not the only program developed during this time. Other programs devised other methods to help identify lipids within samples. One of these techniques, based on fragmentation patterns, was utilized by the program LipidBlast [60].

4.2.3 What Has Occurred Since VaLID Was Developed?

VaLID was, of course, not created in a vacuum and other groups have been working on similar tools for lipidomic bioinformatic development. Development of VaLID began in 2011 and it was first published in early 2013. During that time, other lipidomic tools have been developed. Tools such as LipidHome, and LipidBlast also tackle the problem lipid identification. LipidHome[15] [61] is a database of *in silico* generated lipids from a set of starting parameters, which provides a comprehensive reference for lipid identification. The database is set up in such a way that it organizes the lipids within in multiple different hierarchies. For example, if a user is searching for the lipid PC(16:0/18:0), there are separate pages for the high structural resolution of lipid species itself (e.g., PC(16:0/18:0)), but also for the lower resolution, depicting what we termed the Total Carbon number (i.e., PC(34:0)). LipidHome also provides meta-data, including the formula, mass, number of sub species and any cross-references or papers, for each entry into the database, where available. Because of the multiple hierarchical levels represented, this tool provides information on both the species level for species that

[15] http://www.ebi.ac.uk/metabolights/lipidhome/

have been identified, and at a resolution more apt for high-throughput lipidomics, similar to the third update of VaLID. LipidHome also contains a search tool, based off the precursor masses, resulting in a list of the lipids within the database that fit under each of the masses. As powerful as this program is, unfortunately (as of December 2014) the information for each individual lipid species, specifically the cross-references and papers, seems to have been removed. However, the hierarchical search is still functional, and so the tool can still be used.

Another program, LipidBlast,[16] offers a slightly different approach to lipid identification, however still uses an *in silico* approach. Instead of focusing specifically on the m/z to identify the lipid species, LipidBlast looks at the fragmentation patterns of lipids. By studying many MS/MS spectra of different species with varying carbon chain lengths and number and location of unsaturations, the creators of LipidBlast noticed that the lipids showed a predictable pattern in their MS/MS spectra, allowing them to generate *in silico* a library of MS/MS spectra that researchers can compare to their spectra in order to identify what lipids are in their samples. Their library contains 212,516 spectra covering 119,200 compounds [60]. This program represents another method to identify lipids within lipidomic experiments, but (unlike the case of LipidHome or VaLID) would not be useful with high-throughput experimentations, because MS/MS analysis is not particularly compatible with high-throughput techniques, at the current time.

Other developments in lipidomics have also occurred since VaLID was developed. For example, one group began to tackle the challenge of pathway analysis, specifically by addressing ontology. The goal of gene ontology, the *de facto* standard for functional annotation of gene products, is to unify the representation of genes and their

[16] http://fiehnlab.ucdavis.edu/projects/LipidBlast

products across all species, by using a controlled vocabulary to describe these products. Standardizing ontology is the first step in being able to make comparisons and links between different databases and fields. The group behind the LipidGO[17] tool [62] has applied gene ontology vocabulary to lipid-related terms, beginning the process of creating tools that can efficiently generate biochemical pathway maps specific to lipidomic datasets, addressing the third of Niemelä's challenges [10].

These are just some examples of lipidomic bioinformatic tools that were built around the same time as VaLID. New tools are being created all the time to address gaps of knowledge, and bioinformatic coverage, in this active field.

4.2.4 What Are the Implications for the Field of Lipidomics?

At the beginning of this thesis, I briefly described the evolution of the study of lipids and leading to the current field of lipidomics, once systems biology techniques were applied to studying lipids. I then argued that it was because of the relative youth of the field, compared to other fields such as genomics or proteomics, that there was not extensive bioinformatic coverage for lipidomics. This is now no longer the case. Programs such as VaLID or LipidBlast, and groups such as LipidMAPS have enabled researchers to more adequately study lipids, and there is greater interest in studying lipids than before. This has both led to, and resulted from, the large number of bioinformatic tools being created. Both VaLID and LipidBlast represent a change in the paradigm, from purely curated databases, to *in silico* created ones. While having libraries of curated lipids is ideal, the large variety of theoretically possible lipid species, coupled with technological challenges associated with synthesizing enough standards to populate these

[17] http://compbio.ddns.comp.nus.edu.sg/%7elipidgo/index.php

libraries and databases with the relevant information, makes it infeasible to make these databases comprehensive enough, based on curated data alone. Thus, the future of lipidomic bioinformatics depends on these *in silico* created tools, gradually complemented by further curation to prove biological relevancy.

A future milestone for teams to address is how these new programs can be integrated or overlapped with each other to comprehensively address the issues facing the field.

4.3 Future Directions for Lipidomic Bioinformatics

Every science requires accurate measurement and classification. Lipidomics is a relatively new field in which a current priority is to refine measurement and classification tools. Every laboratory science goes through this phase and once this is complete everyone uses the measurement and analysis methods routinely, and the groundwork that was required to develop the assays is forgotten. Hence, the topic addressed in this thesis forms a necessary but transient stage in building this sub-domain of systems biology.

Once the technology for these measurement and analysis methods has been established, researchers will be able to accurately identify the lipids within their samples and analyse how these lipids change temporally and spatially between test cases. Since the production and metabolism of every aspect of biology is interlinked, researchers will then be able to model how the changes found in lipid profiles affect all aspects of cellular and tissue function throughout the organism. In order to reach this goal, tools will have to be developed that can accurately identify lipids, tools that can map the location and changes in these lipids, tools that can draw from other fields consolidating the results into

a cohesive map, and modelling techniques to determine how all of these changes fit together globally. The tools will have to be developed in such a way that they can communicate with each other. Once these inter-linked tools have been created, the field will be able to finally answer the two major, pressing questions: How many lipid species are there? and How do they interact with other molecules? [11]. *In silico* created databases, and programs like VaLID, are one necessary step towards making this goal a reality.

In short, while the field of lipidomics is moving into its second decade [8, 11, 57] and has made remarkable progress in a brief time, more work is needed to fully appreciate the importance and impact that these oily substances that surround all of our cells have on our biology and on life in general.

References

1. Voet, D., Voet, J.G.: Biochemistry. John Wiley & Sons, USA (2004)

2. Wenk, M.R.: The Emerging Field of Lipidomics. Nature Reviews Drug Discovery 4, 594-610 (2005)

3. Kresge, N., Simoni, R.D., Hill, R.L.: JBC Historical Perspectives: Lipid Biochemistry. Journal of Biological Chemistry Historical Perspectives, H1-H2 (2010)

4. Shevchenko, A., Simons, K.: Lipidomics: Coming to Grips With Lipid Diversity. Nature Reviews 11, 593-598 (2010)

5. Fonteh, A.N., Harrington, R.J., Huhmer, A.F., Biringer, R.G., Riggins, J.N., Harrington, M.G.: Identification of Disease Markers in Human Cerebrospinal Fluid Using Lipidomic and Proteomic Methods. Disease Markers 22, 39-64 (2006)

6. Mapstone, M., Cheema, A.K., Fiandaca, M.S., Zhong, X., Mhyre, T.R., MacArthur, L.H., Hall, W.J., Fisher, S.G., Peterson, D.R., Haley, J.M., Nazar, M.D., Rich, S.A., Berlau, D.J., Peltz, C.B., Tan, M.T., Kawas, C.H., Federoff, H.J.: Plasma Phospholipids Identify Antecedent Memory Impairment in Older Adults. Nature Medicine 20, 415-418 (2014)

7. Kitano, H.: Systems Biology: A Brief Overview. Science 295, 1662-1664 (2002)

8. Li, M., Yang, L., Liu, H.: Analytical Methods in Lipidomics and Their Applications. Analytical Chemistry 86, 161-175 (2013)

9. Han, X.: Neurolipidomics: Challenges and Developments. Frontiers in Bioscience 12, 2601-2615 (2007)

10. Niemelä, P.S., Castillo, S., Sysi-Aho, M., Oresic, M.: Bioinformatics and Computational Methods for Lipidomics. J Chromatogr B Analyt Technol Biomed Life Sci 877, 2855-2862 (2009)

11. Brown, H.A., Murphy, R.C.: Working Towards an Exegesis for Lipids in Biology. Nat Chem Biol 5, 602-606 (2009)

12. Piomelli, D., Astarita, G., Rapaka, R.: A Neuroscientist's Guide to Lipidomics. Nat Rev Neurosci 8, 743-754 (2007)

13. Xu, H., Valenzuela, N., Fai, S., Figeys, D., Bennett, S.A.L.: Targeted Lipidomics - Advances in Profiling Lysophosphocholine and Platelet-Activating Factor Second Messengers. FEBS J 280, 5652-5667 (2013)

14. Karas, M., Hillenkamp, F.: Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons. Analytical Chemistry 60, 2299-2301 (1988)

15. Fenn, J.B.: Electrospray Ionization Mass Spectrometry: How It All Began. Journal of Biomolecular Techniques 13, 101-118 (2002)

16. de Hoffmann, E., Stroobant, V.: Mass Spectrometry Principles and Applications. John Wiley & Sons Ltd, Chichester, West Sussex, England (2007)

17. Han, X., Yang, K., Gross, R.W.: Multi-Dimensional Mass Spectrometry-Based Shotgun Lipidomics and Novel Strategies for Lipidomic Analyses. Mass Spectrometry Reviews 31, 134-178 (2012)

18. Bligh, E.G., Dyer, W.J.: A Rapid Method of Total Lipid Extraction and Purification. Canadian Journal of Biochemistry and Physiology 37, 911-917 (1959)

19. O'Brien, J.S., Sampson, E.L.: Lipid Composition of the Normal Human Brain: Gray Matter, White Matter, and Myelin. Journal of Lipid Research 6, 537-544 (1965)

20. The AOCS Lipid Library, http://lipidlibrary.aocs.org/Lipids/whatdo/index.htm

21. Bergströem, S., Danielsson, H., Samuelsson, B.: The Enzymatic Formation of Prostaglandin E2 from Arachidonic Acid Prostaglandins and Related Factors 32. Biochimica et Biophysica Acta 90, 207-210 (1964)

22. Benveniste, J., Henson, P.M., Cochrane, C.G.: Leukocyte-Dependent Histamine Release from Rabbit Platelets. Journal of Experimental Medicine 136, 1356-1377 (1972)

23. Demopoulos, C.A., Pinckard, R.N., Hanahan, D.J.: Platelet-Activating Factor Evidence for 1-O-Alkyl2-Acetyl-sn-Glyceryl-3-Phosphorylcholine as the Active Component (A new Class of Lipid Chemical Mediators). Journal of Biological Chemistry 254, 9355-9358 (1979)

24. Bennett, S.A.L., Valenzuela, N., Xu, H., Franko, B., Fai, S., Figeys, D.: Using Neurolipidomics to Identify Phospholipid Mediators of Synaptic (Dys)function in Alzheimer's Disease. Front Physiol 4, 1-16 (2013)

25. Suetsugu, S., Kurisu, S., Takenawa, T.: Dynamic Shaping of Cellular Membranes by Phospholipids and Membrane-Deforming Proteins. Physiological Reviews 94, 1219-1248 (2014)

26. McMahon, H.T., Gallop, J.L.: Membrane Curvature and Mechanisms of Dynamic Cell Membrane Remodelling. Nature 438, 590-596 (2005)

27. Kim, H.-Y., Huang, B.X., Spector, A.A.: Phosphatidylserine in the Brain: Metabolism and Function. Progress in Lipid Research 56, 1-18 (2014)

28. Rohacs, T.: Phosphoinositide Regulation of TRP Channels. In: Nilius, B., Flockerzi, V. (eds.) Mammalian Transient Receptor Potential (TRP) Cation Channels, vol. 223, pp. 1143-1176. Springer International Publishing (2014)

29. Holmes, O., Paturi, S., Ye, W., Wolfe, M.S., Selkoe, D.J.: The Effects of Membrane Lipids on the Activity and Processivity of Purified γ-Secretase. Biochemistry 51, 3565-3575 (2012)

30. Lands, W.E.M.: Metabolism of Glycerolipides: a Comparison of Lecithin and Triglyceride Synthesis. Journal of Biological Chemistry 231, 883-888 (1958)

31. Salleh, N.: Diverse Roles of Prostaglandins in Blastocyst Implantation. Scientific World Journal 2014, 1-11 (2014)

32. Aebersold, R., Mann, M.: Mass Spectrometry-Based Proteomics. Nature 422, 198-207 (2003)

33. Luscombe, N.M., Greenbaum, D., Gerstein, M.: What is Bioinformatics? A Proposed Definition and Overview of the Field. Methods of Information in Medicine 40, 346-358 (2001)

34. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayamam, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y.: KEGG for Linking Genomes to Life and the Environment. Nucleic Acids Research 36, D480-D484 (2008)

35. Fahy, E., Cotter, D., Byrnes, R., Sud, M., Maer, A., Li, J., Nadeau, D., Zhau, Y., Subramaniam, S.: Bioinformatics for Lipidomics. Methods in Enzymology 432, 247-273 (2007)

36. Blanchard, A.P., McDowell, G.S., Valenzuela, N., Xu, H., Gelbard, S., Bertrand, M., Slater, G.W., Figeys, D., Fai, S., Bennett, S.A.L.: Visualization and Phospholipid Identification (VaLID): Online Integrated Search Engine Capable of Identifying and Visualizing Glycerophospholipids with Given Mass. Bioinformatics 29, 284-285 (2013)

37. McDowell, G.S.V., P Blanchard, A., Taylor, G.P., Figeys, D., Fai, S., Bennett, S.A.L.: Predicting Glycerophosphoinositol Identities in Lipidomic Datasets Using VaLID (Visualization and Phospholipid Identification) – an Online Bioinformatic Search Engine. BioMed Research International 2014, (2013)

38. McDowell, G.S.V., Blanchard, A.P., Figeys, D., Fai, S., Bennett, S.A.L.: Advancing Lipidomic Bioinformatic Technologies: Visualization and Phospholipid Identification (VaLID) version 3.0. International Work-Conference on Bioinformatics and Biomedical Engineering, Granada, Spain (2014)

39. Fahy, E., Subramarniam, S., Brown, H.A., Glass, C.K., Merrill Jr, A.H., Murphy, R.C., Raetz, C.R.H., Russell, D.W., Seyama, Y., Shaw, W., Shimizu, T., Spener, F., van Meer, G., VanNieuwenhze, M.S., White, S.H., Witztum, J.L., Dennis, E.A.: A Comprehensive Classification System for Lipids. Journal of Lipid Research 46, (2005)

40. Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E.A., Glass, C.K., Merrill Jr, a.H., Murphy, R.C., R.H., R.C., Russell, D.W., Subramaniam, S.: LMSD: LIPID MAPS Structure Database. Nucleic Acids Research 35, D527-D532 (2007)

41. Weininger, D.: SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. Journal of Chemical Information and Modeling 28, 31-36 (1988)

42. http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php

43. Clayden, J., Greeves, N., Warren, S., Wothers, P.: Organic Chemistry. Oxford University Press, Great Clarendon Street, Oxford (2005)

44. Gerlach, H., Laumann, V., Martens, S., Becker, C.F.W., Goody, R.S., Geyer, M.: HIV-1 Nef Membrane Association Depends on Charge, Curvature, Composition and Sequence. Nature Chemical Biology 6, 46-53 (2009)

45. Miyazaki, M., Ntambi, J.M.: Fatty Acid Desaturation and Chain Elongation in Mammals. In: Vance, D.E., Vance, J.E. (eds.) Biochemistry of Lipids, Lipoproteins and Membranes, pp. 191-211. Elsevier, Oxford (2008)

46. Lasiecka, Z.M., Yap, C.C., Vakulenko, M., Winckler, B.: Compartmentalizing the Neuronal Plasma Membrane: From Axon Initial Segments to Synapses. vol. 272. Elsevier Inc, International Review of Cell and Molecular Biology (2009)

47. Bai, F., Witzmann, F.A.: Synaptosome Proteomics. Subcellular Biochemistry 43, 77-98 (2007)

48. Gray, E.G.: Axo-Somatic and Axo-Dendritic Synapses of the Cerebral Cortex: an Electron Microscope Study. Journal of Anatomy 93, 420-433 (1959)

49. Whitehead, S.N., Hou, W., Ethier, M., Smith, J.C., Bourgeois, A., Denis, R., Bennett, S.A.L., Figeys, D.: Identification and Quantitation of Changes in the Platelet Activating Factor Family of Glycerophospholipids Over the Course of Neuronal Differentiation by High-Performance Liquid Chromatography Electrospray Ionization Tandem Mass Spectrometry. Analytical Chemistry 79, 8539-8548 (2007)

50. Ryan, S.D., Whitehead, S.N., Swayne, L.A., Moffat, T.C., Hou, W., Ethier, M., Bourgeois, A.J.G., Rashidian, J., Blanchard, A.P., Fraser, P.E., Park, D.S., Figeys, D., Bennett, S.A.L.: Amyloid-ß$_{42}$ Signals Tau Hyperphosphorylation and Compromises Neuronal Viability by Disrupting Alkylacylglycerophosphocholine Metabolism. PNAS 106, 20936-20941 (2009)

51. Whittaker, V.P., Michaelson, I.A., Kirkland, R.J.A.: The Separation of Synaptic Vesicles from Nerve-Ending Particles ('Synaptosomes'). Biochemical Journal 90, 293-303 (1964)

52. Whittaker, V.P.: Thirty Years of Synaptosome Research. Journal of Neurocytology 22, 735-742 (1993)

53. Budzinski, K.L., Sgro, A.E., Fujimoto, B.S., Gadd, J.C., Shuart, N.G., Gonen, T., Bajjaleih, S.M., Chiu, D.T.: Synaptosomes as a Platform for Loading Nanoparticles into Synaptic Vesicles. American Chemical Society Neuroscience 2, 236-241 (2011)

54. Wislet-Gendebien, S., Visanji, N.P., Whitehead, S.N., Marsilio, D., Hou, W., Figeys, D., Fraser, P.E., Bennett, S.A., Tandon, A.: Differential Regulation of Wild-Type and Mutant Alpha-Synuclein Binding to Synaptic Membranes by Cytosolic Factors. BMC Neuroscience 9, (2008)

55. Wei, J.-W., Yang, L.-M., Sun, S.H., Chiang, C.-L.: Phospholipids and Fatty Acid Profile of Brain Synaptosomal Membrane From Normotensive and Hypertensive Rats. International Journal of Biochemistry 19, 1225-1228 (1987)

56. Breckenridge, W.C., Gombos, G., Morgan, I.G.: The Lipid Composition of Adult Rat Brain Synaptosomal Plasma Membranes. Biochimica et Biophysica Acta 266, 695-707 (1972)

57. Merrill, A.H., Dennis, E.A., McDonald, J.G., Fahy, E.: Lipidomics Technologies at the End of the First Decade and the Beginning of the Next. Advances in Nutrition 4, 565-567 (2013)

58. Watanabe, K., Yasugi, E., Oshima, M.: How to Search the Glycolipid Data in "LIPID BANK for Web", the Newly Developed Lipid Database in Japan. Trends in Glycoscience and Glycotechnology 12, 175-184 (2000)

59. Fahy, E., Sud, M., Cotter, D., Subramaniam, S.: LIPID MAPS Online Tools for Lipid Research. Nucleic Acids Research 35, W606-W612 (2007)

60. Kind, T., Liu, K.-H., Lee, D.Y., DeFelice, B., Meissen, J.K., Fiehn, O.: LipidBlast *in silico* Tandem Mass Spectrometry Database for Lipid Identification. Nature Methods 10, 755-758 (2013)

61. Foster, J.M., Moreno, P., Fabregat, A., Hermjakob, H., Steinbeck, C., Apweiler, R., Wakelam, M.J.O., Vizcaíno, J.A.: LipidHome: A Database of Theoretical Lipids Optimized for High Throughput Mass Spectrometry Lipidomics. PLOS ONE 8, 1-8 (2013)

62. Fan, M., Low, H.S., Zhou, H., Wenk, M.R., Wong, L.: LipidGO: Database for Lipid-Related GO Terms and Applications. Bioinformatics 30, 1043-1044 (2014)

Appendix 1: Published papers

The full text of the papers published during this thesis follow.

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 2 2013, pages 284–285
doi:10.1093/bioinformatics/bts662

Data and text mining

Advance Access publication November 18, 2012

Visualization and Phospholipid Identification (VaLID): online integrated search engine capable of identifying and visualizing glycerophospholipids with given mass

Alexandre P. Blanchard[1,†], Graeme S. V. McDowell[1,†], Nico Valenzuela[1,2,†], Hongbin Xu[1], Sarah Gelbard[1,2], Martin Bertrand[1,2,3], Gary W. Slater[3], Daniel Figeys[1,*], Stephen Fai[3,*] and Steffany A. L. Bennett[1,*]

[1]Ottawa Institute of Systems Biology, CIHR Training Program in Neurodegenerative Lipidomics, Biochemistry, Microbiology, and Immunology, University of Ottawa, Ontario K1H 8M5, [2]Carleton Immersive Media Studio, Azrieli School of Architecture and Urbanism, Carleton University, Ontario K1S 5B6 and [3]Physics, University of Ottawa, Ontario K1H 8M5, Canada

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Establishing phospholipid identities in large lipidomic datasets is a labour-intensive process. Where genomics and proteomics capitalize on sequence-based signatures, glycerophospholipids lack easily definable molecular fingerprints. Carbon chain length, degree of unsaturation, linkage, and polar head group identity must be calculated from mass to charge (m/z) ratios under defined mass spectrometry (MS) conditions. Given increasing MS sensitivity, many m/z values are not represented in existing prediction engines. To address this need, Visualization and Phospholipid Identification is a web-based application that returns all theoretically possible phospholipids for any m/z value and MS condition. Visualization algorithms produce multiple chemical structure files for each species. Curated lipids detected by the Canadian Institutes of Health Research Training Program in Neurodegenerative Lipidomics are provided as high-resolution structures.

Availability: VaLID is available through the Canadian Institutes of Health Research Training Program in Neurodegenerative Lipidomics resources web site at https://www.med.uottawa.ca/lipidomics/resources.html.

Contacts: lipawrd@uottawa.ca

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on September 21, 2012; revised on November 4, 2012; accepted on November 6, 2012

1 INTRODUCTION

The past 10 years have seen remarkable advances in high performance liquid chromatography, electrospray ionization mass spectrometry (MS). Coupled with careful biochemistry enabling the separation of membranes and, in some cases, membrane microdomains, adaptation of these technologies to the study of lipids is permitting comprehensive phospholipid profiling at the molecular level. The emerging field of lipidomics faces three main analytical challenges: (i) a paucity of bioinformatic tools for spectral analysis; (ii) the need for accurate lipid prediction algorithms before experimenters can proceed to empirical validation; (iii) a requirement for visual tools capable of displaying all theoretically possible lipid conformations in 2D and 3D. The most comprehensive tool kit has been developed by the LIPID MAPS Consortium representing >37 000 lipid species. The majority of their curated MS data has been generated using mouse leukemic monocyte-macrophage cells and, while extensive, does not yet represent all biological lipids. To our knowledge, none of the existing open-access web-based engines have capacity to predict identity of every m/z value detected in different MS spectra (Supplementary Table S1). Visualization and Phospholipid Identification (VaLID) is a comprehensive, simple to use, resource enabling rapid prediction of these tissue-specific lipid 'unknowns'.

2 TOOL DESCRIPTION AND FUNCTIONALITY

2.1 Database content

The VaLID database contains exact and average masses (De Laeter *et al.*, 2003) for all theoretically possible: (i) phosphocholines; (ii) phosphoserines; (iii) phosphoethanolamines; (iv) glycerophosphates; (v) glyceropyrophosphates; (vi) glycerophosphoglycerols; (vii) glycerophosphoglycerolphosphates; and (viii) cytidine 5′-diphosphate 1,2-diacyl-*sn*-glycerols. The database includes *lyso*- and phospholipids with 1 to 30 carbons in each chain and up to six *cis* unsaturations. Calculations consider ester, alkyl ether and vinyl ether linkages. Numerical datasets are in Excel™ files.

2.2 Software requirements

VaLID requires a Java™-enabled Internet browser.

2.3 Using VaLID to predict lipid identity

The VaLID search engine predicts lipid identity taking into consideration user-specific MS conditions (Fig. 1). The search engine is coded in Java™. The Excel files are read with JExcelApi. Users choose exact or average mass before inputting their m/z of interest. Simple pull-down menus restrict carbon chain lengths and linkages within phospholipid classes while considering appropriate ion type. Predictions appear in two panels: (i) *Possible Lipids Include* and (ii) *Possible Isomeric Lipids Include*. The first panel returns lipid

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

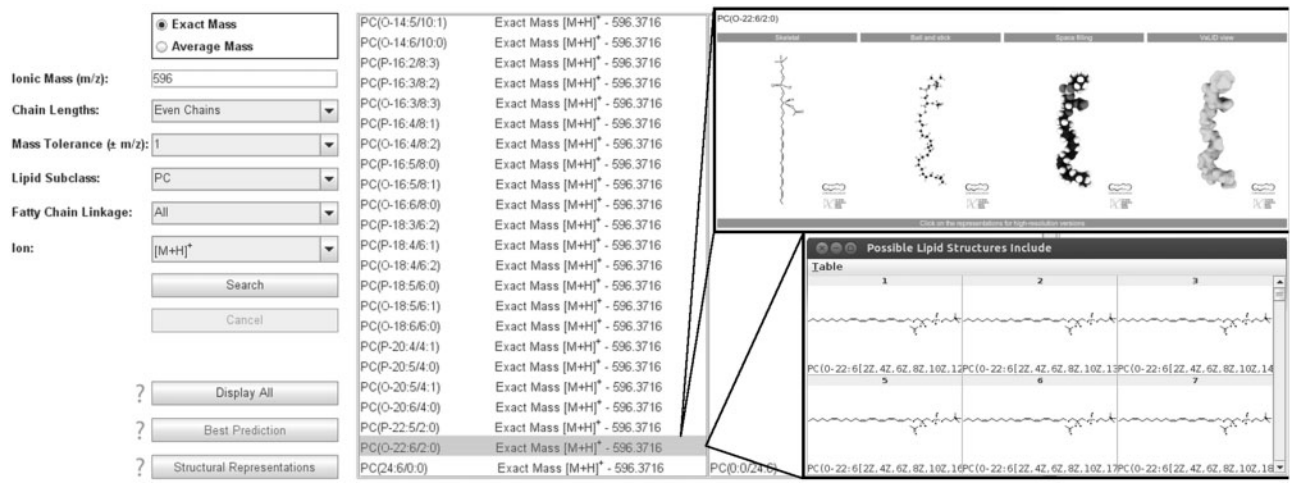


Fig. 1. The VaLID interface

identities with arbitrarily assigned *sn*-1 and *sn*-2 carbon chains. The second panel lists corresponding isomers. List position is ordered by lipid subclass (polar head group) and sorted by ascending (i) m/z; (ii) number of carbons (*sn*-1 chain); (iii) degree of unsaturation (*sn*-1 chain); (iv) number of carbons (*sn*-2 chain); and (v) degree of unsaturation (*sn*-2 chain). Nomenclature adheres to the LIPID MAPS classification system (Fahy *et al.*, 2011). To assist in decision-making, a 'best guess' feature is available whereby lipids in blue are considered 'most likely' based on the prevalence of constituent fatty acids in mammalian cells (Miyazaki and Ntambi, 2008). Lipids in red indicate that (i) the species is part of our Canadian Institutes of Health Research Training Program in Neurodegenerative Lipidomics (CTPNL)-curated database of neural lipids for which (ii) we have generated multiple high-resolution representations. Lipids in black represent theoretically possible combinations.

2.4 Displaying glycerophospholipid structures

To display all theoretically possible phospholipid conformations, users can highlight the lipid of interest and click on the 'Display All' button. The VaLID algorithm is confined to drawing cis double bonds separated by a minimum of two carbons. Using JExcelApi, VaLID calculates where each atom should be in 2D-space and generates 2D representations displayed using ChemAxon's Marvin View 5.5.1.0. Marvin View enables the user to toggle through structural views and save each species in a variety of formats. Chemical structures are drawn in accordance to the standards developed by the LIPID MAPS consortium (Fahy *et al.*, 2011). The 'Best Prediction' button displays only those lipid species found in VaLID's *Predicted to be Common* database. The 'Structural Representations' button displays lipids that are part of VaLID's curated *Structural Representations* database identified in neural tissue by CTPNL researchers. These species can be viewed in high-resolution as (i) 2D skeletal models; (ii) 3D ball and stick models; (iii) space filling models; or (iv) rendered 'VaLID view' models. VaLID view models were assembled in ChemDraw3D®, translated into a 3D model where each atom was marked by an *x*, *y* and *z* coordinate and exported into Autodesk® Maya® v2012. Rigid and dynamic models were derived using Maya® nParticles, converted into smooth polygonal meshes. These meshes were directed to the original *x*, *y* and *z* coordinates and imported as points in space to recapitulate the original molecular structure in an abstracted, organic, form. Resulting VaLID view models are available for download as rigid polygons. They are also available on request fitted with a rig of movable joints between atoms, a process typically used by graphic artists to animate human or animal characters, to facilitate membrane reconstruction and modelling.

3 CONCLUSION

VaLID is a web-based application linking a convenient search engine, a phospholipid database and multiple visualization features for identification and dissemination of large-scale lipidomic datasets. VaLID returns all theoretically possible species based on m/z and user-defined MS conditions. The user is cautioned that VaLID includes lipids (and isomeric bond configurations) that may not be biologically relevant. Investigators are encouraged to mine these lists for species most relevant to their specific biological system for subsequent validation. To assist in decision-making, a 'best guess' feature is available to focus on lipids predicted to be common based on the prevalence of the fatty acid chains in mammalian cells. Every theoretical conformation (in *cis* configuration) for each species can be viewed in 2D and 3D. Curated species can also be downloaded in multiple high-resolution representations for further visualization and model production.

Funding: This resource was funded by CIHR-MOP 89999/CIHR/Institute of Aging-TGF 96121 (to D.F., S.F. S.B.); NSERC CREATE (to D.F., G.S.); Autodesk Research, CFI (to S.F.). A.B., G.M., N.V., S.G., M.B. and H.X. received CTPNL, NSERC CREATE, FRSQ and MITACs awards.

Conflict of Interest: none declared.

REFERENCES

De Laeter,J.R. *et al.* (2003) Atomic weights of the elements. Review 2000. *Pure Appl. Chem.*, **75**, 683–800.

Fahy,E. *et al.* (2011) Lipid classification, structures and tools. *Biochim. Biophys. Acta.*, **1811**, 637–647.

Miyazaki,M. and Ntambi,J.M. (2008) Fatty acid desaturation and chain elongation in mammals. In Vance,D.E. and Vance,J.E. (eds) *Biochemistry of Lipids, Lipoproteins and Membranes*. Elsevier, Oxford, pp. 191–211.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 818670, 8 pages
http://dx.doi.org/10.1155/2014/818670

Research Article

Predicting Glycerophosphoinositol Identities in Lipidomic Datasets Using VaLID (Visualization and Phospholipid Identification)—An Online Bioinformatic Search Engine

Graeme S. V. McDowell,[1,2,3] Alexandre P. Blanchard,[1,2,3] Graeme P. Taylor,[1,2] Daniel Figeys,[2,3] Stephen Fai,[3,4] and Steffany A. L. Bennett[1,2,3]

[1] *Neural Regeneration Laboratory, Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, ON, Canada K1H 8M5*
[2] *Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, ON, Canada K1H 8M5*
[3] *CIHR Training Program in Neurodegenerative Lipidomics, Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, ON, Canada K1H 8M5*
[4] *Carleton Immersive Media Studio, Azrieli School of Architecture and Urbanism, Carleton University, ON, Canada K1S 5B6*

Correspondence should be addressed to Steffany A. L. Bennett; sbennet@uottawa.ca

Received 6 November 2013; Accepted 23 December 2013; Published 20 February 2014

Academic Editor: Tao Huang

The capacity to predict and visualize all theoretically possible glycerophospholipid molecular identities present in lipidomic datasets is currently limited. To address this issue, we expanded the search-engine and compositional databases of the online Visualization and Phospholipid Identification (VaLID) bioinformatic tool to include the glycerophosphoinositol superfamily. VaLID v1.0.0 originally allowed exact and average mass libraries of 736,584 individual species from eight phospholipid classes: glycerophosphates, glyceropyrophosphates, glycerophosphocholines, glycerophosphoethanolamines, glycerophosphoglycerols, glycerophosphoglycerophosphates, glycerophosphoserines, and cytidine 5′-diphosphate 1,2-diacyl-sn-glycerols to be searched for any mass to charge value (with adjustable tolerance levels) under a variety of mass spectrometry conditions. Here, we describe an update that now includes all possible glycerophosphoinositols, glycerophosphoinositol monophosphates, glycerophosphoinositol bisphosphates, and glycerophosphoinositol trisphosphates. This update expands the total number of lipid species represented in the VaLID v2.0.0 database to 1,473,168 phospholipids. Each phospholipid can be generated in skeletal representation. A subset of species curated by the Canadian Institutes of Health Research Training Program in Neurodegenerative Lipidomics (CTPNL) team is provided as an array of high-resolution structures. VaLID is freely available and responds to all users through the CTPNL resources web site.

1. Introduction

The emerging field of lipidomics seeks to answer two seemingly simple questions: How many lipid species are there? What effect does lipid diversity have on cellular function? To address these questions, lipidomics requires a comprehensive assessment of cellular, regional, and systemic lipid homeostasis. This assessment expands beyond lipid profiling to include the transcriptomes and proteomes of lipid metabolic enzymes and transporters, as well as that of the protein targets that affect downstream lipid signalling [1]. Lipidomic analyses also encompass an unbiased mechanistic assessment of lipid function ranging from the physicochemical basis of lipid behaviour to lipid-protein and lipid-lipid interactions triggered by intrinsic and extrinsic stimuli [1]. The first step, however, lies in identifying the molecular identities of the lipid constituents in different membrane compartments.

Recent technological advances in electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) mass spectrometry (MS), coupled to high performance liquid chromatography (LC), allow lipid diversity and membrane composition to be quantified at the molecular level [4–7]. Thousands of unique lipid species across the six major lipid structural categories in mammalian cells (fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, and prenol lipids) and two lipid categories synthesized by other organisms (saccharolipids and polyketides) can now be identified using LC-ESI-MS and, in some cases, MALDI-MS imaging [1, 4, 8]. Yet, with these successes come new challenges. Turning raw MS spectral data into annotated lipidomic datasets is a time-consuming, labour-intensive, and highly inefficient process. Predicting identities of "new" species, not previously curated, is exceedingly difficult. Lipidomic investigations lack essential bioinformatic tools capable of enabling automated data processing and exploiting the rich compositional data present in MS lipid spectra.

The critical first step is to unambiguously assign molecular identities from the MS structural information present in large lipidomic datasets [9]. Where genomics and proteomics capitalize on sequence-based signatures, lipids lack such easily definable molecular fingerprints. Identities must be reconstructed by analysis of (a) lipid mass to charge (m/z) ratios following "soft" ionization ESI and MALDI techniques and (b) defining fragmentation patterns obtained after collision-induced dissociation in various MS modes [7]. Once these molecular identities are predicted, further information about stereospecificity of critical species can then be assessed (e.g., by tandem MS, analysis of *lyso*-form fragment ions, and product ion spectral evaluation) [10–13]. For example, membrane phospholipids are derivatives of *sn*-glycero-3-phosphate with (a) an acyl, an alkyl (ether-linked plasmanyl), or an alkenyl (alkyl-1′-enyl, vinyl ether-linked plasmenyl) carbon chain at the *sn*-1 position; (b) a long-chain fatty acid that is usually esterified to the *sn*-2 position; and (c) a polar headgroup composed of a nitrogenous base, a glycerol, or an inositol unit modifying the phosphate group at the *sn*-3 position. The polar head group defines membership in one of 20 different phospholipid classes (e.g., glycerophosphoserines (PS), glycerophosphoethanolamines (PE), glycerophosphocholines (PC), glycerophosphoinositols (PI), etc.) [14]. Molecular species are further distinguished by individual combinations of carbon residues (chain length and degree of unsaturation) and the nature of each *sn*-1 or *sn*-2 chemical linkage (acyl, alkyl, or alkenyl) to the glycerol backbone. PI(18:0/22:6), for example, defines a lipid with a phosphoinositol polar head group (PI), a fully saturated 18 carbon chain (referred to as :0) ester-linked at the *sn*-1 position, and a 22 carbon chain which is characterized by six unsaturations (indicated by :6) ester-linked at the *sn*-2 position (Figure 1). Immediate PI metabolites (PIP_x) are then produced by carbon-specific phosphorylation of the PI headgroup with unique fatty acyl, alkyl, and/or alkenyl *sn*-1 and *sn*-2 chains (Figures 1 and 2). The tight regulation of PI metabolism and its critical impact on cellular function clearly underlines the importance of these compositional changes (Figure 1). Yet, to date, biological significance of the astonishing number of potentially unique PIs and PIP_xs is unknown. This is primarily due to the challenges associated with unambiguous compositional identification of PIs and PIP_xs in biological membranes [1, 15–19].

Key advances in lipidomic bioinformatics have been led by the LIPID MAPS consortium both in the development of online spectral databases and the reorganization of lipid class ontologies [14, 20]. These toolsets and classification systems have recently been complemented by the *in silico* generation of a searchable library of all theoretically possible MS/MS lipid spectra in different ionization modes (LipidBlast) [21]. Such fundamental toolkits are supported by a growing compendium of targeted spectral tools, reviewed in [6, 7, 20, 22]. Few existing bioinformatic resources, however, provide necessary information on all potential acyl chain inversions (e.g., *sn*-1 versus *sn*-2), critical phospholipid linkages that define lipid function, or theoretically possible double bond positions for every possible species. To address this need, we have developed Visualization and Phospholipid Identification (VaLID)—a web-based application linking a user-friendly online search engine, structural composition database, and multiple visualization features—that is capable of providing users with all theoretically possible phospholipids calculated from any m/z under a variety of MS conditions. VaLID version 1.0.0 was initially restricted to 736,584 unique PS, PE, PC, glycerophosphate (PA), glyceropyrophosphate (PPA), glycerophosphoglycerol (PG), glycerophosphoglycerophosphate (PGP), and cytidine 5′-diphosphate 1,2-diacyl-*sn*-glycerol (CDP-DG) identities (Table 1) [22]. At first release, we did not include the PI family or their bioactive PIP_x metabolites given the significant challenges associated with automating the visualization of all theoretically possible combinations of *sn*-1 and *sn*-2 carbon chain lengths, linkages, and variations in phosphorylation of the phosphoinositol head group. Here, we address this deficit through the development of VaLID version 2.0.0, now coded with an exhaustive PI and PIP_x database, capable of computing and visualizing a total of 1,473,168 theoretically possible phospholipids predicted from any user-inputted m/z value and MS condition. VaLID version 2.0.0 is freely available for commercial and noncommercial use at http://neurolipidomics.ca and http://neurolipidomics.com/resources.html.

2. Materials and Methods

2.1. Programming Language and Packages. VaLID version 2.0.0. was developed using Oracle's Java programming language version 6 and external Java libraries from JExcelApi and structures are displayed within the program by ChemAxon's Marvin View 5.5.1.0. software. The code was written using the IDE Eclipse Kepler, and packaged using the Fat Jar Eclipse version 0.0.31 plugin. VaLID is a web-based Java applet, and thus it requires that Java be both installed and enabled on a user's web browser. The most recent Java security update is recommended, and can be downloaded from http://www.oracle.com/technetwork/java/index.html.

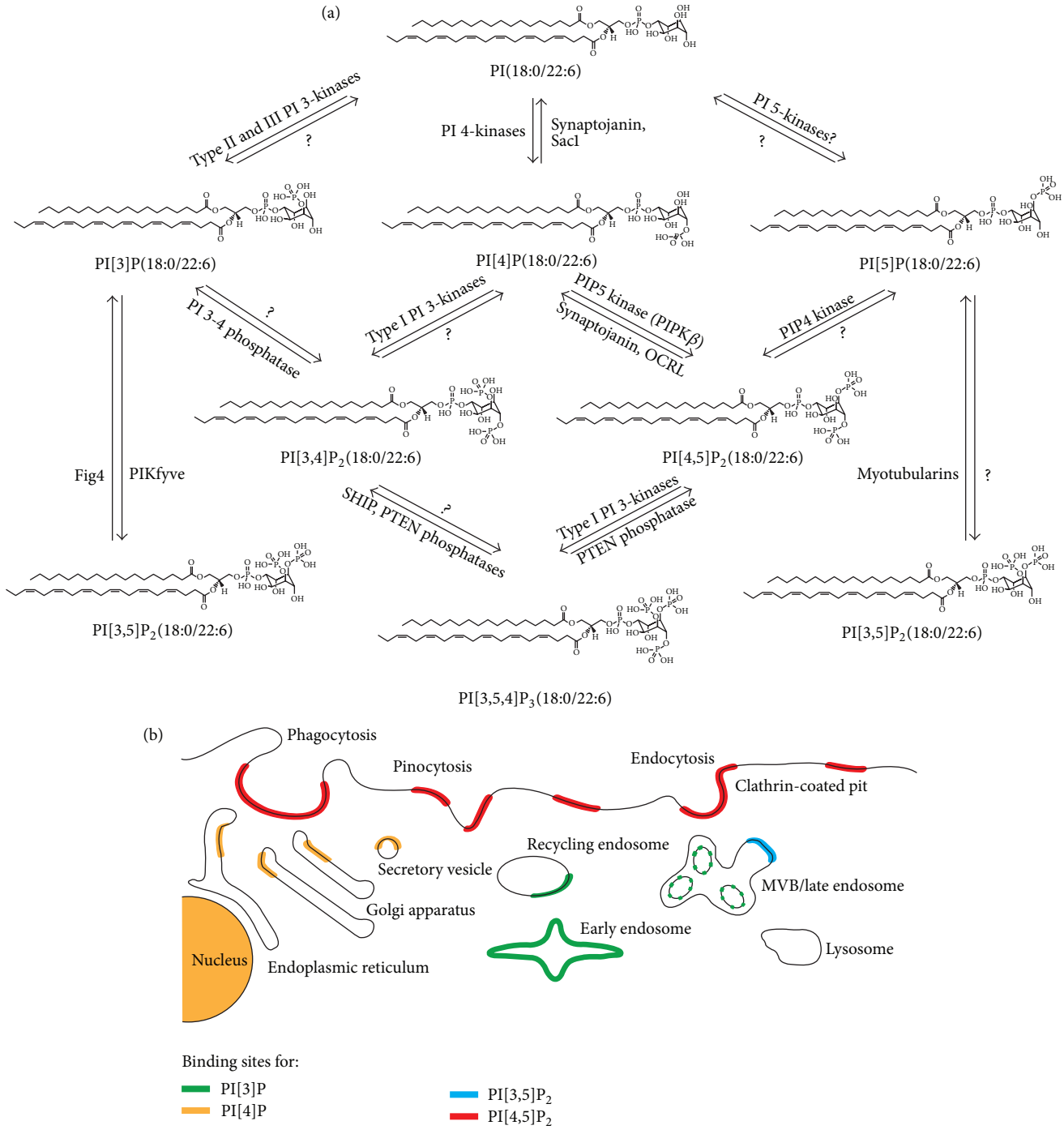


Figure 1: Glycerophosphoinositol (PI) metabolism to PI phosphates (PIP_x). (a) Metabolism of membrane PIs to bioactive PIP_x second messengers. The molecular identity of each species, defined by carbon chain length and linkage to the glycerophospholipid backbone, is predicted to affect signalling specificity in addition to known effects of PI headgroup phosphorylation. (b) Phosphorylation of PIP_x species regulates the localization of different PI-binding proteins and targets them to specific organelles (i.e., lipid-protein interaction). Phosphorylation status and carbon chain length dictate localization and likely restrict functions. Together, structural PIs and their PIP_x second messengers regulate vesicular fusion, exocytosis, and endocytosis as reviewed in (and adapted from) [2, 3].

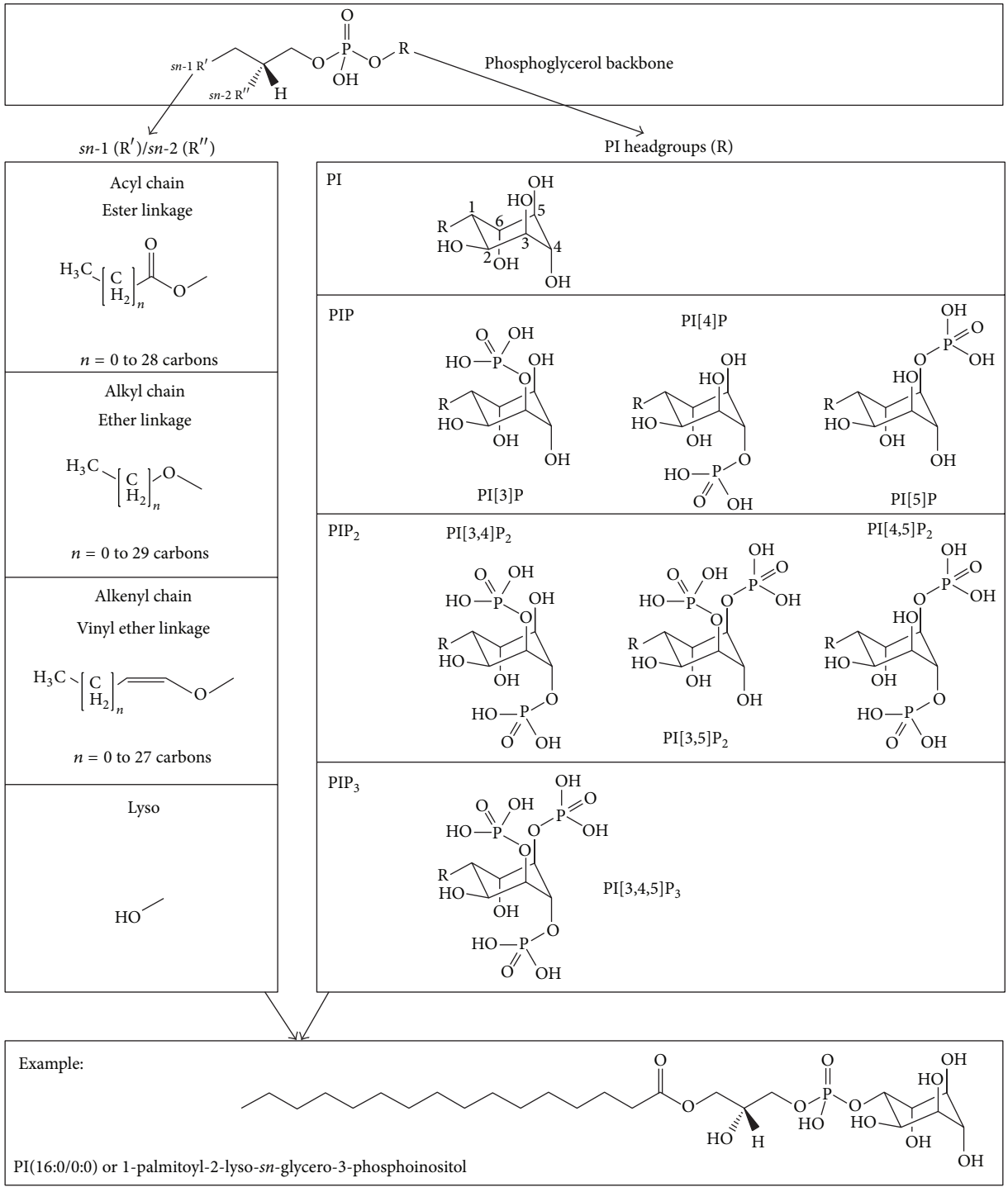


Figure 2: Component and composite structural PI and PIP_x features used to calculate masses. Exact and average masses for all theoretically possible PI and PIP_x species were calculated from the masses of every component possibility: (top panel) the phosphoglycerol backbone, (left panel) *sn*-1 and *sn*-2 hydroxyl residues (lyso-lipids) and *sn-1* and *sn-2* fatty chains ranging from 0 to 30 carbons with up to six unsaturations, considering ester, ether, or vinyl ether linkages to the phosphoglycerol backbone, and (right panel) PI polar headgroups and all biologically relevant phosphorylation possibilities. The bottom panel provides one composite PI example.

Table 1: Total number of species from each subclass that is included in VaLID.

Phospholipid subclass	LIPIDMAPS classification	Abbreviation	Number of species*
Glycerophosphates	GP10	PA	92073
Glyceropyrophosphates	GP11	PPA	92073
Glycerophosphocholines	GP01	PC	92073
Glycerophosphoethanolamines	GP02	PE	92073
Glycerophosphoglycerols	GP04	PG	92073
Glycerophosphoglycerophosphates	GP05	PGP	92073
Glycerophosphoinositols	GP06-09	PI, PIP_x	736584
Glycerophosphoserines	GP03	PS	92073
Cytidine 5′-diphosphate glycerols	GP13	CDP-DG	92073
		Total	1473168

*The calculated number of species does not include lipids formed by changing the position of the double bond beyond those represented in VaLID's structural models. Each lipid *m/z* has been calculated for exact and average masses and can be searched using even and odd carbon chains with mass tolerance ranging from ±0.0001 to ±2 and MS ion modes $[M + H]^+$, $[M + K]^+$, $[M + Li]^+$, $[M + Na]^+$, $[M - H]^-$, or [M (Neutral)].

2.2. The PI and PIP_x Compositional Database. Briefly, the underlying database contains masses of all theoretically possible PI and PIP_x species calculated from both exact and average atomic masses [23]. Component structural masses were first established for: (a) the glycerol backbone, (b) PI polar headgroups with all phosphorylation possibilities, (c) *sn*-1 and *sn*-2 hydroxyl residues (lyso-lipids), (d) *sn*-1 and *sn*-2 fatty chains ranging from 0 to 30 carbons with up to six unsaturations, considering (e) ester, ether, or vinyl ether linkages to the phosphoglyceride backbone (Figure 2). Composite masses were then calculated for every theoretically possible combination. Thus, the underlying database includes all PIs, as well as every acyl, alkyl, and alkenyl variant, for every carbon chain and double bond position, of all mono- (PIP), bis- (PIP_2), and tris- (PIP_3) phosphorylated PI headgroups modified on the hydroxyl group of carbons 3, 4, and/or 5.

2.3. PI and PIP_x Structural Visualizations. We have updated the automated representation drawing feature of VaLID to be able to draw all theoretically possible PI and PIP_x molecular identities. Structures have been restricted to display only cis double bonds separated by a minimum of two carbons. To achieve this goal, the basic structure of the PI backbone was created manually and the atom placement corrected mathematically to match known structures. Slight adjustments to atom placement were further made to improve visibility. The locations of each atom in the headgroup were then established on a Cartesian plane and coded into the software. The automated drawing feature update was integrated into the database and search functions, allowing all PI and PIP_x to be visualized on demand. Chemical structures are displayed using ChemAxon's MarvinView software (Marvin 5.5.1.0, 2011, http://www.chemaxon.com).

3. Results and Discussion

PI and PIP_x are derivatives of *sn*-glycero-3-phosphate with (a) an acyl, an alkyl (ether-linked plasmanyl), or an alkenyl (alkyl-1′-enyl, vinyl ether-linked plasmenyl) carbon chain; (b) a fatty acid commonly esterified but also with possible alkyl or alkenyl linkages to the *sn*-2 position; and (c) a polar headgroup composed of an inositol unit modifying the phosphate group at the *sn*-3 position. Individual species are distinguished by their particular combination of carbon chains (chain length and degree of unsaturation) and by the nature of their *sn*-1 or *sn*-2 chemical linkages (acyl, alkyl, or alkenyl). PI[3, 4, 5]P_3(*O*-16:0/20:4), for example, defines a lipid species with a phosphoinositol polar head group (PI) phosphorylated at the 3rd, 4th, and 5th carbon positions, an ether linkage at the *sn*-1 position (*O*-), 16 carbons at the *sn*-1, and 20 carbons at the *sn*-2 positions, of which the *sn*-1 chain is fully saturated. The number of possible structural and biochemical combinations results in colossal structural diversity; however, PIP_x lipids account for less than 15 percent of the total phospholipid composition in eukaryote cells [24]. The molecular identities of these critical species have yet to be determined in different lipidomes despite emerging evidence that differences in carbon chain length, linkage, and phosphorylation status fundamentally alter biological activity [1, 15–19] (Figure 1).

Here, we enhanced VaLID's capacity to (a) predict identities of glycerophosphoinositol species present in MS spectra from *m/z* under user-defined MS conditions and (b) automatically visualize every theoretically possible PI molecular species at given *m/z*. The updated VaLID interface, showing all of the available search terms, is presented in Figure 3. Since its inception, VaLID was designed to be a comprehensive glycerophospholipid database linking a convenient search engine with visualization features for identification and dissemination of large-scale lipidomic datasets. The intent of this tool was to aid in lipid discovery obtained through multiple MS methodologies and significantly reduce the time required to validate critical phospholipid identities present in target lipidomes. The program initially contained eight phospholipid subclasses, excluding the PI subfamily. In VaLID version 2.0.0, this capacity is now expanded to all theoretically possible PI and PIP_x glycerophospholipids and comprises a total of 1,473,168 unique structures.

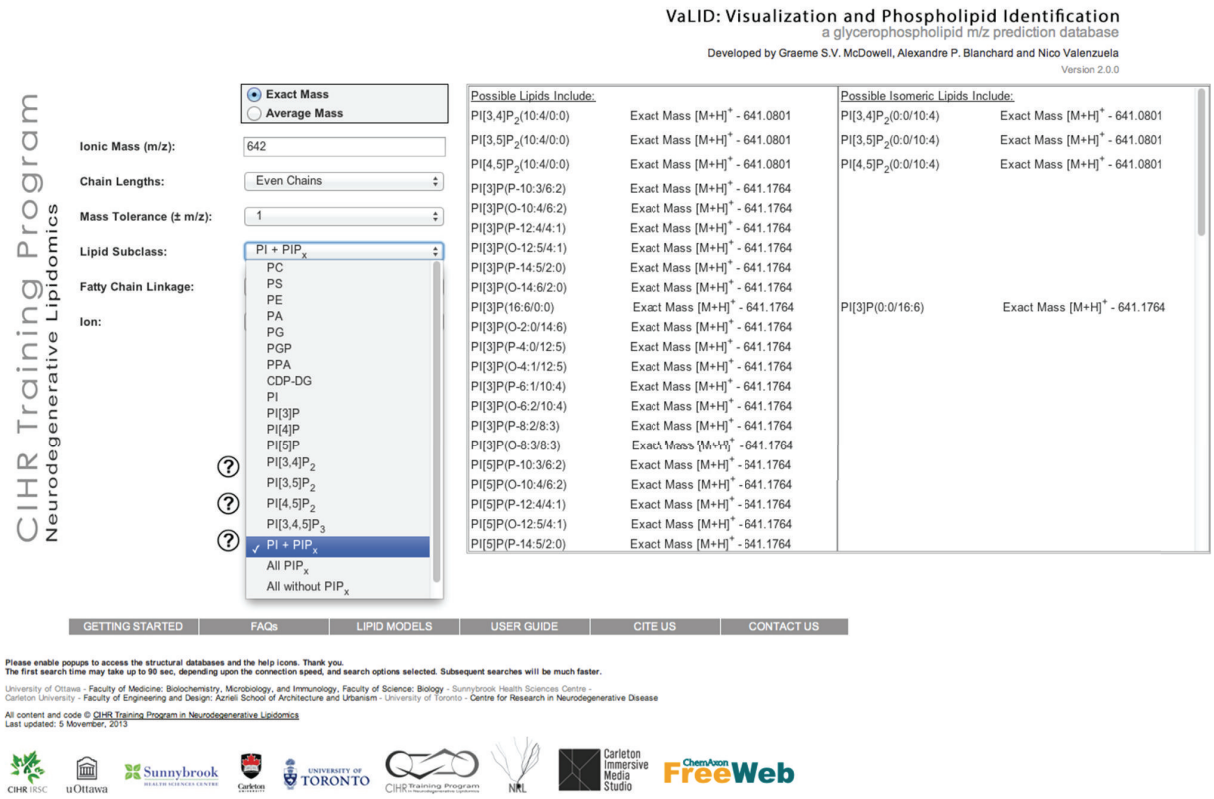


Figure 3: VaLID 2.0.0 interface. The new PI and PIP_x search options are shown for the VaLID interface's drop down menu. Each member of the PI family can be searched individually, as well as in various combinations. PIP_x refers to phosphoinositol mono-, bis-, or tris-phosphate, and can be searched with, or without, PIs.

These additions are meant to provide lipidomic researchers with the additional tools necessary to mine their lipidomes for PI and PIP_x species with specific *m/z* under their particular MS experimental conditions including the ion mode and the lipid subclass. Due to the complexity of the PI superfamily, and to accelerate searching, users can restrict searches to subclasses (PI, PIP_x) or sub-subclasses (PI, PI[3]P, PI[4]P, PI[5]P, PI[3, 4]P_2, PI[3, 5]P_2, PI[4, 5]P_2, PI[3, 4, 5]P_3). For example, if the option PI[3, 4]P_2 is chosen, all molecular species with an inositol backbone phosphorylated only at the 3rd and 4th carbon positions will be provided and VaLID will not return any related PI[3, 5]P_2 or PI[4, 5]P_2 species. The PI +PIP_x option restricts searches to the entire PI superfamily excluding other phospholipid families. The "All without the PIP_x" option returns all of the phospholipids in the database including PI structural precursors with the exception of PIP_x metabolites. Finally, the "All" option returns results from every headgroup. When more than one headgroup is being searched, the program will let the user know how many headgroups have been loaded, and how many are remaining to be loaded.

With respect to the visualization features for PIP or PIP_2, the program will draw the phosphate groups on the inositol ring in the locations that the user specified from the dropdown menu for lipid species selected. As with the other subclasses, choosing the "Display All" button will draw all the theoretically possible structures associated with the selected lipid name. Potential variants in degrees of unsaturation are drawn sequentially in every location along the fatty acid chain, separated by at least two carbons, and in cis configuration. An example of this can be seen in Figure 4. If the selected lipid meets criteria for the "Best Prediction," selecting this option will return only the lipids in VaLID's "Predicted to be Common" database. These species are categorized based on the relative abundance of prevalent fatty acid chains in mammalian cells [25].

4. Conclusions

VaLID is, to our knowledge, the first search engine that has an exhaustive *m/z* and visualization database of all the theoretically possible glycerophospholipids updated here from eight to twelve of the twenty phospholipid subclasses defined by the LIPID MAPS Consortium [26]. The purpose of this update is to facilitate prediction and visualization of the identities of all unknown species, now including all PIs and their metabolites, with given *m/z* and MS condition that may be present in users' lipidomes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

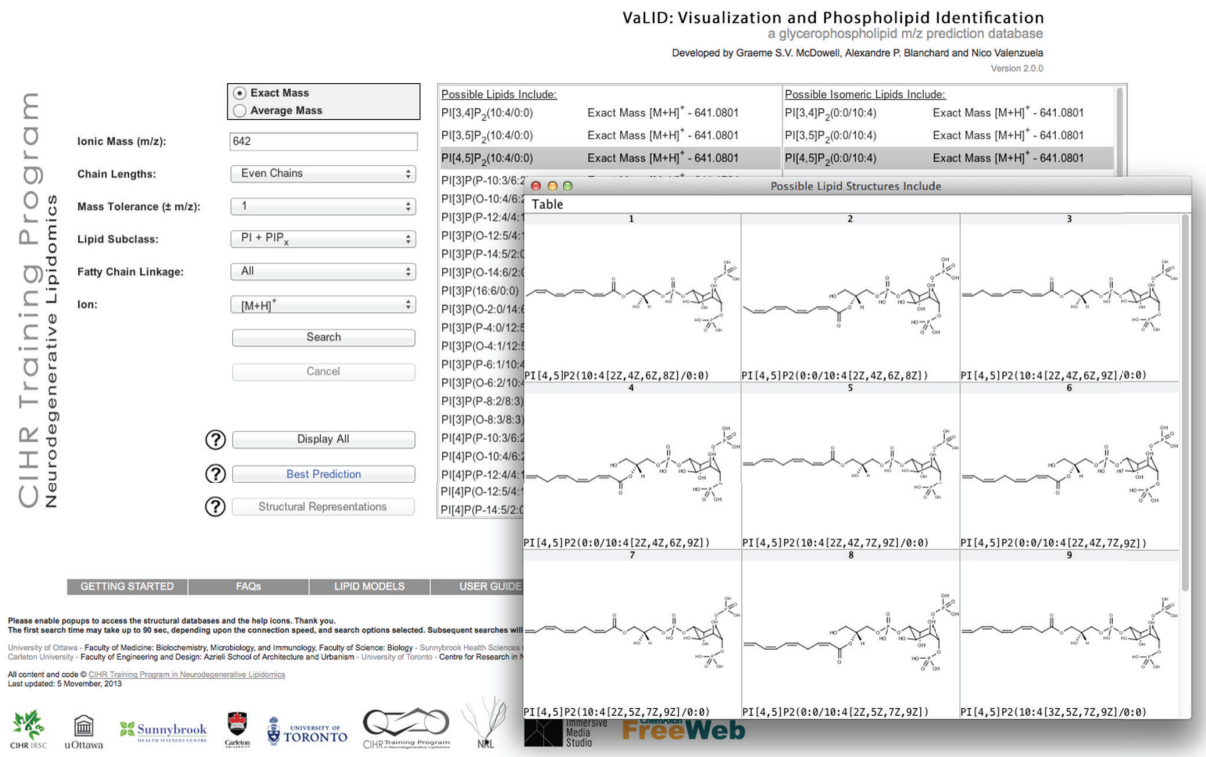


Figure 4: Automated drawing feature of VaLID 2.0.0. An example of a search button, returning all possible PI and PIP_x lipids with *m/z* of 642 (exact mass with a user-defined tolerance of 1 amu), restricted to displaying even carbon chains only, and selecting $[M+H]^+$ ion mode in MS (back panel). The user then selected PI[4, 5]P_2(10:4/0:0) and its *sn*-1/*sn*-2 chain inversion species and pressed "Display All" button. The window labelled "Possible Lipid Structures Include" displays a table containing the possible structures for this lipid, with the restrictions as laid out in the user manual (inset). These drawings can be easily exported for use in publication figures as described in the user manual.

Authors' Contribution

Graeme S. V. McDowell and Alexandre P. Blanchard contributed equally to this work.

Acknowledgments

This resource was funded by the Canadian Institutes of Health Research (CIHR) MOP 89999 to DF and SALB and a Strategic Training Initiative in Health Research (STIHR) CIHR/Training Program in Neurodegenerative Lipidomics (CTPNL) and the Institute of Aging TGF 96121 to DF, SF, and SALB. APB received a FRSQ and CTPNL graduate scholarship; GSVM received a CTPNL graduate scholarship.

References

[1] S. A. L. Bennett, N. Valenzuela, H. Xu et al., "Using neurolipidomics to identify phospholipid mediators of synaptic (dys)function in Alzheimer's Disease," *Frontiers in Physiology*, vol. 4, p. 168, 2013.

[2] C. Le Roy and J. L. Wrana, "Clathrin- and non-clathrin-mediated endocytic regulation of cell signalling," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 2, pp. 112–126, 2005.

[3] L. C. Skwarek and G. L. Boulianne, "Great Expectations for PIP: phosphoinositides as regulators of signaling during development and disease," *Developmental Cell*, vol. 16, no. 1, pp. 12–20, 2009.

[4] D. Piomelli, G. Astarita, and R. Rapaka, "A neuroscientist's guide to lipidomics," *Nature Reviews Neuroscience*, vol. 8, no. 10, pp. 743–754, 2007.

[5] H. A. Brown and R. C. Murphy, "Working towards an exegesis for lipids in biology," *Nature Chemical Biology*, vol. 5, no. 9, pp. 602–606, 2009.

[6] M. Bou Khalil, W. Hou, H. Zhou et al., "Lipidomics era: accomplishments and challenges," *Mass Spectrometry Reviews*, vol. 29, no. 6, pp. 877–929, 2010.

[7] H. Xu, N. Valenzuela, S. Fai et al., "Targeted lipidomics—advances in profiling lysophosphocholine and platelet-activating factor second messengers," *FEBS Journal*, vol. 280, pp. 5652–5667, 2013.

[8] X. Han, K. Yang, and R. W. Gross, "Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses," *Mass Spectrometry Reviews*, vol. 31, no. 1, pp. 134–178, 2012.

[9] P. S. Niemelä, S. Castillo, M. Sysi-Aho, and M. Orešič, "Bioinformatics and computational methods for lipidomics," *Journal of Chromatography B*, vol. 877, no. 26, pp. 2855–2862, 2009.

[10] W. Hou, H. Zhou, M. B. Khalil, D. Seebun, S. A. L. Bennett, and D. Figeys, "Lyso-form fragment ions facilitate the determination of stereospecificity of diacyl glycerophospholipids," *Rapid Communications in Mass Spectrometry*, vol. 25, no. 1, pp. 205–217, 2011.

[11] J. C. Smith, W. Hou, S. N. Whitehead, M. Ethier, S. A. L. Bennett, and D. Figeys, "Identification of lysophosphatidylcholine (LPC) and platelet activating factor (PAF) from PC12 cells and mouse cortex using liquid chromatography/multistage mass spectrometry (LC/MS3)," *Rapid Communications in Mass Spectrometry*, vol. 22, no. 22, pp. 3579–3587, 2008.

[12] S. N. Whitehead, W. Hou, M. Ethier et al., "Identification and quantitation of changes in the platelet activating factor family of glycerophospholipids over the course of neuronal differentiation by high-performance liquid chromatography electrospray ionization tandem mass spectrometry," *Analytical Chemistry*, vol. 79, no. 22, pp. 8539–8548, 2007.

[13] C.-H. Tang, P.-N. Tsao, C.-Y. Chen, M.-S. Shiao, W.-H. Wang, and C.-Y. Lin, "Glycerophosphocholine molecular species profiling in the biological tissue using UPLC/MS/MS," *Journal of Chromatography B*, vol. 879, no. 22, pp. 2095–2106, 2011.

[14] E. Fahy, D. Cotter, M. Sud, and S. Subramaniam, "Lipid classification, structures and tools," *Biochimica et Biophysica Acta*, vol. 1811, no. 11, pp. 637–647, 2011.

[15] U. Igbavboa, J. Hamilton, H.-Y. Kim, G. Y. Sun, and W. G. Wood, "A new role for apolipoprotein E: modulating transport of polyunsaturated phospholipid molecular species in synaptic plasma membranes," *Journal of Neurochemistry*, vol. 80, no. 2, pp. 255–261, 2002.

[16] M. J. Sharman, G. Shui, A. Z. Fernandis et al., "Profiling brain and plasma lipids in human apoe ε2, ε3, and ε4 knock-in mice using electrospray ionization mass spectrometry," *Journal of Alzheimer's Disease*, vol. 20, no. 1, pp. 105–111, 2010.

[17] R. B. Chan, T. G. Oliveira, E. P. Cortes et al., "Comparative lipidomic analysis of mouse and human brain with Alzheimer disease," *The Journal of Biological Chemistry*, vol. 287, no. 4, pp. 2678–2688, 2012.

[18] P. H. Axelsen and R. C. Murphy, "Quantitative analysis of phospholipids containing arachidonate and docosahexaenoate chains in microdissected regions of mouse brain," *Journal of Lipid Research*, vol. 51, no. 3, pp. 660–671, 2010.

[19] S. Osawa, S. Funamoto, M. Nobuhara et al., "Phosphoinositides suppress γ-secretase in both the detergent-soluble and -insoluble states," *The Journal of Biological Chemistry*, vol. 283, no. 28, pp. 19283–19292, 2008.

[20] E. Fahy, D. Cotter, R. Byrnes et al., "Bioinformatics for Lipidomics," *Methods in Enzymology*, vol. 432, pp. 247–273, 2007.

[21] T. Kind, K. H. Liu, Y. Lee do et al., "LipidBlast in silico tandem mass spectrometry database for lipid identification," *Nature Methods*, vol. 10, no. 8, pp. 755–758, 2013.

[22] A. P. Blanchard, G. S. McDowell, N. Valenzuela et al., "Visualization and Phospholipid Identification (VaLID): online integrated search engine capable of identifying and visualizing glycerophospholipids with given mass," *Bioinformatics*, vol. 29, no. 2, pp. 284–285, 2013.

[23] J. R. De Laeter, J. K. Böhlke, P. De Bièvre et al., "Atomic weights of the elements: review 2000," *Pure and Applied Chemistry*, vol. 75, no. 6, pp. 683–800, 2003.

[24] G. Di Paolo and P. De Camilli, "Phosphoinositides in cell regulation and membrane dynamics," *Nature*, vol. 443, no. 7112, pp. 651–657, 2006.

[25] M. Miyazaki and J. M. Ntambi, "Fatty acid desaturation and chain elongation in mammals," in *Biochemistry of Lipids, Lipoproteins and Membranes*, D. E. Vance and J. E. Vance, Eds., pp. 191–211, Elsevier, 2008.

[26] E. Fahy, S. Subramaniam, R. C. Murphy et al., "Update of the LIPID MAPS comprehensive classification system for lipids," *Journal of Lipid Research*, vol. 50, pp. S9–S14, 2009.

Advancing Lipidomic Bioinformatic Technologies: Visualization and Phospholipid Identification (VaLID) version 3.0

Graeme S.V. McDowell[1], Alexandre P. Blanchard[1], Daniel Figeys[1], Stephen Fai[2], and Steffany A.L. Bennett[1]

[1] Ottawa Institute of Systems Biology, Neural Regeneration Laboratory, Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, K1H 8M5, Canada.
{gmcdo092, apbla037, dfigeys, sbennet}@uottawa.ca
[2] Carleton Immersive Media Studio, Azrieli School of Architecture and Urbanism, Carleton University, Ontario, K1S 5B6, Canada.
sfai@cims.carleton.ca
http://neurolipidomics.com/resources.html

Abstract. There is a paucity of bioinformatic tools for spectral analysis capable of assigning and visualizing molecular identities from mass spectrometry-derived structural information. Predicting phospholipid lipid identities is a labour-intensive process given the extreme variability in structure based on permutations of only a few atomic 'building blocks'. Moreover, our ability to visualize all theoretically possible phospholipids present in lipidomic datasets is limited. To address this gap, we created the online lipidomic bioinformatic tool Visualization and Phospholipid Identification (VaLID). This work describes the expansion of this tool to include functional search capacity linking the VaLID 3.0 database of 1,324,224 theoretically possible phospholipids to PubMed and the Human Metabolome Database (HMDB) as well as the inclusion of lipid predictions using nomenclatures now useful for researchers employing a shotgun lipidomic approach. VaLID is freely available at http://neurolipidomics.com/resources.html.

Keywords: Phospholipids • Lipidomics • Database • Mass Spectrometry

1 Introduction

The emerging field of lipidomics seeks to understand how dynamic changes in membrane composition regulate cell function [1]. Neurolipidomics is the study of cellular, regional, and systemic lipid homeostasis in the central nervous system encompassing not only the identification and measurement of individual lipid isoforms but also the mRNA and protein expression profiles of metabolic enzymes and transporters, and the protein targets that affect downstream signalling [1]. Furthermore, a lipidomic analysis includes an unbiased assessment of lipid function ranging from the physico-

chemical basis of lipid behaviour to lipid-protein and lipid-lipid interactions, and the impact of dynamic lipid metabolism on cellular response to intrinsic and extrinsic stimuli [1]. Lipidomics, the systems-level analysis of lipids and their interacting moieties [2], depends upon our ability to answer two seemingly simple questions: How many lipid species are there? And what effect does lipid diversity have on cellular function? Recent advances in high performance liquid chromatography (LC), electrospray ionization (ESI), and matrix-assisted laser desorption ionization (MALDI) mass spectrometry (MS), coupled with new membrane separation and extraction methodologies, now provide us with the means to quantify lipid diversity [3-6]. For example, LC-ESI-MS enables researchers to determine not only phospholipid subclass, but also the number of carbons and the possible number of unsaturations in the fatty acid chains linked by acyl, akyl, or alkenyl linkages to the phospholipid backbone. These and other technological capacities have revived the idea of a highly dynamic membrane, and brought the focus of membrane biology back to their lipid constituents, with determinative roles of lipids in biological processes, such as potent second messenger molecules becoming more apparent [7]. Hundreds to thousands of lipid species can now be profiled in different subcellular membrane compartments [1]. Basic unitary conceptions are being challenged. Diacylglycerol, commonly conceived by cell biologists as a single lipid, is now recognized to be a family of over 50 structurally distinct species each controlling different cellular processes [8, 9].

Yet with this success come new challenges. The lipidomics field faces four formidable roadblocks as elegantly expounded by Niemela et al., [10] : ***(1) Data Processing and Lipid Identification*** – There is a paucity of bioinformatic tools for spectral analysis capable of assigning and visualizing molecular identities from MS-derived structural information; ***(2) Statistical Analysis*** – Lipidomic datasets involve "medium-scale" data ranging from tens to hundreds of lipids per family that, in turn, impact on other related medium-sized lipid networks (i.e., requiring analysis of thousands of "features"). These sizes are difficult to analyze using traditional statistical approaches that classically consider features of less than ten yet are not amenable to genomic/proteomic statistical methodologies where features exceed thousands; ***(3) Pathway Analysis*** – We lack accessible curated databases capable of predicting biochemical, signalling, and regulatory lipid pathways affected by changes in membrane composition; (***4) Modeling Tools*** – We have little to no capacity to rapidly model the impact of altering lipid composition on biological membrane properties and cellular signalling (i.e., lipid interactome pathway analysis).

There is pressing need for bioinformatic interest in lipidomics. We simply do not have the necessary tools to mine our new rich lipid compositional datasets for functional significance. Lipidomics has yet to benefit from the same development of analytical tools available to proteomics and genomics through concerted bioinformatic efforts. The goal of our work is to develop new tools to address the first and third challenges (Lipid Identification and Pathway Analysis). As our technology improves, more precise lipid identification becomes possible [11]. A major issue lies in the identification of isobaric lipids in complex matrices – species with identical mass to charge (m/z) ratios. An example of this property can be seen in Figure 1. Genomics and proteomics can capitalize on sequence-based signatures; lipids lack easily defina

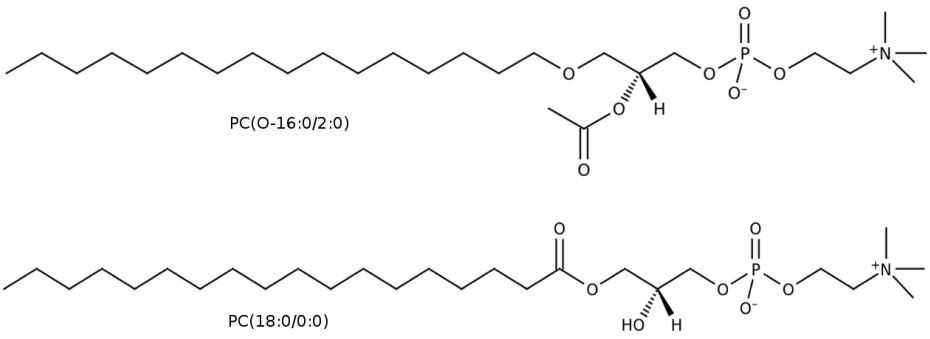


Fig. 1. Isobaric lipids. Both these lipids have the same mass to charge ratio of 523.3638 in $[M+H^+]$ mode where M is mass. Both lipids share a phosphocholine polar headgroup linked to the glycerophospholipid backbone by a phosphor-diester bond. However, the top lipid, a platelet activating factor alkylacylglycerophosphocholine, is defined by a different structure in both carbon chain identity and linkage to the glycerophospholipid backbone, activating different receptors and signalling pathways than the bottom lipid, a lyso-phosphatidylcholine.

ble molecular fingerprints. Identities must be established from structural determinants. Here, bioinformatic tools are urgently needed. For example, the LIPID metabolites and pathways strategy (LIPID MAPS) consortium[1], and LipidBank[2] have created tools that aid in the identification of lipids via MS, and have led the field in standardizing ontologies, notation, and protocols [11, 12]. In the case of LipidMAPS, their lipid databases were created from lipids curated from both literature sources and data generated by their own core laboratories and those of their partners each using specific cell systems [13]. While these databases and the associated search tools are invaluable, not all lipids are represented. Our group, working with neural membranes, has found that many species in our spectra were not present in the LipidMAPS structural databases. Thus, to participate in advancing lipidomic bioinformatics, we created a database and online prediction engine which was designed to be entirely comprehensive – Visualization and Phospholipid Identification (VaLID)[3] [14]. VaLID is web-based application linking a user-friendly search engine to an exhaustive database which contains all theoretically possible phospholipid species combined with different drawing and visualization features. In its first and second iteration, it was designed for use by researchers employing LC-ESI-MS technologies to explore lipid diversity. Here, we describe its expansion to include functional search capacity linking VaLID 3.0 databases to both PubMed[4] and the Human Metabolome Database (HMDB)[5] as

1 www.lipidmaps.org
2 http://lipidbank.jp
3 http://neurolipidomics.com/resources.html
4 http://www.ncbi.nlm.nih.gov/pubmed/
5 http://www.hmdb.ca

well as the expansion of VaLID's prediction engine to include lipid prediction using nomenclatures now useful for researchers employing a shotgun lipidomic approach.

2 Program Description

VaLID is composed of three parts: (1) a group of comprehensive phospholipid databases, (2) multiple search engines enabling lipid prediction, and now basic functional annotation, and (3) various drawing options for lipids within the database. The program is constantly being updated with new bioinformatic features in response to community feedback. We describe here the addition of linking the search engine and database with capacity to query PubMed and the HMDB for known functions and physicochemical properties of target species. To aid researchers employing shotgun lipidomic MS methodologies, we also describe addition of additional lipid nomenclatures to lipid prediction and functional query, defining species by headgroup and total number of carbons, as well as the total number of unsaturations in prediction search.

2.1 Database

The backbone for VaLID is its phospholipid database. Here, we expanded these databases to make the program amenable not only to lipidomic researchers employing LC-ESI-MS technologies but also shotgun lipidomics. To facilitate this transition, the original database was split into separate units, one per phospholipid subclass. The databases were created to be comprehensive (i.e., to contain all the possible lipids within their specific phospholipid subclasses) with one restriction. Species are limited to the realm of biological possibility as they contain only phospholipids with carbon chains ranging in length from 0 to 30 carbons at the *sn*-1 and *sn*-2 positions with combinations of up to six unsaturations in the cis position. The databases also contain species with three different chain linkage options, alkyl, acyl, and alkenyl, representing an ether, ester and vinyl ether bonds, respectively. The average and exact masses for every combination of chain *sn*-1 and *sn*-2 chain lengths, with each of the appropriate linkages are calculated, and added to the exact or average mass [15] of a defining headgroup. Here, the exact mass represents the isotopic mass of the most prevalent isotope of the atom of interest, whereas the average mass is the weighted average of the isotopic masses, given multiple isotopic masses.

The first version of VaLID released in early 2013 contained eight phospholipid subclasses, glycerophosphates (PA), glyceropyrophosphates (PPA), glycerophosphocholines (PC), glycerophosphoethanolamines (PE), glycerophosphoglycerols (PG), glycerophosphoglycerophosphates (PGP), glycerophosphoserines (PS) and cytidine 5'-diphosphate 1,2-diacyl-*sn*-glycerols (CDP-DG). This represented approximately 736,000 unique lipid species [14]. In late 2013, it was coded to include four more phospholipid groups: glycerophosphoinositols (PI), glycerophosphoinositol monophosphates (PIP), glycerophosphoinositol bisphosphates (PIP_2) and glycerophosphoinositiol trisphosphates (PIP_3) and updated to be capable of visualizing a total 1,324,224 theoretically possible phospholipids predicted from any user-inputted m/z

value and MS condition [16]. Now, we describe additional search features and linkage to functional annotation databases.

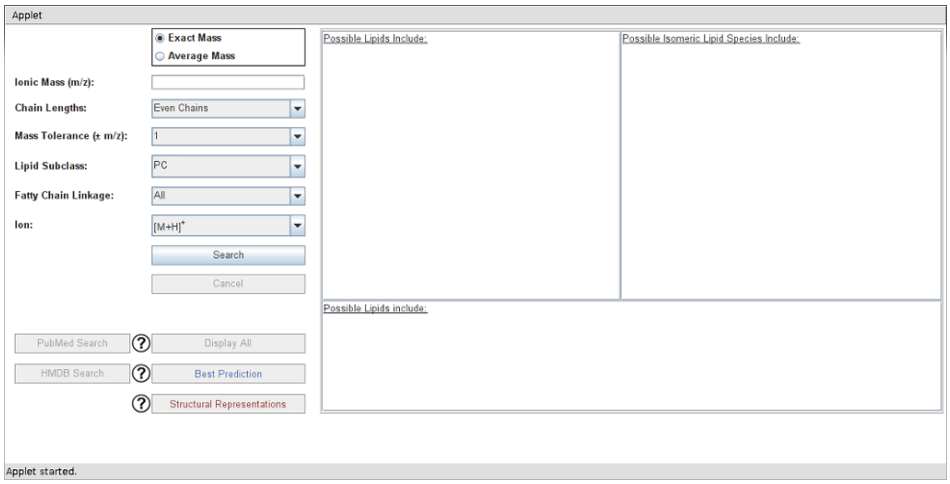


Fig. 2. Graphical Interface of VaLID 3.0

2.2 Search Function

VaLID's graphical interface (Figure 2) allows users to access the databases and return only entries tailored to their particular predetermined specifications (Figure 3). Lipid nomenclature adheres to that developed by the LipidMAPS consortium [12, 17]. Like VaLID 1.0 and 2.0, VaLID 3.0 has multiple selectable fields enabling users to customize their searches based on their particular MS methodologies and prediction requirements: exact or average mass, the ionic mass, even or odd chains, mass tolerance, lipid subclass, fatty chain linkage, and selected ion mode. Users first choose results returned for average or exact mass. The m/z ratio of lipid to be predicted is inputted in the field labelled ionic mass. The mass tolerance represents a range, from 0.0001 to 2 m/z, above and below the desired mass, to accommodate the limit of sensitivity of the user's mass spectrometer. Searches can be restricted to phospholipids with an even number, odd number, or both, carbon chain length at *sn*-1 and *sn*-2 positions. Particular lipid subclasses are defined by target phospholipid headgroup and various combinations of headgroup modification options (Figure 4). Users can specify the type and combination of fatty chain linkages that they wish to search (i.e., an ester (acyl), ether (alkyl) bond at one or both chains, a vinyl ether (alkenyl) at the *sn*-1 position, or all of these options). The MS ion mode employed by the user is inputted, enabling VaLID to predict lipid identity based on appropriate mass. We introduce in VaLID 3.0 a feature to search lipid function and properties within both PubMed and the HMDB (Figure 3). A third box for lipid results, using nomenclature defining headgroup and total number of carbons and possible total number of unsaurations in both chains encompassing all of the lipids returned in the top two boxes, is now returned for users wishing to predict species using these terms (Figure 3).

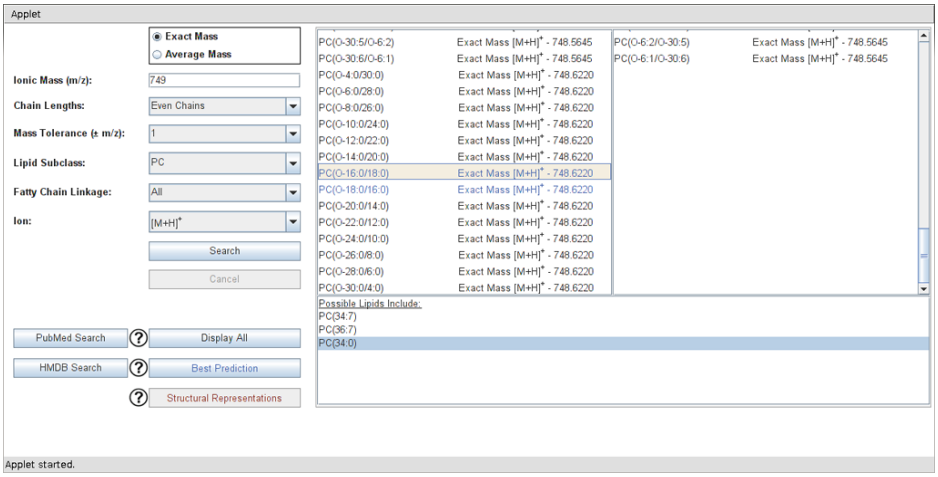


Fig. 3. An example of a search for phosphocholines with mass of 749 within a tolerance of ±1 m/z, selecting for all linkages, but restricted to even chains, from a spectra collected in $[M+H^+]$ ion mode. Possible isomeric lipids (where *sn*-1 and *sn*-2 chains are in alternate positions) theoretically possible are shown in the box on the right hand side. New to version 3.0 are the results labelled "Possible Lipids Include", containing a list of the lipid subclasses, as well as the total number of carbons in both chains, and the number of unsaturations that represent the sum total of all potential species returned in the top right and left boxes. The PubMed and HMDB search buttons are also visible and enabled once a target species is selected in one of the three result boxes.

2.3 Drawing Feature

Every lipid within the database can be drawn via VaLID's visualization feature with curated high-resolutions species identified in neural tissue by our group further presented in a series of rigid models produced using Maya® nParticles as we have described [14]. For example, PC(18:0/18:1) – that is, a glycerophosphocholine with a fatty chain of 18 carbons in length which is fully saturated, in one position, and a fatty chain, 18 carbons long with one unsaturation in the other – has many structural possibilities. Additionally, the predicted species could have *sn*-1 and *sn*-2 chains reversed. The unsaturation in one of the chains could be at any carbon along the chain. The increase in the number of unsaturations and chain length greatly increases the number

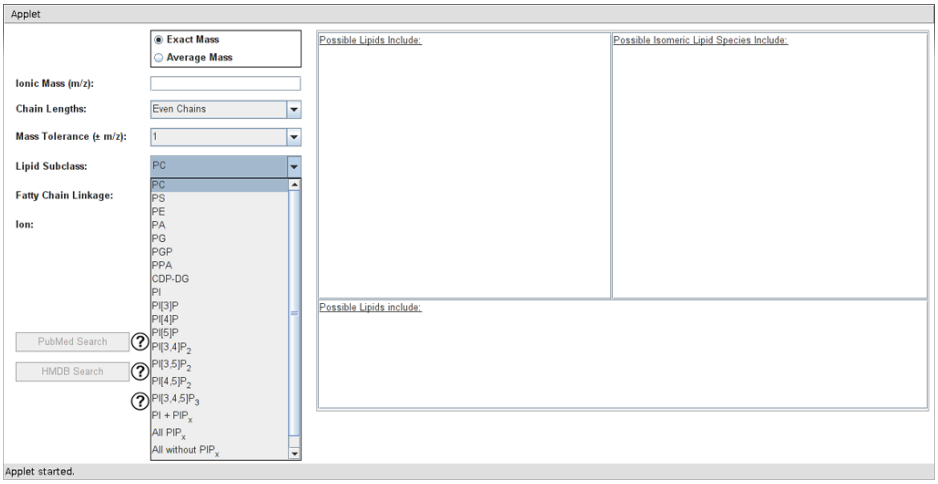


Fig. 4. The phospholipid subclasses searchable in VaLID 3.0.

of possible structures for a given lipid, representing a bioinformatic challenge to phospholipid representation. As such, we created a drawing algorithm that calculates, then draws, the atom location and bond connectivity in a 2D Cartesian plane. The results are displayed using ChemAxon's MarvinView software (Figure 5). The structures were created to match the drawing specifications laid out by the LipidMAPS consortium [12]. Using the features of MarvinView, the user can download any of the structures individually, and save them as a molecular model (.mol) file, which can be opened in multiple chemical drawing tools, such as Marvin or ChemDraw, or saved as an image, such as a PNG file. As described previously [14, 16], there are multiple drawing options available for users within VaLID's drawing component. If a lipid within the "Possible Lipids Include" or the "Possible Isomeric Lipids Include" boxes is selected, and the "Display All" button is pressed, every combination of unsaturations, as described above, will be calculated and drawn. Lipids can also appear in blue or red text. Should a lipid appear in blue font, there is a structure that can be considered more common, or "more likely" to occur. If one of these lipids is highlighted, and the "Best Prediction" button is pressed, the common structure(s) of each individual chain is drawn, as opposed to every combination. For example, if the lipid selected contains "20:4", the *Best Prediction* would return the structure of arachidonic acid, with corresponding positions of double bonds. These fatty chains, with common structures, were identified based on their relative abundance in mammalian cells [16]. Lipids that are returned in red font are lipids which have been curated by the CIHR Training Program in Neurodegenerative Lipidomics (CTPNL) in neural tissue and we provide high-resolution 3D models (VaLID view models) for download by highlighting the lipid and pressing the "Structural Representation" button. These models are derived using Maya® nParticles, converted into smooth polygonal meshes directed to the original x, y, and z coordinates, imported as points in space, recapitulating the original molecular structure developed using MarvinView as described above in an abstracted, organic, form. Resulting VaLID view models are available for download

as rigid polygons. They are also available on request fitted with a rig of movable joints between atoms to facilitate membrane reconstruction and modeling.

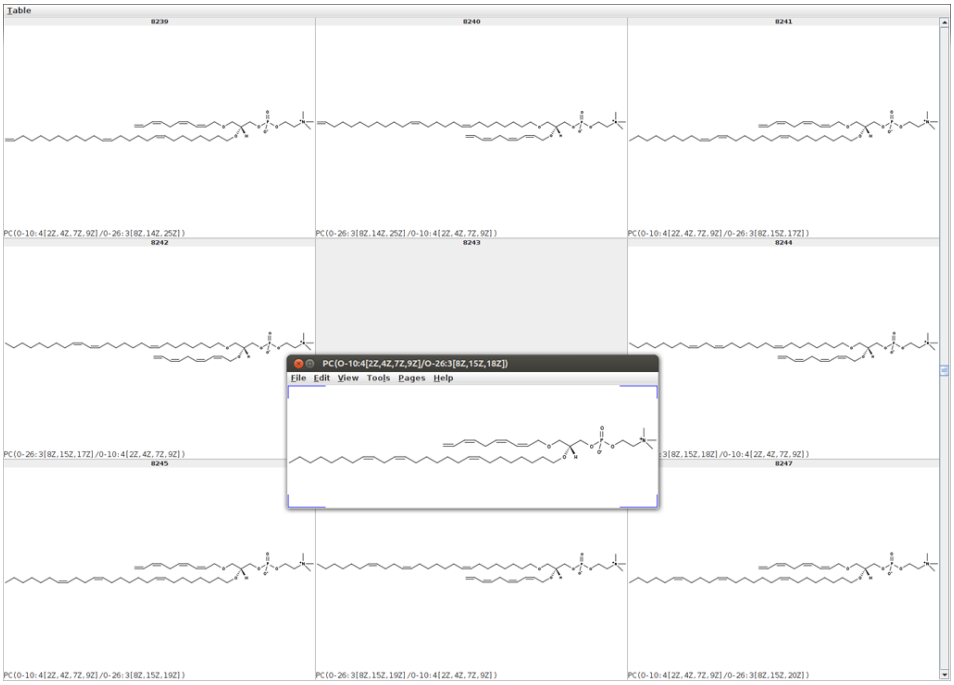


Fig. 5. When the Display All, or Best Prediction buttons are pressed, the structures of the selected lipid are calculated. When the calculations for the structures are completed, a new window pops up containing ChemAxon's MarvinView software, containing a table of all the structures. Each one of these structures can be viewed independently, where they can be saved as a molecular (.mol) file, or as an image.

3 Conclusion

The program VaLID was created initially to aid lipid researchers in predicting lipid identity from LC-ESI-MS spectra. To our knowledge, it is the first lipidomic MS prediction tool that contains all the chemically possible phospholipids up to 30 carbons in both chains, covering twelve subclasses. We describe here the addition in version 3.0 of features making VaLID's prediction capacity useful for researchers employing shotgun lipidomic approaches as well as connection of VaLID's databases to the HMDB and PubMed databases. This additions allow researchers to search existing literature for published information (where available) on 1,324,224 phospholipids. VaLID is made freely available as part of the CIHR Training Program in Neurodegenerative Lipidomics at http://neurolipidomics.com/resources.html and its data-

bases and engines are hosted by the Carleton Immersive Media Studios[6]. Version 3.0 was released at the 2014 IWBBIO 2014 (2nd International Work-Conference on Bioinformatics and Biomedical Engineering) meeting on April 7 2014.

4 Acknowledgements

This resource was funded by the Canadian Institutes of Health Research (CIHR) MOP 89999 to DF and SALB and a Strategic Training Initiative in Health Research (STIHR) CIHR/ Training Program in Neurodegenerative Lipidomics (CTPNL) and Institute of Aging TGF 96121 to DF, SF, and SALB. APB received a FRSQ and CTPNL graduate scholarship; GSVM received a CTPNL graduate scholarship.

[6] http://www.cims.carleton.ca

5 References

1. Bennett, S.A.L., Valenzuela, N., Xu, H., Franko, B., Fai, S., Figeys, D.: Using neurolipidomics to identify phospholipid mediators of synaptic (dys)function in Alzheimer's Disease. Frontiers in physiology 4, 168 (2013)
2. Wenk, M.R.: The Emerging Field of Lipidomics. Nature Reviews Drug Discovery 4, (2005)
3. Piomelli, D., Astarita, G., Rapaka, R.: A neuroscientist's guide to lipidomics. Nat Rev Neurosci 8, 743-754 (2007)
4. Brown, H.A., Murphy, R.C.: Working towards an exegesis for lipids in biology. Nat Chem Biol 5, 602-606 (2009)
5. Bou Khalil, M., Hou, W., Zhou, H., Elisma, F., Swayne, L.A., Blanchard, A.P., Yao, Z., Bennett, S.A.L., Figeys, D.: Lipidomics era: Accomplishments and challenges. Mass Spectrom Rev 29, 877-929 (2010)
6. Xu, H., Valenzuela, N., Fai, S., Figeys, D., Bennett, S.A.L.: Targeted lipidomics - advances in profiling lysophosphocholine and platelet-activating factor second messengers. The FEBS journal 280, 5652-5667 (2013)
7. Zehethofer Nicole, Pinto, D.M.: Recent developments in tandem mass spectrometry for lipidomic analysis. Analytica Chimica Acta 627, (2008)
8. Deacon, E.M., Pettitt, T.R., Webb, P., Cross, T., Chahal, H., Wakelam, M.J., Lord, J.M.: Generation of diacylglycerol molecular species through the cell cycle: a role for 1-stearoyl, 2-arachidonyl glycerol in the activation of nuclear protein kinase C-betaII at G2/M. J Cell Sci 115, 983-989 (2002)
9. Callender, H.L., Forrester, J.S., Ivanova, P., Preininger, A., Milne, S., Brown, H.A.: Quantification of diacylglycerol species from cellular extracts by electrospray ionization mass spectrometry using a linear regression algorithm. Anal Chem 79, 263-272 (2007)
10. Niemela, P.S., Castillo, S., Sysi-Aho, M., Oresic, M.: Bioinformatics and computational methods for lipidomics. J Chromatogr B Analyt Technol Biomed Life Sci 877, 2855-2862 (2009)
11. Andrej, S., Kai, S.: Lipidomics: coming to grips with lipid diversity. Nature Reviews 11, 6 (2010)
12. Fahy, E., Subramarniam, S., Brown, H.A., Glass, C.K., Merrill Jr, A.H., Murphy, R.C., Raetz, C.R.H., Russell, D.W., Seyama, Y., Shaw, W., Shimizu, T., Spener, F., van Meer, G., VanNieuwenhze, M.S., White, S.H., Witztum, J.L., Dennis, E.A.: A comprehensive classification system for lipids. Journal of Lipid Research 46, (2005)
13. Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E.A., Glass, C.K., Merrill Jr, a.H., Murphy, R.C., R.H., R.C., Russell, D.W., Subramaniam, S.: LMSD: LIPID MAPS structure database. Nucleic Acids Research 35, (2006)
14. Blanchard, A.P., McDowell, G.S., Valenzuela, N., Xu, H., Gelbard, S., Bertrand, M., Slater, G.W., Figeys, D., Fai, S., Bennett, S.A.L.: Visualization and Phospholipid Identification (VaLID): online integrated search engine capable of identifying and visualizing glycerophospholipids with given mass. Bioinformatics 29, 284-285 (2013)

15. De Laeter, J.R., Böhlke, J.K., De Bièvre, P., H., H., H.S., P., K.J.R., R., P.D.P, T.: Atomic Weights of the Elements: Review 2000. Pure Applied Chemistry 75, 683-800 (2003)
16. McDowell, G.S.V., P Blanchard, A., Taylor, G.P., Figeys, D., Fai, S., Bennett, S.A.L.: Predicting glycerophosphoinositol identities in lipidomic datasets using VaLID (Visualization and Phospholipid Identification) – an online bioinformatic search engine. submitted (2013)
17. Fahy, E., Sud, M., Cotter, D., Subramaniam, S.: LIPID MAPS online tools for lipid research. Nucleic Acids Research 35, W606-W612 (2007)