Dissertation

submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of

Doctor of Natural Sciences

presented by

Samy Lyes Deghou, M.Sc. Engineering

Oral examination
May 22, 2015

Integration and Analysis Of Large Scale Data In Chemical Biology

Referees Dr. Kiran Raosaheb Patil
Prof. Dr. Robert Russell

Acknowledgments-Danksagung-Remerciements

I devoted 692 active working days to complete this PhD thesis. Countless people have in one way or another contributed to this work, and I would like now to express my gratitude.

The first thank you goes to my supervisor, **Dr. habil. Peer Bork**. You told me on the 04.03.2011 that you would accept me into your group (Friday of the interview week). That's where it all started. I thank you most sincerely for embarking me into this adventure. In a way, you are *the sine qua none* condition to this achievement. You always kindheartedly acquiesced to my requests and questions, and have unfailingly provided me with insightful ideas. You have a good nose for where biological sciences are gonna head, and you have a particular talent for seeing the big picture and always putting things back in the context of a scientific question. These qualities of yours took my head out of irrelevant details more than once, and helped me achieving this work for which I am proud of. So thank you, for enabling this work and chapter of my life.

To my thesis advisory committee members, **Dr. Anne-Claude Gavin**, **Dr. John P Overington**, **Prof. Dr. Robert Russell** and **Prof. Dr. Christian Von Mering**, for providing regular encouragement, critiques and feedbacks during the course of my PhD. To **Dr. Kiran Patil**, **Dr. Wolfgang Huber** and **Prof. Dr. Roland Eils** for accepting without hesitating to be part of my defense committee. I consider myself as extremely lucky to have group leaders of your scientific expertise judging my work.

To **Georg Zeller**. You are a skilled, talented and exemplary scientist and have undoubtedly shaped the most my conception of science. I thank you for teaching me (even though I am a slow learner) the difference between curiosity and "not losing your time on irrelevant tasks", and for giving me a sense of structure in my work.

To **Murat Iskar** for always enthusiastically and helpfully answering my questions, and for your sense of anticipation.

To **Marja Driessen** a.k.a chocolate smuggler. Thank you for having tested half a million times the CART automatic installer, and thank you for believing me each time when I said that it would be working. Thank you for your patience. **Matthew Hayward** and **Simone Li**, thank you for your english expertise applied to this thesis. **Charles**

Girardot, for your eminent galaxy expertise.

Anne-Claude Gavin, **Ivana Vonkova** and **Antoine-Emmanuel Saliba**. Having working in collaboration with you for over 2.5 years almost on a daily basis gave me the feeling of being in an independent group. Together, we have accomplished milestones in the field of protein-lipid interactions and this thesis would definitely not be the same without you. A big thank you to the three of you.

I of course cannot forget **Samuel Stovicek**, the last member of this group. I thank you dearly for not coming to the world too early. The process of writing this PhD thesis would have been much more of a nightmare if you had decided to join us a month earlier.

To my northerner friends, **Mark Rynbeek**, **Ken Haug**. Your eminent software engineering expertise have significantly helped me in the context of the CART project. Thanks for never being -seemingly- annoyed by my questions. Most importantly, I am truly grateful for listening to me and wisely advising me when I experienced rough times at the beginning of my PhD.

A special word to **Janna hastings**. You have influenced me a lot during my time in Cambridge and it has had implications (positive ones) on my PhD. I miss those discussions we used to have in Cambridge's taverns on ontologies, computer science, nature and life in general and I feel verily honored, to be able to have shared that time with you !

Christoph Steinbeck, for having continuously teaching me how to ask yourself whether you're conducting a balanced life. It truly helped me during my PhD.

Mark, **Ken**, **Chris** and **Janna**, thanks again for helping me during those difficult times I had at the beginning of my PhD.

I obviously cannot forget **Alberto Riva**. It is with you that I have done my first steps in the world of computers, and you managed to transform this first time experience into a deep passion, which lead me where I am now (I am not blaming you). Four years have elapsed since I left your group, and I am really happy that we kept it touch. You did pay a lot of attention to my worries and questions 3 years ago, and I am immensely grateful to all your wise advises. I hope to be able to work with you again in the future!

An meine allerersten heidelbergischen Freunde, **Anne Régnier-vigouroux**, **Tim Kees**, **Ivana Dokic** und **Johannes Noack**. Ihr seid der Grund, weshalb ich ursprünglich überhaupt zurück nach Heidelberg wollte. Die 4 Wochen, die ich im Juni/Juli 2009 in eurer Gruppe verbracht habe -wenngleich sie mir wie ein Wimpenschlag erscheinen, wenn man sie mit der Länge eines PhDs vergleicht- zählen zweifellos zu den Besten meiner heidelbergischen Errinerungen. Ich danke euch für diese "Blitz-Einführung", die eine äusserst wichtige Rolle in meiner Entscheidung hierher zu ziehen gespielt hat.

Thanks to the lads (**Francis O'Reilly**, **Dermot Harnett**, **Charlotte Banfield**, a.k.a the lads or the good, the bad and the ugly) for the cracking time in that Wohnungsgemeinschaft of us.

Selbstverständlich darf ich die **Buben** des Triathlonvereins nicht vergessen. **Benj**, danke für die zahllosen sinnfreien Besteigungen des KS, und danke **Jan**, dass du gelegentlich versucht hast so tapfer mitzuhalten. **Janosch**, für deine unermüdlichen Einsätze, die uns unter anderem nach Mallorca geführt haben.

Anita Voigt. Du warst die Erste und bist für eine lange Zeit die Einzige aus meiner täglichen Umgebung geblieben, die mit mir deutsch geredet hat. Ich danke dir von ganzem Herzen dafür. Du hast dadurch viele Türen geöffnet und mir viel ermöglicht.

Lauren, thank you for supporting me during the first half of my PhD. You helped me a lot, and I am grateful for that.
Auré, vieille fripouille. Je pourrais t'écrire un livre, et je commencerais par te remercier pour cette merveilleuse semaine en Italie passé en Avril 2013. Elle arriva au bon moment ! J'éspère qu'on en aura d'autres !

Katharina Zirngibl und meiner zweiten Familie. Ich müsste mich für viel zu viel bei euch bedanken. Also schlicht und ergreifend: danke für alles.

Papa, **maman**. Vous êtes la raison et ma raison pour tout. Vous êtes mon étoile du berger. Continuez à briller, et je continuerai à avancer!

Integration and Analysis Of Large Scale Data In Chemical Biology

Samy Lyes Deghou

Supervisor Dr. Peer Bork
Institute Structural and Computational Biology Unit
European Molecular Biology Laboratory, Heidelberg, Germany

English Summary

The universe of small molecules is huge. Small molecules characterize organic chemical compounds, which are of a much lower molecular weight than macromolecules like proteins or DNA. Small molecules are grouped into different families according to their physico-chemical or functional properties, and they can be either natural (like lipids) or synthetic (like drugs). Only a staggeringly low fraction of the small molecule universe has been characterize, and very little is known about it. For instance, we know that lipids can play the role of scaffolding and energy storage compounds, and that they differently compose biological membranes. However, we don't know if it influences some biological functions, including protein recruitment to membranes and cellular transport.

Chemical biology aims at utilizing chemicals in order to explore biological systems. Advances in synthesizing big chemical libraries as well as in high-throughput screenings have led to technologies capable of studying protein-lipid interactions at large scale and in physiological conditions. Therefore, answering such questions has become possible, but it presents many new computational challenges. For instance, establishing methods capable of automatically classifying interactions as binding or non-binding requiring a minimal interaction with human experts. Making use of unsupervised clustering methods to identify clusters of lipids and proteins exhibiting similar patterns and linking them to similar biological functions.

To tackle these challenges, I have developed a computational pipeline performing a technical and functional analysis on the readouts produced by the high-throughput technology LiMA. Applied to a screen focusing on 94 proteins and 122 lipid combinations yielding more than 10,000 interactions, I have demonstrated that cooperativity was a key mechanism for membrane recruitment and that it could be applied to most PH domains. Furthermore, I have identified a conserved motif conferring PH domains the ability to be recruited to organellar membranes and which is linked to cellular transport functions. Two amino acids of this motif are found mutated in some human cancer, and we predicted and confirmed that these mutations could induce discrete changes in binding affinities in vitro and protein mis-localization in vivo. These results represent milestones in the field of protein-lipid interactions.

While we are progressing toward a global understanding of protein-lipid interactions, data on the bioactivities of small molecules is accumulating at a tremendous speed. In vitro data on interactions with targets are complemented by other molecular and phenotypic readouts, such as gene expression profiles or toxicity readouts. The diversity of screening technologies accompanied by big efforts to collect the resulting data in public databases have created unprecedented opportunities for chemo-informatics work to integrate these data and make new inferences. For instance, is the protein target profile of a drug correlated with a given phenotype? Can we predict the side effects of a drug based on its toxicology readouts? In this context, I have developed CART: a computational platform with which we address major chemo-informatics challenges to answer such questions. CART integrates many resources covering molecular and phenotypical readouts, and annotates sets of chemical names with these integrated resources. CART includes state-of-the-art full-text search engine technologies in order to match chemical names at a very high speed and accuracy. Importantly, CART is a scalable resource that can cope with the increasing number of new chemical annotation resources, and therefore, constitutes a major contribution to chemical biology.

Integration and Analysis Of Large Scale Data In Chemical Biology

Samy Lyes Deghou

Supervisor Dr. Peer Bork
Institute Structural and Computational Biology Unit
European Molecular Biology Laboratory, Heidelberg, Germany

Zusammenfassung
Das Universum von niedermolekularen Substanzen, sogenannter kleiner Moleküle, ist riesig. Kleine Moleküle charakterisieren organische chemische Substanzen, die ein sehr viel geringeres Molekulargewicht besitzen als Proteine oder DNA. Kleine Moleküle sind anhand ihrer physikalisch-chemischen oder funktionellen Eigenschaften in verschiedene Familien gruppiert und können entweder natürlichen (z.B. Lipide) oder synthetischen (z.B. Medikamente) Ursprungs sein. Nur ein äußerst geringer Teil des Universums kleiner Moleküle wurde bis jetzt charakterisiert und das Wissen darüber ist noch sehr begrenzt. Man weiß zum Beispiel, dass Lipide die Rolle von Gerüst- und Energiespeicherstoffen übernehmen können und dass sie biologische Membranen in unterschiedlicher Art zusammensetzen. Es ist jedoch nicht bekannt, ob dies in irgendeiner Weise biologische Funktionen beeinflusst, wie etwa die Rekrutierung von Proteinen zur Membran oder den zellulären Transport. Chemische Biologie zielt darauf ab Chemikalien zu benutzen, um biologische Systeme zu erforschen. Fortschritte in der Synthese großer chemischer Bibliotheken, sowie in Hochdurchsatz-Screenings haben zu Technologien geführt, die in der Lage sind Protein-Lipid Interaktionen im großen Maßstab und unter physiologischen Bedingungen zu untersuchen. Dadurch ist es ermöglicht worden derartige Fragen zu beantworten, dies bedeutet allerdings gleichzeitig neue computergestützte Herausforderungen, wie z.B. die Einführung von Methoden, die automatisch Interaktionen als bindend oder nicht-bindend klassifizieren können und dabei nur einen geringen Austausch mit Experten benötigen. Oder auch die Anwendung unüberwachter Clustertechniken zur Identifizierung von Lipid- und Proteinclustern, die ähnliche Eigenschaften aufweisen und die Verknüpfung dieser mit ähnlichen biologischen Funktionen. Um diese Herausforderungen in Angriff zu nehmen, habe ich eine computergestützte Pipeline entwickelt, die technische und funktionelle Analysen von experimentellen Ausleseergebnissen durchführt, die mit Hilfe der Hochdursatz-Technologie LIMA gewonnen wurden. Angewandt auf einen Screen, der 94 Protein- und 122 Lipidkombinationen aufweist und mehr als 10.000 Interaktionen umfasst, haben wir Kooperativität als Schlüsselmechanismus für Membranrekrutierung nachgewiesen sowie, dass dies auf die meisten PH-Domänen zutrifft. Weitergehend habe ich ein konserviertes Motiv bestimmt, dass PH-Domänen die Fähigkeit verleiht zu Organellenmembranen rekrutiert zu werden und mit zellulären Transportfunktionen verbunden ist. Zwei Aminosäuren dieses Motivs wurden in Patienten einiger Krebsarten mutiert vorgefunden. Wir haben vorausgesagt und bestätigt, dass diese Mutationen in vitro bestimmte Veränderungen der Bindungsaffinität und in vivo falsche Lokalisierungen von Proteinen verursachen können. Diese Ergebnisse stellen einen Meilenstein im Feld der Protein-Lipid Interaktionen dar. Während wir uns immer mehr einem globalen Verständnis von Protein-Lipid Interaktionen nähern, nehmen Daten über bioaktive kleine Moleküle in einer enormen Geschwindigkeit zu. In vitro-Daten über Interaktionen mit Zielmolekülen werden mit anderen molekularen und phänotypischen Ausleseergebnissen ergänzt, wie z.B. Genexpressionsprofilen oder Toxizitätsausleseergebnissen. Die Vielfalt von Screening-Technologien zusammen mit einem großen Bestreben vorhandene Daten in öffentlichen Datenbanken zusammenzuführen, haben bislang unbekannte Möglichkeiten für chemo-informatische Arbeiten geschaffen diese Daten zu integrieren und neue Erkenntnisse aus ihnen zu gewinnen. Korreliert beispielsweise das Proteinzielprofil eines Medikaments mit dem Auftreten eines bestimmten Phänotyps? Ist es möglich die Nebenwirkungen eines Medikaments anhand seiner toxikologischen Ausleseergebnisse vorherzusagen? In diesem Zusammenhang habe ich CART entwickelt: eine computergestützte Plattform, mit der

großen chemo-informatischen Herausforderungen begegnet werden kann, um derartige Fragen zu beantworten. CART integriert viele Datenbanken die molekulare und phänotypische Ausleseergebnisse umfassen und annotiert chemische Namenslisten mit diesen integrierten Datenbanken. CART besitzt eine sich auf dem neusten Stand der Technik befindende Volltext-Suchmaschinentechnologie, um chemische Namen mit einer sehr schnellen Geschwindigkeit und großen Genauigkeit zuzuordnen. CART ist eine skalierbare Ressource, die für die zunehmende Anzahl neuer chemischer Annotationsquellen ausgerichtet ist und somit einen wichtigen Beitrag in der chemischen Biologie leistet.

Contents

I Introduction **ix**

Chemicals . . . xii
Chemical biology . . . xii
The foundations . . . xii
High-throughput screens . . . xiii
Computational analysis of target-based screens . . . xiii
Data integration in chemo-informatics . . . xiv
Summary . . . xvi
Thesis outline . . . xvi

II Development of a computational pipeline to analyze readouts from a high-throughput protein-lipid interaction screen **1**

1 Introduction **4**
1.1 Lipids . . . 4
1.2 Lipid Binding Domains . . . 4
1.3 PH domains . . . 5
1.4 Studying protein-lipid interactions . . . 7

2 Results **9**
2.1 LiMA . . . 9
2.1.1 concept . . . 9
2.1.2 Array design . . . 10
2.1.3 Detection of interactions . . . 10
2.1.4 Readout . . . 12
2.1.5 Performance in high-throughput . . . 14
2.1.6 Summary . . . 15
2.2 Dataset . . . 16
2.2.1 Lipids . . . 16
2.2.2 Proteins . . . 17
2.2.3 Interactions . . . 19
2.3 Preprocessing of LiMA readouts . . . 19
2.3.1 Filtering . . . 19
2.3.2 Reproducibility of the assay . . . 19
2.3.3 Possible bias . . . 20
2.3.4 Extraction of a binding threshold . . . 25
2.3.5 Assessment of binding quality . . . 25

2.3.6 Sensitivty and specificity assessment . . . 30
2.3.7 Number of required manual annotations . . . 31
2.4 Functional analysis of LiMA readouts . . . 32
2.4.1 Clustering analysis . . . 32
2.4.2 A new motif behind the recruitment to Organellar PIPs . . . 35
2.4.3 Lipid cooperativity . . . 42
2.5 Material and Methods . . . 47
2.5.1 Screen settings . . . 47
2.5.2 Technical analysis pipeline . . . 48
2.5.3 Functional analysis pipeline . . . 50
2.6 Discussion . . . 54
2.6.1 LiMA, a novel tool for high throughput screening of protein lipid interactions . . . 54
2.6.2 Systematic analysis of PH domains confirms their ability to interact with membranes 55
2.6.3 Only fraction of PH domains interacts with PIPs localized in ORG membranes . . 55
2.6.4 PH domains interacting with ORG PIPs share common sequence motif . . . 56
2.6.5 Interactions of PH domains with membranes are modulated by cooperation between PIPs and other lipid species . . . 57
2.6.6 Recognition of the same cooperating lipids is shared by functionnally linked PH domain-containing proteins . . . 57
2.6.7 A new path for drug discovery . . . 58
2.6.8 Conclusion . . . 59

III Development of an efficient and scalable chemical analysis platform 60

3 Introduction 63
3.1 Publicly available resources . . . 63
3.2 Toward an integration of the publicly available resources . . . 64
3.3 The example of the genomic field . . . 64
3.4 The challenges . . . 64
3.4.1 Chemical name search . . . 65

4 Results 66
4.1 Features and functionality . . . 66
4.2 Data integration . . . 68
4.2.1 Molecular targets . . . 68
4.2.2 Therapeutic classifications . . . 68
4.2.3 Phenotypic readouts . . . 68
4.3 Name matching . . . 70
4.4 Enrichment . . . 71
4.5 Visualization . . . 73
4.6 Synonyms . . . 74
4.7 Web interface . . . 74
4.8 Standalone version . . . 74
4.9 Benchmarking . . . 75

5 Future directions 77

IV Conclusion **79**

Conclusion **81**

V Supplementary Information **1**

6 Relative to Part II **3**

7 Relative to Part III **30**

List of Publications

This thesis is based on the following manuscripts/publications

Part II

A quantitative liposome microarray to systematically characterize protein-lipid interactions
Saliba, A.E., Vonkova, I, **Deghou, SL** *et al* ***Nature methods***, 2014

Lipid cooperativity as a general membrane-recruitment principle of PH domains
Vonkova, I, Saliba, A.E., **Deghou, SL** *et al* ***Science***, In review

A protocol for the systematic and quantitative measurement of protein-lipid interaction using the liposome microarray-based assay (LiMA)
Saliba, A.E., Vonkova, I, **Deghou, SL** *et al* ***Nature protocols***, Manuscript (First author(s) to be decided)

Part III

CART: an efficient and scalable Chemical Annotation Retrieval Toolkit
Deghou, S., Zeller, G.,Iskar, M., van Noort V., Bork P ***Bioinformatics***, Manuscript

Abbreviations

1p	1 phosphate
BSM	Basic Sequence Motif
Cer	ceramide
DBM	Database Model
DHS	Dihydrosphingosine
DOPI34P2	3,4 biphosphorylated PIP
DOPI35P2	3,5 biphosphorylated PIP
DOPI3P	3 monophosphorylated PIP
DOPI45P2	4,5 biphosphorylated PIP
DOPI45P3	3,4,5 triphosphorylated PIP
DOPI4P	4 monophosphorylated PIP
DOPI5P	5 monophosphorylated PIP
LBD	Lipid Binding Domain
NBI	Normalized Binding Intensity
ORG	Organellar
ORG PIPs	Organellar PIPs
PC	Phosphatidylcholine
PE	Phosphatidylethanolamine
PH domain	Pleckstrin Homology domain
PHS	Phytohydrosphingosine
PIPs	phospholipids
PIs	phosphatidylinositol
PM	Plasmic Membrane
PM PIPs	Plasmic Membrane PIPs
PS	Phosphatidylserine
TAL	Thin Agarose Layer

Part I

Introduction

Part I
Summary

In this part, I will introduce the importance of chemicals as tools for learning about biological systems. I will describe how chemical biology and high-throughput technologies have led to the generation of massive amounts of chemical screening data and I will explain how computational analyses have helped to get a deeper understanding of chemical effects on biological systems.

Introduction

Chemicals

The set of all existing small molecules is colossus. Some estimates of this number approach 10^{60} [1]. On average, one novel substance has been synthesized or isolated every 2.6 seconds since 2008 [2]. At the time when this thesis was written, over 91 millions of unique small molecules were registered in a database maintained by the American Chemical Society [3], making it the most comprehensive chemical repository. Small molecules are chemical substances that may be natural or synthetic and some of them can modulate biological processes [4]. While they are of much lower weight than organic macromolecules like DNA or proteins, lipids and most drugs fall within the definition of small molecules[5]. Only a tiny fraction of the universe of small molecules is characterized, and these have been grouped into different families based on either their physico-chemical or functional properties. The physico-chemical properties of amphiphilic chemicals, a class that includes lipids, force them in aqueous environments to assemble into bilayers, which also form the scaffolds for biological membranes and thus delineate biological organelles and organisms. More than scaffolding compounds, lipids also act as mediators and regulators of many physiological functions, such as energy storage or cell signalling [6]. Another class of chemicals defined both by its functional and physico-chemical properties is the class of drugs. Natural products extracted from plants or bacteria have been used since the dawn of time by healers because of their specific effects on the human organism. As a result of modern pharmacology and medicinal chemistry research, high-throughput technologies have increased the number of potentially bioactive compounds to another dimension [7] [8]. Aside from their value as healing compounds, these small molecules can also be used to perturb physiologically normal organisms, so they can be systematically applied as tools to unravel the molecular responses and ultimately the wiring diagrams of living systems [4] [9].

Chemical biology

The foundations

The twentieth century has seen the raise of molecular biology approaches, which allowed us to understand the chemical reactions fundamentals to living organisms down to the underlying mechanistic principles [10]. Specific knock-outs of certain genes as a means of perturbing a biological system have been pivotal in understanding the role of those genes in the physiology of the genetically altered organism [11]. In the last decade, chemicals have been increasingly used to perturb biological systems [4] [9] [12] as an alternative method to genetic engineering. One advantage of using chemicals over genetic engineering is that it is easier to maintain the temporal control of the system's physiology, as the incubated perturbing chemicals can be washed away to return to an unperturbed state. Pioneer studies have shown that small molecules extracted from coal-tar could interact with specific cells and tissues [13]. Based on

this observation, Paul Ehrlich discovered that another small molecule (methylene blue) was capable of selectively staining nerve cells [14] and postulated for the first time that small molecules could be of great value to study membrane receptors specific to cell types and tissues [15]. By demonstrating that small molecules are endowed with a biological activity, and that they interact with specific targets within an organism, these studies have set the foundations of modern chemical biology. It became clear that small molecules could also help unraveling many other aspects of biological systems, yet it took several years before some of the chemical diversity was gathered in the first chemical libraries. While the first established chemical libraries were the propriety of pharmaceutical companies, later ones became available to academic researchers at the end of the twentieth century, when rapid synthesis of small molecules was invented [16] [17]. Chemical libraries size grew, and with that the hope of getting a more comprehensive understanding of biological systems. However, testing the effects of these chemical libraries, some of which contain up to millions of small molecules, required new high throughput screens performing many assays in parallel[18].

High-throughput screens

High-throughput screening is a technical advance that has made it possible to move beyond studying isolated biological phenomena using a reductionist approach to the comprehensive investigation of biological systems as a whole [19]. This paradigm has been fuelled by advances, mostly in the industrial sector, which enabled the rapid synthesis and screening of big chemical libraries. High-throughput screens of chemical libraries are often divided in two categories, depending on the nature of the readout. High-throughput screens yielding molecular readouts are often referred as target-based screens, and those yielding phenotypic readouts as phenotypic-based screens [4]. These screening efforts engendered a colossus amount of data on chemical interaction profiles and phenotypic effects. While most of the data is still kept proprietary [20], in recent years, high-throughput screening technologies have also been adopted by academic researchers [21], who deposited their data in public databases. For instance, two years after the creation of the Molecular Libraries Screening Centers Network (MLSCN), 256 MLSCN assays spanning more than 140,000 chemicals have been deposited in PubChem [22]. In addition to this, there exist at least 17 publicly available databases focused on biological activities of chemicals [23] [Supplementary Table V.1].

Computational analysis of target-based screens

Target-based screens are utilized to identify ligands of a protein or a group of proteins. A number of techniques have been developed to that purpose and make use of immobilized ligands on solid surface [24]. For instance, Birk et al. successfully utilized such an approach to measure the binding profile of many antibodies to immobilized triazines and screened for 384 compounds to look for those acting as thrombin inhibitors. Other screening setups tend to invert the surface-based method and make use of immobilized proteins [25]. Studies based on these protein arrays have globally mapped interactions in the yeast proteome or discovered novel human protein-protein interactions [26]. Such target-based screens have also led to the postulation of new hypotheses. For instance, are there rules governing the binding of these proteins to the screened chemicals? How do these interactions take part in a particular biological process? Are some proteins more important than others? Are some of these interactions affected in a certain disease? Target-based screens focusing on protein-lipid interactions and yielding quantitative readouts already exist [27] [28] at a throughput suitable for answering such questions. Therefore, there is an immediate need for computational analysis to pair with these technologies in order to address these questions. These analyses encompass assessing the reproducibility of such screens and of detecting possible

bias in the measurement of protein-chemical interactions. Often protein (or lipid) concentration used can vary from one replicate to the other and a precise quantification of the influence of this parameter on the measured readout is essential, as well as a normalization method if any concentration effect is found. Another computational challenge associated with such screens is the automatic classification of interaction as a binding or a non-binding event, ideally requiring minimal interaction with human experts. Automatic analysis assigning a quality index to each interaction has the advantage that subsequent analysis can be restricted to the subset of bona fide interactions. The biggest computational challenge however lies in the functional analysis of such readouts. To identify groups of proteins or lipids exhibiting similar binding patterns, which result in similar biological functions, unsupervised learning methods including hierarchical clustering analysis or principal component analysis can be used. Finding sequence motifs in proteins characteristic of certain binding properties requires sequence analysis of protein clusters. However, this can be a daunting computational task given the low sequence identity of some protein families [29]. As a specific example for the analysis of one such protein lipid interaction screen, I describe in part II of this dissertation how I have addressed these challenges in analysing a novel high-throughput protein-lipid interaction screen.

Data integration in chemo-informatics

While our understanding of protein-lipid interactions is still very incomplete, data on the bioactivities of small drug-like molecules is accumulating at a tremendous speed. In vitro data on interactions with protein targets, including therapeutic targets, off-targets and metabolizing enzymes, are complemented by phenotypic-based screens including multivariate cell-based and in vivo assays, such as expression response of cell lines and animal models to chemical perturbations [30]. The diversity of screening technologies accompanied by big efforts to collect the resulting data in public repositories [31] [32], have created unprecedented opportunities for chemo-informatics work aiming at integrating these heterogeneous data in order to make inferences on the complex effects of chemicals on biological systems. For instance, if each compound of a chemical family or library is annotated on one side in terms of proteins it binds to, and on the other side in terms of phenotypes it induces, then one can attempt to predict whether the recruitment to any specific protein or group of proteins is correlated with a particular propensity to induce a given phenotype, or vice versa (conceptually illustrated in Figure 1). In applying this logic, Rihel et al. brought phenotypic-based screens to the whole-organism level, and demonstrated that clustering chemicals according to their behavioural phenotypes observed in zebrafish was highly correlated of drug's bioactivity in higher organisms including mouse or humans [33]. In applying a similar logic, Fliri et al made use of hierarchical clustering analysis to show that the target profile of some drugs was highly predictive of their side effects [34]. Campillos, Kuhn et al. have demonstrated that the side effect similarity of two drugs could be utilized to predict whether they share a target, which has important implications in drug repurposing [35]. By integrating phenotypic data to known drug-target interactions, Kuhn et al revealed that a substantial proportion of drug side effects are primarily caused by interactions with specific proteins [36]. Associated with the opportunities to integrate these various chemical databases are the big chemo-informatics challenges to develop tools for linking disparate chemical identifiers and naming conventions in order to bridge multiple databases containing information of chemical actions at various scales. These form the basis for subsequent data mining, clustering and visualization. Together integrative computational analysis tools will help pave the way forward to acquiring a better systemic understanding of chemicals' effects on living systems [23].

In this context I present work (in part III of this dissertation) on a software tool developed for the purpose of integrating diverse information on chemical properties collected from various databases with

	PH1	PH2	PH3	PH4	PH5	PH6
Chemical a	0.003	0.020	0.001	0.023	0.343	0.002
Chemical b	0.061	0.001	0.099	0.081	1.001	0.018
Chemical c	3.236	3.018	0.018	0.036	0.146	0.332
Chemical d	0.010	0.130	0.096	1.980	0.130	0.096
Chemical e	0.063	0.008	0.015	0.002	0.008	0.165
Chemical f	0.006	0.035	0.005	0.014	0.017	1.030
Chemical g	1.047	2.009	0.007	0.063	0.039	0.006
Chemical h	0.008	1.016	0.030	0.008	0.016	0.003
Chemical i	1.001	0.221	0.033	1.001	0.025	0.033
Chemical j	1.203	3.067	0.001	0.003	0.007	0.011

Figure 1: Hypothetical illustration of a biological activity matrice . The matrice represents six PH domains whose binding properties have been measured against ten chemicals. A number above 1 implies an interaction between the PH domain and the chemical. If the chemicals highlighted in gray are all capable of inducing a particular phenotype, the PH domain PH1 and PH2 are most likeley responsible for those chemicals to induce this particular phenotype: they are the only PH domain capable of binding all four chemicals. Such biological activity matrices can thus be used to determine which proteins are involved in specific biological processes.

the goal of enabling multi-faceted annotation and enrichment analysis of chemicals. With this work we have addressed a major chemo-informatics challenge related to data on chemical bioactivities being scattered across many different databases. There exists no central repository, in which each chemical substance is referred to by a unique and universally accepted identifier, but there is rather many names for each chemical entity (including brand names) and structural identifiers are generally also not free of ambiguities. Therefore, matching chemical names into a common universe that spans most chemical annotation databases, is a major hurdle for integrative analysis and ideally a name matching algorithm should be accurate and fast enough to handle large sets of chemicals. Importantly, this framework needs to be scalable to cope with the constantly increasing number of new chemical annotations resources.

Summary

Technological advances have led to high-throughput technologies, which when applied to chemical libraries can generate a myriad of multi-parametric readouts, from target profiles to phenotypic outcomes. The availability of these data in public repositories have opened a new area of research in chemical biology, largely driven by computational analysis aiming to unravel the roles played by the plethora of chemical compounds in the context of physiology and diseases of living organisms.

Thesis outline

As such, two layers are being analyzed during this PhD. A first project on large-scale chemical biology was initiated in collaboration with the experimental group of Dr. Anne-Claude Gavin. Her team invented a high-throughput technology systematically charting tens of thousands of protein-lipid interactions in a physiological context [28]. My first contribution was to develop a computational pipeline realizing a technical assessment of this new method by working on a dataset produced by Dr Antoine-Emmanuel Saliba and Dr Ivana Vonkova [28] [37]. Namely, this technical analysis pipeline is meant to calibrate and evaluate the different readouts produced by this technology. My second implication aimed at coupling computational biology approaches to the readouts produced by this technology, in order to establish a lipid-binding landscape of PH domains and to get a deeper understanding of their biology and how lipid regulate cell physiology. A second project focused data integration in the context of chemical-biology and how to make sense of chemical sets [38]. Chemical high-throughput screens are flowering and the need for a computational framework capitalizing chemical resources that can identify common biological themes amongst chemicals urges, which is the focus of the second project [39].

Part II

Development of a computational pipeline to analyze readouts from a high-throughput protein-lipid interaction screen

Part II
Summary

In this part, I present the results of a computational analysis pipeline developed to process and analyze readouts from LiMA [28], a platform enabling the high-throughput study of protein-lipid interactions in physiological conditions. This project contributes to the field of large-scale chemical biology in the sense that it provides a reusable [37] computational platform analyzing readouts produced by a high-throughput technology aiming at understanding how chemicals affect biological systems through their interactions with proteins. The first part of the results focuses on the the technical part of the computational analysis pipeline. These results are published in the LiMA publication Saliba,Vonkova, Deghou et al [28]. The second part of the results focuses on the functional part of the computational analysis pipeline. I applied it to a screen performed by Dr Ivana Vonkova and Dr Antoine-Emmanuel Saliba. The results of this analysis are currently in review in Science (Vonkova, Saliba, Deghou et al [38]). All the data used for this analysis as well as the experimental validations results have been produced by Dr Ivana Vonkova and Dr Antoine-Emmanuel Saliba. I conceived and wrote the entire computational analysis pipeline and produced the results presented here, which were then discussed and confirmed with Dr Ivana Vonkova, Dr Anne-Claude Gavin, Dr Antoine-Emmanuel Saliba and Dr Peer Bork. Lastly, the entire computational analysis pipeline itself will be published as part of the LiMA experiment and analysis workflow publication currently in preparation for Nature Protocol and for which I am writing the computational part.

A quantitative liposome microarray to systematically characterize protein-lipid interactions
Saliba, A.E., Vonkova, I, **Deghou, SL** *et al* ***Nature methods***, 2014

Lipid cooperativity as a general membrane-recruitment principle of PH domains
Vonkova, I, Saliba, A.E., **Deghou, SL** *et al* ***Science***, In review

A protocol for the systematic and quantitative measurement of protein-lipid interaction using the liposome microarray-based assay (LiMA)
Saliba, A.E., Vonkova, I, **Deghou, SL** *et al* ***Nature protocols***, Manuscript
(First author(s) to be decided)

Chapter 1

Introduction

1.1 Lipids

Lipids are chemicals that constitute the keystones of every biological compartment. Eukaryotic cells are composed of thousands of structurally different lipids and 5% of the average Eurkaryotic genome comprises of genes involved in lipid biosynthesis [40]. Biological membranes contain lipids of various size and charge that are organized into a bilayer comprising a hydrophobic core and highly polarized interfacial region [41]. The membranes of cellular organelles are primarily composed of phospholipids (PIPs), glycerolipids, sphingolipids, sterols, and other lipid species in varying concentrations.
Many lipids are produced in the endoplasmic reticulum including; PIPs, glycerol, cholesterol as well as precursors of sphingolipids including ceramide (Cer). Once synthesized these lipids are exported to other organelles, in which they will be modified or directed to the plasma membrane. In the Golgi apparatus other sphingolipids including sphingomyeline, lactoglyceramide and glycosphingolipids are synthesized and transported to the plasma membrane. Phopshatidylinositols (PIPs) head groups contain several phosphorylation sites giving birth to mono, bi and tri phosphorylated PIPs. The diversity of the PIPs and their cellular repartitions make them very good markers of organellar and plasmic membranes. DOPI45P2 is the most abundant PIP and is characteristics of the plasma membrane. DOPI4P is the second most abundant PIP and is located in the membranes of the Golgi apparatus as well as in the plasmic membrane. DOPI35P2 and DOPI3P are on the other side scarcer and solely present in the membranes of lysozomes/endomes and early endosomes, respectively [42].
Eukaryotics cells possess an enormous catalogue of lipids with an almost unique repartition across membranes, which is far away from being completely depicted. It is still very unclear how the lipid composition is maintained and what implications it has on the global organization of the proteome and subsequent cellular functions mediated through protein-lipid interaction, including signalization and protein transport.

1.2 Lipid Binding Domains

A plurality of biological processes, crucial for the cell including cell signaling, vesicle budding, and membrane trafficking occur at the biological membranes. The attention paid to the critical role played by membrane-protein interactions in these processes has become more obvious in the past decade [43] [44]. The first lipid binding domain (LBD) has been discovered in 1989 and since then structural first and then computational analysis have identified at least 10 additional domains conferring to a protein the ability to bind lipids [45] [46]. The different families of LBDs have developed different ways to selectively

target membranes, but most of them must contend with the lipid content and physicochemical properties of plasmic and organellar phospholipid bilayers. Some LBDs like C2 domains and MARCKS ED domains have evolved to selectively target membrane based on their bulk electrostatic properties. Others including PH domains require stereospecific mode of lipid recognition and have a high affinity and selectivity for particular lipid head groups. Finally, some LBDs including FYVE, P40-PX pr P47-PX can also display multivalent lipid head group binding properties, also known as cooperative binding (Figure 1). Proteins capable of targeting membranes often possess more than one LBD in their sequence sand are thus subject to combine several ways of selectively target membranes, could it be via protein-protein interactions (Figure 1) [47] [45].

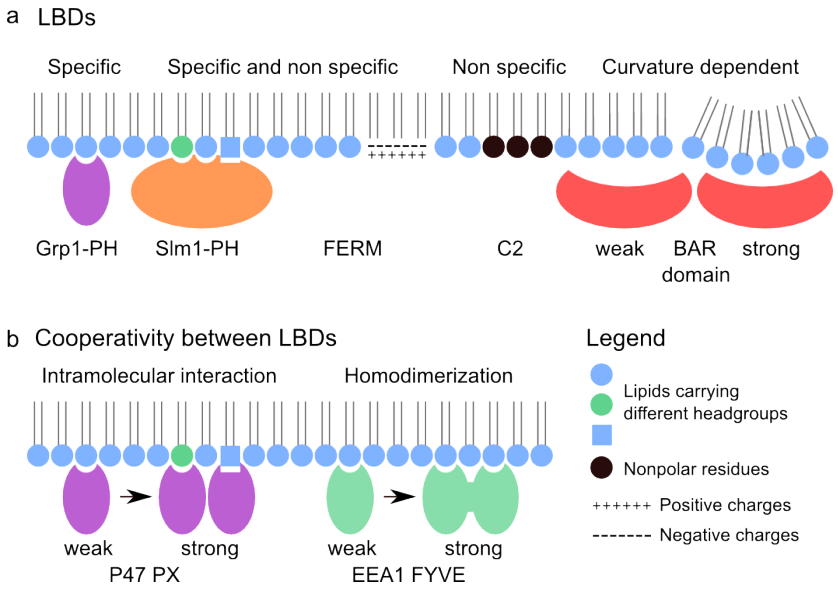


Figure 1: LBDs mode of recruitment to biological membranes (a) Some LBDs can recognize in a specific or unspecific way particular physical or chemical features that characterize the variety of biological membranes. (b) Other LBDs bind membranes by synergistically cooperating with other LBDs.

1.3 PH domains

PH domains are one of the most studied lipid binding domains [48] [49] [50]. They were identified in 1993 as conserved modules of 120 amino acids with a weak degree of homology to a protein kinase C substrate in platelets, the pleckstrin [51] [52]. There are about 250 human proteins known to contain at least one PH domain, making them the most common LBD in the human genome, but also in yeast [53]. PH domains share a common structural core, despite a very high sequence variability (20% sequence identity, [50]). It consists of seven-stranded β cylinders capped at the C-terminal end by an α helix (Figure 3). The loops, which correspond to the regions located within the β strands are extremely variable within subfamilies of PH domains in their structure, length and composition. Many studies have demonstrated that PH domains preferentially bind PIPs [54] [55] [56] [57] [58] [59] and more recent studies have shown their lack of specificity [60] [61] [62]. Their exacerbated binding towards PIPs over other lipid families is thought to be driven by the higher negative charge of the PIPs head groups [60], and even if some PH domains-PIPs interactions appear to be driven by stereospecific interactions, the neighboring electrostatic potential is thought to be primordial in orientating the PH domain toward the membrane [63]. The existence of different modes of interaction and determinism for binding specificities has been

proposed for a handful of PH domains. Namely, the crystal structures of some PH domains including PLCDELTA, PKB, Dapp1, and Grp have led to the identification of residues conferring PH domains the ability to bind certain PIPs. Those residues were baptized "signature motif" Figure 2 [64] [65] [66] [67]. Mutational analyses have indicated that (with the exception of PLCDELTA) the vast majority of the interactions with PIPs head groups are mediated by the signature motif.

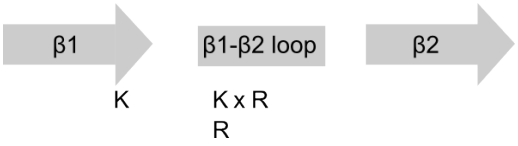


Figure 2: Schematic representation of the Basic Sequence Motif (BSM) required for PH domain to be able to bind PM PIPs.

PH domains have also been shown to cooperatively bind lipids and proteins. The PH domain of the protein β-ARK was shown to hold two binding sites through which the protein could simultaneously and effectively interact with DOPI45P2 and G$\beta\gamma$ in order to be recruited at the plasmic membrane [68]. Few instances of PH domains have also been demonstrated as being able to simultaneously bind two lipids belonging to two different families (PIPs and sphingolipids) [27]. Namely, the interaction strength between the PH domain of Slm1 and the two lipids DOPI45P2 and DHS1P was much higher than what one could have expected from the two individual binding event to DOPI45P2 and DHS1P. The PH domains of the proteins PDK1 and AKT1 also displayed similar cooperative binding mechanisms to the pair of lipids DOPI345P3 and PS [69] [70]. PH domains contain two known binding sites, the canonical and non-canonical binding site. They both flank the β1-β2 loop and are rich in basic amino acids favoring interactions with PIPs head groups (negatively charged) capable of recruiting many lipid ligands [64] [65] [71]. The existence of two binding sites capable of accommodating several ligands could probably explain the three instances of cooperative binding mechanisms that have been observed. PH domains are involved in a myriad of biological processes, including cell signaling, cytoskeletal rearrangements, lipid transport, metabolic processes, protein oligomerization and many others [48]. The binding mechanisms ruling PH domain recruitment to PIPs and cellular membranes is crucial for the fine regulation of these biological processes and the unperturbed continuation of a cell physiology. As a matter of fact, many proteins involved in diseases including diabetes, cardiovascular diseases and cancer are found mutated within their PH domain. In most of these cases, the mutations leads to natural variants either unable to bind a membrane it is supposed to or dramatically increase the binding to a particular membrane [72] [73] [74] [75] [76] [77]. The binding mechanisms dictating the recruitment of PH domains to biological membranes are more complex than what has already been postulated. The small number of proteins that was used to extract the signature motif needs to be increased in order to include PH domains displaying other binding properties. Mutations affecting or changing the membrane specificity of a PH domain clearly point towards residues that are not part of the signature motif and that account for a certain membrane specificity that still remain to be unraveled. PH domains capability to cooperatively bind two different lipids has indeed been mentioned, yet its incidence and implication in terms of biological processes are question that cannot be addressed with the current experimental and computational methods.

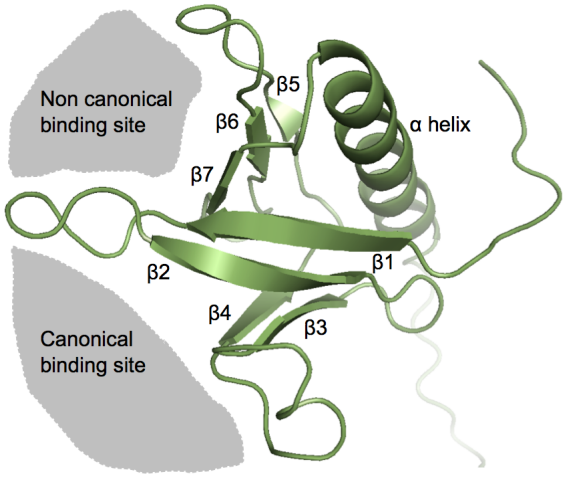


Figure 3: The cartoon structure represents the structure of the Osh3 PH domain (PDB ID: 2D9X), which is composed of seven β sheets and one α helix. The β1-β2 loop delimits the two known binding pockets.

1.4 Studying protein-lipid interactions

Most of what is known in the field of protein-lipid interaction comes from single case study to which all PH domains have been generalized. These findings are based on a very small subset of PH domains. The discovery of the signature motif is a prime example in which the authors used a handful of proteins to extract a motif that they subsequently generalized to all PH domains [78]. Plenty of methods have been developed since then in order to study more specific protein-lipid interactions (Figure 4). None of them fulfill though the criteria of studying them at large-scale, in conditions close to physiological concentrations and with a membrane lipid constitution similar to that of a cell. Moreover, there exist no computational analysis methods for analyzing readouts from protein-lipid interaction screens that could answer some of the questions posed in the introduction. Current in vitro and in vivo methods make use of fluorescent microscopy coupled to mass spectrometry readouts. These methods are known to produce a lot of noise and are not very reproducible. Methods making use of microarrays are well suited to study protein-lipid interactions at large scale. The first study probed the entire yeast proteome [79] and incubated the microarray with five different liposomes containing members of the PIP lipid family. This study represented a first tour de force in which many lipid binding proteins deprived from any known LBD were found, including the Kinase associated domain. This microarray technique offered the possibility to study an entire proteome from one sample, but was extremely restricted with regards to the possible number of testable lipids. Other techniques making use of microarrays directly spotted lipids (PIPS) and to study the P binding properties of yeast PH domains [62]. This analysis unexpectedly showed that few PH domains were capable of binding PIPs. Those that could target PIPs would show no particular specificity. The main weakness of this assay lied in the absence of a quantitative readout and therefore in the impossibility of undertaking any quantitative or statistical analysis of the PH domain's PIPs binding profile. Recently, Gallego et al developed a lipid array accommodating more lipids [27] with the help of which they measured the lipid binding profile of 172 proteins. This high-throughput screen identified a new lipid binding domain, the CRAL-TRIO (cryptic lipid-binding). Additionally, this screen demonstrated that one PH domain (Slm1) was capable to cooperatively bind to DOPI45P2 and DHS1P. The main drawback of this assay lied in the high concentration of lipid of interest, which probably yielded

a very high rate of false positive interactions. Regardless of the inherent drawbacks of each systematic method employed to study protein-lipid interactions, they have all contributed to the discovery of new interactions, new binding domains, or binding properties. However, there exist no technology enabling the precise measurement of lipid-protein interactions under physiological conditions at large scale and mixing several lipid of interest into a single liposome. Such a technology would need to output a quantitative readout representative of the protein dissociation constant, comparable between proteins. Only under these conditions could we conceive statistical and computational methods drawing a precise picture of the PH domain-lipid interaction landscape.

Method		Principle	Readout
In vivo	Ras rescue assay	Ras-signalling deficient cell followed by expression of Ras fused protein	Colony growth
	Genetic interactions	Single/double mutants	
	Live-cell imaging	Cell stimulation-fluorescent protein fusions	Fluorescent microscopy
	Protein-lipid co-purification	Complex assembly in vivo of tag engineered proteins followed by bead recognition of tags	Lipid identification
In vitro	Protein array	Spotted proteins on microarray	Fluorescent signal
	Lipid overlay assay	Spotted lipids on microarray	Radioactivity – immunodetection
	Lipid pull-down	Cell lysate – spotted lipids on beads	Protein identification
Chemical biology	Crosslinking	Cross linking of protein to tag engineered lipids	

Figure 4: High-throughput assays utilized to study protein-lipid interactions

Chapter 2

Results

2.1 LiMA

2.1.1 concept

The amount of systematic studies mentioning the importance of high-throughput approaches in the field of protein-lipid interactions grows rapidly [27] [80] [81]. Yet, protein-lipid interactions remain still poorly characterized on a systematic level because of the lack of methods capable of detecting interactions on a large-scale. We recently introduced Liposome Microarray-based Assay (LiMA) Figure 1 that allows the capture of more than 700 protein lipid interactions per day in a quantitative, automated, multiplexed and high-throughput manner. The core technology of LiMA relies on the in vitro parallel production of artificial membranes, the integration in a microfluidic format and the use of high-throughput fluorescence microscopy to quantify protein recruitment to membranes. LiMA increases by several orders of magnitude the throughput of protein-lipid interaction methods and allows with its paired computational analysis pipeline to envision proteome-wide protein-lipid interaction screens.

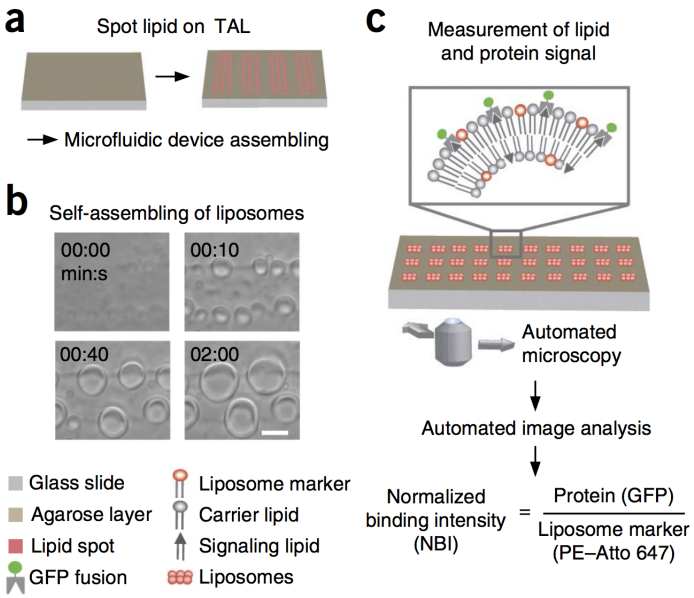


Figure 1: Principle of LiMA (a) The core of LiMA lies in the thin agarose layer containing 120 slots onto which lipid mixtures are automatically spotted. (b) Liposome are composed of the same lipids and differ only in their lipid(s) of interest (characterizing the liposome). The formation of giant liposomes is subsequent to this spotting. The extract containing the tagged protein of interest is then incubated. (c) The automated image analysis follows the incubation and compute one NBI per interaction.

2.1.2 Array design

Artificial liposomes enable the study of lipids under physiological concentrations and within membrane bilayers, but they have many drawbacks, making them difficult to use in a high-throughput manner. They are difficult to scale up [82], unstable and are known to oxidate quickly [83]. LiMA uses some of the liposome preparation methods described by Horger et al., a method enabling the formation of giant liposomes [84]. A defined lipid mixture is applied on a cover glass covered with a thin layer of agarose (TAL), onto which lipid mixtures are spotted. A lipid mixture contains a carrier lipid, the lipid(s) of interest, (characterized the liposome itself) as well as PE linked to poly(ethylene glycol) to help the formation of the liposomes. Finally, each liposome contained a lipid fluorescently labeled (PE-Atto 647) to help during the image-processing step. The detail composition of each mixture used during the screen is available in Supplementary Table V.3. Subsequent to this spotting the TAL is hydrated inducing quick formation of liposomes containing the lipid mixture applied beforehand (Figure 2). The diameter of the liposomes can be adjusted by varying the thickness of the TAL (Figure 1). The geometrical characteristic of the liposomes are stable for a period of 6 hours and they do not diffuse to neighboring spots of the array, thus reducing the risk of cross-contamination (Figure 1). We have applied 110 different lipid mixtures, spanning all principal families and all of them successfully were incorporated into liposomes of equal quality (Figure 3).

2.1.3 Detection of interactions

We produced a miniaturized array containing multiple spots, each holding a particular lipid mixture. We made use of microfluidics techniques in order to streamline the array with automated fluorescence

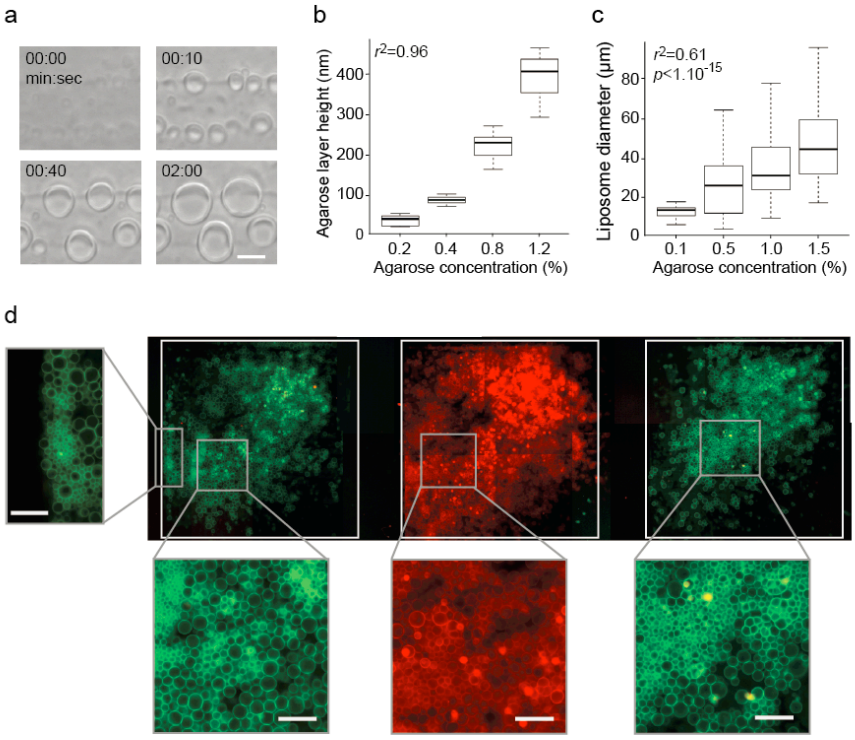


Figure 2: Liposome formation (a) Giant liposome formation time (<2 minutes). (b) TAL height grows with agarose concentration. (c) Liposome diameter grows with agarose concentration. (d) Pictures of giant liposome formed from neighboring spotting points. The enlarged squats (bottom) show that liposomes maintain their position within the spotting area. Cross contamination risks are thus minimized.

microscopy. The lipid mixtures were spotted on the TAL with a Camag thin-layer chromatography spotter. The liposomes were formed upon hydratation of spotted TAL by injecting the assay buffer in to the microfluidic chambers. The liposomes were then incubated with the GFP-tagged protein for 20 minutes. Subsequently, the unbound material was washed away. The amount of GFP-tagged protein recruited by liposomes was characteristic of the interaction between the protein and the lipid of interest of the liposome. This ratio was automatically computed by high-content fluorescent microscopy [85].

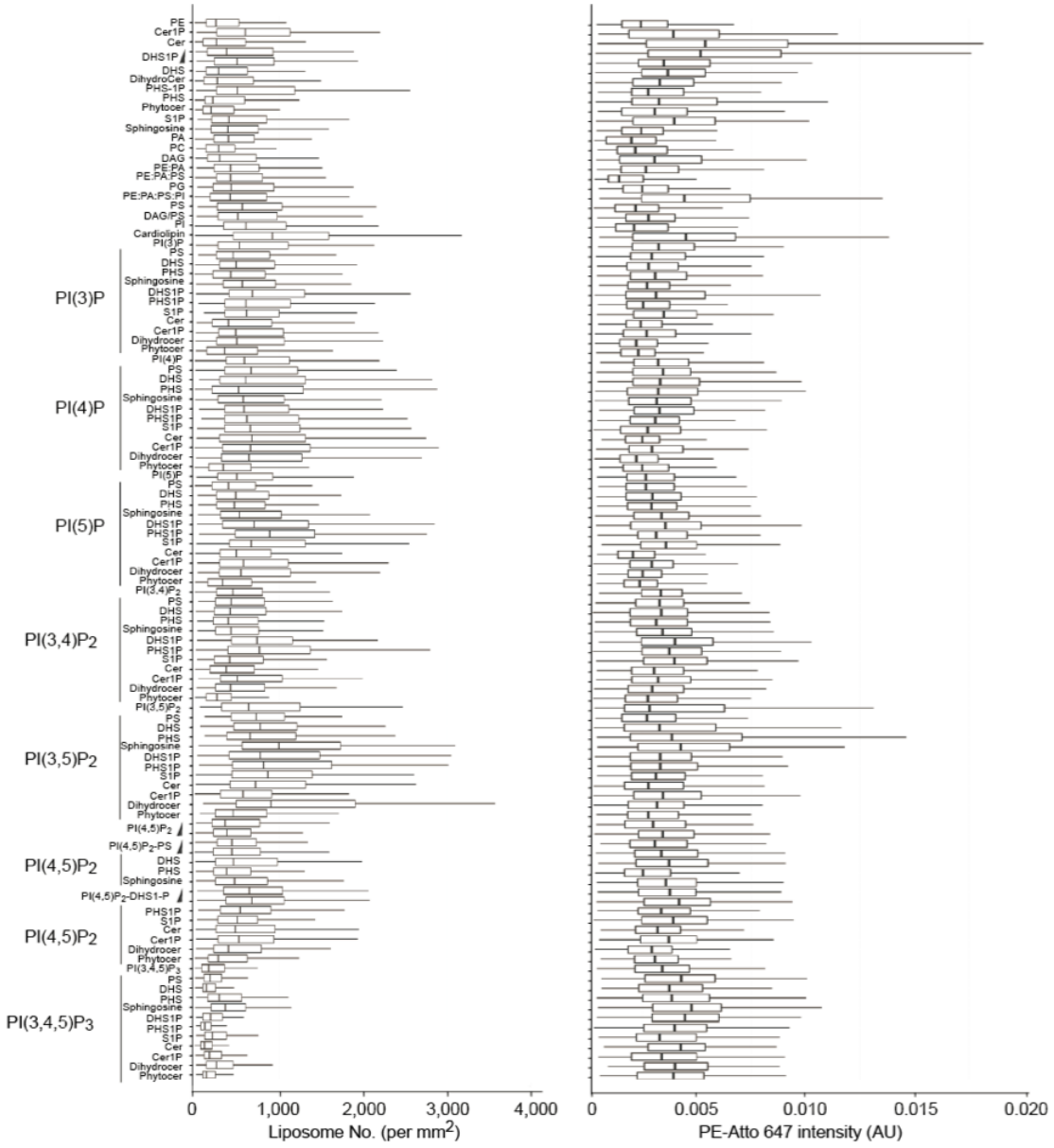


Figure 3: Liposome number (left) Boxplots representing the number of liposome observed for each liposome type and per mm2. (Right) Boxplots representing the liposome fluorescence intensity per liposome type.

2.1.4 Readout

To quantitatively characterize the recruitment of proteins to the formed liposomes, we have developed the normalized binding intensity. Each spot of the array onto which a lipid mixture had been spotted was automatically acquired on two channels: PE-Atto 647, the liposomal marker and GFP, characterizing the amount of protein bound to the liposome. Additionally, several exposure times were acquired in order to detect binders with different affinities and only pixels corresponding to the position of the liposome membranes were retained for the computation of the NBI. The GFP intensities represented the number of proteins recruited to the liposomes and the Addo-647 the available surface for proteins to bind to. The NBI was computed as the ratio of GFP to Atto-647 [See Materials and Methods 2.5.1]. The NBI can be used as a quantitative measure of protein-lipid interactions. LiMA was found to be sensitive to

changes in lipid or protein concentrations, and this was well reflected by the NBI (Figure 4). Protein affinities are also well represented by the NBI. For instance, the ph domain of the protein kinase AKT1 is known to specifically bind to DOPI345P3 and DOPI34P2, and binds to DOPI45P2 with less affinity [86]. The NBI measured with the technology LiMA very well reflected these specificities (Figure 4). The NBI is a quantitative readout that can be sensitive to protein concentration and lipid concentration. Furthermore, it can be used to to compare affinities between different proteins. As a proof of principle, we have measured the NBIs of the two variants of the PH domain of the protein SOS1 (SOS1-HF-WT and SOS1-HF-E108K) to DOPI45P2 and PA. This mutation is linked to the Noonan syndrome [87] and known to increase the affinity to PA [88]. The sensitivity of the technology to detect these slight changes was reflected in the NBIs measured. Surprisingly, an equally important increase of affinity to DOPI45P2 was observed, which was not known before. This might suggest that the two lipid ligands are recruited on the same binding sites (Figure 4).

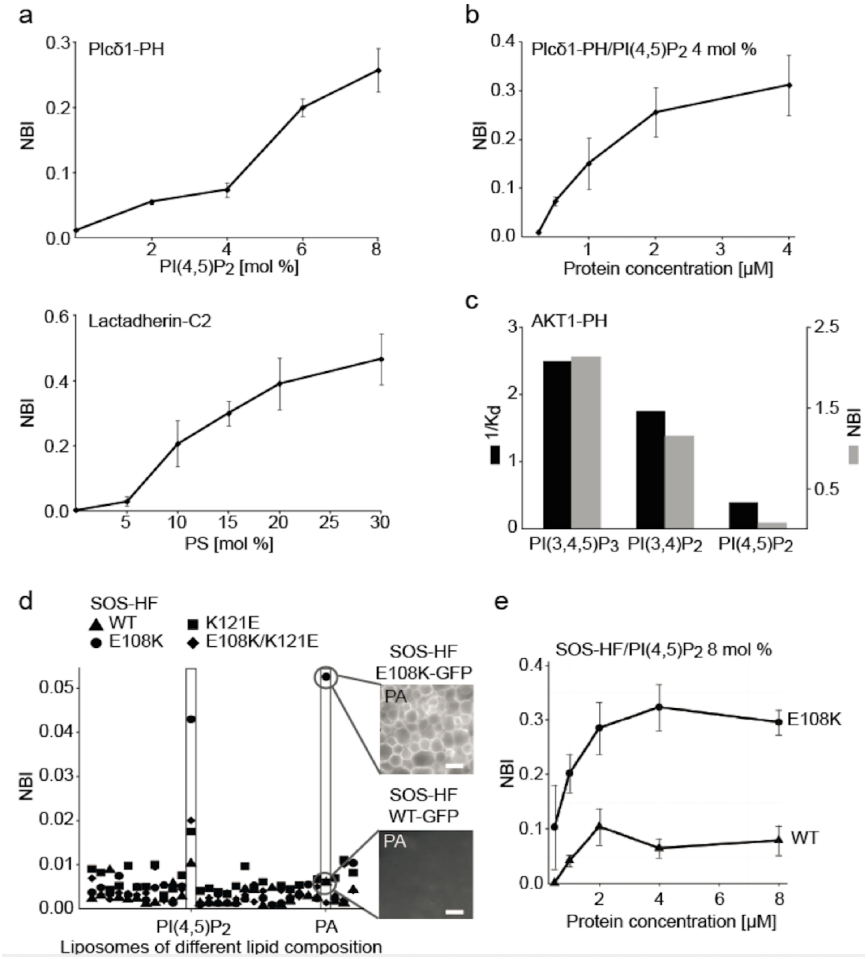


Figure 4: Sensitivity and Specificity of LiMA (a) Sensitivity of LiMA with regard to the lipid concentration : the NBI of a given interaction increases when the concentration of lipid of interest increases. (b) Sensitivity of LiMA with regard to the protein concentration : the NBI of a given interaction increases when the concentration of protein increases. (c) The NBI is directly proportional to the Kd of a protein. (d-e) Sensitivity of LiMA to detect fine alterations of a protein affinity caused by diverse mutations: the disease-associated mutation E108K substantially increased the affinity oft he protein SOS for both PA and DOPI45P2.

2.1.5 Performance in high-throughput

In order to evaluate how well LiMA can perform in a high-throughput screening, we probed seven proteins belonging to the most frequent lipid binding domains encountered in eukaryotes and measured their lipid binding preferences to 30 different liposomes. In total, we measured 300 different protein-lipid combinations (Figure 5). This pilot screen yielded an excellent reproducibility (Figure 5). NBIs reflected small changes in lipid concentrations as well as protein's specificities (Figure 5).

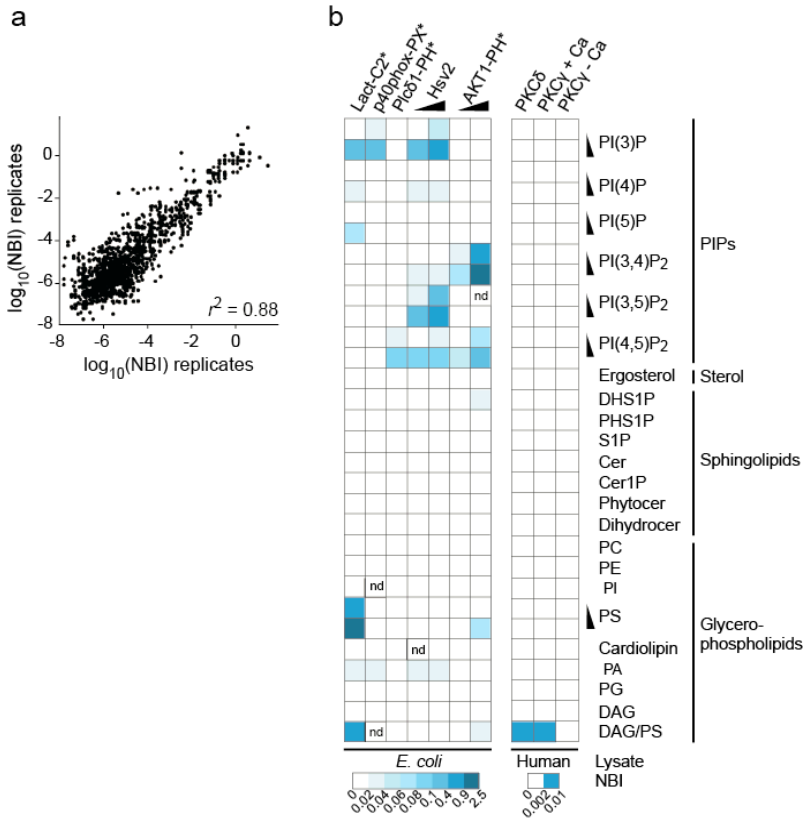


Figure 5: Performance in high-throughput (a) Scatter plot illustrating quantitative reproducibility of NBIs measured on pilot screen interactions (8 LBDs and 30 different liposomes). (b) Heatmap displaying PH domains binding profiles.

2.1.6 Summary

The first step of the computational analysis pipeline have shown that LiMA is a technology that can be applied to tens of thousands of protein-lipid interactions and provides a quantitative readout capable of reflecting protein's specificities. This information is essential to address some of the problems posed in the introduction.

2.2 Dataset

2.2.1 Lipids

This array was further designed to accommodate 122 different types of liposomes. The liposomes contain between one and four lipid of interest selected from a pool of 26 different signalling lipids (Figure 6). For comparison purposes, we decided to include non-physiological analogues that are not synthesized in yeasts or higher eukaryotes [Supplementary Table V.2]. Additionally, given that the exact in vivo lipid concentration is not known [89], we chose to use in vitro concentrations found in other studies studies [90] [91]. Biotinylated phosphatidylethanolamine (PE) was chosen as control lipids and was placed at ten specific positions on the array in order to served as an indicator for the assay quality (Figure 7).

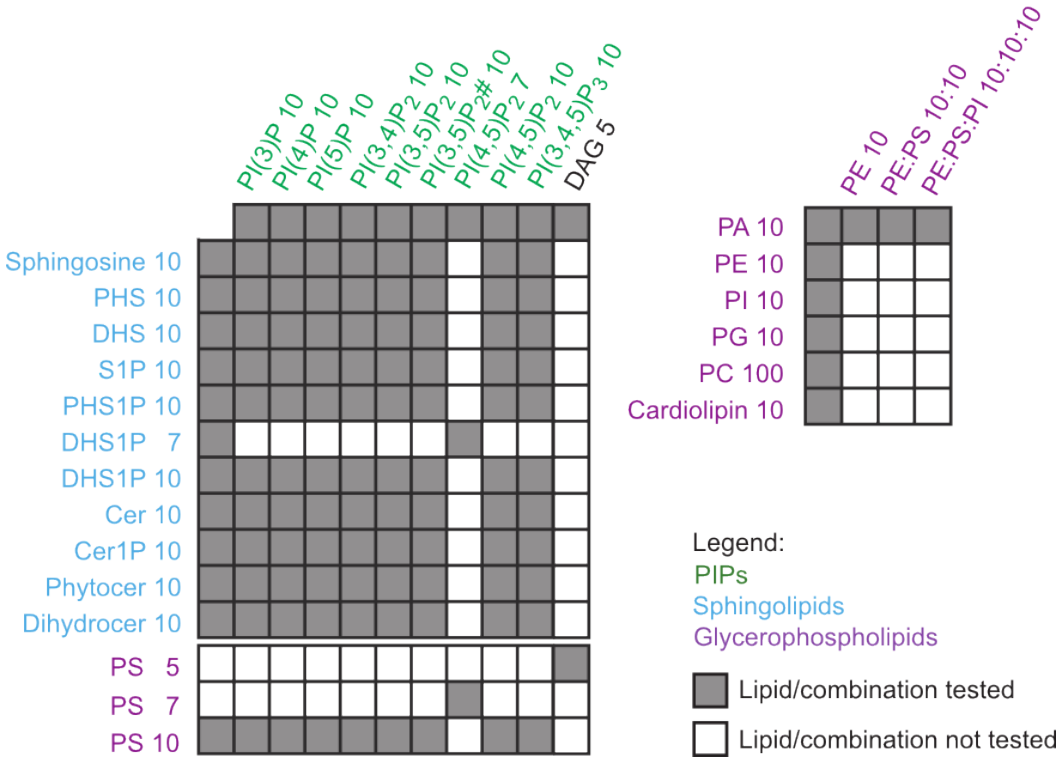


Figure 6: Matrix representing (gray) the lipid combinations selected to compose the 122 liposomes studied along with their concentration in mol % (number written next to the lipid).

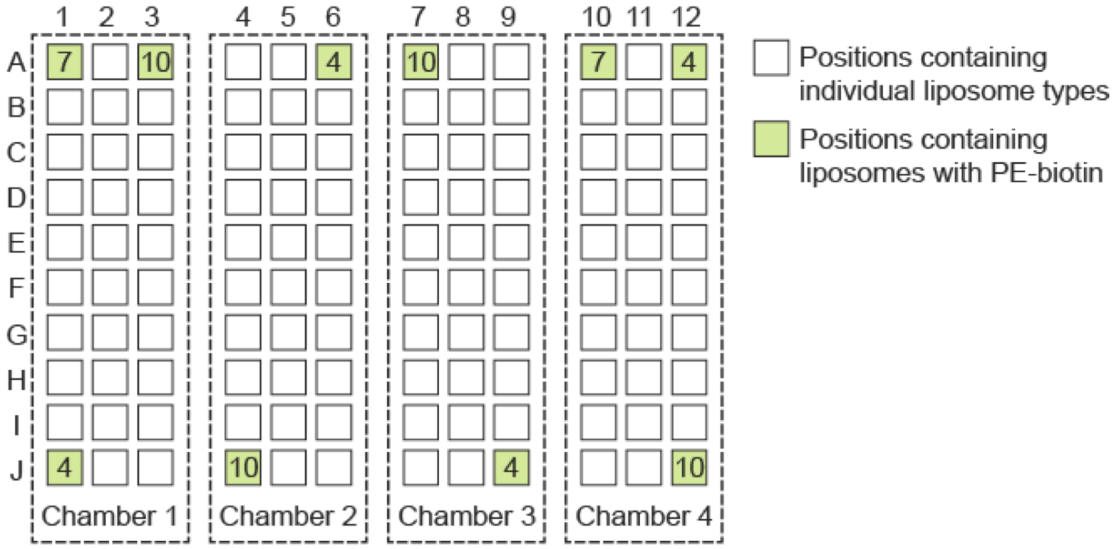


Figure 7: Schematic representation of LiMA's chip. The green spots represent positions onto which control liposomes (with PE-biotin as lipid of interest) were automatically spotted. These liposomes were always spotted on these positions (as opposed to the other liposomes whose positions were shuffled between replicates). The number indicate the concentration of the lipid of interest within the liposome (mol %).

2.2.2 Proteins

This screen focused on all known and predicted PH domains in *Saccharomyces cerevisiae* (58 PH domains from 49 proteins) and their orthologues in a thermophilic fungus, *Chaetomium thermophilum* as identified in [92] (27 PH domains from 24 proteins). We additionally selected six mammalian PH domains and, for comparison purposes, four unrelated LBDs with known and distinct lipid-binding specificities (Figure 8).

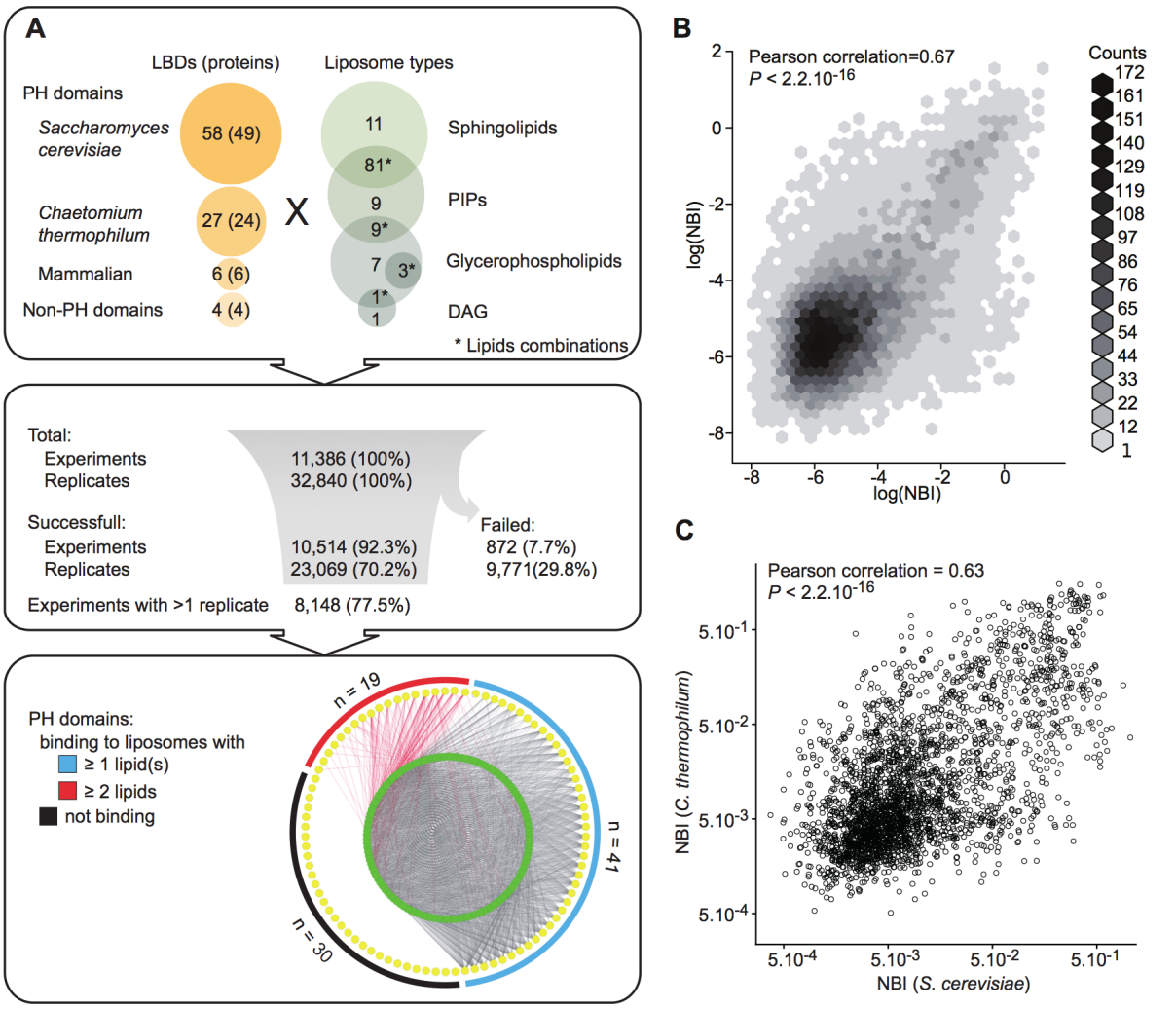


Figure 8: Presentation of the dataset (A) Upper panel: Selection of lipid-binding domains (LBDs, number of sourceproteins in parenthesis) and liposomes covered in the study. Middle panel: Summary of experimental success rate. Lower panel: Summary of PH domain–liposome interactions detected. Inner circles are liposomes, outer circles are PH domains and lines represent high confidence binding events. (B) Quantitative reproducibility of the screen. The Pearson correlation of NBI values is measured for all corresponding replicates for each LBD–liposome experiment. Counts represent the number of measurements per each hexagon in the matrix. (C) Correlation of lipid-binding profiles of 27 pairs of orthologous PH domains from *S.cerevisiae* and *C.thermophilum*. The Pearson correlation of all corresponding normalized binding intensity (NBI) values measured for all orthologous PH domains.

2.2.3 Interactions

The experimentalists yielded 11,386 unique domain-liposome experiments in biological replicates. They have acquired and processed a total of 229,880 images corresponding to 32,840 domain-liposome experiments. They processed a total of 229,880 images that were visually inspected and experiments that were unsuccessful – for example, because of protein precipitation, failure to produce liposomes or to acquire focused images – were labeld as such. This constitutes the unfiltered dataset that still remained to be filtered from the unsuccessful experiments.

2.3 Preprocessing of LiMA readouts

2.3.1 Filtering

All images were visually inspected and unsuccessful experiments – e.g., protein precipitation, failure to produce liposomes or to acquire focused images – were accordingly annotated and removed from the dataset. To facilitate the annotation process, all images were displayed on an internal web page, thanks to kind help of Dr Karl Kugler. Altogether, 29.8% (9,771) of collected data was removed. The final dataset consists of 10,514 unique protein–liposome experiments (92.3% of those designed), comprising 23,069 replicates (i.e. on average 2.7 replicates per protein–liposome pair, Figure 8). The subsequent analysis was based on this final dataset

2.3.2 Reproducibility of the assay

Quantitative reproducibility

The first step of the technical analysis pipeline consists in computing the quantitative reproducibility of the screen. The qualitative reproducibility was measured by plotting the NBIs of the replicates of each interaction against each other and by measuring their correlation. It yielded a pearson correlation coefficient of 0.67 with a P value $< 2.2.10^{-16}$. This indicates a very strong and significant correlation of the NBIs outputted by LiMA which nonetheless seem to display quite a bit of variability (Figure 8). To know whether this variability is synonyms with variability in the biological interpretation of the interactions, we performed a qualitative reproducibility of the assay.

Qualitative reproducibility

The screen benefits from manual annotations that have the advantage of assessing the true value of an interaction regardless of it s associated NBI. The annotator cannot take the replicate variability into account and thus, the annotation is very likely to reflect the true biological information of the picture, that is, if it is a binding event or not. Thus, the qualitative reproducibility assesses the propensity of LiMA to reproducibly output the same biological information for a given interaction regardless of the NBI variation that might be observed between replicates and that could be caused by numerous factors (detailed in paragraph 2.3.3). The overall reproducibility was 91.2% for distinguishing between interacting and non-interacting LBD–liposome pairs, measured on the set of 8,148 experiments for which replicates were available (Figure 9).

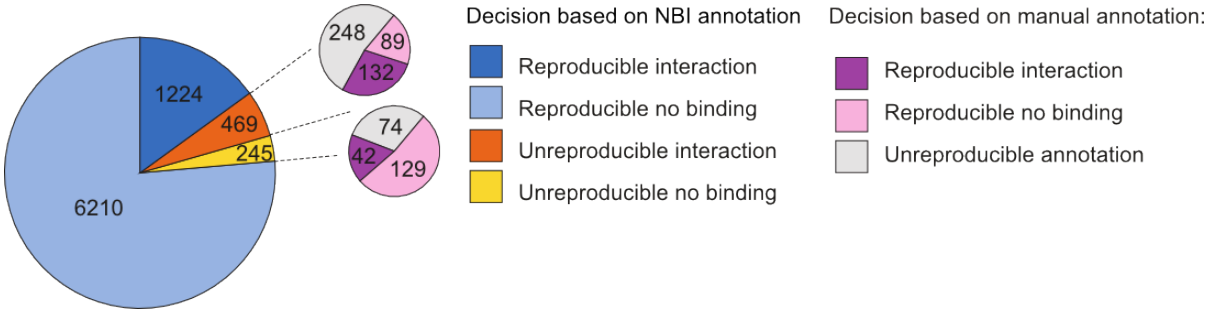


Figure 9: The big pie chart represents how often an interaction yielded NBIs above or below the binding threshold. The small pie charts assesses how often inconsistent NBI derived interpretation of the binding event were in fact reproducibly manually annotated.

2.3.3 Possible bias

The NBI is subject to be influenced by many different parameters, such as the protein concentration, the liposome number per well, the chip itself or the time spent between two spotting. The technical analysis pipeline considers these parameters and alert if one parameter is found to correlate with or to influence the NBI.

Protein concentration effects

For each interaction, the replicate of one protein have been incubated at different concentrations. The protein concentrations used for this screen ranged from 1 to 100 μM [Supplementary Table V.4]. The protein concentration effects were assessed in two different ways. First, we looked at the global correlation between the protein concentration and the NBI collected for all of the screen data (Figure 10). Second, we zoomed in and analyzed the protein concentration effect for each individual interaction (See Materials and Methods 2.5.2 and Figure 11). We found that the majority (92.4%) of these substantial concentration changes did not significantly affect the NBI measured. This indicates that proteins were present in the assay at saturating concentration (or for the proteins that did not show any binding event, that the protein concentration used was inferior to the one required to detect any possible binding event detection). As such, the measured NBIs were mainly dependent on their binding constant and were not influenced by the protein concentration.

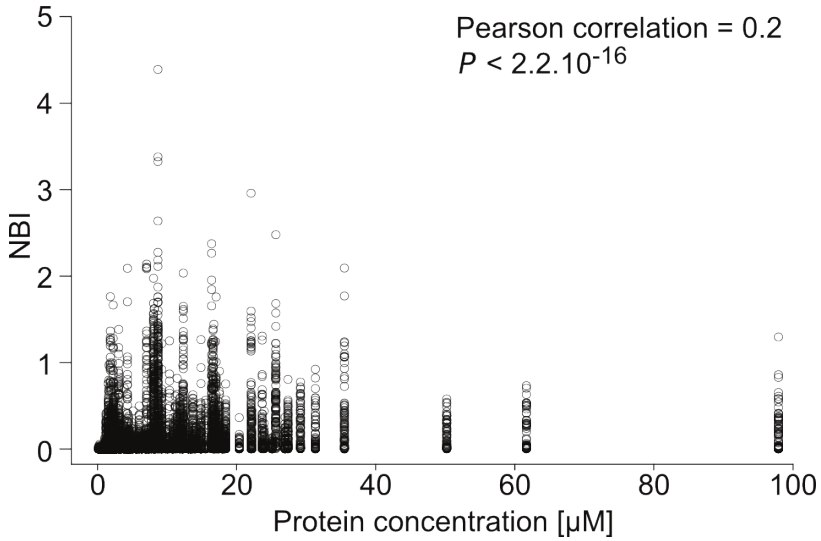


Figure 10: Correlation analysis of LBDs concentration versus NBIs of all experiments performed in the screen.

Lipid concentration effects

Some lipids were present in liposomes at two different concentrations (7 and 10 mol %) : DHS-1P, DOPI45P2, DOPI45P2:DHS-1p and DOPI45P2:DOPS. In order to assess the lipid concentration effects, we collected the NBI values measured for each of each of these lipids (Figure 12). We then computed a Wilcoxon test followed by a Benjamini-Hochberg correction test to assess if the NBI collected at the different lipid concentrations for each of these lipids was statistically significant.

Liposome number effects

The assay comprises 122 liposomes, including 94 mixtures made out of 27 lipids [Supplementary Table V.3]. The formation of liposomes is known to be difficult with certain lipids including DOPI345P3, which is shown in the figure (Figure 3). Given that the number of liposomes obtained can be significantly different from one mixtures to another, we also checked whether these differences also influenced the NBI measured by plotting the NBIs against the liposome numbers(Figure 13). No significant correlation was found, which implied that the NBI did not need to be normalized with regard to the liposome number.

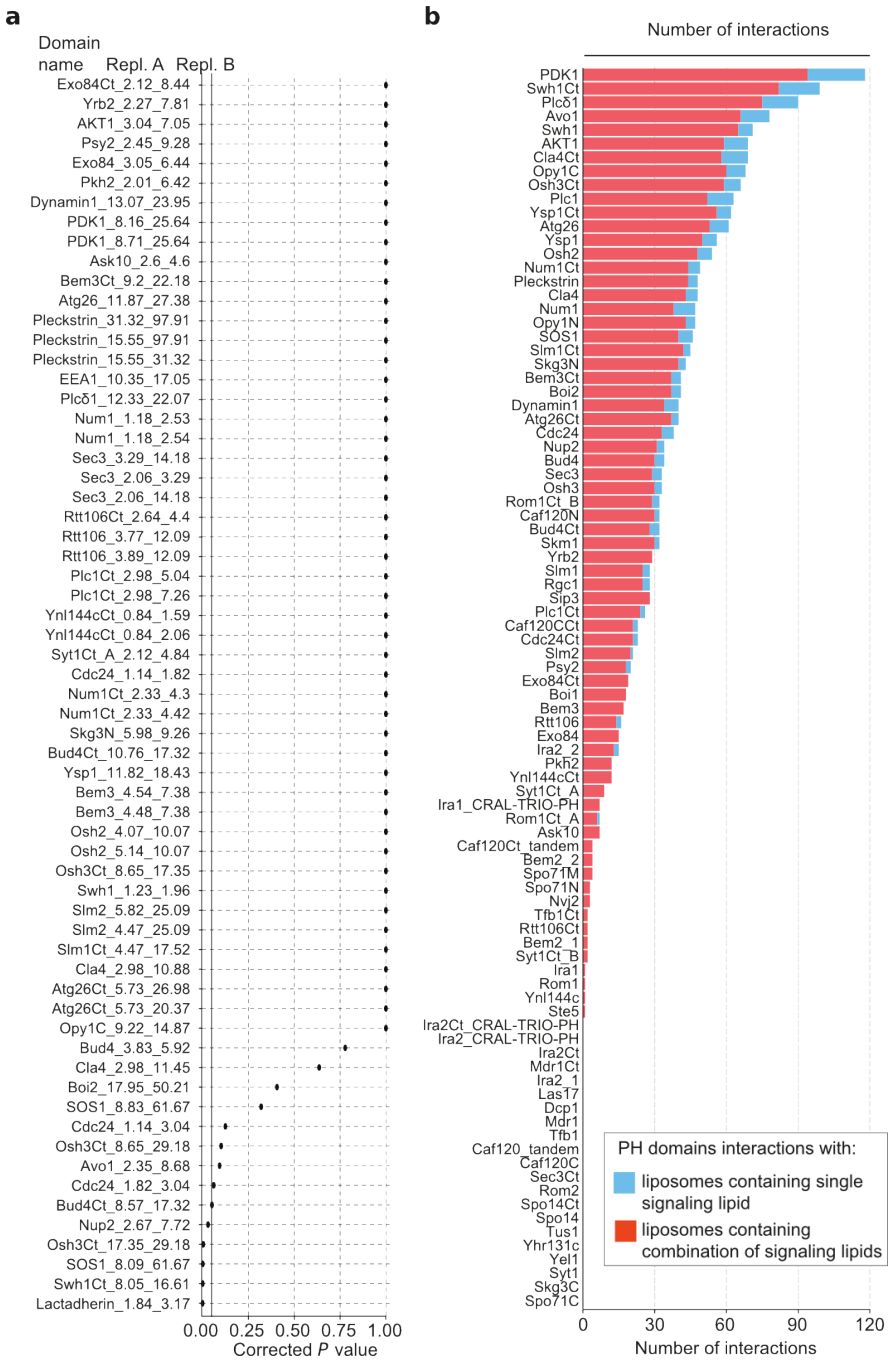


Figure 11: Protein concentration effects (a) Pairwise correlation analysis of NBIs for LBDs with more than 1.5 fold difference in protein concentration between replicates. The NBIs from the corresponding replicates were compared pairwise (Wilcoxon test followed by Bonferroni correction) and corrected P values, reflecting statistical significance of difference in NBIs, were plotted for each pair. P value 0.05 (solid line) was used as a threshold for statistically significant difference. The domain name together with the concentration (μM) of the two corresponding replicates (Repl.) is given. (b) Number of interactions detected per PH domain. The proteins for which data from only one replicate are available are labeled in red, the proteins for which the mean TI of the interactions was ≥ 2 are marked in blue.

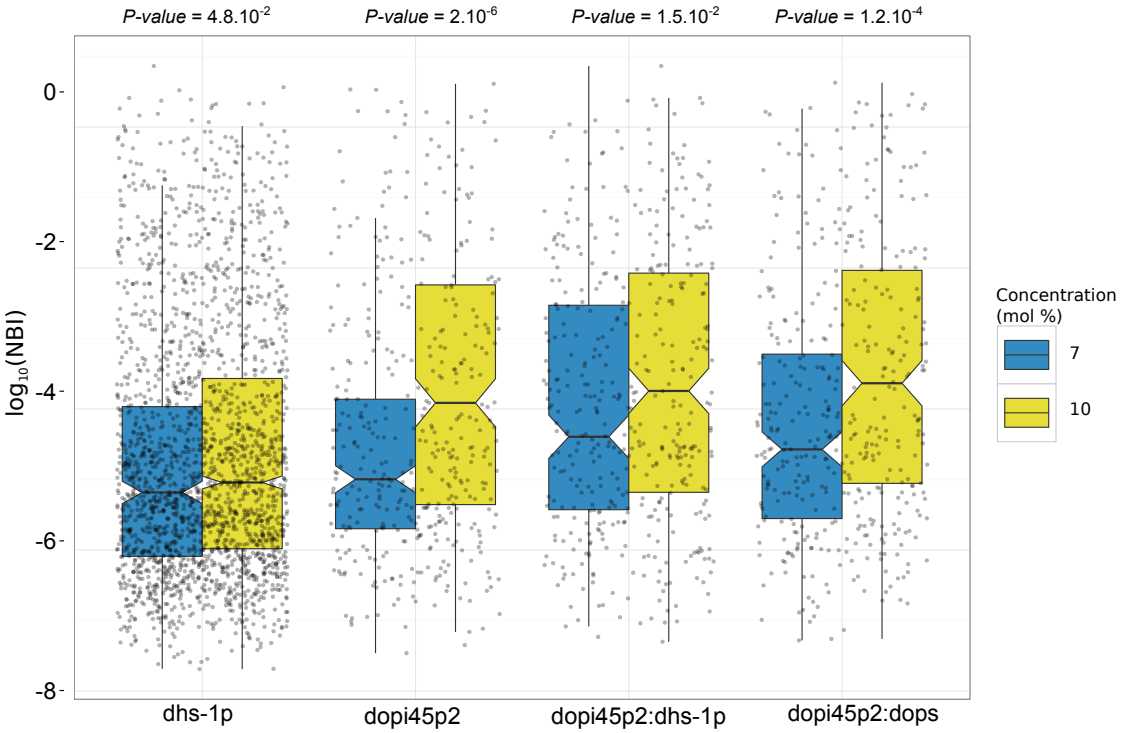


Figure 12: Boxplot representing the influence of the increased concentration of the lipid of interest within four different liposomes on the NBI measured.

Spatial effects

Chip based experiments provide a huge amount of data that can sometimes embody a significant amount of measurement error or bias due to technical issues directly tied to the chip itself. The scientific community has already reported various sources of spatial artifacts [93] [94] [95] which include for instance droplets, scratches, uneven washing or non randomization of the chip [93] [94] [96] [97] [98]. The technical analysis pipeline includes a module to check whether the chips used for the assay are spatially biased. The Figure 14 illustrates the mean NBI collected on each spot of the chip used and demonstrate that chip used were not spatially biased, which implied that no spatial normalization was required.

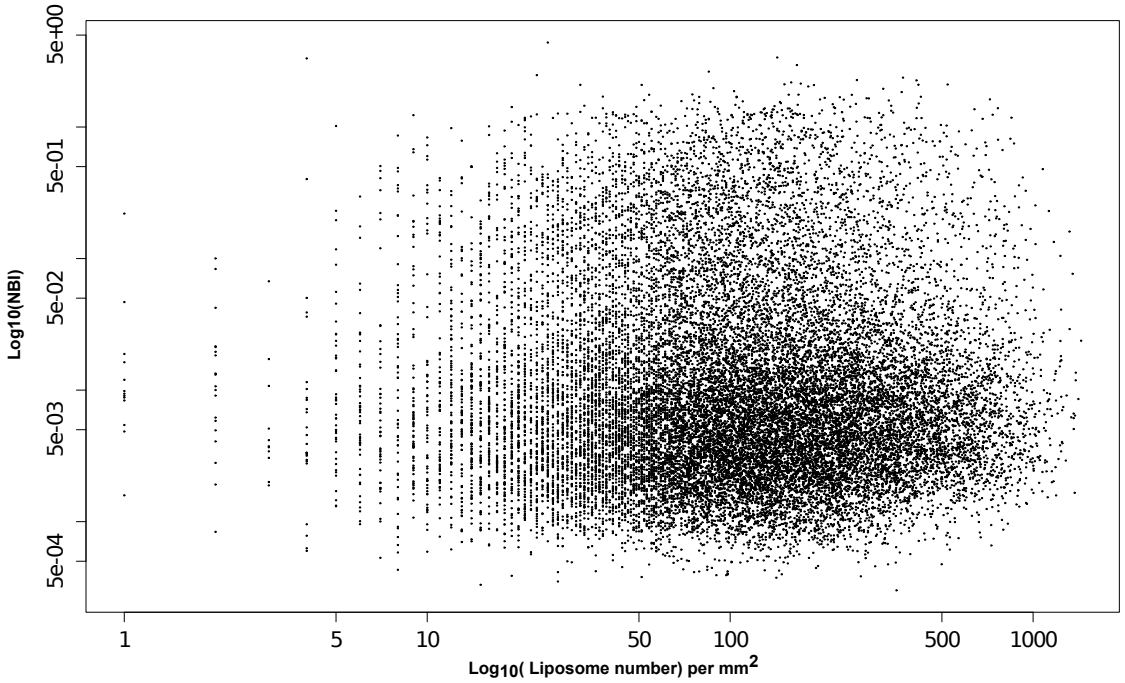


Figure 13: Correlation analysis of liposome number versus NBIs of all experiments performed in the screen.

Effect of the spotting time on the NBIs measured

The placement of the majority of liposome used in this assay was randomly shuffled as suggested in [99], in order to control for spatial bias. Besides, PIPs are known to hydrolyse quickly and the time spent between the measurements of two consecutive replicates could in theory have had an impact on the protein recruitment and hence the NBI measured. We measured the importance of this effect as explained in the paragraph 2.5.2 of Materials and Methods and found that there was no correlation whatsoever between the time spent between the spotting of two replicates and the NBI measured (Figure 15).

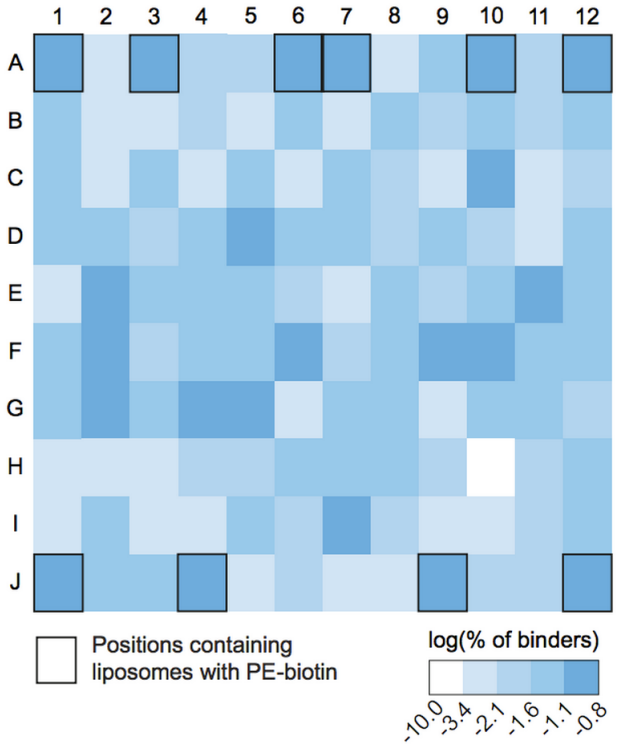


Figure 14: Control for potential position-dependent artifacts on the liposome array. Each cell of the heatmap corresponds to a position on the liposome array and indicates the logarithmic value of the ratio between the number of experiments of NBI > 0.037 (i.e., interactions) and the total number of experiments measured at the particular coordinate. Squared cells represent the positions of the positive control (liposomes containing a PE-biotin lipid), which remained fixed across all replicates.

2.3.4 Extraction of a binding threshold

In order to define a binding event based on the NBI, we performed a ROC analysis using the manual annotation as "golden standard" [100] in order to extract an NBI threshold above which an interaction is predicted as binding event. This threshold is used as interaction predictor for the rest of the data analysis. The binding threshold extraction module allows to select the NBI threshold either by maximizing accuracy (Figure 16) or by maximizing the sum of sensitivity and speficitity(Figure 16). The threshold was set to the NBI value of 0.037 according to the extraction method algorithm chosen [see Materials and Methods paragraph 2.5.2]. This chosen ratio yielded a 75.2% TPR and 3.4% FPR and a recall of 86%, with an AUC of 0.95 (Figure 16). The final NBI value for each domain-liposome experiment was calculated as a mean NBI from all available replicates. The mean NBI values were used for further analysis.

2.3.5 Assessment of binding quality

The assay yields very reproducible readouts (Figure 9) which display nonetheless quite a bit of variability (Figure 10) . Some interactions yielded replicates whose standard error was almost three quarters their mean values. This needed to be taken into account. To this aim, we have developed a Trust Index. The trust index (TI) is computed for each interaction that was measured at least twice. For each interaction between a protein domain and a liposome type, we calculated the mean NBI of the replicates (NBI) as well as the associated standard error (SE_{NBI}). We then calculated the TI of the interaction as such, where Th is the binding threshold (0.037):

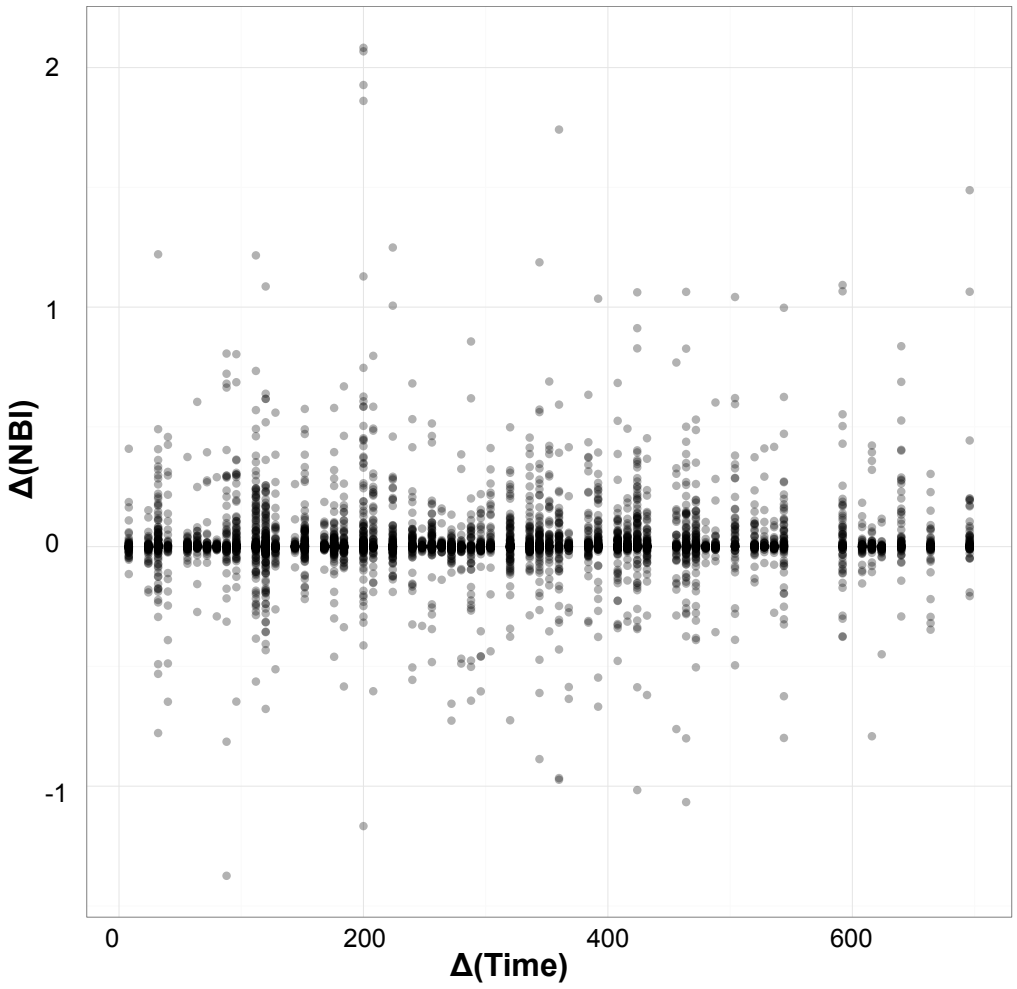


Figure 15: Visualization of the relationship between the time delay in the imaging (x axis)and NBI measurements between replicates (y axis)

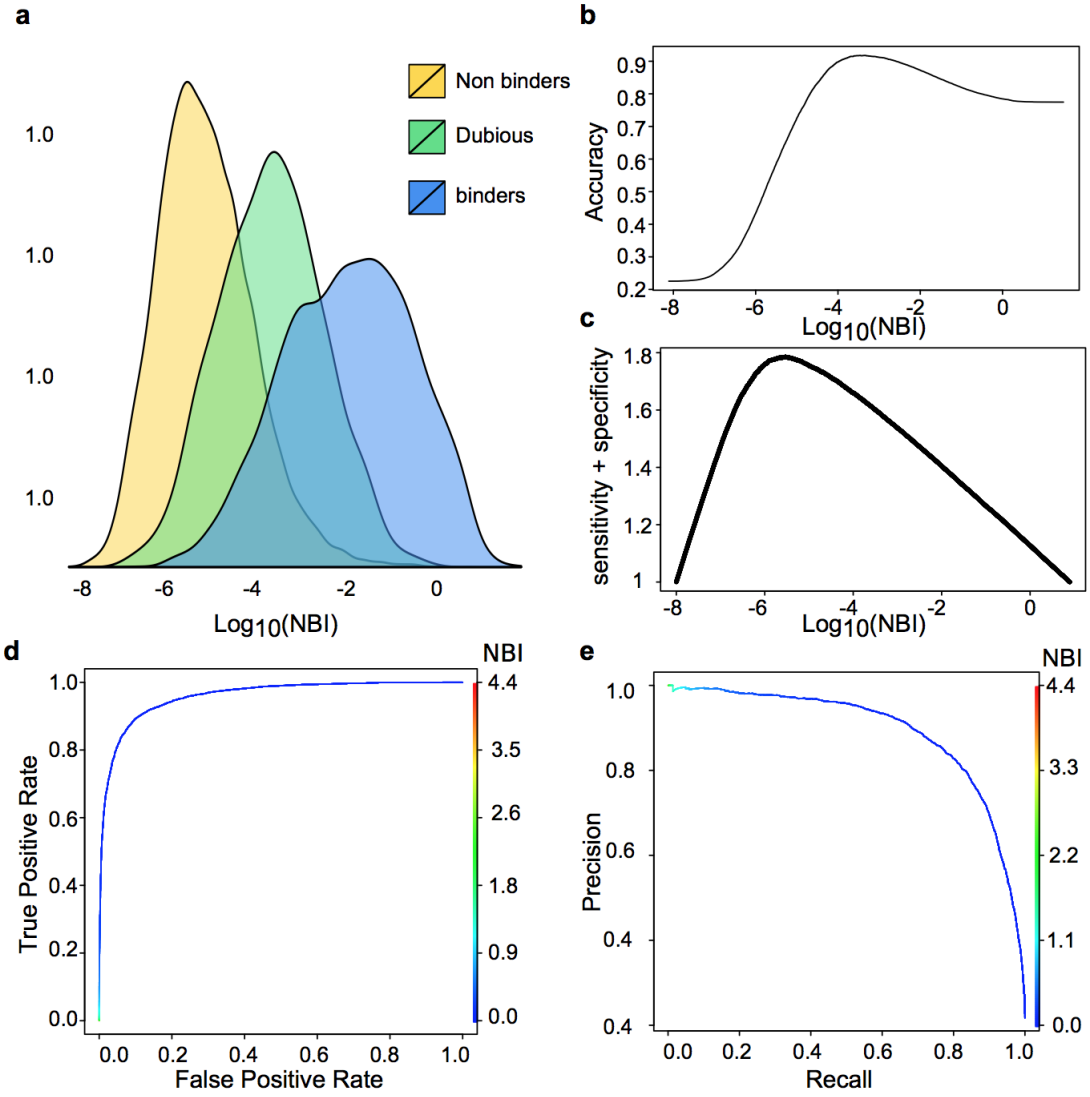


Figure 16: Background estimation and threshold extraction (a) Density plot showing the repartition of NBI of the three population of interactions.(b) Extraction of threshold by maximization of the accuracy. (c) Extraction of threshold by maximization of the sum of sensitivity and specificity. (d,e) ROC curve and precision-recall curve analyses of the NBIs of the screen dataset. The NBI cutoff extracted by maximization of the accuracy yielded an NBI of 0.037. This cutoff yields a true positive rate (recall) of 75.2%, a false positive rate of 3.4%, precision of 86% (dashed lines). AUC (area under curve) of ROC curve was 0.95.

- if NBI > Th : TI $= \sqrt{\frac{SE_{NBI}}{NBI-Th}}$
- if NBI < Th : TI $= -\sqrt{\frac{SE_{NBI}}{Th-NBI}}$

Based on the TI calculation, the experiments have been assigned to four categories:

- TI $\in$]0;2[: high confidence binder
- TI $\in$]2;inf[: low confidence binder
- TI $\in$]-2;0[: high confidence non binder

- TI $\in$]$-$ inf;-2[: low confidence non binder

The Figure 17 illustrates the computation of the TI. and the We provide the table of computed TIs for all the interactions that we probe [Supplementary Table **??**].

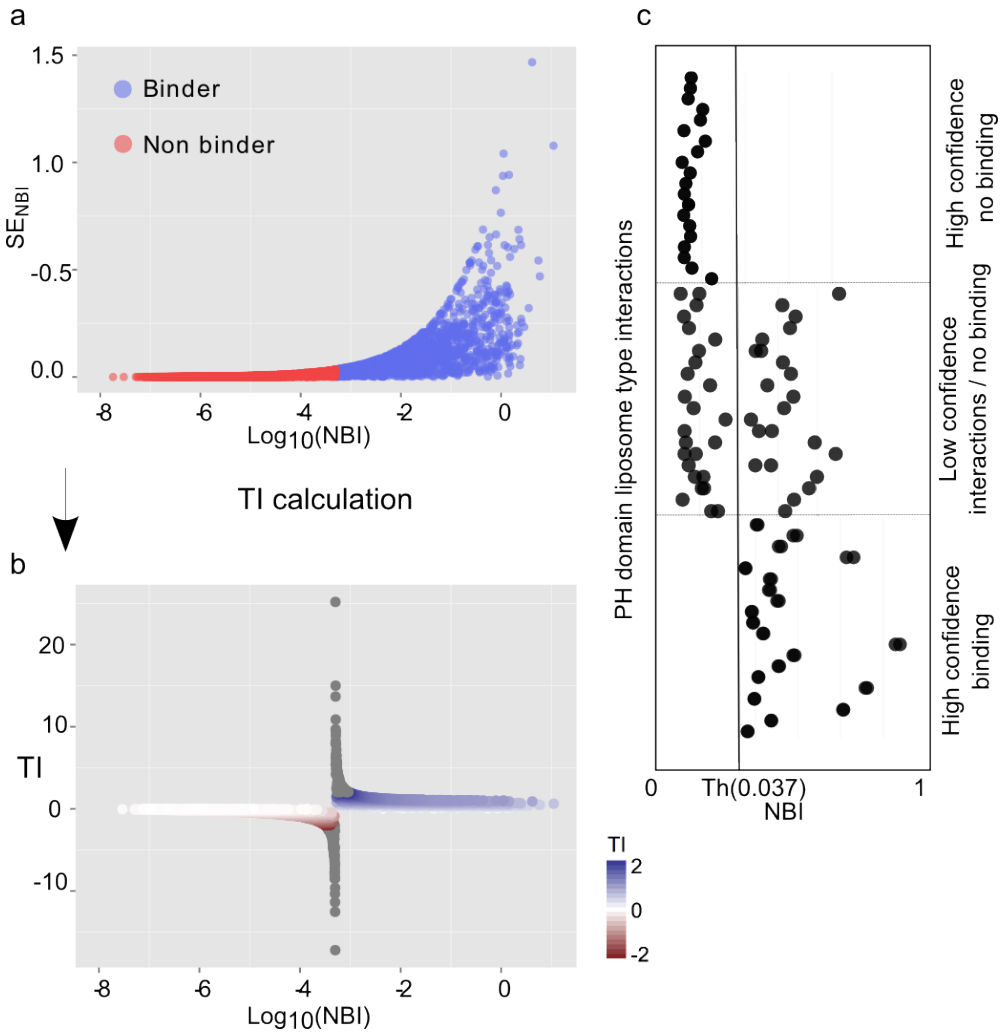


Figure 17: Determination of the trust index (TI) for each domain-liposome type experiment studied. (a) Scatter plot representing the standard error (SE_{NBI}) as a function of the NBI (log10 transformed). (b)Scatter plot represents the TI as a function of the NBI (log_{10} transformed). Interactions yield a positive TI while no binding events yield a negative TI. The closer the TI is to 0, the more confident is the datapoint (interaction or no binding). (c) Boxplot representing randomly picked examples of PH domain-liposomes type experiments which yielded different ranges of TI.

2.3.6 Sensitivty and specificity assessment

The definition of a binding threshold coupled to the TI allows us to make use of our controls by verifying their binding specificities to their known lipids. These should be both the strongest (in terms of NBI) and the best (in terms of TI) binder. The known specific interactions are listed in the supplementary information [Supplementary Table V.6, Supplementary Table V.7, Supplementary Table V.8,]. The Figure 18 confirms that the control proteins do preferably bind their known specific ligands and thereby, confirm that the technology LiMA reproduce very precisely the true biological binding properties of the probed proteins.

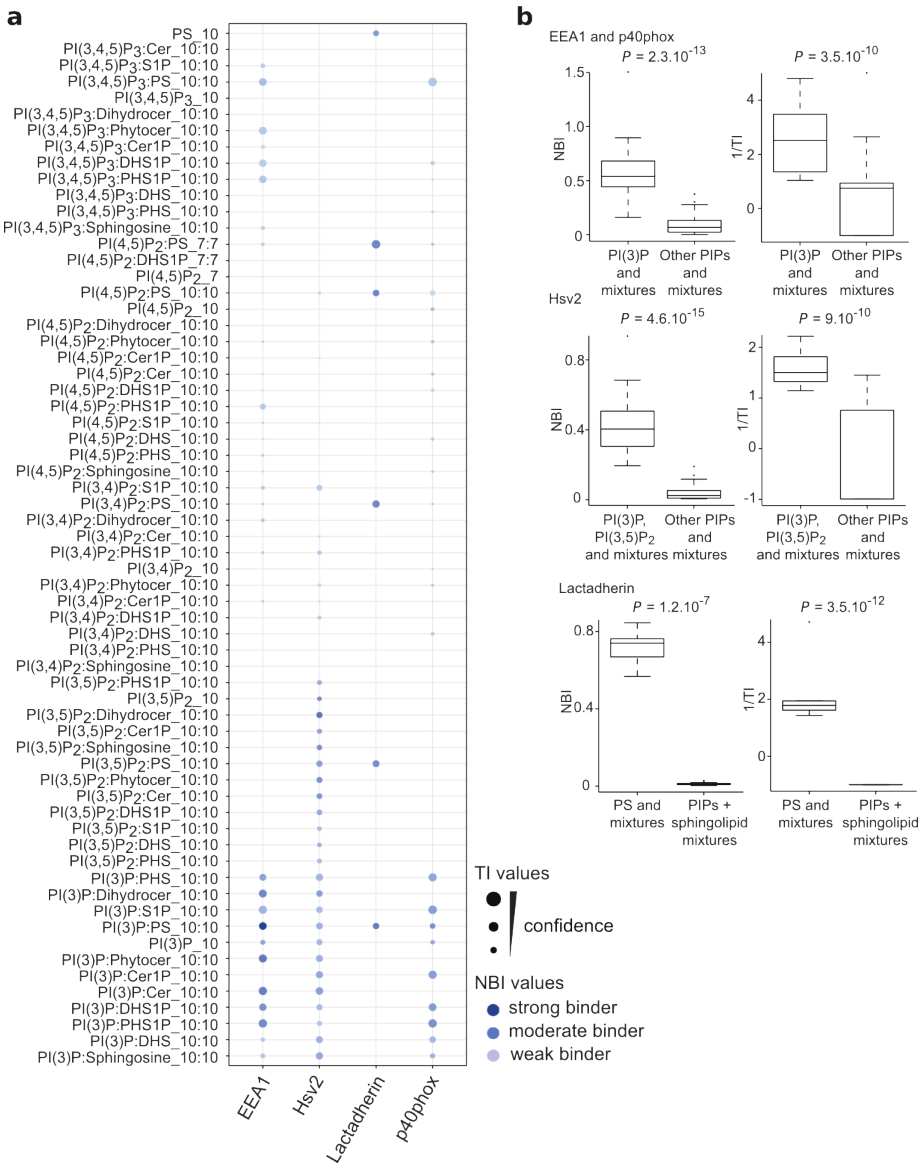


Figure 18: Sensitivity evaluation (a) Assessment of the ability of LiMA to recover the specific lipid-binding profile of four positive control LBDs (EEA1-FYVE, Hsv2, Lactadherin-C2, p40phox-PX). The numbers behind the lipid names indicate lipid concentration (mol %). (b) For the four positive controls shown in panel (a), the boxplots show the NBI and 1/TI for the known specific lipid(s) partners and all the other lipids and lipid mixtures.

2.3.7 Number of required manual annotations

The extraction of the binding threshold has been based on the manual annotation of the entire dataset. Three weeks were necessary to annotate the entire dataset (each annotator approximately annotated 7000 images per day). Screens including more replicates, proteins or liposomes could easily generate millions of pictures and this daunting manual annotation task could then rapidly become unfeasible. To avoid this fastidious task, we show that 500 randomly annotated pictures are enough in order to extract a binding threshold that yields a sufficiently high precision and sensitivity. Beyond this number, the precision and the sensitivity of the extracted binding threshold do not significantly increase (Figure 19).

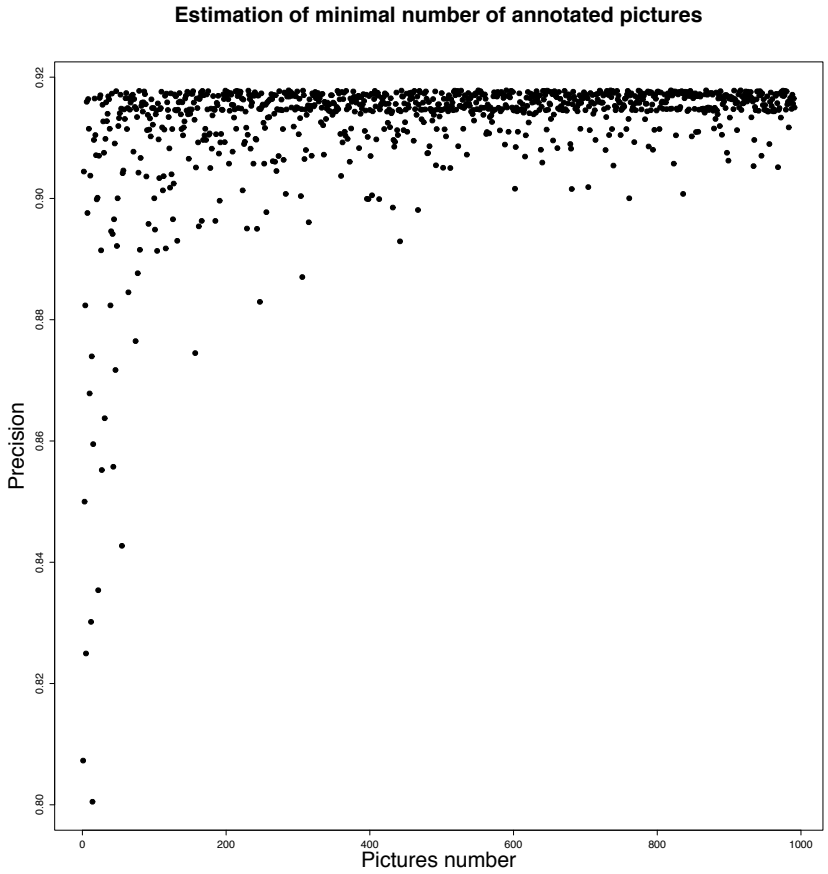


Figure 19: Estimation of the minimal number of annotated pictures required to extract an NBI binding threshold. For each number of picture, an NBI threshold is extracted in order to maximize the accuracy . The rest of the pictures are classified accordingly (as binder or non binder) and those classifications are compared with the manual annotations in order in order to compute the precision yielded by the selected threshold. The precision increases with the number of pictures used in order to extract the NBI binding threshold, but above 500 pictures, the confidence interval of the precision does not seem to increase anymore. Therefore, 500 manual annotations would suffice in order to extract a binding threshold that can be used for the rest of the dataset.

2.4 Functional analysis of LiMA readouts

2.4.1 Clustering analysis

Every PH domain studied in this screen was tested against an array of 122 liposomes. Thus, each PH domain had a fingerprint, characterizing its binding preferences to the universe of lipid tested. Corollarily, every liposome was tested against an array of 94 PH domains, and thus, each liposome had a fingerprint characterizing its binding preferences to the universe of PH domains tested. This screen, making use of LiMA, is the first high-throughput study focusing on protein lipid interactions under physiological concentration and capable of producing a quantitative readout that can be exploited in order to understand PH domains lipid ligand preferences, that is, the basic features giving them the ability to bind particular membranes, otherwise known as the "phosphoinositide code" [60]. The clustering analysis of single lipids of interest liposomes revealed that many more PH domains were capable of binding PIPs containing liposomes as single lipid of interest and with a higher intensity (higher NBI) as well as specificity (lower TI) compared to non PIPs containing liposomes(Figure 20). The clustering analysis of mixture liposomes revealed that PIPs are indeed the dominant PH domains ligands when present in an environment containing multiple lipids that the PH domains can potentially bind. Liposome mixtures containing identical PIPs members displayed a strong propensity to cluster together. This implied that they had a more similar PH domain binding profile (Figure 21 and Figure 22). More noticeably, two clusters seemed to stand out pretty clearly, which respectively contained either liposomes made out of PM PIPs and the other one, liposomes made out of Org. PIPs . The robustness of these clusters has been ascertained by various methods (Figure 22) which all led to the conclusion that these clusters were not random. This undoubtedly suggested that the nature of the recruited PH domains between those two types of liposomes was different. At last, the secondary lipid present in the mixtures also appeared to influence the protein binding profile of the liposome. The hierarchical clustering analysis displayed branches that were enriched in the presence of charged secondary lipids. These formation of sub-clusters did not perturb though the main clusters triggered by the presence of organellar and plasma membrane PIPs. This led to the conclusion that PIPs are the main lipids targeted by PH domains, but that the recruitment can be refined when put in the context of other charged lipids (secondary lipids).

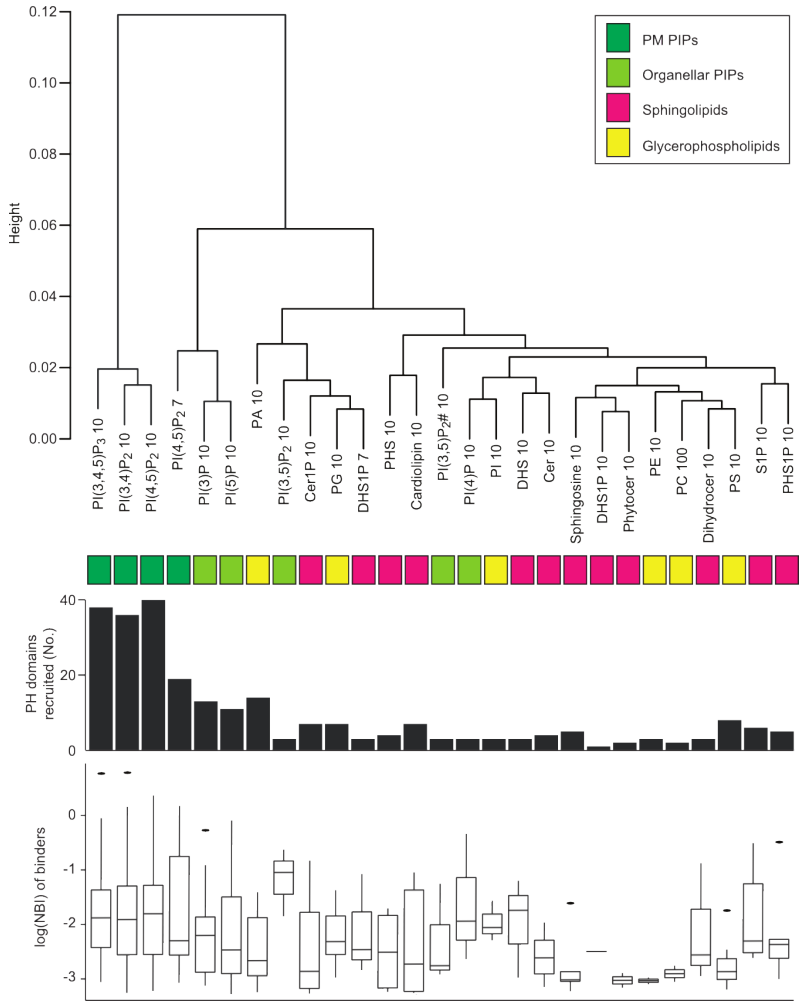


Figure 20: PIPs based liposomes are the preferred ligands (Top) hierarchical clustering of single signalling lipid- containing liposomes according to the similarities in their PH domain-binding profiles. Colours refer to the lipid family. The lipids that are not physiological in *S.cerevisiae* are marked with §. (Middle) barplot represents the total number of PH domains which bound to the liposome whose name is written on the leaf corresponding. (Bottom) boxplot shows the intensity with which the PH domain bound that liposome. The numbers behind the lipid names indicate lipid concentration (mol %) used for each signalling lipid.

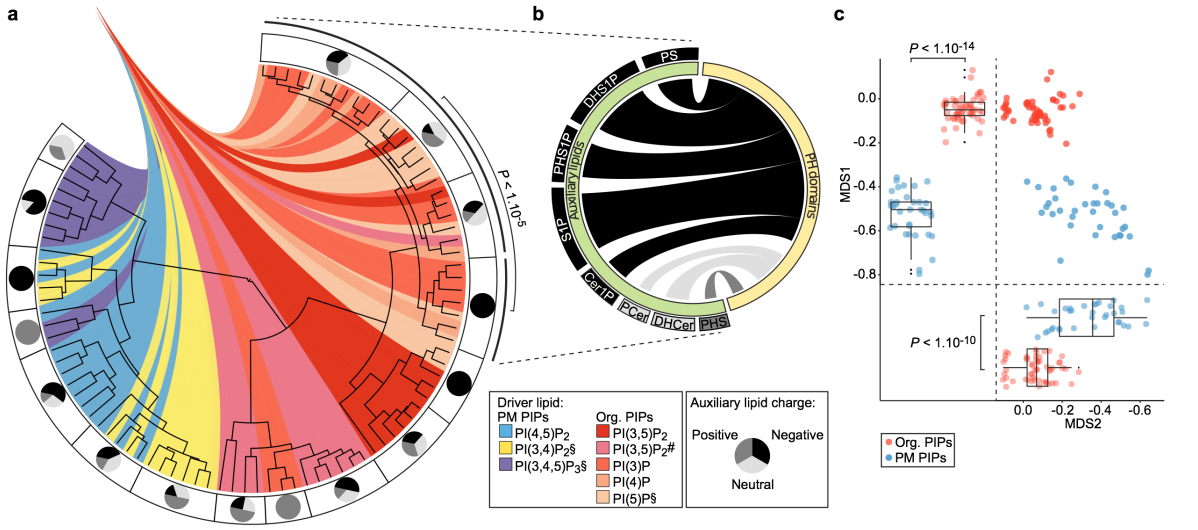


Figure 21: Clustering analysis of liposomes based on similarities in the recruitment of PH domains. (a) Hierarchical clustering of 90 liposome containing at least two lipids of interests.The color of the leaves refers to the PIP contained in the liposome. The pie charts illustrate the representation of charges of the secondary lipids. # indicates an inactive (in our assay) dipalmitoyl variant of DOPI35P2. (b) Effect of the auxilliary lipid's charges on the PH domains-liposomes interactome within the clades highlighted in panel (a). PCer, phytoceramide; DHCer, dihydroceramide. (c), Principal coordinates analysis of the same 90 liposomes mixtures used in (a). The PCoA was performed using the pearson metric based on their their PH domain binding profile using the NBI as a readout. Red dots represent liposome in which the PIP is known to locate in an organellar membrane. Blue dots represent liposomes in which the PIP is known to locate in the plasmic membrane.

2.4.2 A new motif behind the recruitment to Organellar PIPs

The results of the hierarchical clustering analysis concluded that PIPs were the lipid contributing the most to the recruitment of PH domains. This conclusion was supported by the fact that that ORG PIPs based liposomes [DOPI3P, DOPI4P, DOPI5P and DOPI35P2] displayed a completely distinct PH domain binding profile as PM PIPs based liposomes [DOPI34P2, DOPI45P2 and DOPI345P3], and that the presence of other charged lipids did not alter the structure of the observed clusters. A glance at the PIPs binding profile (Figure 24) of the 90 PH domains exhibited a group of 16 proteins capable of binding both ORG PIPs and PM PIPs based liposomes as opposed to another group of proteins that was solely capable of binding PM PIPs based liposomes. In order to find out which amino-acids or positions within the sequence of PH domain could explain the specific binding of ORG PIPs, we first performed a multiple alignment of the 90 PH domains. The sequences were labeled according to their ability to bind ORG PIPs based liposomes. We subsequently employed an algorithm optimized for the detection of sub-family specific residues. In brief, the algorithm loops over every position of the multiple alignment and identifies residues that are different amongst the sub-families present in the alignment, which in this case are the proteins differently labeled according to their ability to bind ORG PIPs based liposomes [101] The algorithm scored every position of the multiple-alignment (Figure 23) and nine positions were retained. More precisely, two positions were located in the β1-β2 loop, three at the end of the β7 strand, and four in the α-helix. These nine positions were those that explained most of the difference between the two groups of labeled proteins in terms of amino-acids content and were thus assumed to be responsible for the specific binding of ORGs PIPs liposomes. This constituted the organellar PIP recognition motif (ORPM) (Figure 24) which were not overlapping with the already described BSM. The BSM (Figure 24) was described as being the Sine qua non condition for PH domains to be able to bind PIPs in general [102].

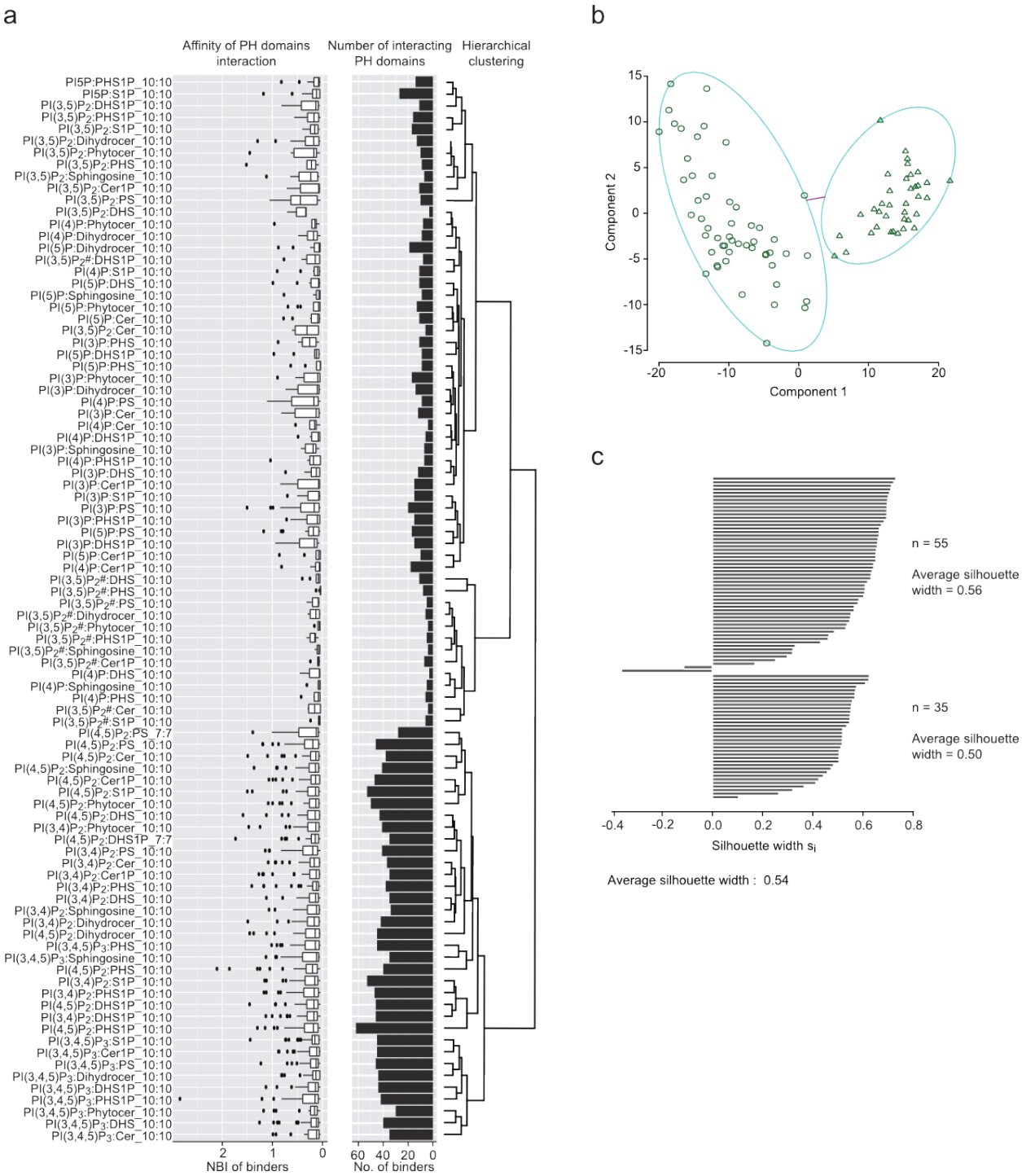


Figure 22: Evaluation of cluster's robustness (a) Detailed Hierarchical clustering analysis of the 90 liposomes containing at least two lipids of interest as described in Figure 25a. The barplot represents the total number of PH domains which bound to the liposome whose name is written on the corresponding leaf. The boxplot shows the intensity with which the PH domain bound that liposome. The numbers behind the lipid names indicate lipid concentration (mol %) used for each signalling lipid. (b) Cluster robustness assessment with k-medoids clustering of the same 90 liposomes using the partition around medoids algorithm. The triangles represent ORG PIPs liposomes and the circles PM PIPs liposomes. (c) Cluster robustness assessment computing silhouette index. Each bar corresponds to a liposome and the group of bar labeled as n=55 corresponds to the ORG PIPs liposomes. Only two liposomes have been found in the wrong cluster.

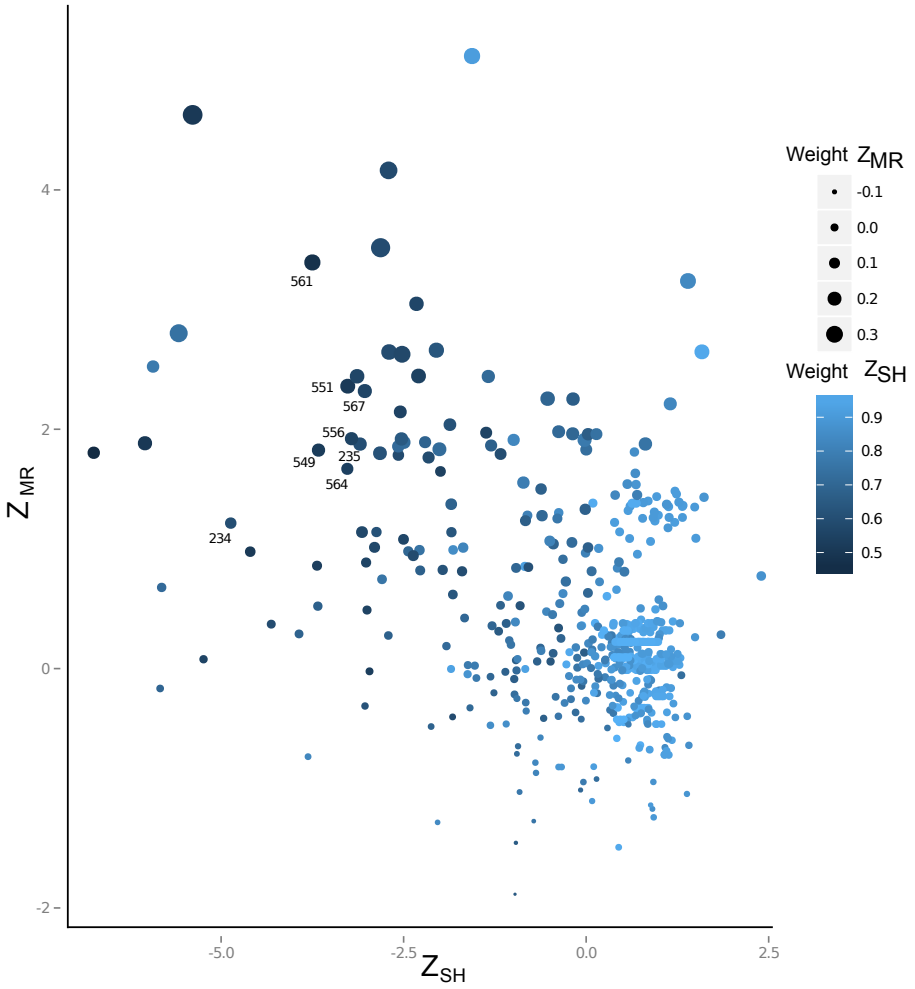


Figure 23: Scatter plot representing the results outputted by the multi harmony algorithm used for the motif extraction. The x axis is the sequence harmony (SH) score, which evaluates how conserved are the residue within each group. It ranges from 0 for completely non-overlapping residue compositions to 1 for identical compositions (in other words, the lower the better). The second statistic is a measure of statistical significance of the SH score (the Z_{SH} score). The Z_{SH} score is calculated by permuting the group labels and running the sequence harmony algorithm for 100 randomizations. The mean and standard deviations of these permuted groups are calculated. A Z_{SH} score of 1 means that the SH score is 1 standard deviation below the random mean (the more negative the Z_{SH} score, the better). The y axis the multi-relief weight (MR score), which evaluates how discriminatory are residues between each group. It ranges from 1 for completely discriminatory residues to 0 for non discriminatory residues (so the higher the better). The fourth statistics is a measure of statistical significance of the MR score (Z_{MR}). The Z_{MR} is calculated by permuting the group labels and running the multi-relief algorithm for 100 randomizations. The mean and standard deviation for these permuted groups are calculated. A Z_{MR} score of 1 means that the MR score is one standard deviation above the random mean (the more positive the Z_{MR} score, the better). At the end, each position of the multiple sequence alignment these four statistics assigned.

The residues are conserved

The identification of the OPRM involved a group of 16 PH domains that all belong either to *S.cerevisiea* or *C.thermophilum*. In order to claim that this motif was universal, we checked for its presence within the entire annotated and reviewed PH domains universe, that is, 1216 PH domains that belong to 72 different species [103]. We performed a multiple alignment of these 1216 sequences as we did for the 90 PH domains and applied a scoring scheme to each PH domain based on the presence or absence of the residues belonging to both motifs (BSM and OPRM) (Figure 24). In order to confirm that the presence of the OPRM amino acids is linked to the propensity of PH domains to bind ORG PIPs, we randomly selected ten PH domains that yielded different scores from the species Homo Sapiens. Five of these PH domains yielded contained most of the twelve amino-acids belonging to both motifs (OSBP1, OSBP2, OSBPL3, OSBPL7, FAPP1), three yielded a medium score (CERT, OSBPL11, OSBPL10), and two with almost no amino-acids belonging to the motifs (OSBPL8, OSBPL5). The confirmatory experiments validated the role of the amino-acids present in both motifs (BSM and OPRM) in the respective recognition of PM PIPs based liposomes and ORG PIPs based liposomes (Figure 24)

Making sense of the motif

In order to make sense of the two described motifs (BSM and ORPM) and understand what were their biological implications, we performed an enrichment analysis on the 50 proteins that yielded the highest score (Figure 24). The most enriched function found was transport between different cellular compartments. This further validates the implication of the ORPM in the specific recognition of ORG PIPs based liposomes (and thus organellar membranes) as proteins required for the non-vesicular transport of lipids, are known to be found at various membrane contact sites and are responsible for the transport of lipids from one membrane to the next ([104] [105] [106] [107])

In vivo validation

In order to confirm the implication of the OPRM in the specific recognition of organellar membranes in vivo, we selected PH domains which yielded different scores (Figure 25) and marked them with a fluoerescent tag (mCherry). As expected, FAPP1 and Swh1 (which both yielded a very high score) did locate at the intracellular membranes, and notably at the Golgi apparatus. These results confirm what has been long observed about lipid transporters, that is, that they are often found to locate to Golgi membranes [105] [108] [109]. PH domains like CERT and Opy1c (which yielded low scores due to the lack of OPRM, but nonetheless presence of BSM) did not –as expected- locate to the organellar membranes but to the plasma membrane.

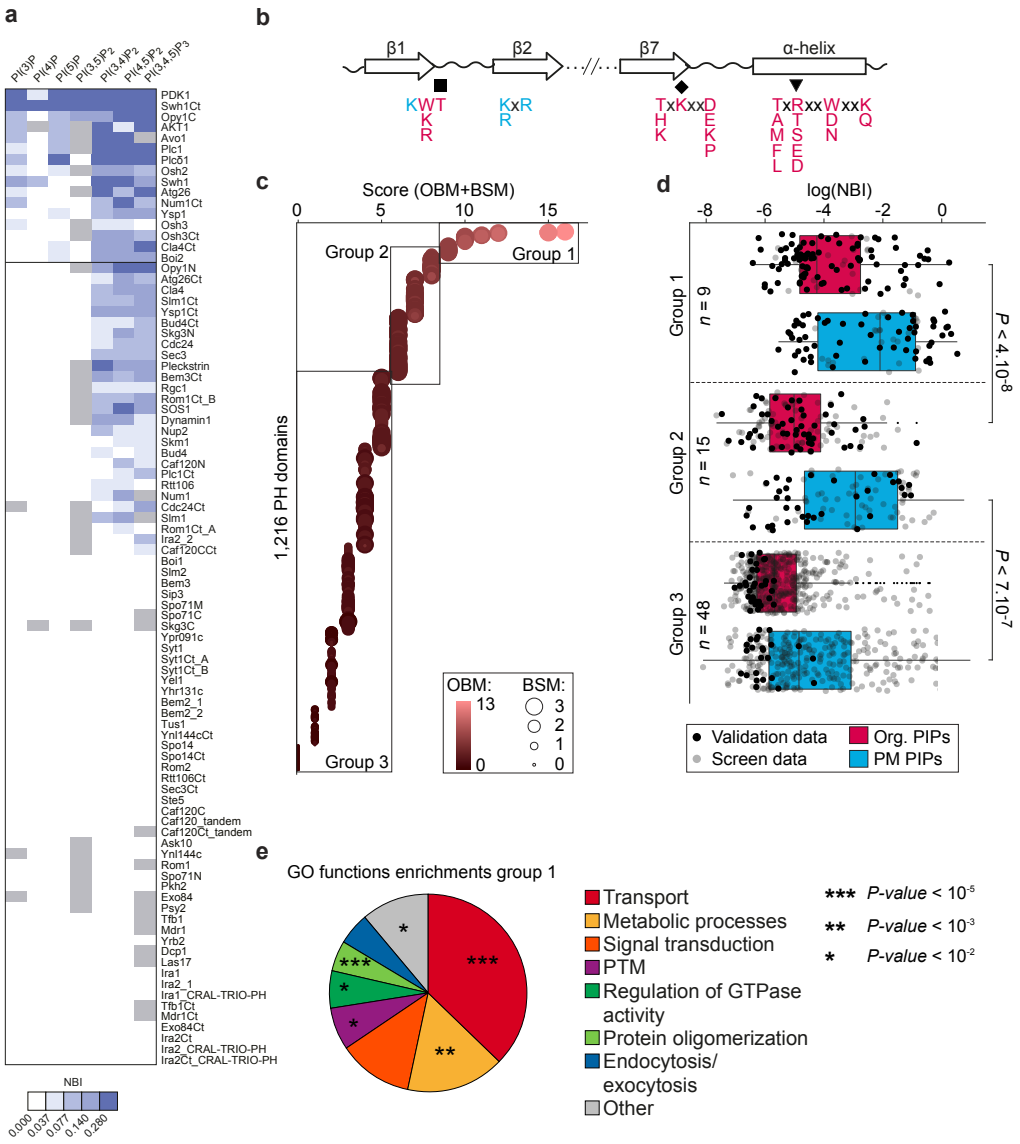


Figure 24: Motif confirmation (a) Heatmap representing the PH domains PIP binding profiles. Columns represent seven liposomes containing one lipid of interest. The four on the left contain lipid of interests that are located in organellar membranes. The three on the right contain lipid of interests that are located in the plasmic membrane. The line delimitates PH domain capable of binding ORG PIPs (foreground) as opposed to those which cannot bind ORG PIPs (background) (b) Representation of a part of the PH domain secondary structure as well as (blue) the BSM and (red) the motif derived from the foreground (c) Ranking of the entire PH domain universe based on the presence or absence of the residues displayed in panel (b). The higher the score, the more residues are found in a given PH domain (d) The boxplot shows NBI values measured for experiments of 72 PH domains (10 human, red dots; 62 yeast, black dots) with either organellar PIPs (pink) or PM PIPs (blue). The PH domains are separated into 3 groups based on the score shown in the plot on the left. n indicates the number of PH domain in each group. The darkness of the dots indicated their annotation as interaction (dark) or no binding (light). ** P value $< 9.10^{-11}$, * P value $< 4.10^{-5}$ (e) Functional enrichment analysis of Group 1 (from panel(c)) using gene ontology annotations.

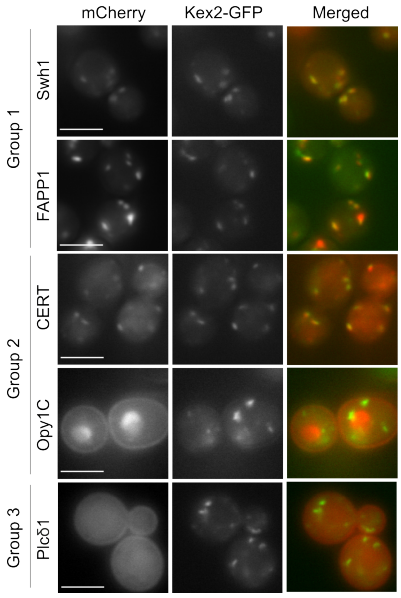


Figure 25: In vivo validation of the motif part 1. The selected PH domains were expressed as mCherry-fusions in yeast strain with endogenously expressed trans-Golgi marker Kex2-GFP. The left column shows localization of the PH domains, middle column the localization of Kex2 protein and right column shows both images merged. The PH domains are separated into three groups that correspond to their scoring in 24. Scale bar, 3μm.

To validate further the role linked to the OPRM, we performed point mutations. We first selected an amino acid located in the β7 strand (lysine), as it appeared to be one of the most conserved amino acids of the OPRM (it was the second most conserved amino acid within the group of 16 proteins of our screen which were capable of binding ORG PIPs based liposomes) (Figure 26). Additionally, the 3D simulations indicated that this amino acid could be directly in contact with the lipid ligand. Lastly, the HMM corresponding to the PH domain family found in the Pfam database indicated that the most common amino acid present at this position is a glutamine. As a consequence, we decided to mutate this amino acid to a glutamine in three different PH domains (Swh1, FAPP1, CERT) and to measure the impact this mutation had on the binding to PIPs-containing membranes [Supplementary Table V.9]. As expected, the mutated proteins lost their ability to strongly bind ORG PIPs based liposomes, whereas the interaction with PM PIPs based liposomes was not affected (Figure 26). The second mutated amino acid was located in the β1- β2 loop, and corresponds to the position which was the most conserved within the group of 16 proteins mentioned earlier (threonine). This amino acid is found mutated in the protein FAPP1 in colorectal carcinoma [110]. This threonine was mutated to an alanine and specifically affected the recruitment to DOPI4P containing liposomes. The recruitment to DOPI45P2 containing liposomes remained unchanged (Figure 26). The third mutated amino acid was an arginine found in the α-Helix. This residue is found mutated in the protein OSBP2 in lung adenocarcinoma [111]. Intriguingly, this residue is located far away from the two possible binding sites and consequently does not appear to be directly involved in the contact with the lipid ligand. Yet, this residue was the most statistically significant residue found by the multi-harmony algorithm [101]. The mutation did though specifically and significantly perturbed the binding of OSBP2 to ORG PIPs based liposomes whereas the recruitment to PM PIPs based liposomes remained unchanged (Figure 26).

All in all, this supports the notion that the recruitment of PH domains to membranes implies the selective recognition of PIP species which is important for the protein function [65] [112].

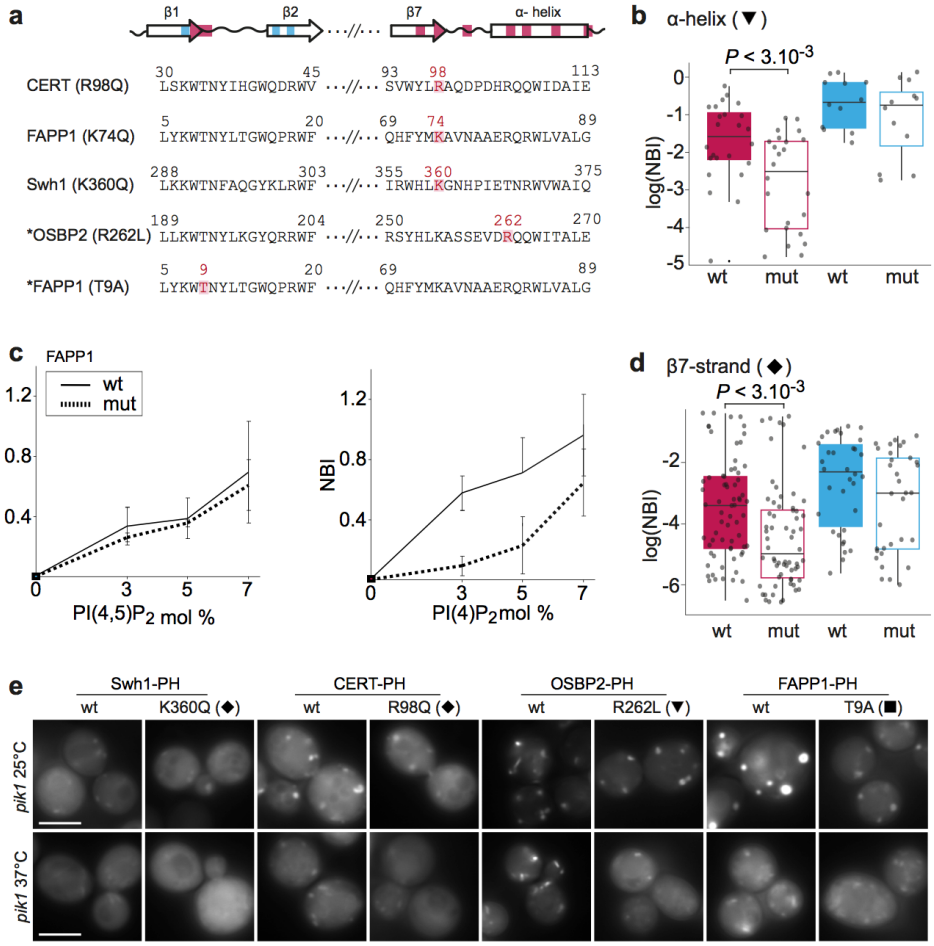


Figure 26: In vivo validation of the motif part 2 (a) Schematic representation of the PH domains secondary structures holding the motif along with five proteins chosen for the point mutation experiments. Highlighted residues represent the mutated residues. (b) Box plot showing the impact of the R262L (in the alpha helix of OSB2-PH) on Org PIPs liposomes (red) and PM PIPs liposomes (blue). (c) Impact of the T9A mutation (in the beta1-beta2 loop of FAPP1) on a PM PIP liposome (DOPI45)P2) (left) and on an ORG PIP liposome (DOPI4P2) (right) (d) Box plot showing the impact of the R98Q (in the beta 7 strand of CERT-PH) on Org PIPs liposomes (red) and PM PIPs liposomes (blue).

2.4.3 Lipid cooperativity

PH domains can bind cooperatively

The data collected from our screen, added to those from anterior studies [27], indicate that the presence or absence of particular lipids around of PIPs within biological membranes can affect the recruitment of PH domains to these membranes. Furthermore, our data indicate that a quarter of the screened PH domains necessitated the presence of several lipids in order to be able to bind these liposomes(Figure 27).

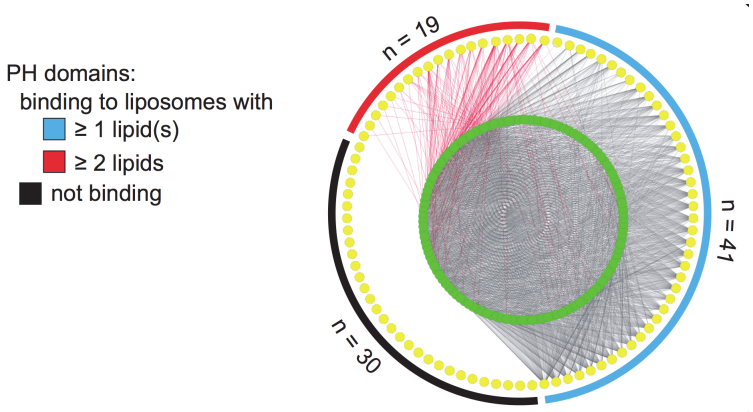


Figure 27: Summary of PH domain–liposome interactions detected. Inner circles are liposomes, outer circles are PH domains and lines represent the 1628 high confidence binding events. The PH domains from which red edges are drawn represent PH domains that could only interact with liposomes containing at least two lipids of interest.

An interaction between a protein and a group of lipid is defined as being cooperative, when the presence of a particular lipid increases the affinity of the binding event between a protein and its lipid ligand. This screen offers thanks to its quantitativeness and physiological conditions the first opportunity to assess whether principles of cooperative binding apply to PH domains. To answer this question, we have developed an algorithm in order to predict if an interaction between a PH domain and a mixture liposome is likely to be of a cooperative nature or not, that is, if the interaction with the mixture liposome is higher than what could have been expected from the two individual signaling lipids. We have focused the data analysis on PIP, sphingolipids and PS based liposomes, and restricted the prediction event to all pairs of physiological interactions. The algorithm predicted that the majority of PH domains was capable of cooperatively bind mixture liposomes (Figure 28) with a high associated trust index. The sensitivity and the specificity of these prediction was assessed by selecting 36 random interactions (predicted as being cooperative or not), we performed dose-response experiments (Figure 28 and Supplementary Table 2), from which 20 out of 25 predicted cooperative interactions were confirmed, and only 2 out of 11 predicted non cooperative interactions were confirmed. The reason that high false negative rate is explained by the screen settings, namely, that most of the lipids were tested at one concentration (10 mol %), which probably was under the threshold of any detectable interaction, or already at saturating concentration. As a matter of fact, the dose-response experiments reveals that 5 out of 9 of the falsely predicted cooperative events behave cooperatively at lower concentrations (< 10 %mol). The incidence of cooperative recruitment of PH domains to liposomes occurred with both PM and ORG PIPs based liposomes, but PM PIPs based liposomes displayed twice as more cooperative binding events. (Figure 28).

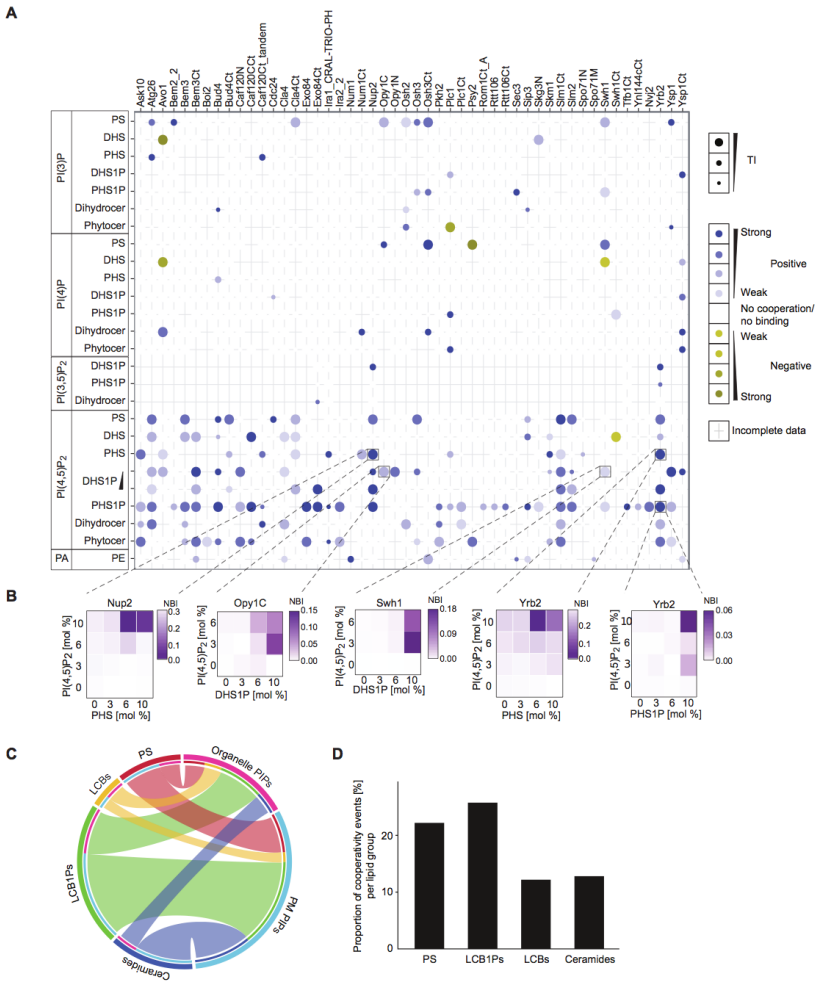


Figure 28: Cooperativity landscape of PH domains (a) Heatmap of cooperative indexes (CI) calculated for PH domains (columns) and membranes containing combinations of physiological signaling lipids (rows; the wedge indicates low and high lipid concentrations, respectively). Only 50 PH domains with high confidence CI > 1 for at least one liposome type are shown. (b) Dose-responses measured for the PH-domain interaction with liposomes containing the indicated concentration of signaling lipids. Values are means (n ≥ 2). (c) The summary of the propensity of different driver lipids (PIPs, right) and auxiliary lipids (left) to cooperate based on data shown in panel (a) Organelle PIPs and PM PIPs: as in Figure 2; Ceramides: Cer, Cer1P, Phytocer, Dihydrocer; LCB1Ps: S1P, DHS1P, PHS1P; LCBs: Sphingosine, DHS, PHS. (d) Cooperative interactions with regard to secondary lipids encountered. The bar plot gives the proportion of cooperative interactions of all experiments performed for each group of auxiliary lipids. LCB1Ps, LCBs and ceramides as in (c).

Impact on membrane recruitment of PH domains

This screen shows for the first time that the incidence of cooperative sensing of lipids is not a sporadic event but an inherent characteristic proper to PH domains, and this has many consequences. We have established a classification with regard to the effects induced in terms of specificity changes by the cooperative binding. The class 1 corresponds to an exacerbation of the PIP specificity induced by the presence of a secondary lipid. The class 2 corresponds to cases in which the presence of a secondary lipid induces a switch in the PIP specificity of the PH domains and 19 PH domains fell into that category

(Figure 29 and Supplementary Table 2). As an example, the PH domain of the kinase AKT1 was very specifically recruited by either DOPI345P3 or DOPI342 based liposomes. The affinity of AKT was yet switched to DOPI45P2 when this lipid was together in a liposome with LCBs (5 folds NBI) (Figure 29). This exacerbated affinity towards DOPI45P2 has already been observed for oncogenic variant of AKT1 carrying a mutation within the β1-β2 loop of its PH domain, triggering a targeting to the plasma membrane [113] and subsequently to its unregulated activation [73]. The PH domain of Bem3 – a GTPase-activating protein (GAP) for Cdc42, a key regulator of polarized cell growth [114] – is not found to be recruited by pure PIPs based liposomes [115]. However, it was found to strongly bind DOPI45P2 or DOPI345P2 based liposomes when a secondary lipid was present, e.g. phosphatidylserine (Figure 29).

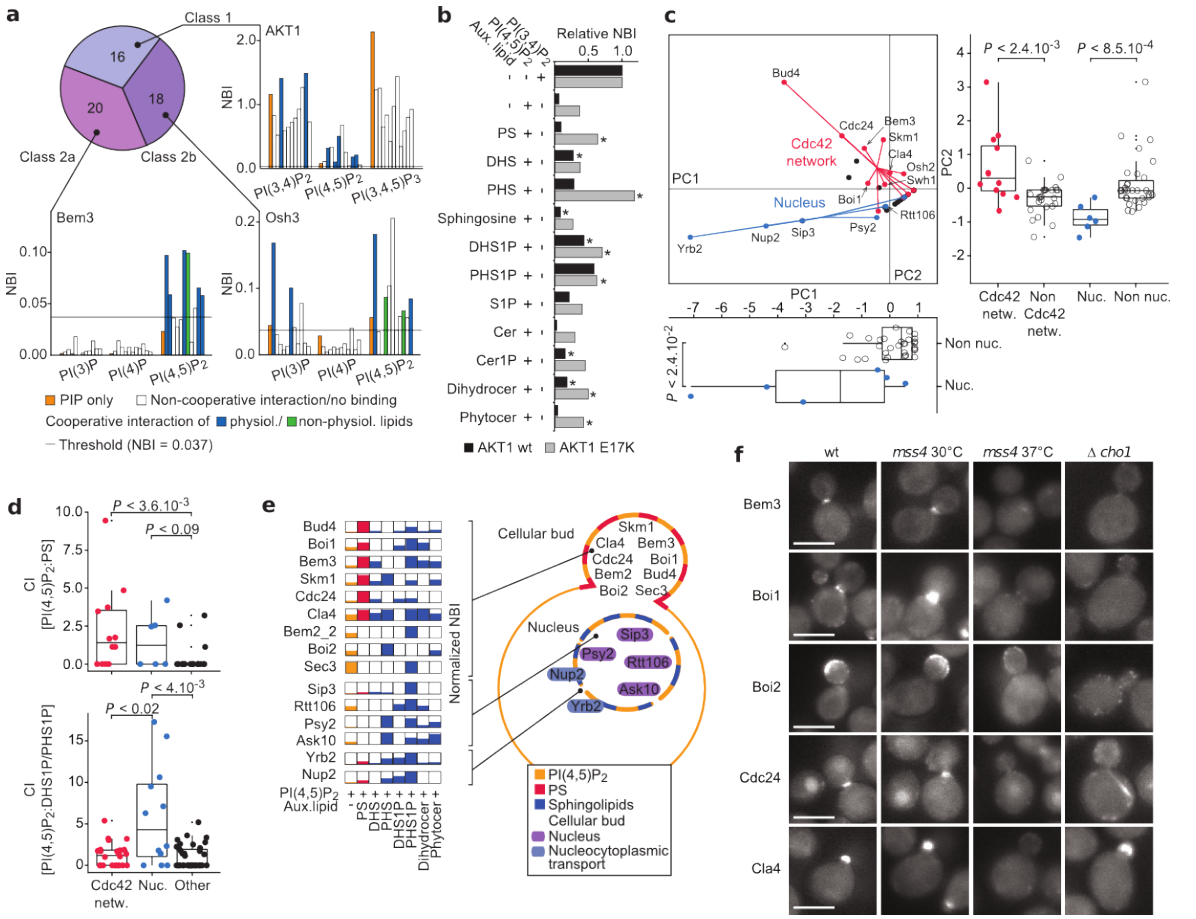


Figure 29: Biological implications of cooperativity binding mechanisms (a) Classification of PH domains in different cooperativity classes according to their cooperativity profile (as illustrated by the surrounding barplots). The number in the pie charts represent how many PH domains belong to the class The surrounding bar plots show the NBIs measured for liposomes containing PIPs alone or mixtures of PIPs with auxiliary lipids. The order of auxiliary lipids in the mixtures is (left to right): PS, DHS, PHS, Sphingosine, DHS1P, PHS1P, S1P, Cer, Cer1P, Dihydrocer, Phytocer . (b) Influence of lipid cooperativity on membrane recruitment of AKT1-PH. Comparison of AKT1-H wild type (wt) and E17K NBIs to membranes of various lipid ompositions. The NBI values for each AKT1 variant are normalized to the value of DOPI34P2 only (relative NBI = 1). Stars indicate cooperative interactions. (c) Principal component analysis of the *S.cerevisiae* PH domains with at least one CI > 1 (n =37). Only CI values for liposomes containing DOPI45P2 with PS/DHS1P/PHS1P were considered. The box plots represent the difference between nucleus and non-nucleus groups (bottom, PC1; right, PC2), and Cdc42-interacting and non-interacting groups (right, PC2). (d) Box plots of the CI values of DOPI45P2:PS (top) or DOPI45P2:DHS1P/PHS1P (bottom) liposomes calculated for the groups of PH domains defined in (c) (e) Proteins targeted by the same cooperating lipid pairs are functionally related. Histograms show NBIs for DOPI45P2 alone or in the presence of cooperating auxiliary lipids (CI > 1). Bars are normalized to the highest value for each individual PH domain. (f) Impact of phosphatidylserine and DOPI45P2 metabolism on the localization of selected GFP fusions in *S.cerevisiae*. δ cho1, phospahtidylserine synthase deletion and mss4ts, thermosentsitive mutant of the phosphatidylinositol-4-phosphate 5-kinase. All scale bars, 3 μm.

Biological impact of cooperativity

So far, we have demonstrated that the majority of PH domains is capable of cooperatively being recruited to lipid membranes, and that the presence of secondary lipids has indeed an implication on the increased or switched PIPs specificity of PH domains. The biological implications of these cooperative mechanisms in terms of cellular location, biological process or molecular mechanism remains yet to be answered. Biological membranes are rich in lipids, and we have seen PH domains recruitment to membranes is finely tunes and determined by the presence of particular lipids. Specific combination of lipids within membrane might as such be good determinant in the exact location of a PH domain in the cell. There exist many biological processes (bud assembly, cell polarization) in yeast which imply the reshuffling, modification and enrichment of particular lipid species in particular cellular component in order to recruit the required protein for that particular process to happen. Corrolarily, some lipids are more prone to be relocated than others, which are more abundant and evenly distributed. DOPI45P2 is for instance continuously and equally dispersed throughout the plasma membrane during the process of cell polarization, whereas PS will be found enriched in areas of polarized growth [116]. The high concentration of PS at polarized growth site is mandatory for the polarization to happen, as it specifically contributes to the binding of Cdc42, which act as chief controller of cell polarity [117]. Cdc42 itself is not known to contain any known lipid binding domain, yet plenty of its interacting partner do. The PH domain analysed in this screen include nine proteins that are known to directly interact with Cdc42, which are all found to locate in the bud compartment (in vivo). For six of these Cdc42 interacting partners, the affinity for DOPI45P2 based liposome was increased when those liposomes also included PS. Remarkably, other PH domains tested known to locate to the bud without being known to directly interact with Cdc42 did not display an increased specificity towards DOPI45P2 based liposomes when put together with PS. Those observations suggested that the specific cooperative binding to DOPI45P2;PS based liposomes is not inherent to bud locating proteins, but appear to be inherent to proteins sharing a particular function. In order to more systematically assess the relevance of the predicted cooperating lipid partners, i.e. if they contribute to a physiological membrane code, we compared the in vitro cooperative profile to physiologically derived in vivo data. We first annotation data on protein–protein interactions provided by STRING [118] and localization SGD [119]. We observed that the proteins displaying similar cooperativity profile with regards to certain lipids are functionally related, or co-localize with their recruiting lipids (Figure 29). For instance, DOPI45P2 and PS – two lipids that accumulate at the sites of polarized growth [116] – most frequently cooperate to recruit PH domains that are known to be part of the Cdc42 interaction network (Figure 29). We additionally made use of live-cell imaging data in order to assess the effect of perturbation of DOPI45P2 or PS metabolism on the cellular localization of Bem3, Boi1, Boi2, Cdc24 and Cla4 fused to GFP (Figure 29). As expected, the results suggested that both lipids are required for the association of Cdc24, Boi2, Cla4, Bem3, and Boi1 with the sites of bud growth and confirms that both lipids might as well cooperate in vivo. Similarly, PH domains present in nuclear proteins also displayed more similar cooperative profile towards combinations of phosphoinositides and phosphorylated LCBs (e.g. DHS1P and/or PHS1P) (Figure 29). These results confirm recent evidences proposing that a nuclear pool of these lipids plays a role in various nuclear functions [120] [121] [122] [123]. Our results confirms that the ability of PH domains to cooperatively bind several lipids and extends it to the majority of PH domains and as being an important of this family. Cooperativity mechanisms can increase or alter the affinity and specificity of PH domains toward PIPs and is linked to the biological function of PH domains.

2.5 Material and Methods

2.5.1 Screen settings

Protein selection and expression

The protein used in the screen were selected using Hidden Markov Models (HMMs) from the Pfam database [124]. Only the species *C.thermophilum* and *S.cerevisiae* were considered. The position of the PH domains within the protein sequences was performed with secondary and 3D structure predictions [125]. Finally, the amino-acid sequences of all PH domains were run through SMART [126] in order to give them an e-value indicating the quality of the PH domain [Supplementary Table V.4]. The lipid binding domains were cloned into a vector (pETM11) and subsequently expressed as N-terminal His6-SUMO3 and C-terminal sfGFP fusions in *E.coli* (BL21 STAR, Invitrogen). Cells were grown at 37°C and proteins were produced over night for 14.5 h before being lysed by sonication.

Liposome preparation

The mixtures used to assemble the liposomes were all composed of one or several lipid of interest, (concentration 10 mol%), a carrier carrier lipid (PC, concentration 95 mol%), a lipid facilitating the autofocusing of the camera during the image acquisition step (PE-Atto 647, concentration 0.1 mol%) and a lipid helping the formation of the liposome (PE-PEG350, concentration 0.5 mol%). The lipid of interest selected for this screen belonged to four different lipid families: DAG, sphingolipids, glycerophospholipids and phosphoinositides. In total, 122 different liposomes containing different lipids of interest were utilized [Supplementary Table V.3]. For technical reasons, two forms of DOPI35P2 differing in their fatty acid tail were used (dioleylphosphate (DOPI3P2) and dipalmitoylphosphate (DPPI35P2). The palmitoyl form was not bound by the specific DOPI35P2 sensor (hsv2). Thus, we solely took into account the results obtained with DOPI35P2 for the rest of the analysis.

Microarray preparation

Lipid mixtures were automatically spotted on the thin agarose layer (TAL). The array was design to hold 120 spots. The spot were all equally sized (800 μm x 800 μm) and equally distant from one another (200 μm). The liposomes (> 5 μm) were fully assembled within 5 minutes following the automatic spotting of the lipid mixtures. Protein extracts were incubated and after 20 minutes washed in order to get rid of unbound proteins. Each protein was tested in three replicate. To this end, three microarrays were prepared and the position of the liposomes was shuffled in order to control for positional and time bias.

Image acquisition

In order to acquire the position of the liposomes as well as the LBDs, the microarray spots were automatically imaged on two different channels. The fluorescent channels Atto547 was meant to acquire the position of the liposomes and was exposed at a single exposure time (3ms). The fluorescent channel GFP, was meant to acquire the position of the LBDs and was imaged at six different exposure times

Image processing

The image processing pipeline relied on the image analysis software Cellprofiler [127]. For each image, the background was removed and the remaining Atto 647 signal was used to delimitate liposome membranes as well as the center of each liposome. The overexposed pixels within the pair GFP/Atto 647 were filtered

out from both images were located in all images and then each image was assigned a mean GFP and a mean Atto647 intensity.

Readout extraction

For each image corresponding to each exposure time, we computed the normalized binding intensity, that is, the ratio GFP/Atto647 divided by the exposure time. This led to constant NBIs over time and constituted our final readout. The computation of the NBIs for each image was performed on a cluster (LSF).

2.5.2 Technical analysis pipeline

Reproducibility

In order to assess the quantitative reproducibility, we plotted the replicates of each interaction against each other (replicate 1 vs replicate 2, replicate 1 vs replicate 3, replicate 2 vs replicate 3). Given that we do not have rank data and are not trying to look for a monotonic relationship but rather a linear relationship between related variables, we have log transformed the NBI and measured the pearson correlation coefficient which yielded 0.67 with a P value $< 2{,}2.10^{-16}$. Two experimentalist as being a binding event, a non-binding event, or a dubious event annotated each image. In total, 229,880 images were flagged as non-binding events, binding eventsor as dubious event. The replicates of each interaction were not necessarily annotated by different experimentalists, which might bias the objectiveness of judgment. To avoid this annotation bias and thus an artificially high qualitative reproducibility, the replicates of an interaction were not annotated at the same time so that the experimentalist does not know what annotation was assigned to the previous image corresponding to the previous replicate. Lastly, the images did not contain any information related to the NBI, so that the experimentalists' judgment can be as independent as possible from parameters proper to the image.

Threshold extraction

Given that each image benefited from a manual annotation (qualitative readout or "golden standard") and a quantitative readout (NBI), we used these two criteria in order to extract an NBI threshold above which an interaction will be classified as binding event. We first draw a ROC curve as well as a precision-recall curve in order to measure what would be the sensitivity, specificity, precision and recall of the threshold chosen. In order to extract a particular threshold, we have chosen to minimize the classification error rate and thus to maximize the accuracy of the prediction. As such, the extracted NBI threshold value was set to 0.037, yielding a sensitivity of 75.2 %, a specificity of 96.6 %, a recall of 81 % and an accuracy of 86 % (Figure 16). The technical analysis pipeline includes the R package Optimal Cutpoints][100] and thus leave to the data analysis the choice of how to select the threshold (that is, maximizing the specificity, maximizing the sensitivity, maximizing the sum of the sensitivity or the specificity etc.)

Number of required annotations

Let N be the total number of interactions and n an integer. For each n that belong to [10;1000], we defined the training set as being the set of n interactions randomly selected from N. We also defined the test set of size m = N-n. For each training set, a binding threshold X was extract as explained in paragraph 2.3.7 and the m interactions of the test set were classified according to that binding threshold : interactions yielding an NBI > X were classified as binders, and interactions yielding an NBI < X were classified as non binders. We then compared those classifications with the actual manual annotations and

computed the precision of the classifications as being the ratio between correctly classified interactions divided by non-correctly classified interactions.

Protein concentration effects

Each interaction was tested in replicates and each replicate was performed at a different protein concentration. Therefore, we could assess the protein concentration effects in two different ways. First, we analyzed if there was a global correlation between the protein concentration and the NBI measured by plotting the NBI collected for each protein concentration probed (Figure 10 and Figure 11). We then measured the pearson correlation between these two hypothetically related variables. Secondly, we analyzed for each interaction if the NBI measured at two different concentrations were significantly different (Figure 10 and Figure 11). In order to assess the statistical significance of these difference, we performed a Kruskal-Wallist one-way analysis of variance test, given that we do not have normally distributed data.

Spatial bias

In this screen, three microarrays were used in order to study the interactions in triplicates. The position of the liposomes was shuffled each time in order to check for spatial bias. The NBI measured at each position of the microarrays were collected and we computed the proportion of binding events measured at this particular spot according to the binding threshold set earlier in the technical analysis pipeline (NBI = 0.037) and represented the microarray as a heatmap (Figure 14). For each row and column we computed a Kruskal-Wallis one-way analysis of variance test to verify some row (or column) significantly contained more binding events than another row (or column).

Effect of the spotting time on the NBIs measured

The experimental time needed to complete the measurement of an array was 3 hours. To assess potential bias in the results due to the imaging time, we took advantage of the fact that we reshuffled all the lipid positions between the replicates. We plotted the difference of NBI measured between the different pairs of replicates, which were measured at different location on the microarray (and thus at different time). We iterated over all interaction and plotted the NBI differences observed vs the time elapsed between two measurements (Figure 15). We also performed an experiment in which we imaged the same liposome array twice, once at time 0 and then two hours later. We did not observe significant changes between the NBIs measured at time 0 and two hours later.

Sensitivity and specificity assessment

In order to assess the sensitivity and the specificity of the binding events, we made used of the proteins (EEA1-FYVE, P40PHOX,HSV2,LACT-C2) for which we know which liposomes they are supposed to bind. For each of these proteins, we measured the NBI and TI collected for their physiological lipids (lipids they are supposed to bind) and non physiological lipids (lipids they are not supposed to bind). We then computed a wilcoxon test in order to evaluate if the NBI and the TI collected for the physiological lipids were respectively significantly higher than the one collected for the non physiological lipids.

2.5.3 Functional analysis pipeline

Clustering analysis

Hierarchical clustering In order to perform a hierarchical clustering analysis of the 122 liposomes based on their protein binding profile using the NBI readout, we proceeded as follow. We first compared the protein binding profile of the heterogeneous liposomes by setting a a matrix M_{het}(dimension PxL_{het}) (where P equates the number of proteins and L_{het} the number of heterogeneous liposomes) whose elements correspond to the average NBI of the replicates measured for the interaction $<p_i,l_j>$. The missing data that were replaced by the average NBI the column they belong to. The second step consisted in applying a log transformation (base 1) of M_{het}. The third step of the clustering analysis consisted in computing the distance matrix M_{dhet} of M_{het}. The distance measure chosen in order to compute the distances between each pair of heterogeneous liposomes (the columns of M_{het}) was the uncentered pearson metric (Figure 22). The R package "dist" was used to this matter [128]. The last step of the hierarchical clustering was to apply an agglomerative algorithm ('complete' algorithm) on M_{dhet}. A second matrix M_{hom} PXL_{hom} (P equates the number of proteins and L_{hom} the number of homogeneous liposomes) was set and an identical approach employed in order to perform a hierarchical analysis of the homogeneous liposomes based on their protein binding profile (Figure 20).

Principal coordinate analysis To verify wether the two main clusters engendered by the PIPs could also be obtained with another method, we performed a principal coordinate analysis. We applied an implementation of the statistical procedure on the same distance matrix M_{dhet}. We quantified the statistical significance of the clusters by performing a Wilcoxon test on their cumulated MDS1 and MDS2 values (Figure 21). The PCoA analysis confirmed the first branching observed in the hierarchical clustering analysis in M_{dhet} separating two clades; one containing exclusively heterogeneous liposomes containing PM PIPs and another clade containing exclusively heterogeneous liposomes containing ORG PIPs.

Cluster validation In order to validate the observed cluster and determine the optimal number of clusters (in other words, to test if these two clusters are not random), we employed two methods. We first used a portioning around medoids algorithm in order to determine the optimal number of cluster using the pam function from the cluster package [129] (Figure 22). We then validated the robustness of these two clusters by computing the their silhouette index using the silhouette function from the cluster package [129](Figure 22).

Sub cluster validation In order to confirm that the secondary lipids could engender sub-clusters within the cluster triggered by the ORG PIPs containing liposomes, we performed a fisher test on the differential content in negative charges within certain clades of the ORG PIPs generated cluster (Figure 21).

Motif search

Definition of foreground and background In order to look for an amino acid sequence that could explain why some PH domains can bind ORG PIPs containing liposomes and other cannot, we first draw the PIPs binding profile of the 90 PH domains 24. We subsequently defined two groups. PH domains capable of interacting at least with one pure ORG PIPs liposome regardless of its TI (the foreground, 16 PH domains) and those which could not (the background, 74 PH domains).

Multiple-alignment and filtering We then performed a multiple sequence alignment of the amino acid sequences of these 90 labeled PH domains (depending on their belonging to the foreground or background). We used the hidden markov model for PH domains in order to align the PH domain sequences with the software HMMER 3 [130] and filtered out two PH domains belonging to the foreground based on their misalignment and high e-value prediction score (PDK1 and Avo1).

Position scoring Given that we wanted to detect amino-acids that could explain binding specificities of proteins within the same protein family, we employed a method which accurately detects subfamily specific residues from a multiple sequence alignment [101]. The algorithm scores compositional differences between n groups of proteins without imposing a high degree of sequence conservation, which is an advantage in our case given that PH domains have less then 25% of sequence identity). In brief, the algorithm loops over each position of the multiple sequence alignment and computes four statistics per position. The first statistics is the sequence harmony (SH) score, which evaluates how conserved are the residue within each group. It ranges from0 for completely non-overlapping residue compositions to 1 for identical compositions (in other words, the lower the better). The second statistic is a measure of statistical significance of the SH score (the Z_{SH} score). The Z_{SH} score is calculated by permutating the group labels and running the sequence harmony algorithm for 100 randomizations. The mean and standard deviations of these permutated groups are calculated. A Z_{SH} score of 1 means that the SH score is 1 standard deviation below the random mean (the more negative the Z_{SH} score, the better). The third statistics is the multi relief weight (MR score), which evaluates how discriminatory are residues between each group. It ranges from 1 for completely discriminatory residues to 0 for non discriminatory residues (so the higher the better). The fourth statistics is a measure of statistical significance of the MR score (Z_{MR}). The Z_{NR} is calculated by permutating the group labels and running the multi-relief algorithm for 100 randomizations. The mean and standard deviation for these permutated groups are calculated. A Z_{MR} score of 1 means that the MR score is one standard deviation above the random mean (the more positive the Z_{MR} score, the better). At the end, each position of the multiple sequence alignment had these four statistics assigned.

Positions selection The computed statistics of all the positions of the multiple alignment were represented in the 4 dimentional scatter plot 2. We selected the positions that yielded a Z_{SH} score < 3, and a Z_{MR} score > 1. These criteria retained 17 positions, which we further filtered. We filtered out positions which show an overlap in overrepresented amino acids in both groups (foreground and background). We also filtered out positions which were not part of the predicted PH domain secondary structure. At the end, we had nine positions which defined the organellar PIP motif (OPM)

Scoring of the 1216 PH domains We downloaded the amino-acid sequences oft he 1,216 reviewed PH domains from the uniprot database. We scored each PH domain according to the presence or absence of the residues present in the newly defined organellar PIP motif (OPM) and those present in the basic sequence motif (BSM). The scoring scheme was the following :

- OPM residues : Three points for the most conserved residues (the threonine residue located in the β1-β2 loop and the lysine residue located at the end oft he β7 strand). All the other residues received one point.
- BSM residues : one point per residue.

Each PH domain could in theory have a maximal score of 16 (maximal score of 13 if the PH domain sequence contained the entire OPM and 3 if it contained the entire BSM).

cooperativity events prediction

Prediction of cooperativity events For each interaction between a protein P and a heterogeneous liposome L_{ab} (with a and b being two lipids of interest), we define NBI_a the NBI observed between P and L_a, NBI_b the NBI observed between P and L_b, NBI_{abo} the NBI observed between P and Lab, and NBI_{abe} the expected NBI_{ab}, that is, the sum of NBI_a and NBI_b. NBI_a, NBI_b, NBI_{abo} and NBI_{abe} have at least two replicate each. For an interaction to be predicted as positively cooperative, the following conditions must be fulfilled (Figure 30):

- $NBI_{abo} > NBI_{abe}$
- The NBI_{abo} has to be superior to the highest NBI observed for NBI_{abe}

For an interaction to be predicted as negatively cooperative, the following conditions must be fulfilled:

- $NBI_{abo} < \max NBI_a; NBI_b$, provided that $NBI_a > 0.037$ or $NBI_b > 0.037$
- The NBI_{abo} has to be inferior to the lowest NBI observed for $\max NBI_a; NBI_b$

Interactions that did not fulfill these criteria were classified as non-cooperative. The computed ratio NBI_{abo}/NBI_{abe} (or $NBI_{abo}/\max NBI_a; NBI_b$ in the case of a negative cooperativity) defines the cooperativity index (CI). This index is completed with the TI computed for the interaction between P and L_{ab} (or between P and L_a or L_b in the case of a negative cooperativity)

Validation of cooperativity events We randomly selected 36 interactions that were classified as positive, negative and non cooperative in order to assess the sensitivity and specificity of these predictions. To do so, we used LiMA In order to assemble liposomes of various concentrations (0,3,6,10 mol%) containing two lipids of interest (DOPI45P2 or DOPI35P2 and DHS1P or PHS1P or PS). Liposomes containing DOPI45P2 and DHS1P as lipid of interest were assemble at a DOPI45P2 concentration of 6 mol %. The PH domain concentration used in the validation experiments were identical to those measured in the screen.

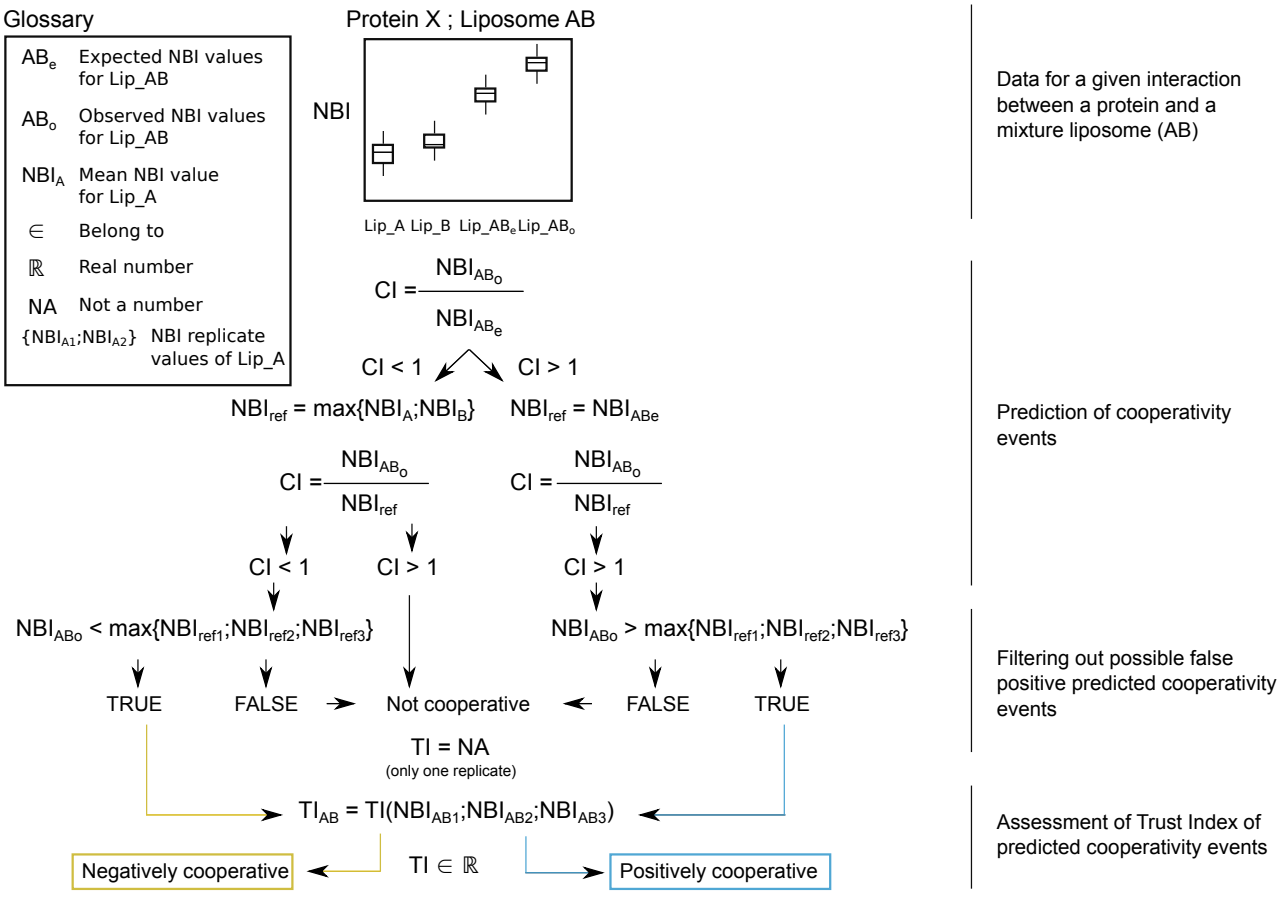


Figure 30: Cooperativity prediction: scheme illustrating the heuristic chosen for the prediction of cooperative binding.

2.6 Discussion

Many cellular processes require the recruitment of protein to specific membranes, which are decorated with distinctive lipids that act as docking sites. In particular, the phosphoinositides (PIPs) form signalling hubs, yet the underlying mechanisms remain elusive. A eukaryotic cell usually process more than 1,000 different lipid species that bear structural, physical and biochemical properties. This field of chemical biology lacks standardized experimental and computational methods that would allow having a comprehensive view of the PH domain binding landscape and to unravel regulatory mechanisms orchestrated by lipids. In this study, I present a new computational analysis pipeline, which in a first time performs a technical assessment of the readouts outputted by the novel liposome microarray-based assay (LiMA). We demonstrate that LiMA produces high quality and quantitative readouts that can be used to capture and compare protein-lipid interactions. In a second time, it performs a functional assessment of the same readouts. We applied this technology to a set of 91 PH domains selected from two species : *C.thermophilum* and *S.cerevisiae*. We designed the LiMA to study a wide range of liposomes containing the main lipid families. We applied the computational analysis pipeline to the readouts collected from this screen and validated the postulate that PIPs are the preferred PH domains ligands and that these binding events are dictated by particular sequence attributes proper to certain PH domains. Moreover, we have generalized the concept of cooperativity as binding principle applied to the majority of PH domains.

2.6.1 LiMA, a novel tool for high throughput screening of protein lipid interactions

To this day, many technologies have been developed in order to study protein-lipid interactions. Some of these methods rely on liposomes and thus imitate very precisely physiological conditions. These approaches usually yield very few false positive or negative interactions but are very difficult to setup and are inconvenient for high-throughput screening . More recent improvements have led to a technology making use of fixed lipids. This method, known as the lipid overlay assay is –as opposed to liposome based assay- easy to setup and can be applied for high throughput screening [27] [60] [62] [131]. However, this method usually yields a high number of false positive interactions due to the very high concentration of lipid of interest, thus creating a non-physiological environment for the studied proteins.

LiMA is a technology marrying both characteristics: high-throughput and physiological conditions. The technology makes use of liposomes in which the lipid(s) of interest studied are within a membrane bilayer. Besides, LiMA makes use of a chip containing 120 spots onto which lipid(s) of interest can be automatically spotted and this at different concentrations, and is thus easy to setup for high-throughput studies. LiMA also sets a quantitative readout (NBI), which is proportional to the dissociation constants (K_d) of a protein to a given ligand. In the future, other readouts could be envisioned and these K_{ds} could even be directly measured by making use of fluorescence imaging technique like two-photon excitation microscopy or by using a different microscope like the total internal reflection fluorescence microscope. In study, the chip settings we chose enabled the study of 120 liposomes. With the proper chip one could assay the protein binding profile of thousands of lipids and screen the entire lipidome and proteome. Future improvements of the technology could also enable the study of curvature sensing by proteins and how it influences their recruitment to biological membranes. Additional adaptations of the protocol could allow the insertion of protein into the liposomes and thereby represent an excellent method to study the recruitment of protein requiring both protein and lipid partners at large scale.

2.6.2 Systematic analysis of PH domains confirms their ability to interact with membranes

In this study, we have made used of the technology LiMA in order the protein-lipid interactions between 91 PH domains and four other LBDs belonging to four species and 122 liposomes covering the most frequent lipid families. We measured more than 11.000 experiments in replicates, which consist in one of the biggest protein-lipid screen. Making use of manual annotation, we have extracted a threshold allowing to distinguish binding events from non binding events. For this study, we used all annotations in order to extract the binding threshold. This has the drawback of having to annotate every single experiment, which cannot be conducted for screens containing hundreds of thousand of interactions. However, the conducted analysis shows that 500 annotated interactions are enough in order to extract a binding threshold yielding an excellent sensitivity and specificity. In this dataset, we have detected 2415 interactions, including 2269 physiological and 1628 of high confidence, capturing interactions with a known Kd ranging from the nanomolar to the micromolar. We chose to focus on PH domains because they are considered as good examples of LBDs. Besides, they have already been used in various studies and this constitutes a good comparison basis for our results. The majority of the previous studies focused on the so called "classical" PH domains, as opposed to PH-like domains. The group to which a PH domain belong depends on the ability of the searching algorithms to recognize them. In this study, we focuse on both types of PH domain and have found that 69 PH domains can bind at least one liposomes. In most of the case, the binding event required the presence of several lipids of interest. Within the PH-like PH domain category, many of them did unexpectedly manage to bind several liposomes (Figure 11 and Figure 27). The binding events found in this study contain a big part of novelty. Indeed, we screened for the first time 20 PH domains belonging to *S.cerevisiae* and 27 belonging to *C.thermophilum*. Another 5 were not found to interact with any lipid in previous studies. 75% of these 52 PH domains managed to bind at least one liposome with a high confidence index, implying that a big fraction of the interaction observed in this screen were novel.

2.6.3 Only fraction of PH domains interacts with PIPs localized in ORG membranes

Our dataset supports the notion that the specificities of the PH domains encompass all seven phosphoinositide species [131] (Figure 24). When clustering liposomes according to their PH domain binding profiles, we observed that those containing phosphoinosidites known to predominantly localize in the PM (DOPI45P2, DOPI34P2, DOPI345P3) and those containing phosphoinositide present in the orga nelles (DOPI3P, DOPI4P, DOPI5P, DOPI35P2) [132] form two clusters that determine the propensity of the two types of membranes to recruit PH domains (Figure 21). The first branching observed in the hierarchical clustering engenders two main clusters. The first one exclusively contains liposomes mixtures containing PIP species found on the plasma membrane. The second one contains PIP which are found on various organellar membranes (but not in the plasma membrane). Those two clusters being driven by the PIPs lipid suggest first that the PIPs lipid is the preferred ligand, and second that lipids located at the plasma membranes and within organellar membranes recruit different sets of proteins. Additionally, we the PM PIPs clade seems to display sub-clusters that are enriched in specific PM PIPs family members (DOPI45P2, DOPI34P2, DOPI345P3). This corroborates the fact that different PM PIPs recruit different PH domains [58] [64] [133]. This observation does not hold for the ORG PIPs cluster. Sub-clusters seemed to be more influenced by the presence of secondary lipids enriched in negative charges. Consistently with earlier findings, many more PH domains were able to bind PM PIPs than ORG PIPs [109] [134] [135] [136] [137]. Noticeably, we have not found a PH domain capable of binding ORG PIPs without

being able of binding PM PIPs, whereas some PH domains were able of exclusively bind members of the PM PIPs family (Figure 24). All in all, most of PH domains can bind PM PIPs and only a subset can recognize ORG PIPs.

2.6.4 PH domains interacting with ORG PIPs share common sequence motif

The PH domains capable of binding ORG PIPs were not probed at a higher concentration than those only capable of binding PM PIPs. Besides, none of the 16 ORG PIPs binders preferentially bound a member of the ORG PIPs. This means that these proteins probably have some attributes conferring them the ability of on top of binding PM PIPs, to also bind ORG PIPs. These attributes (e.g, amino acids) should expectedly be absent from the group of PH domains that can only bind PM PIPs. Some studies have in the past attempted to identify particular amino acids within PH domains responsible for some functions (e.g, binding of particular lipids). These studies suffered from many flaws, which makes the veracity of the proposed amino acids questionable. In some cases, the dataset used considered to few ligands [78] [112], in others, the studies focused datasets including too few PH domains[71]. The small numbers of proteins or lipids used in these studies did not confer them enough statistical power to generalize the function associated with these amino acids to the entire PH domain family. Besides, PH domains are known to have less than 20% of sequence identity, which makes the search of a consensus motif particularly hard without a comprehensive view of the PH domain binding landscape. In this study, we have found nine positions that contains amino-acids enriched in the group of PH domains capable of binding ORG PIPs. Amongst these nine positions, eight are completely new and have never been mentioned in previous studies. Some of these positions are close to the known binding pockets (β1-β2 loop) and others far away (β7 strand and α-helix) and all of them were experimentally confirmed, in vitro and in vivo. Interestingly, two of these positions are found to be mutated in human cancer samples [110] [111]. The human PH domain OSBP2 has a natural variant found at the beginning of the C-terminal α-helix of its PH domain (R262L). This particular position is part of the nine positions we have found and is distant from any reported PH domain's binding pocket. Interestingly, the in vitro and in vivo validation experiments showed that mutations of this position specifically affected the binding to ORG PIPs. The second natural variant found is often mutated within the β1-β2 loop of the FAPP1 PH domain (T9A). This region is known to be involved in the bindinf of DOPI4P [138]. The validation experiments confirmed what was already known. The mutation specifically affected the binding to the DOPI4P liposome in vitro and the localization to the trans Golgi in vivo (Figure 27). Few studies have already mentioned the role that could play amino acids located in the in the β7 strand or in the α helix [136]. For the first time, we confirm this hypothesis but the exact mechanistic implications of these residues remain unknown. One hypothesis is that proteins containing PH domains can form homodimers. It has been shown that some PH domains locate much more at some membranes in dermic forms. Thus, it is assumable that perturbing the homodimerization impacts the recruitment to organellar membranes. An enrichment analysis of proteins showing a high conservation of the OPRN residues indicate that these proteins tend to be involved in oligomerization process much more than other PH domain (deprived from OPRM residues) (Figure 24). If we assume that the PIPs bound by a PH domain is a good marker of its in vivo localization, than the attributes conferring the ability to PH domains to bind these different lipids are primordial for the fulfillment of their cellular functions. When we performed a multiple sequence alignment of the entire PH domain universe and looked for the presence of the 12 residues (BSM and OPRM residues), we found that the OPRM reisdues are found in a much smaller set of PH domains than the BSM. Interestingly, not a single PH domain contained the OPRM residues without containing the BSM residues, whereas the many PH domains solely contained BSM residues. An enrichment analysis revealed that OPRM reidues containing

PH domains were significantly more involved in cellular transport activities compared to the other PH domains (Figure 24). The PH domains enriched in cellular transport activsities typically carry ligands (lipids) from one donor compartment to another acceptor. The fulfillment of this activity require to be able to bind different membranes containing different lipids, even though the exact binding mechanism are still unknown. Some studies pointed out that particular residues might be responsible for the specific recruitment of PH domains to biological membranes, and our data indicate that a single motif might be able to sense both PM and organellar membranes [139]. A question that still remains to be answered is why some PH domains can bind PM membranes without being able to bind organellar membranes, but the other way around does not hold.

2.6.5 Interactions of PH domains with membranes are modulated by cooperation between PIPs and other lipid species

Several studies which focused on a small set of PH domains have already postulated non PIPs lipids can be bound by PH domains. The presence of these "secondary lipids" in the neighborhood of PIPs is assumed in some cases to stabilize the interaction between the PH domains and the PIP. This concept of synergetic binding between a PH domain and several lipids is known as cooperativity [69] [140] [141]. Our data offer for the first time the chance of assessing the occurrence of cooperative events in the PH domain binding landscape. We have demonstrated that cooperative binding is not anecdotal, but rather a general binding principle of PH domains. As several studies in the past have shown, we have found that negatively charged secondary lipids (predominantly LCBP1) account for most of the cooperative events. More than a finely tuned binding, 19 PH domains required the presence of two lipids in a liposomes in order to be able bind them, implying that many PH domains require these cooperative mechanisms in order to bind PIPs containing membranes. Noticeably, the presence of these secondary lipids can sometime confer a particular specificity to a PIPs which the PH domain did not have otherwise. Given that most of previous study were performed with liposomes containing a single lipid of interests, our results answer the long lasting mystery why most PH domains lacked specificity.
The concentration of the lipids of interest in the liposomes most probably plays a role in the amount of false negative cooperative events yielded in our predictions. Most lipids of interest were at high concentration (10 mol %), which is a saturating concentration for many protein-lipid pairs of our screen. Many interactions we classified as non cooperative were contradicted by the validation experiments we performed. In those cases, cooperativity was observed at lower concentration, which we were not able to capture [Supplementary Table 2].

2.6.6 Recognition of the same cooperating lipids is shared by functionnally linked PH domain-containing proteins

Biological membranes are complex and contain a myriad of lipids. Their content is dynamic and varies from one organelle to the next. Although the exact content of each biological membrane has not been entirely depicted yet, one can assume that they fulfill both a structural and a functional role by recruiting particular proteins. In this study, we have used cooperativity profiles as mean of understanding biological function and cellular localization of PH domains. Many proteins studied in this screen are found to locate in vivo the bud and bud neck [142] or to interact with the protein Cdc42, an important actor the the bud formation [117]. Noticeably, most of the PH domains tested interacting with the protein Cdc42 display a very similar cooperativity profile compared with other proteins. They show indeed a stronger binding to DOPI45P2-PS containing liposomes which is in accordance with PS in vivo location, that

is, in areas of polarized growth [116]. Suprisingly, PH domains which were not found to interact with Cdc42 but nonetheless locate to the bud did not exhibit any cooperative event with DOPI45P2-PS. This observation indicates that a strong cooperativity with the pair DOPI45P2-PS seems to be a marker of biological function, that is, Cdc42-regulated cellular functions. In agreement with this hypothesis, we observed that many PH domains from protein locating elsewhere in the cell but still linked to the same biological process also were able to cooperatively bind the pair DOPI45P2-PS. For instance, the PH domain belonging to Slm2 –which was demonstrated to be an actor of the polarized actin assembly- did not bind the liposome containing DOPI45P2 as single lipid of interest. However, the presence of PS induced a binding event [143]. The same observation did not hold for the PH domains belonging to the Osh proteins (Swh1, Swh1-ct, osh2, osh3, osh3-ct). In those cases, the presence of PS only slightly increased the affinity to the liposomes. Additional experiments confirmed that this might be due to an already saturating concentration of DOPI45P2.

The protein Cdc24 contains itself a PH domain and act as chief regulator of Cdc24 activity. Cdc24 locates at two different places: in the nucleus, and in areas of polarized growth. The nuclear membrane contains many lipid species, including PIPs and sphingolipids [122] [144] [145]. Noticeably, most of the PH domains from our screen which are known to locate in the nucleus displayed an exacerbated cooperativity with the liposome containing DOPI45P2-LCB1Ps as lipids of interest.This suggests that the cooperative binding of PH domains to lipids may serve various functions. Lastly, even though *C.thermophilum* and *S.cerevisiae* displayed on average a similar binding landscape (Figure 8), they strongly differed in terms of cooperative landscape. Given that these two organisms are very distinct, we can imagine that these organisms have evolved to contain different lipid mixtures in their biological membranes in order to face different thermic conditions. Alternatively, orthologous proteins probably evolved towards different sets of functions, which could explain the weak sequence identity that orthologous proteins share. These assumptions would require additional analysis regarding the exact lipid compositions of each organism and the implication it has in terms of biological function served, which is not the scope of this study.

2.6.7 A new path for drug discovery

Many lipid binding domains are enzymes, and their specific recruitment to lipids have biological implications, which are often linked to diseases when those binding events are abrogated [73]. Our data have been able to reveal the consequences of many disease-associated mutations on the binding specificities of PH domains. This suggests that protein-lipid interactions are of a great interest for the comprehension of mechanisms underlying certain diseases and that the developed computational methods analysing readouts from such high-throughput chemical biology assays can be employed to aid this understanding. Regardless of the nature of the interaction, lipid-binding domains constitute a relatively new niche to be exploited in drug development. Some lipid binding domains have already been shown to contain a druggable pocket that can be used to perturb protein-lipid interactions [146] and [147]. The discovery of the lipid-binding domain C1 as a druggable target lead to the development of small molecules that are now close to clinical trial [148]. Deregulations in the phosphoinositide-3-kinase pathway have often been encountered in many diseases and represent an attractive set of targets for drug development. Typically, hyperactivations of the PI3K pathway have been linked to cancer, whereas hypoactivation plays an important role in the development of type II diabetes [148]. Many compounds targeting the PI3K pathways have been developed to inhibit downstream targets (typically, Akt1). A recent study has identified two molecules altering interactions between DOPI345P3 and its pleckstrin homology domain containing partners. To do so, the authors screened 50,000 molecules in order to indentify these two molecules, and have tested the anticancer activities of one of these compound (PIT1) in vitro and in vivo. PIT1 revealed to be a good candidate in the treatment of glioblastomas (which show elevated levels of DOPI345P3).

2.6.8 Conclusion

The group of Dr Gavin has created LiMA, the first technology enabling the study of protein lipid interactions under physiological conditions and in a high throughput manner. The computational methods developed here have proved that this technology could produce readouts of high quality for large scale studies. It has also shown that the recruitment of proteins (PH domains) is largely influenced by dominants lipids (PIPs) and that biding affinities and specificities are also frequently determined by the presence of additional lipids. The analysis also revealed cooperativity as a general key mechanism for membrane recruitment and showed that proteins displaying a similar cooperativity profiles do have common biological functions or cellular localizations. The developed methods have also led to the identification of a motif shared by proteins displaying a similar binding profile containing amino-acids found mutated in some human cancer, and thereby enable the interpretation of disease-associated mutations. This opens a new door to identify new classes of drugs specifically perturbing some protein-lipid interactions given that they constitute attractive targets for pharmaceutical drug development. This work provides some milestones in the field of protein-lipid interactions, yet the plethora of lipids constituting cells leaves us far away from understanding them all. Parallely, data on the bioactivities of small drug-like molecules is accumulating at a tremendous speed, and high-throughput assay can now span molecular, phenotypic and organismal readouts. Chemical's therapeutics and off targets are coming from in vitro screens are complemented by high-throughput cell based assays and organismal responses to chemical perturbations. The plurality of these assays paired with a generalized effort to collect the produced data in public databases, have created unprecedented opportunities to develop computational tools aiming at integrating these heterogeneous data in order to make new inferences on the complex effects chemicals can have on biological systems.

Part III

Development of an efficient and scalable chemical analysis platform

Part III
Summary

In this part, I present CART, a Chemical Analysis and Retrieval Toolkit. CART represents a major contribution to the field of large-scale chemical biology in the sense that it is the first scalable computational platform, which annotates chemicals and computes enrichments on these annotations by integrating many chemical centric databases. I am first author the CART publication (manuscript currently in preparation), and have contributed to all aspects of the project. I am the main contributor of source code, have co-designed the pipeline, performed the benchmarks and analysis mentioned in this chapter, and I am co-writing the manuscript.

CART: an efficient and scalable Chemical Annotation Retrieval Toolkit
Deghou, S., Zeller, G.,Iskar, M., van Noort V., Bork P ***Bioinformatics***, Manuscript

Chapter 3

Introduction

3.1 Publicly available resources

The number of small molecules to which humans are exposed has tremendously increased in the last century [149]. These small molecules are found in our food, constitute the drugs we take and can be found in very high concentration in the environment when produced in massive amounts by industries. The various effects of these small molecules on our physiology are far from being unravelled, and while their mechanisms of action are being studied, small molecules can also be used to understand how biological systems function. In chemical biology, the perturbation of a biological system upon chemical treatment can lead to a better understanding of biology. There is a growing awareness of this reality within the scientific and politic community, and this is translated by an increasing number of large-scale studies with the aim to generate comparable and standardized assays as current knowledge is scattered and heterogeneous [150] [151]. These large-scale studies are often made available to the public through open access databases. The number of these databases as well as their content is constantly growing. The program REACH (Registration Evaluation Authorisation and Restriction of Chemicals) constitutes for instance a good illustration. REACH is a European Union regulation of 2006 [152]. REACH demands that chemical companies provide information regarding the compounds they manufacture, import and use. The program has been initially timed to last 12 years. By 2018, around 140 000 chemical substances will be registered. The PubChem database is another prime example of growing chemical centric database. PubChem bioassay contained 800 entries in 2008, almost 500,000 in 2011 and almost 1,200,000 in 2015. PubChem substance has grown from 17 million to more than 100 million entries [153]. The Supplementary Table V.1 lists some of the most cited chemical resources that link chemicals to biological activities and span molecular and phenotypical readouts. The raise of high-throughput screenings has breached into chmical biology, which can now profile tens of thousands of chemicals and provide them with rich and heterogeneous readouts regarding their engendered gene expression profiles, metabolic changes, toxicity profiles, modified cellular phenotypes or target profiles [154] [155] [156]. As a consequence of this, a myriad of open access databases dedicated to chemicals and their associated biological entities emerged in the last 10 years, and the number of these databases as well as their content flourishes dramatically.

3.2 Toward an integration of the publicly available resources

These big efforts have created the opportunity to integrate these various heterogeneous data under the same umbrella and therefore to open the door to new questions and inferences. Characterizing relationships between chemicals and biological entities as well as similarities amongst chemicals helps in drug repurposing, new drug development and elucidation of gene functions by linking drug induced gene expression profiles to phenotypes [30]. Many studies have for instance shown that side effects were primarily caused by the interaction between a drug and a unique protein [36], and that the target profile of drugs was often predictive of their side effects [34]. Alternatively, if each compound of a chemical family or library is annotated on one side in terms of side effects it induces, and on the other side in terms of toxicity it induces, then one can attempt to predict whether which toxicity readouts are most predictive of some side effects. Answering these questions requires an integration of the existing resources, and an analogy with the genomic field helps to understand why this integration is very challenging in the field of chemical biology.

3.3 The example of the genomic field

Within the genomics field it is already possible to gain a better understanding of links between genes and phenotypes. Efforts toward creating a controlled vocabulary to describe the function and localization of gene products as well as the use of common unique identifier to refer to genes helped the process of data integration [157]. It facilitated the implementation of gene enrichment tools [158], which enabled a better understanding of the relationships between the genome of an organism and its phenotype. Although there is a growing demand for such tools in chemical biology, to our knowledge, there exist no software that would enable finding which biological propertiesare shared and/or enriched within a set of chemicals, or simply to have overview of all biological activities that are known of a particular chemical. Many reasons might explain this absence.

3.4 The challenges

First, some chemical entities can have hundreds of synonyms and the absence of officially recognized unique identifiers makes the search of a compound across databases very difficult. The International Union of Pure and Applied Chemistry (IUPAC) [1] has set rules to generate systematic names for chemical compounds. This nomenclature aims at leaving no ambiguity whatsoever when mentioning a given chemical name (either written or spoken). Each chemical name should refer to a single chemical structure. In reality, the interpretation of chemical names is obscured by a tremendous amount synonyms and inaccurate names in usage. Different salts, mixtures and stereoisomers of the same chemical compound often account for the increasing number of names under which chemical compounds (like for instance drugs) can be refered to. Two chemical centric databases could for instance link certain biological activities to two different chemical names (like phthalonitrile and o-dicyanobenzene) refering in fact to the same chemical compound.

[1] http://www.iupac.org/

3.4.1 Chemical name search

Chemical names are often long and complicated, which can make the search for a given chemical whose spelling is not clear very difficult. Searching for a chemical name in a database often needs to take into account possible mistakes (for instance, a user might look for "acetamnophen" when he actually is looking for "acetaminophen"). Not many software programs enables chemical name matching taking into account spelling mistakes (so called "fuzzy search"), and the one providing this option (PubChem, ChemDB [159]) usually limit the search input to one chemical name. Other available online tools like STITCH [160] enable the search of multiple chemical names but prevent fuzzy searches. Additionally, STITCH limits the search input to 200 chemical names. The actual status is that there is no available free open source software that would allow a user to input hundreds, thousands or tens of thousands of chemical names with or without spelling mistakes to get back any information from the database against which the chemical names are matched. This complex problem lies in the fact that many of the existing chemical centric databases rely on relational databases (mostly MySQL or PostgreSQL). Relational databases can store and index vast amount of data and offer some semblance of text-based search functionalities, but they are mostly tuned toward relational operations between tables including joins, and more generally at storing and manipulating structured data. Database Models (DBMs) will perform well when it comes to matching exact chemical names, but are not conceived for fuzzy matching. A fuzzy search for a given chemical name in a big table can often take up to two minutes with a default indexing scheme. Besides, fuzzy searches within a relational database are often tuned for English language and not for chemical names.

Second, open access databases focusing on particular biological activities of chemical entities use their own identifier and rarely cross-reference themselves. Third, chemical entries are not referred to a unique chemical identifier that would help mapping a chemical name to several databases. Fourth, the absence of controlled vocabulary in some databases makes it difficult to identify identical readouts shared by different chemicals.

In the next chapter, I present a computational platform addressing the above-mentioned challenges. It enables the integration of data on chemical bioactivities that are scattered across many different databases.

Chapter 4

Results

We have developed CART, a computational platform that retrieves biological annotations of chemical sets and computes which are enriched. In a nutshell, CART is a software available as a command line tool and via a web interface (http://cart.embl.de) whose input consists in a list of chemical names and the output is a table of enriched biological terms that are associated to these chemical names. CART consists of three modules: Name Matching, Enrichment, Visualization. The Figure 1 represents the CART's workflow.

4.1 Features and functionality

Matching chemical names

CART expects a tab-delimited file as an input for the name matching module. This is the name matching input file. The first column of the name matching input file has to contain chemical names. Optionally, the user can add a second column containing scores proper to each chemical. These are used for rank based statistics. Within that universe, each chemical name is associated to a single Chemical ID (CID). The Name matching module will match each chemical present in the name matching input file against the indexed chemical universe according to different matching algorithms. The CID corresponding to the best match is return. The Name matching modules iterates through the entire list of chemicals and produces the name matching output. This output is then fed to the enrichment module.

Retrieving biological annotations

The enrichment module matches the fetched CIDs to those present in the database selected by the user and get the annotation terms associated to those CIDs provided by the database. Each CID present in the database will thus be assigned to one or several annotation term. Those results are concatenated in a first table, which is the first output of the enrichment module: the annotation table. The enrichment module will compute which of these terms are enriched within the list of chemicals provided by the user (the foreground) by comparing their occurrence to those found either in the database selected by the user, or in a second list of chemicals optionally provided by the user (the background, which also has to be run through the name matching module and then fed to the enrichment module as background). Enriched terms will be printed to a second output file: the enrichment table.

Visualizing results

The two enrichment output files will then be fed to the visualization module, which will generates two html files. One html file containing the enrichment table with links to external resources. One html file rendering an interactive network of the foreground chemicals associated to their enriched terms.

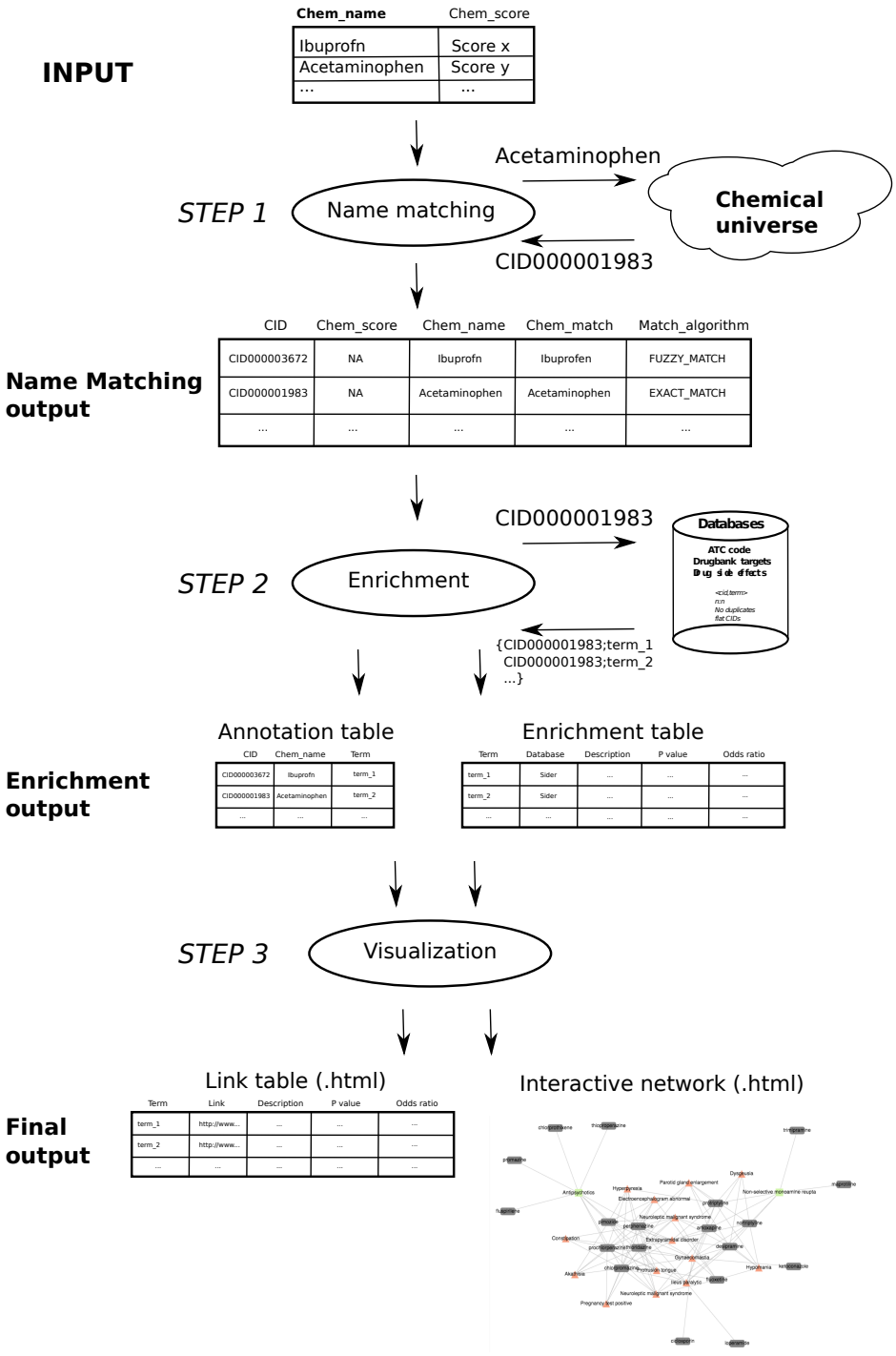


Figure 1: CART's workflow

4.2 Data integration

Chemical annotation data from nine databases were collected and their chemical names were matched to flat Stitch IDs (CIDs) using Name Matching module described in the next section

4.2.1 Molecular targets

Stitch[160]

URL: http://stitch.embl.de
Stitch is the largest resource focused on chemical-protein interactions.It contains interactions for between 300,000 small molecules and 2.6 million proteins from spanning 1133 organisms.

TTD[161]

URL: http://bidd.nus.edu.sg/group/cjttd/
The Therapeutic Target Database (TTD) is a resource containing information relating to therapeutic protein and nucleic acid targets. It also provides information on targeted disease. The database contains information on 2025 targets and 17,816 drugs.

DrugBank[162]

URL: http://www.drugbank.ca/
DrugBank is a database providing chemical, pharmacological and pharmaceutical data. It contains 7759 drug entries and 4282 protein targets (annotated in three main categories : therapeutic targets , metabolic enzyme, transporter).

4.2.2 Therapeutic classifications

Chembl ftc[163]

URL: https://www.ebi.ac.uk/chembl/ftc/
The Functional Therapeutic Classification system is a controlled vocabulary defining and structuring more than 20,000 drugs mechanisms and modes of action.

ChEBI[164]

URL: http://www.ebi.ac.uk/chebi/
The CHemical Entities of Biological Interest (ChEBI) is a database containing a structured vocabulary describing 43216 small molecules in terms of chemical structure and biological function.

4.2.3 Phenotypic readouts

Sider[165]

URL: http://sideeffects.embl.de/
The Side effect resource database (SIDER) contains information on drugs and their side effects. The database contains 996 drugs, 4192 side effects and 99,423 drug-side effects pairs.

DrugMatrix[166]

URL: https://ntp.niehs.nih.gov/drugmatrix/
DrugMatrix contains information on drugs and their induced toxicity in rat. It contains information on 638 drugs and gene expression data generated by extracting RNA from the toxicologically relevant organs and tissues (10,000 gene array).

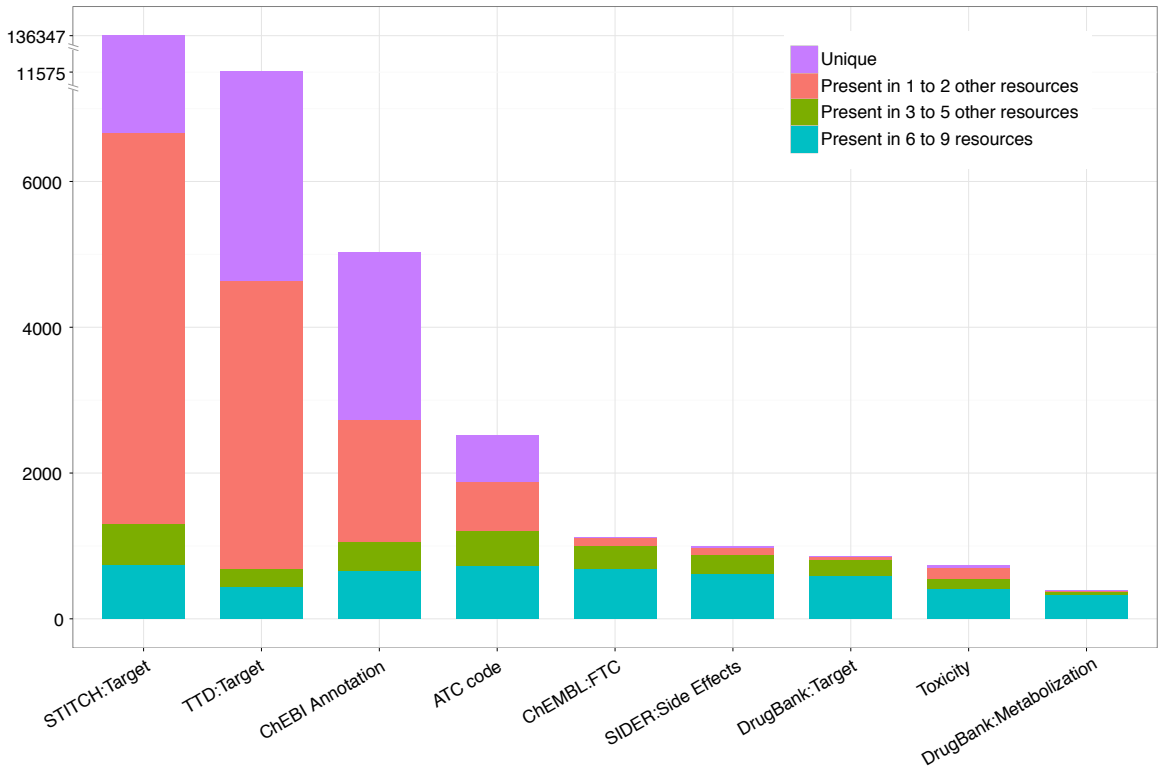


Figure 2: Overview of the resources integrated in CART.

4.3 Name matching

The Name Matching aims at fetching CIDs so that the Enrichment module can use them in order to retrieve annotation terms from any incorporated database.

Input The user input consists of a single file in which each line contains one or two columns. The first column holds the chemical name, and the second column a potential score assigned to the chemical (which might be used for ranked statistics).

Preparing the chemical universe The chemical universe encompasses both STITCH and PubChem. From the STITCH database, we downloaded two source files. The 'alias file' containing a one to one mapping between the CIDs and the chemical names, and the 'synonym file', containing a n to one mapping between the CIDs and the chemical names. The union of these two files needed to be performed for the simple reason that the 'alias file' contained many CIDs which were not found within the 'synonym file'. The 'non flat stitch file' contained thus all CIDs and chemical names present in the STITCH database. We subsequently added each CID as a synonym of itself to the chemical name column both for the 'non flat stitch universe' and 'non flat PubChem universe'. This aims at retrieving CIDs in the event that the user would directly input CIDs instead of chemical names. Finally, the CIDs of both universes were mapped against the flat CID mapping file in order to transform each CID to its parent CID in order to avoid stereochemical and salt form confusion(Figure 3). The final files 'flat stitch universe' and 'flat pubchem universe' were then indexed.

Indexing the chemical universe Chemical name matching is a complex problem for which we have not found any suitable service capable of matching thousands of chemical names against a database in order to retrieve their IDs in an acceptable time and with an acceptable accuracy. This task is hampered by the fact that there exists no officially accepted reference chemical universe that would contain every single compound ever studied and to which the community could refer. Resources like STITCH or PubChem are two initiatives beginning to address this issue. A second obstacle impeding this function lies in the fact that chemicals rarely have a single universally accepted name, making each compound having tens or sometimes hundreds of names, which engender a theoretical huge chemical universe. More importantly, chemical names very often consist of chemical formulas which when inputted by a user are prone to typos, as a results of which, searches for an exact chemical name within a database can often yield no results. Finally, chemical related resources create their own chemical IDs, rarely reference their IDs to other IDs and in the vast majority of cases do not include an exhaustive list of chemical names per ID, making the search for a particular chemical even more difficult. Some of these web services (STITCH) do not provide fuzzy matches, that is, recognizing an inputted chemical name even if it contains some typos. Some of these web services limit the number of possible inputted chemical names in order not to overload their server (STITCH limit the input to 200 chemical names) or disable completely multiple inputs (PubChem). Other web services do not even support fuzzy matches, that is, recognizing a chemical name even if it contains typos. The reason for these restrictions lies in the storing and indexing systems used by these web services. Many of them make use of DBMs including PostgreSQL [1], and even though DBMs can store and index vast amount of data and offer some semblance of text-based search functionalities, they are mostly tuned toward relational operations between tables including joins, and more generally at storing and manipulating structured data. DBMs will perform well when it comes to matching exact chemical names, but are not conceived for fuzzy matching. Thus, we decided to employ a full text search engine in order to store and index our chemical universe. Full text

[1] http://www.postgresql.org/

search engine can offer sub-second search results indicating which indexed documents contain exactly or partially the queried term out of millions or even billions of indexed documents, when DBMs can sometime take over a minute (Figure 4). They offer rich and flexible parameters, algorithms and distance metrics in order to rank the retrieved documents according to their similarity to the input query, which is very difficult to achieve with DBMs. Full text search engines are very well suited for adding, deleting or updating indexed document, which is advantageous to continuously integrate new chemicals that are deposited to PubChem on a daily basis. We have chosen to use the full text search engine library Apache Lucene [2]. It is a scalable, high-performance library entirely written in Java capable of indexing more than 150GB/hour on modern hardware [3], and has very small RAM requirements (only 1MB heap). It uses an incremental indexes as fast as batch indexing and the engendered index size is roughly 25% the size of text indexed. In order to index the chemical universes, we used Solr [4]. Solr is a web application built around lucene which functions as REST-like API. It can be used as an interface between the developer and lucene in order to index documents in lucene. Solr can be deployed in any servlet container and easilly configured through two configuration files (schema.xml, solrconfig.xml).

Matching algorithms Inputted chemical names are matched against the selected universe(s) (indexes) according to three matching algorithm. The exact matching looks for chemicals corresponding exactly to the inputted name and return a NA if no matches has been found. The fuzzy matching tolerates mistakes including missing misspelt chemicals. For instance, if the user inputs 'acetominophen' instead of 'acetaminophen', the name matching will still return the correct CID. The heuristic matching removes part(s) of the chemical names that are found in many chemicals and considered as non-informative. For instance "hcl", "dihydrochloride", "hydrochloride", "salt", "potassium", "dehydrate, "acid", "oxide" and "chloride".

Output The output of the name matching is a tab-delimited file containing five columns. ('name matching [foreground/background] file'). The first column contains the matched CIDs. The second column contains the score provided by the user in the input file. If no score is found in the file, this column contains NA values. The third column contains the chemical names inputted by the user. The fourth column contains the chemical names found in the universe and corresponding to the CID printed in the first column. The fifth column contains the name of the algorithm that identified the CIDs corresponding to the user's chemicals (EXACT MATCH, FUZZY MATCH or HEURISTIC MATCH) (Figure 1).

4.4 Enrichment

Input A first input corresponding to the output of the name matching tool is provided. This is the foreground. Optionally, a second input can be provided and will be considered as the background.

Module The fetched CIDs of the foreground are matched against the collected databases and corresponding annotation terms are retrieved. If background is provided, the same operation is repeated. If no background is provided, the default background is set to the CID found in the database(s) chosen for the enrichment tool, and thus, all the enrichment terms contained in the databases are retrieved. A fisher test is then computed on each annotation term retrieved and a correction method for multiple hypotheses

[2]http://lucene.apache.org/core/
[3]http://people.apache.org/ mikemccand/lucenebench/indexing.html
[4]http://lucene.apache.org/solr/

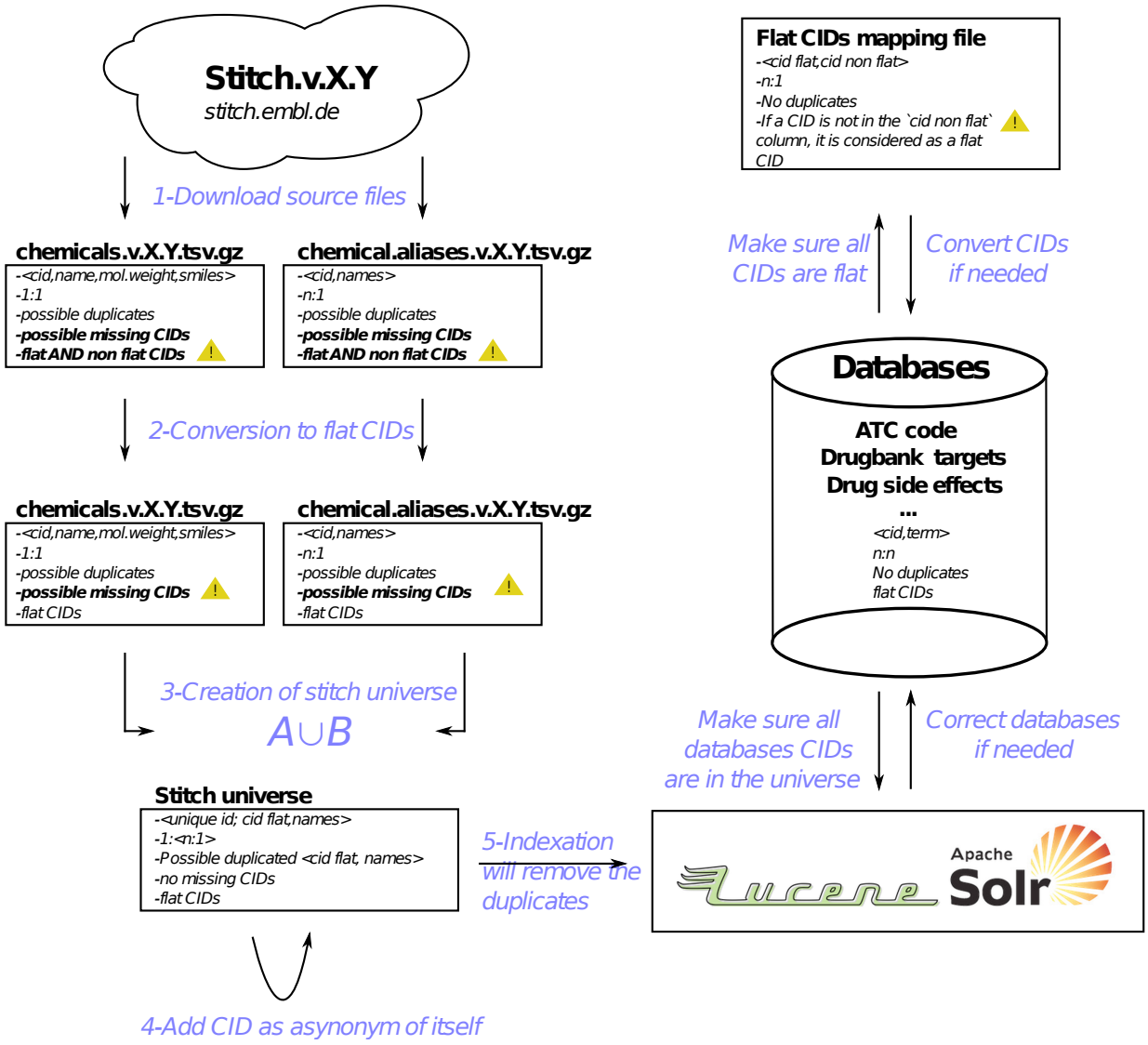


Figure 3: Resources integration process. (Top left)s Two source files are downloaded from the stitch web server. The ‘alias file‘ (named chemical.aliases.v.X.Y.tsv.gz) contains a one to one mapping between the CIDs and the chemical names. The ‘synonym file‘ (named chemicals.v.X.Y.tsv.gz) contains an n to one mapping between the CIDs and the chemical names. The CID of these files are first converted to flat CID using the flat CIDs mapping file (top right). The union of these two newly mapped flat CIDs mapped files is then performed. This is the Stitch universe. We subsequently added each CID as a synonym of itself to the chemical name column. (Top right) Resources were downloaded from URLs mentioned in the paragraph "Data integration" (4.2). The CIDs of each database are first converted to flat CIDs before being integrated to CART. Every CID of the databases has to be included in the indexed chemical universe.

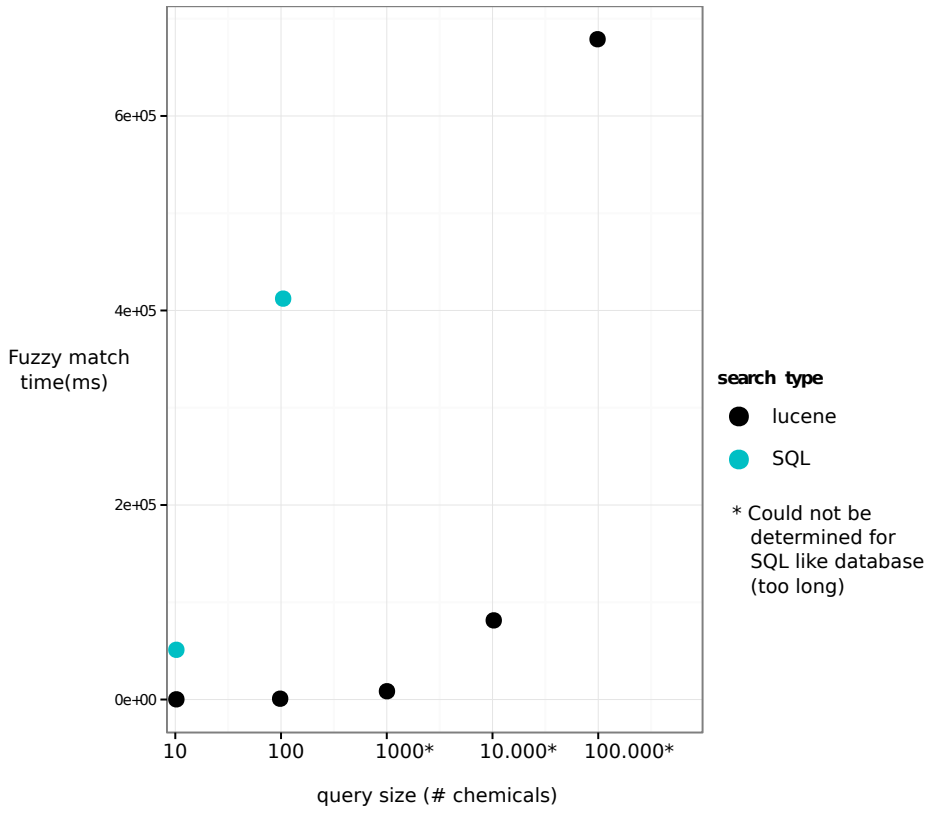


Figure 4: Overview of the lucene full text search engine chemical fuzzy name search performance vs an InnoDB full text index (MySQL). Search time grows linearly for the lucene full teaxt search engine whereas it grows exponentially for the InnoDB full text index.

is applied. Annotation terms yielding a P value inferior to the type I error cut off (by default 0.05) are retained and displayed to the user.

Output Two outputs files are provided. The first output is a tab-delimited file, this is the 'enrichment output file'. It contains the enriched terms the enriched terms along with database of origin, links to external resources, p-values and odd-ratios. The second output is another tab-delimited file, this is the 'annotation output file'. It contains all the annotation terms that have been retrieved for the foreground chemicals using the selected resources (Figure 1).

4.5 Visualization

Input Three input files are required: the two output files from the enrichment module as well as the output of the name matching module corresponding to the foreground.

Output The visualization module generates two .html files. The first one represents a table of the enriched biological annotation terms including URL links redirecting to the databases where the terms

was extracted from. The second .html file represents a dynamic and interactive network representing the links between the chemical from the foreground and their annotation terms. Five layouts are currently included, which algorithmically position the chemicals and biological annotation terms in the graph (Figure 1). The network generation was implemented using the cytoscape.js API [167] and thus benefits of many of the cytoscape.js features.

4.6 Synonyms

Input The required input is the same as the one described in the name matching module

Output The synonym modules uses the same search algorithm described in the Name Matching module and uses PubChem as a default chemical universe. It returns a two tab-delimited

4.7 Web interface

The web interface has been implemented with the galaxy framework. Galaxy is a flexible and modular open source web based platform automatically managing user sessions and workspaces. It keeps track of every file inputted and job ever run. The software is integrated into galaxy and each module can be run independently from each other. The whole workflow can also be run at once. Inputted and outputted files appear in the user's history (or workspace) and can be downloaded and edited, and corresponding jobs can be re-run with the exact same settings, which. Since the history keeps track of a user's action, galaxy is an excellent platform for user to perform reproducible data analysis. Galaxy uses by default a NoSQL database, which is not suitable for a multi user application. Thus, we configured it to work with a PostgreSQL database.

4.8 Standalone version

The software is available at: http://cart.embl.de/static/automatic-installer.sh.
A single shell script takes care of the entire installation process. It downloads the software, resources and sets up the user environment.

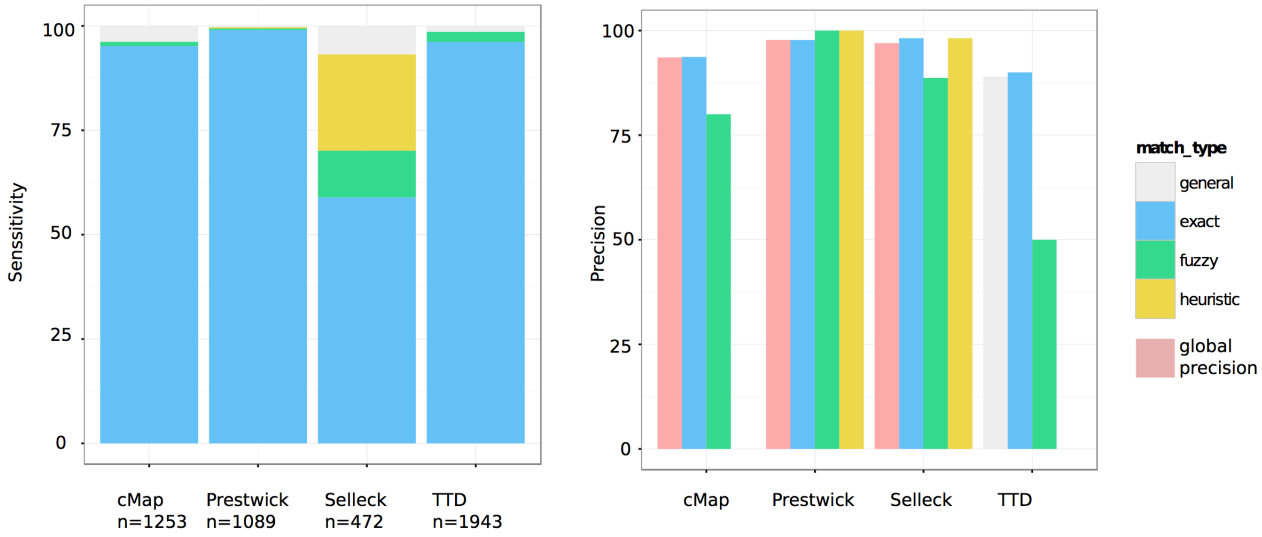


Figure 5: (Left) Overview of the name matching module sensitivity benchmarked against four independent datasets. The name matching modules yields an average of 94% sensitivity and the fuzzy and heuristic match always increase the sensitivity of the search. (Right) Overview of the name matching module precision benchmarked against the same four datasets. The name matching modules yields an average 93% precision.

4.9 Benchmarking

Accuracy The core of CART lies in the name matching module. Fetching the correct CID by matching to the correct chemical is of upmost importance in order not to wrongly annotate a chemical set. Thus, we measured the reliability of the name matching module by evaluating its accuracy. To do so, we have selected four datasets whose chemicals had already been matched to either STITCH CIDs or PubChem CIDs. We subsequently fed the name matching module with the chemical names present in those four datasets and measured the accuracy of matching by comparing the CIDs fetched by the name matching module to those manually matched by the original authors of the dataset (Figure 5). The sensitivity was computed as being the ratio of fetched CIDs over the total number of inputted chemicals and the precision as being the fraction of correctly matched CIDs. On average, the name matching module yielded a 96% sensitivity and 94% precision. In all datasets benchmarked, the fuzzy matching and heuristic matching increased the sensitivity (on average by 12%) and the specificity. In the future, we plan to include an option through which the user will be able to adjust the heuristic matching by setting himself words he wishes to remove from chemical names.

Speed As mentioned earlier, the main motivation for choosing a full text search engine as core of the name matching tool over a DBM lied in the fact that fuzzy matching results can be return within a second depending on the size of the query. We benchmarked the speed of the name matching module on a Macintosh HD 2.8Ghz Intel Core i5 with 8GB of RAM. To this purpose, we generated several datasets of different increasing size by randomly selecting the chemicals from the PubChem universe and run them through the name matching module using the STITCH universe. In a first time, we measured the time taken for it to return the results in order to see whether the trend was linear or exponential. In a second time, we decomposed it per algorithm in order to analyze whether the fuzzy and heuristic matching took significantly longer than the exact matching. Finally, we compared these results to those obtained with a MySQL database having indexed the same universe using a B-tree index (Figure 6). The results of the

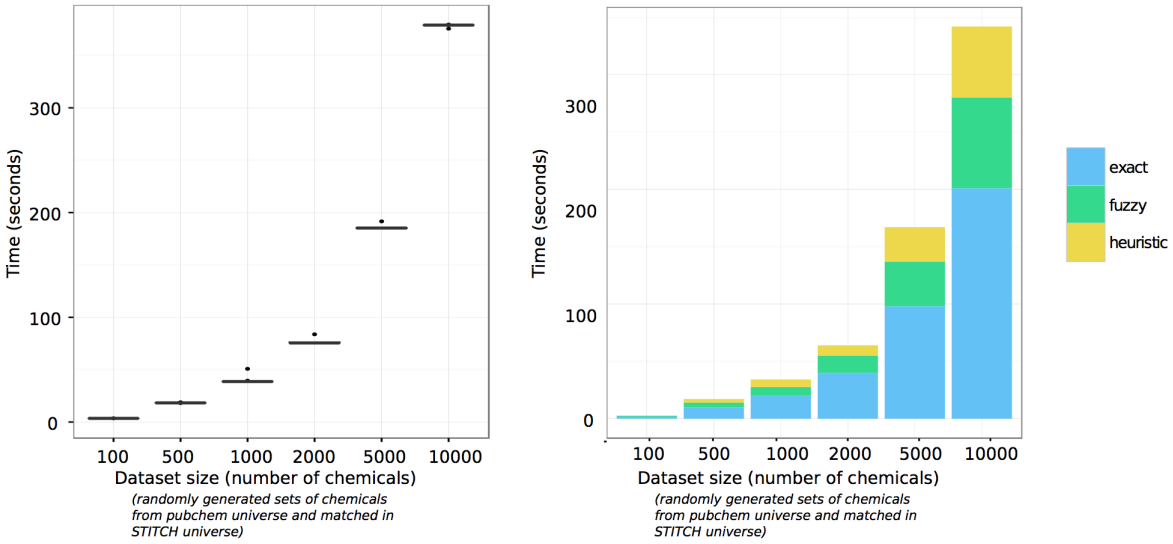


Figure 6: (sLeft) Box plot illustrating the name matching module speed. The x axis represents the dataset size. For each size, 10 random chemical name datasets were generated using the pubchem universe. The y axis represent the time (in seconds) required to match all chemicals using the three chemical name searching algorithms (exact, fuzzy, heuristic). (Right) Detailed overview of the time required by each chemical name searching algorithm using one dataset per size.

benchmarking clearly indicated that the matching speed grew linearly with the size of the query and that the fuzzy and heuristic matching were not significantly slower than the exact matching. When it came to exact matching, DBMs performed slightly better but proved to be unusable when it came to fuzzy and heuristic matches.

Chapter 5

Future directions

CART is a software providing an integrative, scalable, accurate and very fast chemical name matching tool and annotation platform. It can be used in any small molecule high-throughput screening or in any study attempting to make sense of chemical sets. It is available over the web with via an intuitive graphical user interface managing user sessions and is also available as a command line tool for more programmatic uses.

Improvement on the name matching So far, the heuristic matching works with a set of commonly encountered terms that can be removed from the chemical name to increase the likelihood of finding a match. The next version could take into account user specific terms, which should be avoided during the heuristic match of the name matching step. On a technical note, the parallelization of the index building and the name search will also greatly improve the search time performance. Preliminary tests have shown that the search time performance could be increased by a factor 8 (with 10 CPU used as opposed to one in the current release of CART). The parallelization of the name matching step is currently in beta testing. The nature of the name matching could also be extended. CART expects chemical names as an input so that the name matching can match them against the chemical universe chose, which solely consists of chemical names. The scalability and flexibility of the search engine integrated could make possible the integration of chemical structures or experimental outputs. This would be particularly useful for scientists working with high-throughput technologies. Most of the time, high-throughput technologies output particular chemical fingerprints (for instance the NMR or MS profile of a chemical). These could be used as an alternative of chemical names for the ID retrieving step. The advantage of such a matching algorithm would be that it does not assume that the user (typically an experimentalist) knows anything about its chemicals.

More Integration CART includes seven databases from which it extracts biological annotation terms to extend the covered molecular and phenotypic read-outs. One of the first resources to be integrated will be LINCS [168]. Integrating LINCS would enable users having a better understanding of the gene expression profiles and phenotypic readouts induced by more than 10,000 chemicals.

Prediction module The goal of the prediction module is to predict a phenotypic read-out based on another readout. For instance, given a set of toxicological endpoints, can we predict a drug side effect(s)? Can new drug-target connections be inferred from toxicological similarity? Can we use a drug's cytochrome P 450 profile to infer side effect or toxicological endpoints? These predictions would imply measuring the pairwise similarity of inputted chemicals with regards to two different read-outs (for instance, cytochrome P450 and side effects) and to perform a ROC analysis to get the prediction power

of one read-out with regard to a second read-out. This module is still in beta test and will be part of the next CART release.

Part IV

Conclusion

Conclusion

In this dissertation, state-of-the-art computational methods are presented to take the maximum advantage of large scale small molecule data that is increasingly becoming available. I contributed to the field of large scale chemical biology by developing computational methods analyzing and making sense of protein-lipid interactions using readouts produced by LiMA, cutting edge technology with a focus on PH-domains. The results of this analysis pipeline are illustrated in two publications [28] [38]. The analysis pipeline itself is presented in a third publication, whose manuscript is currently in preparation [37], so that it can be reused in any other screen utilizing the technology LiMA with any kind of lipid or protein. Given that a eukaryotic cell usually produces more than 1,000 different lipid species that possess a wide range of structural, physical and biochemical properties, the rules governing protein-lipid interactions are far from being solved, and this work is a first step toward a systematic understanding of these regulatory elements.

The computational tools developed in this context have for instance demonstrated that the membrane recruitment of PH domains necessitates the presence of secondary lipids, which can either increase or decrease the binding affinity of PH domains to the biological membranes. The undertook computational analyses have shown that most PH domains were capable of cooperative recruitment and that it generally translated into similar biological functions (Figure 24). Additionally, the clustering analyses has shown how PH-domain liposome binding data reflected the actual biological repartition of lipids within plasmic and organellar membranes. It led to the identification of a motif conferring PH domains the ability to bind one or both of these membranes and thereby helped to interpret why and how some PH domains were more involved in cellular transport activation than others. The developed methods also helped enabling the interpretation of disease-associated mutations, for the discovered motif was also found to contain amino-acids that are often found naturally mutated in some human cancer biopsies (Figure 22 and Figure 23). We also generalized the concept of cooperative binding to the majority of PH domains, which has important consequences in terms of biological processes (Figure 25).

However, future screens might contain many sources of batch effects [1]. Therefore, some steps would probably need to be generalized in order to apply the computational analysis pipeline to such screens. For instance, the current implementation has been adapted to the screen presented in this study [38], and does not normalize the NBI for protein concentration since no batch effect has been observed on that front. However, other screens whose protein concentrations are not found to be at saturating conditions could realistically generate NBI that are heavily influenced by the protein concentration. These NBIs would need to be normalized in order not to misinterpret some binding events that could lead to false biological conclusions. The consideration of possible spatial bias as a possible source of batch effects would also need to be integrated into future implementations of this computational pipeline. For that purpose, we could envisage applying iterative rank-order normalization methods or neighborhood expectation maximization algorithms to correct for spatial bias, as it has been already applied for gene

[1] Non biological sources of variation

expression microarray data [94] [169].By doing so, LiMA could be employed to study any protein-lipid interaction screen containing sources of batch effects.

Such target-based screens could also be used in combination with chemical libraries to try perturbing protein-lipid interactions that are found exacerbated in some diseases including diabetes and cancer. In that respect, LiMA could represent a new door to identify new classes of drugs specifically perturbing some protein-lipid interactions. Understanding what these perturbing agents have in common and making inferences on their modes of action and their complex effects on biological systems would require efforts on the data integration front, and CART represents a milestone in that respect. CART provides the framework required to find common biological themes amongst chemicals and a suited platform for the scalable integration of chemical centric databases. CART currently works with chemical names, which restricts its use-cases by implicitly expecting a user to know the names of his chemical of interests. Ideally, CART should also be able to recognize structures from SMILES [2], sketches or more interestingly from Mass Spectrometry or Nuclear Magnetic Resonnace signatures. Comparing the bioactivity properties of known or unknown chemicals across all the available databases would enable relevant predictions of chemicals impact on biological processes, and thereby unleash the true potential of computational chemical biology.

I hope my contributions to the field of chemical biology have brought us closer to understand how a tiny fraction of the estimated 10^{18} to 10^{200} [3] small molecules on earth impact biological systems.

[2] A SMILE is a one line representation of a chemical's formula

[3] The author who came up with this number could not be identified: http://www.wisegeek.com/how-many-chemicals-are-there.htm

Part V

Supplementary Information

Chapter 6

Relative to Part II

Table V.1: Summary of the most popular chemical centric databases. The first column indicate the name of the database, the second column the readout associated to each chemical of the database, and the third column the URL under which the database is accessible.

Database	Description	Reference
PubChem	Chemical/bioactivity	https://pubchem.ncbi.nlm.nih.gov/
ChEMBL	Chemical/Bioactivity	https://www.ebi.ac.uk/chembl/
ChEBI	Chemical/Bioactivity	https://www.ebi.ac.uk/chebi
ChemBank	Chemical/bioactivity	http://chembank.broadinstitute.org/
DrugBank	Targets	www.drugbank.ca/
BIDD	Targets	http://bidd.nus.edu.sg/
STITCH	Targets	http://stitch.embl.de/
BIND	Targets	www.bindingdb.org/
ToxScan	Toxicological readout	http://ntp.niehs.nih.gov/results/dbsearch/index.html
SIDER	Side effects	http://sider.embl.de/
KiBank	Structural information	http://www.ncbi.nlm.nih.gov/pubmed/15556481
ZINC	Structural information	http://zinc.docking.org/
GPCRDB	G protein-coupled receptors	www.gpcr.org/7tm/
NucleaRDB	Nuclear Hormone Receptor	www.receptors.org/nucleardb/
NURSA	Nuclear Receptor Signalling	www.nursa.org/
IUPHAR	Pharmacological targets	www.iuphar-db.org/

Table V.2: Table summarizing the 30 different lipids used in this study (28 lipids of interests, 1 PEGylated lipid and 1 biotinylated lipid). The first column indicates the family of the lipid, the second column the name of the lipid, the third column the abbreviation of the lipid used in this study, the fourth column the possible synonyms and the fifth column the CAS number of the lipid.

Family	Name	Abreviation	Synonyms	CAS No
Glycerolipids	1,2-di-(9Z-octadecenoyl)-sn-glycerol	DAG	1,2-dioleoyl-sn-glycerol	24529-88-2
Glycerophospholipids	1-hexadecanoyl-2-(9Z-octadecenoyl)-sn-glycero-3-phosphocholine	PC	POPC	26853-31-6
Glycerophospholipids	1,1',2,2'-tetra-(9Z-octadecenoyl) cardiolipin	Cardiolipin	tetraoleoyl cardiolipin	115404-77-8
Glycerophospholipids	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-phosphoserine	PS	DOPS	90693-88-2
Glycerophospholipids	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-phospho-(1'-rac-glycerol)	PG	DOPG	67254-28-8
Glycerophospholipids	1-hexadecanoyl-2-(9Z-octadecenoyl)-sn-glycero-3-phosphoethanolamine	PE	POPE	26662-94-2
Glycerophospholipids	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-phosphate	PA	DOPA	108392-02-5
Glycerophospholipids	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-phosphoinositol	PI	DOPI	799268-53-4
Glycerophosphoinositol phosphates	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-phospho-(1'-myo-inositol-3'-phosphate)	PI(3)P	DOPI(3)P	1246303-09-2
Glycerophosphoinositol phosphates	1,2-di-(palmitoyl)-sn-glycero-3-phospho-(1'-myo-inositol-4'-phosphate)	PI(4)P	DPPI(4)P	n/a
Glycerophosphoinositol phosphates	1-heptadecanoyl-2-(5Z,8Z,11Z,14Z-eicosatetraenoyl)-sn-glycero-3-phospho-(1'-myo-inositol-4'-phosphate)	PI(4)P*	n/a	475995-51-8
Glycerophosphoinositol phosphates	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-phospho-(1'-myo-inositol-5'-phosphate)	PI(5)P	DOPI(5)P	1246303-10-5
Glycerophosphoinositol phosphates	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-phospho-(1'-myo-inositol-3',4'-bisphosphate)	$PI(3,4)P_2$	$DOPI(3,4)P_2$	799268-54-5
Glycerophosphoinositol phosphates	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-phospho-(1'-myo-inositol-3',5'-bisphosphate)	$PI(3,5)P_2$	$DOPI(3,5)P_2$	799268-55-6
Glycerophosphoinositol phosphates	1,2-di-(palmitoyl)-sn-glycero-3-phospho-(1'-myo-inositol-3',5'-bisphosphate)	$PI(3,5)P_2$#	$DPPI(3,5)P_2$	n/a
Glycerophosphoinositol phosphates	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-phospho-(1'-myo-inositol-4',5'-bisphosphate)	$PI(4,5)P_2$	$DOPI(4,5)P_2$	799268-56-7
Glycerophosphoinositol phosphates	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-[phosphoinositol-3,4,5-trisphosphate]	$PI(3,4,5)P_3$	$DOPI(3,4,5)P_3$	799268-57-8
Sphingolipids	N-(octadecanoyl)-4R-hydroxysphinganine	Phytocer		34354-88-6
Sphingolipids	N-(octadecanoyl)-sphinganine	Dihydrocer	N-(stearoyl)-dihydroceramide	2304-80-5
Sphingolipids	N-(hexadecanoyl)-sphing-4-enine	Cer	N-palmitoyl-D-erythro-sphingosine	24696-26-2
Sphingolipids	N-(hexadecanoyl)-sphing-4-enine-1-phosphate	Cer1P	N-palmitoyl-ceramide-1-phosphate	1246303-22-9
Sphingolipids	Sphing-4-enine	Sphingosine	D-erythro-sphingosine	123-78-4
Sphingolipids	4R-hydroxysphinganine	PHS		388566-94-7
Sphingolipids	Sphinganine	DHS	Dihydrosphingosine	764-22-7
Sphingolipids	(2S,3S,4R)-2-amino-3,4-dihydroxyoctadecyl dihydrogen phosphate	PHS1P		383908-62-1
Sphingolipids	Sphinganine-1-phosphate	DHS1P		19794-97-9
Sphingolipids	Sphing-4-enine-1-phosphate	S1P		26993-30-6
Pegylated lipid	1,2-di-(9Z-octadecenoyl)-sn-glycero-3-phosphoethanolamine-N-[methoxy(polyethylene glycol)-350]	PE-PEG350		474922-90-2
Fluorescent lipid		PE-Atto647		n/a
Biotinylated lipid		PE-Biotin	n/a	384835-54-5

Table V.3: Table summarizing the lipid mixtures used in the liposomes of this study. Each mixture listed in the table is complemented with PC (carrier lipid) as well as PEGylated lipid and a fluorescent lipid as liposome marker. The concentrations indicated in the sixth, seventh, eight and ninth column are in mol %.

Liposome ID	Lipid composition				Lipid concentration			
	Primary lipid	Secondary lipid 1	Secondary lipid 2	Secondary lipid 3	Primary lipid	Secondary lipid 1	Secondary lipid 2	Secondary lipid 3
Lipid_ID_1	PI(3)P				10,0			
Lipid_ID_2	PI(4)P				10,0			
Lipid_ID_3	PI(5)P				10,0			
Lipid_ID_4	PI(3,4)P2				10,0			
Lipid_ID_5	PI(3,5)P2				10,0			
Lipid_ID_6	PI(3,5)P2#				10,0			
Lipid_ID_7	PI(4,5)P2				7,0			
Lipid_ID_8	PI(4,5)P2				10,0			
Lipid_ID_9	PI(3,4,5)P3				10,0			
Lipid_ID_10	Sphingosine				10,0			
Lipid_ID_11	PHS				10,0			
Lipid_ID_12	DHS				10,0			
Lipid_ID_13	S1P				10,0			
Lipid_ID_14	PHS1P				10,0			
Lipid_ID_15	DHS1P				7,0			
Lipid_ID_16	DHS1P				10,0			
Lipid_ID_17	Cer				10,0			
Lipid_ID_18	Cer1P				10,0			
Lipid_ID_19	Phytocer				10,0			
Lipid_ID_20	Dihydrocer				10,0			
Lipid_ID_21	PC				99,4			
Lipid_ID_22	PS				10,0			
Lipid_ID_23	PI				10,0			
Lipid_ID_24	Cardiolipin				10,0			
Lipid_ID_25	PG				10,0			
Lipid_ID_26	PE				10,0			
Lipid_ID_27	PA				10,0			
Lipid_ID_28	DAG				5,0			
Lipid_ID_29	PI(3)P	Sphingosine			10,0	10,0		
Lipid_ID_30	PI(3)P	PHS			10,0	10,0		
Lipid_ID_31	PI(3)P	DHS			10,0	10,0		
Lipid_ID_32	PI(3)P	S1P			10,0	10,0		
Lipid_ID_33	PI(3)P	PHS1P			10,0	10,0		
Lipid_ID_34	PI(3)P	DHS1P			10,0	10,0		
Lipid_ID_35	PI(3)P	Cer			10,0	10,0		
Lipid_ID_36	PI(3)P	Cer1P			10,0	10,0		
Lipid_ID_37	PI(3)P	Phytocer			10,0	10,0		
Lipid_ID_38	PI(3)P	Dihydrocer			10,0	10,0		
Lipid_ID_39	PI(4)P	Sphingosine			10,0	10,0		
Lipid_ID_40	PI(4)P	PHS			10,0	10,0		
Lipid_ID_41	PI(4)P	DHS			10,0	10,0		
Lipid_ID_42	PI(4)P	S1P			10,0	10,0		
Lipid_ID_43	PI(4)P	PHS1P			10,0	10,0		
Lipid_ID_44	PI(4)P	DHS1P			10,0	10,0		
Lipid_ID_45	PI(4)P	Cer			10,0	10,0		
Lipid_ID_46	PI(4)P	Cer1P			10,0	10,0		
Lipid_ID_47	PI(4)P	Phytocer			10,0	10,0		
Lipid_ID_48	PI(4)P	Dihydrocer			10,0	10,0		
Lipid_ID_49	PI(5)P	Sphingosine			10,0	10,0		
Lipid_ID_50	PI(5)P	PHS			10,0	10,0		
Lipid_ID_51	PI(5)P	DHS			10,0	10,0		
Lipid_ID_52	PI(5)P	S1P			10,0	10,0		
Lipid_ID_53	PI(5)P	PHS1P			10,0	10,0		
Lipid_ID_54	PI(5)P	DHS1P			10,0	10,0		
Lipid_ID_55	PI(5)P	Cer			10,0	10,0		
Lipid_ID_56	PI(5)P	Cer1P			10,0	10,0		
Lipid_ID_57	PI(5)P	Phytocer			10,0	10,0		
Lipid_ID_58	PI(5)P	Dihydrocer			10,0	10,0		
Lipid_ID_59	PI(3,4)P2	Sphingosine			10,0	10,0		
Lipid_ID_60	PI(3,4)P2	PHS			10,0	10,0		
Lipid_ID_61	PI(3,4)P2	DHS			10,0	10,0		
Lipid_ID_62	PI(3,4)P2	S1P			10,0	10,0		
Lipid_ID_63	PI(3,4)P2	PHS1P			10,0	10,0		
Lipid_ID_64	PI(3,4)P2	DHS1P			10,0	10,0		
Lipid_ID_65	PI(3,4)P2	Cer			10,0	10,0		
Lipid_ID_66	PI(3,4)P2	Cer1P			10,0	10,0		
Lipid_ID_67	PI(3,4)P2	Phytocer			10,0	10,0		
Lipid_ID_68	PI(3,4)P2	Dihydrocer			10,0	10,0		
Lipid_ID_69	PI(3,5)P2	Sphingosine			10,0	10,0		
Lipid_ID_70	PI(3,5)P2	PHS			10,0	10,0		
Lipid_ID_71	PI(3,5)P2	DHS			10,0	10,0		
Lipid_ID_72	PI(3,5)P2	S1P			10,0	10,0		
Lipid_ID_73	PI(3,5)P2	PHS1P			10,0	10,0		
Lipid_ID_74	PI(3,5)P2	DHS1P			10,0	10,0		
Lipid_ID_75	PI(3,5)P2	Cer			10,0	10,0		
Lipid_ID_76	PI(3,5)P2	Cer1P			10,0	10,0		
Lipid_ID_77	PI(3,5)P2	Phytocer			10,0	10,0		
Lipid_ID_78	PI(3,5)P2	Dihydrocer			10,0	10,0		

Lipid_ID_79	PI(3,5)P2#	Sphingosine			10,0	10,0		
Lipid_ID_80	PI(3,5)P2#	PHS			10,0	10,0		
Lipid_ID_81	PI(3,5)P2#	DHS			10,0	10,0		
Lipid_ID_82	PI(3,5)P2#	S1P			10,0	10,0		
Lipid_ID_83	PI(3,5)P2#	PHS1P			10,0	10,0		
Lipid_ID_84	PI(3,5)P2#	DHS1P			10,0	10,0		
Lipid_ID_85	PI(3,5)P2#	Cer			10,0	10,0		
Lipid_ID_86	PI(3,5)P2#	Cer1P			10,0	10,0		
Lipid_ID_87	PI(3,5)P2#	Phytocer			10,0	10,0		
Lipid_ID_88	PI(3,5)P2#	Dihydrocer			10,0	10,0		
Lipid_ID_89	PI(4,5)P2	Sphingosine			10,0	10,0		
Lipid_ID_90	PI(4,5)P2	PHS			10,0	10,0		
Lipid_ID_91	PI(4,5)P2	DHS			10,0	10,0		
Lipid_ID_92	PI(4,5)P2	S1P			10,0	10,0		
Lipid_ID_93	PI(4,5)P2	PHS1P			10,0	10,0		
Lipid_ID_94	PI(4,5)P2	DHS1P			7,0	7,0		
Lipid_ID_95	PI(4,5)P2	DHS1P			10,0	10,0		
Lipid_ID_96	PI(4,5)P2	Cer			10,0	10,0		
Lipid_ID_97	PI(4,5)P2	Cer1P			10,0	10,0		
Lipid_ID_98	PI(4,5)P2	Phytocer			10,0	10,0		
Lipid_ID_99	PI(4,5)P2	Dihydrocer			10,0	10,0		
Lipid_ID_100	PI(3,4,5)P3	Sphingosine			10,0	10,0		
Lipid_ID_101	PI(3,4,5)P3	PHS			10,0	10,0		
Lipid_ID_102	PI(3,4,5)P3	DHS			10,0	10,0		
Lipid_ID_103	PI(3,4,5)P3	S1P			10,0	10,0		
Lipid_ID_104	PI(3,4,5)P3	PHS1P			10,0	10,0		
Lipid_ID_105	PI(3,4,5)P3	DHS1P			10,0	10,0		
Lipid_ID_106	PI(3,4,5)P3	Cer			10,0	10,0		
Lipid_ID_107	PI(3,4,5)P3	Cer1P			10,0	10,0		
Lipid_ID_108	PI(3,4,5)P3	Phytocer			10,0	10,0		
Lipid_ID_109	PI(3,4,5)P3	Dihydrocer			10,0	10,0		
Lipid_ID_110	PI(3)P	PS			10,0	10,0		
Lipid_ID_111	PI(4)P	PS			10,0	10,0		
Lipid_ID_112	PI(5)P	PS			10,0	10,0		
Lipid_ID_113	PI(3,4)P2	PS			10,0	10,0		
Lipid_ID_114	PI(3,5)P2#	PS			10,0	10,0		
Lipid_ID_115	PI(3,5)P2	PS			10,0	10,0		
Lipid_ID_116	PI(4,5)P2	PS			7,0	7,0		
Lipid_ID_117	PI(4,5)P2	PS			10,0	10,0		
Lipid_ID_118	PI(3,4,5)P3	PS			10,0	10,0		
Lipid_ID_119	DAG	PS			5,0	5,0		
Lipid_ID_120	PE	PA			10,0	10,0		
Lipid_ID_121	PE	PA	PS		10,0	10,0	10,0	
Lipid_ID_122	PE	PA	PS	PI	10,0	10,0	10,0	10,0
Lipid_ID_123	PE-Biotin				4,0			
Lipid_ID_124	PE-Biotin				7,0			
Lipid_ID_125	PE-Biotin				10,0			

Table V.4: Table summarizing the 95 LBDs used in this assay, amongst which 91 are PH domains. The first column indicates the name of the LBD (_1 and _2 are different variants of the same LBD, _A and _B are orthologous of the same protein in S. cerevisiae and C. thermophilum). The second column is the standard name of the protein, the third column indicate the PH domain prediction score (e value) using the SMART database, the fourth column the PMID of the publication in which the domain was discovered. The fifth column includes in the sequence of the domain used in this assay and the sixth column show the median protein concentration used across all replicates (μM). The seventh column indicates the mean TI of the LBD across all its interactions and the eighth column the cooperativity class to which the PH domain belong to. The ninth and tenth columns show cellular localization of the PH domain.

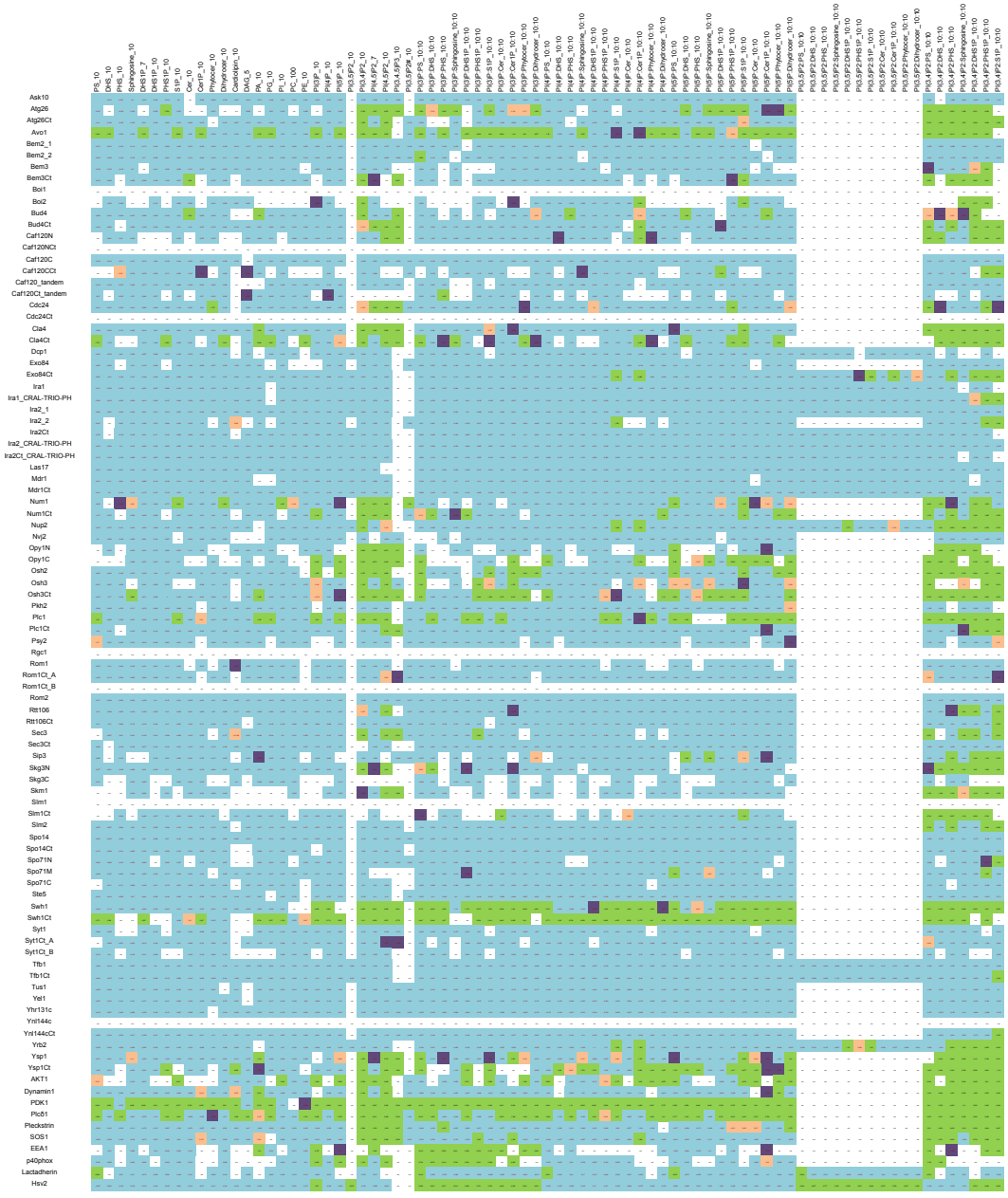
LBD name	Standard name	SMART e-value score (BLAST/NCBI CDS e-value)	Reference (PMID)	Domain sequence (AA)	Median protein concentration [μM]	Interaction with liposomes (0 <TI <2)	Cooperativity class	Annotations of *in vivo* localization (Huh *et al.*, 2003)	Annotations used in Fig. 4
Ask10	Ask10p	7,83E-07	15023338	RKVSDVIYPHMKSPLAKCIKAGYFLKKSELLPTYH QGYFVLTSNYIHEFQSSDFYNLSSSTPNSTKSSA YSSSVSIADTYANANNAKANNHHRQASDVHNSS TTTGGTAGANGIRGIRKKSYLAPIMSIPLNDCTLK DASSTKFVLVGKPTLNENADVRKSSSSTYLSGSS QASLPKYGHETAKIFSKAPFHKFLKGSKPKNKNT KSSELDQFYAAAQKESNNYVTWTFKIVSPEPSEE ELKHFKRWVQDLKNLTSFNDTK	2.99‡	Yes	2a	Cytoplasm	Nucleus#
Atg26	Atg26p	2,03E-14	15023338	KNPRSVKMSGNLNIIRTKLIRSTRYWCVLKNHLFS MYTSSTELYFPVLTIDLREVQKIETQKHTLNGSAT KTFKLYTDESTFKFNADSEFSAKSWVNALKKEQF AAQNSENN	12.84‡	Yes	1	Cytoplasm	Other
Atg26Ct	-	1,05E-16	-	NEVVKSGYLSKCGKRNPKYNRYWFRLKGVVLSY YRDPQDLYFPSGHIDLRYGISASITDKDKEGINFTI	20.37‡	Yes	-	n/a	-
Avo1	Avo1p	42.5 (BLAST: 5E-54)	22505404	DNFQDLFTGAYHKYKVWRRQQMSFINKHERTLA IDGDYIYIVPPEGRIHWHDNVKTKSLHISQVVLVK	8.68‡	Yes	1	n/a	Other
Bem2_1	Bem2p	4.21 (BLAST: 3E-59)	15023338	NTGQTPSLGSVMESNNSARNRRDSRASFSTNR SSVVSNSSHNGVSKKIGGFFRRPFSIGGFNTSSS	3,18	No	-	Bud neck, cytoplasm	-
Bem2_2		22 (NCBI CDS: 5.51E-05)	-	KKIGGFFRRPFSIGGFNTSSSNYSLNSILSQEVSS NKSILPSILPEVDSMQLHDLKPSYSLKTFEIKSIME IINHRNIPAYYYAFKIVMQNGHEYLIQTASSSDLTE	8,27	Yes	2a	Bud neck, cytoplasm	Cdc42 network
Bem3	Bem3p	4,85E-12	15023338	MDDNVKDGSLLLRRPKTLTGNSTWRVRYGILRD DVLQLFDKNQLTETIKLRQSSIELIPNLPEDRFGT RNGFLITEHKKSGLSTSTKYYICTETSKERELWLS AFSDYIDPSQS	4.54‡	Yes	2a	Bud neck, cytoplasm	Cdc42 network
Bem3Ct	-	1,66E-19	-	GGRPYRSGYLTKRGKNFGQWKARYFVLDGPQL KYYETPGGAHLGTIKLRGAQIGKQTNHSNDGSQ GSNEDGDNQYRHAFLILEPKKKDPNSMTKHVLC AESDKERDQWVDALIRWVDYKDP	16.44‡	Yes	2b	n/a	-
Boi1	Boi1p	9,11E-17	15023338	SMQTADCSGWMSKKGTGAMGTWKQRFFTLHG TRLSYFTNTNDEKERGLIDITAHRVLPASDDDRLI SLYAASLGKGKYCFKLVPPQPGSKKGLTFTEPRV HYFAVENKSEMKAWLSAIIKATIDIDTSVPVISS	7,05	No	-	Bud, bud neck, cell periphery, cytoplasm	Cdc42 network
Boi2	Boi2p	1,48E-16	15023338	SISVKEAMKDADFSGWMSKKGSGAMSTWKTRF FTLHGTRLSYFSSTTDTRERGLIDITAHRVVPAKE DDKLVSLYAASTGKGRYCFKLLPPQPGSKKGLTF TQPRTHYFAVDNKEEMRGWMAALIKTTIDIDTSV P	34.08‡	Yes	2b	Bud, bud neck, punctuate composite, cytoplasm	Cdc42 network
Bud4	Bud4p	4,10E-10	15023338	QNIYKEGYLLQDGGDLKGKIENRFFKLHGSQLSG YHEISRKAKIDINLLKVTKVLRNEDIQADNGGQRN FTDWVLFNECFQLVFDDGERITFNAECSNEEKS DWYNKLQEVVELNVFHQ	3.83‡	Yes	2a	Bud neck	Cdc42 network
Bud4Ct	-	3,62E-11	-	ARTWEGHLSQQGDDCPYWRRYFKLVGTKLTA YHEATRQPRATINLANAKRLIDDRRTLLEKGTIDK NGKRRRSAFAEDEEGYMFVEEGFRIRFNNGEVI DFYADSTKEKEEWMRVLGEVIGRDSLL	11.41‡	Yes	2b	n/a	-
Caf120N	Caf120p	3,65E-07	15023338	SPVSPELVPILTLLNAHTHRRYHEGVFLILQDLNN NGTHAARKWKDVYGVLLGTQLALWDAKELAEFT DPSCPVSEKKLKEVASKPTYINLTDATLDNSD NIVMECGKNLTNALVVSTTLKNRYFLQFGNKESF NAWNSAIRLCLYECSSLQEAYTGAFISSRGAKLG	5,35	Yes	2b	Bud, bud neck, cytoplasm	Other
Caf120NCt	-	3,23E-08	-	LQPIFSLLNGHANKLYQEGYFLKLDDQDIRGNLN PDRTWTECFAQLVGTVLSLWDAAELDAAGEDGE VLPKFINLTDASIKMIESLPTKSNDEQPLQNILSIST AGRNRYLLHFNSHHSLIQWTSGIRLAMYEHATLQ EAYTGSLVAG	9,96	n/a	-	n/a	-
Caf120C	Caf120p	not detected (BLAST: 5E-63)	-	YDYKDWVSVRFGAGMPWKRCYAVISQSSSKKK GHFGEINLYENDKVKKNHAMATIVEAKALYAVYP SSPKLIDSSTIIKVVGSVKFEKKESAQEKDVFIMP EKHQAVPSYDTIIRFLIPAMDTFKLYGRPEKLLSSK N	2,09	No	-	Bud, bud neck, cytoplasm	-
Caf120CCt	-	not detected (BLAST: 9E-28)	-	PISEWVRVRFGAGVPWKRCYCVIEPPSEKEYQK AQKEWKKKNPYDRSHGPILKGQIKFFESKKDAE KKKKHQRPIASITDAYSAYAIYPQAKALIDGSTLLKI EGDITIHSEPPSTTEGFVFIMPETHPAVPGFEMLL RFLFPTWDTFGLY	4,93	Yes	2a	n/a	-
Caf120_tandem	Caf120p	3,65E-07	-	SPVSPELVPILTLLNAHTHRRYHEGVFLILQDLNN NGTHAARKWKDVYGVLLGTQLALWDAKELAEFT DPSCPVSEKKLKEVASKPTYINLTDATLDNSD NIVMECGKNLTNALVVSTTLKNRYFLQFGNKESF NAWNSAIRLCLYECSSLQEAYTGAFISSRGAKLG DIRILLTNRKYDYKDWVSVRFGAGMPWKRCYAVI SQSSSKKKGHFGEINLYENDKKVKKNHAMATIVE AKALYAVYPSSPKLIDSSTIIKVVGSVKFEKKESAQ EKDVFIMPEKHQAVPSYDTIIRFLIPAMDTFKLYGR PEKLLSSKN	0,93	No	-	Bud, bud neck, cytoplasm	-
Caf120Ct_tandem	-	3,23E-08	-	LQPIFSLLNGHANKLYQEGYFLKLDDQDIRGNLN PDRTWTECFAQLVGTVLSLWDAAELDAAGEDGE VLPKFINLTDASIKMIESLPTKSNDEQPLQNILSIST AGRNRYLLHFNSHHSLIQWTSGIRLAMYEHATLQ EAYTGSLVAGKGKTLNSINIIMERARQPISEWVRV	1,37	Yes	2a	n/a	-
Cdc24	Cdc24p	3,49E-04	15023338	ISKFGELLYFDKVFISTTNSSSEPEREFEVYLFEKII ILFSEVVTKKSASSLILKKKSSTSASISASNITDNN GSPHHSYHKRHSNSSSSNNIHLSSSSAAAIIHSS TNSSDNNSNNSSSSSLFKLSANEPKLDLRGRIMI MNLNQIIPQNNRSLNITWESIKEQGNFLLKFKNEE TRDNWSSCLQQLIHDLKNEQFKARHHSST	1.82‡	Yes	2b	Nucleus, cytoplasm	Cdc42 network
Cdc24Ct	-	7,48E-04	-	QFGKLLLHGVYTVITGKGDQEKDYEIYLFECILLC CKEVVPGKTKDKRDKARPAQPKVRNKNAKLQLK GRIFMTNVTDVVAMSKPGSYMVQIWWKGDPGV ENFMIKFQNEETMKKWAAGLEQQRKLNAPQP	7,70	No	-	n/a	-
Cla4	Cla4p	5,75E-14	15023338	TSTSKKKSGWVVSYKDDGILSFIWQKRYLMLHDS YVALYKNDKQNDDAILKIPLTSIISVSRTQLKQYCF ELVRCSDRNSVSSGSSSSSLNVSSDSNSKKSIYIA TKTESDLHSWLDAIFAKCPLLSGVSSPTNFTHK	11.16‡	Yes	2b	Bud, cytoplasm	Cdc42 network
Cla4Ct	-	2,48E-09	-	GGLAIIKQGYVGLPSKNPFQTWKSRFMILRKDVL DFHKNENGKHLYKILLADVSVDRVPDPNGDPVF EIRRHSTNQYEAPPGEDQGNGVKATKICVKTDD ELYEWIDCIYARCPGGVSNPVNFSHPV	16,80	Yes	1	n/a	-
Dcp1	Dcp1p	not detected	21119626	SNDSLTYNCGKTLSGKDIYNYGLIILNRINPDNFS MGIVPNSVVNKRKVFNAEEDTLNPLECMGVEVK DELVIIKNLKHEVYGIWIHTVSDRQNIYEQLIKYLLEN EPKDSFA	1,87	No	-	Punctuate composite	-
Exo84	Exo84p	947 (NCBI CDS: 9.62E-08)	21119626	STKGRHILMNSANWMELNTTTGKPLQMVQIFILN DLVLIADKSRDKQNDFIVSQCYPLKDVTVTQEEF STKRLLFKFSNSNSSLYECRDADECSRLLDVIRK AKDDLCDIFHVEEENSKRIRESF	4.74‡	Yes	2a	Bud, bud neck, cell periphery	Other
Exo84Ct	-	1380 (NCBI CDS: 1.98E-06)	-	NSPGRHVVQNAGLWVELDNATYRSRRAMQIFLL NDHILIASRKKRKVDAPNDSRAPMTKLVADRCWH LLDVEIVDMAGPSDSSSGRNKLADAIMIRGGGNE SFIYRTEKPGDPEKASLLLNIRKQVEELRRTLQSE REATNKAKETINYFASR	2.12‡	Yes	2a	n/a	-
Ira1	Ira1p	not detected	16397625	LEVLKDVRVTLHDITLYDKEKQCPCPVSLKIGNKY FQVLHEIPQLYKVTVSNRTFSIKFNNVYKISNLISV	1,83	No	-	Cytoplasm	Other

Ira1_CRAL-TRIO-PH		not detected	16397625	PFVVENREKYPSLYEFMSRYAFKKVDMKEEEED NAPFVHEAMTLDGIQIIVVTFTNCEYNNFVMDSLV YKVLQIYARMWCSKHYVVIDCTTFYGGKANFQKL TTLFFSLIPEQASSNCMGCYYFNVNKSFMDQWA SSVTVENPYLVTTIPRCFINSNTDQSLIKSLGLSG	2,07	Yes	2a	Cytoplasm	-
Ira2_1	Ira2p	not detected	16397625	KVLQDIRVSLHDTLYDEKRNRFTPVSLKIGDIYFQ VLHETPRQYKIRDMGTLFDVKFNDVYEISRIFEVH	1,43	No	-	Cytoplasm	-
Ira2_2		not detected	21119626	VRRNVLNEAKHDDGDIDGDDFYKTTFLLIDDVLG QLGQPKMEFSNEIPIYIREHMDDYPELYEFMNRH AFRNIETSTAYSPSVHESTSSEGIPIITLTMSNFSD RHVDI	8,87	Yes	2a	Cytoplasm	Other
Ira2Ct	-	not detected	-	VVTDTRYVFQPVTRLSRTRGKIEVVIKVGSQYVQ VTTCKKQEVPGFHLSTTINDIFRLGEIEEAPTMIQ TEDDSAFCLRADNGKIVMYFTSPKKADVLQAIRL AKAKYSKDMR	1,72	No	-	n/a	-
Ira2_CRAL-TRIO-PH	Ira2p	not detected	16397625	IYIREHMDDYPELYEFMNRHAFRNIETSTAYSPSV HESTSSEGIPIITLTMSNFSDRHVDIDTVAYKFLQI YARIWTTKHCLIIDCTEFDEGGLDMRKFISLVMGL LPEVAPKNCIGCYYFNVNETFMDNYGKCLDKDN VYVSSKIPHYFINSNSDEGLMKSVGITGQGLKVL	2,66	No	-	Cytoplasm	-
Ira2Ct_CRAL-TRIO-PH	-	not detected	-	HISANTPPAYSRFQDFMIRNAFRNTESYLTARAIY DGGESKDGLSIICILRHIDAESIDYDTLIFCYLKIAS RLWHRPFGLLIDVTGYNGQAEPKDELFQKLEMLT PTELARQLSRIYIYNMNSVSKKCLRRLLRASTRN EASVFSPTNVDYHLIGSLQDLQTHFHLSQLHLPK	1,81	No	-	n/a	-
Las17	Las17p	not detected	21119626	PKASNKIIDVTVARLYIAYPDKNEWQYTGLSGALA LVDDLVGNTFFLKLVDINGHRGVIWDQELYVNFE	1,51	No	-	Punctuate composite	-
Mdr1	Mdr1p	not detected	21119626	RKTAYVYSGRLFLTPHFLVFRDAFDHSSCVLILNI STIKRVERSPSESYEFALLVTLYTGAKVLIQFIGIR YRSEQFCDKLKLNLKENIPNA	0,60	No	-	Cytoplasm	-
Mdr1Ct	-	not detected	-	NRGFGYAGRLHLSESYLCFSTTPSSFLQSASSSS SLLFTGQTHGAGPSGNGFTFPLCAIRRVERLHS QNFQFALAITTWNGISQDAAKDKDRKDLREQRITI QLAGSRQACERFCDGLKKGLRANVVHV	0,33	No	-	n/a	-
Num1	Num1p	2,15E-04	15023338	NEPSIIPALTQTVIGEYLFKYYPRLGPFGFESRHE RFFWVHPYTLTLYWSASNPILENPANTKTKGVAIL GVESVTDFNPYPTGLYHKSIVVTTETRTIKFTCPT RQRHNIWYNSLRYLLQRNMQGISLE	2.19‡	Yes	-	Punctuate composite	Other
Num1Ct	-	2,69E-05	-	GSGTDPRMIQAITQTMIGEYLWKYTRKKTGRGE MSENRHRRYFWVHPYTRTLYWSTRDPSTAGRS ELRAKSVQIEAVRVVNDDNPFPPGLHQKSLIIVSP QRSIKFTCTTGQRHETWFNALSYLLRTANEG	4.30‡	Yes	1	n/a	-
Nup2	Nup2p	not detected	21119626	EEDEVALFSQKAKLMTFNAETKSYDSRGVGEMK LLKKKDDPSKVRLLCRSDGMGNVLLNATVVDSF	2.67‡*	Yes	2b	Nuclear periphery	Nucleus
Nvj2	Nvj2p	1,63E-04	15023338	NKDSDTALQEQILQRTDLKKKQRFFAVLRHGNLF LYKDDSQNANLVHAISLQNRFITIWPRFDELGKEE LPDASLFTKRTCIAIFKNDLVSIDSKNHNVILPHFD PLTSAESNNGDISTNDTTHEYQSQFHSSNQFFLY FDNNMDKEDWYYQLINASKSNSNSLST	1,36	Yes	2a	ER	Other
Opy1N	Opy1p	6,08E-04	15023338	MIAGATAPSSQHEILIASNLIKKPSTSQNKTPTAQS SSGNNGAADGAPQGYHHHHHHHHRHLWWPRTT DHQYWCVLRKNQFAYYKTRDEREAISVIPRFDIL NFKISELDGILTVYTPSKDLIFKFPRGQNEKVGME LMHNWKIALEKFLSSPSGN	12,23	Yes	2b	Cytoplasm	Other
Opy1C		6,21E-10	15023338, 22562153	DFRNAEHQVCSGILYTKVKKKKLFNRAKWQKFN VELTNTSFNLYSFKTGKLKKSIKLDKIIDCIELDNN	11.55‡	Yes	1	Cytoplasm	Other
Osh2	Osh2p	3,25E-13	15023338	ASSKPPTYKGFLKKWTNFAHGYKLRWFILSGDG NLSYYKDQSHVDRPRGTLKVSTCRLHIDSSEKLN	5.14‡	Yes	1	Punctuate composite	Cdc42 network
Osh3	Osh3p	1,21E-12	15023338	QDMLFRVGQGRYLQGYLLKKRRKRLQGFKKRF FTLDFRYGTLSYYLNDHNQTCRGEIVISLSSVSA NKKDKIIIIDSGMEVWVLKATTKENWQSWVDALQ TCFDDQFEIDK	7,25	Yes	2b	n/a	Cdc42 network
Osh3Ct	-	2,16E-09	-	NSTTGIVPTTYHVGVLLKRRRKKGQGYARRFFSL DYITCTLSYYHSRNSSALRGAIPLSLAAIAADERR	17.35‡*	Yes	1	n/a	-
Pkh2	Pkh2p	8,30E+02	21119626	SVFSGKIKKLFHPTSAAETLSSSDEKTKYYKRTIV MTSFGRFLVFAKRRQPNPVTNLKYELEYDINLRQ	6.42‡	Yes	2a	Cytoplasm	Other
Plc1	Plc1p	not detected (NCBI CDS: 4.69E-33)	-	RRRRKFYEFKLINNNGQIIWKDGSKYLELDSVKDI RIGDTASTYQEEVDPKRLRSDSKLWIAIIYKVSNK LKALHVVALNELDFNTFLSCICGLVKLRRELMESI	0,94	Yes	2b	n/a	Other
Plc1Ct	-	2.67 (NCBI CDS: 1.56E-45)	-	VLRQGMPMEKLRIKRGNLKPITLYIDPDAAKIWW DRPHRSPRFIYIDDIAEIRAEDTRQYRLDANLPDS YESRWFTLKLRLPDKQGDKFVHLIAPTEAACEQ	5.04‡	Yes	2b	n/a	-
Psy2	Psy2p	not detected	21119626	VNTEPKRVKVYILENNEWKDTGTGFCIGEVDEG KFAYLVVSDEDSPTETLLKSKLEGNIEYQRQEETL IVWKDLGGKDIALSFEESMGCDTLCEFIVHVQRNI E	5.86‡	Yes	2a	Nucleus	Nucleus
Rgc1	Rgc1p	6,02E-08	15023338	RRLSDIVYPNMKSPLAKCIRVGYLLKKTESSKSFT KGYFVLTTNYLHEFKSSDFFLDSKSPRSKNKPVV EQSDISRVNKDGTNAGSHPSSKGTQDPKLTKRR KGLSSSNLYPISSLSLNDCSLKDSTDSTFVLQGY ASYHSPEDTCTKESSTTSDLACPTKTLASNKGKH QRTPSALSMVSVPKFLKSSSVPKEQKKAKEEANI NKKSICEKRVEWTFKIFSASLEPTPEESKNFKKW VQDIKALTSFNSTQER	8,05	No	-	Cytoplasm	Other
Rom1	Rom1p	66.1 (BLAST: 7E-9)	-	NNEKRKIKHEGLLSRKDVNKTDASFSGDIQFYLL DNMLLFLKSKAVNKWHQHTVFQRPIPLPLLFICPA EDMPPIKRYVTENPNCSAGVLLPQYQTSNPKNAI VFAYYGTKQQYQVTLYAPQPAGLQTLIEKVKQEQ KRLLDETKHITFKGMVQQFF	1,27	No	-	n/a	-
Rom1Ct_A	-	4.12 (BLAST: 1E-95)	-	TRLARYPLLLENVLKYTEPDNPDKEDIPKVLQMIR DLLSRVNAESGKAENRFNLRRLHEQLRFRSPAE RVDLKLTEEGRELVFKTQFKKSPTENPDITAFLFD HAVLLVRIKQVGKTEEYKAYRRPIPLELLSIKEMEE IFPQNGAVKRSSSSLLPALRNTSTEAKKGEGWPI TFRHLGKNGYELTLYATNQSGRQKWLEHIDQAQ QRLRARADFLNRTVISH	4,33	Yes	2b	n/a	-
Rom1Ct_B	-	2,48E-09	-	LDHLGRELLKQGELQRQGSKGVRWVDTHALLFD HYFILAKAFMAKDGSGKKVDVSKEPIPMPLLFLE STNDEPVAKQKSLTAPLTRTAVTGAGSQLMSKSA TMSEKPGADDDGKIIYPFRIKHLGHEIYTLYASSA QERTAWCNAIIEAKTRHARALHAQNAEPF	8,70	No	-	n/a	-
Rom2	Rom2p	not detected (NCBI CDS: 6.55E-48)	12015967	GTVVGDIQFYLLDNMLLFLKAKAVNKWHQHKVF QRPIPLPLLFACPGEDMPALRKYIGDHPDCSGTVI QPEYNTSNPKNAITFLYYGAKQRYQVTLYAAQYA GLQTLLEKIKQGQAAIISKTEMFNV	2,04	No	-	Bud, bud neck, cytoplasm	-
Rtt106	Rtt106p	not detected (NCBI CDS: PH1 5.06E-80; PH2 3.48E-43)	22307274	SETNTIFKLEGVSVLSPLRKKLDLVFYLSNVDGSP VITLLKGNDRELSIYQLNKNIKMASFLPVPEKPNLI YLFMTYTSCEDNKFSEPVVMTLNKENTLNQFKN LGLLDSNVTDFEKCVEYIRKQAILTGFKISNPFVN	3.89‡	Yes	2b	Nucleus	Nucleus
Rtt106Ct	-	not detected (NCBI CDS: PH2 7.45E-31)	-	VSEPVLLEIKDISVVVPQRKKYDICFTQSCLFAKA QGATGPAPGLVYPWIEIEHAFFLPVPDKSGTQYN YILLPRDSYLPTSKASGVSSDRQASLEPLVFTVPA AAPKAGSIKGTAAKTAEAFAESYSCLFHWALTTSL RAAGNESFQLISSDPRVFHSLVKQPHRPNEKAVH VKAFRGSKEGFLFFLPTGILWGFKKPLVFLPLDKI AAVSYTNVLRTTFNLVIELDSDVTAGGNATDKEIE FGMIDQEDYQVINENYVQRHGLAD	2.64‡	Yes	2a	n/a	-
Sec3	Sec3p	194 (NCBI CDS: 8.31E-83)	20062059	HKRNVSRASNSSQTSNFLAEQYERDRKAIINCCF SRPDHKTGEPPNNYITHVRIIEDSKFPSSRPPPD SKLENKKKRLLILSAKPNNAKLIQIHKARENSDGS	3.29‡	Yes	1	Bud, bud neck, cell periphery	Cdc42 network

Sec3Ct	-	610 (NCBI CDS: 2.65E-67)	-	DGSVPETYITHIRITEWQNYPSSPPPPSARAPQY EKPRVIIVAVRKSGRLRVHKSKENANGTFSIGKT WWLDDLQSIESFTSPSANPNLREWARDVGFIVTL GKPYYWEAHSDKEKKFFIASLIKIFNRYTGGRTP	0,56	No	-	n/a	-
Sip3	Sip3p	1,94E-11	15023338	KSPEKSGWLYMKTQVGKPTREIWVRRWCFLKN AVFGMFLLSPSKTYVEETDKFGVFLTNVRYDPEE DRKFCFEVKIFGNKVTEAHDNMSKDITLVFQTSN YLDLKSWLIAFEATKKYVMSIQHDSLE	2,14	Yes	2a	n/a	Nucleus#
Skg3N	Skg3p	6,17E-05	15023338	GDISPELVPIVTLLSAQAHRRYHYGIFLILHDLKTD GTPAARQWEECYGVLLGTQLALWDAKELSDSKN NKNTSTMKKAASRPSFINFTDASVRSLDANDQVII	7.36‡	Yes	1	Bud, bud neck, cell periphery	Other
Skg3C		29.3 (BLAST: 1E-99)	15023338	GDIKVVMADTKFTYEDWVSVRFGTGMPWKRCY AVISPQSGKKKKNSKGSICFYENNKKTKKSNIMT TVVDARALYAVYPSSPILIDTSTIIKLEGFVSFDKSE EPQETNLFIMPEKHQGVPGYDTIIRFLIPAMNAFY LYGRPKGLIANRTDPDSLLFALPTLPHIYYLQVDD VLSLTKDKNYIHWSAADWRNNIVQVLQKKLSKGY KG	0,97	No	-	Bud, bud neck, cell periphery	-
Skm1	Skm1p	9,33E-14	15023338	MKGVKKEGWISYKVDGLFSFLWQKRYLVLNDSY LAFYKSDKCNEEPVLSVPLTSITNVSRIQLKQNCF EILRATDQKENISPINSYFYESNSKRSIFISTRTER DLHGWLDAIFAKCPLLSGVSSPTN	2,86	Yes	2b	n/a	Cdc42 network
Slm1	Slm1p	1,23E-13	15023338, 21119626	EIKSGFLERRSKFLKSYSKGYYVLTPNFLHEFKTA DRKKDLVPVMSLALSECTVTEHSRKNSTSSPNS	12,15	No	-	Punctuate composite	Other
Slm1Ct	-	3,91E-07	-	HYACQEIRAGLLERKSKYLKSYTPGWYVLSPTHL HEFKSADKTQAPVMSLYLPEQKLGSHSEEGSSS NKFILKGRQTGHMHRGHTWVFRAESHDTMMAW YEDIKALTERTPEER	16.42‡	Yes	2b	n/a	-
Slm2	Slm2p	8,71E-14	15023338	NQNDPFTFEVKSGFLEKRSKFLKSYSRGFYVLTP SFLHEFKTPDKHKFSTPLMSIPLVECTVTEHSKKT KSNSEQGKNKFILRTNSNGLIHRGHNWVFKVDS YDDMIEWFGNIKALSSLPNYDDK	5.82‡	Yes	2a	n/a	Other
Spo14	Spo14p	1,97E-03	15023338	GFGKGQGYLVIRSTAKAQGIWRVSHFGKHAFKD MIDRHTTKWFLVRNSYLTYVSDLSSTTPLDVFLID WKFKVRFSGNKNNILDNENEINWIIHDPNLEINDE	1,11	No	-	Cytoplasm	-
Spo14Ct	-	13.2 (BLAST: 6E-32)	-	SYHGKECYLHIQSSKGLDFRRVLTPGKVIARHTR KWFLVRQSYIVCVESPENMNIYDVYLVDSKFQIV SKKSKKQAAEQDDDDDDGDDNIENLDLSTRPQQ KHQKHHTLKLITSERKVKLFSPNQHLIAQFEESIR EMLKNTPW	0,40	No	-	n/a	-
Spo71N	Spo71p	402 (BLAST: 9E-17)	-	KKRTGQILKKEKMLVMVKEAIQNKVPLPNFSENE CFDTRVSERWKEYIVIARSTGRFDPPILLQFYRH RHIPEIEDISSIATKYHRNPLDFFLSRNCIVKFYSS LDKTISIQKPDKRLGGFIDESIEIKKDELKHYSPIKIF ILRCSSIRSSGRWYKFLLESLDRQ	3,53	Yes	2a	n/a	Other
Spo71M		4.92 (BLAST: 1E-115)	15023338	PIANRLRETKSLEGKCLKEPTPIEGFLIRLTDKYG SARTNFGKYSISTAYFFTCENLLFSMKAYRANPPL PIDSMIDDTSTEIEKEEIWKQWKKIPEVYEQQPYP LDTNDHIEWMNCQTTQSEYDSRDFYAFHCFHRR	5,67	Yes	-	n/a	Other
Spo71C		1,36E-06	15023338	NGISLSRPLIQKGPLYQKPHKHSVFSKYYVVLISG FIVLFHCFHRSTTGFAKEVLEYAHYVTIPDDCYLY SGTTTELDLLQRDRTFDEINYGSHALPRVYGDG WRSVEDESSRCFTLWFGTRRALSSNRLQKKGN EKQYTQDYGRQDNNIDPPSAPEADLNNSNVPSN TDKIHFTKKLGVSGKSMVFMARSRQERDLWVMS IYYELERLRRTASTSNSR	0,89	No	-	n/a	-
Ste5	Ste5p	1,47E+03	16847350	ETKTTLPLLRSYFIQILLNNFQEELQDWRIDGDYG LLRLVDKLMISKDGQRYIQCWCFLFEDAFVIAEVD NDVDVLEIRLKNLEVFTPIANLRMTTLEASVLKCT LNKQHCADLSDLYIVQNINSDESTTVQKWISGILN QDFVFNED	1,56	Yes	-	Nucleus, cytoplasm	-
Swh1	Swh1p	4,29E-13	15023338	LHEAPTYKGYLKKWTNFAQGYKLRWFILSSDGK LSYYIDQADTKNACRGSLNMSSCSLHLDSSEKLK FEIIGGNNGVIRWHLKGNHPIETNRWVWAIQGAI RYAKDREILL	1.62‡	Yes	1	n/a	Cdc42 network
Swh1Ct	-	5,66E-13	-	GREAREMRGYLKKWTNYRKGYQLRWFVLEDGV LSYYKHQDDAGSACRGAINMRIAKLHMSPDEKT KFEIIGKSSVKYTLKANHEVEAKRWFWALNNSIQ WCKDQAREEERQRQRNAELLRQAKAEAQALSE AA	16.61‡*	Yes	1	n/a	-
Syt1	Syt1p	1,72E-03	15023338	NSDYKYCKILQMGAIMNLGMPSRKFSIVNSAKIH WKKEFAILTSLGLLICDKMDWINPQMMKDPKSGT TNYIIDFKSGFSFVPGSTIDVYNGLFADRERDSLG KSHFASLVLAYTEHHSTGSHTSNTTAASSSAKHN	1,85	No	-	n/a	-
Syt1Ct_A	-	1,87E-04	-	AAHPGIVDIKITKVGLLWRKDTKKKKTRSPWQEW GAILTGAQLYFFRNTAWVKHLMHQQKDHIKKGH DGDPCIFKPPLDQFKPDALMSTDGAVALMDSSYT KHKHAFLYVRHGGLEEVLLAEDEDDMNDWLAKL NYAAAFRTSGVRM	2.31‡	Yes	-	n/a	-
Syt1Ct_B	-	170 (BLAST: 3E-70)	-	KDRNWTEVFAVIQKGQLSLFSFSPNKSLRNKGR RGPGHNNSLPKGAVVGGGNWQDNATNLGTFSL RQTLASALPPPGYSRTRPYVWALSLPTGAVHLF QVGTPEICKEFVNTVNYWSARLSTHP	1,33	Yes	-	n/a	-
Tfb1	Tfb1p	not detected (NCBI CDS: 1.08E-14)	15909982	MSHSGAAIFEKVSGIIAINEDVSPAELTWRSTDGD KVHTVVLSTIDKLQATPASSEKMMLRLIGKVDESK	2,34	No	-	Nucleus	-
Tfb1Ct	-	not detected (NCBI CDS: 5.94E-22)	-	MSIPRSQTTYKKKEGILTLTEDRKFLIWTPLPATG PPTVSLALDNITNLQQTPPQSAKVILKFTERPRPN	1,96	Yes	2a	n/a	-
Tus1	Tus1p	2,32E-09	15023338	LTYPERKLVLSGTVYKKRDLWLDPTPVYIALLDNC LLITEEISKGETQKYKLIERPIPIDYLSLEKRKIPGT SKQPLRNYSQKEHKSPMHNFSTPINSMRPLLKS SGNHMSTAYGDRKTSNTEISNANPNTDEFSFKIR NTATGESFKFFTESAEVLNQWIDAIMESFKRNAE NHDLN	4,33	No	-	n/a	-
Yel1	Yel1p	35.7 (BLAST: 1E-78)	15023338	LNKDSEWERVKIQVKEGRIFIFKIKPDVKDIIQSSE TDSATIDYFKDISSSYFAYSLLEAEAHVVQDNIIIGS GAMKSNVCNKNTKRKSGNFTVSFPENINGPKLV LEFQTRSVEEAHKFMDCINFWAGRISPVPLTQ	0,95	No	-	Bud neck, cytoplasm	-
Yhr131c	Yhr131cp	43.5 (BLAST: 2E-62)	15023338	NRSWRNFIIEINSTQLNFYHIDESLTKHIRNYSSG ETKSEKEDRIHSDLVHRSDQSQHLHHRLFTLPTR SASEFKKADQERISYRVKRDRSRYLTDEALYKSF TLQNARFGIPTDYTKKSFVLRMSCESEQFLLRFS HIDDMIDWSMYLSIGISV	0,14	No	-	Cytoplasm	-
Ynl144c	Ynl144cp	not detected (BLAST: 3E-24)	15023338	SSRLWNNFILQINSTQINFYSIDDSLTRHIKNYRG GDMFDHSHHSKTASDRHHSARSLLNAFTTKSTY QFDKYDKERICGEIARDEHKFLSDERLFKSYSLQ CAKVGLPIDYSSRDFVLRMRCEGQQFLVQFSHV DELIYWAMYLNMGISLSLDLELR	0,31	No	-	n/a	-
Ynl144cCt	-	7,15E-01	-	STQLEGVFNRKMEIEETTKRAEYRDWRMVYV ELRGTMLNVYSVKKERGWWASKHDGPDISPDN PPWVKKGSLERAYSLLYADAGIAADYRKRRYVIR MRLETDQFLLSCVELSTFVKWLDGIFAAINISAPID ERD	1.59‡	Yes	2a	n/a	-
Yrb2	Yrb2p	not detected	21119626	GEESEECIYQVNAKLYQLSNIKEGWKERGVGIIKI NKSKDDVEKTRIVMRSRGILKVILNIQLVKGFTVQ	2.27‡	Yes	2a	Nucleus	Nucleus
Ysp1	Ysp1p	9,05E-12	15023338	NSEEKGLSGWLYMKTTVGHDPKRVVWVRRWCF LQNNVFGVFSLSPSKTYVEETDKFGILWITVEYLP	13.64‡	Yes	2b	Mitochondrion	Other
Ysp1Ct	-	7,53E-09	-	GAAILEKQGWLFLRVMAGKPVRTTWIRRWYYCR DGVFGWLAQGPNGVLMGDEIGVLLCSARPAVQE ERRFCFEIKTKTQTILLQAETQAQLVEWLEVFETA KNRYIEASMKRHQKSVTLPVG	17,92	Yes	1	n/a	-
AKT1	AKT1	8,50E-17	12176338	MSDVAIVKEGWLHKRGEYIKTWRPRYFLLKNDG TFIGYKERPQDVDQREAPLNNFSVAQCQLMKTE	5.05‡	Yes	1	n/a	-

Dynamin1	Dynamin-1	2,70E-10	7954789	KTSGNQDEILVIRKGWLTINNIGIMKGGSKEYWFV LTAENLSWYKDDEEKEKKYMLSVDNLKLRDVEK GFMSSKHIFALFNTEQRNVYKDYRQLELACETQE EVDSWKASFLRAGVYPERVGDK	13.07‡	Yes	2b	n/a	-
PDK1	PDK1	9.52 (NCBI CDS: 5.99E-67)	15457207	GSNIEQYIHDLDSNSFELDLQFSEDEKRLLLEKQ AGGNPWHQFVENNLILKMGPVDKRKGLFARRR QLLLTEGPHLYYVDPVNKVLKGEIPWSQELRPEA KNFKTFFVHTPNRTYYLMDPSGNAHKWCRKIQE VWRQRYQSHPDAAVQ	9.11‡	Yes	1	n/a	-
Plcδ1	Plcδ1	8,32E-11	8521504	HGLQDDPDLQALLKGSQLLKVKSSSWRRERFYK LQEDCKTIWQESRKVMRSPESQLFSIEDIQEVRM GHRTEGLEKFARDIPEDRCFSIVFKDQRNTLDLIA PSPADAQHWVQGLRKIIHHSGSMDQRQK	16.40‡	Yes	1	n/a	-
Pleckstrin	Pleckstrin	4,28E-22	15698571	DVILKEEFRGVIIKQGCLLKQGHRRKNWKVRKFIL REDPAYLHYYDPAGAEDPLGAIHLRGCVVTSVES	31.32‡	Yes	2b	n/a	-
SOS1	SOS1	5,24E-17	9374522	QQMKGKQLAIKKMNEIQKNIDGWEGKDIGQCCN EFIMEGTLTRVGAKHERHIFLFDGLMICCKSNHG QPRLPGASNAEYRLKEKFFMRKVQINDKDDTNE YKHAFEIILKDENSVIFSAKSAEEKNNWMAALISL QYRSTLE	8.55‡*	Yes	1	n/a	-
EEA1	EAA1	-	9702203	NQDERRALLERCLKGEGEIEKLQTKVLELQRKLD NTTAAVQELGRENQSLQIKHTQALNRKWAEDNE VQNCMACGKGFSVTVRRHHCRQCGNIFCAECS AKNALTPSSKKPVRVCDACFNDLQG	15.07‡	Yes	ND	n/a	-
p40phox	p40phox	-	11684018	AVAQQLRAESDFEQLPDDVAISANIADIEEKRGFT SHFVFVIEVKTKGGSKYLIYRRYRQFHALQSKLE ERFGPDSKSSALACTLPTLPAKVYVGVKQEIAEM RIPALNAYMKSLLSLPVWVLMDEDVRIFFYQSPY DSEQVPQALRR	11,19	Yes	ND	n/a	-
Lactadherin	Lactadherin	-	18160406	CTEPLGLKDNTIPNKQITASSYYKTWGLSAFSWF PYYARLDNQGKFNAWTAQTNSASEWLQIDLGSQ KRVTGIITQGARDFGHIQYVAAYRVAYGDDGVTW TEYKDPGASESKIFPGNMDNNSHKKNIFETPFQA	1.84‡*	Yes	ND	n/a	-
Hsv2	Hsv2p	-	22704557	-	1,52	Yes	ND	n/a	-

Table V.5: Table summarizing the TIs computed for each interaction of the screen. The columns represents the liposomes of the assays and the rows the proteins used in the assay. A green cell indicates a $0 < TI < 2$ (confident binder). An orange cell indicates a $TI > 2$ (non confident binder). A blue cell indicate a $-2 < TI < 0$ (confident non binder). A violet cell indicates a $-2 < TI$ (non confident non binder). A white cell represents an interaction for which not enough replicates were available to compute a TI



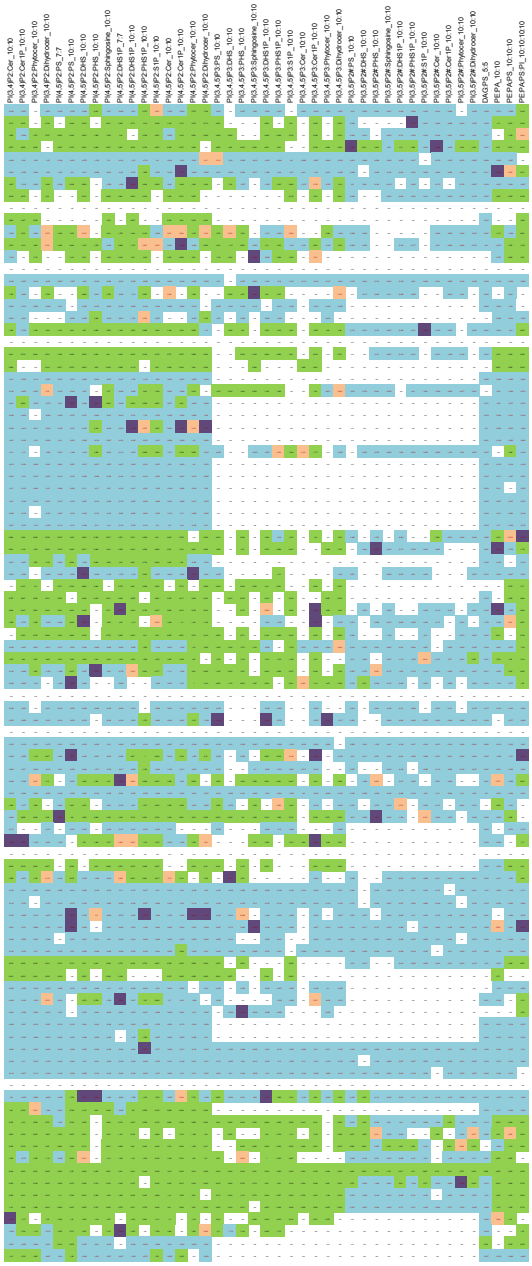

PI(3,4)P2Cer_10:10
PI(3,4)P2Cer1P_10:10
PI(3,4)P2Phytocer_10:10
PI(3,4)P2Dihydrocer_10:10
PI(4,5)P2PS_7:7
PI(4,5)P2PS_10:10
PI(4,5)P2DHS_10:10
PI(4,5)P2Phytocer_10:10
PI(4,5)P2Sphingosine_7:7
PI(4,5)P2DHS1P_10:10
PI(4,5)P2PHS1P_10:10
PI(4,5)P2S1P_10:10
PI(4,5)P2Cer_10:10
PI(4,5)P2Cer1P_10:10
PI(4,5)P2Phytocer_10:10
PI(4,5)P2Dihydrocer_10:10
PI(3,4,5)P3PS_10:10
PI(3,4,5)P3PHS_10:10
PI(3,4,5)P3Sphingosine_10:10
PI(3,4,5)P3DHS1P_10:10
PI(3,4,5)P3PHS1P_10:10
PI(3,4,5)P3S1P_10:10
PI(3,4,5)P3Cer_10:10
PI(3,4,5)P3Cer1P_10:10
PI(3,4,5)P3Phytocer_10:10
PI(3,4,5)P3Dihydrocer_10:10
PI(3,5)P2PS_10:10
PI(3,5)P2PHS_10:10
PI(3,5)P2DHS_10:10
PI(3,5)P2Sphingosine_10:10
PI(3,5)P2DHS1P_10:10
PI(3,5)P2PHS1P_10:10
PI(3,5)P2S1P_10:10
PI(3,5)P2Cer_10:10
PI(3,5)P2Cer1P_10:10
PI(3,5)P2Phytocer_10:10
PI(3,5)P2Dihydrocer_10:10
DAGPS_5:5
PEPA_10:10
PEPAPS_10:10:10
PEPAPS1PL_10:10:10:10

Table V.6: Table summarizing the 45 PH domains that were previously tested for their lipid-binding properties. The first column gives the name of the PH domain as used in this study. The second and third columns indicate if the interaction with lipid was detected in previous studies (referred to with PMID) and in this study with high confidence ($0 < TI < 2$), respectively. The last column summarizes the match of our data with the literature-based benchmark dataset.

Domains previously analysed	Previously known interaction (reference PMID)	Interaction with LiMA (0 < TI < 2)	False positive (fp) True positives (tp) False negative (fn) True negative (tn)
Ask10	yes (15023338)	yes	tp
Atg26	yes (15023338)	yes	tp
Bem3	no (15023338)	yes	fp
Boi2	yes (15023338)	yes	tp
Bud4	yes (15023338)	yes	tp
Caf120N	yes (15023338)	yes	tp
Cdc24	yes (15023338)	yes	tp
Cla4	yes (15023338)	yes	tp
Exo84	no (21119626)	yes	fp
Ira2_2	yes (21119626)	yes	tp
Num1	yes (15023338)	yes	tp
Nup2	yes (21119626)	yes	tp
Nvj2	no (15023338)	yes	fp
Opy1N	yes (22562153)	yes	tp
Opy1C	yes (22562153, 15023338)	yes	tp
Osh2	yes (15023338)	yes	tp
Osh3	yes (15023338)	yes	tp
Pkh2	yes (21119626)	yes	tp
Psy2	yes (21119626)	yes	tp
Sip3	yes (15023338)	yes	tp
Skg3N	yes (15023338)	yes	tp
Skm1	yes (15023338)	yes	tp
Slm2	yes (15023338)	yes	tp
Spo71M	yes (15023338)	yes	tp
Ste5	yes (21119626)	yes	tp
Swh1	yes (15023338)	yes	tp
Yrb2	yes (21119626)	yes	tp
Ysp1	no (15023338)	yes	fp
Bem2_1	yes (15023338)	no	fn
Boi1	yes (15023338)	no	fn
Dcp1	no (21119626)	no	tn
Las17	yes (21119626)	no	fn
Mdr1	yes (21119626)	no	fn
Rgc1	yes (15023338)	no	fn
Rom2	yes (12015967)	no	fn
Skg3C	yes (15023338)	no	fn
Slm1	yes (21119626)	no	fn
Spo14	yes (15023338)	no	fn
Spo71C	yes (15023338)	no	fn
Syt1	yes (15023338)	no	fn
Tfb1	yes (15909982)	no	fn
Tus1	no (15023338)	no	tn
Yel1	yes (15023338)	no	fn
Yhr131c	yes (15023338)	no	fn
Ynl144c	no (15023338)	no	tn

Table V.7: Table summarizing current knowledge on PH domains preferred ligands. The first column indicate the name of the PH domain as used in this study. The second column indicate whether the PH domain was used in the context of its full protein or alone in the litterature. The third column indicates the lipid ligand that was found to be recruited by the PH domain in the litterature. The fourth column indicates the PMID of the publication in which this interaction was found. Cells marked in green indicate interactions that we reproduced and the red ones, interactions that we did not reproduce.

LBD name	LBD/full length protein	Lipid ligand	Reference (PMID)
AKT1	PH	$PI(3,4,5)P_3$	9079675 / 18954143
		$PI(3,4)P_2$	9079675
		$PI(4,5)P_2$	9079675 / 18954143
		PS	21402788
Dynamin1	PH	$PI(3,4)P_2$	9765310
		$PI(4,5)P_2$	9765310
		$PI(3,4,5)P_3$	9765310
PDK1	PH	$PI(3,4)P_2$	21971045 / 9895304
		$PI(3,4,5)P_3$	21971045 / 9895304
		PS	21971045
		$PI(4,5)P_2$	9895304
		PI(3)P	9895304
Plcδ1	PH	$PI(4,5)P_2$	18954143
		$PI(3,4,5)P_3$	18954143
Pleckstrin	PH (C-term)	$PI(3,4)P_2$	15698571
SOS1	PH	PA	17486115
		$PI(4,5)P_2$	9135150
Boi1	PH	$PI(4,5)P_2$	12097146
Boi2	PH	$PI(4,5)P_2$	15023338
		$PI(3,5)P_2$	15023338
		PI(3)P	15023338
		PI(4)P	15023338
Cla4	PH	$PI(4,5)P_2$	15023338
		$PI(3,5)P_2$	15023338
Ira2_2	PH	$PI(3,4,5)P_3$	21119626
		PS	21119626
Num1	PH	$PI(4,5)P_2$	15023338
Opy1N	PH	$PI(4,5)P_2$	22562153
Opy1C	PH	$PI(4,5)P_2$	22562153
Osh2	PH	$PI(4,5)P_2$	15023338
		$PI(3,5)P_2$	15023338
		PI(3)P	15023338
		PI(4)P	15023338
Skm1	PH	$PI(4,5)P_2$	15023338
		$PI(3,5)P_2$	15023338
		PI(3)P	15023338
		PI(4)P	15023338
Ste5	PH	$PI(4,5)P_2$	16847350
Swh1	PH	$PI(4,5)P_2$	15023338
		$PI(3,5)P_2$	15023338
		PI(3)P	15023338
		PI(4)P	15023338
EEA1	FYVE	PI(3)P	9702203
Hsv2	PROPPIN	PI(3)P	22704557
		$PI(3,5)P_2$	22704557
Lactadherin	C2	PS	18160406 / 18187657
p40phox	PX	PI(3)P	11684018

Table V.8: Table summarizing the PH domains for which new high confident ($0 < TI < 2$) interactions were detected. Novel interaction indicates the 34 PH domains that were not previously reported to interact with any membranes. New specificity/mechanism indicates the 26 PH domains for which the interaction with lipids was previously reported and for which we propose additional specificity and/or binding mechanism. The 30 PH domains that did not interact with any liposomes with high confidence in this study are in italics. (n/a: not available)

Domain name	Novelty
Ask10	new specificity/mechanism
Atg26	new specificity/mechanism
Atg26Ct	novel interaction
Avo1	novel interaction
Bem2_1	
Bem2_2	novel interaction
Bem3	novel interaction
Bem3Ct	novel interaction
Boi1	
Boi2	new specificity/mechanism
Bud4	new specificity/mechanism
Bud4Ct	novel interaction
Caf120N	new specificity/mechanism
Caf120NCt	n/a
Caf120C	
Caf120CCt	novel interaction
Caf120_tandem	
Caf120Ct_tandem	novel interaction
Cdc24	new specificity/mechanism
Cdc24Ct	
Cla4	new specificity/mechanism
Cla4Ct	novel interaction
Dcp1	
Exo84	novel interaction
Exo84Ct	novel interaction
Ira1	
Ira1_CRAL-TRIO-PH	novel interaction
Ira2_1	
Ira2_2	new specificity/mechanism
Ira2Ct	
Ira2_CRAL-TRIO-PH	
Ira2Ct_CRAL-TRIO-PH	
Las17	
Mdr1	
Mdr1Ct	
Num1	new specificity/mechanism
Num1Ct	novel interaction
Nup2	novel interaction
Nvj2	novel interaction
Opy1N	new specificity/mechanism
Opy1C	new specificity/mechanism
Osh2	new specificity/mechanism
Osh3	new specificity/mechanism
Osh3Ct	novel interaction
Pkh2	novel interaction
Plc1	novel interaction
Plc1Ct	novel interaction
Psy2	novel interaction
Rgc1	

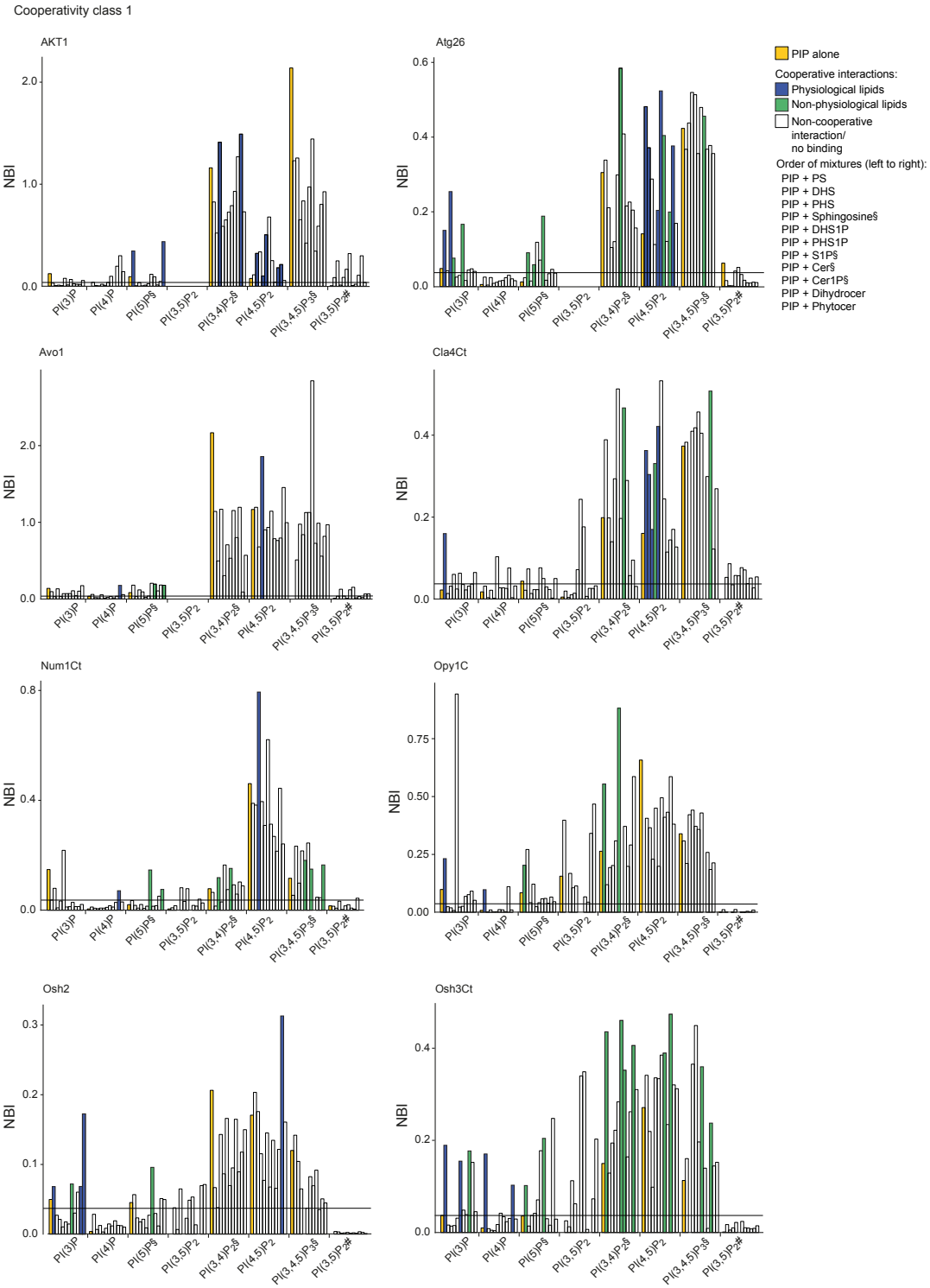
Rom1	
Rom1Ct_A	novel interaction
Rom1Ct_B	
Rom2	
Rtt106	novel interaction
Rtt106Ct	novel interaction
Sec3	new specificity/mechanism
Sec3Ct	
Sip3	new specificity/mechanism
Skg3N	new specificity/mechanism
Skg3C	
Skm1	new specificity/mechanism
Slm1	
Slm1Ct	novel interaction
Slm2	new specificity/mechanism
Spo14	
Spo14Ct	
Spo71N	new specificity/mechanism
Spo71M	novel interaction
Spo71C	
Ste5	novel interaction
Swh1	new specificity/mechanism
Swh1Ct	novel interaction
Syt1	
Syt1Ct_A	novel interaction
Syt1Ct_B	novel interaction
Tfb1	
Tfb1Ct	novel interaction
Tus1	
Yel1	
Yhr131c	
Ynl144c	
Ynl144cCt	novel interaction
Yrb2	novel interaction
Ysp1	novel interaction
Ysp1Ct	novel interaction
AKT1	new specificity/mechanism
Dynamin1	new specificity/mechanism
PDK1	new specificity/mechanism
Plcδ1	new specificity/mechanism
Pleckstrin	new specificity/mechanism
SOS1	new specificity/mechanism

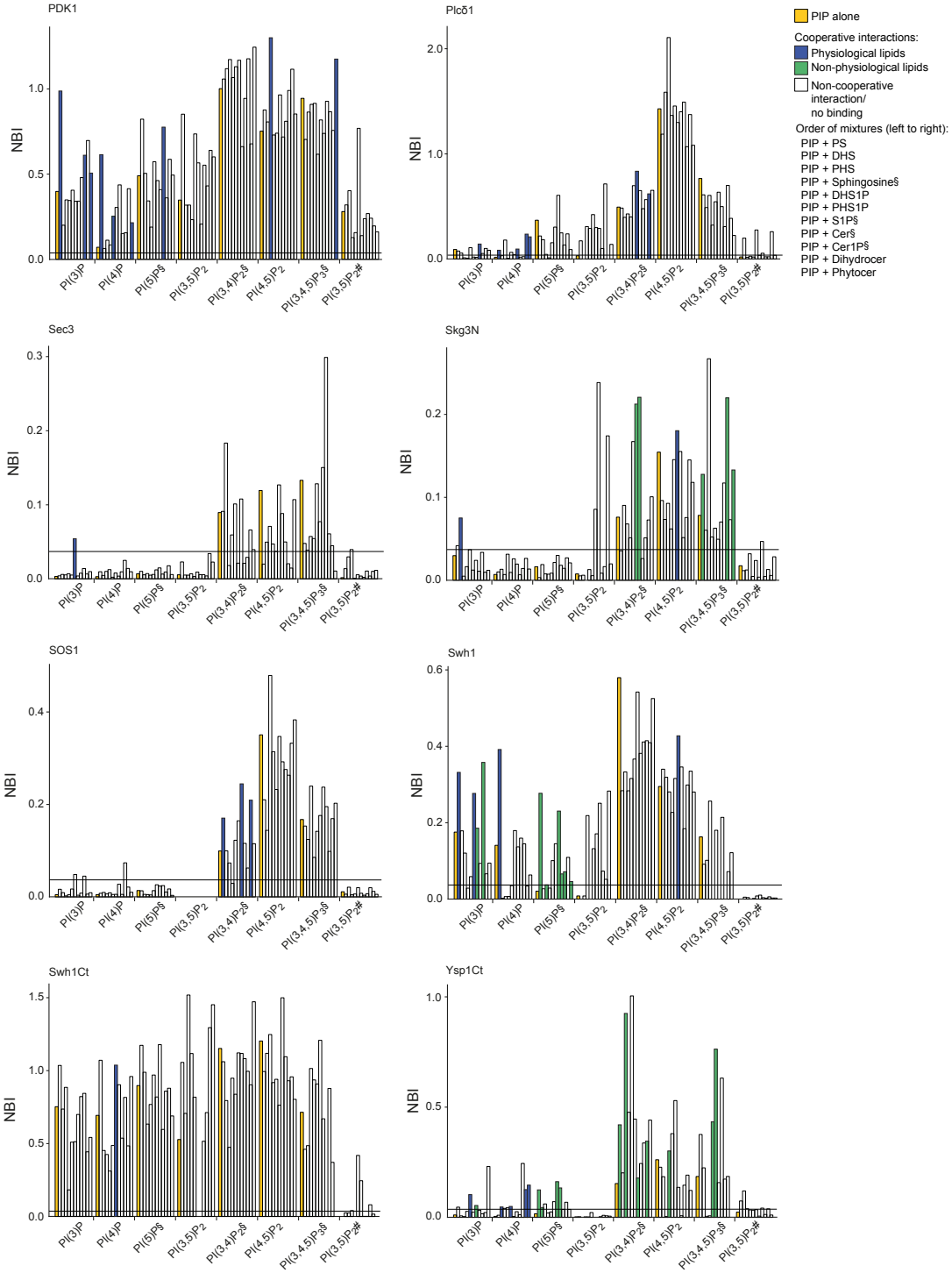
Table V.9: Table listing the PH domains used for the validation experiments. First column indicates the PH domain name as used in this study. The second column indicates its Uniprot ID. The third column the species it belongs to. The fourth column its amino acid sequence. The columns 5 to 11 indicate the protein concentration used in the respective experiments (mentioned by the column header).

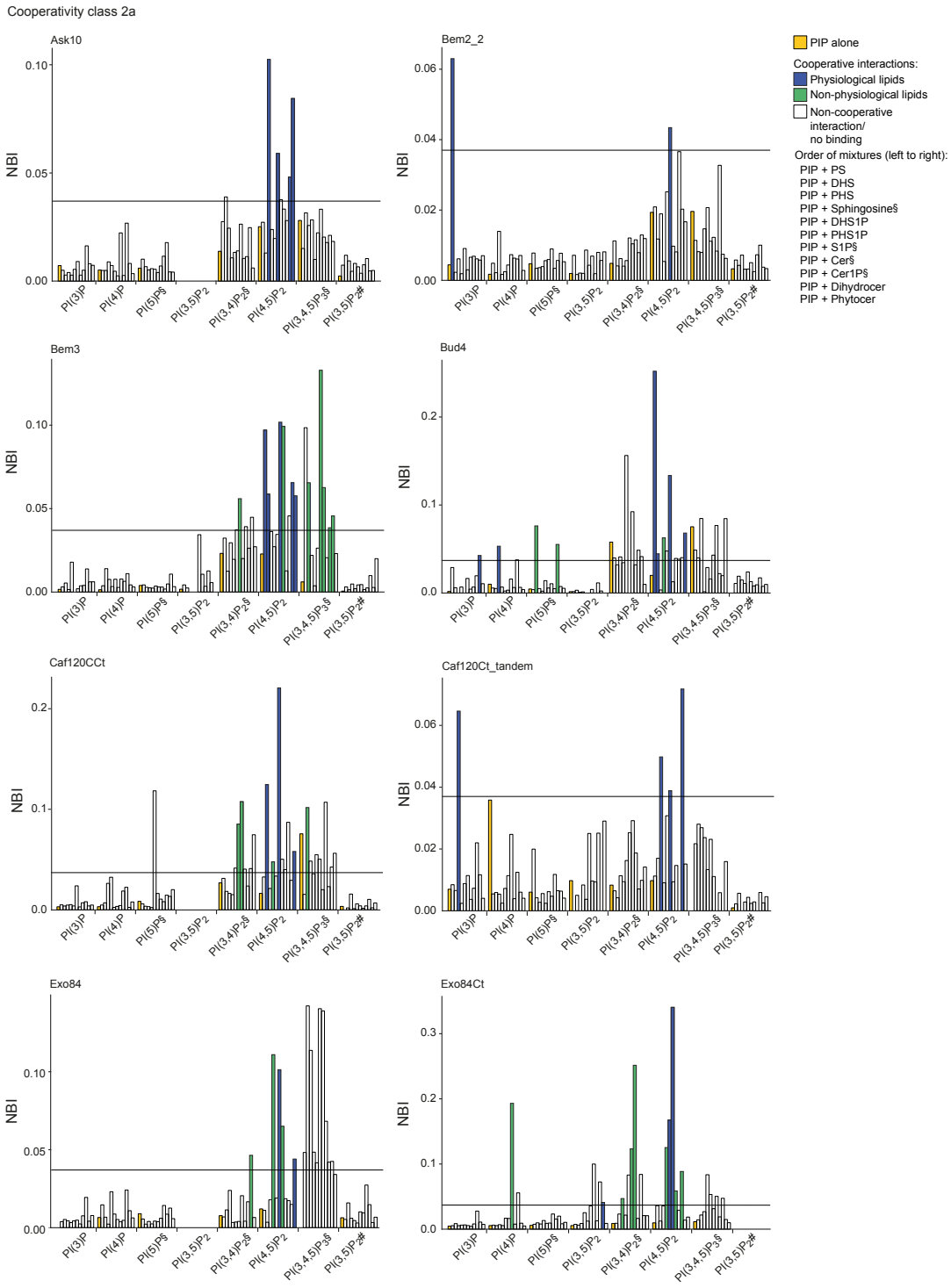
PH domain name	Uniprot ID	Species	PH domain sequence (AA)	Protein concentration [µM]						
				Fig. 5b	Fig. 5d,e	Fig. 5f	Fig. 3b/ Fig. S3b	Fig. 4b	Fig. S1b,c	Fig. S1i
OSBP1	P22059	*Homo sapiens*	GSAREGWLFKWTNYIKGYQRRWFVLSNGLLSYYRSKAEMRHTCRGTINLATANITVED SCNFIISNGGAQTYHLKASSEVERQRWVTALELAKA	1,71	-	-	-	-	-	-
OSBP2	Q969R2	*Homo sapiens*	LDSFEGWLLKWTNYLKGYQRRWFVLGNGLLSYYRNQGEMAHTCRGTINLSTAHIDTED SCGILLTSGARSYHLKASSEVDRQQWITALELAKA	1,87	2,19	-	-	-	6,15	-
OSBP2 R262L			LDSFEGWLLKWTNYLKGYQRRWFVLGNGLLSYYRNQGEMAHTCRGTINLSTAHIDTED SCGILLTSGARSYHLKASSEVDLQQWITALELAKA	-	2,09	-	-	-	5,90	-
OSBPL3	Q9H4L5	*Homo sapiens*	PPVQKGFLLKKRKWPLKGWHKRFFYLDKGILKYAKSQTDIEREKLHGCIDVGLSVMSVK KSSKCIDLDTEEHIYHLKVKSEEVFDEWVSKLRHHRM	1,70	-	-	-	-	-	-
OSBPL5	Q9H0X9	*Homo sapiens*	VVIMADSLKIRGTLKSWTKLWCVLKPGVLLIYKTPKVGQWVGTVLLHCCELIERPSKKDG FCFKLFHPLDQSVWAVKGPKGESVGSITQPLPSSYLIFRAASESDGRCWLDALELALR	1,51	-	-	-	-	-	-
OSBPL7	Q9BZF2	*Homo sapiens*	RQEGHLLKKRKWPLKGWHKRYFVLEDGILHYATTRQDITKGKLHGSIDVRLSVMSINKK AQRIDLDTEDNIYHLKIKSQDLFQSWVAQLRAHR	1,76	-	-	-	-	-	-
OSBPL8	Q9BZF1	*Homo sapiens*	VIVMADWLKIRGTLKSWTKLWCVLKPGVLLIYKTQKNGQWVGTVLLNACEIIERPSKKDG FCFKLFHPLEQSIWAVKGPKGEAVGSITQPLPSSYLIIRATSESDGRCWMDALELALK	1,48	-	-	-	-	-	-
OSBPL10	Q9BXB5	*Homo sapiens*	PALEGVLSKYTNLLQGWQNRYFVLDFEAGILQYFVNEQSKHQKPRGVLSLSGAIVSLSD EAPHMLVVYSANGEMFKLRAADAKEKQFWVTQLRACAK	1,87	-	-	-	-	-	-
OSBPL11	Q9BXB4	*Homo sapiens*	NVYGYLMKYTNLVTGWQYRFFVLNNEAGLLEYFVNEQSRNQKPRGTLQLAGAVISPSD EDSHTFTVNAASGEQYKLRATDAKERQHWVSRLQICTQ	1,91	-	-	-	-	-	-
CERT	Q9Y5P4	*Homo sapiens*	MSDNQSWNSSGSEEDPETESGPPVERCGVLSKWTNYIHGWQDRWVVLKNNALSYYK SEDETEYGCRGSICLSKAVITPHDFDECRFDISVNDSVWYLRAQDPDHRQQWIDAIEQH KTESGYG	1,42	1,42	-	-	-	-	-
CERT R98Q			MSDNQSWNSSGSEEDPETESGPPVERCGVLSKWTNYIHGWQDRWVVLKNNALSYYK SEDETEYGCRGSICLSKAVITPHDFDECRFDISVNDSVWYLQAQDPDHRQQWIDAIEQH KTESGYG	-	1,41	-	-	-	-	-
FAPP1	Q9HB20	*Homo sapiens*	MEGVLYKWTNYLTGWQPRWFVLDNGILSYYDSQDDVCKGSKGSIKMAVCEIKVHSADN TRMELIIPGEQHFYMKAVNAAERQRWLVALGSSKACLTDT	1,60	1,60	2,07	-	-	-	-
FAPP1 T9A			MEGVLYKWANYLTGWQPRWFVLDNGILSYYDSQDDVCKGSKGSIKMAVCEIKVHSADN TRMELIIPGEQHFYMKAVNAAERQRWLVALGSSKACLTDT	-	-	2,11	-	-	-	-
FAPP1 K74Q			MEGVLYKWTNYLTGWQPRWFVLDNGILSYYDSQDDVCKGSKGSIKMAVCEIKVHSADN TRMELIIPGEQHFYMQAVNAAERQRWLVALGSSKACLTDT	-	1,35	-	-	-	-	-
Swh1	P35845	*Saccharomyces cerevisiae*	LHEAPTYKGYLKKWTNFAQGYKLRWFILSSDGKLSYYIDQADTKNACRGSLNMSSCSL HLDSSEKLKFEIIGGNNGVIRWHLKGNHPIETNRWYWAIQGAIRYAKDREILL	-	1,24	-	1,44	-	4,75	0.1-7.0
Swh1 K360Q			LHEAPTYKGYLKKWTNFAQGYKLRWFILSSDGKLSYYIDQADTKNACRGSLNMSSCSL HLDSSEKLKFEIIGGNNGVIRWHLQGNHPIETNRWYWAIQGAIRYAKDREILL	-	1,36	-	-	-	5,05	-
Num1	Q00402	*Saccharomyces cerevisiae*	NEPSIIPALTQTVIGEYLFKYYPRLGPFGFESRHERFFWVHPYTLTLYWSASNPILENPAN TKTKGVAILGVESVTDPNPYPTGLYHKSIVVTTETRTIKFTCPTRQRHNIWYNSLRYLLQR NMQGISLE	-	-	-	1,39	-	7,01	-
Opy1N	P38271	*Saccharomyces cerevisiae*	MIAGATAPSSQHEILIASNLIKKPSTSQNKTPTAQSSSGNNGAADGAPQGYHHHHHHHR HLWWPRTTDHQYWCVLRKNQFAYYKTRDEREAISVIPRFDILNFKISELDGILTVYTPSK DLIFKFPSGQNEKVGMELMHNWKIALEKFLSSPSGN	-	-	-	9,10	-	-	-
Opy1C			DPRNAEHQVCSGILYTKVKKKKLFNRAKWQKFNVELTNTSFNLYSFKTGKLKKSIKLDKII DCIELDNNSKMKNDDTNFALITFDERLSFKAANDQDMVDWIINFKSGILIRKKLKAENI	-	-	-	11,10	-	-	-
Osh3	P38713	*Saccharomyces cerevisiae*	QDMLFRVGQGRYLQGYLLKKRRKRLQGFKKRFFTLDFRYGTLSYYLNDHNQTCRGEIV ISLSSVSANKKDKIIIIDSGMEVWVLKATTKENWQSWVDALQTCFDDQFEDK	-	-	-	5,92	-	-	-
Plcδ1	P10688	*Rattus norvegicus*	HGLQDDPDLQALLKGSQLLKVKSSSWRRERFYKLQEDCKTIWQESRKVMRSPESQLF SIEDIQEVRMGHRTEGLEKFARDIPEDRCFSIVFKDQRNTLDLIAPSPADAQHWVQGLRK IIHHSGSMDQRQK	-	-	-	1,86	-	-	-
SOS1	Q07889	*Homo sapiens*	QQMKGKQLAIKKMNEIQKNIDGWEGKDIGQCCNEFIMEGTLTRVGAKHERHIFLFDGLM ICCKSNHGQPRLPGASNAEYRLKEKFFMRKVQINDKDDTNEYKHAFEIILKDENSVIFSA KSAEEKNNWMAALISLQYRSTLE	-	-	-	6,43	-	-	-
Yrb2	P40517	*Saccharomyces cerevisiae*	GEESEECIYQVNAKLYQLSNIKEGWKERGVGIIKINKSKDDVEKTRIVMRSRGILKVILNIQ LVKGFTVQKGFTGSLQSEKFIRLLAVDDNGDPAQYAIKTGKKETTDELYNIIVKSVPK	-	-	-	2,10	-	-	-
Nup2	P32499	*Saccharomyces cerevisiae*	EEDEVALFSQKAKLMTFNAETKSYDSRGVGEMKLLKKKDDPSKVRLLCRSDGMGNVLL NATVVDSFKYEPLAPGNDNLIKAPTVAADGKLVTYIVKFKQKEEGRSFTKAIEDAKKEMK	-	-	-	3,48	-	-	-
Avo1	Q08236	*Saccharomyces cerevisiae*	DNFQDLFTGAYHKYKVWRRQQMSFINKHERTLAIDGDYIYIVPPEGRIHWHDNVKTKSL HISQVVLVKKSKRVPEHFKIFVRREGQDDIKRYYFEAVSGQECTEIVTRLQNLLSAYRMN HK	-	-	-	9,34	-	-	-
Slm1Ct	G0SAZ0	*Chaetomium thermophilum*	HYACQEIRAGLLERKSKYLKSYTPGWYVLSPTHLHEFKSADKTQAPVMSLYLPEQKLGS HSEEGSSSNKFILKGRQTGHMHRGHTWVFRAESHDTMMAWYEDIKALTERTPEER	-	-	-	15,52	-	-	-
AKT1 E17K	P31749	*Homo sapiens*	MSDVAIVKEGWLHKRGKYIKTWRPRYFLLKNDGTFIGYKERPQDVDQREAPLNNFSVA QCQLMKTERPRPNTFIIRCLQWTTVIERTFHVETPEEREEWTTAIQTVADGLKKQEEEE MDF	-	-	-	-	5,29	6,90	-
AKT1	P31749	*Homo sapiens*	MSDVAIVKEGWLHKRGEYIKTWRPRYFLLKNDGTFIGYKERPQDVDQREAPLNNFSVA QCQLMKTERPRPNTFIIRCLQWTTVIERTFHVETPEEREEWTTAIQTVADGLKKQEEEE MDF	-	-	-	-	-	6,73	0.05-6.0
Cdc24	P11433	*Saccharomyces cerevisiae*	ISKFGELLYFDKVFISTTNSSSEPEREFEVYLFEKIIILFSEVVTKKSASSLILKKKSSTSASI SASNITDNNGSPHHSYHKRHSNSSSSNNIHLSSSSAAAIIHSSTNSSDNNSNNSSSSSLF KLSANEPKLDLRGRIMIMNLNQIIPQNNRSLNITWESIKEQGNFLLKFKNEETRDNWSSC LQQLIHDLKNEQFKARHHSST	-	-	-	-	-	4,55	-
Osh2	Q12451	*Saccharomyces cerevisiae*	ASSKPPTYKGFLKKWTNFAHGYKLRWFILSGDGNLSYYKDQSHVDRPRGTLKVSTCRL HIDSSEKLNFELLGGITGTTRWRLKGNHPIETTRWVNAIQSAIRFAKDKEILNKKKAVP	-	-	-	-	-	5,21	-
Cla4	P48562	*Saccharomyces cerevisiae*	TSTSKKKSGWVSYKDDGILSFIWQKRYLMLHDSYVALYKNDKQNDDAILKIPLTSIISVSR TQLKQYCFELVRCSDRNSVSSGSSSSLNVSSDSNSKKSIYIATKTESDLHSWLDAIFAKC PLLSGVSSPTNFTHK	-	-	-	-	-	6,17	-
Bem3	P32873	*Saccharomyces cerevisiae*	MDDNVKDGSLLLRRPKTLTGNSTWRVRYGILRDDVLQLFDKNQLTETIKLRQSSIELIPNL PEDRFGTRNGFLITEHKKSGLSTSTKYYICTETSKERELWLSAFSDYIDPSQS	-	-	-	-	-	5,45	-
Boi2	P39969	*Saccharomyces cerevisiae*	SISVKEAMKDADFSGWMSKKGSGAMSTWKTRFFTLHGTRLSYFSSTTDTRERGLIDITA HRVVPAKEDDKLVSLYAASTGKGRYCFKLLPPQPGSKKGLTFTQPRTHYFAVDNKEEMR GWMAALIKTTIDIDTSVP	-	-	-	-	-	-	0.2-7.0

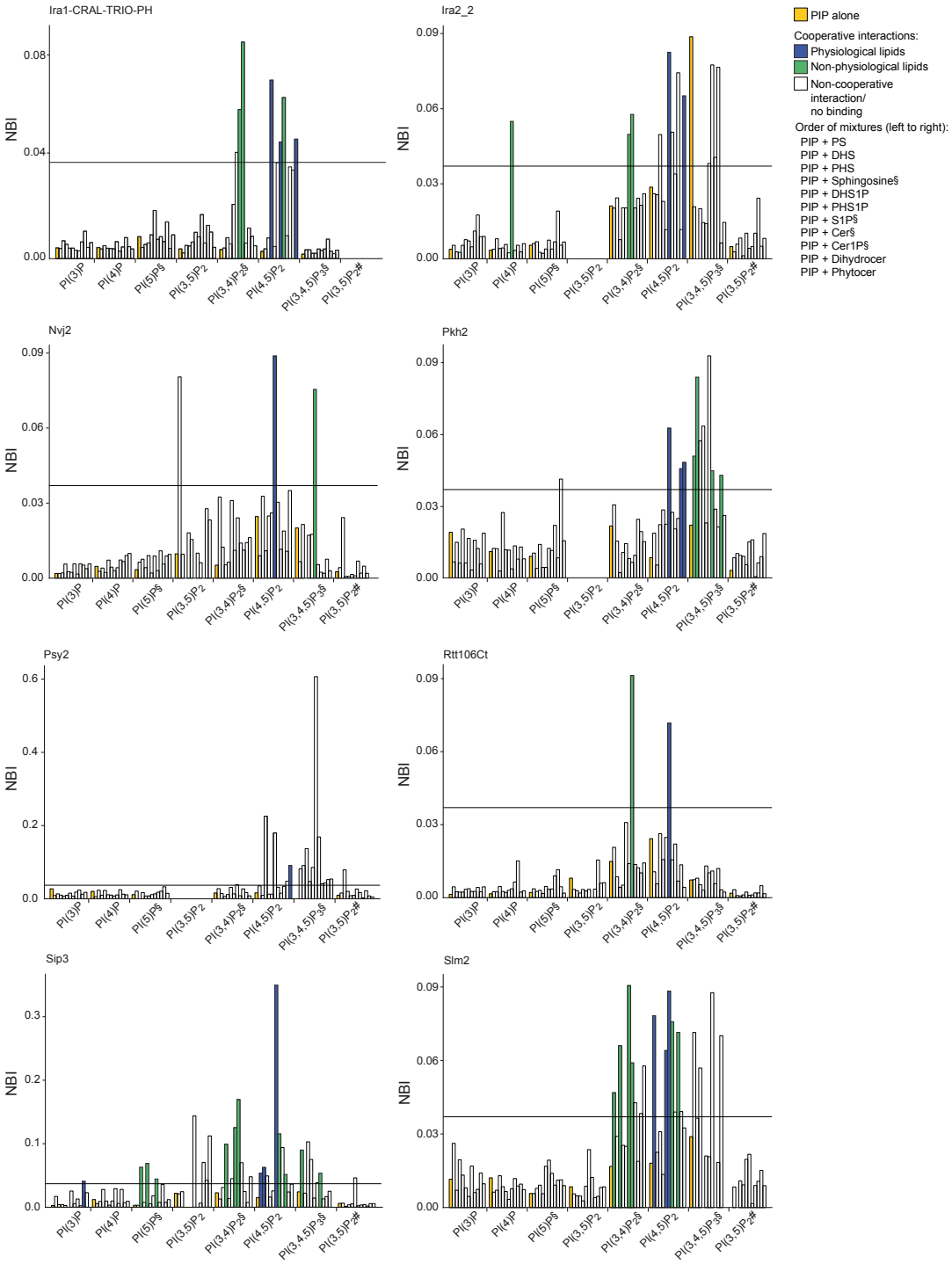
Figure 1: The PH domains are divided into three groups according to their cooperativity class (class 1, 2a and 2b). The barplots show NBI values measured with liposome types composed of single PIP liposomes (yellow bars), or mixtures liposomes (colour/white bars). The colour bars represent predicted cooperative interactions - the combinations composed of lipids physiological in S. cerevisiae in blue, the combinations composed of at least one lipid non-physiological in S. cerevisiae in green, and the white bars indicate non-cooperative interactions or no binding. The horizontal line represents the binding threshold (NBI = 0.037). Only data of PH domains with at least one cooperativity event of physiological lipids are shown. (§) indicates that the lipid is not physiological in S. cerevisiae, # indicates different variants of DOPI35P2 lipid.

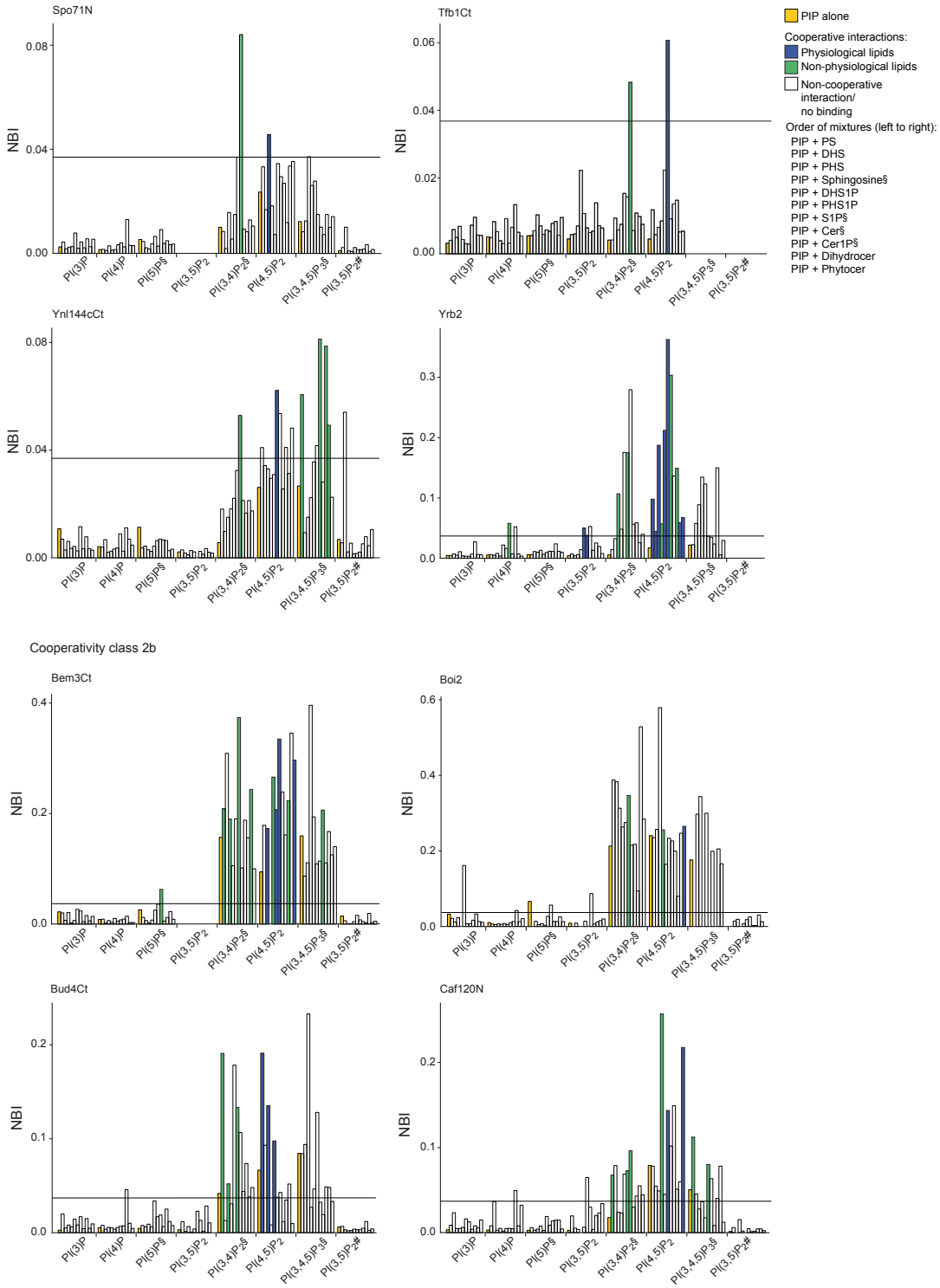
Figure S4

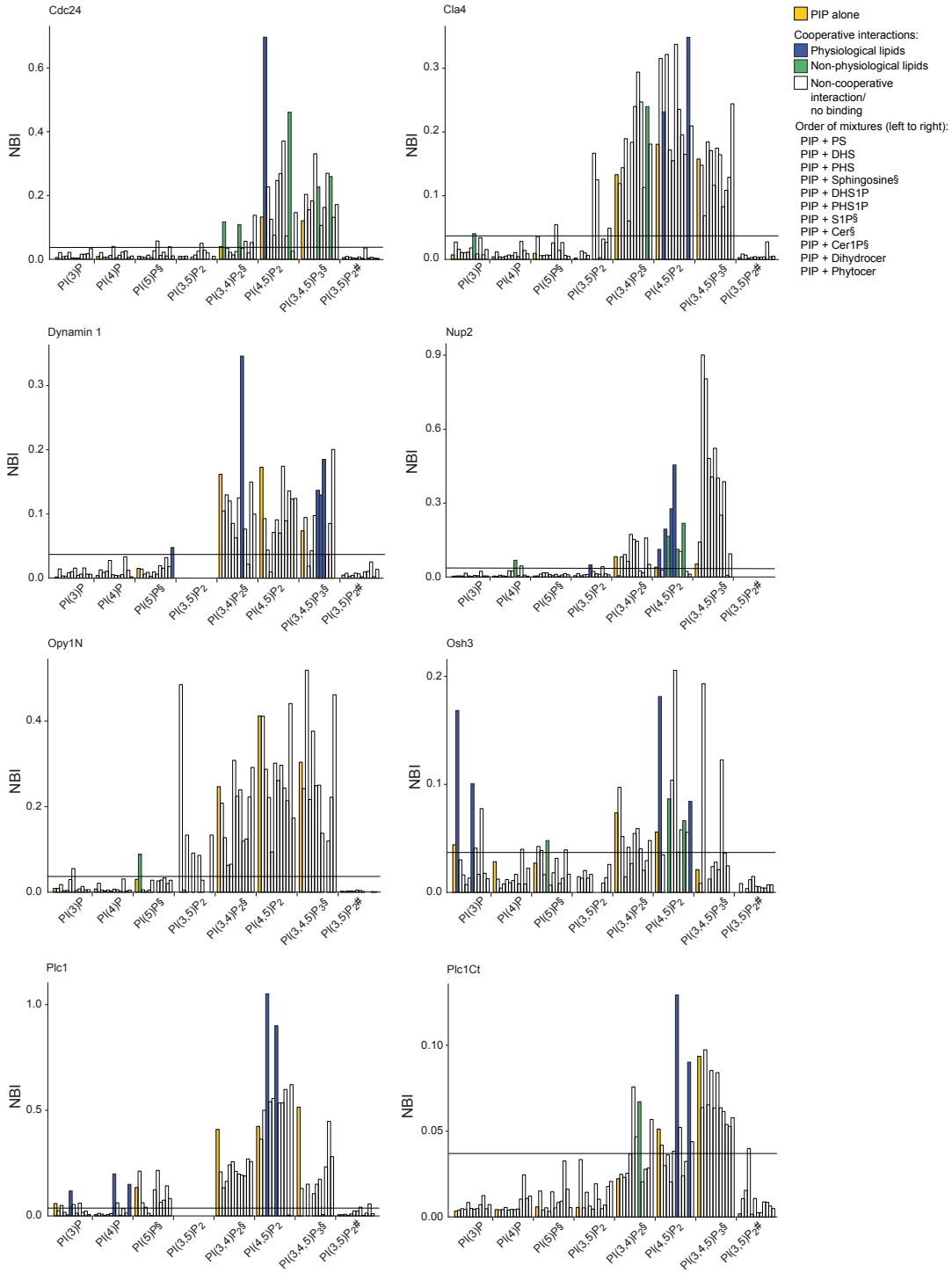

Cooperativity class 1
AKT1
Atg26
Avo1
Cla4Ct
Num1Ct
Opy1C
Osh2
Osh3Ct
NBI
PIP alone
Cooperative interactions:
Physiological lipids
Non-physiological lipids
Non-cooperative interaction/ no binding
Order of mixtures (left to right):
PIP + PS
PIP + DHS
PIP + PHS
PIP + Sphingosine§
PIP + DHS1P
PIP + PHS1P
PIP + S1P§
PIP + Cer§
PIP + Cer1P§
PIP + Dihydrocer
PIP + Phytocer
PI(3)P
PI(4)P
PI(5)P§
PI(3,5)P2
PI(3,4)P2§
PI(4,5)P2
PI(3,4,5)P3§
PI(3,5)P2#


PDK1
Plcδ1
Sec3
Skg3N
SOS1
Swh1
Swh1Ct
Ysp1Ct
NBI
PIP alone
Cooperative interactions:
Physiological lipids
Non-physiological lipids
Non-cooperative interaction/ no binding
Order of mixtures (left to right):
PIP + PS
PIP + DHS
PIP + PHS
PIP + Sphingosine§
PIP + DHS1P
PIP + PHS1P
PIP + S1P§
PIP + Cer§
PIP + Cer1P§
PIP + Dihydrocer
PIP + Phytocer
PI(3)P
PI(4)P
PI(5)P§
PI(3,5)P2
PI(3,4)P2§
PI(4,5)P2
PI(3,4,5)P3§
PI(3,5)P2#


Cooperativity class 2a
Ask10
Bem2_2
Bem3
Bud4
Caf120CCt
Caf120Ct_tandem
Exo84
Exo84Ct
NBI
PI(3)P
PI(4)P
PI(5)P§
PI(3,5)P2
PI(3,4)P2§
PI(4,5)P2
PI(3,4,5)P3§
PI(3,5)P2#
PIP alone
Cooperative interactions:
Physiological lipids
Non-physiological lipids
Non-cooperative interaction/ no binding
Order of mixtures (left to right):
PIP + PS
PIP + DHS
PIP + PHS
PIP + Sphingosine§
PIP + DHS1P
PIP + PHS1P
PIP + S1P§
PIP + Cer§
PIP + Cer1P§
PIP + Dihydrocer
PIP + Phytocer


Ira1-CRAL-TRIO-PH
Ira2_2
Nvj2
Pkh2
Psy2
Rtt106Ct
Sip3
Slm2
NBI
PIP alone
Cooperative interactions:
Physiological lipids
Non-physiological lipids
Non-cooperative interaction/ no binding
Order of mixtures (left to right):
PIP + PS
PIP + DHS
PIP + PHS
PIP + Sphingosine§
PIP + DHS1P
PIP + PHS1P
PIP + S1P§
PIP + Cer§
PIP + Cer1P§
PIP + Dihydrocer
PIP + Phytocer


Spo71N
Tfb1Ct
Ynl144cCt
Yrb2
Cooperativity class 2b
Bem3Ct
Boi2
Bud4Ct
Caf120N
NBI
PI(3)P
PI(4)P
PI(5)P§
PI(3,5)P2
PI(3,4)P2§
PI(4,5)P2
PI(3,4,5)P3§
PI(3,5)P2#
PIP alone
Cooperative interactions:
Physiological lipids
Non-physiological lipids
Non-cooperative interaction/ no binding
Order of mixtures (left to right):
PIP + PS
PIP + DHS
PIP + PHS
PIP + Sphingosine§
PIP + DHS1P
PIP + PHS1P
PIP + S1P§
PIP + Cer§
PIP + Cer1P§
PIP + Dihydrocer
PIP + Phytocer


Cdc24
Cla4
Dynamin 1
Nup2
Opy1N
Osh3
Plc1
Plc1Ct
NBI
PIP alone
Cooperative interactions:
Physiological lipids
Non-physiological lipids
Non-cooperative interaction/ no binding
Order of mixtures (left to right):
PIP + PS
PIP + DHS
PIP + PHS
PIP + Sphingosine§
PIP + DHS1P
PIP + PHS1P
PIP + S1P§
PIP + Cer§
PIP + Cer1P§
PIP + Dihydrocer
PIP + Phytocer
PI(3)P
PI(4)P
PI(5)P§
PI(3,5)P2
PI(3,4)P2§
PI(4,5)P2
PI(3,4,5)P3§
PI(3,5)P2#

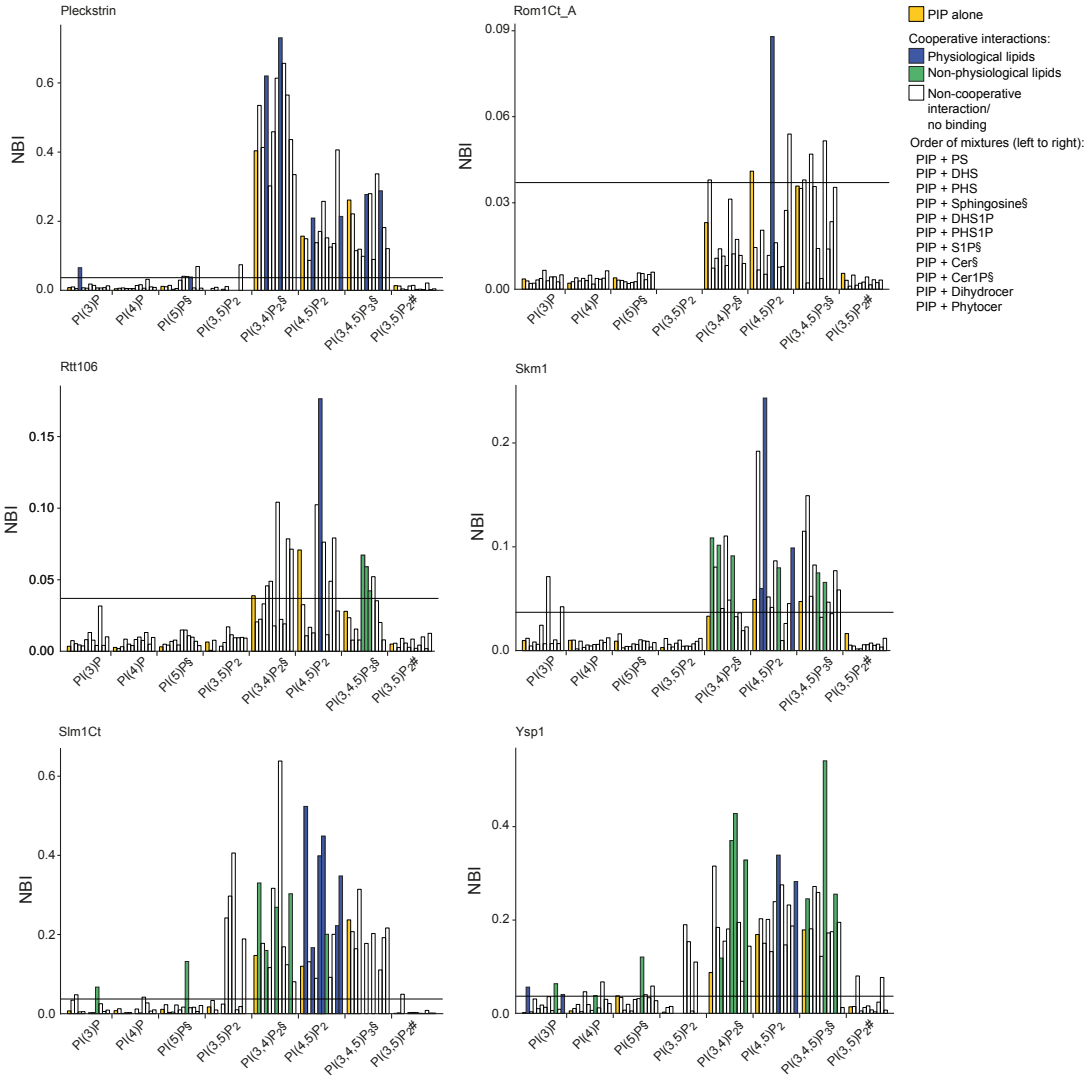
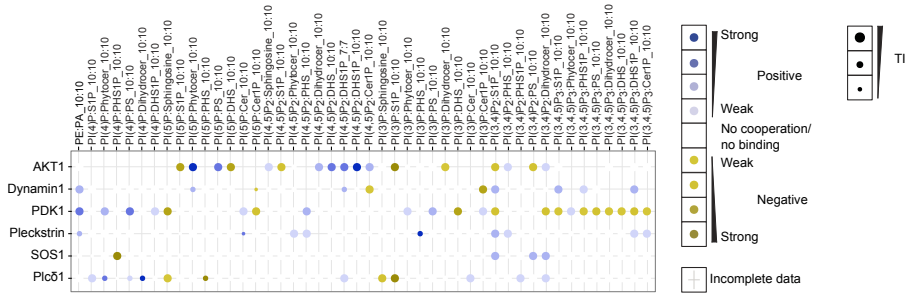

Pleckstrin
Rom1Ct_A
Rtt106
Skm1
Slm1Ct
Ysp1
NBI
PIP alone
Cooperative interactions:
Physiological lipids
Non-physiological lipids
Non-cooperative interaction/ no binding
Order of mixtures (left to right):
PIP + PS
PIP + DHS
PIP + PHS
PIP + Sphingosine§
PIP + DHS1P
PIP + PHS1P
PIP + S1P§
PIP + Cer§
PIP + Cer1P§
PIP + Dihydrocer
PIP + Phytocer
PI(3)P
PI(4)P
PI(5)P§
PI(3,5)P2
PI(3,4)P2§
PI(4,5)P2
PI(3,4,5)P3§
PI(3,5)P2#

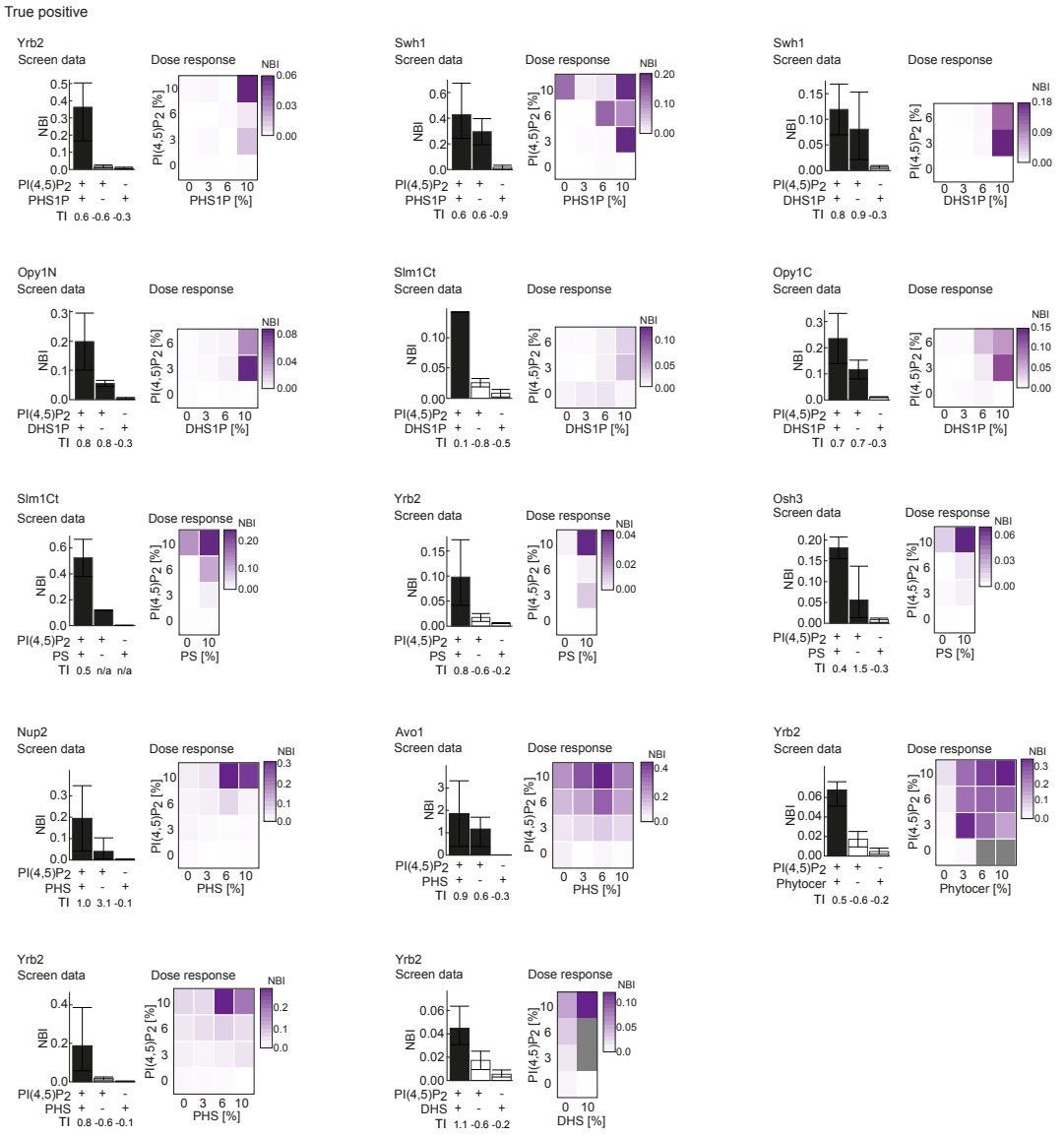
Figure 2: (A) Landscape of cooperating lipids in the targeting of mammalian PH domains to liposomal membranes. (B) Validation of interactions of PH domains with cooperating lipids using dose response experiments. The table gives the results of dose response experiments (heatmaps) for 31 selected PH domain-lipids combination pairs and compares them with the results obtained from the PH domain screen (barplots and TI values). The results are divided into four groups: true positive, true negative, false positive and false negative. Each cell in the heatmaps gives the NBI value (violet) measured for the PH domain interaction with liposomes containing the indicated concentration of signalling lipids. The grey colour indicates missing data. Values are mean ($n \geq 2$). The bar plots show mean NBI values (+/- s.d.) measured in the PH domain screen for each liposome type containing indicated signalling lipids or their combinations (black, NBI ≥ 0.037; white, NBI < 0.037). The concentration of signalling lipids in liposomes used in the PH domain screen was 10 mol %, except the combination of DOPI45P2 and DHS1P, for which data from liposomes containing 7 mol % of both lipids are shown. The TI values for each experiment is given under the bar plots. In the group of false negative, the PH domains recognizing cooperating lipids at lower lipid concentrations are marked with a star.

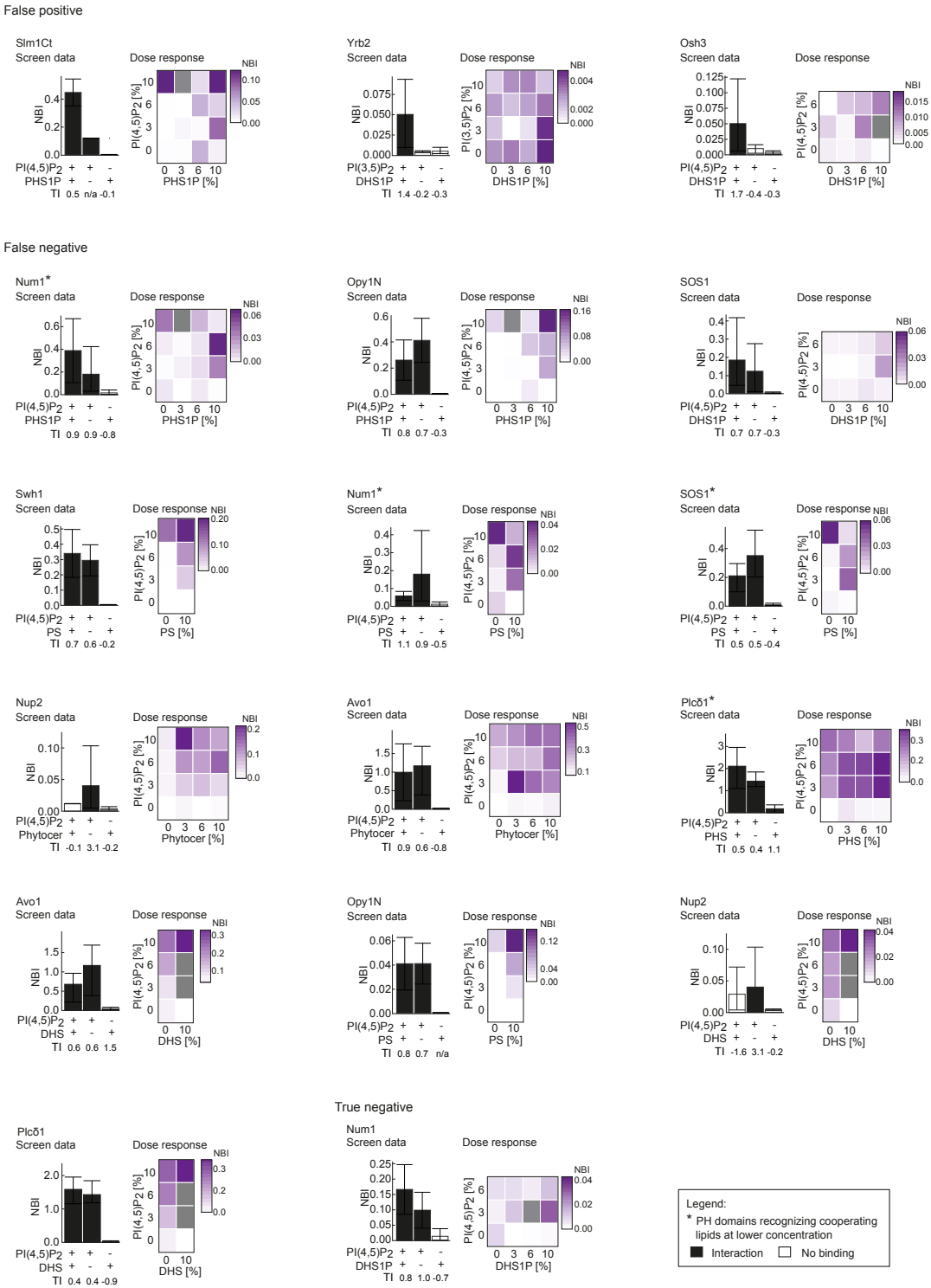
Figure S3

A


PE:PA_10:10
PI(4)P:S1P_10:10
PI(4)P:Phytocer_10:10
PI(4)P:PHS1P_10:10
PI(4)P:PS_10:10
PI(4)P:Dihydrocer_10:10
PI(4)P:DHS1P_10:10
PI(4)P:Sphingosine_10:10
PI(5)P:S1P_10:10
PI(5)P:Phytocer_10:10
PI(5)P:PS_10:10
PI(5)P:PHS_10:10
PI(5)P:DHS_10:10
PI(5)P:Cer_10:10
PI(4,5)P2:Sphingosine_10:10
PI(4,5)P2:S1P_10:10
PI(4,5)P2:Phytocer_10:10
PI(4,5)P2:PHS_10:10
PI(4,5)P2:Dihydrocer_10:10
PI(4,5)P2:DHS_10:10
PI(4,5)P2:DHS1P_7:7
PI(4,5)P2:DHS1P_10:10
PI(4,5)P2:Cer1P_10:10
PI(3)P:Sphingosine_10:10
PI(3)P:S1P_10:10
PI(3)P:Phytocer_10:10
PI(3)P:PHS_10:10
PI(3)P:Dihydrocer_10:10
PI(3)P:DHS_10:10
PI(3)P:Cer_10:10
PI(3)P:Cer1P_10:10
PI(3,4)P2:S1P_10:10
PI(3,4)P2:PHS_10:10
PI(3,4)P2:PHS1P_10:10
PI(3,4)P2:PS_10:10
PI(3,4)P2:Dihydrocer_10:10
PI(3,4,5)P3:S1P_10:10
PI(3,4,5)P3:Phytocer_10:10
PI(3,4,5)P3:PHS1P_10:10
PI(3,4,5)P3:PS_10:10
PI(3,4,5)P3:Dihydrocer_10:10
PI(3,4,5)P3:DHS_10:10
PI(3,4,5)P3:DHS1P_10:10
PI(3,4,5)P3:Cer1P_10:10
AKT1
Dynamin1
PDK1
Pleckstrin
SOS1
Plcδ1
Strong
Positive
Weak
No cooperation/ no binding
Weak
Negative
Strong
Incomplete data
TI

B


True positive
Yrb2
Screen data
Dose response
PI(4,5)P2 + + -
PHS1P + - +
TI 0.6 -0.6 -0.3
PHS1P [%]
Swh1
PI(4,5)P2 + + -
PHS1P + - +
TI 0.6 0.6 -0.9
PHS1P [%]
Swh1
PI(4,5)P2 + + -
DHS1P + - +
TI 0.8 0.9 -0.3
DHS1P [%]
Opy1N
PI(4,5)P2 + + -
DHS1P + - +
TI 0.8 0.8 -0.3
DHS1P [%]
Slm1Ct
PI(4,5)P2 + + -
DHS1P + - +
TI 0.1 -0.8 -0.5
DHS1P [%]
Opy1C
PI(4,5)P2 + + -
DHS1P + - +
TI 0.7 0.7 -0.3
DHS1P [%]
Slm1Ct
PI(4,5)P2 + + -
PS + - +
TI 0.5 n/a n/a
PS [%]
Yrb2
PI(4,5)P2 + + -
PS + - +
TI 0.8 -0.6 -0.2
PS [%]
Osh3
PI(4,5)P2 + + -
PS + - +
TI 0.4 1.5 -0.3
PS [%]
Nup2
PI(4,5)P2 + + -
PHS + - +
TI 1.0 3.1 -0.1
PHS [%]
Avo1
PI(4,5)P2 + + -
PHS + - +
TI 0.9 0.6 -0.3
PHS [%]
Yrb2
PI(4,5)P2 + + -
Phytocer + - +
TI 0.5 -0.6 -0.2
Phytocer [%]
Yrb2
PI(4,5)P2 + + -
PHS + - +
TI 0.8 -0.6 -0.1
PHS [%]
Yrb2
PI(4,5)P2 + + -
DHS + - +
TI 1.1 -0.6 -0.2
DHS [%]
NBI
PI(4,5)P2


False positive
Slm1Ct
Screen data
Dose response
Yrb2
Osh3
False negative
Num1*
Opy1N
SOS1
Swh1
Num1*
SOS1*
Nup2
Avo1
Plcδ1*
Avo1
Opy1N
Nup2
True negative
Plcδ1
Num1
NBI
PI(4,5)P2
PI(3,5)P2
PHS1P
DHS1P
PS
Phytocer
DHS
PHS
TI
PHS1P [%]
DHS1P [%]
PS [%]
Phytocer [%]
DHS [%]
PHS [%]
Legend:
* PH domains recognizing cooperating lipids at lower concentration
Interaction
No binding

Chapter 7

Relative to Part III

Table V.1: Table summarizing the arguments required by the name matching module. The first column indicates the argument name, the second column indicates whether the argument is mandatory, the third column gives an example of input value, the fourth column mentions whether a default value is already included within the module and the fifth column gives a short description of what kind of value is expected to be given.

Argument name	REQUIRED?	Example	Default value	Description
-n	YES	fg_input	No default value	The absolute path of a chemical name file properly formatted
-o	YES	nm_fg_output	No default value	Name matching output (corresponding to background)
-u	OPTIONAL	STITCH	STITCH	Universe used for the name matching
-a	OPTIONAL	TRUE	TRUE	Fuzzy name matching algorithm
-e	OPTIONAL	TRUE	TRUE	Heuristic name matching algorithm
-s	OPTIONAL	TRUE	FALSE	Additional synonym output
--verbose	OPTIONAL	2	2	Verbose level

Table V.2: Table summarizing the arguments required by the enrichment module. The first column indicates the argument name, the second column indicates whether the argument is mandatory, the third column gives an example of input value, the fourth column mentions whether a default value is already included within the module and the fifth column gives a short description of what kind of value is expected to be given.

Argument name	REQUIRED?	Example	Default value	Description
-f	YES	nm_fg_output	No default value	name matching output corresponding to foreground
-b	OPTIONAL	ALL	ALL	Either name matching output corresponding to background, or "ALL"
-d	YES	drug-side-effects	No default value	Database form which biological terms will be extracted
-o	YES	enr-enr-output	No default value	Enrichment module output containing the enriched biological terms
-p	YES	enr-ann-output	No default value	Enrichment module output containing the chemical/biological terms annotations
-m	OPTIONAL	fisher	fisher	Method used to perform the enrichment
-a	OPTIONAL	0.05	0.05	Enrichment significance level
-c	OPTIONAL	FDR	FDR	FDR correction method
--versbose	OPTIONAL	2	2	Verbose level

Table V.3: Table summarizing the arguments required by the visualization module. The first column indicates the argument name, the second column indicates whether the argument is mandatory, the third column gives an example of input value, the fourth column mentions whether a default value is already included within the module and the fifth column gives a short description of what kind of value is expected to be given.

Argument name	REQUIRED	Example	Default value	Description
-i	YES	enr-enr-output	No default value	The enrichment table outputted by the enrichment module
-o	YES	res-viz-tab.html	No default value	File name (.html) where the table containing links to external resources will be generated

Bibliography

[1] R. S. Bohacek, C. McMartin, and W. C. Guida, "The art and practice of structure-based drug design: a molecular modeling perspective", *in*: *Med Res Rev* 16.*1* (Jan. 1996), pp. 3–50.

[2] Wired, *Humans have found or made 50 million different chemicals here on Earth*, 2009, URL: `http://www.wired.com/2009/09/humans-have-made-found-or-used-over-50-million-unique-chemicals/`.

[3] American Chemical Society, *CAS REGISTRY - The gold standard for chemical substance information*, 2014, URL: `http://www.cas.org/content/chemical-substances`.

[4] B.R. Stockwell, "Chemical genetics: ligand-based discovery of gene function", *in*: *Nature reviews. Genetics* (2000).

[5] C.M. Dobson, "Chemical space and biology", *in*: *Nature* (2004).

[6] E. Fahy et al., "Update of the LIPID MAPS comprehensive classification system for lipids", *in*: *Journal of lipid research* (2009).

[7] S. Subramaniam et al., "Bioinformatics and systems biology of the lipidome", *in*: *Chemical reviews* (2011).

[8] S.L. Schreiber, "Target-oriented and diversity-oriented organic synthesis in drug discovery", *in*: *Science (New York, N.Y.)* (2000).

[9] B.R. Stockwell, "Frontiers in chemical genetics", *in*: *Trends in biotechnology* (2000).

[10] S. Brenner, "The genetics of Caenorhabditis elegans", *in*: *Genetics* (1974).

[11] S. Legg England and K.G. Götestam, "The nature and treatment of excessive gambling", *in*: *Acta psychiatrica Scandinavica* (1991).

[12] B.R. Stockwell, "Chemical genetic screening approaches to neurobiology", *in*: *Neuron* (2002).

[13] Travis AS., "Perkin's mauve: ancestor of the organic chemical industry", *in*: *Technol Cult* (1990).

[14] J. Parascandola, "The theoretical basis of Paul Ehrlich's chemotherapy", *in*: *Journal of the history of medicine and allied sciences* (1981).

[15] A.S. Travis, "Science as receptor of technology: Paul Ehrlich and the synthetic dyestuffs industry", *in*: *Science in context* (1989).

[16] R. E. Dolle and K. H. Nelson, "Comprehensive survey of combinatorial library synthesis: 1998", *in*: *J Comb Chem* 1.*4* (1999), pp. 235–282.

[17] R. E. Dolle, "Comprehensive survey of chemical libraries yielding enzyme inhibitors, receptor agonists and antagonists, and other biologically active agents: 1992 through 1997", *in*: *Mol. Divers.* 3.*4* (1997), pp. 199–233.

[18] S.A. Sundberg, "High-throughput and ultra-high-throughput screening: solution- and cell-based approaches", *in*: *Current opinion in biotechnology* (2000).

[19] E. Voit, A.R. Neves, and H. Santos, "The intricate side of systems biology", *in*: *Proceedings of the National Academy of Sciences of the United States of America* (2006).

[20] C. Harrison, "GlaxoSmithKline opens the door on clinical data sharing", *in*: *Nature reviews. Drug discovery* (2012).

[21] P. Nisen and F. Rockhold, "Access to patient-level data from GlaxoSmithKline clinical trials", *in*: *The New England journal of medicine* (2013).

[22] H.G. Roider et al., "Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network", *in*: *BMC bioinformatics* (2014).

[23] M. E. Weiss, P. Ryan, and L. Lokken, "Validity and reliability of the Perceived Readiness for Discharge After Birth Scale", *in*: *J Obstet Gynecol Neonatal Nurs* 35.*1* (2006), pp. 34–45.

[24] N. Winssinger et al., "Profiling protein function with small molecule microarrays", *in*: *Proceedings of the National Academy of Sciences of the United States of America* (2002).

[25] G. MacBeath, "Protein microarrays and proteomics", *in*: *Nature genetics* (2002).

[26] A. Espejo et al., "A protein-domain microarray identifies novel protein-protein interactions", *in*: *The Biochemical journal* (2002).

[27] O. Gallego et al., "A systematic screen for protein-lipid interactions in Saccharomyces cerevisiae", *in*: *Molecular systems biology* (2010).

[28] A.E. Saliba et al., "A quantitative liposome microarray to systematically characterize protein-lipid interactions", *in*: *Nature methods* (2014).

[29] Galperin MY Koonin EV, *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics, Chapter 4*, Boston: Kluwer Academic, 2003.

[30] M. Iskar et al., "Drug discovery in the age of systems biology: the rise of computational approaches for data integration", *in*: *Current opinion in biotechnology* (2012).

[31] Y. Wang et al., "PubChem: a public information system for analyzing bioactivities of small molecules", *in*: *Nucleic acids research* (2009).

[32] M. Prendergast et al., "Chorioamnionitis, lung function and bronchopulmonary dysplasia in prematurely born infants", *in*: *Arch. Dis. Child. Fetal Neonatal Ed.* 96.*4* (July 2011), F270–274.

[33] J. Rihel et al., "Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation", *in*: *Science (New York, N.Y.)* (2010).

[34] A. F. Fliri et al., "Analysis of drug-induced effect patterns to link structure and side effects of medicines", *in*: *Nat. Chem. Biol.* 1.*7* (Dec. 2005), pp. 389–397.

[35] M. Campillos et al., "Drug target identification using side-effect similarity", *in*: *Science (New York, N.Y.)* (2008).

[36] M. Kuhn et al., "Systematic identification of proteins that elicit drug side effects", *in*: *Molecular systems biology* (2013).

[37] Antoine-Emmanuel Saliba. et al., "A protocol for the systematic and quantitative measurement of protein-lipid interaction using the liposome microarray-based assay (LiMA)", *in*: *In preparation for Nature Protocol* ().

[38] Ivana Vonkova*. et al., "Lipid cooperativity as a general membrane-recruitment principle of PH domains", *in*: *Science, in review* ().

[39] S. Deghou et al., "CART: an efficient and scalable Chemical Annotation Retrieval Toolkit", *in*: *In preparation for Nature Protocol* ().

[40] E. Fahy et al., "A comprehensive classification system for lipids", *in*: *Journal of lipid research* (2005).

[41] G. van Meer, "Cellular lipidomics", *in*: *The EMBO journal* (2005).

[42] T.G. Kutateladze, "Translation of the phosphoinositide code by PI effectors", *in*: *Nature chemical biology* (2010).

[43] M.N. Teruel and T. Meyer, "Translocation and reversible localization of signaling proteins: a dynamic future for signal transduction", *in*: *Cell* (2000).

[44] R.V. Stahelin, "Lipid binding domains: more than simple lipid effectors", *in*: *Journal of lipid research* (2009).

[45] M.A. Lemmon, "Membrane recognition by phospholipid-binding domains", *in*: *Nature reviews. Molecular cell biology* (2008).

[46] J.H. Hurley, "Membrane binding domains", *in*: *Biochimica et biophysica acta* (2006).

[47] J.G. Carlton and P.J. Cullen, "Coincidence detection in phosphoinositide signaling", *in*: *Trends in cell biology* (2005).

[48] M.A. Lemmon and K.M. Ferguson, "Signal-dependent membrane targeting by pleckstrin homology (PH) domains", *in*: *The Biochemical journal* (2000).

[49] M.A. Lemmon and K.M. Ferguson, "Molecular determinants in pleckstrin homology domains that allow specific recognition of phosphoinositides", *in*: *Biochemical Society transactions* (2001).

[50] M.A. Lemmon, K.M. Ferguson, and C.S. Abrams, "Pleckstrin homology domains and the cytoskeleton", *in*: *FEBS letters* (2002).

[51] R.J. Haslam, H.B. Koide, and B.A. Hemmings, "Pleckstrin domain homology", *in*: *Nature* (1993).

[52] B.J. Mayer et al., "A putative modular domain present in diverse signaling proteins", *in*: *Cell* (1993).

[53] E. S. Lander et al., "Initial sequencing and analysis of the human genome", *in*: *Nature* 409.*6822* (Feb. 2001), pp. 860–921.

[54] T.F. Franke et al., "Direct regulation of the Akt proto-oncogene product by phosphatidylinosito 3,4-bisphosphate", *in*: *Science (New York, N.Y.)* (1997).

[55] P. Garcia et al., "The pleckstrin homology domain of phospholipase C-delta 1 binds with high affinity to phosphatidylinositol 4,5-bisphosphate in bilayer membranes", *in*: *Biochemistry* (1995).

[56] J.E. Harlan et al., "Pleckstrin homology domains bind to phosphatidylinositol-4,5-bisphosphate", *in*: *Nature* (1994).

[57] J.K. Klarlund et al., "Signaling by phosphoinositide-3,4,5-trisphosphate through proteins containing pleckstrin and Sec7 homology domains", *in*: *Science (New York, N.Y.)* (1997).

[58] M.A. Lemmon et al., "Specific and high-affinity binding of inositol phosphates to an isolated pleckstrin homology domain", *in*: *Proceedings of the National Academy of Sciences of the United States of America* (1995).

[59] K. Salim et al., "Distinct specificity in the recognition of phosphoinositides by the pleckstrin homology domains of dynamin and Bruton's tyrosine kinase", *in*: *The EMBO journal* (1996).

[60] J. M. Kavran et al., "Specificity and promiscuity in phosphoinositide binding by pleckstrin homology domains", *in*: *J. Biol. Chem.* 273.*46* (Nov. 1998), pp. 30497–30508.

[61] L.E. Rameh et al., "A comparative analysis of the phosphoinositide binding specificity of pleckstrin homology domains", *in*: *The Journal of biological chemistry* (1997).

[62] J.W. Yu et al., "Genome-wide analysis of membrane targeting by S. cerevisiae pleckstrin homology domains", *in*: *Molecular cell* (2004).

[63] J.P. DiNitto, T.C. Cronin, and D.G. Lambright, "Membrane recognition and targeting by lipid-binding domains", *in*: *Science's STKE : signal transduction knowledge environment* (2003).

[64] K.M. Ferguson et al., "Structural basis for discrimination of 3-phosphoinositides by pleckstrin homology domains", *in*: *Molecular cell* (2000).

[65] S. E. Lietzke et al., "Structural basis of 3-phosphoinositide recognition by pleckstrin homology domains", *in*: *Mol. Cell* 6.*2* (Aug. 2000), pp. 385–394.

[66] C.C. Thomas et al., "High-resolution structure of the pleckstrin homology domain of protein kinase b/akt bound to phosphatidylinositol (3,4,5)-trisphosphate", *in*: *Current biology : CB* (2002).

[67] K.M. Ferguson et al., "Structure of the high affinity complex of inositol trisphosphate with a phospholipase C pleckstrin homology domain", *in*: *Cell* (1995).

[68] J. A. Pitcher et al., "Pleckstrin homology domain-mediated membrane association and activation of the beta-adrenergic receptor kinase requires coordinate interaction with G beta gamma subunits and lipid", *in*: *J. Biol. Chem.* 270.*20* (May 1995), pp. 11707–11710.

[69] B.X. Huang et al., "Phosphatidylserine is a critical modulator for Akt activation", *in*: *The Journal of cell biology* (2011).

[70] N. Lucas and W. Cho, "Phosphatidylserine binding is essential for plasma membrane recruitment and signaling function of 3-phosphoinositide-dependent kinase-1", *in*: *The Journal of biological chemistry* (2011).

[71] D.F. Ceccarelli et al., "Non-canonical interaction of phosphoinositides with pleckstrin homology domains of Tiam1 and ArhGAP9", *in*: *The Journal of biological chemistry* (2007).

[72] J.R. Bayascas et al., "Mutation of the PDK1 PH domain inhibits protein kinase B/Akt, leading to small size and insulin resistance", *in*: *Molecular and cellular biology* (2008).

[73] J.D. Carpten et al., "A transforming mutation in the pleckstrin homology domain of AKT1 in cancer", *in*: *Nature* (2007).

[74] K. Hussain et al., "An activating mutation of AKT2 and human hypoglycemia", *in*: *Science (New York, N.Y.)* (2011).

[75] M.J. Lindhurst et al., "A mosaic activating mutation in AKT1 associated with the Proteus syndrome", *in*: *The New England journal of medicine* (2011).

[76] D.J. Rawlings et al., "Mutation of unique region of Bruton's tyrosine kinase in immunodeficient XID mice", *in*: *Science (New York, N.Y.)* (1993).

[77] J.D. Thomas et al., "Colocalization of X-linked agammaglobulinemia and X-linked immunodeficiency genes", *in*: *Science (New York, N.Y.)* (1993).

[78] S.J. Isakoff et al., "Identification and analysis of PH domain-containing targets of phosphatidylinositol 3-kinase using a novel in vivo assay in yeast", *in*: *The EMBO journal* (1998).

[79] H. Zhu et al., "Global analysis of protein activities using proteome chips", *in*: *Science* 293.*5537* (Sept. 2001), pp. 2101–2105.

[80] P. S. Aguilar et al., "A plasma-membrane E-MAP reveals links of the eisosome with sphingolipid metabolism and endosomal trafficking", *in*: *Nat. Struct. Mol. Biol.* 17.*7* (July 2010), pp. 901–908.

[81] P. Zhang et al., "Proteomic identification of phosphatidylinositol (3,4,5) triphosphate-binding proteins in Dictyostelium discoideum", *in*: *Proc. Natl. Acad. Sci. U.S.A.* 107.*26* (June 2010), pp. 11829–11834.

[82] P. Walde et al., "Giant vesicles: preparations and applications", *in*: *Chembiochem* 11.*7* (May 2010), pp. 848–865.

[83] A. Jesorka and O. Orwar, "Liposomes: technologies and analytical applications", *in*: *Annu Rev Anal Chem (Palo Alto Calif)* 1 (2008), pp. 801–832.

[84] K. S. Horger et al., "Films of agarose enable rapid formation of giant liposomes in solutions of physiologic ionic strength", *in*: *J. Am. Chem. Soc.* 131.*5* (Feb. 2009), pp. 1810–1819.

[85] B. Neumann et al., "High-throughput RNAi screening by time-lapse imaging of live human cells", *in*: *Nat. Methods* 3.*5* (May 2006), pp. 385–390.

[86] M. Frech et al., "High affinity binding of inositol phosphates and phosphoinositides to the pleckstrin homology domain of RAC/protein kinase B and their influence on kinase activity", *in*: *J. Biol. Chem.* 272.*13* (Mar. 1997), pp. 8474–8481.

[87] M. Tartaglia et al., "Gain-of-function SOS1 mutations cause a distinctive form of Noonan syndrome", *in*: *Nat. Genet.* 39.*1* (Jan. 2007), pp. 75–79.

[88] K. K. Yadav and D. Bar-Sagi, "Allosteric gating of Son of sevenless activity by the histone domain", *in*: *Proc. Natl. Acad. Sci. U.S.A.* 107.*8* (Feb. 2010), pp. 3436–3440.

[89] G. van den Bogaart et al., "Membrane protein sequestering by ionic protein-lipid interactions", *in*: *Nature* 479.*7374* (Nov. 2011), pp. 552–555.

[90] E. Boura and J. H. Hurley, "Structural basis for membrane targeting by the MVB12-associated Î²-prism domain of the human ESCRT-I MVB12 subunit", *in*: *Proc. Natl. Acad. Sci. U.S.A.* 109.*6* (Feb. 2012), pp. 1901–1906.

[91] K. Moravcevic et al., "Kinase associated-1 domains drive MARK/PAR1 kinases to membrane targets by binding acidic phospholipids", *in*: *Cell* 143.*6* (Dec. 2010), pp. 966–977.

[92] S. Amlacher et al., "Insight into structure and assembly of the nuclear pore complex by utilizing the genome of a eukaryotic thermophile", *in*: *Cell* 146.*2* (July 2011), pp. 277–289.

[93] M. Suárez-Fariñas et al., "Harshlight: a "corrective make-up" program for microarray chips", *in*: *BMC bioinformatics* (2005).

[94] P. Neuvial et al., "Spatial normalization of array-CGH data", *in*: *BMC bioinformatics* (2006).

[95] J.S. Song et al., "Microarray blob-defect removal improves array analysis", *in*: *Bioinformatics (Oxford, England)* (2007).

[96] Y.H. Yang et al., "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation", *in*: *Nucleic acids research* (2002).

[97] G. Rigaill et al., "ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays", *in*: *Bioinformatics (Oxford, England)* (2008).

[98] J.M. Cairns et al., "BASH: a tool for managing BeadArray spatial artefacts", *in*: *Bioinformatics (Oxford, England)* (2008).

[99] A. Koren, I. Tirosh, and N. Barkai, "Autocorrelation analysis reveals widespread spatial biases in microarray experiments", *in*: *BMC Genomics* 8 (2007), p. 164.

[100] T. Sing et al., "ROCR: visualizing classifier performance in R", *in*: *Bioinformatics* 21.*20* (Oct. 2005), pp. 3940–3941.

[101] B. W. Brandt, K. A. Feenstra, and J. Heringa, "Multi-Harmony: detecting functional specificity from sequence alignment", *in*: *Nucleic Acids Res.* 38.*Web Server issue* (July 2010), pp. 35–40.

[102] K. Moravcevic, C.L. Oxley, and M.A. Lemmon, "Conditional peripheral membrane proteins: facing up to limited specificity", *in*: *Structure (London, England : 1993)* (2012).

[103] M. Magrane and U. Consortium, "UniProt Knowledgebase: a hub of integrated protein data", *in*: *Database (Oxford)* 2011 (2011), bar009.

[104] G. D'Angelo et al., "Glycosphingolipid synthesis requires FAPP2 transfer of glucosylceramide", *in*: *Nature* (2007).

[105] K. Hanada et al., "Molecular machinery for non-vesicular trafficking of ceramide", *in*: *Nature* (2003).

[106] K. Maeda et al., "Interactome map uncovers phosphatidylserine transport by oxysterol-binding proteins", *in*: *Nature* (2013).

[107] T.A. Schulz et al., "Lipid-regulated sterol transfer between closely apposed membranes by oxysterol-binding protein homologues", *in*: *The Journal of cell biology* (2009).

[108] A. Godi et al., "FAPPs control Golgi-to-cell-surface membrane traffic by binding to ARF and PtdIns(4)P", *in*: *Nature cell biology* (2004).

[109] T.P. Levine and S. Munro, "Dual targeting of Osh1p, a yeast homologue of oxysterol-binding protein, to both the Golgi and the nucleus-vacuole junction", *in*: *Molecular biology of the cell* (2001).

[110] K. Tomczak, P. Czerwi?ska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge", *in*: *Contemp Oncol (Pozn)* 19.*1A* (2015), pp. 68–77.

[111] M. Imielinski et al., "Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing", *in*: *Cell* 150.*6* (Sept. 2012), pp. 1107–1120.

[112] W. S. Park et al., "Comprehensive identification of PIP3-regulated PH domains from C. elegans to H. sapiens by model prediction and live imaging", *in*: *Mol. Cell* 30.*3* (May 2008), pp. 381–392.

[113] K.E. Landgraf, C. Pilling, and J.J. Falke, "Molecular mechanism of an oncogenic mutation that alters membrane targeting: Glu17Lys modifies the PIP lipid specificity of the AKT1 PH domain", *in*: *Biochemistry* (2008).

[114] R.C. Aguilar et al., "Epsin N-terminal homology domains perform an essential function regulating Cdc42 through binding Cdc42 GTPase-activating proteins", *in*: *Proceedings of the National Academy of Sciences of the United States of America* (2006).

[115] J.W. Yu and M.A. Lemmon, "All phox homology (PX) domains from Saccharomyces cerevisiae specifically recognize phosphatidylinositol 3-phosphate", *in*: *The Journal of biological chemistry* (2001).

[116] G.D. Fairn et al., "Phosphatidylserine is polarized and required for proper Cdc42 localization and for development of cell polarity", *in*: *Nature cell biology* (2011).

[117] S. Etienne-Manneville, "Cdc42–the centre of polarity", *in*: *Journal of cell science* (2004).

[118] D. Szklarczyk et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored", *in*: *Nucleic acids research* (2011).

[119] R. Nash et al., "Expanded protein information at SGD: new pages and proteome browser", *in*: *Nucleic acids research* (2007).

[120] C.A. Barlow, R.S. Laishram, and R.A. Anderson, "Nuclear phosphoinositides: a signaling enigma wrapped in a compartmental conundrum", *in*: *Trends in cell biology* (2010).

[121] A. E. Lewis et al., "Identification of nuclear phosphatidylinositol 4,5-bisphosphate-interacting proteins by neomycin extraction", *in*: *Mol. Cell Proteomics* 10.*2* (Feb. 2011), p. M110.003376.

[122] N.C. Lucki and M.B. Sewer, "Nuclear sphingolipid metabolism", *in*: *Annual review of physiology* (2012).

[123] K. Viiri, M. Mäki, and O. Lohi, "Phosphoinositides as regulators of protein-chromatin interactions", *in*: *Science signaling* (2012).

[124] A. Bateman et al., "The Pfam protein families database", *in*: *Nucleic Acids Res.* 30.*1* (Jan. 2002), pp. 276–280.

[125] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server", *in*: *Bioinformatics* 16.*4* (Apr. 2000), pp. 404–405.

[126] J. Schultz et al., "SMART: a web-based tool for the study of genetically mobile domains", *in*: *Nucleic Acids Res.* 28.*1* (Jan. 2000), pp. 231–234.

[127] M. R. Lamprecht, D. M. Sabatini, and A. E. Carpenter, "CellProfiler: free, versatile software for automated biological image analysis", *in*: *BioTechniques* 42.*1* (Jan. 2007), pp. 71–75.

[128] R Core Team, *R: A Language and Environment for Statistical Computing*, ISBN 3-900051-07-0, R Foundation for Statistical Computing, Vienna, Austria, 2013, URL: `http://www.R-project.org/`.

[129] Martin Maechler et al., *cluster: Cluster Analysis Basics and Extensions*, R package version 1.14.3 — For new features, see the 'Changelog' file (in the package source), 2012.

[130] C. Mosquera et al., "[Preventing the recurrence of febrile seizures: intermittent prevention with rectal diazepam compared with continuous treatment with sodium valproate]", *in*: *An. Esp. Pediatr.* 27.*5* (Nov. 1987), pp. 379–381.

[131] S. Dowler et al., "Identification of pleckstrin-homology-domain-containing proteins with novel phosphoinositide-binding specificities", *in*: *The Biochemical journal* (2000).

[132] G. van Meer, D.R. Voelker, and G.W. Feigenson, "Membrane lipids: where they are and how they behave", *in*: *Nature reviews. Molecular cell biology* (2008).

[133] G. Di Paolo and P. De Camilli, "Phosphoinositides in cell regulation and membrane dynamics", *in*: *Nature* (2006).

[134] M. Nakamura-Kubo et al., "The fission yeast pleckstrin homology domain protein Spo7 is essential for initiation of forespore membrane assembly and spore morphogenesis", *in*: *Molecular biology of the cell* (2011).

[135] A. Roy and T.P. Levine, "Multiple pools of phosphatidylinositol 4-phosphate detected using the pleckstrin homology domain of Osh2p", *in*: *The Journal of biological chemistry* (2004).

[136] T. Sugiki et al., "Structural basis for the Golgi association by the pleckstrin homology domain of the ceramide trafficking protein (CERT)", *in*: *The Journal of biological chemistry* (2012).

[137] K. Yamada et al., "Identification and characterization of splicing variants of PLEKHA5 (Plekha5) during brain development", *in*: *Gene* (2012).

[138] J. He et al., "Molecular basis of phosphatidylinositol 4-phosphate and ARF1 GTPase recognition by the FAPP1 pleckstrin homology (PH) domain", *in*: *The Journal of biological chemistry* (2011).

[139] S. Lev, "Non-vesicular lipid transport by lipid-transfer proteins and beyond", *in*: *Nature reviews. Molecular cell biology* (2010).

[140] J. D. Knight and J. J. Falke, "Single-molecule fluorescence studies of a PH domain: new insights into the membrane docking reaction", *in*: *Biophys. J.* 96.*2* (Jan. 2009), pp. 566–582.

[141] C.L. Lai et al., "Molecular mechanism of membrane binding of the GRP1 PH domain", *in*: *Journal of molecular biology* (2013).

[142] W.K. Huh et al., "Global analysis of protein localization in budding yeast", *in*: *Nature* (2003).

[143] M. Fadri et al., "The pleckstrin homology domain proteins Slm1 and Slm2 are required for actin cytoskeleton organization in yeast and bind phosphatidylinositol-4,5-bisphosphate and TORC2", *in*: *Molecular biology of the cell* (2005).

[144] R.W. Ledeen and G. Wu, "Nuclear sphingolipids: metabolism and signaling", *in*: *Journal of lipid research* (2008).

[145] Z.H. Shah et al., "Nuclear phosphoinositides and their impact on nuclear functions", *in*: *The FEBS journal* (2013).

[146] K. Segers et al., "Design of protein membrane interaction inhibitors by virtual ligand screening, proof of concept with the C2 domain of factor V", *in*: *Proceedings of the National Academy of Sciences of the United States of America* (2007).

[147] P.M. Blumberg et al., "Wealth of opportunity - the C1 domain as a target for drug development", *in*: *Current drug targets* (2008).

[148] L.C. Cantley, "The phosphoinositide 3-kinase pathway", *in*: *Science (New York, N.Y.)* (2002).

[149] D.O. Carpenter, K. Arcaro, and D.C. Spink, "Understanding the human health effects of chemical mixtures", *in*: *Environmental health perspectives* (2002).

[150] A. Gaulton and J. P. Overington, "Role of open chemical data in aiding drug discovery and design", *in*: *Future Med Chem* 2.*6* (June 2010), pp. 903–907.

[151] S. Orchard et al., "Minimum information about a bioactive entity (MIABE)", *in*: *Nat Rev Drug Discov* 10.*9* (Sept. 2011), pp. 661–669.

[152] S. O. Hansson and C. Ruden, "Priority setting in the REACH system", *in*: *Toxicol. Sci.* 90.*2* (Apr. 2006), pp. 304–308.

[153] Y. Wang et al., "PubChem's BioAssay Database", *in*: *Nucleic acids research* (2012).

[154] M. M. Savitski et al., "Tracking cancer drugs in living cells by thermal profiling of the proteome", *in*: *Science* 346.*6205* (Oct. 2014), p. 1255784.

[155] Y. Tang et al., "Differential determinants of cancer cell insensitivity to antimitotic drugs discriminated by a one-step cell imaging assay", *in*: *Journal of biomolecular screening* (2013).

[156] U.D. Vempati et al., "Metadata Standard and Data Exchange Specifications to Describe, Model, and Integrate Complex and Diverse High-Throughput Screening Data from the Library of Integrated Network-based Cellular Signatures (LINCS)", *in*: *Journal of biomolecular screening* (2014).

[157] M. Ashburner et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium", *in*: *Nature genetics* (2000).

[158] d.a. .W. Huang, B.T. Sherman, and R.A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources", *in*: *Nature protocols* (2009).

[159] J. H. Chen et al., "ChemDB update–full-text search and virtual chemical space", *in*: *Bioinformatics* 23.*17* (Sept. 2007), pp. 2348–2351.

[160] M. Kuhn et al., "STITCH 4: integration of protein-chemical interactions with user data", *in*: *Nucleic Acids Res.* 42.*Database issue* (Jan. 2014), pp. D401–407.

[161] F. Zhu et al., "Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery", *in*: *Nucleic Acids Res.* 40.*Database issue* (Jan. 2012), pp. D1128–1136.

[162] V. Law et al., "DrugBank 4.0: shedding new light on drug metabolism", *in*: *Nucleic Acids Res.* 42.*Database issue* (Jan. 2014), pp. D1091–1097.

[163] S. Croset, J. P. Overington, and D. Rebholz-Schuhmann, "The functional therapeutic chemical classification system", *in*: *Bioinformatics* 30.*6* (Mar. 2014), pp. 876–883.

[164] P. de Matos et al., "Chemical Entities of Biological Interest: an update", *in*: *Nucleic Acids Res.* 38.*Database issue* (Jan. 2010), pp. D249–254.

[165] M. Kuhn et al., "A side effect resource to capture phenotypic effects of drugs", *in*: *Mol. Syst. Biol.* 6 (2010), p. 343.

[166] B. Ganter et al., "Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action", *in*: *J. Biotechnol.* 119.*3* (Sept. 2005), pp. 219–244.

[167] K. Ono, B. Demchak, and T. Ideker, "Cytoscape tools for the web age: D3.js and Cytoscape.js exporters", *in*: *F1000Res* 3 (2014), p. 143.

[168] C. Liu et al., "Compound signature detection on LINCS L1000 big data", *in*: *Mol Biosyst* 11.*3* (Mar. 2015), pp. 714–722.

[169] E. A. Welsh et al., "Iterative rank-order normalization of gene expression microarray data", *in*: *BMC Bioinformatics* 14 (2013), p. 153.