



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

A Systems Biological Approach to Parkinson's Disease

Katharina Friedlinde Heil



Doctor of Philosophy

Institute for Adaptive and Neural Computation

School of Informatics

University of Edinburgh

2017

Abstract

Parkinson's Disease (PD) is the second most common neurodegenerative disease in the Western world. It shows a high degree of genetic and phenotypic complexity with many implicated factors, various disease manifestations but few clear causal links. Ongoing research has identified a growing number of molecular alterations linked to the disease.

Dopaminergic neurons in the substantia nigra, specifically their synapses, are the key-affected region in PD. Therefore, this work focuses on understanding the disease effects on the synapse, aiming to identify potential genetic triggers and synaptic PD associated mechanisms. Currently, one of the main challenges in this area is data quality and accessibility.

In order to study PD, publicly available data were systematically retrieved and analysed. 418 PD associated genes could be identified, based on mutations and curated annotations. I curated an up-to-date and complete synaptic proteome map containing a total of 6,706 proteins. Region specific datasets describing the presynapse, postsynapse and synaptosome were also delimited. These datasets were analysed, investigating similarities and differences, including reproducibility and functional interpretations.

The use of Protein-Protein-Interaction Network (PPIN) analysis was chosen to gain deeper knowledge regarding specific effects of PD on the synapse. Thus I generated a customised, filtered, human specific Protein-Protein Interaction (PPI) dataset, containing 211,824 direct interactions, from four public databases. Proteomics data and PPI information allowed the construction of PPINs. These were analysed and a set of low level statistics, including modularity, clustering coefficient and node degree, explaining the network's topology from a mathematical point of view were obtained.

Apart from low-level network statistics, high-level topology of the PPINs was studied. To identify functional network subgroups, different clustering algorithms were investigated. In the context of biological networks, the underlying hypothesis is that proteins in a structural community are more likely to share common functions. Therefore I attempted to identify PD enriched communities of synaptic proteins. Once identified, they were compared amongst each other. Three community clusters could be identified as containing largely overlapping gene sets. These contain 24 PD associated genes. Apart from the known disease associated genes in these communities, a total of 322 genes was identified. Each of the three clusters is specifically enriched for specific biological processes and cellular components, which include neurotransmitter secretion, positive regulation of synapse assembly, pre- and post-synaptic membrane,

scaffolding proteins, neuromuscular junction development and complement activation (classical pathway) amongst others.

The presented approach combined a curated set of PD associated genes, filtered PPI information and synaptic proteomes. Various small- and large-scale analytical approaches, including PPIN topology analysis, clustering algorithms and enrichment studies identified highly PD affected synaptic proteins and subregions. Specific disease associated functions confirmed known research insights and allowed me to propose a new list of so far unknown potential disease associated genes. Due to the open design, this approach can be used to answer similar research questions regarding other complex diseases amongst others.

Lay Summary

Parkinson's Disease is a brain disease with extreme consequences for patients, their families and carers. Treatment only moderates the symptoms and the number of patients is growing on a daily basis.

Many research projects identified dysfunctioning intracellular processes mainly located in a specific part of brain cells. This part, the synapse, is in charge of transporting information and its dysfunction leads to known disease symptoms such as tremor, shuffling gait and less known non-motor symptoms such as enhanced sweating. Parkinson's can currently only be diagnosed at a stage when brain-cells are dying, making it very hard to treat the disease effectively. Another challenge are the very individual symptoms the disease provokes in patients. A number of dysfunctions are known to appear in the brain cells of patients, but not all of them can be found in all individuals.

Therefore this thesis aims towards gaining better understanding of specific disease causes. New knowledge could then help to develop better treatment or even a disease cure. To work towards this aim different systems biological analytical steps were carried out. 418 genes which have shown to be affected in Parkinson's Disease patients were identified. The synapse was analysed and around 6,500 genes were identified in this brain-cell region.

To understand the disease influence on the synapse, so called large-scale approaches are required. Protein-Protein-Interaction Networks were used to analyse how proteins interact and allow to identify gene groups which are in charge of specific synaptic functions. Parkinson's Disease associated genes could be located in the network. By doing so three gene groups with an unexpected, significantly high number of disease associated genes were identified. Apart from the disease genes these contained a set of other genes which were analysed in-depth. It was possible to determine their overall function which is affected under disease conditions. Amongst others the release of neurotransmitters, the main component of information exchange between brain cells as well as structural aspects, guaranteeing protein interactions and their full functionality could be identified.

The set of around 150 specific genes can now be used to i) set up more targeted experiments, ii) help to identify different disease types and iii) develop new treatments. Overall it would not have been possible to obtain these results without the use of large-scale analytical approaches. Hence this work highlights their potential and promising application in the research field of complex diseases.

Acknowledgements

It's been four years since I embarked on this journey and many things have happened, changed, moved forward, and developed. It was not always easy and often felt like a ride on a roller-coaster - during a day, a week, a month, and just generally over the whole four years. Many moments and experiences helped me to keep going, but it was really the people I encountered on this journey who helped the most. Some accompanied me over the entire time, others in specific moments, and everybody has left an imprint on my PhD experience and shaped me - to be the person I am today. Therefore I owe all of you a big, big thank you!

Said that, there are a number of people I would like to thank in particular. **Douglas Armstrong**, you've been my supervisor at the University of Edinburgh. Your vision of the scientific future encouraged and motivated me. Thinking about the potential impact my/our research can have has helped me through many moments of struggle, specifically, when the "simple things" proved to be more challenging than expected. **Oksana Sorokina**, you have accompanied me, especially through the first years of my studies. Thank you for the introduction to network biology and kappa modelling. **Colin Mclean**, our chats about the concepts and maths behind networks were interesting and helpful and I am also specifically grateful about all your advice during the writing up stage. It has been great to get access to some of your code, learn from it, and use it to answer similar, but different, research questions. Your patience was highly appreciated and helped me to grow my knowledge in theoretical network analysis. Thank you **David Sterratt** for letting me work with you over many, many months. Having a joint project showed me how active exchange can help to solve problems, advance together and motivate each other. I am grateful to have been part of the "review-paper team" and look forward to the day when we submit the work for publication. **Emilia Wysocka**, you have also been part of that team and your critical thoughts and questions often helped me to rethink concepts that I might have taken for granted. Our discussions highly stimulated my thinking process and I hope that I did not distract you too often from your project while asking questions that were slightly off track. **Grant Robertson**, it has been very valuable to share my research with you. I've been uplifted by the constructive conversations that could often shed light over my questions. It was also good to see, that work was not always completely focused on my research topic.

I am also very thankful to my second supervisor, **Jeanette Hellgren Kotaleski** at KTH in Stockholm, who, together with **Olivia Eriksson**, always found time to listen and give helpful advice. It was great to be welcomed by you and the group during my

visits at KTH and SciLifeLab - two places that got to be my second research home. Getting the opportunity to present my work during your group meetings was extremely helpful to put things into perspective and confirm or slightly redefine the main focus of my project.

Overall I am very grateful for all the time, patience and positive words you, my supervisors and research colleagues, shared with me during the last four years! Your engagement made it possible for me to grow and work on my own ideas. I wish you all the best for the future.

It has been a great experience to spend my time as a PhD student at several universities. Thanks to the Erasmus Mundus EuroSPIN programme for making this possible. The international connections were an enriching experience for neuroinformatics research. EuroSPIN also offered a great environment and I would like to thank all my fellow students, who have supported me through many moments of my PhD. Without you I would have missed all the lunch-conversations in Stockholm, our self-organised workshop at NCBS in Bangalore and lots of in-depth insights into many aspects of the brain and memory research. I am convinced, that great careers are waiting for you and I would love to see our paths cross again in the future.

The Neuroinformatics DTC was another great group with open, enthusiastic and inspiring people. Additional travel support allowed me to attend summer schools, workshops and conferences. These played a substantial role in my scientific development and often remind me of the importance of my research project and the impact it can have on Parkinson's Disease patients' lives. Thanks to **Marc van Rossum**, **Alison Edie** and the whole **DTC cohort** for support, encouragement, and an open ear, as well as lots of motivation.

And there was another part that made my PhD experience unique. With my background in biomedical sciences I never thought it possible to be able to play an active role as a mentor or tutor in the School of Informatics. Thanks to **Paul Ardin**, who somehow knew what I was looking for, I got the chance to move all the way up to be a Teaching Assistant. **Jane Hillston**, **Henry Thompson**, and **Garry Ellard**, it was specifically through you that I gained so much insight into the "back-end" of university teaching. I am still inspired by your dedication to SDP and the way you engage with students to give them the best possible learning experience. Being inspired and challenged at the same time is definitely an experience that I will take away with me and I am very happy to do so.

I am very glad that I took on a role as a student representative. I was lucky to

be introduced as a student member of the Research Training Committee of the College of Science and Engineering, which later on opened me the doors to the Research Experience Committee. Seeing that things can change if they are well addressed and represented encouraged me to remain active in those roles until the very end of my PhD. Thanks to the student cohort to trust me as their representative and thanks to the committee members to value the students' input and react upon our ideas and thoughts. It's been a great contribution to my personal development, adjacent to the purely scientific experiences.

But as mentioned in the beginning it wasn't only the university environment that supported and encouraged me on the way.

Nathalie Dupuy - I am quite certain that we have spend at least half of our total time in Edinburgh together - mostly in the office, our office! Thank you for being who you are! I don't think that I would have made it all the way until today without our coffees (and espresso martinis), your smiles, hugs, stories and support - in any possible way! Merci pour tout! I am already looking forward to many more moments together!

Often the Informatics Forum felt like my real Edinburgh home - especially office 2.53. A big thank you goes to a number of people who became great companions and friends: **Martino, Nathalie, Emilia, Paul, Xin, Alba, Jinli, Maciej** (thanks for the proof-reading), **Valentina, Sahar, Aga, David, Emma** (thanks for the proof-reading), **Balazs, Aleks, Irene, Joe, Jon, Ksenia, Marzena, Miha, Scott, Victoria, Wioleta** and many more. Thank you for listening, cheering me up and making life in the Forum so much more fun - it has been a pleasure to share many moments, conversations and lunches with you.

I would like to thank the **CSE** for keeping my body in shape and more importantly my mind healthy. Your staff and programme are irreplaceable - thanks for all the hours of sweat and fun!

Capital Communicators Toastmasters - you taught me how to present in a clear and entertaining, professional way - and not only my research. I loved all the moments during my journey with you - it's been fun and inspiring, and you are a great group of people to be part of, whom I will deeply miss. **Renée** you've been a great mentor - thank you for all your support. It is an honour to learn from you and your experiences - I wish you the best of luck for your next adventures.

Finally there are more people that supported me throughout the roller-coaster ride of my PhD. Thank you, **Julia**, for making 20 Carnegie Court our home and **Eleana** for coming back to Edinburgh and sharing many, many unforgettable moments and

whiskies with me. **Younghwa** and **Annie** (and the IAD) - thank you for all your support and the fun we had.

Ronja, Jonas, Vera, Charly, Rafa, Simone, and Lore, you always knew how to make me laugh and cheer me up. **Olga** and **Fer**, thank you for making my time in Stockholm even better. **Luisa** - sometimes I have a feeling that you know me better than I do - thanks for helping me to put things into perspective and always be there when needed. It's a pleasure to have you in my life!

Family and **Friends** - ihr wisst wer ihr seid: Danke, dass ihr mich von klein auf dazu ermuntert habt Fragen zu stellen. Ohne eure positiven Worte und Begeisterung wäre es um einiges schwieriger gewesen anzukommen - wo ich heute bin. Ihr habt mich geprägt und seid somit auch ein Teil von mir.

An dieser Stelle bleibt meine Familie, ganz besonderen **Juli** - danke für deine Ehrlichkeit, deine Abenteuer und Inspiration - ich bin stolz auf dich! **Mama** und **Papa** - danke - für eure unendliche Liebe, konstante Motivation und herzengute Unterstützung. I could write a never-ending list of compliments - but it is all summarised in a few words: thank you for believing in me, giving me all the support I needed, and sharing so many moments of this journey with me. I am looking forward towards many more to come.

And last but not least: **Salva** - muchas gracias por creer en mí - ayudándome a mantener el foco en mi proyecto de doctorado y explorar el mundo a mi lado. I am grateful for every supportive and reflective word and look forward to our joint adventures to come.

Declaration

I declare that this thesis was composed by myself and that the work contained herein is my own except where explicitly stated otherwise in the text. This work has only been submitted for a joint PhD degree between KTH, Stockholm and the University of Edinburgh within the regulations of the Erasmus Mundus EuroSPIN programme and for no other degree or professional qualification except as specified.

(Katharina Friedlinde Heil)

20th March 2018

Before starting to present my work I would like to share one of the phrases that accompanied me through much of my life and was highly important and motivating throughout my PhD.

In the words of Confucius, I encourage you all to believe in yourself and your endeavours:

“The journey is the destination”

“El camino se hace al andar”

“Der Weg ist das Ziel”

Contents

1	Introduction	1
1.1	Parkinson's Disease	1
1.1.1	Pathology	2
1.2	The Synapse	8
1.2.1	The Synaptic Proteome	9
1.3	Systems Biology	10
1.4	Protein-Protein-Interaction Networks	11
1.4.1	Statistical Network Analysis	12
1.4.2	Network Clustering	14
1.5	Functional Gene Set Analysis	15
1.5.1	Testing for Enrichment	16
1.5.2	Functional Annotations	17
1.5.3	topGO and topONTO	18
1.6	Objectives	20
2	Methods	23
2.1	General Programming	23
2.1.1	Venn Diagrams	23
2.1.2	Computing Environment	24
2.2	Annotations and Mappings	24
2.2.1	Mapping File Generation	25
2.3	Enrichment Analysis	26
2.3.1	Hypergeometric Testing	26
2.3.2	Gene Set Enrichment: topGO and topONTO	28
2.3.3	Multiple Testing Correction	29
2.4	Protein-Protein-Interaction Network Analysis	30
2.4.1	Network Clustering Algorithms	32

3	Finding a Parkinson’s Disease Core Dataset	35
3.1	Objective	35
3.2	Material	36
3.2.1	Data Types	37
3.2.2	Data Sources	37
3.3	Results	40
3.3.1	PD associated genes studied in literature	40
3.3.2	PD associated genes based on expression data	40
3.3.3	PD associated genes with genetic and/or manually curated ev- idence	43
3.3.4	Meta-analysis	49
3.3.5	Summary	51
3.4	Discussion	53
4	Protein-Protein Interaction Data	57
4.1	Objective	57
4.2	Introduction and Data Processing	58
4.2.1	Protein-Protein Interactions	58
4.2.2	Data Format	60
4.2.3	Databases	61
4.2.4	Data Curation	63
4.3	Results	65
4.3.1	Data Analysis and Cross-Comparison	65
4.3.2	The final, joint, human PPI dataset	72
4.4	Discussion	76
5	The Synaptic Proteome and Parkinson’s Disease	81
5.1	Objective	81
5.2	Introduction and Material	83
5.2.1	Proteomic Studies	83
5.3	Results	83
5.3.1	Synaptic Proteome Datasets	83
5.3.2	Protein Coverage and Data Consistency	87
5.3.3	Top Coverage Genes	92
5.3.4	Regional Synaptic Properties	96
5.3.5	PD and the Synapse	103

5.3.6	PD Affected Functions	106
5.4	Discussion	109
6	Synaptic Protein-Protein-Interaction Network Analysis and PD	113
6.1	Hypothesis and Objective	113
6.2	Material and Methods	114
6.3	Results	115
6.3.1	Synaptic Protein-Protein-Interaction Networks	115
6.3.2	Network Clustering	122
6.3.3	PD Enriched Communities	127
6.3.4	Synaptic PD Affected Functions	133
6.3.5	Summary	143
6.4	Discussion	143
7	Discussion	149
7.1	Data Consistency	149
7.2	Proteomic Datasets	151
7.3	Protein-Protein-Interaction Networks and PD	152
7.4	Systems Biology and PD Research	154
7.5	Synaptic Dysfunctions and PD	155
7.6	Future Research Perspectives	157
7.6.1	Clathrin Mediated Endocytosis - a Dynamic Model	159
7.6.2	Disease in Computational Models of Neurons	160
7.7	Conclusion	161
A	Literature based Parkinson's Disease associated genes	163
B	MI-IDs	165
C	Extended Overview of Synaptic Proteomic Studies	169
D	Additional Protein-Protein-Interaction Networks	173
E	Core PD associated gene sets	177
F	Enriched Gene Ontology terms in the top three PD enriched clusters	179
G	Clathrin Mediated Endocytosis - a Dynamic Model	187

H Disease in Synaptic Models	211
I Acronyms	249
List of Figures	251
List of Tables	255
Bibliography	261

Chapter 1

Introduction

1.1 Parkinson's Disease

Parkinson's Disease (PD) is the second most common neurodegenerative disease in the Western world (De Lau and Breteler, 2006) and its underlying causes are far from understood. Due to the growing improvement in the treatment of cancer and other lethal diseases, neuronal disease is becoming more prevalent and currently about ten million people worldwide suffer from the condition (European Parkinson's Disease Association¹). In the US 0.01% of the population under the age of 45 and 1.2 - 4.38% over the age of 65 are diagnosed with PD (numbers consider regional variability) (Kowal et al., 2013). Usually symptoms appear between the age of 62 and 70 (Muangpaisan et al., 2011). Apart from the impact on personal health and well being the estimated financial burden in the US in 2010 was around eight million USD medical cost directly attributed to PD and another 14 million social cost incurred by the PD affected population. Additionally about six million USD were associated with reduced employment, lost work days due to illness, formal care and others (Kowal et al., 2013). Numbers in Europe are expected to be proportionally similar.

As presented in the World Health Organisation report, "Neurological Disorders: Public Health Challenges"² (2006), one of the dangers associated with neurodegenerative disease is the lack of communicable conditions and diagnosis. The main known cause of PD and its symptoms is the degeneration of dopaminergic neurons in the substantia nigra pars compacta in the midbrain. The progressive degeneration remains

¹<http://www.epda.eu.com/>

²http://www.who.int/mental_health/neurology/neurological_disorders_report_web.pdf

largely unnoticed by the patient and can not yet be specifically detected. Hence diagnosis is only possible at a very advanced disease stage, when neurons are already irreversibly destroyed.

Recent evidence is accumulating and indicates that synapses play a key role in the degenerative process (Lüscher and Isaac, 2009). Neuronal connectivity, based on synapses, was identified to be fundamental for a healthy brain. Hence, the gradual loss of synapses and deteriorated synaptic plasticity precede neuronal dysfunction and cell death, implying neurodegeneration (Knight and Verkhatsky, 2010).

This leads to motor and non-motor symptoms. Motor dysfunctions include bradykinesia (decreased movement), rest tremor and rigidity. Non-motor functions are depression, cognitive impairment sleep disturbances and failure of cognitive abilities such as memory and decision making (Magrinelli et al., 2016). Overall, disease development and symptoms are very patient specific and depend highly on underlying causes. A cure is currently not available and medication only moderates and alleviates symptoms allowing for improved quality of life (Chen and Pan, 2014; Bredesen et al., 2006).

In order to find better treatment it is crucial to know disease causing dysfunctions and have a better disease understanding. The following section introduces known details about the PD pathology.

1.1.1 Pathology

PD is considered a complex disease, with a number of dysfunctions associated with it, all of which lead to the degeneration of dopaminergic neurons in the substantia nigra, pars compacta (Dexter and Jenner, 2013). Recent years allowed to identify more and more molecular alterations significantly associated with the development of PD. These can be found in different patients and distinct combinations.

The familial (inherited) form of PD (~10% of the cases) made it possible to identify genetic alterations associated with the disease (Spatola and Wider, 2014). These explain about 30% of the familial and between 3-5% of sporadically occurring PD cases (Klein and Westenberger, 2012). Even though these numbers seem relatively small they are a great source for research (Bonifati, 2014).

Additionally, a large number of non-genetic cases exist. These can occur due to random genetic variants or other molecular dysfunctions. Figure 1.1 shows the central dogma of molecular biology, indicating different molecular levels that can be affected and lead to disease manifestation. Part A shows different cellular units and part B

describes the union of those. Genetic modifications are reflected on the DNA. Changes in RNA expression or altered protein levels are noticed on exome, transcriptome or proteome level. The latter can also lead to disease manifestation but are more difficult to detect. Since alterations on one level are not always directly propagated to the next level (e.g. from genome to exome), RNA and protein level changes are not apparent from genomic information and require alternative detection approaches.

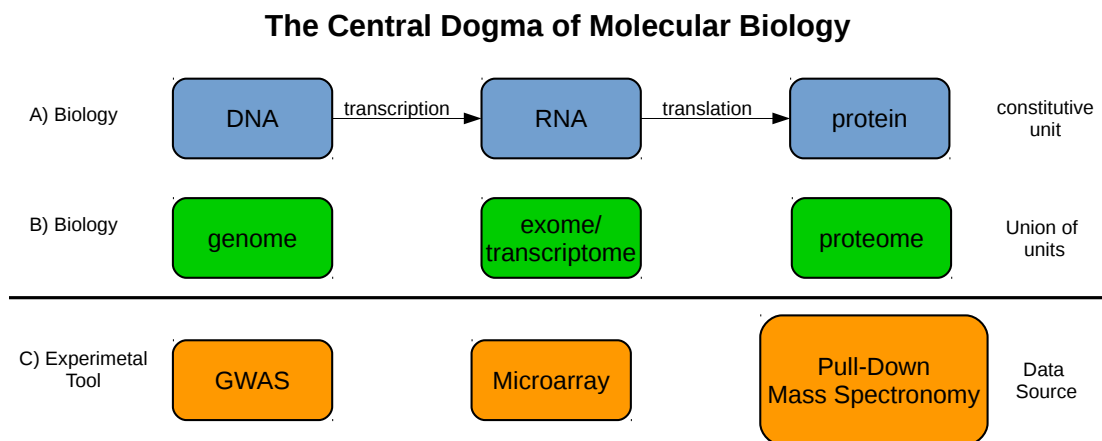


Figure 1.1: The Central Dogma of Molecular Biology. Individual constitutive units are visualized (part A) as well as the union of all the units (part B). Part C shows examples of experimental tools that can be used to study the different levels of information.

As indicated in Figure 1.1 C GWAS studies supply information regarding heritability in genetic regions (also not completely correctly referred to as mutations), addressing alterations in the genome. Microarray studies identify disease associated changes in the exome, transcriptome and proteome. These modifications can be detected with a combination of pull-down analysis and mass spectrometry amongst others.

Compared to more traditionally used techniques, these all cover a large part or all of the genome, transcriptome or proteome. This is specifically beneficial for uncovering unsuspected disease associated alterations without targeting them based on prior knowledge which was often the case in previously available, small-scale studies focusing on individual proteins.

To gain an overall insight of the disease, large-scale analysis of results covering all levels of the molecular machinery need to be considered and combined. Generally the number of large-scale studies is increasing but results are most frequently considered individually. Combining knowledge covering information describing distinct disease

aspects is necessary and crucial to shed light over unknown connections amongst dysfunctions and the complete disease picture.

1.1.1.1 Affected Subsystems and Pathways

Even though there has not been any large-scale combinatorial study systematically analysing similarities and differences between PD related results, individual studies identified a number of PD affected molecular functions, also referred to as pathways. All of them can contribute to the PD complexity. Based on current knowledge dysfunctions appear in different combinations, leading to the complex set of PD geno- and phenotypes (Thenganatt and Jankovic, 2014). A major effort has been made and a PD map³ was created, being under constant curation and expansion (Fujita et al., 2014). It presents a great source highlighting affected pathways and Figure 1.2 shows the published visualization of the interactive tool. Based on this overview and additional studies the following paragraphs briefly introduce affected systems.

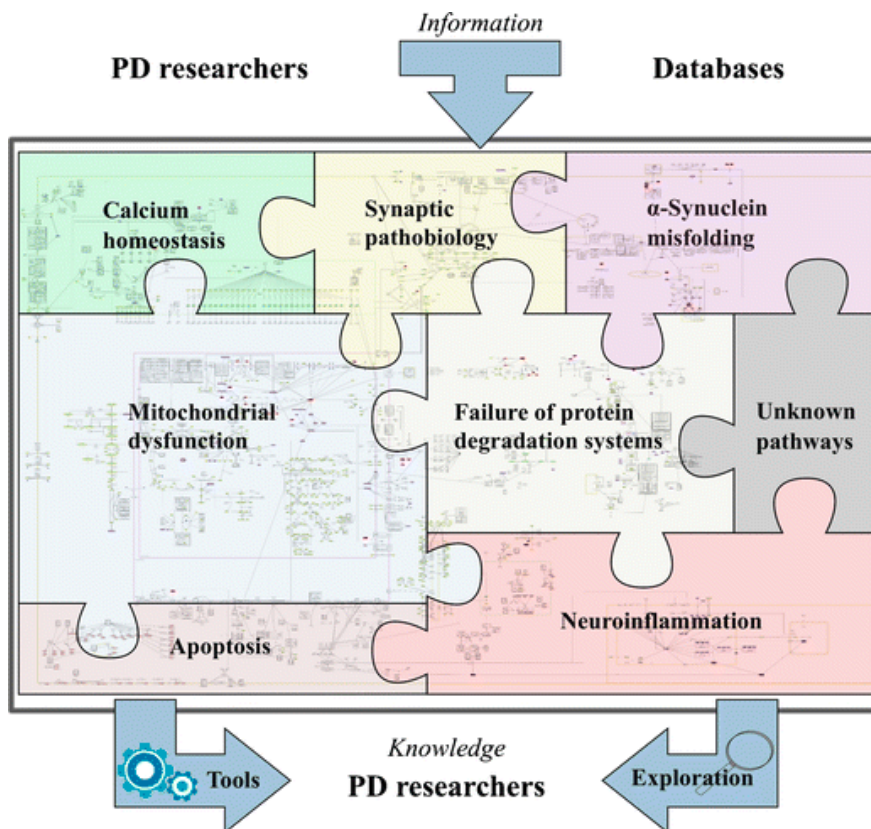


Figure 1.2: The concept of PD map and its visualization (taken from Fujita et al. (2014)).

³<http://minerva.uni.lu/MapView/>

Alpha-synuclein misfolding leads to the appearance of Lewy Bodies, a major component of PD. Although a lot of research has been carried out, the exact function of alpha-synuclein remains unknown (Breydo et al., 2012). Previously, links between the presence of Lewy Bodies and neuronal plasticity responses, enzyme regulation, transporters, and neurotransmitter vesicles and others were established (Uversky, 2008). Overall, Lewy Bodies are found in the majority of PD affected brains (Wakabayashi et al., 2007), but can also be an indicator for other types of Lewy Body dementia. Environmental factors, oxidative stress, mitochondrial dysfunction, genetic factors and dysfunction of the ubiquitin-proteasome system have been proposed to trigger the misfolding of alpha-synuclein leading to the Lewy Body formation. Initially it was proposed that alpha-synuclein and Lewy Bodies are cytotoxic, nevertheless a direct link with neuronal cell death could not be shown (Wakabayashi et al., 2007). Hence the real impact of alpha-synuclein misfolding is still elusive even though a link to PD is widely accepted. Understanding the role of Lewy Bodies in disease development could help to target them during disease treatment.

Apoptosis is specifically associated with PD in late stages of the disease development. It has been proposed that a proapoptotic environment in the nigrostriatal region of PD patients induces neuronal cell death (Lev et al., 2003). Neuronal cell death has also been classified as an active process referred to as a programmed cell death. Compared with induced cell death it seems to involve slightly different pathways than traditional apoptotic ones. Some affected functions are shared by both processes, but programmed cell death also requires ATP and shows a number of associated molecular alterations (Venderova and Park, 2012). Some of these can explain the link to PD. Hence, more detailed insight into the process could help to reduce the speed of neurodegeneration and overall disease progression.

Calcium homeostasis has been shown to be dysregulated in PD patients. Since calcium plays a ubiquitous role in cells it influences different PD associated pathways. Within dopaminergic neurons calcium is related to mitochondrial functionality, oxidative stress and lysosomal activity (Schapira, 2013). Furthermore, its role is key in the transmission of depolarizing signal and contributes to synaptic activity (Cali et al., 2014). All of these effects show major disease links and further knowledge might help to counteract energy dysregulation.

Failure of the protein degradation system can be a cause for protein accumulation within cells. This characteristic is specifically associated with age-related diseases, including PD. Under healthy conditions misfolded or not required proteins are degraded. One of the main systems in charge of such processes is the ubiquitin-proteasome system (Cook et al., 2012). In cases of disruption, misfolded proteins accumulate, potentially leading to cell death. Additionally, a direct link with the accumulation of alpha-synuclein (Martins-Branco et al., 2012) has been proposed. Unravelling concrete dysfunctions in the system can help to better understand links to PD.

Mitochondrial dysfunction can influence brain cells in PD patients in different ways. It can affect the cells through mitochondrion dependent programmed cell death or necrosis (Perier et al., 2012). Additionally it was possible to link complex 1 of the mammalian electron transfer chain to the PD pathology (Greenamyre et al., 2001). Its dysfunction leads to depressed rates of ATP synthesis possibly inducing graded mitochondrial depolarization and causing a decrease in intracellular ATP/energy levels. This lack of cellular energy will ultimately lead to cell death. Avoiding these processes could counteract the manifestation of PD.

Neuroinflammation has been linked to PD in several occasions. Distinct triggers for the inflammatory process are known and range from immunological challenges through bacterial or viral infections to injury such as stroke and others (Tansey and Goldberg, 2010). All of these alterations lead to an increase in the blood brain barrier permeability allowing filtration of lymphocytes and macrophages into the brain. Identifying substructures related to the immune response in affected brain regions of PD patients is another direct link of neuroinflammation with neuronal cell death (Hirsch et al., 2012). Such a pathway could be classified as an “autoimmune” response. A better understanding could help to prevent emergence of these processes.

Synaptic Vesicle Cycling and recycling has been linked to PD in several occasions. Failing to transport information, in form of neurotransmitters e.g. to the synaptic membrane can lead to a lack of information and postsynaptic triggers. This leads to a synaptic dysfunction inducing cell death (Esposito et al., 2012).

Although it remains questionable whether the presented processes are direct PD causes or consequences of dopaminergic cell loss, knowing about them can help to identify the

disease causing ones amongst them. This raises hope to be able to identify and diagnose PD in earlier disease stages. Gaining this knowledge might also allow to establish more specific “disease-subtypes”, depending on dysfunctioning systems, reflecting the underlying disease pathology. Some of the traditionally known PD subtypes are introduced in the next section.

1.1.1.2 PD-Subtypes

Traditionally PD is divided into a familial and sporadic form. This division depends mainly on the family history (and possibly traceable mutations) which can provide evidence for the familial form. Amongst familial cases around 30% are known to be based on genetic dysfunctions. This number decreases to 3-5% in sporadic cases. Some of the most well known genetic causes are linked to genes such as *LRRK2* and *SNCA* (Li et al., 2014; Siddiqui et al., 2016).

Furthermore, classic subtypes are described based on the disease phenotype and distinguish between either akinetic-rigid or tremor-dominant. Other research identified differently defined large clusters of symptoms. These specify patients with “old-” versus “young-age-at-onset” and “rapid-” versus “slow-disease-progression” (van Rooden et al., 2011; Eggers et al., 2014). Based on the variety of affected pathways further disease subtypes may emerge in the future.

Additionally, recent findings lead to the hypothesis that PD should be considered a syndrome rather than a single disease (Caligiore et al., 2016). As such, “PD” currently describes the “ultimate” disease phenotype, caused by a wide range of affected underlying subsystems (Fujita et al., 2014). Individually or jointly affected subsystems could be classified as PD subtypes. Given the diversity of subsystems it is also very likely that those could be referred to as different diseases, especially when earlier diagnosis becomes possible. Hence this reflects additional support for the importance of identifying, classifying and separating causes which can trigger the disease outbreak individually. Apart from identifying subtypes this would also allow earlier diagnosis and more specific treatment.

Overall, many individual PD associated pathways are relatively well understood. Major efforts have been made to understand these individually. However, few studies have been carried out to capture the complete disease picture.

Systems biological approaches are the tool of choice to tackle the presented problem. For best results data quality is of highest importance. The next sections covers respective details.

1.2 The Synapse

Synapses are part of neurons. As such they make up a large part of the (mammalian) brain. The synapse is key to cell-cell communication, allowing to transmit information from one cell to another. Chemical synapses (Yuste, 2015) can be split into three main compartments. These are the presynapse, postsynapse and synaptic cleft (Figure 1.3). More recently glial cells (astrocytes amongst others) are considered as part of the synapse as well. These surround presynapse, synaptic cleft and postsynapse, generating a micro-environment. Their specific role is not yet understood, but the concept of the “tetrapartite synapse” is gaining growing recognition with a large body of literature showing a role of glial cells in all essential brain functions (Dieterich and Kreutz, 2016). Nevertheless, glial cells are beyond the scope of this study.

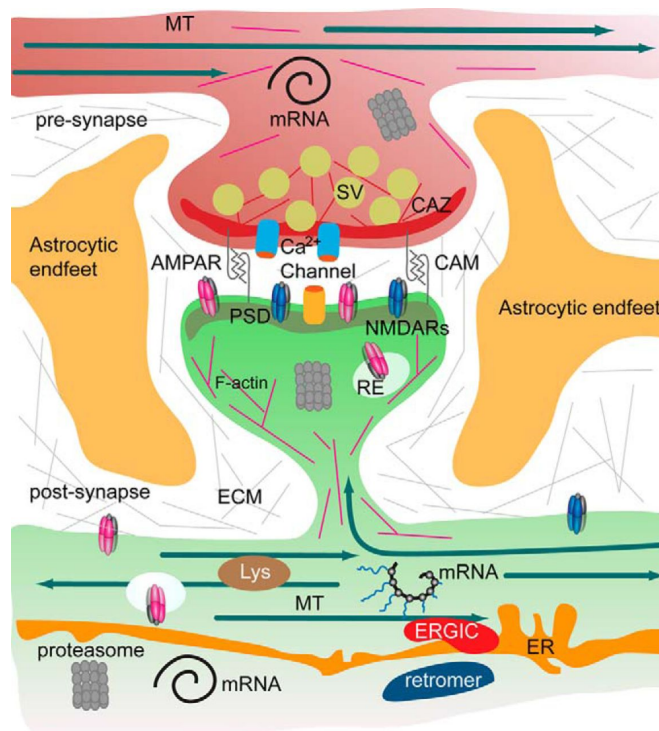


Figure 1.3: The tetrapartite synapse of principal neurons, consisting of the pre- and postsynaptic compartment, synaptic cleft, astrocytic endfeet, and extracellular matrix. The tightly regulated protein composition in the different regions can be seen. SV stands for synaptic vesicle (taken from Dieterich and Kreutz (2016)).

Reflected by the anatomical composition of a synapse, signals are transmitted from the presynapse to the postsynapse. An incoming presynaptic electric signal triggers synaptic vesicles to locate on the presynaptic membrane. In the following step they

release specific neurotransmitters and spread these into the synaptic cleft. There, they bind to receptors, integrated in the postsynaptic membrane of the receiving neuron. These binding reactions trigger signalling cascades inside the postsynapse, translating the incoming signal into a variety of processes. This “two-component” interaction has a vast number of regulatory points (Di Maio, 2008). Many aspects of the information transmission process can be altered and adjusted, making the synapse a highly adaptable system which is reflected in its cell type specificities. Apart from its versatility this complexity makes the synapse very hard to study and susceptible to disease with dysfunctions which are hard to identify.

Apart from the purely anatomical description, the term synaptosome is widely used when considering synapses. This is mainly due to experimental tissue preparation techniques (Laßek et al., 2015) which established the term as the unit of extracted tissue. It summarises the synapse as introduced earlier as well as additional components in the presynaptic terminal, such as mitochondria and synaptic vesicles as well as extracellular matrix proteins (Laßek et al., 2015).

Recent advances in experimental techniques, such as high throughput proteomic studies, gave access to extended synaptic datasets. The next section describes how to obtain and process such datasets.

1.2.1 The Synaptic Proteome

Proteomic studies aim towards identifying all proteins transcribed and translated in a tissue or region. Thus one of the initial challenges in such an experimental setup is to obtain the material of interest. Tissue preparation for a synaptic sample is challenging and initially based on the synaptosome. It is the key structure, isolated from brain tissue (Sokolow et al., 2012; Dieterich and Kreutz, 2016). A number of experimental protocols are available to obtain proteomic data, all starting with tissue homogenate as the raw material. If desired, density centrifugation is used to separate pre- from postsynaptic material and other cells. Optionally antibodies or other tags can be used to specifically target proteins from one of the synaptic regions. Once proteins are extracted, these are purified and mass-spectrometric analysis is used to identify them. Analytical data analysis is carried out and generates information of the entire analysed proteome. Depending on centrifugation steps and the purpose of the analysis, some studies analyse the full synaptosome, consisting of the entire synaptic region (Whittaker et al., 1964; Sokolow et al., 2012; Dieterich and Kreutz, 2016) or focus on the

presynaptic (Boyken et al., 2013; Grønberg et al., 2010) and/or postsynaptic proteome (Fernández et al., 2009; Bayés et al., 2012) individually.

One needs to keep in mind that results might not be fully complete and always only reflect the set of proteins present in the extracted sample at the point of tissue extraction. Therefore proteomic datasets are specific to a certain developmental state and time.

Even though experimentally identified units are proteins, it is more convenient to work with gene identifiers. This facilitates to exchange information coming from different species and avoids bias towards specific protein isoforms. This is specifically the case since mass spectrometry and analytical steps are not yet detailed enough to separate protein isoforms with high precision when reading large samples.

1.3 Systems Biology

Systems biology is a still relatively young field, but has been growing rapidly in recent years. Very often it is associated with large-scale analysis which is not intrinsically true. In general, systems biology addresses any topic on a “systems level” including experimental and/or theoretical approaches. The field aims towards gaining a high-level overview of a given system, considering data availability (Kitano, 2002b,a) and combining suitable approaches.

Often such approaches have proven to be challenging since data are supplied in different formats and certain analytical tests require very specific information and data input. Several initiatives have been put in place to assist endeavours towards facilitating data accessibility, usage and interpretation. One of them is the “FAIR” data-use principal (Wilkinson et al., 2016). FAIR stands for: **F**indable, **A**ccessible, **I**nteroperable and **R**eusable and aims towards generating more easily exchangeable data to allow the whole scientific community to benefit from it.

In the presented study differences in data annotation presented recurring challenges. Depending on the situation they were solved in different ways, largely working towards the use of accepted standards. Some of the mapping steps could not be completely automatised and required additional manual steps. This thorough approach lead to results following accepted standards, making the data more valuable. Their use in further experiments, a wide range of analyses, and amongst the research community is highly beneficial.

1.4 Protein-Protein-Interaction Networks

Network analysis, also network theory or graph analysis studies structures emerging in directed or undirected networks. Such networks can be defined as graphs consisting of “nodes”, also referred to as “vertices”, and “edges” connecting those nodes. Nodes and edges can have attributes such as names and weights adding further information to the network. The Euler’s solution to the “Seven Bridges of Königsberg problem” is seen as the first proof in network theory (Newman, 2003). Since then amongst others physics, computer science, engineering, biology, economics and sociology apply network approaches to unravel insights into e.g. the World Wide Web, social, epistemological or gene regulatory and metabolic networks through the use of various analytical approaches.

In this work nodes are proteins and edges their interactions. These are specifically referred to as Protein-Protein Interactions (PPIs) and a growing amount is available in publicly accessible databases. A detailed introduction can be found in Chapter 4.

Given a biological context network analysis shows growing impact in a number of areas. Protein-Protein-Interaction Networks (PPINs) visualize complex biological interaction patterns and aim to identify molecularly similar subgroups (Xia et al., 2014; Wang et al., 2010; Pizzuti and Rombo, 2014). Neuronal networks (Paliwal and Kumar, 2009) strive towards describing processes such as memory formation and signal transmission. Gene regulatory networks highlight regulatory and control relationships between proteins and genes or vice versa (Emmert-Streib et al., 2014). Other approaches are available and more will likely be added in the coming years.

With the increase in data availability the number of analytical approaches is constantly growing. This points towards the power of network analysis, if correctly applied.

For the purpose of this work the focus is on PPINs. Proteins are the functional units of cells and synapses. To carry out their functions they need to interact between each other. Some proteins undergo interactions with many others, whereas others with very few. This means that a protein can have a central, connective position or play a highly specific role involved in one function. This is just one example when PPINs are an attractive analytical tool to unravel and point out such properties.

Various network measures exist to describe distinct network properties. Some of these cover general measures referring to the entire network, and others focus on node or edge specific properties, characterising these individually. These can be referred

to as low-level statistics. Additionally high-level network statistics address aspects on a more general network level such as its division into communities. A network's structure is often referred to as topology (Davis et al., 2015), reflecting properties of the presented data. In general, all the presented measures can be used to characterise and compare networks between each other. The following two sections focus on the low- and high-level concepts and respective analytical approaches.

1.4.1 Statistical Network Analysis

Different types of low-level network analytical approaches exist. These range from values describing overall network properties to information specific to individual nodes or edges (Bliss et al., 2014). Both types have different advantages and disadvantages and serve distinct purposes. Overall, these measurements give a general idea of the network structure, whereas node or edge specific values can supply information regarding their individual role in the network.

For a general overview the underlying principles behind the statistical approaches are introduced:

Clustering Coefficient is a measure describing the degree to which nodes in a network tend to cluster together (Soffer and Vázquez, 2005). This gives a first insight into the possibility of finding clearly defined network substructures and reflects a property of the entire network.

(Network) Density “ D ” is defined as the ratio of the number of edges (“ E ”) that appear in the network of interest, compared to the number of possible edges between all nodes (Pavlopoulos et al., 2011). This measure indicates how densely connected networks are, pointing towards the connectivity of its components.

Node Degree is a measure describing the number of connections a node has with other nodes. In other words, it is the number of edges adjacent to a node. Nodes with a large number of connections (relative to the connectivity in the network and the total number of nodes) are hubs. In biological networks they often play a role as key connectors and regulators between different pathways. Hubs are often multi-domain proteins, likely involved in a versatile set of functions (Patil et al., 2010). Generally two types of hubs exist, and can appear with different references in literature. Transient or date hubs participate in single interactions

at a time whereas obligate or party hubs undergo multiple interactions simultaneously (Ran et al., 2013; Han et al., 2004). Depending on the type of hub, their removal implies different consequences, but overall it leads to crucial changes in the network structure. Hence hub nodes play important roles and are generally located very centrally in the network. On the contrary, the low degree nodes are found in the network periphery and their removal does not normally cause drastic effects on the network itself.

Apart from identifying prominent positions in the network the node degree distribution can give insights into the heterogeneity of a network. If the node degree distribution can be fit to a power law distribution the network is considered “scale-free”. This implies a long tail, power-law distribution of the node degree with few highly connected nodes and an exponentially larger number of weakly connected nodes. In biological terms this means that the probability of a substrate to react with x other substrates decays as a power law (Ravasz et al., 2002; Barabási and Albert, 1999). Based on these properties scale-free analysis reflects network topology regarding the connectedness between network nodes. This analysis can help to confirm if a network has a topology generally known for biological networks.

Betweenness (Centrality) is a centrality measure based on the number of shortest paths passing through a node (Freeman, 1977; Brandes, 2001). It describes the control a node has over a network, based on the “amount” of information that passes through it. This can also be described as the amount of information that “flows” over a certain node. Higher betweenness scores stand for higher centrality, monitoring communications between other nodes in the network. Considering PPINs such insight is specifically useful, since nodes with a high betweenness value are highly frequented and can assist in information exchange between different pathways (Vidal et al., 2011). On the contrary, nodes with a low betweenness score, are also referred to as “bottlenecks” or “gate-keepers” since information can get “stuck” or is purposefully delayed, by not being forwarded rapidly to other nodes in the network. Such detail helps to better characterise individual network nodes.

Using the introduced measures to classify and analyse networks often allows one to draw further conclusions which are based on certain combinations of the network statistics. For example, the scale-free nature of many biological networks (Barabási

and Albert, 1999) refers to properties such as high degree nodes, indicating hubs, sometimes linked to disease related genes. Overall, most of the measures can be used to draw biological and functional conclusions and provide tools assisting the comparison of networks against each other.

1.4.2 Network Clustering

After having analysed the PPINs as they emerge based on the PPI pattern, further interest lies in identifying network substructures. So called clustering algorithms divide networks into communities, aiming towards identifying the “best”, most realistic division of network nodes. To do so, a number of approaches are available, all aiming towards grouping more closely connected nodes together by separating them from less closely connected ones. Given a biological context, these communities consist of genes likely sharing similar functions or being “jointly” affected by the same disease.

Identifying “close connectedness” between a set of genes compared to others is one of the main challenges in the field of network analysis and the number of available techniques is constantly growing. Such techniques are referred to as network clustering and use so called clustering algorithms. Some of the approaches are based on node betweenness scores, the shortest walk between nodes and other measures such as modularity scores of the network (Brandes et al., 2008). Modularity “ Q ” is a measure describing the number of edges falling within a given gene group less than the expected fraction if genes were allocated at random or alternatively in an equivalent network (Newman, 2006b; Ravasz et al., 2002).

Apart from the spinglass approach (introduced later on), clustering algorithms used in this work are modularity based. An overview, as well as underlying computational principles, can be found in Table 1.1.

1.4.2.1 Analysis of clustered PPINs

Once networks are clustered their structures can be compared. Due to the large datasets and complex emerging community constellations it remains a major challenge to compare network clustering results amongst each other. This makes it hard to identify the right or best algorithm to represent a dataset, since it might not even exist. Certain statistical tests are available to e.g. test for the robustness of a clustering result, supporting emerging network structures. Overall, it is necessary to be aware of the remaining gaps and drawbacks PPIN clustering analysis contains.

Table 1.1: Community clustering algorithms used to divide networks into communities.

Algorithm Name	Reference	Principle	Additional Comments
Fast-greedy	Clauset et al. (2004)	hierarchical agglomerative algorithm with greedy optimization approach	one of the first algorithms for large networks with reasonable compute time, vaguely based on Newman (2004)
Infomap	Rosvall and Bergstrom (2008)	information theory approach - minimizes the expected description length of a random walker trajectory	seeks optimal community structure by compressing a descriptive "information flow" between nodes in the network
Louvain	Blondel et al. (2008)	heuristic algorithm, based on modularity optimization in a hierarchical way	reassigns community to nodes in an iterative manner, works with very large networks and short computation times
Spectral	Mclean et al. (2016); Newman and Girvan (2004)	spectral based modularity clustering with fine-tuning step	eigenvectors and eigenvalues are used to describe the network.; only available in C++ and as a cytoscape app; especially powerful in detecting network communities that are enriched in similar biological functions
Spinglass	Reichardt and Bornholdt (2006); Traag and Bruggeman (2009)	community detection is equivalent to identifying the ground state of a infinite range spin glass	minimization of the spin glass with the spin state representing the community indices; allows to detect overlap and hierarchy in community structure

Nevertheless, the use of PPINs greatly supports the identification of patterns and biologically similar subgroups amongst larger datasets. Depending on the research question, adjustments can be made to fine-tune analytical steps and to best benefit from the results.

1.5 Functional Gene Set Analysis

Very often genes are not analysed individually, but as a group, since gene sets can show common properties. To identify these properties, information regarding the property of interest needs to be available for all genes in the set of interest as well as a background set.

Gene sets as well as properties of interest can vary largely. Examples include genes specifically expressed in a tissue or cell type of choice compared to all protein coding genes in the human genome or a subgroup of expressed genes in a tissue of interest, compared to all genes expressed in the same. Properties range from previously identified gene-disease associations, functional descriptions of genes or their spatial expression within a cell and many others (Fury et al., 2006).

In any given scenario the main question is to identify if a given number of genes with a certain property found in a gene set of interest is higher than expected by chance. Such a situation can be described as an over-representation or enrichment of a property

amongst genes in a set. To calculate this probability a number of factors need to be taken into account and statistical tests are available to carry out exactly this analysis. The next section introduces the details.

1.5.1 Testing for Enrichment

To confirm over-representation of genes with a certain property, Fisher's exact test or a hypergeometric test are commonly used. Since both are known to be equivalent (Rivals et al., 2007), a detailed example and description of the hypergeometric enrichment test is given.

To identify a non-random accumulation of genes associated with a specific property in a gene set four key numbers need to be considered. These are:

1. The number of genes in a full dataset, also considered as the background dataset, N . Given the interest in a specific group of proteins, the background could either contain all human protein coding genes or a specific gene set of interest, e.g. all genes expressed in the synaptic proteome, also referred to as the synaptic proteome.
2. The number of genes n in the subset of the full dataset which is tested for enrichment. This is referred to as the "gene set of interest" and could be any subset of the background set N . Examples are all genes in the presynapse, or a specific set of proteins expressed in the synapse, e.g. a network community.
3. The number of genes associated with a certain property in the full dataset, T . This can either be the number of genes associated with a specific disease, function or spatial component amongst others.
4. The number of genes t which represent a subset of T found in n . This refers to the number of genes associated with the studied property (T) that are also present in the gene list of interest (n).

Based on these numbers a 2×2 contingency table can be constructed and the probability of encountering the exact number of hits t of interest in a set of genes n associated with a property T , given a background N , can be calculated. Section 2.3.1 introduces the formula and further details.

If this probability is less than a certain threshold (e.g. $p < 0.05$), the dataset is regarded as enriched for the tested property (Rivals et al., 2007), or alternatively genes

with the property of interest are considered as over-represented in the gene set of interest.

1.5.2 Functional Annotations

Based on the introduced principle, the property of interest can be a specific function, a process, a spatial component, a disease, etc. Over recent years large initiatives have curated functional annotations for human protein coding genes.

The **Reactome** (Croft et al., 2014; Fabregat et al., 2016) database for example is a “free, open-source, curated and peer reviewed pathway database”⁴. It associates genes to molecular pathways also supplying a full overview of dependencies between involved proteins.

As previously mentioned, **gene-disease association** information is of considerable interest for this study. Several databases such as ClinVar (Landrum et al., 2014) and the Human Gene Mutation Database (HGMD) (Stenson et al., 2014) amongst others store such information. Standardised disease identifiers are supplied by the Disease Ontology consortium (Schriml et al., 2011) which aims towards developing a “standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts”⁵. Further details are addressed in Chapter 3. Even though databases exist, in theory any (self-generated) disease-gene-association dataset can be used as a source of information.

Another initiative is **Gene Ontology (GO)** focusing on functional terms. Considering GO (Ashburner et al., 2000), gene associated properties are also referred to as traits. GO aims towards developing an “up-to-date, comprehensive, computational model of biological systems”⁶. Therefore it covers three key areas: (i) Biological Processes (ii) Molecular Functions and (iii) Cellular Components. Data in all of those ontologies are publicly accessible and follow a directed acyclic graph (DAG) structure. It means that terms relate to each other in a tree structure, moving from very generic terms describing functions such as “metabolic process” (GO:0008152) to more specific ones like “positive regulation of L-dopa biosynthetic process” (GO:1903197).

Depending on the analytical questions a study can address a specific level of detail along the annotation tree. Since this information is deposited in publicly available

⁴<http://www.reactome.org/>

⁵<http://disease-ontology.org/>

⁶<http://www.geneontology.org/>

databases, gene annotations of interest can be easily obtained and used for analysis on a large scale. Their combination with above mentioned enrichment tests can be used to classify gene sets regarding their overall function.

A number of tools have become available to test for enrichment given a gene and background as well as trait dataset of interest. The next section introduces two available tools.

1.5.3 topGO and topONTO

topGO and topONTO are computational environments (both available in R) allowing to carry out enrichment studies. The first developed topGO package (Alexa et al., 2006) introduces a way to directly work with gene-trait association information from GO. Based on this, the R package topONTO (He and Simpson, 2017b) was developed. It provides a more flexible environment which allows to work with ontologies other than GO. Both packages facilitate functional enrichment analysis for gene sets of interest given a self defined background gene set. The Fisher Exact test is used to identify enrichment and tested traits can be retrieved directly from GO.

To benefit from the hierarchical tree structure of ontologies, topGO implemented a number of more advanced analytical approaches. For best results specific algorithms, considering the ontology structure, are put in place. topONTO inherits these algorithms making them available for the use with other ontologies as well. The next section introduces the technique and specific algorithm which was chosen in the work presented.

1.5.3.1 The topGO elim algorithm

Enrichment results depend on different aspects of available annotation information. Based on the ontology structure the analysis can be adjusted to a desired level of detail along the ontology tree. For example, a gene can be tagged with the term “transmission across chemical synapses” which is a relatively broad description. But more specific tags such as “trafficking of AMPA receptors” are also available. The higher the term is located in the hierarchy, the more genes are associated with it. For example, 212 genes are associated with “transmission across chemical synapses”, but only 31 genes are specifically described as related to “trafficking of AMPA receptors”. Considering the relationship between those numbers the 31 genes associated with “trafficking of AMPA receptors” are also amongst the 212 genes specified in the “transmission across chemical synapses” category. Hence, different levels in annotation detail are an

important feature to be considered during enrichment analysis.

To retrieve most specific and refined terms among significantly enriched ones, the `elim` algorithm proposed by (Alexa et al., 2006) was used. Through consideration of the ontology tree structure it is possible to target the enrichment analysis towards a desired information specificity. Since a child term is potentially more interesting than its more generic ancestors, the `elim` algorithm computes significance of a term depending on its child terms. More specifically functional enrichment analysis is carried out starting from the lowest level traits in the ontology tree. If the lowest leaf terms of a branch are not significantly enriched in the gene set of interest traits on the next up-stream level are tested. Once the gene set of interest is significantly enriched based on a certain ontology term, genes associated with the enriched term are deleted from all gene-trait sets upstream of the enriched one. In other words, genes associated to a child term of a trait of interest are disregarded in future enrichment tests. Hence, once a gene trait association contributed to an enriched term of a gene set that gene is no longer considered for additional contribution to enrichment of a functional parent term. In summary, this approach guarantees to identify the most specific enriched functional term for the gene set tested.

It is also advantageous that the `elim` algorithm keeps track of the number of enrichment tests that are carried out while analysing the whole ontology tree. This information is crucial to correct for multiple testing which is explained in the next section.

Before moving on, it should be pointed out that the ONTO-Suite Miner, underlying the `topONTO` package (He and Simpson, 2017a) is also a great tool to extract e.g. gene-disease association information retrieved from distinct databases in an automated manner. Specific disease identifiers (“DOIDs”) can be used to screen loaded datasets e.g. from Ensembl Variation (EnsVar), Gene Reference into Function (GeneRIF) and Online Mendelian Inheritance in Man (OMIM), all providing a link between gene alterations and disease (see Section 3.2.2 above for more details).

1.5.3.2 Correction for Multiple Testing

A common challenge when carrying out large-scale analytical tests is correction for multiple testing. The main interest in using such a correction is to ensure that obtained significance values were not influenced by the number of comparisons made throughout the repetitive testing process (Shaffer, 1995; Al-Shahrour et al., 2004).

Therefore different, more or less stringent approaches are available. One of the most traditional and very strict correction approaches is referred to as Bonferroni cor-

rection (Bonferroni, 1935). It considers the exact number of elements in the tested dataset to correct the initial p-value.

An alternative less stringent and well studied approach was published by Benjamini and Yekutieli (Benjamini and Yekutieli, 2001). Instead of considering the exact number of elements, the number of comparisons, also representing number of “degrees of freedom” are used for correction. Additionally the method controls for the false discovery rate, describing the expected proportion of false discoveries amongst the rejected tests. This makes it less stringent than the family-wise error rate, giving increased sensitive to detect enriched traits.

This work concentrates on the second, less strict approach which was proven to be sufficient enough to correct initial p-values, without being overly strict and could lead to the loss of interesting results.

1.6 Objectives

It is of great interest to gain further understanding of the complexity behind PD. Having a more detailed overview would be highly beneficial to better diagnose the disease, identify biomarkers and develop more targeted treatment. Therefore this study includes data from several sources addressing distinct molecular details. PPINs were chosen to shed light over the data structure and gain detailed gene specific, as well as more general, functional insights to the data and disease.

Data quality and an open development of the analytical workflow are of highest importance and propose open approaches to be used to answer similar research questions in the future.

More specifically, it has been widely suspected that PD, as it is diagnosed today, can be the result of different causal dysfunctions which might be considered disease subtypes or even represent individual diseases. Hence dividing affected genes into groups is of great interest and might represent disease types which are associated with crucial dysfunctions in distinct cellular regions and affect diverse molecular functions.

Hence the main interest of the presented analysis is to gain more in-depth functional information about genes associated with PD and detect highly affected gene groups acting together. Since these are often triggering the disease manifestation in the synapse they indicate likely promising areas for further detailed research in the PD field. The identification of their overall function can contribute more insight pinning down largely affected functional areas.

Finally, this work aims towards presenting a set of known and new PD associated genes, as well as synaptic dysfunctions.

Figure 1.4 shows a general overview of the different pieces of work presented in the following chapters.

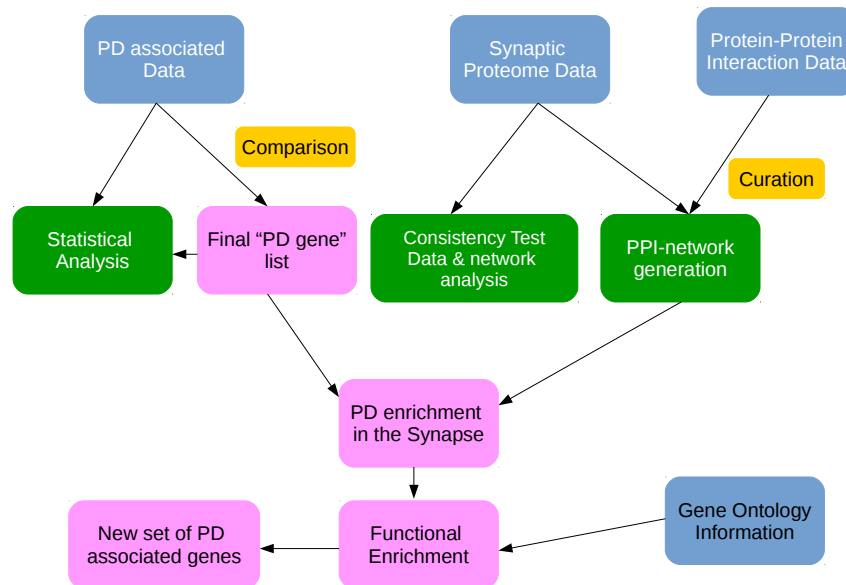


Figure 1.4: Work presented in this thesis. Blue boxes refer to data, orange boxes indicates processes, green boxes highlight analytical steps and outcomes are shown in magenta boxes.

Chapter 2

Methods

This chapter introduces methods of general relevance to the work presented and used in more than one occasion. Methods used only in a specific chapter are introduced there.

2.1 General Programming

Most of the analytical work presented in this work was carried out with one of the programming languages: Python (2.7.13) (van Rossum, 1995)¹ or R (3.4.1) (R Core Team, 2017)². A number of different packages were used, these are listed in the specific sections.

2.1.1 Venn Diagrams

Venn diagrams were generated using either the python package `matplotlib-venn`³ or the R library `Vennerable` (Swinton, 2013). Most of the four way diagrams were generated with R, but generally the tool of choice was defined based on where the data were previously processed and loaded.

¹<http://www.python.org>

²<https://www.R-project.org/>

³<https://github.com/konstantint/matplotlib-venn>

2.1.2 Computing Environment

The Edinburgh Compute and Data Facility (ECDF), *eddie*⁴ supplies a high-performance computing facility which was used for computationally heavier processes. It also allowed to parallelise computations if required.

2.2 Annotations and Mappings

Traditionally the discovery of a new gene or protein allowed researchers to name those. Over the years this led to genes and proteins with multiple names used in different publications not always being connected. In 1957 an international committee published recommendations for genetics symbols and their nomenclature (Tanaka, 1957). The Edinburgh Human Genome Meeting formalised those in 1979 publishing “full guidelines for human genome nomenclature” (HGNC, 1979). Now every known gene is specifically identified through a name, symbol and ID. Protein nomenclature underlies similar efforts and is closely linked to gene nomenclature. Human genes can be mapped to human proteins and vice versa, allowing the use of one single identifier (ID) for analysis. Due to variability in gene to protein transcription and translation, mappings are not always direct, one-to-one. Hence, one gene can encode several proteins. Additionally one protein can be encoded by several genes, which can be explained by having several copies of the same gene in the genome, all leading to the same protein product.

Genes and proteins can also be mapped across species. All species tend to follow a similar annotation structure, and structural and sequential similarities are reflected in gene and protein names. This allows mapping of e.g. mouse gene IDs to human gene IDs which is especially beneficial when experiments are carried out in different species.

It is also possible to use databases such as Ensembl when evolutionary relationship is not evident. These rely on more advanced methods to ensure gene correspondence across species (Herrero et al., 2016).

⁴<http://www.ed.ac.uk/information-services/research-support/research-computing/ecdf/high-performance-computing>

2.2.1 Mapping File Generation

Due to variability in data and experimental procedures not all used sources present results using the same gene or protein ID. To allow working with unique IDs, a mapping table was generated to map GeneSymbols (Uniprot Gene IDs) to Entrez Gene IDs and vice versa. Data to generate the mapping were obtained from UniProt (UniProt Consortium et al., 2017) and National Center for Biotechnology Information (NCBI) (NCBI, 2016). Three annotation files were used. The used Uniprot idmapping (selected tab file⁵ was obtained from the ftp server⁶ which can be accessed via the “previous release” repository⁷ in the moment of writing the thesis. NCBI data were also downloaded via an ftp server⁸. Two files were used to include an intermediate mapping step. The gene2accession file⁹ supplies a number of mappings, including the “NCBI protein accession” which is mapped to the “Uniprot Accession” which can be used in combination with a further mapping file¹⁰. Figure 2.1 shows an overview of data, mapping steps and the outcome. Inter-species mapping was carried out using NCBI and UniProt homology mapping flat-files^{11 12}. In the cases where no results were found homology mappings from the Mouse Genome Database¹³ were consulted. Homology mapping was mainly carried out by Colin Mclean.

All three datasets were downloaded on March 24th 2017 and analysis was specific for human gene taxID (9606) associated genes. For best mapping outcome the two NCBI mapping files were combined based on the “NCBI protein accession”. The columns “Entrez ID”, “NCBI Protein Accession”, “EntrezGeneName” and “UniprotAccession” were extracted from the raw files and kept for further analysis. Based on the NCBI mapping the data were merged with the Uniprot mapping file to obtain Uniprot IDs. This mapping step was possible through cross-linking the Entrez ID as well as the Uniprot Accession.

Based on this approach it was possible to map 33,345 Entrez Gene IDs to Uniprot

⁵idmapping_selected.tab.gz

⁶ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/

⁷ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2017_03/knowledgebase/knowledgebase2017_03.tar.gz

⁸<ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/>

⁹gene2accession.gz

¹⁰gene_refseq_uniprotkb_collab.gz

¹¹ftp://ftp.ncbi.nlm.nih.gov/pub/homology_maps

¹²ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/

¹³http://www.informatics.jax.org/downloads/reports/HOM_AllOrganism.rpt

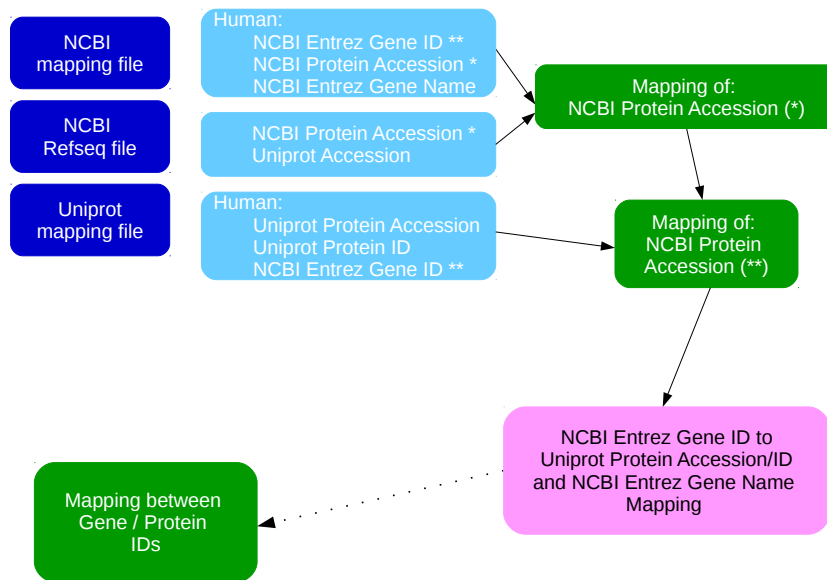


Figure 2.1: Overview of the mapping approach, showing input data and the obtained output. Dark blue boxes indicate raw data, light blue highlights the columns of interest in the respective files. Green boxes refer to processes and the magenta box highlights the outcome. */** highlight the information that was cross-linked between the files.

IDs (also referred to as “Uniprot entry name”). 19,377 Entrez Gene IDs and 32,879 Uniprot IDs are part of that dataset. An additional mapping file was obtained, including 57,606 Entrez Gene IDs, being mapped to 57,557 Entrez gene names.

Depending on the dataset size a manual mapping step followed the automatised mapping, to identify outstanding hits. Most of these used gene aliases as the reference gene name and could not be mapped automatically.

For the purpose of this work the Entrez Gene ID is used as the unique gene identifier. It will be referred to as Entrez ID.

2.3 Enrichment Analysis

2.3.1 Hypergeometric Testing

Hypergeometric testing is a statistical approach that calculates the probability of over-representation of a “trait” of interest in a subgroup of a larger group (Fury et al., 2006). Thus statistical testing has been extensively used in large-scale studies to identify over-represented or underrepresented “traits” such as specific genes and/or functions. To be

able to calculate potential over-representation a background gene set is required. Trait information for all genes needs to be available. Given the context of this work, the groups consist of varying numbers of genes. Traits can vary widely from genes previously associated to a disease, genes associated to a specific biological process, molecular function or cellular component. They can also be manually defined or retrieved from databases.

A number of different scenarios can require enrichment testing. Understanding numbers larger than expected at random e.g. disease associated genes in a gene subset (e.g. in a specific tissue, given a genome background), identifying a significant overlap between two gene lists or detecting general over-representation of a functional term amongst genes in a set are just some examples.

The hypergeometric probability $h(t; N, n, T)$ describes the probability of encountering the exact number of hits of interest in the data subset, given a defined background set as well as a set of trait-associated genes.

Four numbers are required to compute the probability:

1. the number of elements in the full dataset, also considered as the background dataset (“ N ”),
2. the number of elements in the subset of the full dataset which shall be tested for enrichment (“ n ”),
3. the number of elements of interest, e.g. associated with a certain trait, in the full dataset (“ T ”) and
4. the number of elements of interest in the data subset of interest (“ t ”).

This information allows to compute a probability (p-value) of how likely it is to observe a given distribution of items. It can be calculated in the following way:

$$h(t; N, n, T) = \frac{\binom{T}{t} \binom{N-T}{n-t}}{\binom{N}{n}} \quad (2.1)$$

To describe the probability of finding the exact number of items of interest (“ t ”) in the subset (“ n ”) or more, the cumulative hypergeometric probability is used. It is the sum over the hypergeometric probabilities:

$$h(t \leq t; N, n, T) = \sum_{x=0}^t \frac{\binom{T}{x} \binom{N-T}{n-x}}{\binom{N}{n}} \quad (2.2)$$

Since one is interested in over-representation of a certain trait, relative to the total number of elements of interest the probabilities of seeing between “ t ” and “ T ” hits needs to be calculated. This is done as follows:

$$h(t \leq T; N, n, T) = \sum_{x=t}^T \frac{\binom{T}{x} \binom{N-T}{n-x}}{\binom{N}{n}} \quad (2.3)$$

Predefined functions to compute the (cumulative) hypergeometric probability are available in R. The `dhyper` function (Johnson et al., 2005) was selected for calculations in this work. Depending on the research question and analytical setup the one-tailed Fisher’s exact test is commonly applied. It is known to be equivalent to the hypergeometric test (Rivals et al., 2007).

2.3.2 Gene Set Enrichment: `topGO` and `topONTO`

Functional enrichment analysis is one of the examples where hypergeometric testing is required. `topGO` (Alexa et al., 2006) supplies an environment to carry out a number of enrichment test. Apart from the enrichment testing itself, it contains a range of algorithms guiding the testing approach. Since functional enrichment information used in these setups relies on data structured in form of directed acyclic graphs enrichment testing can be adjusted to consider this information. Since `topGO` is only able to access information from the Gene Ontology (GO) database, `topONTO` (He and Simpson, 2017b) was built to load ontologies and use the tools provided by `topGO` for the analysis of other gene-trait information sets. Thus `topONTO` has been used for the analysis in this work, also allowing to specify desired GO versions.

The presented work uses the one-tailed Fisher exact test, equivalent to the hypergeometric test. Section 1.5.1 explains the underlying test principle and Section 2.3.1 introduces technical details. As outlined in Section 1.5.3.1 the `elim` algorithm was used.

For a detailed overview of the full `topGO` and `topONTO` analysis, a step by step protocol is presented.

1. All available trait-gene mappings are retrieved from the source database. In the presented case, traits were GO terms associated with human Entrez IDs.
2. Once the database content was retrieved, lists of all Entrez IDs with an associated GO term of the specific subclasses (Biological Process, Molecular Function, Cellular Component) were generated.

3. Depending on the defined enrichment background, genes with an associated GO term were selected.
4. The gene subset to be tested for enrichment was imported and genes present in the background dataset were identified.
5. Once all data was prepared in the correct format, GO-data objects were generated. At this stage, the specific GO subclass is defined and according datasets are used.
6. The generated GO-data object is used to perform desired enrichment testing. At this stage enrichment test and algorithm are chosen.
7. Results can be accessed and visualized in data tables, as GO-graphs and word clouds amongst others.

A number of steps require specific data formats, and details can be found in the `topGO`¹⁴/`topONTO`¹⁵ documentation. These documents also include installation details.

Apart from being a great tool to carry out functional enrichment studies, `topONTO` can be used to extract sets of gene associated to terms in the loaded ontology. Hence genes associated to a specific disease (based on human Disease Ontology Identifiers (DOIDs)) can be automatically obtained.

2.3.3 Multiple Testing Correction

Different programming languages implement functions to carry out multiple testing correction. Since `topGO/topONTO` are run in R the *p.adjust* function from the `R-stats` package was used. Corrected p-values were calculated for all originally obtained ones.

When using the *p.adjust* function, the correction test type can be selected. The Benjamini and Yekutieli correction was used (Benjamini and Yekutieli, 2001). Apart from having medium stringency, it is relatively accessible regarding the specification of the number of tests that need to be corrected for. Given the use of the `elim` algorithm, it was possible to extract that number and integrate it for multiple testing correction.

Other available correction alternatives include the classical, very stringent Bonferroni correction, the Benjamini & Hochberg approach (also referred to as “fdr”/“false discovery rate detection”) as well as Holms, Hochbergs and Hommels individual methods.

¹⁴<https://bioconductor.org/packages/release/bioc/html/topGO.html>

¹⁵<https://github.com/hxin/topOnto>

2.4 Protein-Protein-Interaction Network Analysis

A number of visualization and analysis tools are available to generate and work with Protein-Protein-Interaction Networks (PPINs). *igraph* (Csardi and Nepusz, 2006)¹⁶ was chosen for network generation, clustering and analysis. This work uses the R implementation of *igraph*. Some of the final visualizations were obtained with *cytoscape* (Shannon et al., 2003)¹⁷.

All networks presented in this work are based on Protein-Protein Interaction (PPI) lists, so called edge-lists or edge-tables. These are generated based on a gene list of interest and contain information of two interacting genes (using a gene ID of choice). The curated, direct, human PPIs list (Chapter 4) was used to extract internal PPIs. This means that interactions are only considered if both interactors are present in the supplied gene list of interest. The Entrez ID was used as the unique identifier.

The edge list is the minimum requirement to build a network. For further detailed information a node list or node table can be supplied. This table contains information regarding (all) nodes (genes) in the network. The additional information can include gene names, disease association or other values of interest.

Based on this information, *igraph* can generate PPINs. Additionally, it provides a number of tools to directly compute statistical network measures. The following section describes concepts of global and local parameters and statistics which are used in this work.

Betweenness (Centrality) C_B reflects the number of shortest paths passing through a node. High betweenness centrality scores indicate that a lot of information passes by and/or is processed by a node (Freeman, 1977; Brandes, 2001). This highlights a node's centrality and often means that it is a form of communication centre between different network regions. To obtain the betweenness centrality of a node V in a graph $G : (V, E)$ with V nodes and E edges the following steps need to be taken: (i) the shortest path between each pair of nodes (xy) is computed; (ii) the fraction of shortest paths passing through the node in question (V) is determined; (iii) the final value is the sum over all the fraction values for all node pairs. More formally it can be represented as follows:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{xy}(v)}{\sigma_{xy}} \quad (2.4)$$

¹⁶<http://igraph.org>

¹⁷<http://www.cytoscape.org/>

σ_{xy} refers to the total number of shortest paths from nodes x to y and $\sigma_{xy}(v)$ is the number of those passing through node v . The betweenness can be normalised if required.

Closeness Centrality C_C quantifies the normalized average length of the shortest paths (geodesics) through a given node of interest. It is calculated by dividing the number of all nodes in the network by the sum of the shortest paths through all nodes in the network. More formally it is defined as follows:

$$C_C(x) = \frac{N}{\sum_y d(y,x)} \quad (2.5)$$

$d(x,y)$ is the distance between nodes x and y and N is the total number of nodes in the graph. This formula applies to large networks, where the difference between N and $N - 1$ is inconsequential. Nodes with a low value are separated from others by short geodesics. This might highlight better access to information or more direct influence at other vertices.

(Network) Density describes how dense a graph is based on the number of edges that are appearing in the graph. It is defined as the proportion of edges in the graph compared to all possible edges between any two nodes in the network. More formally this means:

$$\text{Network Density} = \frac{\text{actual edges in graph}}{\text{potential edges in graph}} \quad (2.6)$$

The number of potential edges in the graph is calculated as follows:

$$\text{Network Density} = \frac{n * (n - 1)}{2} \quad (2.7)$$

n refers to the number of nodes in the network. The measure is proportional to the maximum amount of all possible edges appearing between any two nodes in the graph (Wasserman and Faust, 1994).

Diameter describes the longest geodesic in a graph. It is basically the “longest shortest path” which can be found in the network, connecting two nodes amongst each other. It is identified through comparison of all geodesics in a graph.

Global Transitivity/Clustering coefficient are two terms, used equivalently. In an undirected graph, they describe the ratio of closed triangles appearing in a network, compared to connected triangles in the graph (the direction of the edges is ignored). In other words, it is a measure of probability that the adjacent vertices of a vertex are connected (Barrat et al., 2004).

A triangle can be described as a sequence of nodes x, y, z, x which are connected as follows: (x, y) , (y, z) and (z, x) . Global Transitivity is then calculated as follows:

$$T = \frac{\text{number of closed triangles}}{\text{total number of possible triangles}} \quad (2.8)$$

Node Degree is the number of connections a node shows. In other words it can be referred to as the number of edges connected to a node. The maximum node degree is the largest node degree appearing in a network. Generally nodes with a large degree are referred to as hubs.

Scale Free Network Analysis requires information regarding the node degree of all nodes in a network. Based on that, an alpha value describing the exponent of a fitted power-law distribution of the node degree can be obtained. A long tail, power-law distribution of the node degree points towards a so called scale-free network which is commonly seen in a biological context (Ravasz et al., 2002; Barabási and Albert, 1999).

Depending on the required insights different values were computed and analysed. For further higher-level analysis the next section outlines required steps.

2.4.1 Network Clustering Algorithms

Once PPINs are generated (Section 2.4), clustering algorithms can be applied to divide the network into communities. Many clustering algorithms are included in `igraph`. The ones used in this study include:

- fastgreedy (Clauset et al., 2004),
- infomap (Rosvall and Bergstrom, 2008),
- louvain (Blondel et al., 2008) and
- spinglass (Reichardt and Bornholdt, 2006; Newman and Girvan, 2004; Traag and Bruggeman, 2009).

However other algorithms are not implemented and need to be used separately. One of them is the

- spectral clustering algorithm (Newman, 2006a; Mclean et al., 2016).

A C++ and cytoscape implementation is deposited on sourceforge¹⁸ and can be directly downloaded¹⁹. The README files includes installation and usage instructions.

Since most of the above algorithms are based on modularity optimisation the underlying mathematical calculation is introduced. As a measure, describing the number of edges falling within a given gene group less than the expected fraction if genes were allocated at random, modularity Q is usually defined via a symmetric modularity matrix. Elements of this matrix can be represented as follows:

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \quad (2.9)$$

A_{ij} refers to the number of edges between nodes i and j , k_i and k_j is the number of edges of these nodes and $m = \frac{1}{2} \sum_i k_i$. Based on this the modularity Q is calculated with the following formula:

$$Q = \frac{1}{4m} s^T B s \quad (2.10)$$

with s being the column vector of elements s_i in the matrix which indicates if node i belongs to group 1 or group 2 under the tested conditions (Newman, 2006b).

¹⁸<https://sourceforge.net/projects/cdmsuite/>

¹⁹https://sourceforge.net/projects/cdmsuite/files/CDMSuite_cpp_v1r1/

Chapter 3

Finding a Parkinson's Disease Core Dataset

3.1 Objective

Parkinson's Disease (PD) is considered a complex disease affecting a number of pathways and showing large diversity in phenotypes. To gain better understanding of how PD affects the human cellular machinery, a core dataset covering genes significantly associated with the disease is needed. This chapter analyses published datasets describing PD, all of which address different disease aspects. Available sources were identified to extract PD associated genes and proteins. The raw data included information describing i) direct and indirect influences of mutations on genes and ii) effects of protein expression changes on the cellular machinery.

An additional aim was to understand the impact of a gene or protein alteration on the disease picture itself. This could help to classify alterations as 1) disease causal or 2) "consequential". The two aspects can be linked to the disease genotype or phenotype. Identifying links between underlying dysfunctions would allow classification of PD into different types.

Hence, this chapter aims to define a key set of genes associated with PD. An overview of the workflow including aims, analytical approaches and conclusions can be seen in Figure 3.1.

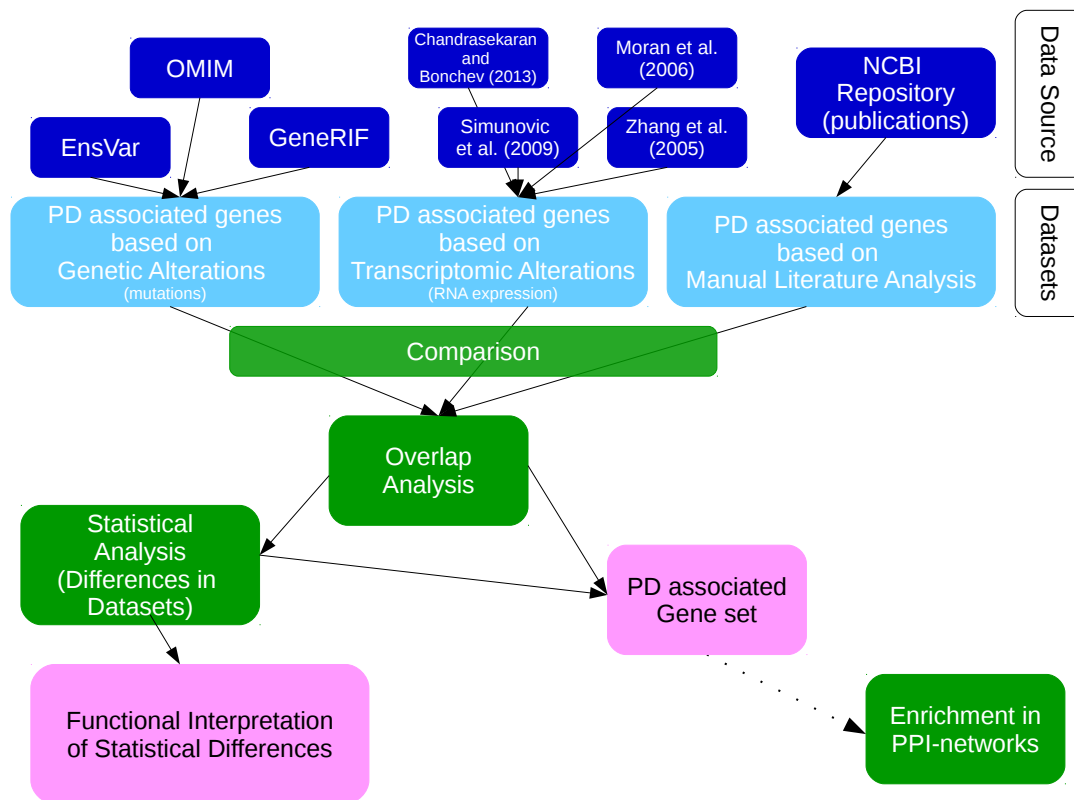


Figure 3.1: Work presented in Chapter 3, focusing on the workflow used to analyse data and generate a combined PD dataset. Dark blue boxes refer to published data, light blue boxes are generated datasets, green boxes describe processes and magenta boxes show outcomes.

3.2 Material

Over the years many researchers dedicated time and effort towards shedding light on the cause of PD. Significant effort has been put into collecting data which was made available to the research community mostly via databases. A number of (bioinformatics) approaches were used to retrieve desired information such as genes and proteins significantly associated with PD.

For the purpose of this study the unique identifier was chosen to be the human (gene) Entrez ID (further referred to as Entrez ID). In cases where only protein symbols or alternative identifiers were used, these were mapped to the respective Entrez ID.

3.2.1 Data Types

Figure 1.1 C shows different experimental approaches facilitating insight into distinct levels of the cellular machinery. These tools are designed to obtain large-scale data sets. Nevertheless results can also contain false positive hits. Due to our interest in the effects of PD on the genome as well as the exome, transcriptome (and proteome) data obtained via the following approaches were used.

Genome Wide Association Study (GWAS) are the most commonly used large-scale tool to obtain information regarding Single Nucleotide Polymorphisms (SNPs) associated to e.g. a disease. They are applied to analyse genetic alterations found in a given population as compared to a reference one. Once SNPs are identified these can be associated to nearby genes.

Microarray Studies can be used to identify differences in RNA and protein expression in a specific tissues between samples of interest and control. Statistical methods are used to define significant changes between the two datasets. Results include protein expression levels that change significantly under e.g. disease conditions. This study uses transcriptomic microarray study results as a data source.

Manually Curated Data Individual publications identify a number of protein- and gene-disease associations. The experimental objective can vary, but usually a small number of proteins are addressed. Through the screening of individual publications a further gene-disease set can be obtained. A number of databases collect such expert curated information and are publicly available.

3.2.2 Data Sources

The National Center for Biotechnology Information (NCBI) (NCBI, 2016) offers access to a large amount of information using publicly available publications and datasets. For the purpose of identifying PD associated genes a literature search, followed by manual curation was carried out. Papers published roughly in the last 10 years, based on personal recommendations and NCBI searches including: “(Parkinson’s Disease [Title]) AND (“2006/01/01” [Date - Publication] : “2014/01/01” [Date - Publication])” helped to identify such studies.

Manual publication search also allowed the identification of key studies focusing on PD associated genes identified based on transcriptomic changes (see the 2nd level of

detail in Figure 1.1 B). Most of those have deposited raw data in the Gene Expression Omnibus (GEO). Directly querying GEO helped to identify further deposited studies of interest.

GEO is a public repository with functional genomics data¹, accepting array and sequence based results (Edgar et al., 2002). Many publications share their data by publishing it on a platform such as GEO. Functional genomics data includes microarray and sequence-based technologies, both high-throughput approaches. Data can be accessed by specific accession numbers and can be cited in publications. Publication title, authors and keywords are accepted as search terms to query the database and obtain studies of interest.

Compared to information based on transcriptomic changes a larger number of resources hosting mutation-based gene-disease alterations are available. These either associate mutations to diseases and establish a mutation to gene relationship afterwards or directly associate genes to diseases. The former are widely based on GWAS or other mutational studies, whereas the latter rely on (manually curated) text annotation. The `topONTO` R package, allows mutation- and gene-disease information to be extracted from a number of databases. This highly facilitates their use and accessibility (more details in Section 1.5.3). The following databases can be queried via `topONTO` and were used in this study:

Ensembl Variation (EnsVar) contains information regarding genetic differences between individuals (Chen et al., 2010)². Available data describe sequence and structural variance, including SNPs (specific to one single nucleotide in the genome), insertions or deletions (of one or several nucleotides) as well as copy number variations (indicating the in- or decrease in the copy number of a given genomic region). Mutations are annotated, depending on their position on the genome. To understand potential links with gene coding or regulatory regions altered positions are mapped to gene location. EnsVar obtains data for human genetic information from six different sources (see entries type “variant” in the table at the online source³). These are: (i) “ClinVar”⁴ (Landrum et al., 2014) (ii) “COSMIC”⁵ (Forbes et al., 2015) (iii) “dbSNP”⁶ (Sherry et al., 2001)

¹<https://www.ncbi.nlm.nih.gov/geo/>

²<http://www.ensembl.org/info/genome/variation/index.html>

³http://www.ensembl.org/info/genome/variation/sources_documentation.html

⁴<https://www.ncbi.nlm.nih.gov/clinvar/>

⁵<http://cancer.sanger.ac.uk/cosmic>

⁶<https://www.ncbi.nlm.nih.gov/SNP/>

(iv) “ESP”⁷ (Exome Variant Server, 2012) (v) “HGMD-PUBLIC”⁸ (Stenson et al., 2014) and (vi) “PhenCode”⁹ (Giardine et al., 2007). Detailed information describing the six reference databases can be found in Table 3.1.

Table 3.1: Databases which EnsVar retrieves its data from.

Database - full name	Database - abbreviation	Data Description	Reference
Catalogue Of Somatic Mutations in Cancer	ClinVar	aggregates information about genomic variation and its relationship to human health; focus on medically important variants and phenotypes	Landrum et al. (2014)
Catalogue Of Somatic Mutations in Cancer	COSMIC	information about somatic mutations in human cancer	Forbes et al. (2015)
Database of Short Genetic Variations	dbSNP	hosted by NCBI; contains small genetic variations < 50 base pairs (bp)	Sherry et al. (2001)
NHLBI Exome Sequencing Project	ESP	focus on heart, lung and blood disorders; next-generation sequencing data of human protein coding regions is used to discover novel gene/mechanism-disease associations	Exome Variant Server (2012)
Human Gene Mutation Database	HGMD-PUBLIC	free and public version (slightly less up-to-date); supplies information about (published) gene lesions underlying and/or causing human inherited disease	Stenson et al. (2014)
PhenCode	PhenCode	aims towards better understanding of relationships between genotype and phenotype in humans, specifically focusing on clinical data; information combination from various locus-specific mutation databases with genome sequence data and evolutionary history	Giardine et al. (2007)

Gene Reference into Function (GeneRIF) (Jimeno-Yepes et al., 2013)¹⁰ provides a simple mechanism for researchers to integrate functional annotations of genes to genes listed in the NCBI EntrezGene “Gene” database (Maglott et al., 2005). Thereby it enriches available information, through e.g. functional terms or disease association. It is based around an open system where scientists can submit information for the wider community. A peer-reviewed publication is required to support any description which requires experimental and not only computational evidence. Additional information can be accessed with, e.g., text mining tools to

⁷<http://evs.gs.washington.edu/EVS/>

⁸<http://www.hgmd.cf.ac.uk/ac/index.php>

⁹<http://phencode.bx.psu.edu/>

¹⁰<https://www.ncbi.nlm.nih.gov/gene/about-generif>

obtain genes with descriptions of interest.

Online Mendelian Inheritance in Man (OMIM) , the Online Catalogue of Human Genes and Genetic Disorders¹¹ (McKusick, 1998; Amberger et al., 2009) contains text based annotations regarding all known Mendelian disorders for more than 15,000 human genes. OMIM focuses on the relationship between phenotype and genotype. Text-mining approaches can be used to extract information such as gene-disease links of interest.

Raw data containing information about PD associated genes is supplied as digital supplementary material (folder: “PD-associated-data”, the README file contains detailed information about the individual files).

3.3 Results

3.3.1 PD associated genes studied in literature

To obtain an overview of existing PD research, a manually curated dataset of genes associated with the disease was generated. A representative, rather than exhaustive list of publications, based on recommended papers and references, was gathered (see NCBI search as specified in Section 3.2.2). Publications were considered individually and a set of 52 Entrez IDs were extracted, several of them appearing in more than one publication. The identified genes of interest are: *ACMSD*, *ADORA2A*, *APP*, *ATP13A2*, *BST1*, *CACNA1D*, *CALB1*, *CALM1*, *CALR*, *CCDC62*, *CCL5*, *CDH8*, *DGKQ*, *EIF4G1*, *GAK*, *FGF20*, *GBA*, *GIGYF2*, *GPR37*, *HIP1R*, *HSPA4*, *HTRA2*, *ICAM1*, *ITGA8*, *MCC*, *LAMP2*, *LAMP3*, *LRRK2*, *MAPT*, *MCCCI*, *NOS2*, *NR4A2*, *NSF*, *PANK2*, *PPARGC1A*, *PARK2*, *PARK7*, *PARK12*, *PARK16*, *PINK1*, *RAB25*, *SLC25A48*, *SLCO3A1*, *SNCA*, *SNCAIP*, *STK39*, *SYT11*, *TMEM163*, *UCHL1*, *UNC13B*, *VPS35*, *WNT3*. Appendix Table A.1 presents an overview, including human Entrez IDs and reference sources. This set represents the in-house generated set of PD associated genes.

3.3.2 PD associated genes based on expression data

Altered gene expression, which is detected as changes in level of expression at the transcriptome or exome level, may have a direct impact on the proteome or interfere

¹¹<https://www.omim.org/>

in intracellular regulatory processes. Such changes can influence both disease development as well as the overall disease picture which are both of considerable interest to the research community.

A first PD microarray meta-analysis was published by Chandrasekaran and Bonchev (2013). Raw data is available via the GEO. Two years later a second statistical meta-analysis of human brain transcriptome data by Glaab and Schneider (2015) analysed a larger number of public microarray gene expression datasets to identify significantly affected pathways in PD patients.

Three original case-control studies, as well as the first meta-analysis were considered as references for this work and details of the original studies can be seen in Table 3.2.

Table 3.2: Four gene expression microarray studies, used to obtain PD associated differentially expressed genes. "Publication" refers to the study, "Brain Region" describes the tissue that was analysed, "Array Type" gives information about the array (all Affymetrix human GeneChips covering the whole human genome). The significance threshold shows p-value and fold-change information applied during original data analysis. "Associated Genes" shows the number of genes identified in the study, "Additional Information" contains further details, "Mapped Genes to Entrez ID" refers to the number of genes which were extracted from the study and successfully mapped to a unique Entrez ID and "Sample Size" refers to the number of samples (PD cases/controls) tested in the study.

Publication	Brain Region	Array Type	Significance Threshold	Associated genes	Additional Information	Mapped genes to Entrez ID	Sample Size (PD/control)
Chandrasekaran and Bonchev (2013)	post-mortem brain samples	UI33A, UI33B	p-value < 0.01	267 Entrez IDs and Gene Symbols	full gene list was supplied after contacting the author	267	30/24
Moran et al. (2006)	post-mortem medial and Lateral Substantia Nigra	UI33A, UI33B	fold change > 1 - fold	570 differentially expressed genes - top 21 differentially expressed genes, also confirmed in a second study; top 25 records based on gene sequences and mapping to regions of established PD linkage were supplied	data (Gene Symbols) manually copied from publication	43	32/15
Simunovic et al. (2009)	post-mortem (isolated) Substantia Nigra dopamine neurons	UI33A	p-value < 0.005; 50 top records: fold-change > 3; 375 genes fold-change > 1.5	1048 records in total - filtered for top scoring ones	data (Gene Symbols) manually copied from supplementary file	335	11/11
Zhang et al. (2005)	post-mortem Brodmann's Area 9, Putamen, Substantia Nigra	UI33A	p-value < 0.05; fold-change was > 1	published top 50 records for each brain area and combined top changed genes list	data (Uni Gene IDs and Gene Symbols) manually copied from supplementary Table II	85	15/15

All samples were post mortem, and information regarding the tested brain region from which human tissue was obtained, the used microarray chip (all Affymetrix human GeneChips covering the whole genome) as well as a summary of the findings are

presented. The number of significantly differently expressed genes associated with PD per study can be found in column “Mapped Genes to Entrez IDs”.

Based on the four studies, 667 unique genes (based on Entrez IDs) were identified to be significantly differently expressed in PD patient's brains compared to healthy controls. As visualized in Figure 3.2 the majority of genes is only associated with PD in one study. None of the records could be replicated in all four studies, and only 57 genes (~8.5% of all genes detected via microarray studies) are found in at least two datasets.

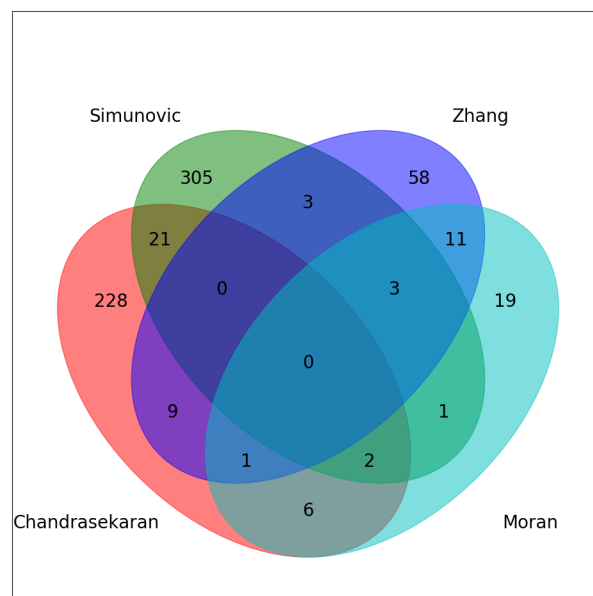


Figure 3.2: Venn Diagram showing the overlap of genes significantly associated with PD (based on Entrez ID count). The different coloured ellipses represent Entrez IDs that have been associated with PD based on a microarray expression study. The four compared studies are: Chandrasekaran and Bonchev (2013) (red), Simunovic et al. (2009) (green), Zhang et al. (2005) (blue) and Moran et al. (2006) (turquoise). Numbers in overlapping regions indicate genes found in one or more studies.

3.3.3 PD associated genes with genetic and/or manually curated evidence

The genome of many PD patients shows alterations. As outlined in Section 1.5.3, topONTO was the tool of choice to retrieve PD associated genes. This decision was based on the principle of retrieving reviewed, curated and high quality disease-gene association data. The ability to retrieve information from several databases as well as meta-data allowed for best possible data screening, filtering and curation. The coming paragraphs explain required steps.

The Disease Ontology (DO)¹² was developed to associate unique identifiers to human diseases (Schriml et al., 2012; Kibbe et al., 2014), referred to as Disease Ontology Identifiers (DOIDs), and was used to obtain a list of genes associated with PD. The unique DOID for PD is “14330”. It describes PD as a “synucleinopathy” which is classified as a “neurodegenerative disease”. Seven child terms are associated to PD and can be seen in Figure 3.3.

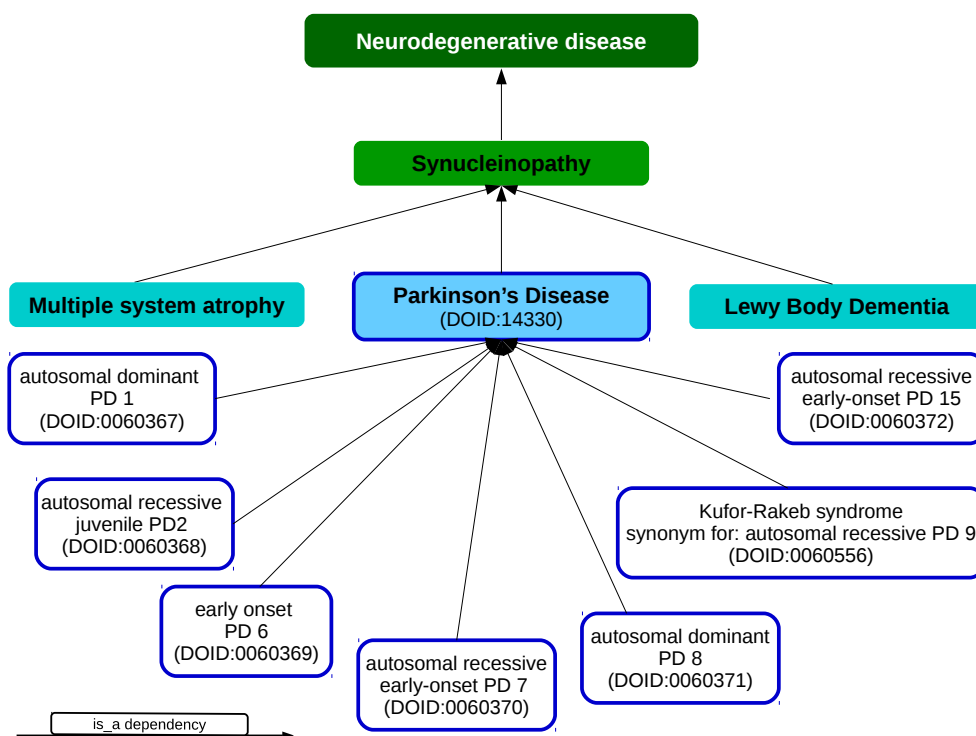


Figure 3.3: DO graph showing PD with its parent and child terms. DOIDs indicated in brackets are the official disease identifiers, used to extract associated genes. PD as well as all subtypes, indicated in boxes with blue borders are used in the analysis.

¹²<http://disease-ontology.org/>

Since top_{ONTO} queries data from different sources: (i) EnsVar (ii) GeneRIF and (iii) OMIM, these were considered separately and jointly. Depending on the source, gene-disease associations are only made if the exact DOID is associated with the gene, but not the parent term. This means that some genes associated with a disease subtype are not associated with the parent disease term (PD) itself. To obtain a full gene set individual searches were carried out for all PD subtype DOIDs, as identified in the DO tree.

After obtaining the results, data was manually checked. A number of irregularities were spotted. E.g. it was not possible to directly map the Entrez ID 401884 to a gene symbol. Closer analysis showed that it refers to a discontinued NCBI entry which is now Entrez ID: 147081 and included in the dataset. Therefore the discontinued record was deleted from the list. All other entries identified by top_{ONTO} could be confirmed and Table 3.3 shows that 635 genes were associated with at least one of the DOIDs. Source database specific information is shown in different columns. As indicated by the numbers in brackets, only two disease subtypes show genes specifically associated with them, but not directly to PD. In both cases, autosomal recessive juvenile PD 2 and autosomal recessive early onset PD 7, all hits were retrieved from the GeneRIF database.

Table 3.3: Number of genes associated with PD based on the top_{ONTO} query. Columns refer to the different source databases and “Disease (Subtype)” refers to the different PD subtypes (see Figure 3.3). Numbers in brackets refer to the number of genes associated with only the disease subtype but not PD itself.

Disease (Subtype)	Ensembl Variation	GeneRIF	OMIM	All Sources (joint)
PD	290	372	19	620
autosomal dominant PD 1	0	5	0	5
autosomal recessive juvenile PD 2	0	22 (14)	0	22
early onset PD 6	0	3	1	3
autosomal recessive early onset PD 15	0	2	0	2
autosomal recessive early onset PD 7	0	7 (2)	1	7
autosomal dominant PD 8	0	3	0	3
Kufor Rakeb syndrome	0	0	0	0
All Disease Types (joint)	290	388	21	635

Databases follow distinct annotation approaches for generating the different gene sets. To identify the overlap between PD associated genes depending on the reference source which detected them, Figure 3.4 shows their overlap.

As illustrated, the number of genes appearing in all three databases is rather small,

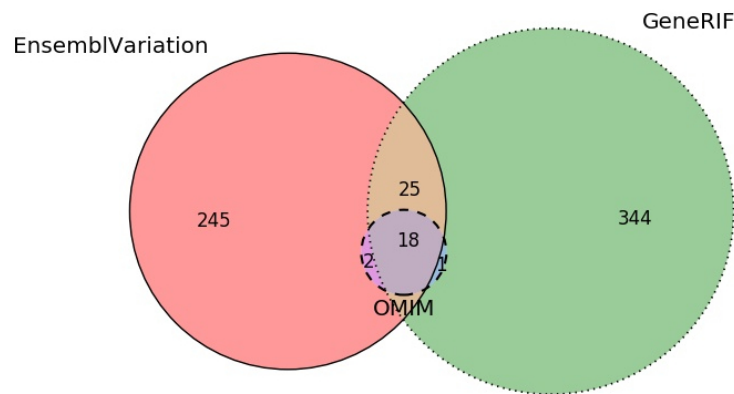


Figure 3.4: Venn Diagram showing the overlap of genes significantly associated with PD based on data retrieved with topONTO; PD subtype term results are included.

containing only 18 records. This phenomena can be explained by the differences in nature of the annotation mechanisms of the databases and distinct focus on the reference data.

As outlined in the Section 3.2.2, OMIM is based on text information regarding Mendelian disorders. OMIM genes are directly associated with PD or its respective subtypes. Due to well curated expert annotation information the data is of highest quality. With 21 PD associated genes based on OMIM data, this group makes the smallest part of topONTO queried results. All identified genes are also found in at least one of the other two databases, confirming the high annotation quality.

GeneRIF is also based on an open system, describing genes in terms of function and gene-disease association. A total of 388 genes are associated with PD based on the GeneRIF database. Compared to OMIM, annotation terms vary more widely. GeneRIF uses the following “term_names” to retrieve PD associated genes: “park1”, “park15”, “park2”, “park6”, “park7”, “park8”, “parkinson disease”, “parkinson disease (parkinson’s disease)” and “parkinson’s disease”. Manual curation showed that the “park-x” terms refer more likely to the *PARK* genes and are often analysed in a different context. In several cases text containing a “park-x” associated genes describes associations to e.g. a cancer risk factor. For best data quality all genes associated with PD based on a “park-x” term were excluded from the GeneRIF gene set. Additionally one discontinued entry was identified (Entrez ID 23707) and removed. This reduces the GeneRIF dataset from 388 to 372 genes, all of which are associated with either “parkinson disease” or “parkinson’s disease”.

As Figure 3.8 illustrates, those 15 “park-x” associated entries were only present in

the GeneRIF database.

Overall, PD is a well defined term. Since it represents an unambiguous concept text-mining approaches tend to be highly reliable, and errors are unlikely. Additionally the above filtering step makes GeneRIF data more concise by only targeting PD as a disease and not genes associated by name. Nevertheless negative associations can be part of the descriptions. For example a text annotation such as: “gene A is NOT related to PD” or “there is NO association of gene A to PD”. Hence GeneRIFs text annotations were manually filtered, by reviewing all entries containing either “NOT” or “NO”. Based on manual inspection of the descriptions, 10 genes were excluded due to only negative association to PD (Entrez IDs: *CIQA*, *CIQB*, *DRD4*, *GSTA4*, *HCRT*, *HFE*, *IL10*, *PSMC1*, *STX6* and *TLR9*. All those were only referenced in the GeneRIF database. Some more negative associations were identified, but they all showed at least one positive association based on a different reference and were kept in the disease-gene datasets.

This filtering step leaves a total of 362 genes retrieved from the GeneRIF data base. The overlap with the other datasets and final numbers are addressed later on (Figure 3.8).

EnsVar is based on identifying genetic alteration, such as SNPs. In this case, `topONTO` extracts SNP-disease associations. Based on the SNP location disease associated genes are identified.

This approach is required due to the raw data EnsVar supplies. Associating SNPs to genes is not an easy task and `topONTO` uses a relatively straight forward approach. EnsVar supplies the position of a SNP on the genome. That information is compared with gene locations (gene coding sequences (CDSs)) were retrieved from Ensembl gene, via biomart). Genes associated with a SNP are (i) overlapping with the SNP position, (ii) the closest upstream or (iii) the closest downstream gene relative to the SNP location. Figure 3.5 illustrates the three situations.

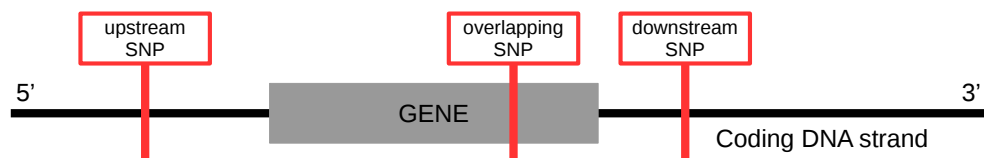


Figure 3.5: SNP-gene association classification of SNPs extracted from the EnsVar database, based on their position on the genome relative to the gene.

Figure 3.6 shows that around one third of all previously identified genes contain at least one SNP that is overlapping with the CDS (“overlapping SNP”). Another third of the genes is associated with PD by virtue of being either the closest up- or downstream one (“upstream SNP”, “downstream SNP”). Four genes are identified with an overlapping as well as up- and downstream SNP linked to the them. These are: *LRRK2*: leucine-rich repeat kinase 2 (120892), *CRHR1*: corticotropin releasing hormone receptor 1 (1394), *TMEM175*: transmembrane protein 17 (84286) and *MAPT*: microtubule associated protein tau (4137), all very well known to be associated with PD. Observing three different SNPs associated with those genes might confirm their high impact on the disease, meaning that distinct genetic alterations impact on the same gene.

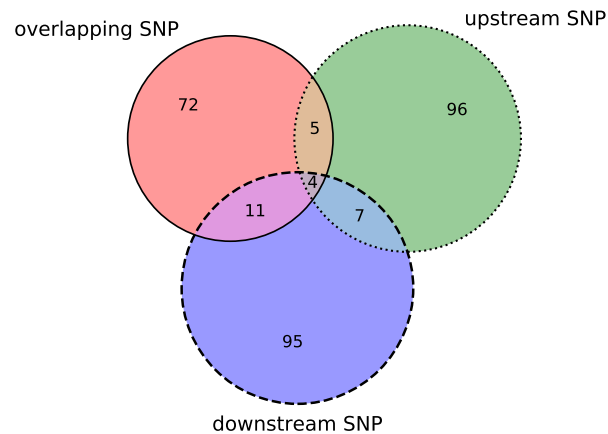


Figure 3.6: Venn diagram showing genes associated with PD based on a SNP (derived from EnsVar). The different circles indicate the relative position of the SNP to the gene (Figure 3.5). Numbers refer to gene numbers and one gene can be affected by several SNPs.

Spatial information of the SNP relative to the gene is of interest since it can contain more details about the potential effect on the affected gene. Figure 3.7 shows a more detailed gene overview, visualizing the CDS. Based on the Ensembl glossary definition the CDS only consists of protein coding sequences, exons. Nevertheless the spatial comparison only considers start and stop position of the CDS meaning that an overlapping SNP can also be located in an intronic, 5' or 3' UTR (untranslated) region.

The exact SNP location determines the effect on the protein. A genetic variant (SNP) in an exon can directly affect the protein, through alteration of its DNA sequence

and the produced protein. Alterations in an intronic or 5' and 3' UTR do not always have to show a direct effect, but most likely do so. Overall SNPs lying in a gene coding region are relatively well studied and their effect on the protein can be analysed more easily compared to SNPs located elsewhere. In fact, it is almost impossible to screen the effect of SNPs not overlapping with exons in the CDSs in an automated manner. Mutations allocated in an intron might not show a direct effect on the protein but can do so, e.g. through the generation of additional splicing sites, and hence modify the gene product and transcribed protein. SNPs in the 5' and 3' UTR are more likely to show effects on the gene expression regulation if at all.

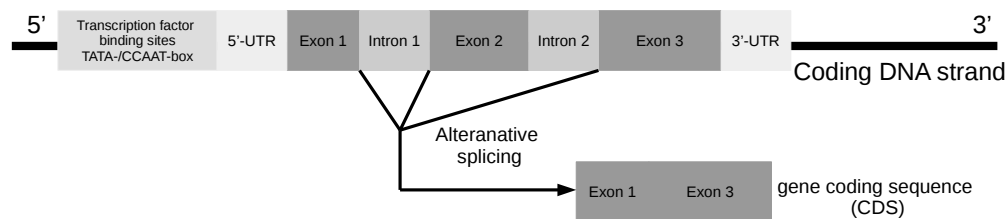


Figure 3.7: Schematic illustration of gene components, highlighting the gene coding sequence.

SNPs that are located up- or downstream of the associated gene can show a different level of influence, or not influence the currently associated genes at all. Due to the complexity of the genetic code it is possible that the up- and downstream regions of a gene contain sensitive regulatory regions. If these are affected by the SNP, gene expression levels can be influenced. This is true if e.g. the transcription factor binding sites, such as a TATA-/CCAAT-box is affected (Figure 3.7). Since such details are not widely annotated, they were not considered for further data filtering, but can be considered in individual cases.

Based on this knowledge it was decided to currently only consider genes showing mutations with at least one overlapping SNP for further studies. This means that 198 PD associated genes based on non-overlapping SNPs are excluded.

After analysing gene-disease associations separately, highly trustable datasets were combined. These are:

- all 21 positive entries from the OMIM search,
- all 92 genes associated with overlapping SNPs (EnsVar) and
- the filtered list of 362 GeneRIF results.

The joint set is constituted of a total of 418 genes. The core PD associated gene set is highlighted in Figure 3.8. Numbers labelled with a * belong to the set of 418 genes. This will be the reference set of PD associated genes based on best possible data curation methods and highest data quality.

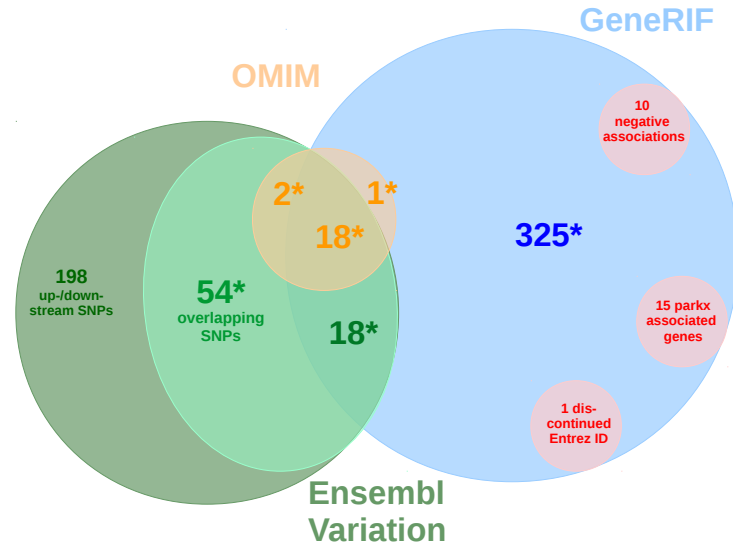


Figure 3.8: Venn diagram summarising all the filtering steps of the data retrieved with topONTO. Gene numbers with a star refer to gene sets that are part of the final PD associated gene set.

3.3.4 Meta-analysis

After having identified, verified and cleaned all PD associated gene sets the overlap between the datasets was analysed. Three gene sets were considered:

- the manually curated literature list with 52 PD associated genes (Section 3.3.1),
- the full list of 667 PD associated genes based on microarray studies (Section 3.3.2) and
- the fully filtered list of 418 genes associated with PD based on a filtered topONTO query result (Section 3.3.3).

A total of 1055 unique genes were identified. As Figure 3.9 shows, different sources lead to different results, showing a very small overlap between the datasets.

Only 10 genes appear in all three sources. These are: *APP*, *ATP13A2*, *HTRA2*, *MAPT*, *NSF*, *PARK7*, *PINK1*, *SNCA*, *UCHL1*, *WNT3*. Table 3.4 shows an overview

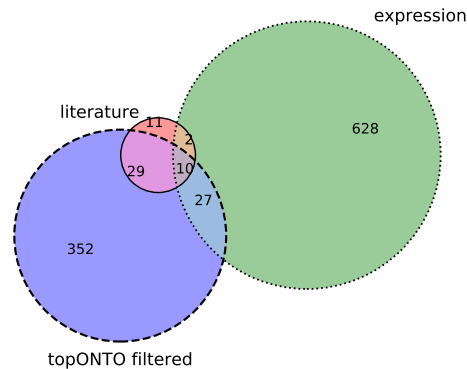


Figure 3.9: Venn diagram showing the overlap of PD associated gene sets retrieved from different sources.

of these with additional detailed information. Regarding the genetic evidence, all but one gene are found in EnsVar, with an overlapping SNP. *APP* is the one gene not to be found in EnsVar, but showing a very strong disease link based on its GeneRIF annotation. Four genes are also found in the OMIM database, further confirming the high disease association and genetic link.

Genes associated with PD based on expression differences show that *NSF* was found in three out of the four considered studies. Seven records were found by Simunovic et al. (2009) and four by Chandrasekaran and Bonchev (2013). Genes *MAPT*, *SNCA*, *PARK7* and *PINK1* have also been identified in more than one manually annotated study.

11 genes appear only in the PD literature dataset. These are (in alphabetical order of gene name): *ACMSD* (reference pubmed ID: 22438815, 21812969), *CACNA1D* (23771339), *CALR* (23771339), *HIP1R* (22438815, 22786590), *ICAM1* (18044695), *MCC* (22438815), *RAB25* (22438815), *PANK2* (22806825), *SLCO3A1* (21812969), *SLC25A48* (21812969), *TMEM163* (22438815). Another two PD literature based associated genes only overlapped with genes showing PD association based on altered transcription. These are: *CALM1* (23771339) and *CDH8* (21812969).

Due to the large differences between the PD associated gene sets the significance of their overlap was analysed. A hypergeometric test was used to identify significance in the overlap of PD associated genes based on the filtered topONTO query, compared to the PD associated genes based on the four microarray studies. A p-value was calculated, as described in Section 2.3.1. The background dataset contained 20,000 genes

Table 3.4: PD associated genes referenced in all three sources (ordered numerically by Entrez ID). “Genetic Evidence” can be “E” for EnsVar, “G” for GeneRIF or “O” for OMIM; “Microarray Study” is “C” (Chandrasekaran and Bonchev, 2013), “M” (Moran et al., 2006), “S” (Simunovic et al., 2009) or “Z” (Zhang et al., 2005). “Literature Reference” lists the pubmedID of the paper(s) where the PD link was recorded.

Entrez ID	Gene Symbol	Genetic Evidence	Microarray Study	Literature Reference
351	<i>APP</i>	G	S	22438815
4137	<i>MAPT</i>	E, G, O	S	22438815, 22806825
4905	<i>NSF</i>	E	M, S, Z	21812969
6622	<i>SNCA</i>	E, G, O	C, S	23380027, 20495568, 22438815, 21812969, 22786590, 21412835
7345	<i>UCHL1</i>	E, G, O	S	23380027
7473	<i>WNT3</i>	E	C	21812969
11315	<i>PARK7</i>	E, G	S	23418303, 23380027, 20495568, 22581678, 21812969
23400	<i>ATP13A2</i>	E, G	C, M	23380027
27429	<i>HTRA2</i>	E, G, O	C	20495568
65018	<i>PINK1</i>	E, G	S	23380027, 20495568, 22581678, 21812969

(all human protein coding genes, (Ezkurdia et al., 2014)) and the dataset of interest with 418 PD associated genes (core PD associated gene list). The sub-sample considered contains 667 PD associated genes based on microarray studies. 37 genes were found in both datasets. The hypergeometric test showed that the probability of finding 37 successes or more in the sub-sample given the indicated background and dataset size is: $P(X \geq 37) = 2.30 \times 10^{-08}$. This is a highly significant overlap, indicating that the PD associated datasets are not unspecific, but potentially cover different disease aspects.

3.3.5 Summary

Important differences in the PD associated datasets, obtained depending on input data and applied methods, were identified. A decision regarding which datasets to take forward as a reference was needed. As stated previously, the objective of this part of the study was to obtain the most concise PD gene set possible.

The results presented highlight the differences between available datasets. This variety reflects the complexity of PD and confirms what was outlined in the Central Dogma of Molecular Biology (Figure 1.1 A).

Since the datasets presented may cover distinct disease aspects, disease states, disease types or others a core PD datasets of interest needed to be selected. Genomic alterations are generally associated with underlying disease causes and triggers, whereas changes in the transcriptome and subsequently proteome most likely reflect the dis-

ease phenotype. Additionally, manually reviewed and annotated data is more secure and linked to a direct disease association. Based on this understanding, the most reliable dataset describing directly disease associated PD genes was selected.

The PD associated genes based on literature search nicely coincide with parts of the other datasets, with a tendency towards genes retrieved through the `topONTO` query. Eleven genes have not been identified in other studies. This gives an interesting insight and shows that several individually identified genes seem to have a disease link that has not yet been discovered in any large-scale study. This phenomena might be explained by the very detailed focus of experimental studies addressing pathways very specifically. Such studies are also able to detect genes/proteins with a rather minimal influence on the disease which can not be picked up in more generic settings.

Raw data from microarray studies was available, but time was too short to reproduce the published findings and confirm consistency in significance of the detection levels. Due to the variability in the published results, especially regarding the significance thresholds (Table 3.2), it was decided to use this dataset as a “wider” description of the disease picture, with a tendency towards capturing the disease phenotype and pathology. The fact that its overlap with the genes obtained via `topONTO` is significant, based on hypergeometric testing, proves a clear link of both datasets with PD. Hence, the possible indirect relationship between the two datasets that might emerge based on molecular regulatory mechanisms, adds great comparative value to the gene set obtained through microarray experiments. It could be considered a great source for comparison with new hypothesized PD associated genes later on in this study. Furthermore the minimal overlap of results between datasets was surprising. This finding can be caused by different aspects. Tissue extraction, preparation, experimental setups, and data analysis can be some of them. Additionally the results might cover different disease types or stages. Therefore all identified genes might play a role in PD but more studies are required to confirm these links. To maintain a general focus on disease causative genes and given the low coverage for most of the records identified in the microarray studies, this dataset was excluded from the key PD associated genes.

In summary, the key PD associated gene set includes PD associated genes deposited in curated databases and the ones with an overlapping SNP in their CDS. Regarding data retrieved from `EnsVar`, `GeneRIF` and `OMIM` (all obtained via `topONTO` a number of filtering steps were taken to obtain the “best possible” dataset (Section 3.3.3)). This leads to a core gene set which will be considered for further analysis in the following chapters. The “key PD associated gene set” contains 418 PD associated genes. 37

and 12 genes in the set overlap with the PD microarray evidence dataset and the PD literature dataset respectively. 10 genes can be found in all three datasets (Figure 3.9).

For comparative purposes, the set of 1055 PD associated genes is relatively broad but of good value to be used to gain a more general PD overview. It also represents a valuable reference source, especially with regard to the microarray records. These might likely capture PD associated genes having the potential to explaining the disease “phenotype”, supporting functional conclusions at a later stage of this study.

3.4 Discussion

A lot of effort has been put into identifying and understanding underlying causes of PD. A growing number of publications use large-scale approaches to gain wide understanding of this complex, neurodegenerative disease. Experimentalists mainly ask specific questions and analyse individual candidate processes in detail. This study aimed to identify a key set of genes describing the genetic and molecular PD complexity on a large scale. The endeavour was addressed through integration of PD associated data from different databases, covering specific disease aspects.

Data was retrieved and filtered to obtain the most concise and comprehensive dataset covering different aspects of the disease. Even though datasets seem quite different a small, but significant overlap between datasets capturing PD associated genes based on genetic and curated information versus gene expression alterations was identified. This indicates that datasets are specifically describing common aspects of the disease, but very likely mutations are not directly reflected on the gene expression level. Different approaches are possibly capturing distinct disease aspects and are biased towards the detection sensitivity they have towards a certain set of disease associated genes.

Different insights can also emerge due to differences in experimental material and analysed tissue. Extraction, preparation and further conditions such as post-mortem tissue processing, biopsy techniques and the use of specific tissue parts can highly influence results. This can also be considered a positive point, since it might capture a wide range of PD subtypes covering a large variety of factors to be considered to capture the full disease picture. Nevertheless this means that current datasets can show biased results, depending on their strengths and weaknesses of the underlying detection approach, leading to the encountered differences.

Overall differences between `topONTO` and microarray results are likely to reflect

the influence of the genes on the PD genotype and phenotype.

The literature based dataset only supplies very superficial insights and one would suspect a bias towards genes and proteins that are more easily studied and experimentally analysed. They might also show a longer detection history and/or well known disease links. This dataset is far from complete but shows higher overlap with genes with a genetic and manually curated link to PD. It is difficult to judge if genes are analysed based on previous experimental findings, their appearance in large GWAS studies or based on therapeutic potential. In any case this approach gave a good first insight to the field and could suggest experimental approaches, mostly addressing genes showing mutations and possibly triggering the disease without directly influencing the disease phenotype.

Considering the vast differences between the results obtained in the different microarray studies, obtaining experimental material is a crucial step to ensure data quality. When working with human brain tissue, this is specifically challenging, since it involves collecting post-mortem samples. Depending on tissue extraction procedures and timing, certain intracellular degradation processes might have become active in brain cells and influence the results. Nevertheless this is a phenomena that affects all studies of this type and can be counteracted by guaranteeing a maximum time between death and obtaining the tissue. Differences in the dataset size reported in the distinct studies can be partly explained by individual significance thresholds, the material, detection and analytical sensitivity. Further aspects influencing microarray expression study analysis are analytical procedures and thresholds. Since these highly influence the results they introduce bias and complicate cross comparison of results amongst studies. Sample and data quality differences can emerge at various levels also considering technical setups. The selection of test samples and controls is specifically challenging considering complex diseases, such as PD. The joint consideration of patients with potentially distinct disease types might lead to confusion of disease associated gene expression differences. Similar effects can happen in case controls are affected by undetectable alterations highly specific to individuals and modifying significance in obtained results.

All these challenges represent limitations of the use of microarray expression study results in other projects, not diminishing the information it can contribute to the disease understanding as an additional reference and/or as support to distinct research questions. They reflect an additional point why this data was excluded in the final PD associated gene set

In the curated datasets, the annotation specificity of GeneRIF adds an additional level of information, but also a source of error. Human annotation and interpretation error can lead to false positive records in databases. Mapping genes only to PD subtypes but not to the parent PD term might lead to their exclusion. This also reflects the properties of the GeneRIF database, where associations between gene and diseases are not automatically propagated to parent nodes in the ontology tree. Nevertheless, in-depth detailed insight and associations considering disease subtypes can be very valuable and need to be considered when carrying out large-scale studies. To counteract these phenomena the presented study explicitly included information covering all disease subtypes.

Using data with a genetic link retrieved from EnsVar is the most secure approach to use when drawing conclusions regarding the PD genotype. The further filtering step, screening results for genes with CDS overlapping SNPs additionally support this link and should be considered before drawing conclusive decisions. For further certainty a number of other aspects could be considered. The aforementioned (in-) direct effect of SNPs and their effect on the affected gene, apart from the SNP position, can be influenced by the number of alterations found in one gene. This number would also need to be interpreted depending on the gene length. This is just one additional example of the drawbacks in using relatively direct gene-disease association approaches. A number of approaches are being developed to describe the effect of genetic alterations on a gene in a score based system. Such an approach can further support gene-disease links and improve data quality which could be considered in future studies.

As far as can be ascertained, this is the first study directly comparing disease associated data retrieved from different sources and capturing distinct PD aspects on a large scale and in this level of detail.

The PD map (Fujita et al., 2014)¹³ is a joint effort addressing a similar question and trying to build a set of “all” genes associated with PD. The data is publicly available and accessible in a very interactive way. Nevertheless included genes are not all evidence based, or evidence is not shared with the user, making it very hard to understand the strength of a gene-PD link in depth. As this chapter shows, such information can have a large influence in the reproducibility of results and likely the type of gene-disease-link. Therefore the presented manually curated gene set was preferred for the available details.

The findings also illustrate that in the case of PD, as probably with any other com-

¹³<http://minerva.uni.lu/MapViewer/>

plex neurodegenerative disease, datasets should be treated individually when required (which is not possible in the case of the PD map). Doing so will avoid wrong conclusions being drawn, and allows detection of “real patterns”, focusing on e.g. the genotype or phenotype of a disease.

Overall, a key advance presented in this chapter is the combination and thorough analysis of data from multiple data sources and different levels in the biological machinery, defining the disease picture. More importantly manual curation steps were performed to confirm and/or discard initially detected gene associations to PD, leading to a high quality set of PD associated genes.

Even though a gene set was obtained, further detailed disease insights are missing. The complex disease pattern suggests that a range of different PD subsystems are affected by several genes in the dataset. Those are very likely to be connected and/or influencing each other, making the picture even more complex. Since individual gene analysis does not allow such links to be detected, these need to be analysed and understood on a large scale. For this purpose Protein-Protein Interactions are of most interest, as they allow cross-links to be established between affected genes, likely involved in the disease. The coming chapter introduces the concepts behind such an approach and following chapters introduce their effective use in Protein-Protein-Interaction Networks and draw first conclusions regarding the PD complexity.

Chapter 4

Protein-Protein Interaction Data

4.1 Objective

Even though an increasing number of databases supply Protein-Protein Interaction (PPI) information in a standardised format, most researchers obtain their PPI data from just one of the available ones, running the risk of ignoring crucial information deposited in other repositories. In order to avoid such loss of information this chapter merges content from multiple databases. This endeavour allows differences and similarities among datasets used, to be identified. These insights can then be used to set adequate filters to obtain a clean, human-only unified reference PPI dataset. Processes such as mapping of gene and protein identifiers as well as joining datasets are performed using available bioinformatics tools and methods.

Hence, this effort aims towards obtaining a clean, curated, human-only PPI set, which is a valuable source for further analysis, guaranteeing highest quality of results and bridging the gap between data deposited in different databases. Figure 4.1 shows an overview of sources, techniques and the outcomes of this chapter.

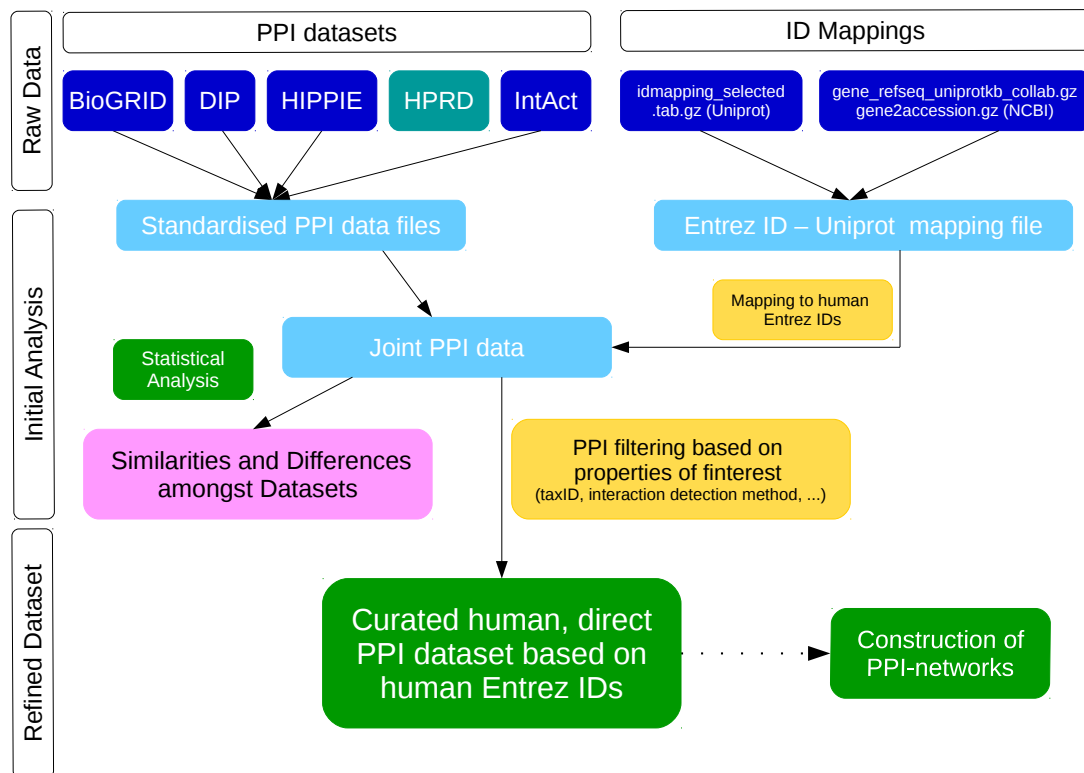


Figure 4.1: Overview of work presented in Chapter 4. Input databases are shown in dark blue boxes (turquoise represents a special case addressed in the text and excluded in the final PPI set). Light blue stands for newly generated and curated datasets. A yellow box refers to processes, leading to an analytical result (pink boxes). Green boxes represent outputs of this chapters analysis or future results.

4.2 Introduction and Data Processing

4.2.1 Protein-Protein Interactions

Interactions of proteins as well as interactions between proteins and other biomolecules, are crucial for any process happening within and between cells. Only interactions can trigger signals, initiate enzymatic processes and release interaction cascades, through activation, inhibition, or other processes. Some of the most common ones appear between e.g. enzyme and inhibitor or antibody and antigen but a large number of other interactions are required for cells to fully function.

Two types of PPI interactions can be easily distinguished: transient or stable ones (Perkins et al., 2010). Transient interactions lead to specific effects in a short time width, whereas stable ones usually lead to more permanent multiprotein complex for-

mation. For example, the formation of a clathrin cage, involved in endocytosis, as well as most reactions in signalling pathways are transient interactions (Ozbabacan et al., 2011). In the clathrin case cages build up, carry out their function and disassemble. Since this is not a permanent state, interactions are considered as transient. In contrast the ribosome, a large macromolecular complex and the gamma-aminobutyric acid type A receptor (*GABAA*) rely on highly stable and permanent PPIs. In both cases it is the joint protein complex that carries out a function and can only do so as a fully assembled union. Such complexes can vary in size and most often act as molecular machines in living systems. Their constitution is referred to as quaternary structure, describing interacting domains and structural relationships between individual proteins (Yu et al., 2006).

The transition between the two types is difficult to define and it is relatively hard to experimentally identify the interaction type. Nevertheless depending on detection approach used, physical interactions between protein pairs can be classified based on standardised interaction types. This point is addressed and discussed later on in this chapter (Section 4.3.1).

An additional challenge is the specific identification of binary interactions, based on only two elements, as opposed to the presence of two elements in one complex, not undergoing a direct interaction. Some databases maintain information including this distinction, but the majority do not. Often, if experiments supply data based on complexes these are “spoke expanded”, meaning that all possible protein pairs in a complex are considered as PPIs (Gingras and Raught, 2012). In particular this is specifically the case when retrieving data through so-called pull-down experiments. This practice might not be the best approach but is widely accepted and frequently used on a large scale.

Various high-throughput techniques exist to extract PPIs on a large scale. Amongst these are yeast-two-hybrid, pull-down and co-localization studies (Berggård et al., 2007). Experiments can be carried out with proteins from different species and even cross species, meaning that for example human proteins are expressed in a mouse cell line (*in-vitro*). Different approaches are more suited to identify certain types of PPIs. An exhaustive, recent review explains the different techniques (Wetie et al., 2013). Additionally, computational approaches can predict human PPIs based on structural similarities or occurrence in other species such as mouse and rat. These rely on homology and interolog mapping (Folador et al., 2014).

Nowadays most published PPI studies submit their data to at least one of the major

databases, or databases identify new publications and include PPIs in their repositories. To combine data from multiple sources additional analysis is required. The following section explains the most commonly used, current standard format put in place to store data and facilitate its analysis.

4.2.2 Data Format

The mitab25 format (following PSI-MI standards) allows researchers to access published PPI data in an easy and automated way, allowing direct integration into workflows. HUPO, the Proteomics Standards Initiative introduced the PSI-MI TAB format for data storage and interchange in a tab delimited format (Hermjakob et al., 2004a; Kerrien et al., 2007)¹². As such it is under constant review to serve the scientific community as required. To follow the format's standards a minimum of 15 standard columns are required with each of them containing predefined content. An overview is given in Table 4.1. Further columns can be added but are not required.

Table 4.1: 15 standard mitab columns together with their content, including an example (randomly selected, not consistent between different columns).

Column	Content	Example
1 & 2	interactors	entrez gene/locuslink:84665; uniprotkb:P49418
3 & 4	alternative IDs	biogrid:124185; entrez gene/locuslink:MYPN; intact:EBI-7121510; uniprotkb:Q75MK5; intact:MINT-109264
5 & 6	interactor aliases	entrez gene/locuslink:CMD1DD(gene name synonym); psi-mi:amph_human(display_long); uniprotkb:AMPH(gene name); psi-mi:synj1_human(display_long); uniprotkb:Synaptic inositol 1,4,5-trisphosphate 5-phosphatase 1(gene name synonym)
7	interaction detection method	psi-mi:"MI:0018"(two hybrid); psi-mi:"MI:0084"(phage display)
8	first author, reference publication	"Bang ML (2001)"; Cestra et al. (1999)
9	publication identifier	pubmed:11309420; mint:MINT-5211933
10 & 11	taxon ID of interactors	taxid:9606; taxid:9606(human); taxid:9606(Homo sapiens)
12	interaction type	psi-mi:"MI:0407"(direct interaction)
13	source database	psi-mi:"MI:0463"(biogrid); psi-mi:"MI:0471"(MINT)
14	interaction identifier	biogrid:117; intact:EBI-7121552; mint:MINT-16056
15	confidence score (if available)	-; intact-miscore:0.56

Some of the meta-data can be standardised through the use of MI-IDs. MI-IDs

¹<http://www.psidev.info/molecular-interactions>

²<https://psicquic.github.io/MITAB25Format.html>

are widely used identifiers, defined in different ontologies, specific to the column content and available for information supplied in mitab25 file columns 7 and 12-14. The “source database” for example can be referred to with a (standardised) name or an MI-ID. Many of such mappings can be retrieved from the Ontology lookup service³ (Jupp et al., 2015).

This meta-data can be used to filter interactions based on their properties which can lead to a more specific set of PPIs of interest. Based on the standard mitab25 format these filtering steps can extract interactions, which are e.g.

- (i) detected in one specific species (through the use of taxIDs, columns 10 and 11),
- (ii) obtained via the use of a specific experimental approach (defined by interaction detection method, column 7),
- (iii) described with a specific interaction type (specified in column 12) and/or
- (iv) extracted from a specific source database (listed in column 13).

These steps allow the level of data “cleanliness” and “certainty” to be increased and enable users to set personal PPI data filters based on their needs and research purpose.

4.2.3 Databases

Based on the growing amount of data, major databases gather published PPIs and make the data accessible. Five major PPI databases are, in alphabetical order (i) BioGRID (Stark et al., 2006; Chatr-aryamontri et al., 2016) (ii) Database of Interacting Proteins (DIP) (Xenarios et al., 2000; Salwinski et al., 2004) (iii) HIPPIE (Schaefer et al., 2012; Alanis-Lobato et al., 2016) (iv) Human Protein Reference Database (HPRD) (Prasad et al., 2009) and (v) IntAct (Hermjakob et al., 2004b; Orchard et al., 2013). Table 4.2 introduces them in more detail, including references, primary protein identifiers and first and most recent release dates. It can be seen that all sources, apart from HPRD supply data following mitab25 standards (see above) (Hermjakob et al., 2004a; Kerrien et al., 2007). Unfortunately the HPRD data format is incompatible with the mitab25 standards. Furthermore the last release was updated in 2010, meaning that the information contain is outdated and all entries are now most likely covered by other databases. HIPPIE lacks two mitab25 standard columns (describing the “interaction type” and “interaction identifiers”) but the data can still be combined with the other sources.

³<http://www.ebi.ac.uk/ols/index>

Table 4.2: Five of the most commonly used PPI databases. “Main Identifier” refers to the identifier used for the interactors (given in columns one and two of the psimi25 standard format files). Most recent release refers to the point of writing (May 2017).

Database Name	Reference	First available	Most recent release	Main Identifier (data format)
BioGRID	Stark et al. (2006); Chatr-aryamontri et al. (2016)	2002 (monthly release)	May 2017	Entrez ID (<i>mitab 25</i> format)
DIP	Xenarios et al. (2000); Salwinski et al. (2004)	1999 (irregular releases)	February 2017	Uniprot ID / Uniprot Accession ID (<i>mitab 25</i> format)
HIPPIE	Schaefer et al. (2012); Alanis-Lobato et al. (2016)	2011 (irregular releases)	June 2016	Entrez ID (<i>mitab 25</i> , but missing two columns: “Interaction Type” and “Interaction Identifiers”)
HPRD	Prasad et al. (2009)	2003 (irregular releases)	April 2010	Uniprot ID / Entrez ID (tab delimited xml format)
IntAct	Hermjakob et al. (2004b); Orchard et al. (2013)	2002 (monthly release)	April 2017	Uniprot ID / Uniprot Accession ID (<i>mitab 25</i> format)

Hence for the purpose of this study HPRD was excluded. Based on the selected databases Table 4.3 shows the exact data releases that were used in this work. Links to the online data repositories, specifying the datasets and releases used are listed. To analyse the development of database content and its consistency, data belonging to two different releases was considered. Data were downloaded directly from the online repositories, and are accessible via FTP servers. Since the previous HIPPIE release was published more than 18 months earlier only one HIPPIE dataset was considered.

In summary, a number of PPI databases gather protein interaction information and make it publicly available to the wider research community. The use of a uniform data format allows information to be compared and combined, while maintaining highest quality.

Based on those efforts it became easier to access and use the data in a solitary and joint manner. Nevertheless content in different repositories still varies. Cross analysis and strict filtering can help to obtain the most concise PPI dataset possible. Based on available resources and including only a minimum of false positive interactions, this chapter aims to obtain such a dataset. Details regarding the data cleaning and joining process can be found in the next section.

Table 4.3: Overview of the four databases of choice. All supply data in the mitab25 format. Two different, recent releases, per database are listed together with the data-file names.

Database	URL Reference	Dataset/File name	Version (release date)
BioGRID	https://thebiogrid.org/download.php	BIOGRID-ALL-3.4.139.mitab.zip	2016/08
		BIOGRID-ALL-3.4.147.mitab.zip	2017/03
DIP	http://dip.mbi.ucla.edu/dip/download?mst=1:2:1&tbs=0:0	dip20160731.txt.gz	2016/07/31
		dip20170205.txt.gz	2017/02/05
HIPPIE	http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/HIPPIE-current.mitab.txt	releasev2.0	2016/06/24
		releasev2.0	2016/06/24
IntAct	ftp://ftp.ebi.ac.uk/pub/databases/intact	psimitab/intact.zip	2016/08/01
		psimitab/intact.zip	2017/03/02

4.2.4 Data Curation

The use of datasets from different sources made it necessary to merge information. One of the key challenges was the mapping between gene and protein IDs that were used as primary identifiers by the different databases (see Table 4.2). This discrepancy might be one of the main reasons why most other researchers limit themselves to using data retrieved from only one of the listed resources. Nevertheless using information from all different sources broadens the insight and allows a more complete dataset to be obtained.

Therefore a number of mapping and merging approaches were carried out. BioGRID and HIPPIE use Entrez IDs as their primary interactor identifier, but DIP and IntAct base their primary interactor identifiers on Uniprot IDs or Uniprot Accession IDs. To obtain consistency, the following steps were taken after having downloaded the data:

1. Raw data files from each database were considered individually. All files were read and headers were checked to confirm the mitab25 format. Data were filtered for the taxID of interest (“9606” for human). 15 standard mitab25 columns were printed in database specific output files, using official mitab25 headers. This step was also used to extract some statistical insights (e.g. the percentage of human

PPIs compared to the full dataset).

2. Each of the preprocessed files was considered individually. Columns 1 and 2, containing the interactors, were read. Columns were checked individually. If the column content was identified as an Entrez ID, it was kept (based on an internal comparison with the mapping table, see Section 2.2.1 for more details). If the supplied interactor was not an Entrez ID it was first checked against a list of UniprotIDs. In case of successful mapping, the Entrez ID was obtained and used. Alternatively the supplied interactor ID was checked against Uniprot Accession IDs and mapped to the corresponding Entrez ID (for mapping tables see Section 2.2.1). In some cases a UniprotID mapped to a number of different human Entrez IDs, meaning that the same protein is encoded by different genes. In those cases all possible Entrez IDs were added to the PPI set. This phenomena can influence the number of total PPIs before and after the mapping step.

In case one of the two, or both, interactor identifiers could not be mapped to an Entrez ID the interaction was not included. If mapping was required, the “original” identifier was moved to columns 3 or 4 (“alternative IDs”) respectively.

In an additional step, other columns were processed simultaneously to reflect the following standards:

- (a) Columns 7 and 12, the “interaction detection method” and “interaction type” were standardised to identifiers in the “MI:number” format.
- (b) Columns 9, 10 and 11, containing pubmed IDs as well as taxIDs of both interactors were cleaned to contain only respective numeric identifiers.
- (c) Column 13, containing information about the source database was translated to the database’s MI-ID. Database name to ID mappings were obtained from the Ontology lookup service (Jupp et al., 2015) and an overview table can be seen in Appendix Table B.2.

Depending on the database additional filtering steps were applied and will be introduced in more detail later on, since these are based on intermediate results.

In summary, this step generated database specific mitab25 files with unified identifiers, Entrez IDs, and standardised column content. In columns with multiple entries, the pipe (“|”) separator was used as a divider. If protein IDs were used and mapped to several Entrez IDs these were added to the dataset as new rows,

maintaining information in the additional columns. These data can be used to identify similarities and differences between the repositories, and allow for further data processing.

3. Having translated all data to unique identifiers, data from different sets were joined. During this data merge interactors were first kept in the original column order, but later sorted by ascending Entrez ID (Entrez ID interactor 1 < Entrez ID interactor 2). Where an interaction occurred multiple times information in any of the other columns was joined and separated by “|”. Ordering of the identifiers results in a file with unique PPIs, excluding mirrored duplicates (a-b and b-a are merged to one interaction, where “ID-a < ID-b”).
4. Based on further intermediate results, other filtering steps were undertaken. These addressed certain columns and extracted full rows depending on the filtering criteria. Details are covered in the Results section (Section 4.3).

The joint dataset was analysed to understand and identify similarities and differences between data sources and to detect potential bias and underlying patterns in the data.

4.3 Results

To gain a deeper understanding of published PPIs, individual datasets obtained from four distinct databases (Section 4.2.3) were processed and analysed. Through specific data filtering a joint set was obtained. This chapter shows intermediate results, outlines filter settings and introduces the “final” PPI set further used in this study. Statistical analysis is also presented.

4.3.1 Data Analysis and Cross-Comparison

Individual datasets were analysed regarding the number of PPIs they contain. A first filter was set to consider only human PPIs. Two different data releases were considered to track development over time. Table 4.4 and Table 4.5 show source-specific overviews regarding the number of PPIs contained in the different databases. Table 4.4 refers to the most up-to-date data as of August 2016, whereas Table 4.5 presents most up-to-date datasets available in May 2017. Numbers increase slightly with the newer release, but relative proportions of data in the different databases remain the

same. Overall both tables show that the number of deposited PPIs varies widely between databases. If not further specified this chapter refers to numbers based on the most up-to-date data in March 2017 (Table 4.5).

In terms of numbers, BioGRID contains the largest amount of total entries: ~1.4 million. It is followed by IntAct with ~0.7 million entries, HIPPIE with ~ 0.27 million PPIs and DIP with ~76 thousand interactions.

Table 4.4: Overview of PPIs obtained from four databases (August 2016). Numbers represent PPI counts based on the row count in the file. Some PPIs may occur multiple times and duplicates such as (a-b and b-a are counted separately). “Human” means that both interactors were associated with the human taxid (9606). “Unique” PPIs represents the unique number of PPIs (filtered for mirrored duplicates). Direct interactions were obtained by filtering for direct-only interaction types.

	BioGRID	DIP	HIPPIE	IntAct
Rows in file (PPIs)	1,030,500	76,796	273,927	650,097
Human PPIs	298,823	5,537	273,927	122,049 + 49,005 (unassigned pubmedID - rows) + 73,999 (spoke expansion - rows)
Percent human PPIs	28.99	7.21	100	18.77
PPIs mapped to human Entrez ID	298,745	6,010	272,431	117,571
Unique PPIs mapped to human Entrez ID	216,887	5,967	271,815	61,627
Unique, direct PPIs mapped to human Entrez ID	184,648	5,957	0	60,959

Table 4.5: Overview of PPIs obtained from four databases (March 2017). See caption Table 4.4.

	BioGRID	DIP	HIPPIE	IntAct
Rows in file (PPIs)	1,381,962	76,881	273,927	718,180
Human PPIs	305,924	5,569	273,927	125,147 + 81,256 (unassigned pubmedID - rows) + 76,939 (spoke expansion - rows)
Percent human PPIs	22.13	7.24	100	17.42
PPIs mapped to human Entrez ID	305,847	6,041	272,431	120,224
Unique PPIs mapped to human Entrez ID	221,419	5,998	271,815	62,136
Unique, direct PPIs mapped to human Entrez ID	188,945	5,988	0	61,458

After filtering the data for human-only interactions, with both interactors associated with the human taxID (9606), it was confirmed that HIPPIE only contains human specific interactions. Considering the other databases (see Table 4.5), BioGRID contains the highest proportion of human-only PPIs (~22%), followed by IntAct (~17.5%) and DIP (~7%). The number of human IntAct PPIs splits up into: (i) 81,256 interactions which lack an associated pubmed ID, (ii) 67,939 spoke expanded interaction,

leaving (iii) 125,147 interactions following this study's requirements. Studies without a pubmed ID referenced an IMEx⁴ number instead. It is a joint effort across molecular data databases to provide curated information, e.g. from PPI databases. All 81,256 interactions were retrieved from two studies. These were published in 2016 and 2017 with IMexx identifiers "IM-25054" and primary pubmedID "unassigned1312"⁵ as well as "IM-25472" and "unassigned1304"⁶ respectively. It was surprising that the PPI records did not show an associated pubmedID. Detailed investigation showed that the study published in 2016 was associated with a pubmedID (27173435) (Boldt et al., 2016), based on the data description qualifier "see-also". Such a link was not available for the study published in 2017. To maintain reproducibility, consistency and allow for automated and controlled data curation, PPIs retrieved from those references were not considered.

Furthermore, spoke expanded PPIs are likely to include a high percentage of false positive entries. These were identified and excluded from the PPI set to maintain best possible data quality. After identifying those "special" cases IntAct data were filtered accordingly.

In the next step, and to obtain comparable datasets all "interactor identifiers" were mapped to human Entrez IDs. BioGRID and HIPPIE entries were already supplied as such. Nevertheless all entries were compared with records in the mapping file and only kept if both interactor Entrez IDs could be confirmed. In the case of IntAct and DIP, Uniprot IDs or Uniprot Accession IDs were mapped to Entrez IDs. As Table 4.4 and Table 4.5 illustrate most of the entries could be mapped to human Entrez IDs. Considering data in DIP it can be seen that the number of "PPIs mapped to human Entrez IDs" is higher than the number of "human PPIs". This is due to multiple mappings of single UniprotIDs to several Entrez IDs. One example is uniprotkb Q13748, representing the Tubulin alpha-3C/D chain, which maps to two Entrez IDs: 7278 (tubulin alpha 3c), and 113457 (tubulin alpha 3d). Similarly uniprotkb P86479, proline-rich protein 20C maps to six different Entrez IDs, all being different forms of the proline rich 20 genes. This intermediate step allowed further analysis carried out in subsequent steps.

Directionality in PPIs is very hard to measure and will not be considered in this work. Some PPIs are listed in both directions (a-b and b-a). In some cases this is caused by diverging meta-data associated with the entries. To obtain unique PPI counts

⁴<http://www.imexconsortium.org>

⁵<http://www.ebi.ac.uk/intact/interaction/EBI-11901113>

⁶<http://www.ebi.ac.uk/intact/interaction/EBI-13150962>

so called “mirrored duplicates” were filtered and equal PPIs were joined, keeping all meta-data. Table 4.4 and Table 4.5 show the consequences on PPI numbers after this filtering step. Remaining PPIs are referred to as “unique”. All further statistical analysis reflects only uni-directional (“unique”) interactions.

A further crucial aspect when considering PPIs and using information for data analysis is the fact that interactions can be direct or indirect. Considering the use of PPI information in this study, and since most analytical approaches rely on physical protein interactions, these should be direct. Therefore direct experimental evidence is crucial to guarantee a minimum certainty of an interaction happening under *in vivo* conditions. To obtain such information the “interaction type” was considered (column 12 of the psimi25 PPI file). Figure 4.2 shows details about interaction types associated with PPIs in the joint dataset and retrieved from any of the four source databases.

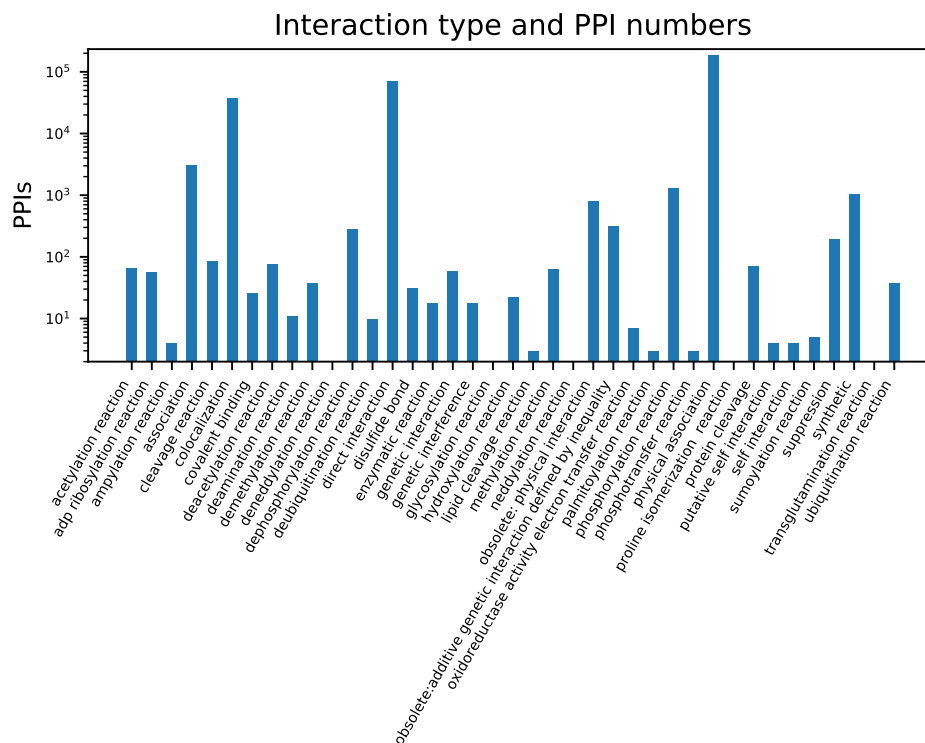


Figure 4.2: PPIs based on a certain “interaction type”. Data from all four source databases, not filtered for direct interactions is visualized (“unique PPIs mapped to human IDs” in Table 4.5). The x-axis shows the interaction type in alphabetical order.

It can be seen that a number of PPIs are considered interactions, based on protein ‘colocalization’ as well as “genetic interaction”. These interaction types, classified

as non-physical, contain a high probability of not being direct, instead proteins were detected in a shared cellular location or via “genetic interaction”. Therefore these classifiers were considered as “non-direct” interactions.

After having excluded these two interaction terms an interaction type ontology (Jupp et al., 2015) was used to identify a set of “direct” interaction type identifiers. Figure 4.3 shows an overview of interaction types part of the ontology tree. This confirmed that “colocalization” (MI:0403), “genetic interaction” (MI:0208) (including “suppression” (MI:0796) and “synthetic” (MI:0794) amongst others) and “predicted interaction” (MI:1110) were classified as being very likely to refer to non-direct interaction sets. Hence interactions only based on such evidence were excluded from the final PPI set.

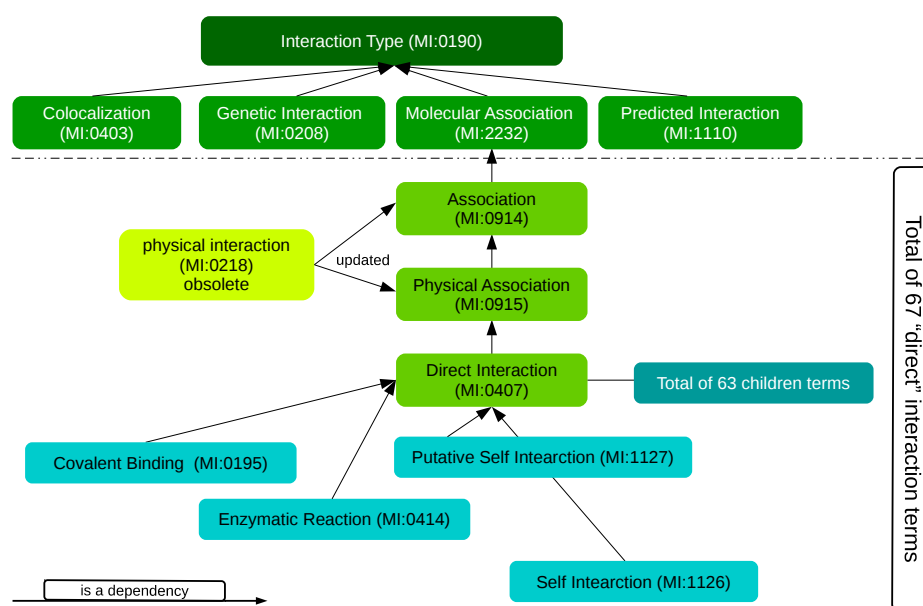


Figure 4.3: Tree structure of the interaction type ontology branch and respective MI-ID.

The term “molecular association” (MI:2232) was considered as indicating direct interactions and addressed in more detail. With respect to the ontology it is followed by the interaction types “association” (MI:0914), “physical association” (“MI:0915”) and “direct interaction” (“MI:0407”). The “direct interaction” term contains 63 specific child-terms, all indicating different subtypes of direct interactions. It was also discovered that some of the source data references the obsolete interaction type “physical interaction” (MI:0218), which was updated to “association” and “physical association”. After checking all the individual child terms referring to specific “direct interaction” types it was decided to include them all to the final list classifying direct

type”, the overlap of unique PPIs deposited in HIPPIE and the other databases was investigated. Figure 4.5 shows that 44,547 out of 271,815 PPIs are only found in HIPPIE. On the contrary, the remaining ~84% (227,268) of the HIPPIE PPIs appear in at least one of the other databases. Hence interaction type information and other meta-data can be obtained from there. Given the interest of eliminating as many false positive PPIs as possible it is preferred to not consider PPI records without “interaction type” information. This lead to the exclusion of PPIs unique to the HIPPIE database.

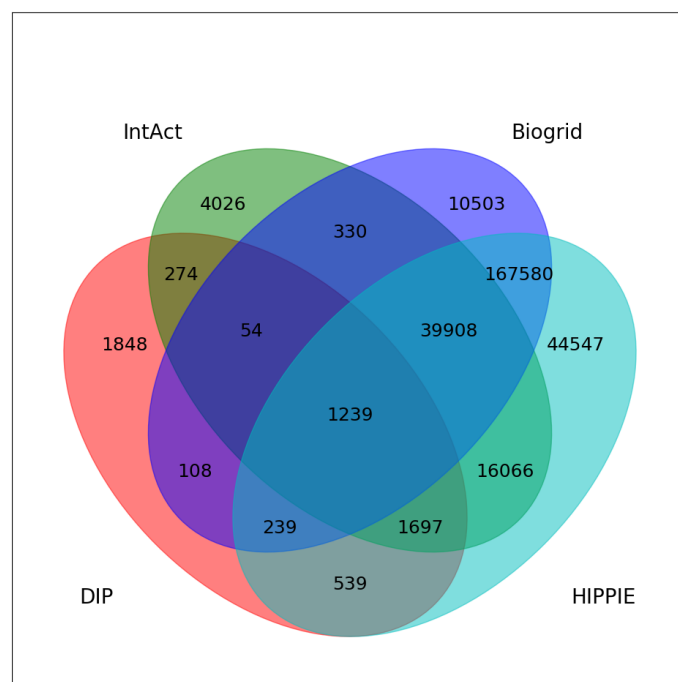


Figure 4.5: Overlap of PPIs supplied by the different databases.

To understand the development of available data, two of the most recent releases for each database were compared. Table 4.4 and Table 4.5 show that all databases but HIPPIE exhibit an increase in the number of stored PPIs. HIPPIE data were not updated between August 2016 and March 2017. Since the data were not included in the final dataset this was not further investigated. Numbers of PPIs obtained from other sources increased by varying numbers of interactions, depending on the annotation methods used by the source databases. This insight highlights the need of constant data-updating to maintain best data quality.

4.3.2 The final, joint, human PPI dataset

Based on the above insights a final PPI dataset was generated and can be found in the digital supplementary material (folder: “PPI-data”). Examples showing the raw data format as well as the final, human, unique, direct PPI lists and a README file are available. Table 4.6 shows a total of 353,294 PPIs, available with Entrez IDs. 288,958 of those are unique, and again, 211,824 are unique and direct. A number of other statistics are included in the table.

Table 4.6: PPI count depending on different filter settings.

Joint PPI set	number of PPIs (August 2016)	number of PPIs (May 2017)
Human PPIs	347,898	353,294
Human unique PPIs	284,169	288,958
Human unique direct PPIs	207,175	211,824
Number of pubmed-id references	31,386	32,271
Number of interaction detection methods	160	161
Number of interaction types	32	33
Number of source types	14	14

Hence, after applying all outlined filtering steps, the final PPI set only contains interactions with a pubmed reference and, when information was available, excludes any spoke expanded interactions. As far as can be ascertained this is the currently in existence most concise and complete PPI dataset excluding as many false positive records as possible, with a low removal rate of real PPIs. This is a major step forward towards highest data quality. The dataset was further analysed and the results are as follows.

Figure 4.6 shows the overlap of PPIs appearing in the three remaining sources. This figure also highlights the data diversity in the different databases, with only 1,264 PPIs appearing in all three source databases.

Based on that insight information listed in the “source database(s)” column of the final PPI dataset was analysed. All original sources are shown in Figure 4.7, with BioGRID and IntAct as the main references. Mint, DIP and the Uniprot knowledge base follow, together with 9 additional sources appearing with small numbers of PPIs associated with them. These additional PPI databases are mostly smaller efforts concentrating on subsets of available data. They were not considered individually since the selected databases integrate information deposited in those smaller ones. A potential explanation for the high numbers of PPIs provided by BioGRID is their relabelling of the data source to “biogrid” independently of the original data source.

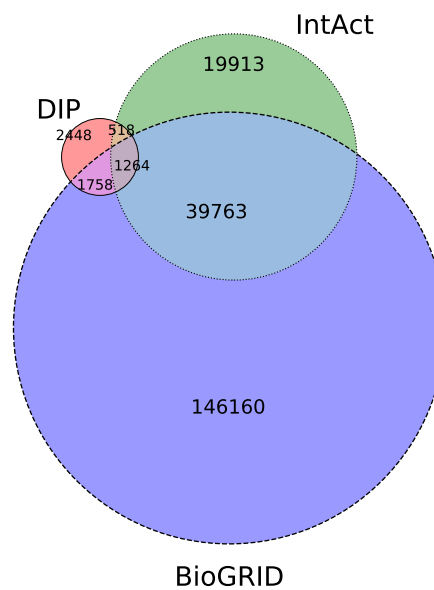


Figure 4.6: Overlap of PPIs between the three main databases that were kept to retrieve the final (human, unique, direct PPI dataset).

Based on this insight Figure 4.8 shows in how many “source database(s)” a PPI appears. The majority, almost 170,000 (~80%) PPIs are only listed in one source database, with a maximum of six (out of 14 possible databases, see Figure 4.7).

To further test PPI detection coverage, PPIs are checked for the number of different publications they appear in. As indicated in Table 4.6, 32,271 different pubmed IDs describe 211,824 PPIs. Figure 4.9 highlights that most of the PPIs are found in only one publication (~190,000 PPIs, corresponding to ~90%), with the remaining ~10% confirmed in two or more. This percentage is higher when considering PPIs that appear in multiple source databases. Around 20% of the PPIs appear in two databases or more, showing that databases pick up interactions from the same publications. Regarding the coverage based on publications, a maximum of 389 references for one single PPI can be detected. The interaction between *MDM2*, the MDM2 proto-oncogene (Entrez ID: 4139) and *TP53*, tumor protein p53 (Entrez ID: 7157), has been described in 389 publications. This might highlight a key role and ubiquitous presence of the interaction, but also points towards a very highly studied interaction, leading to the high, observed value.

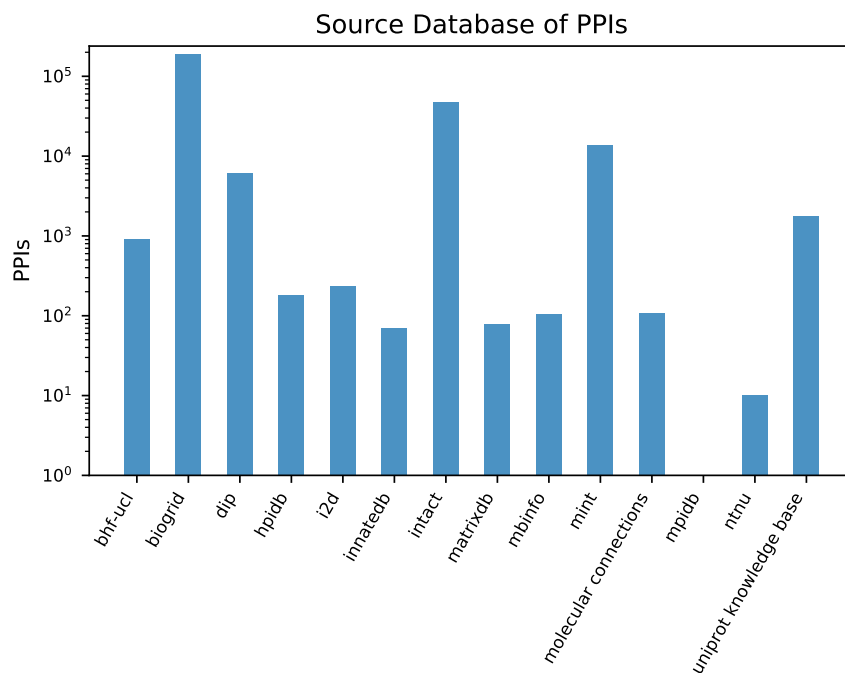


Figure 4.7: Human, unique, direct PPIs found in different source databases (based on information in the “source database” column of the mitab25 files). The x-axis shows the source database in alphabetical order.

In conclusion, the analysis presented leads to a high confidence dataset which can be used for future studies. Source and interaction type of PPIs offer a good understanding of data quality. Furthermore additional filtering can be applied depending on the user’s needs and requirements.

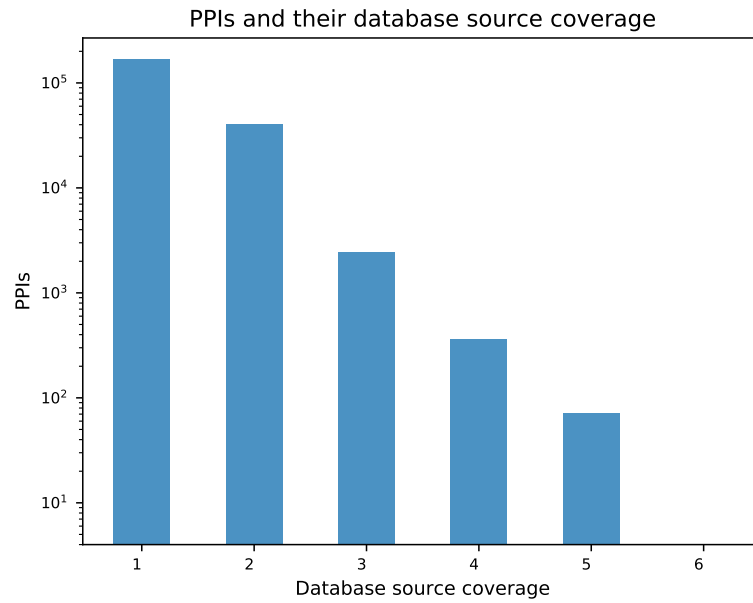


Figure 4.8: Human, unique, direct PPI coverage in different source databases (based on “source database” count).

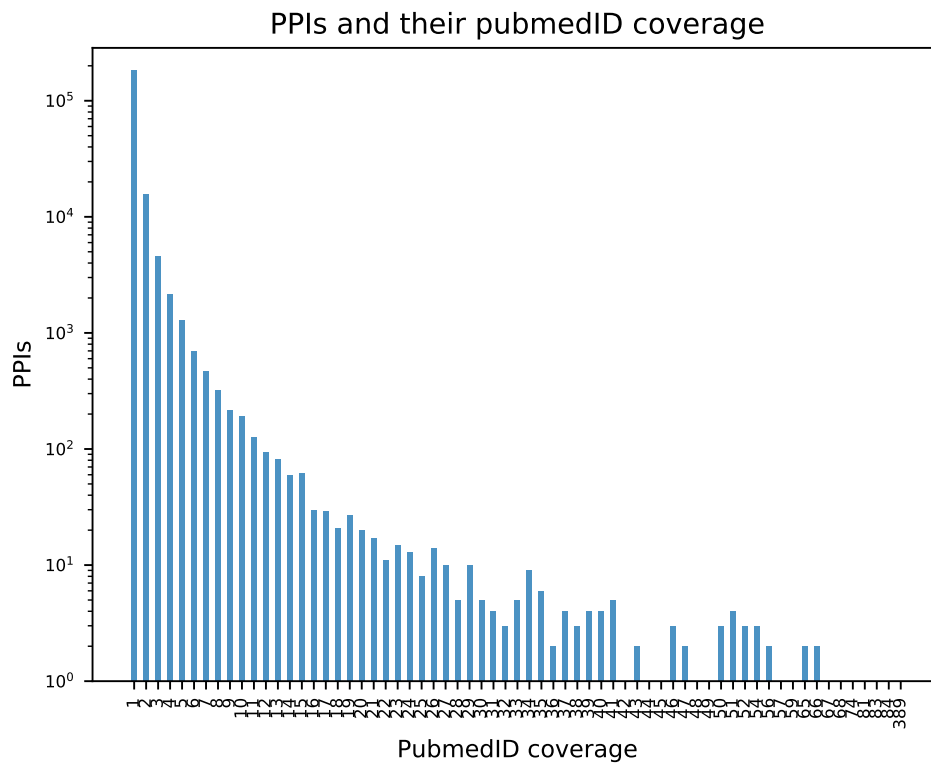


Figure 4.9: Human, unique, direct PPI coverage in different publications (based on “publication identifier” count).

4.4 Discussion

A growing number of databases supply PPI information. However their content varies widely and most often individual studies use data from only one resource.

Based on the use of the accepted PSI-MI XML 2.5 standard, the mitab25 format allows combination and comparison of data from different sources. Nevertheless databases use different primary identifiers for interactors. The human Entrez ID was chosen as the main identifier in this study and two mapping files (Entrez IDs to Uniprot IDs as well as Entrez ID to Uniprot Accession IDs) were used to make the mapping process as complete as possible. Checking all Entrez IDs for their presence in the mapping file guaranteed the exclusion of outdated or replaced records.

The use of Entrez IDs unique to genes is debatable, since it does not reflect post-translational modifications such as splicing and other events. These can lead to protein isoforms which are not reflected in Entrez IDs. Nevertheless they were considered the best option guaranteeing consistency and avoiding divergent results due to different protein isoforms. Detecting specific PPIs based on exact protein variants is almost impossible given current technologies. For consistency between data and annotation methods the use of gene IDs seemed to be more precise and avoided biases towards more easily detectable transcription variants of a protein.

Considering the PPI type and defining a set of direct interaction types reduced the number of total interactions, generating the final, best possible interaction dataset, given the defined interests. Depending on data use, this step seems crucial and avoids including too many false negative records, possibly leading to wrong conclusions.

The large differences between data deposited in different databases demands attention. Therefore combining information from distinct sources is important and allows complete PPI datasets to be assembled to the best of current knowledge. It also exposes how different data gathering, extraction and annotation strategies used by the different databases are, none of which can be considered as the only right or wrong approach. The information combination hence enriches the final PPI dataset, a more complete one, compared to ones considering only single sources.

These are just some aspects highlighting the improved quality of the presented dataset. Considering and including as much (high-quality) data as possible allows for optimal and most reliable results. Since the presented dataset is straight-forward to generate and shared with the scientific community it could help other researchers to easily benefit from the full range of available PPI data for their studies, without having

to personally join information from different sources.

A similar effort was made by the IMEx consortium (Orchard et al., 2012). It aims to supply a non-redundant PPI set spanning a range of organisms. In some cases data are still repetitive and a number of primary resources have not yet been integrated. Furthermore BioGRID data are not included and contain a large part of the analysed data, since it does not overlap with information available via DIP and IntAct. Additionally *psicquic*, which is compatible with IMEx supplies a data query interface which allows to retrieve interactions from IMEx associated databases. An R package supplying an interface to *psicquic* is available via bioconductor⁷ and linked to the HUPO Proteomics Standard Initiative (HUPO-PSI). A list of linked PPI databases can be found online⁸. During the exploration of options of this project, there were times when some of the advertised linked source databases were not reachable via *psicquic*. Furthermore computation times to obtain the desired datasets were very long, given the large size of the full human PPI set. Additionally it was preferred to have direct access to the meta-data instead of having the *psicquic* tool act as the intermediate filter. Working with raw data also allowed its in-depth analysis, such as cleaning and standardising meta-data. This highlights the increased quality and interoperability of the presented PPI dataset. Nevertheless *psicquic* is a good tool, especially when working with smaller datasets. It allows researchers to obtain PPI information deposited in different sources making their datasets more complete.

Overall the implemented approach allowed for higher flexibility in updating data, obtaining statistics and being able to access all meta-data at any point. It was possible to generate the best possible and most up-to-date PPI dataset for human, unique, direct interactions. The ease of rerunning the data extraction and combination pipeline allows constant use of most up-to-date data, directly supplying statistics describing most recent changes to the database content. This again illustrates the quality of the dataset as well as its maintain- and traceability including all available meta-data. Extending or modifying existing studies based on updated PPI data gets easier since the presented process is more transparent and PPI source relationships can be easily obtained, which is not the case using for example the *psicquic* tool. Furthermore the flexibility of filtering options and ease of use makes the pipeline and PPI set a valuable source for the wider research community.

As more studies become available it seems that not all PPIs have been discov-

⁷<https://bioconductor.org/packages/release/bioc/html/PSICQUIC.html>

⁸<http://www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml>

ered yet. Coverage is still low with new studies identifying additional PPIs instead of rediscovering already known ones. Hence the (human) PPI set still appears to be incomplete. The increase in PPIs between August 2016 and May 2017 confirms this phenomenon. Future studies will show if the total number of PPIs stagnates, indicating a saturated PPI dataset. More targeted PPI analysis could also show that the full set of interactions occurring in certain cell types or tissues, at specific developmental stages is already completely understood, but remains challenging to be identified.

When using PPI data, given a specific background, it has to constantly be considered that none of the databases supply information regarding the tissue, cell type, cell compartment, or developmental stage in which an interaction occurs or was recorded. In addition many interactions have been detected based on experiments carried out under artificial, experimental conditions, using varying setups. Even though human proteins were used, interactions might have been discovered *in-vitro* and not *in-vivo*, and partially in cells derived from different organisms (Rao et al., 2014). Other challenges are presented when considering the definition of a direct interaction and the way they can be identified in experiments. For example pull-down experiments do not only detect direct interactions, but include first and second order interactors if they are part of interaction complexes (Zhang et al., 2008). First order interactions are direct whereas second order interactions are indirect, with additional proteins between the two proteins identified. Only the IntAct database includes information of those so called “spoke expanded” interactions. Other databases do not supply such information but include PPIs derived from pull-down experiments. Excluding all records retrieved from pull-down experiments seemed too strict, since they represent a reliable, well studied and high quality source. Nevertheless specifically labelled records (from the IntAct database) were excluded to maintain data quality. New technologies may be able to produce and confirm this information in the future, allowing for more precise filtering and leading to even more concise PPI datasets.

Notwithstanding the uncertainty of confirming the presence of an interaction in a given tissue, or during a specific developmental stage, the use of proteomic data are of high value. Knowing the set of expressed genes in a tissue of interest allows a more precise PPI set to be generated. This does still not eliminate the problem of temporal expression patterns, but excludes proteins and their interactions that are not expected to occur in the tissue at all. With improving proteomic experimental setups this challenge might be counteracted and data could be retrieved capturing different developmental stages, enabling the comparison between the proteomes as well as their interaction

patterns e.g. at different developmental stages. Some first efforts have been made, including multi-scale modelling. However to gain confidence in the results, detection limits for proteomics still need to improve.

The growing availability of PPI data also led to an increase in analytical tools for large-scale study. The potential of PPI data, is to unravel patterns among connected proteins on a large scale, which is of high value and part of an expanding field.

Therefore the next two chapters focus on the identification of a proteomic dataset and the use of PPI data. Chapter 5 identifies and introduces the proteomic datasets of interest and Chapter 6 combines the proteomes with PPIs and uses network analysis approaches, including clustering techniques, to gain a deeper insight into the data. This can help to find answers to various scientific questions. This work focuses specifically on the effects of Parkinson's Disease (PD) in the synapse.

Chapter 5

The Synaptic Proteome and Parkinson's Disease

This chapter covers work that was part of a joint project. All analytical results presented in this chapter were obtained by myself. Screening and annotating of published synaptic proteomic studies was carried out by Colin Mclean and Oksana Sorokina, both from Douglas Armstrong's research group at the University of Edinburgh. This work is currently in preparation to be published with above mentioned co-authors and will be submitted shortly.

5.1 Objective

This chapter aims to identify the most up-to-date proteomic datasets describing the presynapse, postsynapse, synaptosome and the entire synapse. Once required mapping steps were carried out, individual datasets were joined to generate "regional" synaptic reference proteomes, specific to the presynapse, postsynapse, synaptosome as well as the synapse. Extracted data were compared to identify reference datasets containing all expressed proteins in the synapse.

Since the regional datasets emerged through data-fusion, based on different publications, protein detection coverage relative to the year of first detection is presented. This helped to determine data quality and allowed to carry out meta-analysis of the final proteomic datasets. In this way similarities and differences between protein-sets expressed in distinct synaptic regions were identified.

To understand synaptic region specific and overall functions, enrichment analysis was used. Through the use of gene-trait annotation information it aims to identify

over-represented specific traits in a subset of a given dataset. These traits can be molecular functions, biological processes and cellular components common amongst genes uniquely expressed in a region of choice. This approach helps to gain region specific insights as well as identify common synaptic specific functions.

Furthermore Parkinson's Disease (PD) associated genes were compared with the synaptic proteome. By doing so a set of synaptic PD associated genes was identified. Consequentially this also leads to a set of PD associated genes not expressed in the synapse. It was of interest to extract common properties of genes in the two PD gene lists, as this new knowledge can then help identify main disease-affected cellular functions, pathways and components in the synapse and other tissues.

Figure 5.1 shows an overview of described approaches, including data input, applied processes and outputs of this chapter.

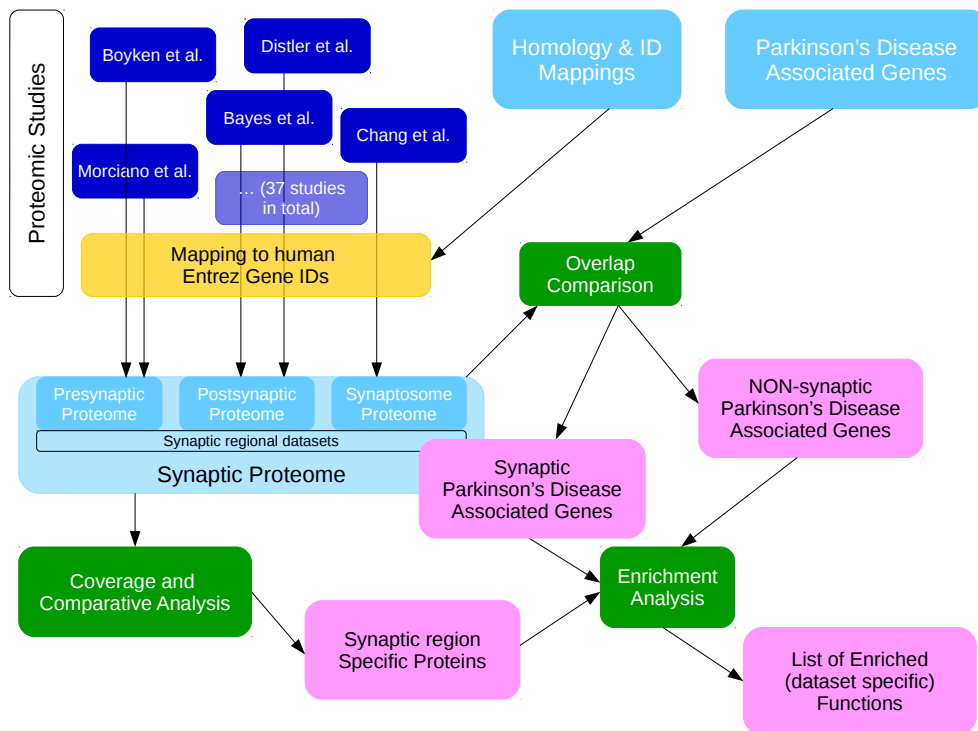


Figure 5.1: Overview of used data, analytical processes and outcomes of Chapter 5. Dark blue boxes refer to published data, light blue boxes are generated datasets, yellow boxes refer to analytical steps, green boxes describe processes and magenta boxes show outcomes.

5.2 Introduction and Material

A cell's phenotype and characteristics are defined by the expressed genes and consequently transcribed proteins. The set of expressed proteins in a cell is referred to as its proteome (Figure 1.1 B). In recent years a number of tissue and cell type specific proteomes have been identified, analysed and published. Thanks to the increase in large-scale experimental approaches the number of published studies rises constantly. This leads to larger, very likely more complete proteomic datasets resulting in higher statistical power to draw significant data-based conclusions.

Since the synapse is thought to be the main cellular region affected by PD, hosting disease-causing alterations it was addressed in this study. These alterations can provoke a number of dysfunctions, ultimately triggering the degeneration of dopaminergic neurons leading to the disease manifestation.

5.2.1 Proteomic Studies

The selection of proteomic studies was carried out in a joint effort and based on the research groups expertise in the field. Individual publications were studied and supplied protein identifiers were retrieved. Therefore data were extracted from a variety of file-types, such as .txt, .csv and .pdf. To obtain consistent identifiers these needed to be mapped between the different species mouse, rat and human. Mappings between protein and gene or gene and gene identifiers were carried out as specified in Section 2.2.1.

5.3 Results

5.3.1 Synaptic Proteome Datasets

Nine presynaptic, 22 postsynaptic and six synaptosome proteomic studies were identified. Two studies contain two datasets each: Distler et al. (2014) in the postsynaptic set and Cohen et al. (2013) in the synaptosome set. These were considered separately, leading to 23 postsynaptic and seven synaptosome datasets.

Data were extracted and all identified genes were mapped to human Entrez IDs. Homology information from published annotation files was used and verified through manual checks. If an entry could not be mapped it was discarded. For all the studies,

Table 5.1 contains further information such as the organism the data were obtained from as well as the number of identified genes (mapped to human Entrez IDs).

Table 5.1: Synaptic proteome publications and respective datasets used in this study. “# genes” refers to the number of proteins, mapped to human Entrez IDs identified in the study. Studies are sorted based on presynapse, postsynapse, synaptosome and ascending depending on the year of publication. Studies highlighted with * contain two datasets. More details can be found in the Appendices C.1, C.2, C.3.

Study	Year	Reference	Region	Species	# genes
MORCIANO	2005	Morciano et al. (2005)	presynapse	rat	85
BURRE	2006	Burré et al. (2006)	presynapse	rat	157
MORCIANO	2009	Morciano et al. (2009)	presynapse	rat	308
GORINI	2010	Gorini et al. (2010)	presynapse	mouse	49
GRONBORG	2010	Grønberg et al. (2010)	presynapse	rat	613
BOYKEN	2013	Boyken et al. (2013)	presynapse	rat	414
WILHELM	2014	Wilhelm et al. (2014)	presynapse	rat	1158
BRINKMALM	2014	Brinkmalm et al. (2014)	presynapse	mouse	68
WEINGARTEN	2014	Weingarten et al. (2014)	presynapse	mouse	467
WALIKONIS	2000	Walikonis et al. (2000)	postsynapse	rat	29
PENG	2004	Peng et al. (2004)	postsynapse	rat	325
SATOH	2002	Satoh et al. (2002)	postsynapse	mouse	45
YOSHIMURA	2004	Yoshimura et al. (2004)	postsynapse	rat	435
FARR	2004	Farr et al. (2004)	postsynapse	rat	71
JORDAN	2004	Jordan et al. (2004)	postsynapse	mouse and rat	390
LI	2004	wan Li et al. (2003)	postsynapse	rat	137
TRINIDAD	2005	Trinidad et al. (2005)	postsynapse	mouse	234
CHENG	2006	Cheng et al. (2006)	postsynapse	rat	288
COLLINS	2006	Collins et al. (2006)	postsynapse	mouse	717
DOSEMESI	2006	Dosemeci et al. (2006)	postsynapse	rat	113
DOSEMESI	2007	Dosemeci et al. (2007)	postsynapse	rat	548
TRINIDAD	2008	Trinidad et al. (2008)	postsynapse	mouse	2150
SELIMI	2009	Selimi et al. (2009)	postsynapse	mouse	61
FERNANDEZ	2009	Fernández et al. (2009)	postsynapse	mouse	292
BAYES	2010	Bayés et al. (2011)	postsynapse	human	1441
BAYES	2012	Bayés et al. (2012)	postsynapse	mouse	1545
SCHWENK	2012	Schwenk et al. (2012)	postsynapse	unknown	34
DISTLER PSD1*	2014	Distler et al. (2014)	postsynapse	mouse	3545
DISTLER PSD2*	2014	Distler et al. (2014)	postsynapse	mouse	2092
BAYES	2014	Bayés et al. (2014)	postsynapse	human	1141
UEZU	2016	Uezu et al. (2016)	postsynapse	mouse	1111
FOCKING	2016	Föcking et al. (2016)	postsynapse	human	2026
FILIOU	2010	Filiou et al. (2010)	synaptosome	mouse	2778
DAHLHAUS	2011	Dahlhaus et al. (2011)	synaptosome	mouse	638
ZIV synapse*	2013	Cohen et al. (2013)	synaptosome	rat	185
ZIV full*	2013	Cohen et al. (2013)	synaptosome	rat	2447
BIESEMANN	2014	Biesemann et al. (2014)	synaptosome	mouse	157
CHANG	2015	Chang et al. (2015)	synaptosome	human	2076
DISTLER	2014	Distler et al. (2014)	synaptosome	mouse	4475

Further details, including an extended description of the datasets, can be seen in

Appendices C.1 (presynapse), C.2 (postsynapse) and C.3 (synaptosome). The appendix tables also include details regarding experimental approaches used in the original studies. Distler et al. (2014) is the only study supplying data for more than one region. The publication contains data describing the synaptosome as well as the postsynapse.

Data was retrieved from the individual publications and can be found in the digital supplementary material (folder: “synaptic-proteome-data”). It contains one file for the presynaptic, postsynaptic and synaptosome proteome. These specify the different studies and respective proteins in the set.

Having a unique identifier (human Entrez IDs) for all genes is of great value and allows efficient comparison of datasets amongst each other. Based on the growing number of available studies, the increase in detected proteins was analysed. Figure 5.2 visualizes the increase in identified presynaptic (5.2a), postsynaptic (5.2c) and synaptosome (5.2b) proteins over the time since the first proteomic study was published. Those different datasets will be referred to as region specific synaptic datasets or proteomes. Figure 5.2d shows the dataset growth considering the union of all proteins in the three regional datasets. This dataset will be referred to as the joint synaptic proteome.

All four plots show an increase in the number of proteins. Presynaptic data started to be published in the early 2000’s and show three peaks in protein numbers, one when first published (2004), a second around 2010 and an additional increase in recent years, leading to a current total of 1,867 presynaptic proteins (Figure 5.2a). Studies addressing the synaptosome have only been published starting from around 2010. Based on available experimental approaches, those already contained more than 2,500 proteins. By 2016 the number of identified synaptosome proteins grew to 5,862 proteins (Figure 5.2b).

The postsynaptic proteome in comparison has been studied for almost 20 years. Initially smaller datasets were identified and a first significant increase in identified protein numbers can be seen in the early 2000’s. This is followed by another peak leading to around 2,500 identified proteins which stabilises nicely around 2010. More recent studies lead to another rise in protein numbers leading to a current total of 5,053 postsynaptic proteins (Figure 5.2c).

Considering the synaptosome as an individual unit might be subject to discussion, since, by definition, it contains all presynaptic and postsynaptic proteins. Hence the union of all three datasets was also considered. It shows constant data growth up to the

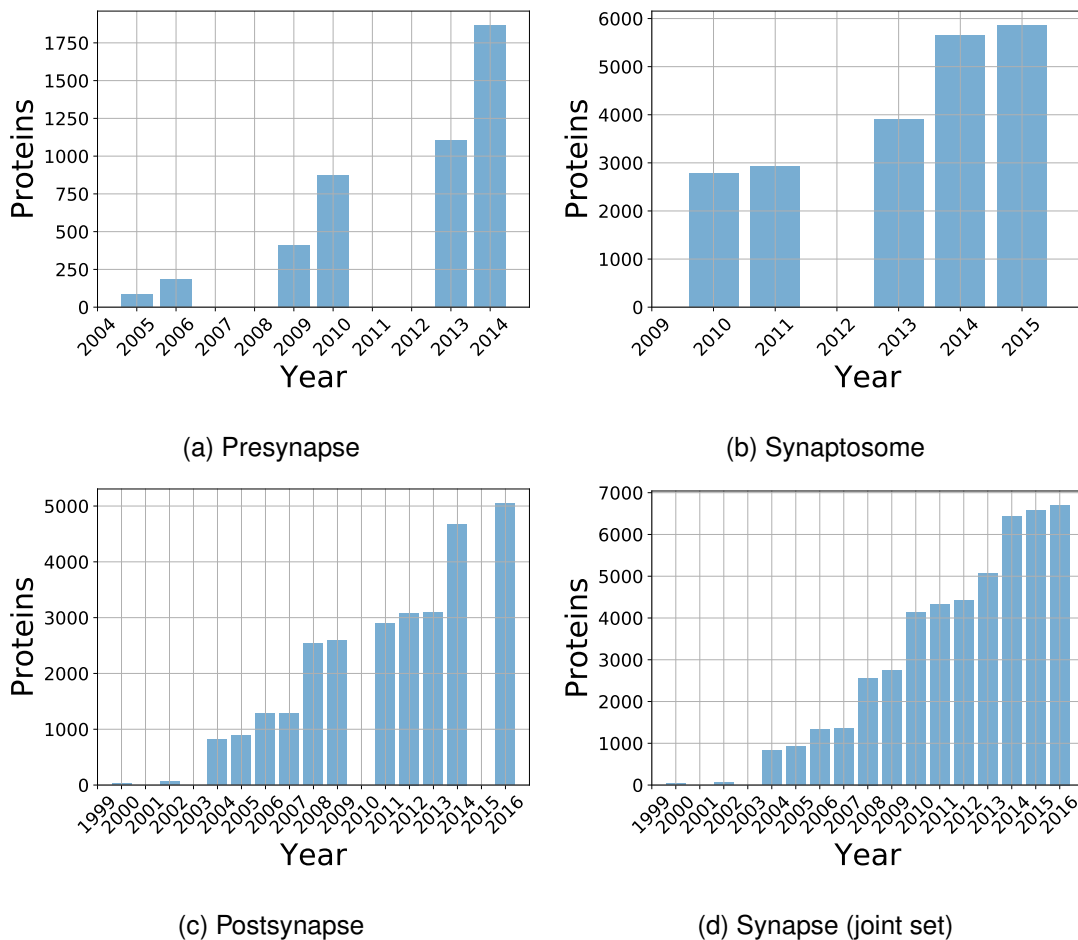


Figure 5.2: Increase in unique synaptic proteins, identified in different studies over the indicated years. Years without a bar reflect that no new data were added in those years.

current number of 6,706 proteins. The increase in identified proteins follows roughly the steps indicated for the postsynaptic proteome (increasing in the early 2000's, 2010's and in recent years (Figure 5.2d).

The time-frames of dataset growth are similar in all cases. The clear steps might reflect advances in experimental setups and analytical approaches which allowed detection of by then undiscovered proteins.

The postsynaptic and synaptic datasets show relatively clear plateaus in protein growth. This stagnation might indicate that the full datasets for those regions are almost identified. Regarding the presynapse and synaptosome no plateaus can be observed. Hence it is likely that an additional number of proteins, possibly already included in the current synaptic proteome will be associated with the presynapse based on future studies.

The next section addresses the presence of proteins found in different studies as

well as their functions.

5.3.2 Protein Coverage and Data Consistency

To test whether newer studies lead to a stabilization of total protein numbers in the regional sets their coverage amongst different studies was analysed. Due to improved experimental setups, it is suspected that newer studies re-detect proteins which had already been identified in previous analysis. Such findings would confirm synaptic proteome sets quality and consistency. Coverage is referred to as the number of studies in which a specific protein has been detected.

Figure 5.3 shows the number of genes detected with a certain coverage (blue bars). Since the number of studies varies among different datasets, total numbers can be misleading. Therefore Figure 5.3 also visualizes the coverage on a proportional scale (red bars). Both visualizations show that the majority of proteins have a coverage of 1 (for all four datasets: presynapse, postsynapse, synaptosome and the total synapse). This phenomenon is prominent in the presynaptic and synaptosome datasets, showing a steep dip of protein numbers associated with coverage 1 and 2.

When considering the percentage bars a slightly different impression can be obtained. The postsynaptic, as well as the joint synaptic dataset, contain a larger proportion of proteins identified in two studies or more: 68% and 75% respectively. This leaves 32% and 25% of the proteins identified in a single study. Considering the presynapse and synaptosome, slightly more than 60% and almost 45% of the proteins are found in one study only which is likely due to the lower number of total published studies addressing those regions.

Overall this approach does not consider the moment of first detection of a protein. Figures 5.4, 5.5, 5.6 and 5.7 reflect this fact, as well as protein detection coverage in studies published after first detection. Proteins are only associated with a year if they were newly detected. Each protein is represented once and multiple detections in other studies are visualized based on the coverage colour code. For example a postsynaptic protein detected in 2004 is one of almost 800 others first detected in 2004 (Figure 5.5). It was detected in at least one of five studies (see Figure 5.5, “studies” in x-axis label). Total studies indicates how many studies were published in the given year (here 2004) and thereafter (until 2016). Assuming the scenario that the protein of interest was detected in two studies published in 2004 and another five thereafter it is one of the proteins counted towards the “coverage seven” colour code in the barplot.

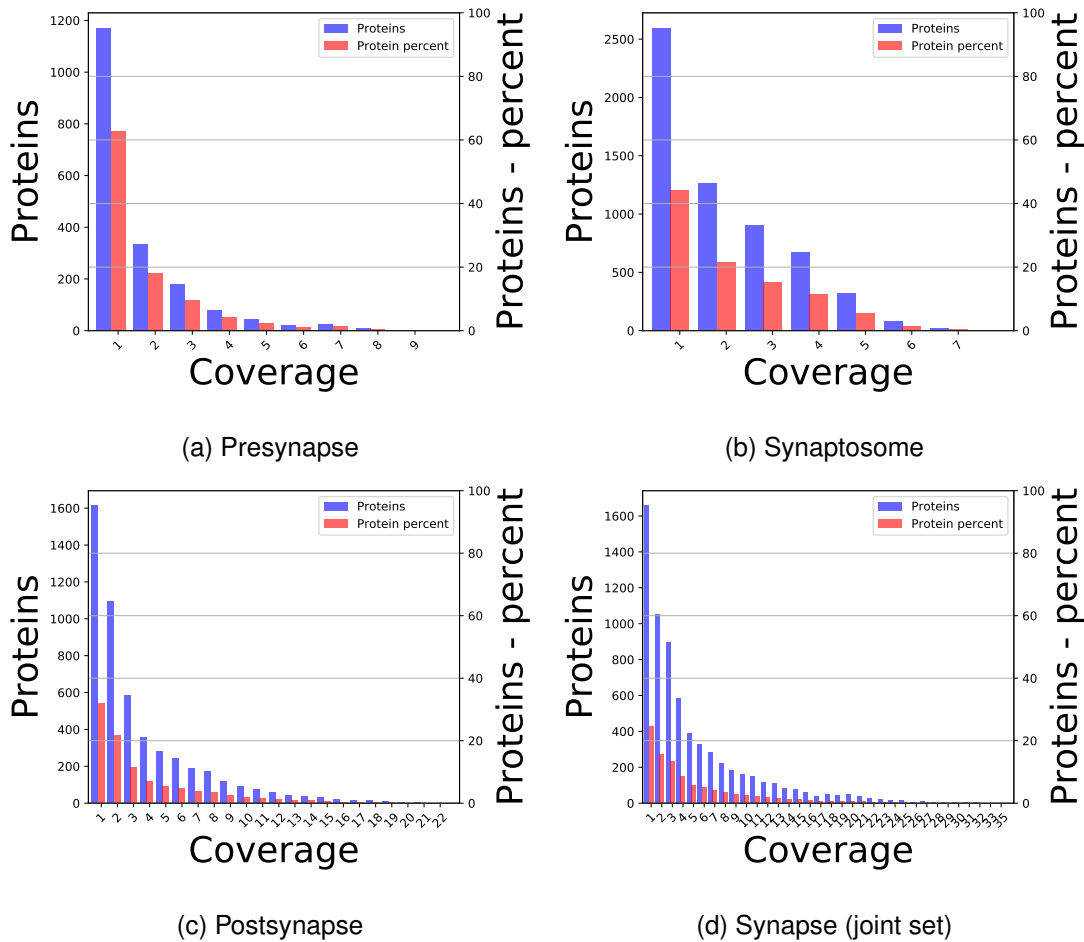


Figure 5.3: Number of proteins found in the regional proteomic studies and their coverage in a specific number of studies (blue bar). The red bar indicates the percentage of proteins identified with the respective coverage relative to the studied dataset.

This approach helps to highlight proteins detected for the first time at a later point, possibly due to advances in experimental technologies and not “penalising” their later detection through lower coverage, compared to “long standing single coverage” proteins that have not been re-detected in a large number of follow-up studies. An example could be the almost 100 presynaptic proteins first published in 2009 (see lowest, completely transparent part of the bar, Figure 5.4). These have not been re-detected in any of the following six studies published thereafter, whereas the remaining approximately 120 proteins were found in at least one more future study.

Hence Figures 5.4, 5.5, 5.6 and 5.7 confirm the peaks in dataset growth, indicating an increase in protein numbers in 2004, around 2010 and 2014. This supports the hypothesis that more advanced, fine-grained experimental techniques lead to the discovery of until then undetected proteins. Improvements might have occurred at sev-

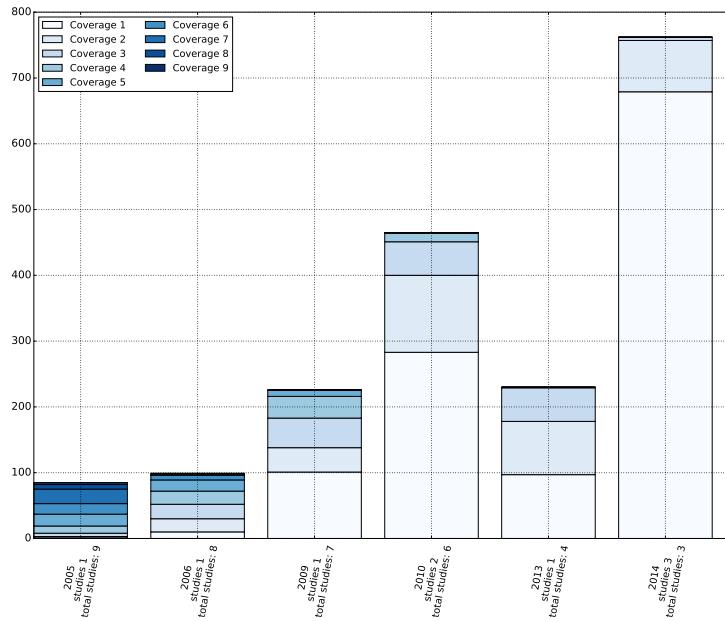


Figure 5.4: Coverage of presynaptic proteins, relative to year of first detection. X-axis label contains the year, number of studies published in that year (“studies”) and the number of studies published in the year and all coming years (“total studies”).

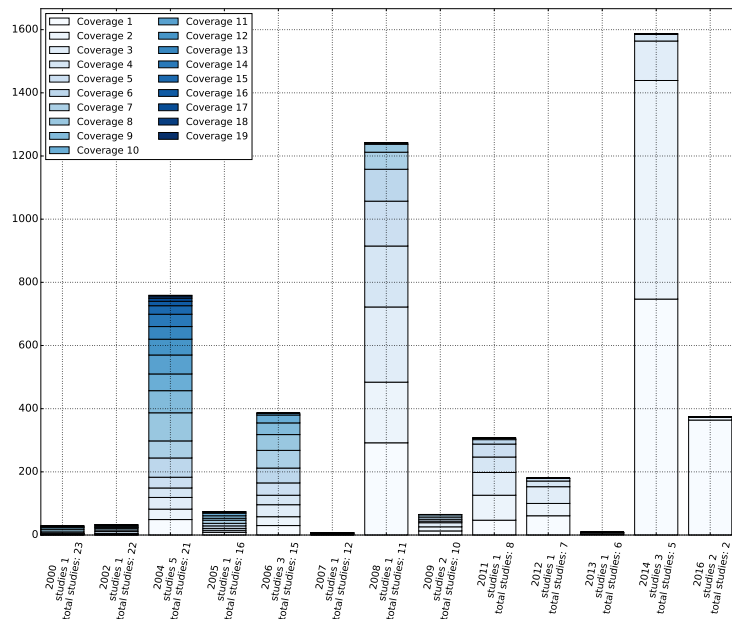


Figure 5.5: Coverage of postsynaptic proteins, relative to year of first detection. X-axis label contains the year, number of studies published in that year (“studies”) and the number of studies published in the year and all coming years (“total studies”).

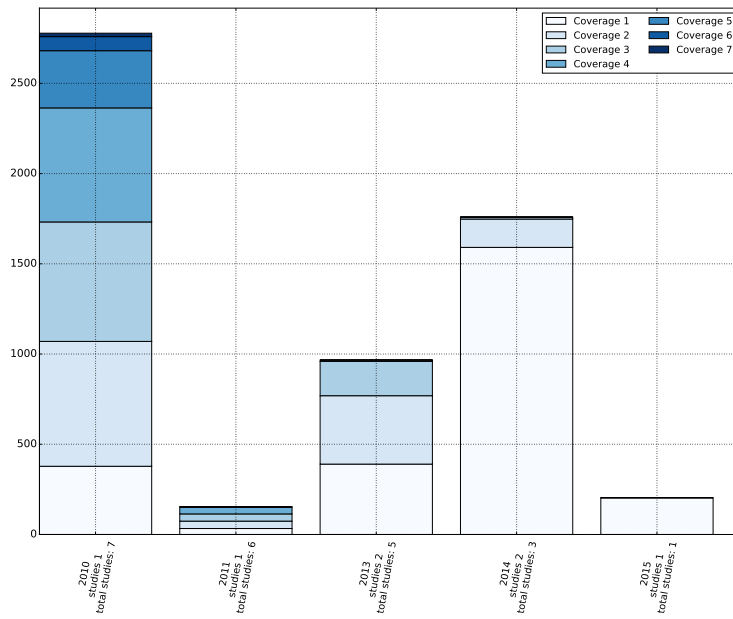


Figure 5.6: Coverage of synaptosome proteins, relative to year of first detection. X-axis label contains the year, number of studies published in that year (“studies”) and the number of studies published in the year and all coming years (“total studies”).

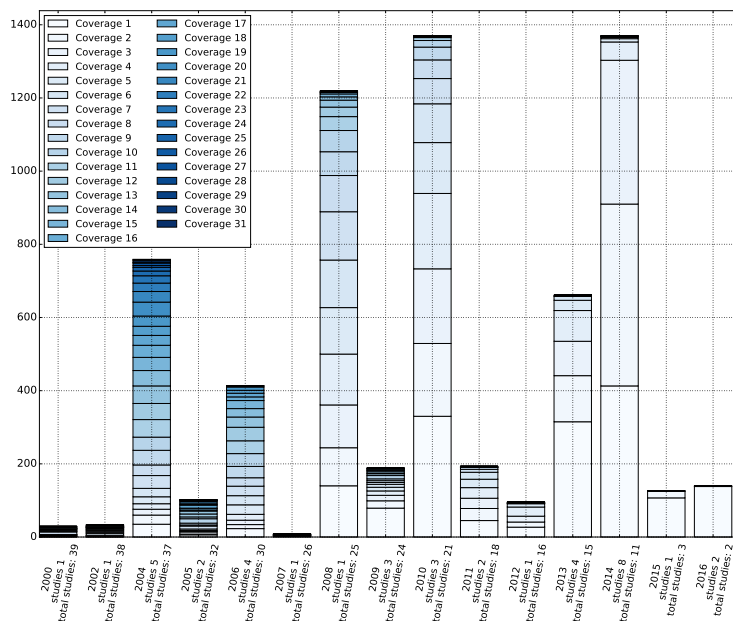


Figure 5.7: Coverage of joint synaptic proteome proteins, relative to year of first detection. X-axis label contains the year, number of studies published in that year (“studies”) and the number of studies published in the year and all coming years (“total studies”).

eral levels of the experimental pipeline, ranging from tissue extraction and preparation (meaning the extraction of proteins) to mass-spectrometry accuracy and data analysis. Hence 2008 seems to be a key year when first larger scale experiments were published. Figure 5.5 shows more than 1,200 newly identified proteins and Figure 5.4 around 700 in 2009 and 2010. Further evidence for advanced technologies is the first large dataset describing the synaptosome which contains more than 2,500 proteins (Figure 5.6).

Considering the plateau interpretation mentioned earlier and initially pointing towards having identified all proteins specific to any of the regions it can be seen that specifically studies in 2014 uncovered a number of new proteins for all the datasets. It is therefore questionable if the total numbers of proteins expressed in the different synaptic regions are already saturated or not.

The ongoing detection of additional synaptic proteins seems to reflect the continuing growth and refinement of the synaptic proteome. It may be that a number of presynaptic proteins remaining to be discovered, and/or rediscovered in future studies. Considering the synaptosome it seems reasonable to focus on the joint dataset including more studies. Even though higher coverage is seen in postsynaptic proteins, the joint synaptic proteome reflects the biological definition of the synaptosome and summarises all currently available knowledge, ensuring best possible data quality.

This analysis allows to draw first conclusions considering protein coverage, relative to the first detection of proteins. This is of considerable interest, since higher coverage increases the probability that a protein is really expressed in a region of interest. This adds certainty to a dataset, making it more credible. It also gives the opportunity to delete false positive records amongst the data. Setting a fixed filter should nevertheless consider when a protein was first discovered, allowing a certain number of equally potent studies (considering experimental approaches) to re-discover the protein of interest before excluding it from the set based on a fixed coverage threshold. Therefore having a “relative” coverage measure would allow to set flexible filters to reflect experimental advances and to obtain the best possible synaptic proteome datasets. Since there are no clear cut-offs yet, this chapter assumes the risk of having false positive records in the data, compared to discarding any of the real records.

Coverage information was subsequently used to gain insights into the functional roles of key proteins in the regional datasets.

5.3.3 Top Coverage Genes

Coverage does not only guarantee data quality, but can also give first hints towards central proteins in a dataset. In contrary to low coverage, high coverage can point towards highly expressed proteins, which are widely and easily detectable. These might have key roles in the respective tissues, even though this is not guaranteed simply due to high coverage. A list of these records can highlight similarities and differences between the regional datasets. Therefore proteins with maximum coverage in the different datasets were identified. The maximum coverage of proteins varies depending on the number of available studies.

Table 5.2, Table 5.3 and Table 5.4 show the top coverage records in the presynapse, postsynapse and synaptosome (second top coverage is included if only one top coverage record was found). Table 5.5 shows the top eleven proteins from the joint synaptic dataset. These overlap partly with top coverage genes in regional proteomes. The results do not consider the first detection year of proteins.

A close look at the top coverage genes in the joint synaptic proteome shows that one presynaptic gene appears in the joint synaptic top records. This number increases when considering postsynaptic and synaptosome top-coverage genes, with five and six genes present among the top-coverage genes in the joint synaptic dataset. In three cases synaptic top coverage genes are also amongst the top records in both, the postsynapse and synaptosome. Only two of the joint synaptic top records are not amongst the top coverage genes in any of the regional datasets.

The next paragraphs highlight some of the functions of the top coverage genes. Such insight presents a first biological interpretation of likely dominating functions in the distinct synaptic regions.

Three genes show top coverage in the presynaptic dataset (Table 5.2). The *VAMP2* protein confirms the importance of presynaptic vesicle to membrane fusion which is crucial for synaptic information transmission. *ATPIA3* and *GNAO1* might not intuitively associate to the presynapse, but are consistently detected. *ATPIA*, a sodium-potassium-pump (member of the P-type cation transport ATPases), plays a role in maintaining electrochemical concentration gradients. This is important for all neuron related regions and other sources have confirmed associations between *ATPIA* and the axon and synapse (Blom et al., 2016). Additionally links to PD have been previously postulated (Blanco-Arias et al., 2009). *GNAO1* is a member of the signal-transducing guanine nucleotide-binding (G) protein family (Murtagh et al., 1991) and was shown

to be implicated in ion channel regulation. This functionality might be more likely associated with the postsynapse but also contributes to maintain ion gradients across the whole synapse.

Table 5.2: Genes detected with top coverage in the presynaptic proteome (ordered by coverage and alphabetically by gene name).

Gene Name	Gene Name long	Entrez ID	Coverage
<i>ATP1A3</i>	ATPase Na+/K+ transporting subunit alpha 3	478	9
<i>GNAO1</i>	G protein subunit alpha o1	2775	9
<i>VAMP2</i>	vesicle associated membrane protein 2	6844	9

Overall the three presynaptic top-coverage genes are involved in key functions in the synapse. The mainly postsynaptic G-protein associated functionality as well as the sodium-potassium-pump properties associated with the presynapse could highlight the ubiquitous presence of some synaptic proteins amongst the different regions. This is specifically true for *GNAO1* which is also amongst the top-coverage genes in the joint synaptic dataset. Alternatively the detection of intuitively postsynaptic genes in the presynaptic proteome highlights the difficulty of extracting tissue specific to either the pre- or postsynapse. Nevertheless the identified results are of great interest, but it should be considered that presynaptic and postsynaptic expression specificity might diffuse between the synaptic regions.

In 22 out of the 23 postsynaptic datasets *DLG4*, a scaffolding protein was detected (Table 5.3). This highlights the complex structure of the postsynapse and the necessity to hold proteins in place and finely position them. Scaffolding proteins such as *DLG4* make this possible. Another six proteins have been identified in 21 datasets. Two of them are the well studied *CAMK2A* and *CAMK2B*, both members of the serine/threonine protein kinase family as well as its Ca²⁺/calmodulin-dependent protein kinase subfamily (Coultrap and Bayer, 2014). Calcium signalling is crucial for plasticity, specifically in glutamatergic synapses and significantly linked to memory and its formation including long-term potentiation (LTP) (Voglis and Tavernarakis, 2006). This can explain the central role of those proteins in the (post-)synapse. Additionally *SYNII* plays a role in synaptogenesis and neurotransmitter disease modulation (Cruceanu et al., 2012). This gene might be more naturally associated with the presynapse. Nevertheless it seems to play a central role in the postsynapse or likely amongst the whole synapse. This again underlines the hypothesis that tissue separation between the pre- and postsynapse remains very challenging.

Table 5.3: Genes detected with top coverage in the postsynaptic proteome (ordered by coverage and alphabetically by gene name).

Gene Name	Gene Name long	Entrez ID	Coverage
<i>DLG4</i>	discs large MAGUK scaffold protein 4	1742	22
<i>CAMK2A</i>	calcium/calmodulin dependent protein kinase II alpha	815	21
<i>CAMK2B</i>	calcium/calmodulin dependent protein kinase II beta	816	21
<i>GAPDH</i>	glyceraldehyde-3-phosphate dehydrogenase	2597	21
<i>INA</i>	internexin neuronal intermediate filament protein alpha	9118	21
<i>SPTBN1</i>	spectrin beta, non-erythrocytic 1	6711	21
<i>SYN2</i>	synapsin II	6854	21

Considering the synaptosome (Table 5.4), calcium related processes as well as vesicle and synapse specific proteins are detected with a high coverage.

19 genes show the top coverage of seven in the synaptosome. *CAMK2A* and *CAMK2B* are amongst the top coverage hits, as they are in the postsynapse. Similarly *DLG4* as well as *DLG2* are found, confirming the necessity of distinct scaffolding proteins. A number of the proteins seem to appear in pairs. This indicates that different variants of the proteins are identified with a high coverage. Such a finding might highlight that those form part of one complex, requiring both genes to be expressed for full functionality. *STXBP1* and *STXBP5* are one example. Both of them are involved in the synaptic vesicle cycle, specifically vesicle-membrane fusion and carry out their full functionality by interacting with other proteins such as *STX1*. *STXBP5* plays a (negative) regulatory role in exocytosis and neurotransmitter release (Joshi and Whiteheart, 2017) and *STXBP1* might determine intracellular fusion specificity (Archbold et al., 2014). It has been proposed that both proteins compete for *STX1* binding¹. *SV2B* and *SV2A* are members of the synaptic vesicle proteins 2 family (*SV2*) associated with the regulation of vesicle trafficking and exocytosis. Additionally *SV2A* interacts with *SYT1*, enhancing low frequency neurotransmission in quiescent neurons (Bartholome et al., 2017). Overall this shows the centrality of synaptic vesicle cycling which features a prominent role in the synaptosome as well as the presynapse.

With regard to protein pairs these are either functionally dependent from each other or reflect two protein variants. These may carry out similar functions in distinct brain regions or cell types or actively compete with each other. Overall top coverage genes detected in the synaptosome are involved in central synaptic processes. Similar findings can be made in the joint synaptic proteome and are addressed below.

¹<http://www.uniprot.org/uniprot/Q5T5C0>

Table 5.4: Genes detected with top coverage in the synaptosome proteome (ordered by coverage and alphabetically by gene name).

Gene Name	Gene Name long	Entrez ID	Coverage
<i>AP2M1</i>	adaptor related protein complex 2 mu 1 subunit	1173	7
<i>ATP6V1A</i>	ATPase H ⁺ transporting V1 subunit A	523	7
<i>CADPS</i>	calcium dependent secretion activator	8618	7
<i>CAMK2A</i>	calcium/calmodulin dependent protein kinase II alpha	815	7
<i>CAMK2B</i>	calcium/calmodulin dependent protein kinase II beta	816	7
<i>CTNNA2</i>	catenin alpha 2	1496	7
<i>DLG4</i>	discs large MAGUK scaffold protein 4	1742	7
<i>DLG2</i>	discs large MAGUK scaffold protein 2	1740	7
<i>NSF</i>	N-ethylmaleimide sensitive factor, vesicle fusing ATPase	4905	7
<i>PPFIA3</i>	PTPRF interacting protein alpha 3	8541	7
<i>SH3GL2</i>	SH3 domain containing GRB2 like, endophilin A1	6456	7
<i>SNAP25</i>	synaptosome associated protein 25	6616	7
<i>STXBP1</i>	syntaxin binding protein 1	6812	7
<i>STXBP5</i>	syntaxin binding protein 5	34957	7
<i>SV2A</i>	synaptic vesicle glycoprotein 2A	9900	7
<i>SV2B</i>	synaptic vesicle glycoprotein 2B	9899	7
<i>SYNGR3</i>	synaptogyrin 3	9143	7
<i>SYP</i>	synaptophysin	6855	7
<i>SYT1</i>	synaptotagmin 1	6857	7

The 11 top coverage proteins in the joint synaptic proteome are listed in Table 5.5. The maximum coverage is 35 (out of 38 datasets) and more proteins are found with a coverage of 32 and 31. The first six records in particular overlap with top coverage proteins in the postsynapse or synaptosome. These proteins cover calcium/calmodulin related functions as well as vesicle cycling and fusion related roles. *SYN1* and *SEPT5* are not amongst the top coverage genes in any of the regional datasets, but appears in the joint synapse.

SYN1 forms part of synaptic vesicles. It has been shown to be involved in neural development, synaptic neurotransmission as well as plasticity (Fassio et al., 2011). Its coverage in the regional datasets is relatively high, but is not amongst the top hits (8, 26 and 6 in the presynapse, postsynapse and synaptosome). *SYN1* together with *SYN2* additionally appear to be part of another protein pair encoding neuronal phosphoproteins, associated with the synaptic vesicle surface. Interactions between the two proteins have been identified and point towards their complex formation (Hosaka and Südhof, 1999). This finding can also be used to expand the identification of key genes,

Table 5.5: Genes with top coverage, detected in the joint synaptic proteome (ordered by coverage and alphabetically by gene name). “pre”, “post” and “synapt” refer to the presynapse, postsynapse and synaptosome proteomes respectively.

Gene Name	Gene Name long	Entrez ID	Coverage	top enriched in other dataset(s)
<i>CAMK2A</i>	calcium/calmodulin dependent protein kinase II alpha	815	35	post, synapt
<i>INA</i>	internexin neuronal intermediate filament protein alpha	9118	33	post
<i>NSF</i>	N-ethylmaleimide sensitive factor, vesicle fusing ATPase	4905	33	synapt
<i>SYN2</i>	synapsin II	6854	32	post
<i>SYT1</i>	synaptotagmin 1	6857	32	synapt
<i>DLG4</i>	discs large MAGUK scaffold protein 4	1742	32	post, synapt
<i>SYN1</i>	synapsin I	6853	32	-
<i>STXBP1</i>	syntaxin binding protein 1	6812	31	synapt
<i>GNAO1</i>	G protein subunit alpha o1	2775	31	pre
<i>CAMK2B</i>	calcium/calmodulin dependent protein kinase II beta	816	31	post, synapt
<i>SEPT5</i>	septin 5	5413	31	-

based on other proteins, likely forming part of the same reaction complex.

In general this illustrates how dataset comparison can confirm known principles and lead to new insights of key genes in large datasets. The presented results show that top coverage genes might not always be as region specific as expected, based on their role, but give a notion of overall important and dominating functions in the synapse and its specific regions. Therefore individual analysis and interpretation of protein functions within the synapse can help to pin down central functions of synaptic regions. This analysis considered all protein in the respective datasets. To understand region specific properties the regional datasets were compared.

5.3.4 Regional Synaptic Properties

After having analysed the individual datasets as well as the joint synaptic proteome, these were compared with each other. Figure 5.8 visualizes the overlap in terms of common genes. A total of 6,706 synaptic proteins were identified in at least one study (mapped to a human Entrez ID). This represents the size of the synaptic proteome as introduced earlier (Section 5.3.1).

The intersection of the three sets contains 1,478 proteins, being 22% of all identified proteins. These are present in all three datasets and likely carrying out general cellular functions.

Overall it can be seen that the synaptosome contains most of the proteins (5,862), including almost all genes expressed in either the presynapse or postsynapse. Only

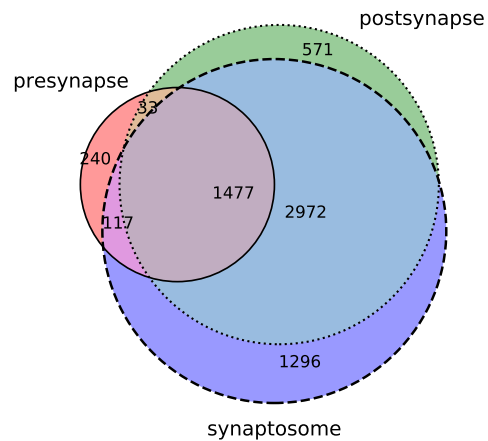


Figure 5.8: Overlap of unique, human Entrez IDs of the genes identified in the presynapse, postsynapse and synaptosome proteome. All genes are included (minimum coverage = 1).

844 of the total number of synaptic proteins were not found in the synaptosome (~12% of the joint synaptic set).

The postsynapse is the second largest set with 5,053 proteins in total, followed by the presynapse, containing 1,867 proteins. The number of proteins uniquely expressed in each of the three regions is relatively small. 1,296 (22%), 571 (11%) and 240 (12%) proteins are specifically expressed in the synaptosome, postsynapse and presynapse respectively (percentages are relative to the total number of proteins in the regional dataset).

Since protein coverage in the different datasets may reflect data quality (Section 5.3.2) it was of interest to consider changes in overlap when only considering proteins found in a minimum of two or three studies. Figures 5.9a and 5.9b show the respective venn diagrams and Table 5.6 shows a numeric overview. It can be seen that removing low coverage genes does not automatically remove all region specific records. It reduces the parts specific to the presynapse and synaptosome, whereas the postsynaptic proportion rises. This is due to the fact that far less presynapse and synaptosome specific studies were available compared to postsynaptic ones. This phenomena re-confirms that higher detection coverage can increase data quality and consistency, but previously explained first-detection times should to be considered.

Even though region specific unique gene sets are small, they very likely contain

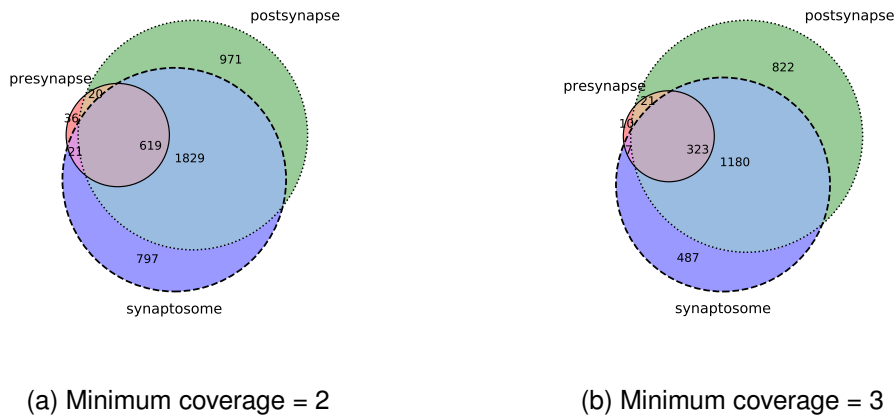


Figure 5.9: Overlap of unique human Entrez IDs of the genes identified in the presynapse, postsynapse and synaptosome proteomes (minimum coverage of considered proteins as indicated above).

Table 5.6: Number of genes in the synaptic regional proteomes, filtered for coverage.

Coverage	Presynapse	Postsynapse	Synaptosome	Joint synaptic set
1	1,867	5,053	5,862	6,706
2	696	3,439	3,266	5,452
3	361	2,346	1,997	2,850

genes associated primarily to functions, typical to the different synaptic regions. To identify those, functional enrichment analysis was carried out. Gene Ontology (GO) annotation terms were used to test gene sets uniquely expressed in the presynapse, postsynapse, synaptosome and amongst all three datasets. Enrichment was carried out for Biological Process, Molecular Function and Cellular Component associated terms (retrieved from the GO database). The Fisher exact test, the *topGO* elim algorithm and Benjamini and Yekutieli multiple testing correction were used. Results based on different background datasets were compared and using the full synaptic proteome as a reference background turned out to give most representative results which are presented in this chapter.

Few functional terms were enriched for region-unique datasets. Table 5.7 summarises the results and the following paragraphs interpret the findings.

Regarding uniquely presynaptic proteins “neurotransmitter biosynthetic process” (GO:0042136) stands out as the one enriched Biological Process. Considering Molecu-

Table 5.7: Significantly enriched functional GO terms of the gene sets specifically expressed in only one of the three regional synaptic proteome datasets. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, `elim` algorithm and Benjamini and Yekutieli multiple testing correction; significance p-value threshold was set to 0.05.

Gene Ontology Type	Presynapse (p-value)	Postsynapse (p-value)	Synaptosome (p-value)
Biological Process	neurotransmitter biosynthetic process (4.40×10^{-02})	negative regulation of interleukin-10 production (4.90×10^{-03})	mRNA splicing, via spliceosome (6.2×10^{-03})
Molecular Function	serine-type endopeptidase activity (3.34×10^{-02})	protein heterodimerization activity (3.30×10^{-02})	-
Cellular Component	transcription elongation factor complex (1.10×10^{-02})	integral component of peroxisomal membrane (5.51×10^{-03})	nucleoplasm (1.5×10^{-03})

lar Function terms, “serine-type endopeptidase activity” (GO:0004252) was identified, and “transcription elongation factor complex” (GO:0008023) is the enriched Cellular Component. All these terms overlap with well known presynaptic functions. Neurotransmitters are crucial for cell-cell communication and as the top-coverage genes indicate their transport in vesicles is also a highly central presynaptic function (Section 5.3.2). Regarding the Molecular Function “serine-type endopeptidase activity” is a specific form of a catalytic activity, assisting to initiate other interactions. This mostly happens through the modification of a protein, converting it into its active form. Identifying the “transcription elongation factor complex” indicates the high activity of the presynaptic region, requiring large amounts of newly generated protein to maintain its functionality. The elevated need of proteins such as neurotransmitters and transport related factors could explain the enrichment of proteins supporting protein production specifically in the presynapse.

With regard to the postsynapse specific genes (Table 5.7), a common Biological Process is “negative regulation of interleukin-10 production” (GO:0032693). Interleukin-10 (IP-10), also referred to as C-X-C motif chemokine 10 (*CXCL10*) and is a small cytokine belonging to the CXC chemokine family. Its expression is usually triggered by IFN-gamma as a response to pathogens. Since this does not seem to be a brain specific functionality, the expression pattern of IP-10 was reviewed. It could be seen that IP-10 is highly expressed in the fetal brain² which might explain its enriched appearance in the postsynapse. Another explanation is around the theory that pathogens tend to affect synaptic vesicles or cytokines to enhance the pathogens reproduction.

²<http://biogps.org/#goto=genereport&id=3627>

Therefore a range of (synaptic) vesicle proteins are associated to pathogenic and viral terms which appear amongst enrichment analysis results (Franco and Shuman, 2012). Regarding prominent Molecular Function, “protein heterodimerization activity” (GO:0046982) stands out. The dimerization of heterodimers is closely connected with G-protein coupled receptors (GPCRs) which communicate external postsynaptic signals into the cell to trigger downstream actions. The full functionality of a range of GPCRs is only given when two of them dimerize which makes the heterodimerization activity a crucial postsynaptic process. Furthermore the “integral component of peroxisomal membrane” (GO:0005779) is the enriched cellular component amongst proteins uniquely expressed in the postsynapse. Peroxisomes are cell organelles, involved in the fatty acid catabolism and hosting highly important enzymes, participating in the energy metabolism (Wanders and Waterham, 2006). Additionally they have been shown to synthesize ether phospholipids which are critical for normal mammalian brain function. This role could explain their over-representation in the postsynapse.

The last region specific gene set describes the synaptosome. As Table 5.7 shows, “mRNA splicing via spliceosome” (GO:0000398) stands out as the enriched Biological Process. Splicing is crucial to generate mature mRNA which is consequentially translated into a protein. Through alternative splicing it can also lead to different mature mRNA products. Since the synapse is a highly active region hosting many processes at the same time, it requires well functioning protein production activity. Additionally it might indicate that different synapses produce distinct protein splice variants, requiring high spliceosome activity. No Molecular Function was identified as significantly enriched and the “nucleoplasm” (GO:0005654) is the enriched Cellular Component for the synaptosome specific proteins. The nucleoplasm comprises all nuclear proteins other than the chromosomes. Again this could point towards highly elevated protein production in the synapse which requires transcription of DNA as well as their transport through the nuclear membrane. These are functions covered by proteins in the nucleoplasm.

Alternatively to the established hypotheses the discovered terms might be artefacts occurring due to contamination of analysed samples. Experimental spot-checks as well as additional studies might help to find additional proof for the association or help to discard it.

Apart from the dataset specific functions, common functions covered by all three datasets were investigated. The 1,477 genes expressed in all three synaptic regions were analysed. Figure 5.10 displays the enriched terms. Compared to the region unique

datasets, the number of enriched terms is higher. This might be due to the fact that this dataset is larger as well as an increase in processes specific to the synapse itself.

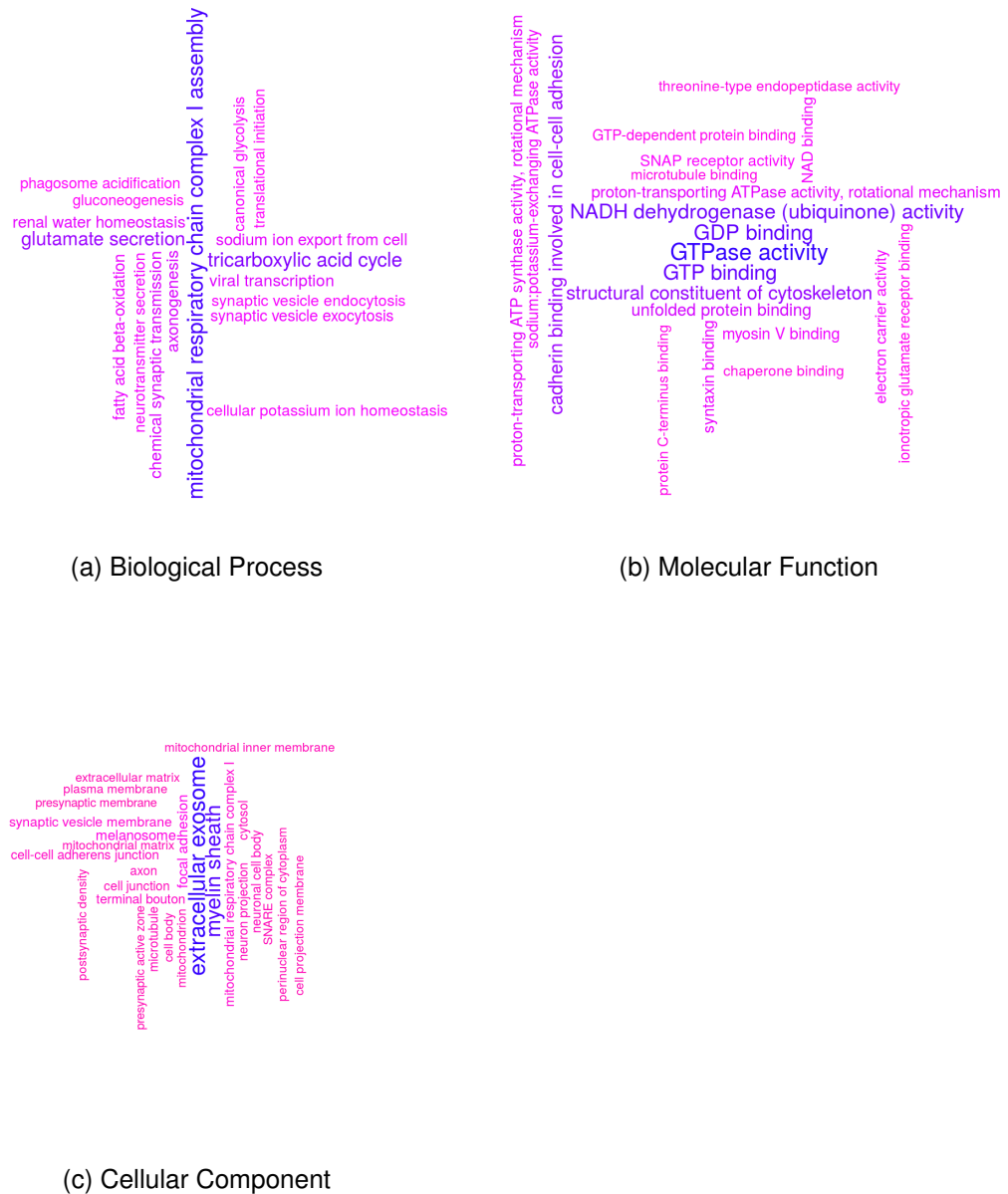


Figure 5.10: GO enrichment of the set of genes expressed in all three datasets (presynapse, postsynapse and synaptosome). Enrichment was tested compared to the whole synapse as a background. Results for different GO ontologies are shown. Fisher test, the `elim` algorithm and Benjamini and Yekutieli multiple testing correction were used. Colour gradient (violet to blue) and size (small to large) reflect significance of the terms.

Some of the significantly enriched Biological Processes are “mitochondrial respiratory chain complex I assembly” (GO:0032981), “glutamate secretion” (GO:0014047),

“neurotransmitter secretion” (GO:0007269) as well as “synaptic vesicle endocytosis” (GO:0048488) and “synaptic vesicle exocytosis” (GO:0016079) (Figure 5.10a). Mitochondria generate intracellular energy which is ubiquitously required in the synapse. The central role of energy production has been seen in several studies and links between a lack of energy to neurodegenerative diseases are getting more and more clear (Beal, 1998). Synaptic vesicles are the main transport media involved in information transmission and also amongst the top coverage genes in the synaptosome and joint synaptic proteome (Tables 5.4 and 5.5). Their prominent functionality amongst the joint synapse dataset shows their clear importance even though synaptic vesicles might intuitively be associated with the presynapse.

Enriched Molecular Function terms contain “GTPase activity” (GO:0003924) and “GTP” and “GDP binding” (GO:0005525, GO:0019003) as well as “NADH dehydrogenase (ubiquinone) activity” (GO:0008137) and “structural constituent of cytoskeleton” (GO:0005200) (Figure 5.10b). These are processes mainly involved in energy rich functions as well as protein generation and structural intracellular management. Hence, the terms provide further evidence of the high energy consumption of the synapse and indicate that energy related processes are very prominent amongst the most prevalent ones in the region. To manage parallel processes the cytoskeleton has a crucial role in assisting the transport of components in the synapse. Identifying enriched terms associated to the cytoskeleton and its structure confirms the importance of spatial intra-synaptic organisation for full functionality.

The “extracellular exosome” (GO:0070062) as well as the “myelin sheath” (GO:0043209) are two enriched cellular components, confirming that the genes found in all three regional datasets are describing the synapse, or more generally neurons themselves (Figure 5.10c).

This shows that gene set enrichment analysis of heterogeneous datasets is a great tool to obtain detailed and general functional descriptions of gene sets. Results should always be cautiously analysed but reveal general patterns.

After having analysed the data from the synapse level they were used for further studies in PD. Putting dataset completeness over the removal of possible false positives, the full dataset (coverage = 1) was used to locate PD associated genes in the synapse.

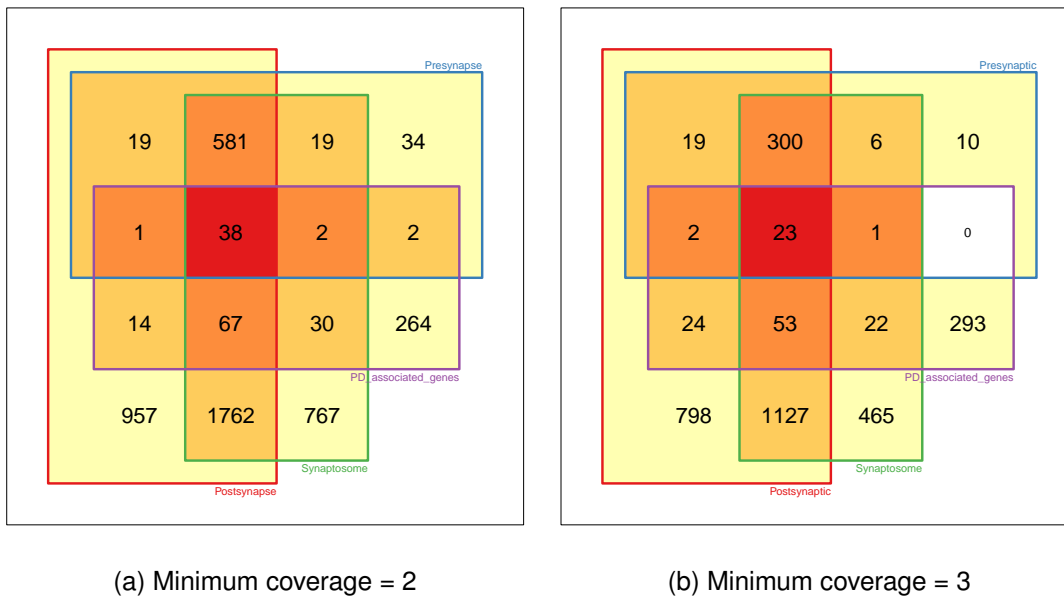


Figure 5.12: Overlap of the three regional synaptic proteomes with PD associated genes (adjusted coverage of synaptic proteins in all regional datasets).

associated genes was 418 when testing for enrichment given all human protein coding genes as the reference background and 205 when considering the synaptic proteome as a background datasets. The significance threshold was set to a p-value of: 0.05.

It can be seen that the synaptic proteome is significantly enriched for PD associated genes, meaning that the number of PD associated genes found amongst the synaptic proteome genes is higher than expected, if all the PD genes were distributed equally over the genome. The corrected p-value is 2.66×10^{-11} (Table 5.8 column “synapse”, row “unique hypergeometric enrichment (genome background)”). To test if each of the synaptic regions is specifically enriched, the overlap of PD associated genes with regional proteomes was analysed individually. PD genes uniquely found in region specific gene sets were analysed for enrichment in the respective datasets. The number of proteins specifically expressed in the region of interest were considered as the gene set of interest.

Genes unique to the postsynaptic and synaptosome proteome are significantly enriched for PD associated genes, compared to the “genome background”. P-values for the presynapse and compared to the synaptic background are very close to the 0.05 significance threshold (see Table 5.8).

The intersection of the three regional synaptic proteomes is enriched for PD associated genes relative to both genetic background datasets. This points towards an overall

general significant enrichment of the disease associated genes in the synapse, rather than specifically for any of the regional set.

Table 5.8: Overlap of PD associated genes with regional synaptic proteomes. “unique” refers to disease genes only overlapping with the indicated regional dataset and “total” refers to all the PD associated genes found in the respective proteome. Hypergeometric testing was carried out considering the full genome as a background (all human protein coding genes, referred to as “genome background”) as well as the synaptic proteome. Grey numbers indicate that the significance threshold of 0.05 was not reached.

	presynapse	postsynapse	synaptosome	all three	synapse
number of unique proteins	240	571	1,296	1,478	6,706
unique PD count	5	13	34	71	205
unique hypergeometric enrichment (genome background)	5.65x10 ⁻⁰¹	4.16x10 ⁻⁰¹	1.01x10 ⁻⁰¹	3.05x10 ⁻¹¹	2.66x10 ⁻¹¹
unique hypergeometric enrichment (synapse background)	8.64x10 ⁻⁰¹	9.00x10 ⁻⁰¹	8.65x10 ⁻⁰¹	1.85x10 ⁻⁰⁵	
total number of proteins	1,867	5,053	5,862	-	-
total PD count	81	162	186	-	-
total hypergeometric enrichment (genome background)	1.53x10 ⁻¹⁰	5.87x10 ⁻¹⁰	2.14x10 ⁻¹¹	-	-
total hypergeometric enrichment (synapse background)	1.61x10 ⁻⁰⁴	1.22x10 ⁻⁰¹	8.50x10 ⁻⁰²	-	-

To confirm this hypothesis a second calculation included all PD associated genes found in the regional subsets. The entire regional presynaptic, postsynaptic and synaptosome proteomes were tested for enrichment of PD associated genes. In this scenario, compared to the synaptic proteome as the background, all regional proteomes show PD enrichment with a p-value of 0.01 or lower. Enrichment compared to the genome background is significant with p-value < 1.5x10⁻¹⁰.

Combining all these insights the analysis confirms that the synapse is highly enriched for PD associated genes. The full synapse as well as the regional sets show significant disease enrichment. Nevertheless this analysis does not supply evidence for the disease affecting a specific synaptic region. Therefore it is very likely that PD associated genes found in the synapse affect a number of shared functions. Chapter 6 focuses on identifying potential commonalities in the function of disease associated proteins in the different regions of the synapse.

Apart from the synaptic PD associated genes, 213 of the original list were not found in any of the synaptic proteomes. Identifying common functionalities or regional expression patterns of these was of further interest and is addressed in the next section.

5.3.6 PD Affected Functions

After having identified a set of PD associated genes specific to the synapse and a second set of non-synaptic PD associated genes, functional roles of the proteins were investigated. The synaptic and non-synaptic PD associated gene sets were analysed separately regarding their functionality. GO enrichment for the two sets was carried out considering Biological Process, Molecular Function and Cellular Component terms. Different background lists were used for each set. For both lists the full set of human protein coding genes was one of them. The second background list was the synaptic proteome for the synaptic PD associated gene set and the “rest of the genome” (all human protein coding genes apart from the ones part of the synaptic proteome) for non-synaptic PD associated genes. Overall results are very similar, but seem slightly more targeted towards synaptic or non-synaptic functions when using the more specific background sets.

Results for the synaptic PD associated genes are based on the synaptic background dataset and are presented in Table 5.9 (column 1, “Synapse”). Enrichment was calculated using the Fisher exact test, `elim` algorithm and Benjamini and Yekutieli multiple testing correction. Top enriched Biological Process terms include “dopamine biosynthetic process” (GO:0042416) as well as “response to drug” (GO:0042493). Before applying multiple testing correction “clathrin coat disassembly” (GO:0072318) is also enriched for this gene set (p-value: 3.5×10^{-5} before and 0.39 after correction), this will be addressed again in Section 7.6.1. Finding enrichment of dopaminergic biosynthetic processes associated to genes in the set is likely related to the predominant effects of PD in dopaminergic neurons.

“Receptor binding” (GO:0005102) is the one enriched Molecular Function indicating PD effects on information transmission between neurons. Enriched Cellular Component terms are “neuronal cell body” (GO:0043025), “terminal bouton” (GO:0043195), and “axon” (GO:0030424). All these terms confirm known PD affected cellular functions and hint towards more general neuronal functions to be affected as well. Other enriched terms are “blood microparticle” (GO:0072562) and “platelet alpha granule lumen” (GO:0031093). These terms hint towards neuroinflammatory processes linked to PD (Hirsch et al., 2012) which allows and is reflected through enhanced access of immune response related particles into the brain.

Enriched functions of non-synaptic PD associated genes reveal different functional terms as the ones found enriched in PD associated genes expressed in the synapse. The

results consider enrichment compared to a background dataset including all human protein coding genes but the ones part of the synaptic proteome (Table 5.9, column “elsewhere”).

“Negative regulation of neuron apoptotic process” (GO:0043524) shows up as an enriched Biological Process term. This points towards a possible failure in the regulation of apoptotic processes as an aetiological factor. It could indicate a mechanism that fights against neuron loss, induced through apoptosis, specifically under disease conditions. Enriched Molecular Function terms are “peptidoglycan binding” (GO:0042834), “transcription factor binding” (GO:0008134), “protein heterodimerization activity” (GO:0046982) and “enzyme binding” (GO:0019899) amongst others. Finding those less brain specific terms amongst the enriched ones points towards non-synaptic PD affected regions and processes. Four mammalian peptidoglycan recognition proteins have been identified that actively recognise components, usually external to the human body, such as bacteria (Dziarski, 2004). Several studies uncovered their versatile activity against distinct bacterial strains (Bobrovsky et al., 2016) and showed a link to the chlamydial two-component stress response system. Identifying such a term amongst the ones enriched in non-synaptic PD associated genes hints towards elevated expression of related genes due to enhanced cellular defence mechanism activity. The other three terms are all related to “activating” processes. Gene transcription initiation could be enhanced to produce defensive or replacement proteins due to the dysfunction of others. Heterodimerization as well as enzyme binding can both be reactions to activate specific processes. The combination of these may indicate that non-synaptic PD associated genes are involved in generative processes influencing the cellular protein composition and contributing to the PD phenotype.

Two of the enriched Cellular Component terms amongst non-synaptic PD associated genes are “neuron projection” (GO:0043005) and “cell body” (GO:0044297). Based on the GO definitions neuron projection refers to the prolongation of a process extending from a nerve cell. This could be axons or dendrites. Cell body on the contrary describes the portion of a cell bearing surface projections from axons and dendrites, but excluding all cell projections. This combination of terms provides evidence for the role of PD associated genes linked to information and signal transmission and reception. This is a crucial synaptic function and its dysregulation can lead to neuron loss. “Lewy body” (GO:0097413) is another enriched term confirming the specificity of the dataset containing PD associated genes and highlighting their presence outside the synapse, but still in the brain. Even though Lewy Bodies are part of

other pathologies as well their combination with other enriched terms fits well into the PD pathology.

Table 5.9: Functional GO enrichment of PD associated genes expressed in the synapse and elsewhere. The gene sets of interest were enriched compared to all synaptic genes (“synapse”) as well as all human protein coding genes, apart from the ones expressed in the synapse (“elsewhere”). Results were obtained using the Fisher exact test, elim algorithm and Benjamini and Yekutieli multiple testing correction; significance p-value threshold was set to 0.05 (representation in alphabetical order).

Gene Ontology Type	Synapse	“Elsewhere”
Biological Process (p-value)	dopamine biosynthetic process (1.85x10 ⁻⁰²) response to drug (1.85x10 ⁻⁰²)	negative regulation of neuron apoptotic process (4.49x10 ⁻⁰³)
Molecular Function (p-value)	receptor binding (1.15x10 ⁻⁰²)	BH3 domain binding (1.80x10 ⁻⁰²) copper ion binding (8.68x10 ⁻⁰³) enzyme binding (8.68x10 ⁻⁰³) growth factor activity (1.69x10 ⁻⁰²) identical protein binding (8.68x10 ⁻⁰³) peptidoglycan binding (8.68x10 ⁻⁰³) protein homodimerization activity (1.69x10 ⁻⁰²) transcription factor binding (1.80x10 ⁻⁰²) ubiquitin protein ligase binding (8.68x10 ⁻⁰²)
Cellular Component (p-value)	axon (2.60x10 ⁻⁰²) blood microparticle (10 ⁻⁰² 2.64x10 ⁻⁰⁴) neuronal cell body (1.8010 ⁻⁰³) perinuclear region of cytoplasm (8.65x10 ⁻⁰³) platelet alpha granule lumen (6.31x10 ⁻⁰³) terminal bouton (5.10x10 ⁻⁰⁴)	cell body (6.86x10 ⁻⁰⁴) extracellular space (1.94x10 ⁻⁰²) integral component of plasma membrane (4.76x10 ⁻⁰²) Lewy Body (1.56x10 ⁻⁰²) membrane raft (1.94x10 ⁻⁰²) neuron projection (6,74x10 ⁻⁰⁶)

Overall this points towards non-synaptic effects of PD to be influencing neuron projection which could explain a part of the disease phenotype, affecting patients motor and movement difficulties as well as non-motor symptoms of the disease.

In summary and based on enrichment analysis results, on the one hand synapse specific PD associated genes have been proven to affect the dopaminergic system as well as receptor binding, specifically in terminal boutons. On the other hand non-synaptic PD associated genes are associated with apoptotic processes and affect neuron

projection. This could indicate that PD causal dysfunctions appear predominantly in the synapse. These would then project their effects, creating the PD pathology outside the synapse and brain to more distal body parts.

5.4 Discussion

The growing number of synaptic proteome studies allowed the generation of joint datasets describing the presynapse, postsynapse, synaptosome and the full synapse. Publicly accessible data are a great source for high data quality.

Nevertheless the data-joining process was not always straight forward and a number of challenges were faced. General issues encountered during data extraction and mapping were based on how information is presented by authors. In some cases supplementary information was in non-machine readable formats (e.g. .pdf-format), requiring manual annotation which is very time consuming and can be error prone.

Once all data were transformed to be machine readable, original identifiers needed to be mapped to one identifier of choice. In this work the human Entrez ID was chosen. Due to proteomics data obtained from non-human species protein IDs needed to be mapped between species and from protein to gene identifiers. Therefore mapping information was used, but at times manual fine-tuning steps were required. The encountered challenges highlight common problems of bioinformatics researchers working with information obtained in different species and from different sources. All these points might explain why many researchers stick to the use of individual sources avoiding data mapping and comparison. By doing so one full published synaptic proteome is used and necessary mapping steps are avoided. Nevertheless this approach carries a high risk of losing valuable additional information contained in distinct studies.

The presented regional proteome sets are hence the currently most complete synaptic proteomic datasets. The use of these “complete” proteomes is encouraged and should guarantee best possible data quality.

The human Entrez ID was chosen since the main application area of the datasets focus on a human perspective and the ID is considered a very stable source. In the context of this work, the role of PD associated genes was investigated. Furthermore a far larger amount of human Protein-Protein Interaction (PPI) data are available which will be combined with the synaptic proteome data described in Section 5.3.1. Therefore gene identifiers of the proteomes were mapped to Entrez IDs.

Striking size variation appears between the different published proteomic datasets.

Presynaptic studies for example identified between 49 (Gorini et al., 2010) and 1,158 proteins (Wilhelm et al., 2014). Regarding postsynaptic studies between 34 (Schwenk et al., 2012) and 3,545 proteins (Distler et al., 2014) were detected. The number of proteins identified in studies addressing the synaptosome ranges from 157 (Biesemann et al., 2014) to 4,475 (Distler et al., 2014). These differences are partly due to the analysed tissue portion. In some cases only a specific receptor complex, membrane channel or other structural parts were analysed, compared to e.g. the entire presynapse. Detection potential and sample size also increased in recent years. More advanced experimental techniques, material, and machines allowed large-scale screens leading to larger datasets.

The increase in data availability allowed to study multiple detection of synaptic genes in different studies. Considering the year of first detection of a protein allowed for a more detailed picture regarding the interpretation of protein detection coverage. This is specifically the case for more recently detected proteins which might have only be identified due to more advanced experimental techniques. Keeping this in mind can help to classify single coverage proteins differently, e.g. assigning lower credibility to a protein first detected in the early 2000's and never again, compared to a firstly discovered protein in 2015 thanks to more advanced experimental approaches.

Given all these insights the total number of synaptic proteins seems extremely large. It was initially intended to identify a key synaptic proteome. Nevertheless more data is required to estimate the size of the different synaptic proteomes and identify the exact set of genes part of these.

With regard to the protein abundance it needs to be remembered that numbers of proteins used in this Chapter refer to individual protein entities, not considering the copies of these, present in a cellular region. Additionally the data presented contains proteins expressed in the synapse at any given time. This does not mean that all of these are present in the synapse simultaneously, with variations depending on developmental stages amongst others. Hence numbers presented should not be considered as a total count of proteins in the synaptic regions, but rather present the diversity of proteins in the synapse. To estimate such a total count of proteins in the synapse, spatial constraints could possibly be considered. Nevertheless differences in protein size complicate such an endeavour.

The synaptosome by definition comprises the whole synapse including presynapse and postsynapse as well as other cell organelles such as synaptic vesicles and mitochondria. This explains why a portion of the synaptosome proteome does not overlap

with neither the pre- nor postsynaptic proteome. The (small) portions of presynaptic and postsynaptic proteins which do not appear in the synaptosome proteome might be due to low expression levels, hindering their detection in a larger dataset. Alternatively they might be detected in future studies. Nevertheless given the pre- and postsynapse specific genes those could be used to identify region specific functionalities.

Key biological functions of proteins in the presynapse and postsynapse vary largely, so does the current proteome size (1,867 proteins versus 5,053 proteins respectively). Nevertheless a large part of the presynaptic proteome (~80%) overlaps with the postsynaptic one, with only ~20% specific to the presynapse. Due to the larger size of the postsynapse, only ~30% of postsynaptic genes overlap with the presynapse and ~70% are specific to the postsynapse.

Based on the available data it is possible that a larger number of presynapse specific proteins still remain to be identified. If it turns out that the current information is correct, showing very low numbers of region specific proteins, this confirms that functional specificity of a cellular region can emerge and be explained by a relatively small amount of proteins.

To gain insights into similarities and differences in regional synaptic datasets the top coverage genes were considered. In this way genes with the highest detection coverage are shown to have well characterised synaptic (region specific) functions. Presynapse, postsynapse, and synaptosome genes with maximum coverage hint towards different functionalities. Caution needs to be taken since these results might be biased, based on specific detection methods targeting those proteins, as being highly, and specifically expressed in the synapse.

Enrichment studies supported the distinct functional focus of genes unique to the different regional proteomes. Even though region specific sets are relatively small, the analysis was able to confirm the main known roles of proteins specific to the different synaptic regions.

Considering the presence of PD associated genes in the synapse revealed that only ~50% of these are expressed in any of the synaptic regions. Nevertheless the overlap between disease associated genes and the synaptic proteome is significant (based on hypergeometric testing). More detailed analysis could not identify any of the region specific sets as overly enriched. Hence the synapse itself was proven to be highly and ubiquitously affected by PD.

The division of PD associated genes into a synaptic and non-synaptic group generated two datasets possibly representing different aspects of the disease. The synapse is

still believed to be a key cellular region where PD manifests itself and shows molecular alterations. To better understand the role of PD associated genes in the synapse functional enrichment was carried out. This confirmed known details linking PD to synaptic functions, such as receptor binding. Very likely all these could be considered disease triggering dysfunctions.

Functional enrichment of PD associated genes not found in the synapse revealed more generic and pathology related pathways, known to be affected in PD patients and representing consequences of synaptic PD associated dysfunctions. Finding enriched functional terms associated to signal releasing as well as signal receiving cellular components might have been suspected but has not yet been shown on a large scale. Even though this might not facilitate the search for drug targets it could point towards distinct affected cellular regions given different tissues or similar.

Overall, the analysis presented confirms known PD effects on the synapse and other cellular pathways. It is a first proof that using large-scale analytical approaches, such as functional enrichment analysis can help to shed light over complex research questions. Yet the aim is to obtain more specific results and uncover potentially still unknown disease links and causes. Therefore and to further investigate the influence of PD on the synapse more in-depth network analytical approaches are used. The following chapter presents Protein-Protein-Interaction Networks and clustering algorithms used to divide datasets into subgroups. Together with functional enrichment analysis this helps to gain better and deeper insight into specifically affected intracellular synaptic regions and pathways associated to PD.

Chapter 6

Synaptic Protein-Protein-Interaction Network Analysis and PD

6.1 Hypothesis and Objective

The complexity of Parkinson's Disease (PD) is reflected at various intracellular levels and affects a number of different functions. Therefore, it can be hypothesized that different sets of PD associated genes affect specific pathways. Most of the key, causal dysfunctions are suspected to be found in synapses. In this way it is suspected that several cellular functions are affected but via different molecular mechanisms.

To test this hypothesis it is intended to identify molecular pathways embedded within the synapse and enriched with PD associated gene. Figure 6.1 illustrates the overall workflow.

Curated synaptic proteomic datasets are used (Chapter 5) and combined with Protein-Protein Interaction (PPI) information (Chapter 4) to generate Protein-Protein-Interaction Networks (PPINs). An ongoing challenge in regard to PPINs is to identify patterns and substructures in analysed datasets. Based on different mathematical approaches clustering algorithms are able to identify highly connected network groups, referred to as communities.

In this chapter five different clustering algorithms were used to identify sets of synaptic genes showing an over-representation of PD associated genes. Functional enrichment analysis is applied to characterise the genes functions.

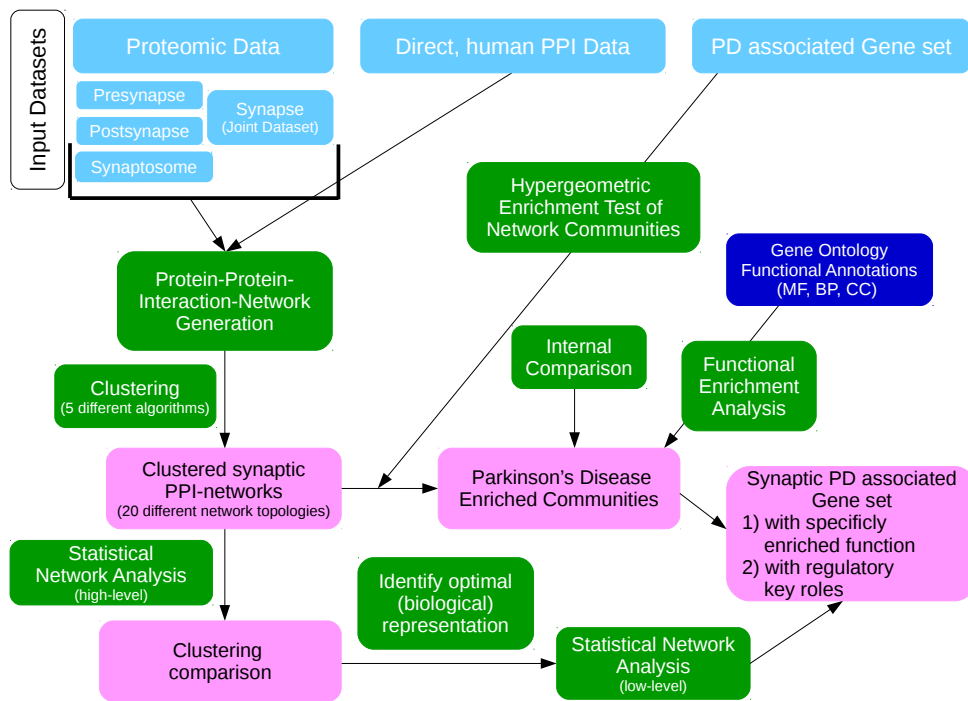


Figure 6.1: Overview of data, processes, and outcomes of Chapter 6. Dark blue boxes refer to published data, light blue boxes are generated datasets, green boxes describe processes and magenta boxes show outcomes.

6.2 Material and Methods

To gain understanding of a complex disease such as PD, large-scale analytical techniques, as well as high quality datasets, are highly helpful. Previous chapters set the baseline for the analysis presented in this chapter by obtaining clean and reliable datasets.

Section 1.4 introduces the principles of network analysis and specifically focuses on PPINs. The concepts of network generation, clustering algorithms and analytical methods are explained in Section 2.4.1.

Principles of (functional) enrichment studies can be found in Section 2.3. Considering functional enrichment with respect to networks, the main gene sets in the network are: i) all genes in the network, representing the background dataset and ii) genes in any of the network communities being the subgroup of genes to be tested for enrichment of a trait of interest.

All the analysis presented in this chapter can be run via a number of scripts. A master script allows the full analysis to be performed in one go. Information required are

the list of genes of interest as well as a PPI list. If the latter is not given, human, direct PPIs will be used. Should one want to test for gene-disease enrichment, gene-disease association data needs to be supplied. A number of parameters require command line input to adjust the analysis. All necessary scripts can be found here¹. A README file is supplied for more information.

6.3 Results

6.3.1 Synaptic Protein-Protein-Interaction Networks

Proteins mediate biological function, and in the majority of the cases they do so by interacting with each other. Hence using PPINs enhances understanding of interactions and emerging sub-structures amongst synaptic proteins. To characterise networks and identify similarities and differences statistical methods are applied.

Static, undirected PPINs of the regional and joint synaptic proteome datasets (Section 5.3.1) were built. Therefore human, internal, direct PPIs were used (Section 4.3.2). Table 6.1 shows some of the parameters obtained after initial analysis and describes the four networks.

Table 6.1: Overview statistics of the PPINs of the presynaptic, postsynaptic, synaptosome and joint synaptic proteome. Number of genes refers to the number of proteins in the proteome (mapped to human Entrez IDs). Nodes are proteins and edges PPIs. “bcc” stands for biggest connected component. “Clustering Coefficient” refers to the global measure. “Density”, “Diameter” and “Power-law Alpha” values are overall network measures. Details can be found in Section 2.4.

Dataset	Number of Genes	Total Nodes	Total Edges	Nodes (bcc)	Edges (bcc)	Max Degree	Clustering Coefficient	Density	Diameter	Power-law Alpha
presynapse	1,867	1,582	9,092	1,551	9,063	281	0.0892	0.0075	8	2.5714
postsynapse	5,053	4,583	47,152	4,562	47,132	690	0.0655	0.0045	8	2.5326
synaptosome	5,862	5,380	58,974	5,356	58,951	796	0.0643	0.0041	7	2.5135
joint synapse	6,706	6,094	69,545	6,068	69,520	893	0.0608	0.0037	7	2.5283

Differences between the number of genes and number of total nodes in the network might indicate that not all expressed proteins undergo interactions with other proteins in the dataset. Alternatively this can point out weaknesses of the PPI set, meaning

¹<https://github.com/KFHeil/thesis>

that it is incomplete. This point also touches upon the possibility that the PPI set contains false positive connections emerging through experimental sample contamination amongst others. Similarly the number of nodes and edges in the biggest connected component (Table 6.1, “bcc” columns) is smaller than the total numbers. This can be explained by small numbers of proteins interacting amongst each other, but not with the majority of other proteins in the dataset, the biggest connected component. These non-connected subgroups are of minimal size and were not further considered in this study.

The global clustering coefficient, also referred to as transitivity, ranges between 0.06 - 0.09. This measure describes the modular network topology, ranging between 0 - 1. High(er) values indicate “full connectedness” amongst network nodes, whereas lower values stand for sparsely connected networks. The observed values indicate that all synaptic networks are sparsely connected (Hwang et al., 2006).

To further analyse the connection pattern between the proteins, the degree of network nodes was analysed. The degree of a node is the number of edges linked to it and is often related to its centrality (Section 2.4). Maximum degree in the four presented networks ranges between 281 up to 893 interactions for single nodes. Top ranking records are listed in Table 6.2. Apart from the maximum degree, the degree distribution was analysed and fit to a power law distribution. The alpha value (Table 6.1) describes the fit of the data and ranges around 2.5. This indicates a heavy-tailed degree distribution of nodes in the networks, meaning that they are scale free (Section 2.4). From a biological point of view, this means that some very highly connected hubs, high degree nodes, appear alongside an exponentially increasing number of nodes with very low node degree.

Other measures listed in the table include network density which is very low in all four cases. It defines the percentage of edges appearing in the network, compared to all possible edges, not consider the PPI data, but assuming, that an interaction can occur between any two nodes in the network. An additional measure is the diameter, the longest geodesic in the graph. It describes the longest shortest path between two random nodes in the network. In the networks presented it is either seven or eight.

In summary, sparse network connectivity and scale-free degree distribution indicate that all four networks represent biological interaction patterns and reflect a known structure for large biological datasets (Ravasz et al., 2002; Barabási and Albert, 1999).

To gain a more detailed insight into key proteins in the network, node specific values were analysed. Together with the previously introduced node degree, betweenness

scores were calculated. Betweenness is another approach to gain detailed information about the role of nodes in the network and their relationship amongst each other. Table 6.2 shows the top 10 nodes with highest node degree and betweenness score.

Table 6.2: Top 10 nodes with maximum degree and highest betweenness score in the four different networks. “deg” refers to degree and “btw” to betweenness. Numbers in parenthesis refer to the rank. Grey scaled numbers are outside the top 10; “-” indicates missing genes in the respective datasets. A PD link is indicated in the last column.

Entrez ID	gene name acronym	gene name	pre deg	pre btw	post deg	post btw	synaptosome deg	synaptosome btw	synapse deg	synapse btw	PD associated
351	<i>APP</i>	amyloid beta precursor protein	281 (1)	241364 (1)	690 (1)	1066084 (1)	796 (1)	1457415 (1)	893 (1)	1853558 (1)	YES
7316	<i>UBC</i>	ubiquitin C	188 (2)	91665 (2)	446 (9)	347519 (7)	498 (8)	429978 (7)	543 (8)	548387 (7)	NO
8452	<i>CUL3</i>	cullin 3	-	-	688 (2)	554111 (3)	796 (2)	747686 (3)	840 (2)	841525 (3)	NO
2885	<i>GRB2</i>	growth factor receptor bound protein 2	165 (4)	82041 (3)	370 (13)	327951 (8)	396 (14)	373534 (9)	432 (16)	453681 (9)	NO
1994	<i>ELAVL1</i>	ELAV like RNA binding protein 1	-	-	548 (3)	869141 (2)	747 (3)	1284024 (2)	821 (3)	1563969 (2)	NO
1956	<i>EGFR</i>	epidermal growth factor receptor	-	-	548 (4)	530706 (4)	607 (4)	669252 (4)	645 (4)	773317 (4)	NO
2335	<i>FN1</i>	fibronectin 1	-	-	540 (5)	318622 (9)	580 (5)	368808 (10)	615 (7)	443895 (10)	NO
7514	<i>XPO1</i>	exportin 1	-	-	501 (6)	446903 (6)	557 (6)	575025 (5)	622 (5)	729977 (5)	NO
9820	<i>CUL7</i>	cullin 7	-	-	455 (8)	193952 (12)	509 (7)	249195 (11)	535 (10)	261247 (16)	NO
10482	<i>NXF1</i>	nuclear RNA export factor 1	-	-	480(7)	454101 (5)	-	-	616 (6)	619534 (6)	NO
4343	<i>MOV10</i>	Mov10 RISC complex RNA helicase	-	-	-	-	495 (9)	545787(6)	538 (9)	492472 (8)	NO
10987	<i>COP55</i>	COP9 signalosome subunit 5	-	-	424 (10)	168899 (16)	469 (10)	202796 (17)	508 (12)	254144 (17)	NO
26270	<i>FBXO6</i>	F-box protein 6	-	-	403 (11)	249059 (10)	462 (11)	377352 (8)	479 (13)	425790 (11)	NO
51547	<i>SIRT7</i>	sirtuin 7	169 (3)	69948 (4)	-	-	-	-	518 (11)	322055 (12)	NO
7534	<i>YWHAZ</i>	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta	152 (5)	53014 (5)	360 (16)	204559 (11)	382 (16)	238447 (13)	413 (17)	288584 (14)	YES
55832	<i>CAND1</i>	cullin associated and neddylation dissociated 1	150 (6)	35132 (8)	374 (12)	94333 (31)	426 (12)	123213 (29)	452 (14)	142042 (32)	NO
8266	<i>UBL4A</i>	ubiquitin like 4A	142 (7)	46881 (7)	223 (30)	170233 (14)	229 (36)	66085 (63)	244 (40)	84657 (54)	NO
7415	<i>VCP</i>	valosin containing protein	138 (8)	48896 (6)	267 (20)	68334 (39)	337 (21)	217333 (14)	354 (22)	247929 (18)	NO
3320	<i>HSP90AA1</i>	heat shock protein 90 alpha family class A member 1	129 (9)	34686 (9)	330 (17)	190131 (13)	360 (18)	211259 (15)	403 (18)	280273 (15)	NO
988	<i>CDC5L</i>	cell division cycle 5 like	120 (10)	34566 (10)	-	-	361 (17)	203703 (16)	385 (20)	2.03939 (22)	NO

The top 10 degree and top 10 betweenness proteins were identified. Most of the high degree nodes are also the top betweenness nodes. This might indicate that the central nodes are not just highly connected (hubs), but also transmit crucial information between different synaptic processes (high betweenness).

Postsynapse, synaptosome and joint synapse top ranking nodes overlap largely, whereas the presynaptic top nodes vary. Most of the nodes show a high degree in the other networks, but are not amongst the the top 10 since other nodes in these networks take over the top 10 positions. These differences indicate region specific functionalities.

Two of the top degree nodes are present in all four datasets. These are the amy-

loid beta precursor protein (*APP*) as well as ubiquitin C (*UBC*). *APP* is the gene with highest node degree and highest betweenness score, also showing a link to PD based on a Gene Reference into Function (GeneRIF) annotation (Compta et al., 2011; Aasly et al., 2012; Irwin et al., 2013). *APP* encodes a cell surface receptor and transmembrane precursor protein. Its primary function seems to be unknown, but it has been associated with iron export (specifically in Alzheimer's Disease) (Duce et al., 2010), synapse formation regulation (Priller et al., 2006) and neural plasticity (Turner et al., 2003). Nevertheless *APP* is mostly known as a the precursor protein of beta amyloid. As such, it is cleaved and can form the basis of the amyloid plaques found in the brains of Alzheimer's Disease patients.

UBC encodes the polyubiquitin precursor protein. Conjugated ubiquitin monomers or polymers can have various roles within a cell. Depending on the composition, ubiquitination processes are linked to protein degeneration, DNA repair, cell cycle regulation, kinase modification, endocytosis and the regulation of other cell signalling pathways (Kleiger and Mayor, 2014). Even though it does not activate a heat-shock response, its expression is enhanced during stress, providing extra ubiquitin to assist the ubiquitin system and remove damaged or unfolded proteins (Ryu et al., 2007; Tsigiotis et al., 2001). This versatile and protective functionality might explain its central role in all synaptic regions.

A second protein in the list has been associated with PD. The *YWHAZ* gene, can be found amongst the top 20 in all four regional datasets (top 10 in the presynapse). It is linked to PD based on a GeneRIF annotation (Ostrerova et al., 1999). The tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta (*YWHAZ*) gene encodes a protein which belongs to the 14-3-3 protein family. Also being referred to as 14-3-3 σ (protein name), it binds to phosphorylated serine/threonine motifs of target proteins and influences these in various ways. It is involved in signal transduction, apoptosis, cell cycle, cell growth and others (Rüenauver et al., 2014; Aghazadeh and Papadopoulos, 2016). Apart from a link to insulin level regulation and a predicted link to cancer, a link with Alzheimer's Disease was suggested previously (Qureshi et al., 2013).

The growth factor receptor bound protein 2 (*GRB2*) is the only gene showing a top 10 betweenness score in all four datasets with the node degree between rank 10 and 20. This might highlight its important role in cross-pathway communication and information flow regulation. The protein encoded by *GRB2* was originally detected as a binding partner of the growth factor receptor which then forms complexes with

proline-rich protein regions (Oda et al., 2005). Proteins containing such regions vary and are involved in a number of pathways, mostly facilitating signal transduction and cell communication (Lowenstein et al., 1992). More recent studies showed that it also forms complexes with protein tyrosine kinases, receptor tyrosine kinases, phosphatases, adaptors and intracellular scaffolds and can act as a key control point in the *MAPK* signalling (Ahmed et al., 2015). In this context *MAPK* influences information transmission from receptors to cell nucleus (McCain, 2013) supporting the hypothesis proposing its role in information flow and transmission regulation, based on the high betweenness score.

Similarly the F-box protein 6 (*FBXO6*) appears amongst the top 10 betweenness score genes in the postsynapse and synaptosome, and ranks 13 in the joint synaptosome. Nevertheless it is not present in the presynapse. This points towards a central role implicated in information transmission mainly in the postsynapse. *FBXO6* encodes a member of the F-box protein family which constitute the ubiquitin protein ligase complex (*SCF*). In addition to its role in the ubiquitin system it also seems to play a role in endoplasmatic reticulum stress-responses (Chen et al., 2016).

Three genes among the top 10 are specifically expressed in only one of the regional datasets. The sirtuin 7 gene (*SIRT7*), specifically expressed in the presynapse, is a homolog to *Sir7* in yeast. Its functions in human is still undetermined, but the yeast counterpart is involved in epigenetic gene regulation. More recently it was suggested to interact with the human RNA Polymerase I and II to carry out regulatory functions in chromatin remodelling (Tsai et al., 2012).

The nuclear RNA export factor 1 (*NXF1*) is the only postsynaptic specific top 10 node degree gene. *NXF1* is known to form complexes with *NXT1* and functions as a carrier between the nucleus and cytoplasm. The complex predominantly binds symmetric RNA substrates such as the *CTE-RNA* motif which are part of retroviruses amongst other (Aibara et al., 2015).

MOV10 is the only synaptosome specific gene amongst the top 10 degree nodes. The Mov10 RISC complex RNA helicase is part of the the RNA-induced silencing complex (*RISC*) and enhances its gene silencing function (Robb and Rana, 2007). As such it has been associated with the inhibition of retrotransposition (Goodier et al., 2012).

Another five presynaptic (*CAND1*, *UBL4A*, *VCP*, *HSP90AA1*, *CDC5L*) and seven postsynaptic, synaptosome and joint synaptic genes (*CUL3*, *ELAVL1*, *EGFR*, *FNI*, *XPO1*, *CUL7*, *COPS5*) are found amongst the top 10 records.

In summary all the presented proteins play central roles in the synapse. Hence their alteration very likely lead to crucial dysfunctions possibly triggering disease or cell death. The two identified PD associated genes are not amongst the most well known PD triggers, meaning that if solely affected they might not cause the disease outbreak. Based on this principle, specifically complex diseases, do not tend to affect the most central proteins in a cellular region, but a number of highly connected proteins. Their combined dysfunction can then lead to different diseases.

To gain a better insight into the most central PD associated genes, Table 6.3 shows the 10 PD associated genes with the highest node degree and their rank amongst all nodes in the network. For these, the respective betweenness scores including their rank are also provided. Overall the top 10 PD associated genes are amongst the top 109 node degree and top 150 betweenness score nodes in the networks.

Table 6.3: Synaptic PD associated genes with a top 10 degree value and their betweenness scores (together with the overall rank in the respective network). The table is sorted by coverage in the different datasets and based on the first available node degree based on the table columns. “degree” refers to node degree and “btw” to betweenness. Numbers in parenthesis refer to the rank. Grey numbers are outside the top 10; “-” indicates missing genes in the respective datasets.

Gene Name	pre degree	pre btw	post degree	post btw	synapt degree	synapt btw	synapse degree	synapse btw
<i>APP</i>	281 (1)	241364 (1)	690 (1)	1066084 (1)	796 (1)	1457415 (1)	893 (1)	1853558 (1)
<i>YWHAZ</i>	152 (5)	53014 (5)	360 (16)	204559 (11)	382 (16)	238447 (13)	413 (17)	288584 (14)
<i>HSPA8</i>	81 (20)	23019 (14)	151 (64)	49376 (56)	174 (62)	67866 (56)	188 (63)	83081 (57)
<i>TARDBP</i>	-	-	199 (33)	29841 (121)	217 (38)	36381 (140)	228 (44)	40748 (157)
<i>LRRK2</i>	-	-	193 (35)	62417 (42)	194 (48)	63978 (66)	213 (50)	82473 (60)
<i>AKT1</i>	-	-	172 (46)	73603 (36)	184 (52)	86859 (39)	192 (61)	96846 (43)
<i>PTEN</i>	-	-	149 (66)	48542 (61)	160 (77)	56553 (83)	172 (82)	65191 (89)
<i>WWOX</i>	-	-	140 (73)	33321 (103)	160 (75)	34787 (145)	170 (85)	43687 (150)
<i>NEDD4</i>	-	-	133 (86)	47641 (62)	149 (95)	67956 (55)	167 (92)	82186 (62)
<i>ABL1</i>	67 (27)	16002 (23)	-	-	184 (53)	89283 (38)	195 (60)	103213 (40)
<i>GSK3B</i>	46 (59)	11081 (50)	119 (99)	44515 (68)	134 (109)	55065 (87)	141 (119)	60889 (99)
<i>ATF2</i>	-	-	140 (76)	43096 (74)	-	-	156 (108)	53502 (116)
<i>SNCA</i>	64 (30)	14085 (28)	108 (120)	31315 (116)	131 (117)	49294 (100)	133 (137)	52410 (120)
<i>CSNK2B</i>	47 (55)	11922 (43)	109 (116)	37432 (90)	124 (124)	57562 (81)	134 (131)	70544 (78)
<i>RAB7A</i>	47 (56)	10060 (53)	97 (139)	34118 (102)	100 (177)	36551 (139)	106 (192)	40308 (160)
<i>HSPA4</i>	44 (61)	7645 (64)	116 (104)	31763 (113)	131 (115)	38350 (134)	147 (112)	48880 (134)
<i>MAPT</i>	40 (69)	5869 (86)	70 (221)	11467 (295)	76 (262)	13797 (337)	77 (314)	14099 (417)
<i>DLG4</i>	33 (90)	12803 (37)	62 (271)	42157 (77)	62 (345)	44762 (112)	64 (399)	54964 (111)

The top two records have already been detected amongst the overall top 10 node degree genes in the respective networks and information can be found above. A third gene is amongst the top 10 records in all four datasets. The heat shock protein family A (Hsp70) member 8 (*HSPA8*) is a constitutively expressed member of the heat shock protein 70 family. As a chaperone it binds to polypeptides facilitating correct

folding. Additionally it has been shown to function as an ATPase in the disassembly of clathrin-coated vesicles. In this role it is specifically active during transport of membrane components through the cell (Daugaard et al., 2007).

Apart from these three, the glycogen synthase kinase 3 beta (*GSK3B*) is among the top 10 degree nodes in presynapse, postsynapse and synaptosome. It is involved in neuronal cell development and body pattern formation as well as the energy metabolism. Furthermore it has been shown to influence phosphorylation and accumulation of tau and alpha-synuclein (Credle et al., 2015).

The ABL proto-oncogene 1, non-receptor tyrosine kinase (*ABL1*) is ubiquitously expressed and linked to cell cycle functions. It is the one gene that is specific to the presynapse, synaptosome and joint synaptic dataset but not expressed in the postsynapse. As a tyrosine kinase it is involved in cell division, adhesion, differentiation and stress response functions (Paul and Mukhopadhyay, 2004). Alterations in these functions can lead to neuronal degeneration which might explain the proto-oncogenic role of *ABL1* (Wang, 2014).

Activating transcription factor 2 (*ATF2*) is specifically expressed in the postsynapse (as well as the joint synaptosome) and binds to the DNA as part of the leucine zipper family. As such it is associated with various different functions including transcription, histone acetylation and DNA damage response (Desai et al., 2014).

No major evidence for region specific appearance of these two genes could be found but they might be involved in so far unspecified regional processes.

Two of the most prominent PD associated genes: *LRRK2* and *SNCA* are also amongst the top 10 degree nodes. This could point towards a very central role which also allowed their early detection and genetic based disease link.

Apart from these another five genes (*TARDBP*, *AKT1*, *PTEN*, *WWOX*, *NEDD4*) are amongst the postsynaptic, synaptosome and joint synaptic genes and five more amongst the presynaptic top 10 degree records (*CSNK2B*, *RAB7A*, *HSPA4*, *MAPT*, *DLG4*). These are all also present in the other regional datasets but not amongst the top PD associated degree nodes.

Another way to interpret node degree and betweenness scores is in a reverse combination. A combination of a high betweenness score and low degree value (or vice versa) is a prominent support for network modularization. Nodes with a low degree and high betweenness score for example seem to act as a connector between different pathways by separating the two from each other, but allowing communication between them (Koschützki and Schreiber, 2008). One such example is *DLG4*. Even though it

shows relatively low ranking node degree values its betweenness scores rank amongst the top records in the networks. As a scaffolding protein (with top detection coverage in the postsynapse) it likely plays a key role in a range of functions. The ability to connect other proteins can allow information exchange between pathways that are usually separated, explaining the high betweenness score.

6.3.2 Network Clustering

To gain a more “high-level” insight into the substructures within the PPINs, clustering algorithms were used to divide the networks into communities. These represent densely connected network regions which often contain proteins sharing a biological function (Brun et al., 2004). For an overview of all analytical steps taken in this section, Figure 6.2 shows an overview.

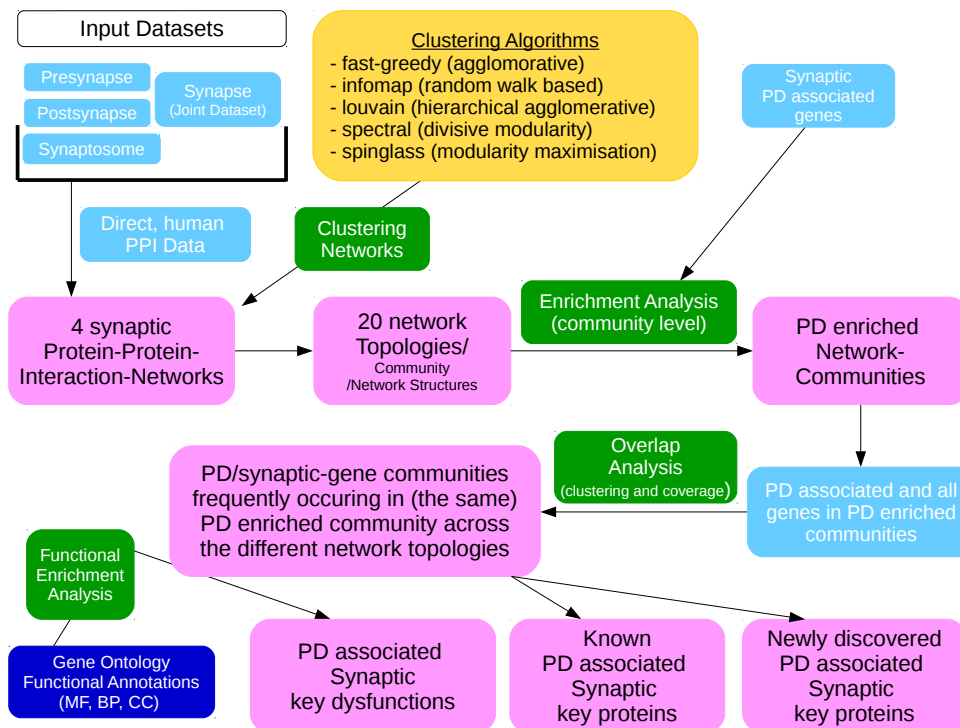


Figure 6.2: Detailed overview including network clustering, enrichment, and key-protein as well as community detection processes in Section 6.3.2. Dark blue boxes refer to published data, light blue boxes are generated datasets, yellow boxes refer to analytical tools, green boxes describe processes and magenta boxes show outcomes.

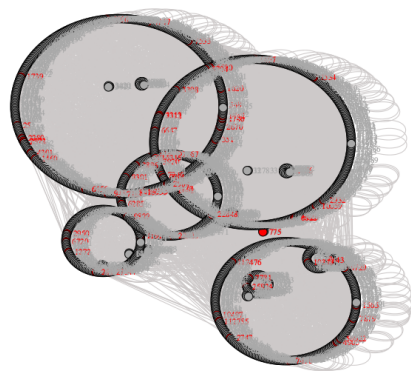
This section considers presynaptic, postsynaptic, synaptosome and joint synaptic PPINs. All four PPINs were analysed using five different clustering algorithms. These

were: fast greedy, infomap, louvain, spectral and spinglass (Section 1.4.2 introduces details and Section 2.4.1 contains technical information). Hence five different topologies emerge for each of the four networks, leading to a total of 20 differently clustered topologies.

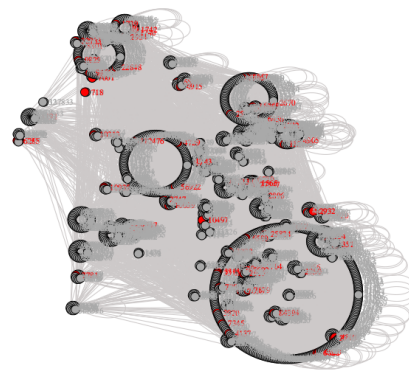
Figures 6.3 and 6.4 show results for the presynapse and the joint synaptic proteome. Visualizations of the postsynapse and synaptosome are very similar to the joint synaptic proteome and can be found in Appendix D (Figures D.1 and D.2). The sub-figures visualize networks based on distinct clusterings: (a) fast greedy, (b) infomap, (c) louvain, (d) spectral and (e) spinglass algorithm. PD associated genes were located in the networks and are highlighted in red. The Figures are included to provide a visual impression of the networks topologies and emerging community structure. These schematics highlight the differences in community number and size amongst the different networks and algorithms. In addition, in certain cases, PD associated genes (highlighted in red) tend to be accumulating in a specific network region, but not in others. To gain a better overview of the presented networks, Table 6.4 summarises key statistics of the differently clustered networks.

Given the focus of this study, significantly PD enriched network communities were identified. Hypergeometric testing was used (Mclean et al., 2016) for this purpose which finds network communities with a higher number of disease associated genes than expected by chance (compared to a random allocation, given the network environment). The significance threshold for disease enrichment was set to a p-value of 0.05. Table 6.4 also includes the number of PD enriched communities in the different networks.

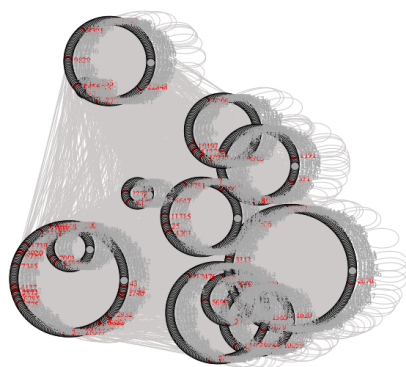
As Figures 6.3 and 6.4, as well as Table 6.4 show, clustering algorithms divide networks differently. Irrespective of the clustering algorithm the presynaptic dataset differs slightly from the others due to its smaller size. The other three networks are relatively similar. The following sections highlights general properties of emerging structural topologies based on the use of different network clustering algorithms (in alphabetic order).



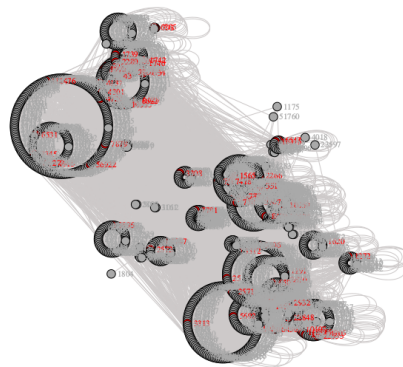
(a) fast greedy



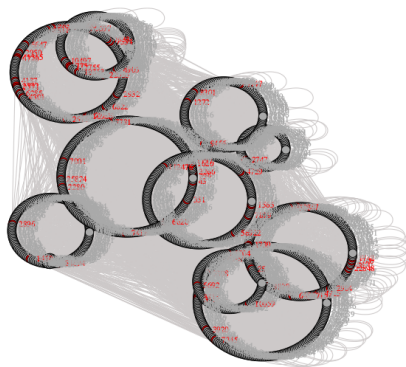
(b) infomap



(c) louvain



(d) spectral



(e) spinglass

Figure 6.3: Presynaptic PPINs. Different clustering algorithm results are highlighted. Red coloured nodes represent PD associated genes. Grey “background” shows network edges.

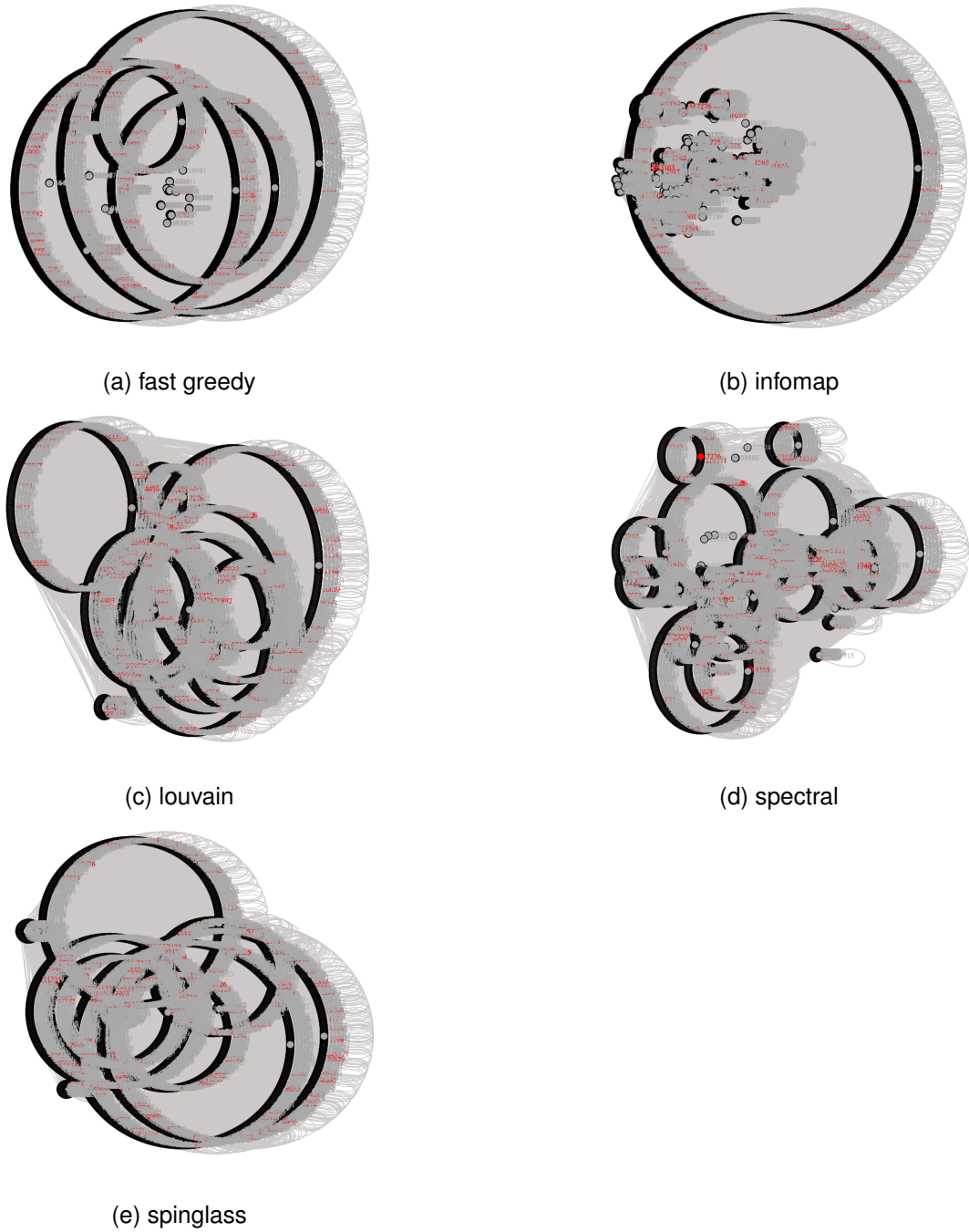


Figure 6.4: Joint synapticPPINs. Different clustering algorithm results are highlighted. Red coloured nodes represent PD associated genes. Grey “background” shows network edges.

Table 6.4: Results obtained from the clusterings of the networks of the different regional datasets and using different clustering algorithms. Columns 3, 4, 5, 9 and 10 refer to the number of respective communities. “smaller 4” and “larger 200” refers to the number of nodes per community. Remaining columns refer to the number of nodes per community.

Regional Data	Clustering Algorithm	communities	communities with PD genes	PD enriched communities (smaller 4 in parenthesis)	maximum community size	minimum community size	average community size	communities smaller 4	communities larger 200
presynapse	fast greedy	17	9	0	404	2	91	4	3
postsynapse	fast greedy	25	5	0	1349	2	182	9	0
synaptosome	fast greedy	48	9	1 (1)	1939	2	112	23	3
synapse	fast greedy	26	8	0	2061	2	233	8	4
presynapse	infomap	108	38	1	250	2	14	14	1
postsynapse	infomap	199	47	7 (1)	1645	2	23	28	1
synaptosome	infomap	235	57	6	1918	2	23	34	2
synapse	infomap	258	62	8	2155	2	24	27	2
presynapse	louvain	12	12	2	235	25	129	0	2
postsynapse	louvain	12	10	0	780	4	380	0	9
synaptosome	louvain	13	12	0	1005	4	412	0	10
synapse	louvain	13	12	1	977	4	467	0	9
presynapse	spectral	53	26	2 (1)	165	1	29	19	0
postsynapse	spectral	67	29	1	400	1	68	33	8
synaptosome	spectral	95	35	(3)	427	1	56	53	12
synapse	spectral	94	32	4 (1)	472	1	65	56	12
presynapse	spinglass	11	10	1	221	13	141	0	2
postsynapse	spinglass	15	11	0	767	9	304	0	8
synaptosome	spinglass	12	10	0	1002	4	446	0	8
synapse	spinglass	14	11	0	1115	6	433	0	7

The fast-greedy algorithm produces some large, with up to ~2,000 genes, and some very small communities. It appears as though PD associated genes allocate in similar communities, but only one community in the synaptosome PPIN is actually enriched for PD.

The infomap algorithm produces few, very large communities. Only two are larger than 200 nodes, but a maximum of 2,155 genes is found in one community. Additionally a large number of small communities can be found. Nevertheless the clustering results indicate a relatively large number of communities (between 1-8 per network) being enriched for PD associated genes, in other words showing an over-representation of PD associated genes. Only one PD enriched community in the network contains less than four nodes.

Moving on with the louvain clustering results, no community is smaller than four genes and a large number of big communities, with more than 200 nodes, exists. This also confirms the large average community size, ranging between 130 and 467 genes.

The spectral clustering results lead to maximum community sizes between 165 and 472 nodes. The average community size ranges between 29 and 68 genes. A number of communities with more than 200 genes exist, but none of them are extremely large.

The clustering also leads to the maximum number of small communities. A number of PD enriched communities can be found in all four networks, with some being small and in some sense isolating the disease associated genes from others.

The spinglass algorithm seems to perform very similar to the louvain algorithm. Very large communities emerge with large average sizes and only one (presynaptic) PD enriched community.

Overall the analysis shows the differences in the clustering results. It appears that the increased size of the postsynaptic, synaptosome and joint synaptic proteomes make it harder to identify biologically meaningful and disease enriched communities, compared to the smaller presynaptic PPIN. Considering the different clustering algorithms it seems that the spectral and infomap approach show the most useful division of the data into communities, specifically when considering PD enrichment. As previously shown the emerging communities are very precise considering functional similarities (McLean et al., 2016). Since there is no established and straight forward technique to best classify clustering results and/or compare them amongst each other all generated communities were considered in the next step. The coming Section 6.3.3 studies PD enriched communities further, with the aim to identify a significantly affected synaptic region.

6.3.3 PD Enriched Communities

Since one of the main objectives of this work was to identify most PD affected synaptic subregions, significantly disease enriched communities were extracted (p-value < 0.05). 41 PD enriched communities were identified and can be seen in Table 6.5. Information regarding community sizes and PD enrichment (corrected p-values) are included. Seven of the identified communities contain less than four genes and were not considered further. The remaining 34 were specifically addressed regarding their similarities and differences.

The enriched communities were analysed regarding the regional synaptic dataset as well as the clustering algorithm. Only communities with a minimum of four genes were considered. One PD enriched community emerges after fast greedy and spinglass clustering in the synaptosome and presynaptic network respectively. The louvain clustering leads to a little more with three enriched communities. Overall the spectral and infomap clustering results show the highest numbers of enriched communities.

These insights reflect a previous observation that the spectral and infomap algo-

rithms seem to be most appropriate to divide large PPINs into biologically interpretable communities containing proteins with common functions (Section 6.3.2) (McClean et al., 2016).

Table 6.5: PD enriched communities in the different networks based on one of the four datasets and one of the five clustering algorithms (p -value < 0.05 , after multiple testing correction). All enriched communities are listed, irrespective of their size. Rows are ordered based on dataset and algorithm. “synapse” refers to the joint synaptic proteome. Grey font highlights communities with less than four genes.

Regional Data	Clustering Algorithm	Community number	Genes in community	PD associated genes in community	PD enrichment p-value (corrected)
presynapse	infomap	90	4	2	1.23×10^{-02}
presynapse	louvain	3	25	4	2.69×10^{-02}
presynapse	louvain	11	209	25	2.78×10^{-06}
presynapse	spectral	10	3	2	6.11×10^{-03}
presynapse	spectral	69	17	3	4.02×10^{-02}
presynapse	spectral	72	79	11	6.81×10^{-04}
presynapse	spinglass	3	183	20	1.49×10^{-04}
postsynapse	infomap	10	72	6	3.35×10^{-02}
postsynapse	infomap	22	37	7	1.91×10^{-04}
postsynapse	infomap	54	12	2	5.98×10^{-02}
postsynapse	infomap	69	10	2	4.26×10^{-02}
postsynapse	infomap	81	10	1	4.26×10^{-02}
postsynapse	infomap	84	9	2	3.48×10^{-02}
postsynapse	infomap	126	7	2	2.12×10^{-02}
postsynapse	infomap	177	3	1	3.32×10^{-03}
postsynapse	spectral	55	55	6	9.70×10^{-02}
synaptosome	fast greedy	12	11	1	4.85×10^{-02}
synaptosome	fast greedy	24	3	1	3.15×10^{-03}
synaptosome	infomap	9	78	6	4.22×10^{-02}
synaptosome	infomap	27	32	6	5.14×10^{-04}
synaptosome	infomap	76	12	2	5.70×10^{-02}
synaptosome	infomap	88	10	2	4.06×10^{-02}
synaptosome	infomap	101	10	1	4.06×10^{-02}
synaptosome	infomap	137	8	1	2.63×10^{-02}
synaptosome	spectral	48	1	1	3.27×10^{-02}
synaptosome	spectral	60	1	1	3.27×10^{-02}
synaptosome	spectral	85	1	1	3.27×10^{-02}
synapse	infomap	12	68	6	2.09×10^{-02}
synapse	infomap	35	25	3	4.41×10^{-02}
synapse	infomap	60	15	3	1.10×10^{-02}
synapse	infomap	71	14	3	9.00×10^{-03}
synapse	infomap	96	10	3	3.20×10^{-02}
synapse	infomap	98	12	2	5.43×10^{-02}
synapse	infomap	102	11	2	4.62×10^{-02}
synapse	infomap	200	6	1	1.40×10^{-02}
synapse	louvain	6	75	7	2.51×10^{-03}
synapse	spectral	3	148	9	4.58×10^{-02}
synapse	spectral	27	150	9	4.92×10^{-02}
synapse	spectral	54	55	10	7.11×10^{-06}

Table 6.5: PD enriched communities in the different networks based on one of the four datasets and one of the five clustering algorithms (p -value < 0.05 , after multiple testing correction). All enriched communities are listed, irrespective of their size. Rows are ordered based on dataset and algorithm. “synapse” refers to the joint synaptic proteome. Grey font highlights communities with less than four genes.

Regional Data	Clustering Algorithm	Community number	Genes in community	PD associated genes in community	PD enrichment p-value (corrected)
synapse	spectral	97	1	1	3.18×10^{-02}
synapse	spectral	100	4	2	5.73×10^{-03}

Considering differences amongst regional datasets, six presynaptic, eight postsynaptic, seven synaptosome and 13 full synapse communities are amongst the enriched ones (with a minimum of four genes). This relatively even distribution can also be observed amongst the communities detected after clustering with the infomap and spectral algorithm.

Taking these findings into account, there was no obvious trend towards one of the datasets being specifically associated to PD. Therefore all PD enriched communities with at least four genes were compared amongst each other. Figure 6.5 shows the 71 PD associated genes in the 34 significantly enriched communities (x-axis). The heatmap shows clusters of communities based on the genes shared between them, following a hierarchical, agglomerative clustering approach (implemented in the Python `seaborn.clustermap` package²). The labelling of the y-axis reflects the dataset (presynaptic, postsynaptic, synaptosome or joint synaptic proteome) as well as the clustering algorithm (fast greedy, infomap, louvain, spinglas, spectral) and the number of the enriched community in the respective (network) topology. Apart from considering PD associated genes only, all 819 genes in the enriched communities were considered, clustered and visualized. Figure 6.6 shows the results.

The clustering of PD associated genes in PD enriched communities (Figure 6.5) leads to three prominent “gene blocks”. These are highlighted in a green, blue and red box and will be referred to as Cluster 1, Cluster 2 and Cluster 3. Clusters represent PD associated genes appearing together in a number of communities in the differently clustered networks. Similarly three blocks can be identified considering the similarity between the full communities (Figure 6.6). Closer examination reveals that independently of considering only PD associated genes in enriched communities, or all genes in enriched communities Cluster 1, 2 and 3 contain similar network communities.

²<http://seaborn.pydata.org/generated/seaborn.clustermap.html>

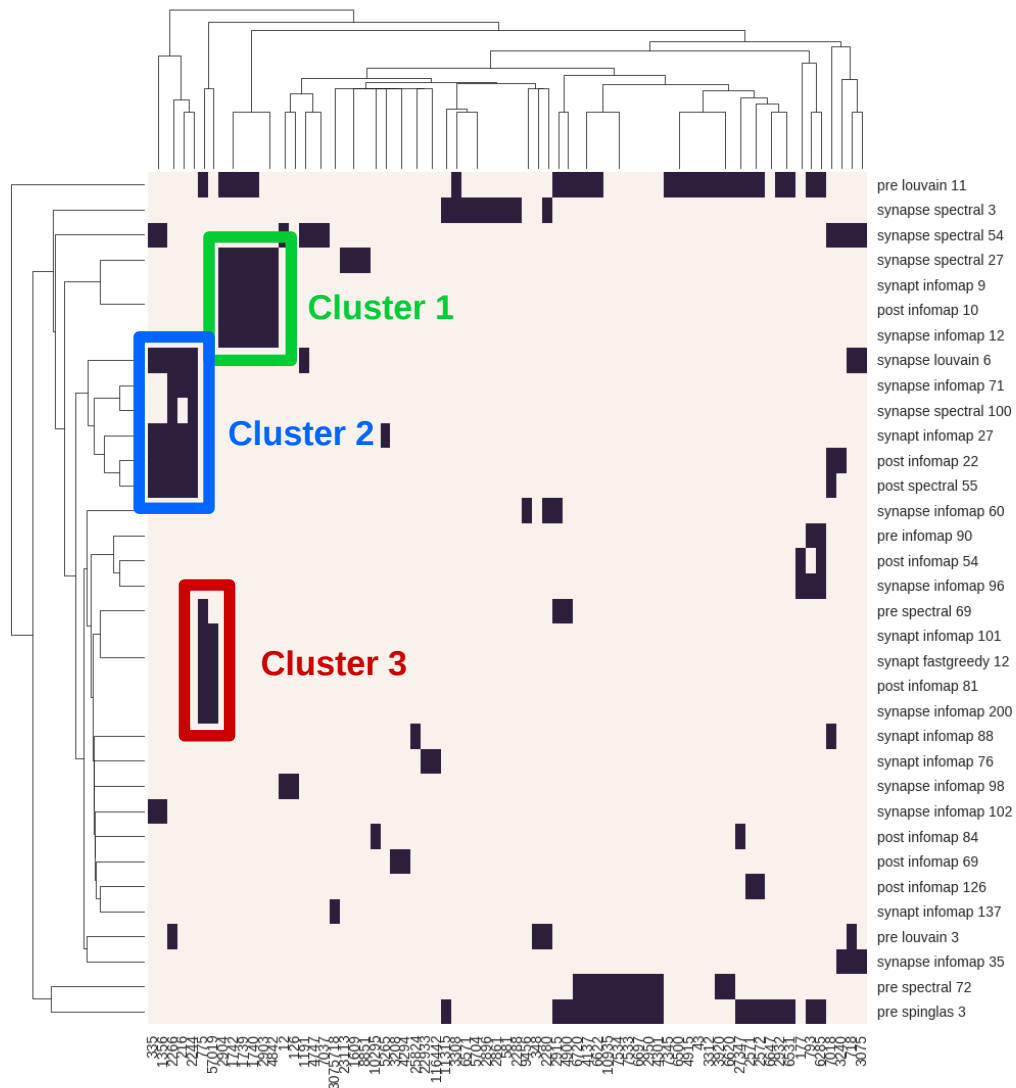


Figure 6.5: Clustering highlighting the overlap of PD associated genes in significantly PD enriched communities. The x-axis shows Entrez IDs and the y-axis indicates the dataset, algorithm and community number in which the community was found to be enriched for PD associated genes. “pre”, “post”, “synapt” and “synapse” refer to the presynaptic, postsynaptic, synaptosome and joint synaptic proteome.

This consistency supports credibility of the community structure. Hence even though clustering algorithms are based on different principles they all detected highly similar PD enriched communities containing the same set of PD associated genes. Genes in these clusters show a high probability of being associated with PD, as well as influencing its development and manifestation.

A closer look at the communities in the clusters suggests that these were mostly

found in either postsynaptic or full synaptic PPIN communities. Nevertheless presynaptic communities show enrichment as well. Overall enriched communities emerged in networks based on different regional datasets. This may indicate ubiquitous effects of PD on the synapse, not targeting a specific synaptic region. As suspected based on the overall number of enriched communities emerging through the different clusterings, the spectral and infomap algorithm are preferred best.

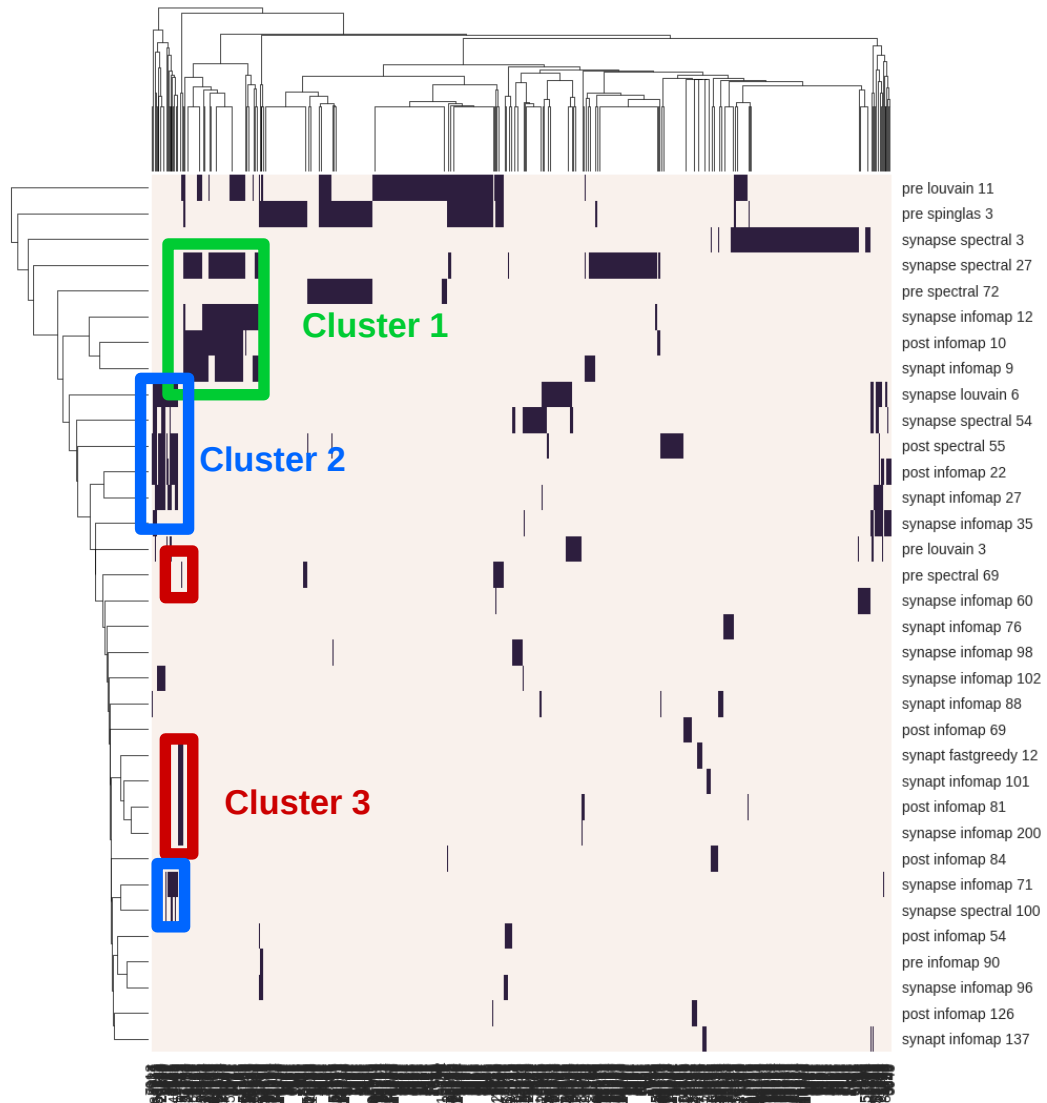


Figure 6.6: Clustering highlighting the overlap of all genes in significantly PD enriched communities (including all community genes). x- and y-axis labelling are as in Figure 6.5.

To investigate the clustered communities in more detail, Table 6.6 shows a summary of these. Information overlaps with Table 6.5 but contains further community

cluster information and the colour code corresponds to the ones in Figures 6.5 and 6.6.

Table 6.6: Three clusters of PD enriched communities. Cluster numbers and colour code as in Figures 6.5, 6.6 and 6.7. “Community Number” refers to the community in the original network; “Genes” refers to the number of genes in the community; “Communities in cluster” refers to the colour coded clusters; columns 8-12 refer to community counts in total and unique for the three clusters.

Cluster	Dataset	Algorithm	Community Number	Genes	PD genes	Communities in cluster	Genes in any community in cluster	unique genes in any community in cluster	average genes per community	PD genes in any community in cluster	unique PD genes in any community in cluster
1	synapse	spectral	27	150	9	4	368	172	92	27	9
1	synapse	infomap	12	68	6	4	368	172	92	27	9
1	post	infomap	10	72	6	4	368	172	92	27	9
1	synaptosome	infomap	9	78	6	4	368	172	92	27	9
2	post	infomap	22	37	7	5	217	113	53.25	30	11
2	synapse	louvain	6	75	8	5	217	113	53.25	30	11
2	synapse	infomap	71	14	3	5	217	113	53.25	30	11
2	post	spectral	55	55	6	5	217	113	53.25	30	11
2	synaptosome	infomap	27	32	6	5	217	113	53.25	30	11
2	synapse	spectral	100	4	2	5	217	113	53.25	30	11
3	pre	spectral	69	17	3	5	54	37	13.5	11	4
3	synapse	infomap	200	6	2	5	54	37	13.5	11	4
3	synaptosome	infomap	101	10	2	5	54	37	13.5	11	4
3	post	infomap	81	10	2	5	54	37	13.5	11	4
3	synaptosome	fast greedy	12	11	2	5	54	37	13.5	11	4

For an even better understanding and to identify the potential new PD associated gene sets, coverage of individual genes amongst the different communities in the clusters was analysed. Figures 6.7 a and b show the coverage of PD associated genes, as well as all genes in the 15 enriched communities belonging to one of the three clusters.

As Table 6.6 and Figure 6.7 show, there is variability in the total number of genes in the clusters. To confirm the overlap of genes between the communities in each of the clusters with more detail, their coverage was analysed. Figure 6.7 a shows that a substantial number of PD associated genes in the different clusters, illustrated with different colour bars, appear in more than only one community of the cluster, meaning that the coverage is higher 1. Similarly Figure 6.7 a) highlights that a substantial proportion of genes appears in more than one community in the respective cluster.

More precisely, Cluster 1 is the largest one with a total of 172 unique genes out of which less than half (82) appear in only one community. Cluster 2 contains 113 unique genes with a slightly higher proportion (slightly less than two thirds) of genes detected in only one community. With 37 unique genes out of which 31 have only been detected once, Cluster 3 is the least consistent one.

For best data consistency and to identify a core key target gene set, genes were only considered further if they appeared in at least two of the four, five or six communities

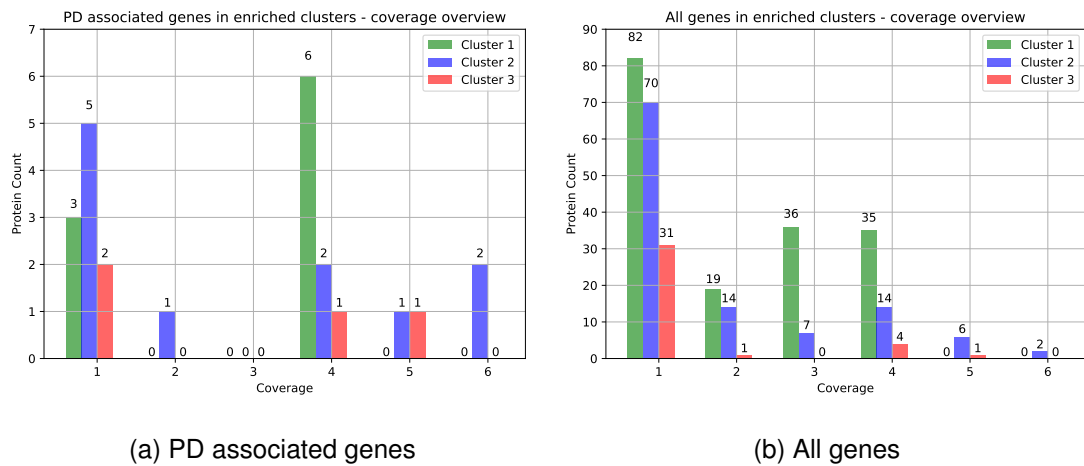


Figure 6.7: Coverage of different proteins in the PD enriched communities in the three enriched community clusters. Colours represent clusters. The x-axis indicates the coverage. Genes are associated to the coverage based on the number of PD enriched communities in the respective cluster they appear in.

per cluster. Hence Cluster 1 contains 90, Cluster 2, 43 and Cluster 3 six core genes, including six, six and two PD associated ones amongst them (Figure 6.7). Appendix E contains the full list of genes in these clusters.

Based on these insights properties of the gene sets as well as individual genes in the three clusters were further investigated.

6.3.4 Synaptic PD Affected Functions

After having identified most PD affected synaptic regions, these were analysed regarding their predominant overall functions. Gene Ontology (GO) enrichment analysis (for the three categories Biological Process, Molecular Function and Cellular Component) was carried out. This was initially done for each individual community in the enriched clusters. The background dataset was chosen to be the joint synaptic proteome and the analysis was carried out with `topONTO` using the Fisher exact test, `elim` algorithm and Benjamini and Yekutieli multiple testing correction. This allowed identification of very specific enriched terms, being found in the lower levels of the ontology trees, guaranteeing the optimum and most specific insight into the joint and dominating functionality of the genes in the communities.

Enriched terms for individual communities in a cluster were compared and overlapped largely. Every term appearing in at least two of the communities in a cluster was taken forward, and included in the final results. The following paragraphs address

the three clusters individually.

6.3.4.1 Cluster 1

Cluster 1 consists of four communities and contains 90 genes appearing in two or more of them. For a better overview, the 90 genes are visualised in Figure 6.8. It can be seen that all genes undergo interactions with each other. There are 246 internal PPIs in total.

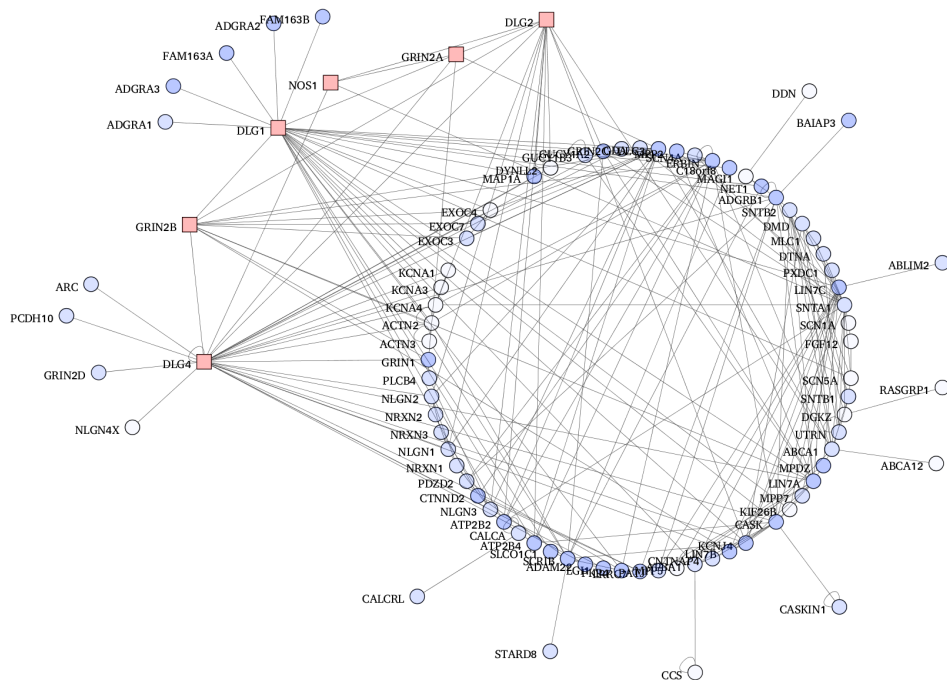


Figure 6.8: Genes in Cluster 1 (minimum coverage of two). Opacity represents the coverage. Red squares are PD associated genes.

As highlighted in Figure 6.8 six PD associated genes can be found in Cluster 1. Table 6.7 lists these, as well as three additional ones with a coverage of 1. Two to three different functional areas are covered by the PD associated genes in this cluster.

DLG1, *DGL2*, *DLG4* encode scaffolding proteins, playing a main role in the structural organisation of proteins and facilitating their full functionality. *DLG1* is required for “normal development”, playing a major role in signal transduction, cell proliferation and synaptogenesis. It is also referred to as *SAP-97* (Howard et al., 2010). *DGL2* and *DLG4* encode proteins which heteromultimerize to form the membrane-associated guanylate kinase (MAGUK) and are also referred to as *PSD-93* and *SAP-102*. As such

Table 6.7: PD associated genes in Cluster 1. Ordered by coverage and Entrez ID.

Gene ID	Gene Name	Short Gene Description	Coverage
1739	<i>DLG1</i>	DLG1 discs large MAGUK scaffold protein 1	4
1740	<i>DLG2</i>	DLG2 discs large MAGUK scaffold protein 2	4
1742	<i>DLG4</i>	DLG4 discs large MAGUK scaffold protein 4	4
2903	<i>GRIN2A</i>	GRIN2A glutamate ionotropic receptor NMDA type subunit 2A	4
2904	<i>GRIN2B</i>	GRIN2B glutamate ionotropic receptor NMDA type subunit 2B	4
4842	<i>NOS1</i>	NOS1 nitric oxide synthase 1	4
1609	<i>DGKQ</i>	diacylglycerol kinase theta	1
8851	<i>CDK5R1</i>	cyclin dependent kinase 5 regulatory subunit 1	1
23113	<i>CUL9</i>	cullin 9	1

they interact with the postsynaptic membrane, being recruited into NMDA receptor and potassium channel clusters. In these regions they form a scaffold for the clustering of receptors, ion channels and associated signalling proteins (Oliva et al., 2012; Sun and Turrigiano, 2011).

Furthermore two glutamate ionotropic receptors are amongst the PD associated genes in Cluster 1. *GRIN2A* and *GRIN2B* both encode for proteins of the N-methyl-D-aspartate (NMDA) receptor family, also referred to as *GluN2A* and *GluN2B*. These receptors are both ligand- and voltage dependant and involved in long-term potentiation and synaptic transmission efficacy, showing links to specific memory types and learning. These functionalities are regulated based on Ca^{2+} influx into the postsynapse (Paoletti et al., 2013). Apart from their joint properties *GRIN2B* specifically acts as an agonist binding site for glutamate (Hu et al., 2016).

Additionally *NOS1* is amongst the six PD associated genes. Nitric oxide synthase 1 belongs to the family of nitric oxide synthases, synthesizing nitric oxide from L-arginine (Stuehr, 2004). Nitric oxide has been linked to neurodegenerative disease since it adopts a neurotransmitter like role inducing neurotoxicity (Dawson and Dawson, 1996).

In summary, the PD associated genes are associated with functions generally known to be linked with PD. In addition, apart from generic terms, some findings point towards much more concrete dysfunctions. A slight focus towards postsynaptic dysfunctions can be detected based on information in Cluster 1.

To understand the overall function of all genes in Community 1 common enriched functions were identified. Table 6.8 shows GO terms enriched amongst all genes in the PD associated communities from the Biological Process, Molecular Function and Cellular Component ontologies. Appendix Table F.1 shows the GO terms, IDs as well

as short definitions of the terms. These were retrieved from QuickGO³ via GONUTS⁴.

Table 6.8: GO terms enriched in at least two communities of Cluster 1 (alphabetical order); significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, `elim` algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values not supplied since different in distinct enriched clusters).

Biological Process	Molecular Function	Cellular Component
GDP metabolic process	cell adhesion molecule binding	basolateral plasma membrane
gephyrin clustering involved in postsynaptic density assembly	extracellular-glutamate-gated ion channel activity	bicellular tight junction
GMP metabolic process	guanylate kinase activity	cell junction
ionotropic glutamate receptor signaling pathway	ionotropic glutamate receptor binding	dendritic spine
maintenance of epithelial cell apical/basal polarity	L27 domain binding	dystrophin-associated glycoprotein complex
negative regulation of peptidyl-cysteine S-nitrosylation	neurexin family protein binding	exocyst
neurotransmitter secretion	neuroligin family protein binding	juxtaparanode region of axon
positive regulation of excitatory postsynaptic potential	NMDA glutamate receptor activity	MPP7-DLG1-LIN7 complex
positive regulation of synapse assembly	PDZ domain binding	myelin sheath abaxonal region
positive regulation of synaptic vesicle clustering	scaffold protein binding	neuron projection
postsynaptic density protein 95 clustering		NMDA selective glutamate receptor complex
protein localization to basolateral plasma membrane		postsynaptic density of dendrite
receptor localization to synapse		postsynaptic membrane
regulation of grooming behaviour		presynaptic membrane
regulation of sodium ion transmembrane transport		presynapse
vocalization behaviour		sarcolemma
		synapse
		T-tubule
		voltage-gated potassium channel complex
		Z disc

The following paragraphs highlight some of the enriched functions that stand out in the context of PD and the presented analysis.

³<https://www.ebi.ac.uk/QuickGO/>

⁴https://gowiki.tamu.edu/wiki/index.php/Main_Page

As previously highlighted by the functions of individual PD associated genes, “positive regulation of synapse assembly” and “receptor localization to synapse”, appear amongst the enriched Biological Processes. Finding “scaffold protein binding” amongst the enriched Molecular Function terms confirms the functional role of proteins in Cluster 1 with respect to scaffolding proteins. These findings support the hypothesis that scaffolding proteins, and more generally the spatial organisation of genes, is affected in brain cells of PD patients.

Similarly “ionotropic glutamate receptor signalling pathway” as well as “neurotransmitter secretion” are amongst the enriched Biological Process terms. This is confirmed through the Molecular Function terms “ionotropic glutamate receptor binding” as well as “NMDA glutamate receptor activity” and the Cellular Component “NMDA selective glutamate receptor complex”. Hence, it seems quite likely that NMDA receptors can be highly affected in PD patients. Their role in the disease pathology can also be confirmed by the use of glutamatergic receptors as therapeutic targets (Johnson et al., 2009; Hallett and Standaert, 2004).

A number of other overall affected pathways appear. Enriched Cellular Component terms largely focus around the “synapse”. More specifically they including terms such as the “presynaptic membrane” as well as “postsynaptic membrane”. Together with the term “cell junction”, this supports the theory that PD has a substantial influence on synaptic information transmission.

Further terms based on the Biological Process ontology include terms related to (intra-) cellular structure. “gephyrin clustering involved in postsynaptic density assembly”, “positive regulation of synaptic vesicle clustering” as well as “protein localization to basolateral plasma membrane” confirm the possible alteration of structure related processes in the brain cells of PD patients.

Overall, enriched terms partly overlap with the known PD associated genes. The generation of an extended set of so far non-PD associated genes indicates their role in disease affected processes and makes these genes potential next targets to investigate regarding their link to PD and their potential use as a drug targets or biomarkers.

Apart from the presented PD associated genes, eight genes in Cluster 1 could be found in at least one of the analysed PD expression data studies presented in Section 3.3.2. These are: *FBG12* (2257), *DTNA* (1837), *NRXN2* (9379), *GUCY1B3* (2983), *APBA1* (320), *ATP2B2* (491), *ERBIN* (55914), *PKP4* (8502). This overlap is an indication of having identified a highly PD affected synaptic gene set.

6.3.4.2 Cluster 2

Cluster 2 consists of six communities which contain 43 genes with a minimum coverage of two. The 43 genes undergo 74 internal interactions and Figure 6.9 visualises these. Six PD associated genes with a minimum coverage of two can be found in the set as well as another five appearing in only one of the communities. Table 6.9 lists these.

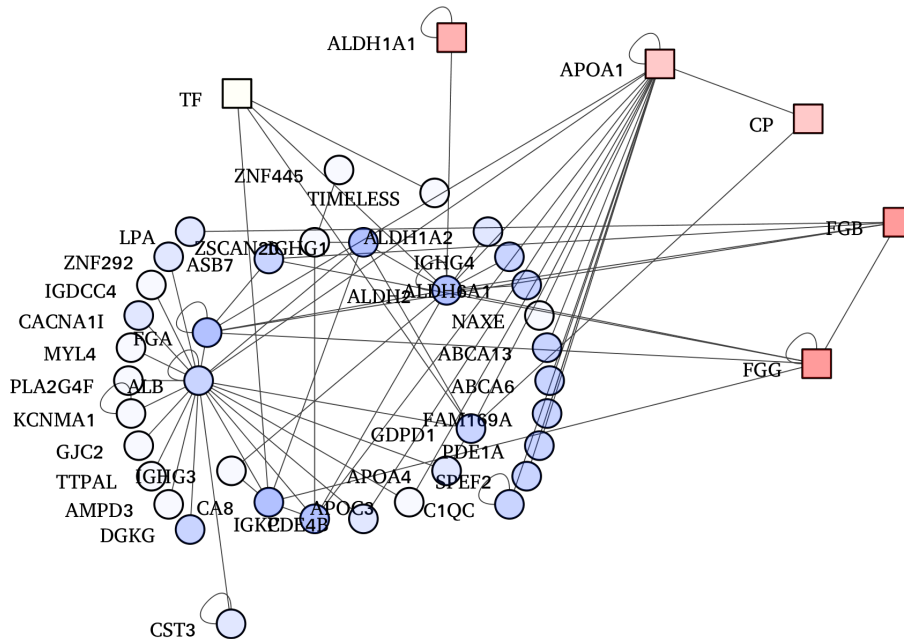


Figure 6.9: Genes in Cluster 2 (minimum coverage of two). Opacity represents the coverage. Red squares are PD associated genes.

It seems that PD associated genes in Cluster 2 are linked to four different functionalities which are presented in the next paragraphs.

FGB and *FGG* are the beta and alpha chain of fibrinogen, a blood-borne glycoprotein comprised of three pairs of non-identical polypeptide chains. Fibrinogen is a protein which is highly involved in the formation of blood clots. It is cleaved by thrombin and its cleavage products have been associated with cell adhesion and cell spreading. Furthermore they showed vasoconstrictor and chemotactic activities. Since these functions do not seem specific to the synapse the expression pattern of *FGB* and *FGG* was investigated. Both of the protein products are highly expressed in the liver, but expression for *FGG* was confirmed in the presynapse, postsynapse and synapto-

Table 6.9: PD associated genes in Cluster 2. Ordered by coverage and Entrez ID.

Gene ID	Gene Name	Short Gene Description	Coverage
2244	<i>FGB</i>	fibrinogen beta chain	6
2266	<i>FGG</i>	fibrinogen gamma chain	6
216	<i>ALDH1A1</i>	ALDH1A1 aldehyde dehydrogenase 1 family member A1	5
335	<i>APOA1</i>	APOA1 apolipoprotein A1	4
1356	<i>CP</i>	CP ceruloplasmin	4
7018	<i>TF</i>	transferrin	2
718	<i>C3</i>	complement C3	1
1191	<i>CLU</i>	clusterin	1
3075	<i>CFH</i>	complement factor H	1
3240	<i>HP</i>	haptoglobin	1
5265	<i>SERPINA1</i>	serpin family A member 1	1

some proteome, while *FGB* is expressed in the postsynapse and synaptosome with a coverage of five or higher.

ALDH1A1 is a gene involved in the alcohol metabolism and *APOA1* is the major component of the high density lipoprotein *HDL* in the plasma. Its transcript is involved in promoting cholesterol efflux from tissues to the liver. Both of these genes are also highly expressed and transcribed in the liver.

The remaining two genes ceruloplasmin (*CP*) and transferrin (*TF*) are also prominent in the liver. Ceruloplasmin binds most of the copper in the plasma and is involved in peroxidation of Fe(II)transferrin to Fe(III)transferrin. Its dysfunction leads to iron accumulation inducing tissue damage and neurologic abnormalities. Transferrin itself acts as an iron transporter from the intestine and reticuloendothelial system as well as the liver parenchymal cells, to all proliferating cells in the body. Transferrin has also been associated with PD in at least one of the PD expression studies (Section 3.3.2). Additionally *APP* the gene with highest node degree and betweenness score in all networks is linked to iron export (Section 6.3.1).

It seems as though a number of PD associated genes are highly related to liver functions. This might be surprising since all of these genes are also expressed in the synapse. To confirm these findings Table 6.10 shows Biological Process, Molecular Function and Cellular Component terms enriched in at least two communities in Cluster 2. Appendix Table F.2 shows the respective GO terms, IDs, as well as short definitions of the terms (retrieved from QuickGO via GONUTS).

Compared to the first cluster, far less functional terms are enriched for genes in Cluster 2. This might be due to the smaller size or more diverse functionality of genes

Table 6.10: GO terms enriched in at least two communities of Cluster 2 (alphabetical order); significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, `elim` algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values not supplied since different in distinct enriched clusters).

Biological Process	Molecular Function	Cellular Component
complement activation, classical pathway	immunoglobulin receptor binding serine-type endopeptidase activity	blood microparticle external side of plasma membrane fibrinogen complex immunoglobulin complex, circulating platelet alpha granule lumen

leaving less terms enriched. Nevertheless, the results partly confirm functionalities detected based on PD associated genes in this cluster.

“Complement activation, classical pathway” is the enriched Biological Process term. It is a component of the innate immune system (Schlachetzki and Winkler, 2015) and has previously been linked to PD. This could point towards an autoimmune reaction leading to the cell death of neurons (specifically in PD patients). Immune system related terms can be found amongst the enriched Molecular Functions as well. “Serine-type endopeptidase activity” as well as “immunoglobulin receptor binding” have been identified. Quite often immune responses are triggered from the liver which hosts a large number of natural killer and natural killer T cells (Racanelli and Rehmann, 2006). Hence, this finding overlaps with the regional specificity highlighted, based on the functionality of PD associated genes.

Considering the enriched Cellular Components “immunoglobulin complex, circulating” confirms identified Biological Process and Molecular Function terms. Furthermore the functionality of PD associated genes is reflected by “blood microparticle”, “platelet alpha granule lumen” and “fibrinogen complex”. All these terms associate with the neuroinflammatory pathway which leads to an increase in brain barrier permeability. This can lead to the detection of these terms in association to PD (Section 1.1.1.1).

These terms highlight a possible different explanation to the manifestation and/or the underlying causes of PD. A direct link between fibrinogen levels in elderly Japanese-American men and PD prevalence could be shown (Wong et al., 2010) and well as overall elevated fibrinogen levels in PD patients (Lu et al., 2014). Furthermore, a link

between the immune system and PD has been discussed, and this analysis supplies a more concrete set of 6 PD associated genes in addition to another 37 genes, with a potential high impact and link to the disease.

6.3.4.3 Cluster 3

Cluster 3 consists of five communities and has the smallest number of genes. Out of 37 only six genes show a coverage of two or higher. These are connected via six internal PPIs. Figure 6.10 gives an overview and Table 6.11 lists all PD associated genes in Cluster 3. Two of these are found in more than two communities and two others only in one.

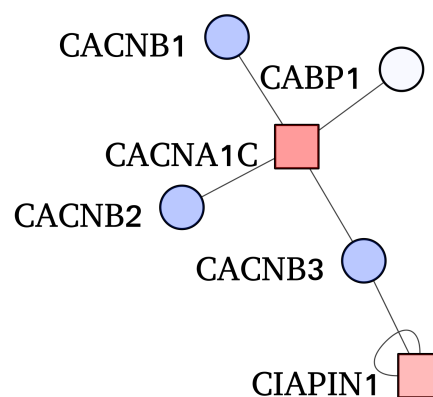


Figure 6.10: Genes in Cluster 3 (minimum coverage of two). Opacity represents the coverage. Red squares are PD associated genes.

Table 6.11: PD associated genes in Cluster 3. Ordered by coverage and Entrez ID.

Gene ID	Gene Name	Short Gene Description	Coverage
775	<i>CACNA1C</i>	CACNA1C calcium voltage-gated channel subunit alpha1 C	5
57019	<i>CIAPIN1</i>	CIAPIN1 cytokine induced apoptosis inhibitor 1	4
2915	<i>GRM5</i>	glutamate metabotropic receptor 5	1
4900	<i>NRGN</i>	neurogranin	1

One of the PD associated genes in Cluster 3 is *CACNA1C* which encodes an alpha-1 subunit of a voltage-dependent calcium channel. As such it is involved in membrane depolarization and Ca^{2+} influx into neurons. As a member of the receptor subfamily 1 its main roles are integration of synaptic input in neurons and synaptic transmission (Catterall, 2011). More generally Ca^{2+} levels are also crucial to maintain energy homeostasis which is highly important to maintain a healthy cell state and alterations

have been linked to PD previously (Hurley and Dexter, 2012). Additionally the gene *CACNB3* (784) is one of the other genes in the community and found in at least one of the analysed expression studies identifying PD associated genes (Section 3.3.2). It is involved in the regulation of voltage-dependent calcium channels confirming the importance of a link between Ca^{2+} and PD.

Furthermore *CACNA1C* has been linked to a large number of other neurodegenerative diseases (Lee et al., 2016) such as schizophrenia, bipolar disorder and others. First studies have proposed to use it as a drug target, as a key member of calcium channels (Imbrici et al., 2013).

The second PD associated gene is the cytokine induced apoptosis inhibitor 1, *CIA-PINI*, which points towards another PD affected pathway: apoptosis. Apoptotic processes are also regulated based on Ca^{2+} levels (Pinton et al., 2008), which might explain the link between the two genes appearing in the same communities across the analysed networks.

Based on these brief insights relying on the PD associated genes overall gene functions in Cluster 3 were studied. Given the small cluster size it was uncertain if significant results could be obtained. Table 6.12 shows GO Biological Process, Molecular Function, Cellular Component GO terms found in at least two communities in the cluster. For further details Appendix Table F.3 shows the respective GO terms with their IDs as well as short definitions of the term. These were retrieved from QuickGO via GONUTS.

Table 6.12: GO terms enriched in at least two communities of Cluster 3; significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, `elim` algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values not supplied since different in distinct enriched clusters).

Biological Process	Molecular Function	Cellular Component
neuromuscular junction development	high voltage-gated calcium channel activity	L-type voltage-gated calcium channel complex

The enriched functions are “neuromuscular junction development” (Biological Process), “high voltage-gated calcium channel activity” (Molecular Function) as well as “L-type voltage-gated calcium channel complex” (Cellular Component). Two of the terms relate to calcium channels, and the third the neuromuscular junction. Since the neuromuscular junction relies on Ca^{2+} input a clear link between the three terms is

proposed. Additionally Ca^{2+} levels are linked to apoptosis and intracellular energy regulation. Therefore full functionality of related processes is crucial to maintain cells in a healthy state.

The case presented is an example how PD associated genes can help to identify specific disease affected pathways. Even though individual PD associated genes are not specific to a function, knowing their close interactors, found in the network communities, can point towards the affected pathways.

Furthermore and similar to Cluster 2, even though these terms were found based on (brain) synaptic gene communities, they might reflect the effects of PD on general cellular functions. The enrichment of neuromuscular activity can also point towards peripheral dysfunctions in the body of patients’.

6.3.5 Summary

Thanks to the curated core PD gene set and a combination of PPIN analysis, clustering algorithms, as well as functional (GO) enrichment this chapter could highlight key synaptic gene sets and functions affected by PD.

Being able to divide large datasets into connected subgroups is a key contribution to current research needs and developments. The ability to identify concrete gene groups as well as a core and extended gene set is a great step forward. This not only confirms and highlights most affected (synaptic) PD associated functions, but supplies beneficial information to a number of additional research questions. Hence new hypotheses can be established and further experimental studies can specifically target these newly proposed genes to gain more in-depth understanding.

Overall this confirms that the use of PPINs is a powerful tool to shed light on complex biological questions involving large datasets. Furthermore, results supply details for more efficient, in-depth and highly targeted follow-up research, specifically in the field of complex diseases.

6.4 Discussion

The use of PPINs is a growing area of research especially in the biological and medical field. Even though standardised procedures are not yet available this work illustrates their potential.

The large variety and possible uncertainty regarding data quality at several levels

decreases the predictive power of the results. Therefore, any additional data curation steps contribute to the quality. The use of proteomic data together with a curated list of PPIs allows for a best possible network representation of the data of interest. Even though this might not guarantee that all predicted interactions happen at the same time and/or are actually valid, the use of proteomic data highly rises the quality of the network providing information about proteins experimentally detected in the region studied.

The use of local network measures gives first insights to the PPIN structure and some of the most and least central proteins. Nevertheless, especially in large networks, these values can be influenced by properties of various unnaturally behaving nodes. Hence, best characterisation is achieved when analysing these statistical measures in a joint manner.

For more in-depth insight it is recommended to work with node or edge specific characteristics. These might be hard to analyse individually, but considering nodes and edges with extreme values can help to identify genes with key roles amongst the data. Using such a measure can also be misleading or biased since more studies focusing on a specific gene, e.g. due to its importance in disease, can lead to biased results and artefacts in the data. Again, the combination of different measures can help to add certainty to observed results. An alternative approach to analyse node and edge specific parameters is to consider the correlation between two values. For example, this can help to identify nodes with extreme values for two parameters giving more specific insight, which was illustrated in the case of the the connecting role of *DGLA*, supporting information flow without being in a top central network position (Section 6.3.1).

Generally it could be seen that the smaller presynaptic network seems to show slightly different properties compared to the larger postsynaptic, synaptosome and joint synaptic ones. This highlights the potential impact of size on network measure and a suggested need to normalise values before comparing them with each other. A similar phenomena can be observed considering the PPIN topology emerging after network clustering. Compared to the very node specific approaches outlined above, network clustering identifies highly connected gene sets. In recent years this has proven to be a profoundly beneficial tool. Especially addressing complex questions related to large datasets, such as the case presented, has been facilitated.

Clustering algorithms rely on distinct analytical principles leading to varying community compositions. Classifying the quality of a clustered network is an ongoing

challenge which might never be answered. One of the reasons is the subjective definition of a “good cluster” as well as its mathematical definition and the lack of a ground truth dataset to be used for comparison. Visualization is an additional challenge which can be addressed in different ways, either focusing on details or the overall pattern, but no ideal solution is available. Therefore statistical approaches are even more important.

With regard to identifying the most adequate clustering algorithm additional measures can be used. Community robustness for example describes the probability of specific nodes to belong to one or another community. Such information indicates how stable a network topology is, contributing to its detailed understanding, but hard to apply to 20 different network topologies. In addition it leads to the same question concerning network cluster cross-comparison. One possible approach to identify the most representative clustering of a network is the use of the cumulative distribution function of network consensus matrices. Information regarding the proportion of ambiguously clustered pairs can also help to identify the best fit. This approach is presented in the draft in preparation presenting the synaptic proteome (Section 5).

Overall, using a range of different clustering approaches complicates the thorough analysis of the quality of individual topologies and might not be the most economic choice. Nevertheless, the results presented showed that the approach was efficient given the addressed research question. The convergence of results emerging due to different clusterings, based on different mathematical concepts and different datasets, describing distinct synaptic regions, is promising. These findings support the credibility of obtained outcomes.

Therefore, the presented combinatorial approach leading to a set of core communities, gathering in three clusters could be a recommended choice for use in similar studies. The combination of different clustering algorithms can be considered as a varied and multi-angle approach towards interpreting one and the same dataset.

PPINs are static representations of PPIs assigning nodes to only one specific community. Given a dynamic cellular background this is most likely not the case. Furthermore networks only contain one representation of each protein which ignores the fact that many proteins are involved in distinct cellular functions. Such details can be included in a network model by assigning probabilities to nodes reflecting the likelihood to belong to a community. Robustness studies, as mentioned previously, can be used to calculate such values. Integrating these aspects could be a beneficial extension of the presented results, addressing networks individually.

Using network communities as cellular groups and applying enrichment tests is key

to identify disease enriched synaptic structures. The use of hypergeometric testing is a well established approach to do so. Multiple testing correction adds to the credibility of the obtained significance. Another aspect could be the application of permutation tests. These would also consider community robustness and increase the certainty of encountered results even more.

Currently available literature has not yet considered the comparison of communities emerging from differently clustered networks to investigate disease enrichment in specific cellular regions. Hence, the identification of three key sets containing a significant number of PD associated genes is a valuable achievement. Using a coverage threshold to identify a core gene set for each cluster is another way to focus on potential key genes with a strong link to PD. These steps allow to fine-tune future research questions based on the set of identified genes. Hence, the extended gene set is a valuable references to verify new research outputs.

It might be asked why none of the most traditional and well known PD associated genes appear in the disease enriched communities. One possible explanation could be that these are able to trigger the disease by themselves not leading to enrichment of the affected community. Such a behaviour might be due to their central role in a synaptic pathway. Nevertheless most often disease complexity emerges due to a combination of molecular dysfunctions affecting one pathway inducing functional alterations. This can lead on to the question why not all PD associated genes grouped and allocated in disease enriched communities. Technical and phenomenological points might explain this. Challenges associated to clustering algorithms, not always leading to the most representative communities, difficulties obtaining a concise set of disease associated genes and multiple testing approaches can influence the results. In regard to phenomenological reasons some of the disease genes can be rather consequential, meaning that they show a close link to the disease phenotype. In such cases the effects can be quite diverse not specifically accumulating in equal pathways and showing enrichment.

The interpretation of the dominating functions amongst genes in a network community using GO terms is of considerable interest. Available enrichment analysis tools are used and allow for specific adjustments to obtain best results. In many cases results confirm, that the joint functionality of genes in the set aligns with the one of the known PD associated gene. In addition and since individual gene functions are often very versatile, knowing direct interactors allows to specify concrete affected cellular functions. Results including cases such as rather unexpected disease associated pathways intuitively happening in other organs, such as the liver, can lead to reconsid-

eration of long-standing assumptions and lead to new hypotheses. Hence, readjusting the research focus based on insights obtained through network analysis is just another benefit of PPIN analysis.

In summary, approaches presented contribute largely to the ongoing development in the growing field of Systems Biology and integrated medical research. Combining these endeavours with further experimental and clinical studies can lead to breakthrough results in the coming years.

Chapter 7

Discussion

This chapter discusses critical steps, challenges and findings. Relevant ideas for future extensions of this work and more in-depth analysis are addressed. Finally key contributions to the research field are highlighted.

During the execution of the presented research a number of challenges were faced. Most of them could be solved through alternative methodologies or re-consideration of the underlying research question. Some of these steps contributed insights that are worth sharing for consideration in future studies.

7.1 Data Consistency

Describing a disease including as much detail as possible is a crucial step to understand its diverse, genetic origin and effect on individuals. Recent experimental and technological advances have allowed a wide range of gene-disease associations to become publicly available. A good example therefore is the Gene Expression Omnibus (GEO) database (Edgar et al., 2002) which gives access to a vast quantity of raw and published high-throughput data, covering gene expression microarray experiments from almost 20 000 published manuscripts (Barrett et al., 2012). Even though such a data repository sounds like a great source, in practise it can be notoriously difficult to access and re-analyse the stored data. Lack of standards covering the use of analytical programs, significance thresholds, data formats, and other complications mean that meta-analysis down-stream of the available data can be a challenge, proving very difficult at times. This phenomena leads to concerns about data-quality. This is specifically acute where if published results have not been reproduced.

Genome Wide Association Study (GWAS) studies have similar issues. However, in

this research area standards are more commonly accepted and followed, meaning that raw experimental data is more easily accessible. This makes data more comparable across publications and platforms, enhancing their credibility and allowing for downstream analysis.

When raw experimental data of interest is obtained, one of the fundamental requirements to identify similarities and differences between the datasets is a common, unique identifier. This is when annotations become very important and the role of “data mapping” becomes crucial. Genes and proteins have different, non interchangeable identifiers. Generally every protein is encoded by a single gene, meaning that the up-stream dependency of a protein to a gene can be clearly identified; often several proteins can be transcribed from a single gene, since splicing and a number of post-transcriptional modifications may induce additional variety. Depending on the experimental techniques used, post-transcriptional modifications may be missed, especially when working with large-scale approaches. It is therefore often adequate to consider gene identifiers as the unique reference identifier ID for all genes and proteins to avoid any potential bias. Such steps might limit detailed analysis, but guarantee consistency, reducing the number of false positive records, and allowing data consistency and usability in the future.

Using single identifiers facilitates further down-stream data analysis, as well as providing a single reference point. National Center for Biotechnology Information (NCBI) Entrez IDs were chosen for this purpose. These guarantee stable gene references which are widely used amongst the community (Maglott et al., 2010). Considering that not all data used in this study was retrieved from human samples, cross-species mapping was required as well. Publicly available homology information was used for this endeavour, which facilitated the use of all available data (without restrictions) from the original species (mouse, rat or human). This was particularly necessary when building the synaptic proteome datasets. Moreover the available human specific Protein-Protein Interaction (PPI) set was larger than that for mouse and rat. Since every mapping step shows a slight risk of inaccuracy, minimizing the required mappings was important.

In summary, these challenges in data consistency are currently addressed in ongoing efforts involving a wide range of researchers and research fields. The scientific community is working towards adapting many standardised formats for data and associated meta data in a guided way. This will help facilitate interoperability of datasets, reusability of data as well as its findability and accessibility. With a comprehensive im-

plementation of such standards, as proposed by the FAIR initiative (Wilkinson et al., 2016), many research questions could be answered more efficiently.

7.2 Proteomic Datasets

Proteomic data can provide detailed insight into the molecular constitution of a tissue region under investigation. As pointed out in Section 7.1 data quality is important to obtain best possible, most reliable, and consistent research results. While building the synaptic proteome datasets, Entrez IDs were used as unique identifiers for individual genes and proteins. This guarantees consistency and uniformity across datasets and best possible use of all available information, for example across studies, sources and species. In addition it facilitates data comparison with the set of Parkinson's Disease (PD) associated genes (Section 3.3.5).

Most difficulties related to the generation of proteomic datasets reside around experimental setups. Some of these are common challenges in the field of proteomics and include tissue extraction, homogenization and mass-spectrometry analysis. Therefore the joining together and comparing of various studies can help lead to increase in data quality and credibility. Analysis of the joint dataset can also highlight differences between available data, and point towards possible false positive records. Specifically due to the very large total number of proteins identified in the synaptic proteomes defining a core set was considered. However, using coverage as an approach to cut-off and identify a core proteomic dataset as described in Section 5.3.2, can lead to problems. For example genes detected with newer, more fine-grained technologies can be penalised or unwillingly excluded. Considering the first detection approach for filtering could help to prevent this.

However, a more generic approach can assist in the endeavour of filtering the full gene list reflecting detection consistency, and hence protein presence in the respective tissue region. The postsynaptic proteome which has been analysed in 23 published studies (Chapter 5) could be "cleaned" with such an approach. For the presynaptic proteome, further datasets are needed before being able to define a representative cut-off considering the year of first detection of a protein. The idea of a relative cut-off, depending on e.g. the year of first publication as proposed here has not been addressed in published literature so far, but could lead to more consistent results and a smaller, more representative synaptic proteome.

Although no additional filtering steps were applied to the generated proteomic

datasets for the purpose of this study, the large total numbers of proteins in the synaptic proteomes suggest that the total number of identified proteins exceeds the real one. The high number of proteins detected in only one of the synaptic studies proposes that these proteins might not actually be found in any of the synaptic regions. Therefore a combination of protein coverage and detection year together with other parameters should be considered to contributing towards identifying a core synaptosome providing even higher data quality.

Working with the large number of proteins in the proteomic datasets, ranging between 1,867 and 6,706 proteins in the presynaptic and joint synaptic proteome, required large-scale analytical techniques. The use of such is one way to gain insight into the data structure and role of PD in this region. Therefore Section 7.3 discusses the application of large-scale analytical techniques, in the context of Protein-Protein-Interaction Networks (PPINs).

7.3 Protein-Protein-Interaction Networks and PD

Datasets associated with a complex diseases, such as PD, tend to contain relatively large numbers of genes which are not obviously interacting or affecting similar biological pathways. These properties represent a challenging territory for in-depth analysis. Under such conditions the use of data-driven models and representations, machine learning techniques and tools, and statistical analysis and interpretation is of great value to gain new biological insights. This research focused on the use of PPINs as a promising tool to represent and analyse complex data encapsulating disease information, for example related to PD. Data-driven models, such as the PPINs proposed, can also assist in testing hypotheses emerging from experimental studies, usually focusing on single, specific disease mechanisms.

To ensure data-quality, PPIs used to generate the PPINs, were mined and internally assessed (Section 4.3.1). This lead to the best possible set of human, direct PPIs for the purpose of this work. Adding this extra effort to the workflow helped to reveal differences between the main PPI databases and how they obtain their interaction information. The steps taken whilst generating the PPI list (Section 4.3.2) helped to identify the final filter settings. For example the exclusion of non-direct interactions helped to ultimately obtain a high quality dataset.

In light of using PPI data in a human focused analysis, filtering only direct and human specific PPIs to build networks is highly recommended. The number of available

human PPIs seems relatively “complete” which might not be the case for other species. For example the mouse PPI set currently only contains ~15,500 specific PPIs, compared to ~200,000 human PPIs. When working with purely mouse specific data it is hence recommendable to make use of data obtained in other species, for a more comprehensive dataset. Nevertheless, not having to fall back on cross-species mapping, as when working with human-only data, increases quality of the network structure and allows for better predictions.

The use of species specific information can generally help to avoid bias in the PPINs. Additionally using proteomic datasets to build the networks imposes further quality standards, for example, by only considering those proteins found in datasets expressed in the tissue region of interest. By doing so however, a number of verified PPIs are excluded from the network, due to not having detected one or both of the interactors in the analysed proteome. The PPIN can therefore only represent a limited spatial and temporal overview. More advanced and fine-tuned experimental techniques are required to capture spatial and temporal changes in the proteome to obtain more fine-grained insights and construct specific networks.

The current procedure proposed during this research, using proteomic data and applying filtering criteria to the mined PPIs, shows the best quality control for network building and insightful analysis currently available. This procedure has led to identify key proteins in the networks, and in the context of PD has helped to confirm the most central disease genes.

Apart from that there was further interest uncovering the underlying complex structure of PD related genes in the synaptic PPINs. Such knowledge can help to identify disease subtypes, disease manifestation mechanisms, and help to make predictions regarding potential biomarkers. The use of machine learning techniques, such as community detection algorithms, was crucial in this aim. Choosing an appropriate clustering algorithm for a specific PPIN was no trivial task. This is specifically true when working with large datasets as addressed in this work (Section 6.3.2).

Another additional challenge is how to compare clustering results between algorithms, when no ground truth data exists for the PPINs. Clustering algorithms employ different machine learning techniques, metrics, parameters and values, when dealing with how to divide a network into communities. This diversity leads to largely varying community sizes amongst other network parameters. Hence trying to identify what constitutes the “best” clustering for a specific network plays an important role. Different research questions might require more coarse- or fine-grained groupings (commu-

nities) to advance understanding of the data, and potentially answer open questions.

The choice of which clustering algorithm to use needed to be taken. Due to the network size of up to 6,068 nodes and 69,520 edges, which is considered large for a biological system, the chosen algorithm had to be able to break down the structure into communities, in a reasonable computing time. Since there was no obvious best choice, a set of clustering algorithms also seemed a good way to address the question of “best” clustering. Based on results obtained none of them seemed to outperform the others. Hence, all algorithms and communities were considered with equal probability and taken forward for PD and functional enrichment analysis.

In summary, this part of the research showed the use of PPINs as being crucial for analysis of large proteomic datasets and disease. The difficulty of finding the most adequate algorithm was bypassed through the combination of several what specifically addressed the research question. It turns out that such an approach lead to results, regarding the question of interest, which were both highly consistent and promising. Hence this supports the idea that synaptic gene communities are associated to PD independently of the applied clustering strategy. In general, this result should be transferable to detecting communities specific to other diseases as well.

7.4 Systems Biology and PD Research

Studying complex diseases is a non-trivial task, especially given continuous growth of experimental data, such as GWAS and microarray expression results, and genetic information associated with the disease. Such complexity requires a systems approach, the use of PPINs for example, to gain an understanding of (i) biological processes active in the synapse and (ii) relationship between datasets.

Since complex diseases are caused by a combination of genes and their dysfunctions, these genes are often affecting similar pathways or cellular regions. Hence understanding which disease associated genes act together and identify their closest interactors is a promising approach to deepen knowledge about the diseases and guide future research endeavours.

Clustering PPINs to communities is one way to identify such gene communities. Based on the known PD associated genes in the communities, it is possible to identify gene sets showing a significant over-representation of PD, and hence reveal possibly new links between known and unknown PD associated genes.

Since communities shed light on the nearest neighbours of known PD associated

genes in disease affected synaptic regions, they can also help minimise the number of potential key PD associated genes required to alter and/or are involved in the alterations of disease related functions. Hence identifying subsets of genes associated to a specific disease, tissue, or function is of great benefit to understand the disease.

Furthermore the overall function of all genes in the enriched community can be studied. The use of Gene Ontology (GO) enrichment is a powerful approach here, even though the vast ontology structure makes it difficult to use. Using annotated gene-function information, as deposited in the GO database, can assist the identification of non-random functional similarities between genes in a set. Various techniques are available, and the focus on more specific functional terms, found in lower levels of the ontology tree, seems to be the most beneficial approach. Therefore the `elim` algorithm was chosen (Alexa et al., 2006).

The enrichment of highly specific GO terms could be found, confirming functionalities based on known PD associated genes. Additionally, functional enrichment results were able to specify further PD associated pathways, which have not yet been addressed in detail as being associated to PD.

As a further development to functional enrichment studies and to fine-tune functional enrichment results other concepts are being developed. One of them is the possible clipping of the GO tree to obtain a more targeted, for example neurology specific test environment (Geifman et al., 2010). Such advances can contribute interesting, mostly similar, results.

In summary, the potential of using PPINs to answer complex biological questions is promising, but its benefits are only just being discovered. Unravelling densely connected gene groups allows for further analytical steps such as gene-set specific enrichment studies. This combination of large-scale analytical approaches such as clustering, disease, and functional enrichment was then applied to study the set of PD associated genes (Section 3.3.5).

7.5 Synaptic Dysfunctions and PD

Apart from affecting multiple molecular functions, complex diseases affect cells at different levels. Joining available data was crucial to make better predictions, but was also a very challenging task. Varying quality controls and standards was the main issue in generating a final PD associated gene set.

This often meant excluding possibly useful information, which did not reach a

certain quality threshold. Nevertheless, a valuable core and extended gene disease association dataset could be generated, which showed that some of the later identified genes had been linked to PD previously. This could be confirmed by the detection of some genes associated to PD, found in experimental expression data, and amongst gene sets enriched for PD in network clusters. This also underlines the hypothesis that (i) distinct experimental approaches capture different facets of the disease and (ii) distinct cellular levels are affected in specific ways by the disease manifestation. This combination and flexibility reflects different cellular dysfunctions making up the disease genotype or forming part of its phenotype.

The final PD associated gene set was concise and robust enough to work with, allowing further knowledge to be gained. Therefore results provide a proof-of-concept, confirming known disease-gene links and finding new candidates. This is specifically true since this work put a large effort into identifying PD associated genes from different sources, and combining these in a comprehensive way.

One might think this replicates the work presented in the PD-map (Fujita et al., 2014). This in-depth analysis provides a great tool to visualize PD affected molecular functions and synaptic regions, but does not allow key sets of disease associated genes to be directly and computationally accessed. Information regarding direct or indirect links or associations are not available, making it almost impossible to directly use the supplied data in the context of this work. Therefore the manually curated core gene set was used.

The exercise of combining publicly available knowledge, describing PD gene associations, allowed to show significant overlap of data from distinct sources, confirming their PD specificity by highlighting their disease focus.

Once the core PD gene set was defined, further analysis using a systems biological approaches could be considered. Network analytical techniques were able to pin-point main affected PD associated pathways (Section 6.3.4). Even though initial clustering results varied based on the chosen clustering algorithm, comparative analysis and results were consistent and reassuring. This suggests the presented strategy could be used for other studies as well.

With respect to PD three main gene sets were highlighted as being highly influenced. A number of postsynaptic functionalities include receptor localization, receptor signalling pathways and neurotransmitter secretion (Sections 1.1.1.1 and 6.3.4.1). Furthermore, liver associated pathways were found associated to PD associated genes in the enriched Cluster 2 (Section 6.3.4.2). Such abnormalities in liver enzymes have

been linked to PD as early as 1991 (Tanner, 1991). More specifically fibrinogen was studied and findings confirmed a link between elevated fibrinogen levels and PD in men, older than 75 years of age. These pathways point towards an inflammatory reaction linked to the disease and could also be classified as an autoimmune response. Recent research has also proposed the liver drug UDCA to treat PD patients showing first protective effects on nerve cells in a genetic mouse model (Mortiboys et al., 2015).

Having verified not just the affected processes, but also a concrete set of genes linked to known PD associated ones, allows further gene targets to be researched. The identification of known disease associated functions is a proof-of-principle for the presented idea. Nevertheless the real advance lies in the generation of concrete, relatively concise gene sets. Proposed disease target genes, highly linked to known PD associated ones are a substantial source for future advances. Follow-up studies can now analyse the role of these genes from various perspectives advancing knowledge in their detailed functionality and impact on molecular dysfunctions linked to PD. These can consider the genes to be (i) so far unsuspected disease triggers, (ii) potential drug targets, due to their close link and influence on disease associated genes or (iii) biomarkers, allowing better diagnosis.

7.6 Future Research Perspectives

Every study is limited by time and resources and often interesting follow-up questions arise at more advanced stages of a project. In the study presented, the developed concept could be used to answer other similar research questions.

This thesis covers a range of aspects which come together to shed light over complex diseases given an affected tissue. This “pipeline” allowed me to identify core disease associated gene sets as well as their dysfunctions, helping to uncover unknown or unsuspected links between diseases and pathways as well as cross-pathway interaction. The use of a range of network parameters combined with curated datasets guaranteed best possible quality of analysis and results.

Given the flexibility of the pipeline one might wonder why the analysis was not carried out for a second reference tissue and/or another neurodegenerative disease. Results for either of the scenarios could confirm the specificity of the results to PD and highlight more general neurological effects. The biggest challenge was the remarkable importance of data quality, considering the set of disease associated genes as well as the proteome of interest. No equally high standard datasets could be easily obtained,

but future studies should take these proposals into account.

Apart from questions related to complex diseases the approach is able to cover other aspects. For example, protein datasets, containing proteomic data from different tissues could be used to associate e.g. disease to a specific organ. More phenomenological trait information, such as obesity markers, retrieved from the Human Phenotype Ontology (Köhler et al., 2017), amongst others, can give insights into underlying molecular mechanisms.

Due to current data availability, presented results describe “an average synapse”, not considering subtypes or alike. PD mostly affects dopaminergic neurons in the substantia nigra pars compacta. Building a PPIN of proteins expressed in specifically these neurons can enhance insights and results would be more specific. For example the Allen Brain Atlas (Hawrylycz et al., 2012, 2014)¹ supplies detailed insights for human brain tissues. Nevertheless the data are based on a very small sample size of only three individuals (at time of study). The Human Protein Atlas (Uhlén et al., 2015)² might be another data source. Brain region specific tissues are not yet available in this repository but cell type specific data is increasing. A major drawback in both cases is the supplied data type. Information is based on gene-expression data, which based on the central dogma of molecular biology (Figure 1.1), does not allow concrete prediction of the proteome since it disregards the transcription step. Hence available data does not reflect standards of a proteomic dataset and can influence the network topology and ultimately obtained results. Upcoming technologies are addressing these challenges which might be overcome in the next decade. Therefore, data from the before mentioned sources supply good starting points to explore the benefits gained through more specific datasets.

Two additional aspects need consideration. Variation in the proteome based on developmental stages (temporal aspect) and spatial restrictions. The proteome changes over the course of cell and tissue development and possibly under disease conditions. Furthermore spatial intracellular division likely prevents a number of PPIs from occurring. This can be explained by cellular compartments such as the nuclear or endoplasmic reticulum membrane acting as physical barriers. Apart from gaining more specific insight based on more precise data the simulation and comparison of several proteomes can illustrate the development of for example disease effects on a system such as the synapse. The field of multi-scale modelling (Hirakis et al., 2015) is cur-

¹<http://www.brain-map.org/>

²<http://www.proteinatlas.org/>

rently exploring advances in this field and has shown first promising results.

Furthermore PPINs are static representations of protein interactions. Protein abundance, interaction probabilities and the fact that some proteins are more likely to belong to two communities than to one can not be reflected. Nevertheless the possibility to divide networks and identify specifically disease enriched gene sets is of substantial value. This makes PPINs a very powerful, currently available, tool in the area of systems biology allowing to formulate new, more specific research hypotheses.

To address some remaining challenges two side projects were carried out alongside this PhD research and illustrate additional areas of advancement. PPINs allowed to identify a potential new key set of disease associated genes and functional associations to consider more closely in future studies (Section 6.3.4 and Appendix E). Hence apart from focusing on a regional, proteome level, the impact of PD on the synapse was addressed from a more low-level perspective. To consider the implications of the disease on a whole system more perspective needed to be gained. The impact of PD on neuronal information transmission requires such a higher level systems understanding, modelling the behaviour of the entire synapse or neuron. Such a model also allows better analysis of drug effects and might allow the detection for disease biomarkers since it considers an entire system.

Both, the lower and higher level angle were addressed in shared side projects, and are introduced in the next sections.

7.6.1 Clathrin Mediated Endocytosis - a Dynamic Model

As Section 5.3.6 shows, one of the PD associated subsystems is the “clathrin coat assembly” cycle. It relates to the larger system of Clathrin Mediated Endocytosis (CME), which involves around 30 proteins undergoing about 60 internal PPIs (McMahon and Boucrot, 2011). Eight of the 30 proteins have been previously linked to PD. These are the proteins and genes *actin*, *auxilin*, *cortactin*, *endophilin*, *EPS15-EPS15R*, *GAK*, *HIP1R*, and *HSC70* which are part of the pathway, as well as *NSF* and *LRRK2* showing a link to the clathrin light chain as well as cytoskeletal signal transmission respectively.

To better understand the system and analyse implications of individual disease associated genes, a dynamic CME model was generated. This work was carried out together with Oksana Sorokina, Anatoly Sorokin and Douglas Armstrong. One of the elaborated models (which was implemented by myself; “model 2”) contains individual clathrin molecules and describes their interaction leading to the formation of clathrin

coated vesicles in a detailed way. A rule-based modelling approach, specifically kappa modelling (Danos et al., 2008)³ was used to simulate the system in detail.

Even though a large amount of research has previously focused on the CME system from an experimental as well as computational point of view, the rule-based model allows for relatively easy extension. Hence, a more detailed view of a PD affected process was generated. This is a first step towards simulating the effect of dysfunctions on the system which can for example predict the impact of PD on CME or vice versa. Some of the aforementioned genes have already been added to preliminary models and need to be further explored.

A joint draft is currently in preparation, proving the concept of using a rule-based modelling approach to simulate the dynamic CME system. The current version can be found in Appendix G.

For a broad understanding of disease mechanisms, not only more in-depth insights are required. Broader system understanding is necessary to grasp the effect of changes in individual proteins on a whole system such as a whole neuron and its role in information transmission.

7.6.2 Disease in Computational Models of Neurons

A large number of computational models of neurons are available. Most of these focus on slightly different research questions which contain topics around information transmission between neurons. Due to the lack of standardised modelling languages and common nomenclature it is very challenging to systematically analyse and compare models and identify their key components.

In a joint effort, together with Emilia Wysocka and David C. Sterratt, both University of Edinburgh, a number of synaptic models were analysed and a key set of modelled synaptic genes could be identified. A publication is in preparation and the current version can be found in Appendix H.

The annotation of synaptic models and extraction of entities appearing in these models was a non-trivial task. Once a set of modelled genes was identified it was compared to a list of genes commonly found in seven neurodegenerative diseases. The study showed that a large number of disease associated genes are not modelled in the analysed computational models of neurons. The availability of the list of modelled genes together with a set of disease associated genes can now help to identify most

³<http://dev.executableknowledge.org/>

suitable available models to extend and gain more disease related insights.

Thus this study showed that considering higher system-level models, by first identifying their general focus and modelled elements, can help to find the best fit pre-existing model to be extended to answer specific disease related questions. Studying the effect of drugs and identifying disease biomarkers is highly facilitated by such a system and should be considered for future analysis.

7.7 Conclusion

To gain better understanding of a complex disease, which shows its main effect at the synapse, a highly regulated brain region, proved challenging. Nevertheless the curation of high quality datasets, systems biological approaches and in-depth statistical and functional analysis allowed me to obtain new insights into the disease.

The use of PPINs as a mathematical representation of complex diseases helped me to identify PD associated gene sets, and their closely connected interactors, showing a novelty which has not yet been addressed in PD research. The use of PPINs added detailed, functional insight and moved from the analysis of single disease genes to gene sets. This helps to show the joint impact of several genes on disease development and manifestation.

Network analysis cannot hope to solve all research problems in their entirety, but can contribute to speeding up the process of finding answers to complex questions; specifically by advancing and better defining future research questions and directions. Hence this study is a proof-of-concept, highlighting the need of large-scale techniques, to address detailed research questions related to complex disease.

The static nature of interaction networks did not prove a major obstacle, supported by the fact that numerous slightly different datasets as well as clustering algorithms lead to very similar results. It might be argued that the combination of datasets and clustering algorithms was needless since similar results were obtained (Section 6.3.3). Nevertheless it is the combination and mutual confirmation of results which strengthens the outcomes and make them more trustworthy.

Considering PD specifically three new functionally defined gene sets were identified (Section 6.3.3). These contain known PD associated genes and a number of reliably linked others. The core datasets are available to be considered in future research into underlying disease causes, disease subtypes, biomarkers, and drug targets. Identifying known functional GO terms related to spatial organisation, involving scaffolding pro-

teins and vesicle cycling confirms the potential of the enrichment approach (Section 6.3.4.1). New hints towards immune response related functions should drive research to emphasize on these fields in upcoming PD related studies (Section 6.3.4.2).

Overall, this work presents a combination of systems biological approaches, including the use of PPINs and functional enrichment studies. It was possible to identify new potential gene sets and their overall function, playing key roles in PD. Such findings are very beneficial in the light of further understanding the disease complexity.

In summary, existing techniques were used to confirm known and unravel unknown details regarding PD. Results obtained are very promising and further development of presented ideas and findings can lead to major contributions in theoretical, experimental and clinical PD research. Available knowledge was enriched and extended and future analysis of many other open challenges, not only related to PD, but other complex diseases can be made more targeted and efficient.

Appendix A

Literature based Parkinson's Disease associated genes

Table A.1: Genes manually identified to be linked to PD in reviewed papers (ordered alphabetically by Gene Name short). PMCID shows the reference where the gene-disease association was identified.

Entrez ID	Gene Name short	PMCID
130013	<i>ACMSD</i>	22438815, 21812969
135	<i>ADORA2A</i>	24032478
351	<i>APP</i>	22438815
23400	<i>ATP13A2</i>	23380027, 2650009
683	<i>BST1</i>	22438815, 21812969, 22786590
776	<i>CACNA1D</i>	23771339
793	<i>CALB1</i>	23771339
801	<i>CALM1</i>	23771339
811	<i>CALR</i>	23771339
84660	<i>CCDC62</i>	22438815, 21812969
6352	<i>CCL5</i>	21048992
1006	<i>CDH8</i>	21812969
1609	<i>DGKQ</i>	22438815
1981	<i>EIF4G1</i>	20495568
26281	<i>FGF20</i>	20495568
2580	<i>GAK</i>	22438815, 21812969, 22786590
2629	<i>GBA</i>	23380027, 20495568, 22438815, 21812969
26058	<i>GIGYF2</i>	20495568
2861	<i>GPR37</i>	23251443
9026	<i>HIP1R</i>	22438815, 22786590
3308	<i>HSPA4</i>	23380027
27429	<i>HTRA2</i>	20495568
3383	<i>ICAM1</i>	18044695
8516	<i>ITGA8</i>	22438815

3920	<i>LAMP2</i>	23380027
27074	<i>LAMP3</i>	22438815, 21812969, 22786590
120892	<i>LRRK2</i>	3035023, 23380027, 20495568, 22438815, 21812969, 22786590
4137	<i>MAPT</i>	22438815, 22806825
4163	<i>MCC</i>	22438815
56922	<i>MCCC1</i>	22438815, 21812969
4843	<i>NOS2</i>	23744073
4929	<i>NR4A2</i>	24126627
4905	<i>NSF</i>	21812969
80025	<i>PANK2</i>	22806825
5071	<i>PARK2</i>	3035023, 23380027, 20495568, 22581678, 21812969
11315	<i>PARK7</i>	3035023, 23418303, 23380027, 20495568, 22581678, 21812969
677662	<i>PARK12</i>	17068789
100359403	<i>PARK16</i>	22438815, 21812969
65018	<i>PINK1</i>	3035023, 23380027, 20495568, 22581678, 21812969
10891	<i>PPARGCIA</i>	23380027
57111	<i>RAB25</i>	22438815
153328	<i>SLC25A48</i>	21812969
28232	<i>SLCO3A1</i>	21812969
6622	<i>SNCA</i>	3035023, 23380027, 20495568, 22438815, 21812969, 22786590, 21412835
9627	<i>SNCAIP</i>	23127794
27347	<i>STK39</i>	22438815, 21812969, 22786590
23208	<i>SYT11</i>	22438815, 21812969, 22786590
81615	<i>TMEM163</i>	22438815
7345	<i>UCHL1</i>	23380027
10497	<i>UNC13B</i>	21812969
55737	<i>VPS35</i>	2426182, 22806825
7473	<i>WNT3</i>	21812969

Appendix B

MI-IDs

Table B.1: MI IDs specifying all direct interaction types, used to filter PPIs (ordered alphabetically based on the description).

Interaction MI-ID	Description
0192	acetylation reaction
0557	adp ribosylation reaction
0193	amidation reaction
1143	aminoacylation reaction
1148	ampylation reaction
0914	association
0882	atpase reaction
1139	carboxylation reaction
0194	cleavage reaction
0195	covalent binding
0197	deacetylation reaction
1310	de-ADP-ribosylation reaction
0985	deamination reaction
1140	decarboxylation reaction
0198	defarnesylation reaction
0199	deformylation reaction
0200	degeranylation reaction
0558	deglycosylation reaction
0871	demethylation reaction
0201	demyristoylation reaction
0569	deneddylation reaction
0202	depalmitoylation reaction
0203	dephosphorylation reaction
0568	desumoylation reaction
0204	deubiquitination reaction
1027	diphtamidation reaction
0407	direct interaction
0408	disulfide bond
0572	dna cleavage

0701	dna strand elongation
0414	enzymatic reaction
0206	farnesylation reaction
0207	formylation reaction
0209	geranylgeranylation reaction
0559	glycosylation reaction
0883	gtpase reaction
0210	hydroxylation reaction
1250	isomerase reaction
0211	lipid addition
0212	lipid cleavage reaction
0213	methylation reaction
1251	methylmalonyl-CoA isomerase reaction
0571	mrna cleavage
0214	myristoylation reaction
0567	neddylation reaction
0910	nucleic acid cleavage
0986	nucleic acid strand elongation reaction
0881	nucleoside triphosphatase reaction
0945	oxidoreductase activity electron transfer reaction
0216	palmitoylation reaction
1146	phospholipase reaction
0971	phosphopantetheinylation
0217	phosphorylation reaction
0844	phosphotransfer reaction
0915	physical association
1237	proline isomerization reaction
0570	protein cleavage
1127	putative self interaction
0902	rna cleavage
0987	rna strand elongation
1126	self interaction
1327	sulfurtransfer reaction
0566	sumoylation reaction
0556	transglutamination reaction
0220	ubiquitination reaction
1230	uridylation reaction
0218	obsolete: physical interaction

Table B.2: MI IDs specifying source databases of PPIs (orderd alphabetically based on description).

Database MI-ID	Description
2166	ai
0575	alliance for cellular signaling
2165	bar
1332	bhf-ucl
0462	bind
1123	bindingdb
1108	biocarta
1105	biocyc
0463	biogrid
0967	chembl
1063	consensuspathdb
0464	cygd
0465	dip
0466	ecocyc
0936	emdb
1331	evidence ontology
1116	genemania
0448	gene ontology
2017	heterogen
1335	hpidb
0468	hprd
0670	imex
0974	innatedb
0469	intact
0585	intenz
0461	interaction database
0923	irefindex
1262	i2d
0470	kegg
2012	kegg compound
0917	matrixdb
1222	mbinfo
2164	mind
0471	mint
0459	mmdb
1263	molecular connections
0903	mpidb
1264	ntnu
1124	pathwaycommons
1106	pathways database
0472	pdbe
0806	pdbj
1107	pid
0467	reactome

0460	rcsb pdb
1115	spike
1014	string
1117	topfind
0486	uniprot knowledge base
1098	uniprot/swiss-prot
1099	uniprot/trembl
1114	virhostnet
0805	wwpdb

Appendix C

Extended Overview of Synaptic Proteomic Studies

Table C.1: Presynaptic proteome publications and respective datasets. “Count” shows the number of proteins, mapped to human Entrez IDs found in the study.

Study	Year	Reference	Short Description	Species	Method	Description	Count
MORCIANO	2005	Morciano et al. (2005)	synaptic vesicle	rat	Co-IP with SV2, MALDI-TOF-MS and 2D BAC/SDS-PAGE	Synaptic vesicle proteins from nerve terminal proteome	85
BURRE	2006	Burré et al. (2006)	synaptic vesicle	rat	1-D SDS-PAGE & nano-LC ESI-MS/MS, or 2-D SDS-PAGE & (BAC)/SDS-PAGE, or SDS (dSDS)-PAGE & MALDI-TOF-MS	Synaptic vesicle proteins	157
MORCIANO	2009	Morciano et al. (2009)	synaptic vesicle	rat	IP, MALDI-TOF-MS, nanoLC ESI MS/MS and 2D BAC/SDS-PAGE	Using a monoclonal antibody against synaptic vesicle protein 2 we immunopurified a presynaptic compartment containing the active zone with synaptic vesicles docked to the presynaptic plasma membrane as well as elements of the presynaptic cytomatrix	308
GORINI	2010	Gorini et al. (2010)	presynaptic	mouse	CO-IP, MALDI-TOF-MS and MASCOT	Presynaptic vesicle recycling: We used co-immunoprecipitation followed by mass spectrometry or western blotting to investigate the synaptic protein network for the candidate proteins BKCa, dynamin-1, SNAP-25, syntaxin-1A, and VAMP-2	49
GRONBORG	2010	Grønberg et al. (2010)	synaptic vesicle	rat	ICAT SCX, iTRAQ	Synaptic vesicle proteome (glutamatergic and GABA synapses)	613
BOYKEN	2013	Boyken et al. (2013)	presynaptic	rat	LS/MS-MS, iTRAQ	Synaptic vesicle docking and endocytosis	414
WILHELM	2014	Wilhelm et al. (2014)	synaptic vesicle	rat	quantitative IP and LS/MS-MS	Synaptic vesicle cycle proteins from synaptic buttons, synaptosome compartmentalised	1158
BRINK-MALM	2014	Brinkmalm et al. (2014)	presynaptic	mouse	IP with SNAP 25 and LC-MS/MS	Presynaptic (SNARE complex)in AD	68

Table C.1: Presynaptic proteome publications and respective datasets. "Count" shows the number of proteins, mapped to human Entrez IDs found in the study.

Study	Year	Reference	Short Description	Species	Method	Description	Count
WEINGARTEN	2014	Weingarten et al. (2014)	presynaptic	mouse	CoIP with SV2, SDS-PAGE and LC-MS/MS	Presynaptic active zone	467

Table C.2: Postsynaptic proteome publications and respective datasets. "Count" shows the number of proteins, mapped to human Entrez IDs found in the study.

Study	Year	Reference	Short Description	Species	Method	Description	Count
WALIKONIS	2000	Walikonis et al. (2000)	postsynaptic	rat	MALDI-TOF-MS and SDS-PAGE	Forebrain PSD	29
PENG	2004	Peng et al. (2004)	postsynaptic	rat	liquid chromatography and LC-MS/MS	Forebrain PSD	325
SATOH	2002	Satoh et al. (2002)	postsynaptic	mouse	2D LC/MS/MS	Forebrain PSD	45
YOSHIMURA	2004	Yoshimura et al. (2004)	postsynaptic	rat	2D LC/MS/MS	Forebrain PSD	435
FARR	2004	Farr et al. (2004)	MGLUR5	rat	co IP	mGluR5 interacting complex (co IP)	71
JORDAN	2004	Jordan et al. (2004)	postsynaptic	mouse and rat	nanoflow HPLC and LC-MS/MS	Brain PSD	390
LI	2004	wan Li et al. (2003)	postsynaptic	rat	2D gel/ICAT and matrix-assisted laser desorption ionization-time of flight, LC-MS/MS	Brain PSD	137
TRINIDAD	2005	Trinidad et al. (2005)	postsynaptic	mouse	Nano-LC-ESI-QTOF MS/MS	Forebrain PSD	234
CHENG	2006	Cheng et al. (2006)	postsynaptic	rat	LC-MS/MS	PSD identifications found in both cerebellum and forebrain	288
COLLINS	2006	Collins et al. (2006)	postsynaptic, MASC/NR2B, MASC/NR1, NRC-MASC, AMPA, postsynaptic-consensus	mouse	1D SDS-PAGE and LC-MS/MS, immunoprecipitation with antibody against NR2B, NR1 or against Gria2 (AMPA), SDS-PAGE, LC-MS/MS SDS-PAGE, LC-MS/MS	G2C PSD dataset, immunopurification for NR2B, NR1 or GRIA2 (AMPA), NRC/MASC complex = total from NR1 + NR2B above; Consensus PSD calculated by Collins et al, more than 2 mentionings in 6 published proteomic studies and 119 individual papers	717
DOSEMESI	2006	Dosemeci et al. (2006)	postsynaptic	rat	2D LC/MS/MS	PSD from hippocampus	113
DOSEMESI	2007	Dosemeci et al. (2007)	postsynaptic	rat	LC-MS/MS; LC-MS/MS, IP with PSD95	Cortex PSD, PSD95 protein complex	548
TRINIDAD	2008	Trinidad et al. (2008)	postsynaptic	mouse	Nano-LC-ESI-QTOF MS/MS	PSD relative quantification of expression and phosphorylation status from: cortex, hippocampus, midbrain, cerebellum	2150
SELIMI	2009	Selimi et al. (2009)	postsynaptic	mouse	1-D SDS-PAGE, matrix-assisted laser desorption/ionization (MALDI) quadrupole/time-of flight (QqTOF) MS and MALDI ion trap (MALDI-IT) tandem MS (MS/MS)	PSD from parallel fiber/purkinje cell synapse	61
FERNANDEZ	2009	Fernández et al. (2009)	TAP-PSD-95-CORE, TAP-PSD-95-FULL	mouse	TAP tag, LC-MS/MS	TAP-PSD-95 pull-down core list and full list	292
BAYES	2010	Bayés et al. (2011)	consensus postsynapse, full postsynapse	human	LC-MS/MS	Human neocortex (hPSD) biopsy PSD consensus and full list	1441
BAYES	2012	Bayés et al. (2012)	consensus postsynapse, full postsynapse	mouse	LC-MS/MS	Cortex PSD consensus	1545

Table C.2: Postsynaptic proteome publications and respective datasets. “Count” shows the number of proteins, mapped to human Entrez IDs found in the study.

Study	Year	Reference	Short Description	Species	Method	Description	Count
SCHWENK	2012	Schwenk et al. (2012)	AMPA	unknown	multiepitope and target knockout-controlled affinity purifications (ME-APs), blue native polyacrylamide gel electrophoresis (BN-PAGE) and nano-LC MS/MS	AMPA receptor complex	34
DISTLER PSD1	2014	Distler et al. (2014)	postsynaptic I	mouse	LS/MS-MS with iTRAQUDMSE, ISOQuant	mouse hippocampus PSD	3545
DISTLER PSD2	2014	Distler et al. (2014)	postsynaptic II	mouse	LS/MS-MS with iTRAQUDMSE, ISOQuant	mouse hippocampus PSD	2092
BAYES	2014	Bayés et al. (2014)	postsynaptic, MASC	human	SDS-PAGE and nanoLC-MS/MS	PSD (post-mortem neocortex samples and biopsy tissue), MASC (post mortem neocortex samples and biopsy tissue)	1141
UEZU	2016	Uezu et al. (2016)	postsynaptic, PSD95 (ePSD), iPSD, iPSD	mouse	in vivo affinity purification approach BioID (iBioID) + streptavidin-based affinity purification and mass spectrometry (MS), DLG4-BirA, collybistin(Arhgef9)n-BirA and InSyn1-(BirA) & gephyrin-BirA, InSyn1-coIP, LS-MS	pilot iPSD: BirA, PSD95-BirA, gephyrin- BirA, ePSD, iPSD, frontal cortex and hippocampus of C57BL/6mice, InSyn1 pulldown	1111
FOCKING	2016	Föcking et al. (2016)	postsynaptic	human	Label free LC-MS	supragenual (BA24) - anterior cingulate cortex (ACC)	2026

Table C.3: Synaptosome proteome datasets and respective publications. “Count” shows the number of proteins, mapped to human Entrez IDs found in the study.

Study	Year	Reference	Short Description	Species	Method	Description	Count
FILIOU	2010	Filiou et al. (2010)	synaptosome	mouse	Isoelectric focusing (IEF) and MS	Synaptosome and phosphoproteome	2778
DAHLHAUS	2011	Dahlhaus et al. (2011)	synaptosome	mouse	MALDI-MS and iTRAQ	Synaptic proteome from mouse visual cortex	638
ZIV synapse	2013	Cohen et al. (2013)	synaptosome	rat	Stable Isotope Labeling with Amino acids in Cell culture (SILAC), mass spectrometry (MS), Fluorescent Non-Canonical Amino acid Tagging (FUNCAT)	Synaptic protein turnover rates	185
ZIV full	2013	Cohen et al. (2013)	synaptosome	rat	Stable Isotope Labeling with Amino acids in Cell culture (SILAC), mass spectrometry (MS), Fluorescent Non-Canonical Amino acid Tagging (FUNCAT)	Synaptic protein turnover rates	2447
BIESEMANN	2014	Biesemann et al. (2014)	synaptosome	mouse	Fluorescence Activated Synaptosome Sorting (FASS), LS/MS-MS, Mascot	VGLUT1/VENUS knock-in mice, glutamatergic synaptosomes	157
CHANG	2015	Chang et al. (2015)	synaptosome	human	SCX fractionation with SWATH analysis	Synaptic proteome from hippocampus and motor cortex in autopsy brain for Alzheimer's disease and control	2076
DISTLER	2014	Distler et al. (2014)	TOTAL	mouse	LS/MS-MS with iTRAQUDMSE, ISOQuant	mouse hippocampus PSD	4475

Appendix D

Additional Protein-Protein-Interaction Networks

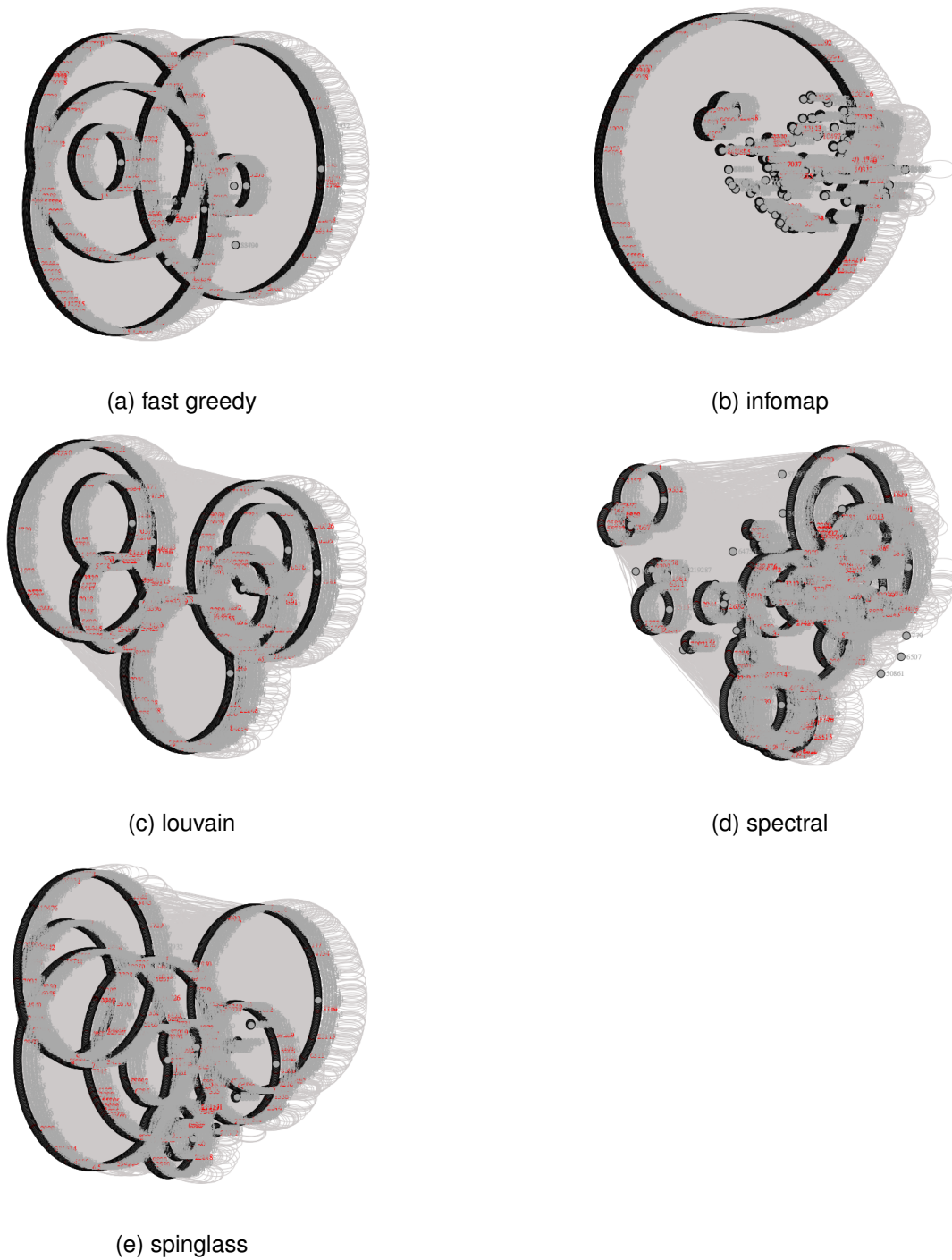


Figure D.1: Postsynaptic PPINs. Different clustering algorithm results are highlighted. Red coloured nodes represent PD associated genes. Grey “background” represents network edges.

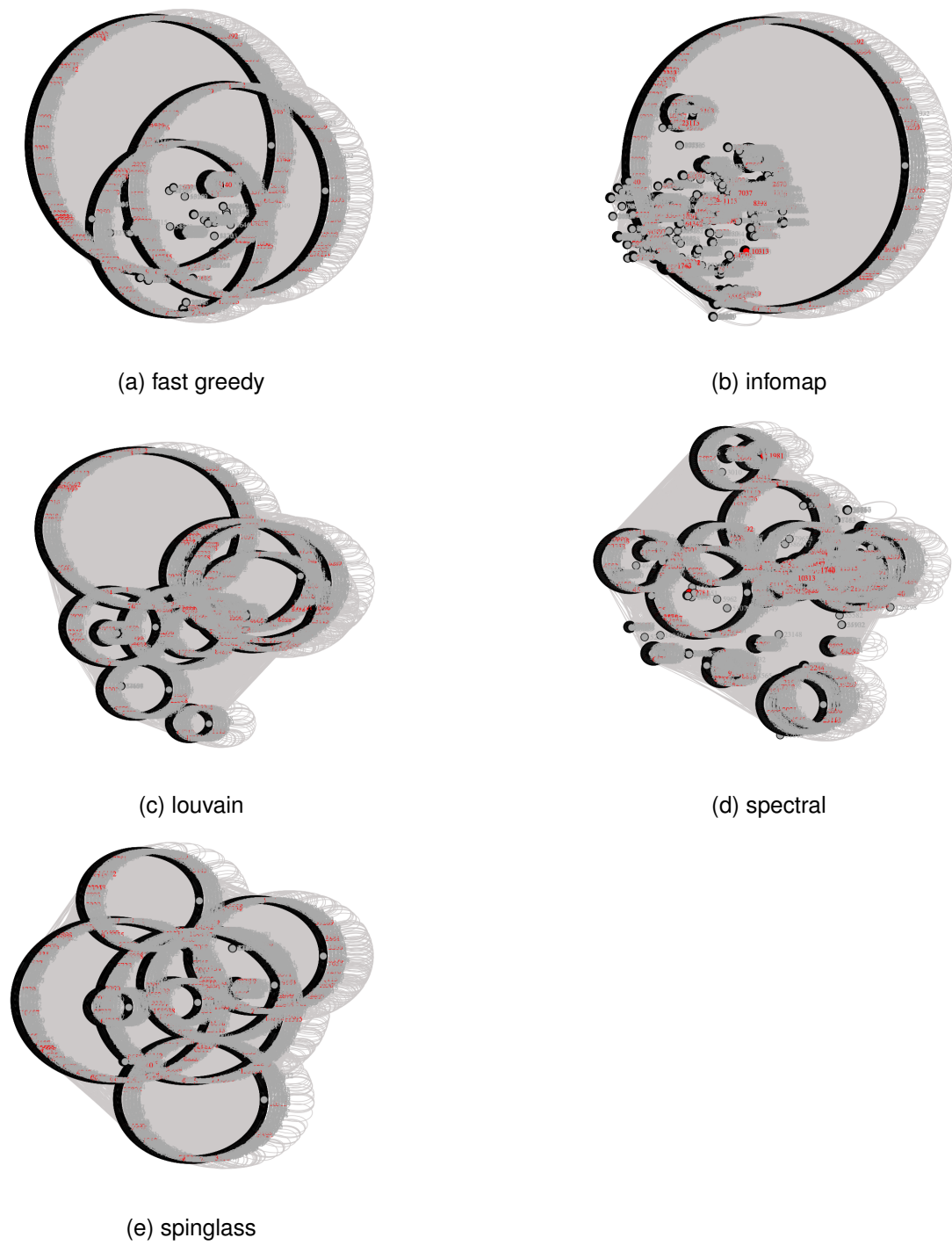


Figure D.2: Synaptosome PPINs. Different clustering algorithm results are highlighted. Red coloured nodes represent PD associated genes. Grey “background” represents network edges.

Appendix E

Core PD associated gene sets

Cluster 1 (sorted by Entrez ID)

19 (*ABCA1*), 88 (*ACTN2*), 89 (*ACTN3*), 320 (*APBA1*), 491 (*ATP2B2*), 493 (*ATP2B4*), 575 (*ADGRB1*), 796 (*CALCA*), 1501 (*CTNND2*), 1739 (*DLG1*), 1740 (*DLG2*), 1741 (*DLG3*), 1742 (*DLG4*), 1756 (*DMD*), 1837 (*DTNA*), 2257 (*FGF12*), 2902 (*GRIN1*), 2903 (*GRIN2A*), 2904 (*GRIN2B*), 2905 (*GRIN2C*), 2906 (*GRIN2D*), 2977 (*GUCY1A2*), 2983 (*GUCY1B3*), 3736 (*KCNA1*), 3738 (*KCNA3*), 3739 (*KCNA4*), 3761 (*KCNJ4*), 4130 (*MAP1A*), 4355 (*MPP2*), 4842 (*NOS1*), 5332 (*PLCB4*), 6323 (*SCN1A*), 6329 (*SCN4A*), 6331 (*SCN5A*), 6640 (*SNTA1*), 6641 (*SNTB1*), 6645 (*SNTB2*), 7402 (*UTRN*), 8502 (*PKP4*), 8525 (*DGKZ*), 8573 (*CASK*), 8777 (*MPDZ*), 8825 (*LIN7A*), 8938 (*BAIAP3*), 9211 (*LGII*), 9223 (*MAGII*), 9369 (*NRXN3*), 9378 (*NRXN1*), 9379 (*NRXN2*), 9615 (*GDA*), 9754 (*STARD8*), 9973 (*CCS*), 10125 (*RASGRPI*), 10203 (*CALCRL*), 10207 (*PATJ*), 10276 (*NET1*), 11336 (*EXOC3*), 22871 (*NLGN1*), 23037 (*PDZD2*), 23109 (*DDN*), 23209 (*MLC1*), 23237 (*ARC*), 23265 (*EXOC7*), 23513 (*SCRIB*), 25960 (*ADGRA2*), 26154 (*ABCA12*), 29919 (*CI8orf8*), 53616 (*ADAM22*), 53919 (*SLCO1C1*), 54413 (*NLGN3*), 55083 (*KIF26B*), 55327 (*LIN7C*), 55914 (*ERBIN*), 57502 (*NLGN4X*), 57524 (*CASKIN1*), 57554 (*LRRC7*), 57555 (*NLGN2*), 57575 (*PCDH10*), 60412 (*EXOC4*), 64130 (*LIN7B*), 64398 (*MPP5*), 84435 (*ADGRA1*), 84448 (*ABLIM2*), 85445 (*CNTNAP4*), 140735 (*DYNLL2*), 143098 (*MPP7*), 148753 (*FAM163A*), 166647 (*ADGRA3*), 221749 (*PXDC1*), 642968 (*FAM163B*)

Cluster 2 (sorted by Entrez ID)

213 (*ALB*), 216 (*ALDH1A1*), 217 (*ALDH2*), 272 (*AMPD3*), 767 (*CA8*), 335 (*APOA1*), 337 (*APOA4*), 345 (*APOC3*), 714 (*CIQC*), 1356 (*CP*), 1471 (*CST3*), 1608 (*DGKG*), 2243 (*FGA*), 2244 (*FGB*), 2266 (*FGG*), 3500 (*IGHG1*), 3502 (*IGHG3*), 3503 (*IGHG4*),

3514 (*IGKC*), 3778 (*KCNMA1*), 4018 (*LPA*), 4329 (*ALDH6A1*), 4635 (*MYL4*), 5136 (*PDE1A*), 5142 (*PDE4B*), 7018 (*TF*), 7579 (*ZSCAN20*), 8854 (*ALDH1A2*), 8911 (*CACNA1I*), 8914 (*TIMELESS*), 23036 (*ZNF292*), 23460 (*ABCA6*), 26049 (*FAM169A*), 57165 (*GJC2*), 57722 (*IGDCC4*), 79183 (*TTPAL*), 79925 (*SPEF2*), 128240 (*NAXE*), 140460 (*ASB7*), 154664 (*ABCA13*), 255189 (*PLA2G4F*), 284161 (*GDPD1*), 353274 (*ZNF445*)

Cluster 3 (sorted by Entrez ID)

775 (*CACNA1C*), 782 (*CACNB1*), 783 (*CACNB2*), 784 (*CACNB3*), 9478 (*CABP1*), 57019 (*CIAPIN1*)

Appendix F

Enriched Gene Ontology terms in the top three PD enriched clusters

Table F.1: GO terms enriched in at least two communities in Cluster 1 (alphabetical order of GO terms); significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, `elim` algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values available upon request since different in distinct enriched clusters).

GO term	GO term ID	GO term definition
Biological Process		
GDP metabolic process	GO:0046710	The chemical reactions and pathways involving GDP, guanosine 5'-diphosphate
gephyrin clustering involved in postsynaptic density assembly	GO:0097116	The clustering process in which gephyrin molecules are localized to distinct domains in the postsynaptic density as part of postsynaptic density assembly. Gephyrin is a component of the postsynaptic protein network of inhibitory synapses
GMP metabolic process	GO:0046037	The chemical reactions and pathways involving GMP, guanosine monophosphate
ionotropic glutamate receptor signaling pathway	GO:0035235	A series of molecular signals initiated by glutamate binding to a glutamate receptor on the surface of the target cell, followed by the movement of ions through a channel in the receptor complex. Ends with regulation of a downstream cellular process, e.g. transcription
maintenance of epithelial cell apical/basal polarity	GO:0045199	The maintenance of the apicobasal polarity of an epithelial cell
negative regulation of peptidyl-cystein S-nitrosylation	GO:1902083	ny process that stops, prevents or reduces the frequency, rate or extent of peptidyl-cysteine S-nitrosylation
neurotransmitter secretion	GO:0007269	The regulated release of neurotransmitter from the presynapse into the synaptic cleft via calcium regulated exocytosis during synaptic transmission

positive regulation of excitatory postsynaptic potential	GO:2000463	Any process that enhances the establishment or increases the extent of the excitatory postsynaptic potential (EPSP) which is a temporary increase in postsynaptic potential due to the flow of positively charged ions into the postsynaptic cell. The flow of ions that causes an EPSP is an excitatory postsynaptic current (EPSC) and makes it easier for the neuron to fire an action potential
positive regulation of synapse assembly	GO:0051965	Any process that activates, maintains or increases the frequency, rate or extent of synapse assembly, the aggregation, arrangement and bonding together of a set of components to form a synapse
positive regulation of synaptic vesicle clustering	GO:2000809	Any process that activates or increases the frequency, rate or extent of synaptic vesicle clustering
postsynaptic density protein 95 clustering	GO:0097119	The clustering process in which postsynaptic density protein 95 (PSD-95) molecules are localized to distinct domains in the cell membrane. PSD-95 is mostly located in the post synaptic density of neurons, and is involved in anchoring synaptic proteins
protein localization to basolateral plasma membrane	GO:1903361	A process in which a protein is transported to, or maintained in, a location within a basolateral plasma membrane
receptor localization to synapse	GO:0097120	Any process in which a receptor is transported to, and/or maintained at the synapse, the junction between a nerve fiber of one neuron and another neuron or muscle fiber or glial cell
regulation of grooming behaviour	GO:2000821	Any process that modulates the frequency, rate or extent of grooming behavior
regulation of sodium ion transmembrane transport	GO:1902305	Any process that modulates the frequency, rate or extent of sodium ion transmembrane transport
vocalization behaviour	GO:0071625	The behavior in which an organism produces sounds by a mechanism involving its respiratory system
Molecular Function		
cell adhesion molecule binding	GO:0050839	Interacting selectively and non-covalently with a cell adhesion molecule
extracellular-glutamate-gated ion channel activity	GO:0005234	Enables the transmembrane transfer of an ion by a channel that opens when extracellular glutamate has been bound by the channel complex or one of its constituent parts
guanylate kinase activity	GO:0004385	Catalysis of the reaction: ATP + GMP = ADP + GDP
ionotropic glutamate receptor binding	GO:0035255	Interacting selectively and non-covalently with an ionotropic glutamate receptor. Ionotropic glutamate receptors bind glutamate and exert an effect through the regulation of ion channels
L27 domain binding	GO:0097016	Interacting selectively and non-covalently with a L27 domain of a protein. L27 is composed of conserved negatively charged amino acids and a conserved aromatic amino acid. L27 domains can assemble proteins involved in signaling and establishment and maintenance of cell polarity into complexes by interacting in a heterodimeric manner
neurexin family protein binding	GO:0042043	Interacting selectively and non-covalently with neurexins, synaptic cell surface proteins related to latrotoxin receptor, laminin and agrin. Neurexins act as cell recognition molecules at nerve terminals
neuroligin family protein binding	GO:0097109	Interacting selectively and non-covalently with a member of the neuroligin protein family, neuronal cell surface proteins that mediate synapse formation

NMDA glutamate receptor activity	GO:0004972	An cation channel that opens in response to binding by extracellular glutamate, but only if glycine is also bound and the membrane is depolarized. Voltage gating is indirect, due to ejection of bound magnesium from the pore at permissive voltages
PDZ domain binding	GO:0030165	Interacting selectively and non-covalently with a PDZ domain of a protein, a domain found in diverse signaling proteins
scaffold protein binding	GO:0097110	Interacting selectively and non-covalently with a scaffold protein. Scaffold proteins are crucial regulators of many key signaling pathways. Although not strictly defined in function, they are known to interact and/or bind with multiple members of a signaling pathway, tethering them into complexes
Cellular Component		
basolateral plasma membrane	GO:0016323	The region of the plasma membrane that includes the basal end and sides of the cell. Often used in reference to animal polarized epithelial membranes, where the basal membrane is the part attached to the extracellular matrix, or in plant cells, where the basal membrane is defined with respect to the zygotic axis
bicellular tight junction	GO:0005923	An occluding cell-cell junction that is composed of a branching network of sealing strands that completely encircles the apical end of each cell in an epithelial sheet; the outer leaflets of the two interacting plasma membranes are seen to be tightly apposed where sealing strands are present. Each sealing strand is composed of a long row of transmembrane adhesion proteins embedded in each of the two interacting plasma membranes
cell junction	GO:0030054	A cellular component that forms a specialized region of connection between two or more cells or between a cell and the extracellular matrix. At a cell junction, anchoring proteins extend through the plasma membrane to link cytoskeletal proteins in one cell to cytoskeletal proteins in neighboring cells or to proteins in the extracellular matrix
dendritic spine	GO:0043197	A small, membranous protrusion from a dendrite that forms a postsynaptic compartment - typically receiving input from a single presynapse. They function as partially isolated biochemical and an electrical compartments. Spine morphology is variable including "thin", "stubby", "mushroom", and "branched", with a continuum of intermediate morphologies. They typically terminate in a bulb shape, linked to the dendritic shaft by a restriction. Spine remodeling is thought to be involved in synaptic plasticity
dystrophin-associated glycoprotein complex	GO:0016010	A multiprotein complex that forms a strong mechanical link between the cytoskeleton and extracellular matrix; typical of, but not confined to, muscle cells. The complex is composed of transmembrane, cytoplasmic, and extracellular proteins, including dystrophin, sarcoglycans, dystroglycan, dystrobrevins, syntrophins, sarcospan, caveolin-3, and NO synthase
exocyst	GO:0000145	A protein complex peripherally associated with the plasma membrane that determines where vesicles dock and fuse. At least eight complex components are conserved between yeast and mammals
juxtaparanode region of axon	GO:0044224	A region of an axon near a node of Ranvier that is between the paranode and internode regions

MPP7-DLG1-LIN7 complex	GO:0097025	A heterotrimeric protein complex formed by the association of MMP7, DLG1 and either LIN7A or LIN7C; regulates the stability and localization of DLG1 to cell junctions
myelin sheath abaxonal region	GO:0035748	The region of the myelin sheath furthest from the axon
neuron projection	GO:0043005	A prolongation or process extending from a nerve cell, e.g. an axon or dendrite
NMDA selective glutamate receptor complex	GO:0017146	An assembly of four or five subunits which form a structure with an extracellular N-terminus and a large loop that together form the ligand binding domain. The C-terminus is intracellular. The ionotropic glutamate receptor complex itself acts as a ligand gated ion channel; on binding glutamate, charged ions pass through a channel in the center of the receptor complex. NMDA receptors are composed of assemblies of NR1 subunits (Figure 3) and NR2 subunits, which can be one of four separate gene products (NR2A-D). Expression of both subunits are required to form functional channels. The glutamate binding domain is formed at the junction of NR1 and NR2 subunits. NMDA receptors are permeable to calcium ions as well as being permeable to other ions. Thus NMDA receptor activation leads to a calcium influx into the post-synaptic cells, a signal thought to be crucial for the induction of NMDA-receptor dependent LTP and LTD
postsynaptic density of dendrite	GO:0014069	An electron dense network of proteins within and adjacent to the postsynaptic membrane of the dendrite of asymmetric synapses. Its major components include neurotransmitter receptors and the proteins that spatially and functionally organize them such as anchoring and scaffolding molecules, signaling enzymes and cytoskeletal components
postsynaptic membrane	GO:0045211	A specialized area of membrane facing the presynaptic membrane on the tip of the nerve ending and separated from it by a minute cleft (the synaptic cleft). Neurotransmitters cross the synaptic cleft and transmit the signal to the postsynaptic membrane
presynaptic membrane	GO:0042734	A specialized area of membrane of the axon terminal that faces the plasma membrane of the neuron or muscle fiber with which the axon terminal establishes a synaptic junction; many synaptic junctions exhibit structural presynaptic characteristics, such as conical, electron-dense internal protrusions, that distinguish it from the remainder of the axon plasma membrane
presynapse	GO:0098793	The part of a synapse that is part of the presynaptic cell
sarcolemma	GO:0042383	The outer membrane of a muscle cell, consisting of the plasma membrane, a covering basement membrane (about 100 nm thick and sometimes common to more than one fiber), and the associated loose network of collagen fibers.

synapse	GO:0045202	The junction between a nerve fiber of one neuron and another neuron, muscle fiber or glial cell. As the nerve fiber approaches the synapse it enlarges into a specialized structure, the presynaptic nerve ending, which contains mitochondria and synaptic vesicles. At the tip of the nerve ending is the presynaptic membrane; facing it, and separated from it by a minute cleft (the synaptic cleft) is a specialized area of membrane on the receiving cell, known as the postsynaptic membrane. In response to the arrival of nerve impulses, the presynaptic nerve ending secretes molecules of neurotransmitters into the synaptic cleft. These diffuse across the cleft and transmit the signal to the postsynaptic membrane
T-tubule	GO:0030315	Invagination of the plasma membrane of a muscle cell that extends inward from the cell surface around each myofibril. The ends of T-tubules make contact with the sarcoplasmic reticulum membrane
voltage-gated potassium channel complex	GO:0008076	A protein complex that forms a transmembrane channel through which potassium ions may cross a cell membrane in response to changes in membrane potential
Z disc	GO:0030018	Platelike region of a muscle sarcomere to which the plus ends of actin filaments are attached

Table F.2: GO terms enriched in at least two communities in Cluster 2 (alphabetical order of GO terms); significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, `elim` algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values available upon request since different in distinct enriched clusters).

GO term	GO term ID	GO term definition
Biological Process		
complement activation, classical pathway	GO:0006958	Any process involved in the activation of any of the steps of the classical pathway of the complement cascade which allows for the direct killing of microbes, the disposal of immune complexes, and the regulation of other immune processes
Molecular Function		
immunoglobulin receptor binding	GO:0034987	Interacting selectively and non-covalently with one or more specific sites on an immunoglobulin receptor molecule
serine-type endopeptidase activity	GO:0004252	Catalysis of the hydrolysis of internal, alpha-peptide bonds in a polypeptide chain by a catalytic mechanism that involves a catalytic triad consisting of a serine nucleophile that is activated by a proton relay involving an acidic residue (e.g. aspartate or glutamate) and a basic residue (usually histidine)
Cellular Component		
blood microparticle	GO:0072562	A phospholipid microvesicle that is derived from any of several cell types, such as platelets, blood cells, endothelial cells, or others, and contains membrane receptors as well as other proteins characteristic of the parental cell. Microparticles are heterogeneous in size, and are characterized as microvesicles free of nucleic acids
external side of plasma membrane	GO:0009897	The leaflet of the plasma membrane that faces away from the cytoplasm and any proteins embedded or anchored in it or attached to its surface
fibrinogen complex	GO:0005577	A highly soluble, elongated protein complex found in blood plasma and involved in clot formation. It is converted into fibrin monomer by the action of thrombin. In the mouse, fibrinogen is a hexamer, 46 nm long and 9 nm maximal diameter, containing two sets of non-identical chains (alpha, beta, and gamma) linked together by disulfide bonds
immunoglobulin complex, circulating	GO:0042571	An immunoglobulin complex that is secreted into extracellular space and found in mucosal areas or other tissues or circulating in the blood or lymph. In its canonical form, a circulating immunoglobulin complex is composed of two identical heavy chains and two identical light chains, held together by disulfide bonds. Some forms are polymers of the basic structure and contain additional components such as J-chain and the secretory component
platelet alpha granule lumen	GO:0031093	The volume enclosed by the membrane of the platelet alpha granule

Table F.3: GO terms enriched in at least two communities in Cluster 3 (alphabetical order of GO terms); significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, `elim` algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values available upon request since different in distinct enriched clusters).

GO term	GO term ID	GO term definition
Biological Process		
neuromuscular junction development	GO:0007528	A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of a neuromuscular junction
Molecular Function		
high voltage-gated calcium channel activity	GO:0008331	Enables the transmembrane transfer of a calcium ion by a high voltage-gated channel. A high voltage-gated channel is a channel whose open state is dependent on high voltage across the membrane in which it is embedded
Cellular Component		
L-type voltage-gated calcium channel complex	GO:1990454	A type of voltage-dependent calcium channel responsible for excitation-contraction coupling of skeletal, smooth, and cardiac muscle. 'L' stands for 'long-lasting' referring to the length of activation

Appendix G

Clathrin Mediated Endocytosis - a Dynamic Model

Anatoly Sorokin^{1*}, Katharina F. Heil^{2*}, J. Douglas Armstrong² and Oksana Sorokina²

*These authors contributed equally to this work.

¹ - Institute of Cell Biophysics, RAS, Russia

² - Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh

Rule-based modelling provides an extendable framework for comparing candidate mechanisms underpinning clathrin polymerisation.

Anatoly Sorokin^{1,2}, Katharina F Heil³, J Douglas Armstrong³ and Oksana Sorokina^{3*}

*Correspondence to Oksana.Sorokina@ed.ac.uk

Author Affiliations:

1. Institute of Cell Biophysics, RAS, Pushchino, 142290, Russia
2. Moscow Institute of Physics and Technology, 141700, Dolgoprudnyi, Moscow Region, Russia
3. School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom

Abstract:

Polymerisation of clathrin is a key process that underlies clathrin-mediated endocytosis. Clathrin-coated vesicles are responsible for cell internalization of external substances required for normal homeostasis and life –sustaining activity. There are several hypotheses describing formation of closed clathrin structures. According to one of the proposed mechanisms cage formation may start from a flat lattice buildup on the cellular membrane, which is later transformed into a curved structure. Creation of the curved surface requires rearrangement of the lattice, induced by additional molecular mechanisms. Different potential mechanisms require a modeling framework that can be easily modified to compare between them. We created an extendable rule-based model that describes polymerisation of clathrin molecules and various scenarios of cage formation. Using Global Sensitivity Analysis (GSA) we obtained parameter sets describing clathrin pentagon closure and the emergence/production and closure of large-size clathrin cages/vesicles. We were able to demonstrate that the model can reproduce budding of the clathrin cage from an initial flat array.

Introduction

Clathrin is the major protein component of clathrin-mediated endocytosis (CME)^{1,2}. Due to its particular shape and (auto-) polymerization capacity, clathrin is believed to induce the cell membrane to adopt a vesicular shape. A range of different mechanisms have been proposed for this process³⁻⁵, from a few minimalistic ones propose that clathrin polymerization alone is sufficient to generate buds in a planar membrane⁶ to the consensus that describe the orchestrated action of additional proteins and signaling cascades on the intracellular side of the membrane, so that ~30 proteins directly participate in the various steps of endocytosis^{1,7-9}.]

The structural properties of clathrin have been extensively investigated with respect to their role in vesicle formation. Usually a clathrin molecule is composed of one heavy (~190 kDa) as well as one light chain (~25 kD) and is about 475 Ångström (Å) in length¹⁰. Within the cell clathrin exists in a form of trimers (triskelia), consisting of three clathrin molecules (three heavy and three light chains respectively), where individual clathrin monomers are referred to as “legs”. Deviating from the normal 1:1 ratio between light and heavy

chain several studies have also revealed the existence of triskelia with fewer light chains. Triskelia formation itself does not seem to be influenced by a loss of light chain molecules¹¹, but regulatory control of vesicle formation and cargo selection have been proposed.

Due to its internal trimeric structure every single clathrin molecule in the triskelia complex can polymerize with another clathrin molecule from a different clathrin triskelia. Hence every triskelia is able to undergo interactions with three further triskelia. This leads to the formation of dimers and trimers, which can grow to construct large polymers. However, in a normal biological context, hexagonal and pentagonal shapes are among the most frequently observed^{12,13}. Specific combinations of these shapes induce the formation of the typical vesicle closed spherical structure. Normally, closed structures contain 12 pentagonal faces and $(N-20)/2$ hexagonal faces. The fixed relative numbers between pentagonal and hexagonal faces are based on geometric constraints, given the clathrin structure and minimal flexibility of the trimer legs. Based on the number (N) of triskelia different sphere sizes can emerge, three of which are well defined: The mini-coat, hexagonal-barrel and soccer ball¹³.

Since its discovery in 1975¹⁴, significant attention has been focused on the mechanism of clathrin polymerisation. It was highlighted in¹ that understanding CME is not possible without proper knowledge of its key process, the clathrin cage formation. Although it was experimentally shown that clathrin self-assembles following pH decrease from 8 to 6.5¹⁵ or under bivalent cation administration¹⁶, to obtain biologically realistic vesicle shapes the participation of external regulatory proteins is likely critical¹.

A range of computational models for clathrin self-assembly exists that describes the formation of clathrin cages^{12,13,17-19}, or pits and vesicles^{13,15,20}. Early models considered the association of 3-valent polymers with equi-reactive binding sites from the Flory's theory point of view with²⁰ or without²¹ allowance for intramolecular loop formation. These studies dissected the dependence of the solution/gel phase transition linked to the critical concentration of the monomer on the equilibrium constants of different steps of the polymerisation process. In the early theoretical models of multivalent condensation, the term "gel" was used to describe the situation when the majority of agents participate in one global complex. There are two phases in such system: a solution consisting of many small complexes and monomers, and a gel, composed of one global complex and a few free monomers. The formation of the global complex is a key phase transition in the systems dynamics. Prior to gel formation, the dynamics of the system are driven by bi-molecular reactions (when two complexes form a bigger one, or a monomer attaches to the complex). After gel formation, the dynamics are driven by uni-molecular reactions within the complex. The key finding of Falk and Thomas²⁰ is that before the transition to the gel phase, uni-molecular reactions are negligible.

In particular, it was shown by Pastan and Willingham¹⁵, that the critical concentration of clathrin, sufficient for the phase transition was 30 mg/ml. Taking into account that the triskelia molecular mass is about 640kDa, this value corresponds to the molar concentration of 46 μ M, or approximately 55000 triskelia per eukaryotic cell.

More recent studies examined the assembly of 5- and 6- member rings in parallel with investigation of how different physical triskelia characteristics might impact on cage formation. These characteristics include

triskelia rigidity²¹, their asymmetry¹⁷, emergent tension during cage closure²² and the effects of superficial membrane tension²³. These studies provide approximations of binding energy between the chains of the neighbouring clathrin triskelia¹⁷.

The polymerisation process alone presents a significant challenge for mechanistic modeling, as the number of molecular species, which have to be described, grows exponentially with the number of available monomers. Rule-based modeling^{24–26} provides a viable solution allowing a network-free simulation technique^{27–29}. It uses ‘lumped’ reaction rules to concisely represent molecule interactions. One can assume the rules as implicit combinations of different reactions into classes, where all the members of the same class perform a common transformation. This modeling approach is generally exploited for large-scale biochemical systems to overcome combinatorial complexity and it has previously demonstrated its effectiveness in simulations of ligand-receptor complex polymerization²⁵.

Here we present a suite of rule-based models of clathrin polymerisation with increasing complexity, starting from a very basic model where the molecule has three equally reactive binding sites to a more advanced model reproducing realistic triskelia clathrin structure. We examined the correspondence of each model’s behavior with the existing theoretical models while sampling from a wide range of parameter values.

We found that although the basic model exactly reproduces Flory’s findings, it is unable to provide the amounts of 5- and 6- member rings required for cage formation and, therefore, it fails to reproduce clathrin vesicle formation. A revised model with a more realistic clathrin structure that explicitly supports predominant closure of pentagons and hexagons allows 3D cage formation and permits the evolution of flat 2D clathrin patches into a 3D cage structures by shifting the ratio of the pentagon/hexagon dissociation constants.

Methods

Models and simulation

We used the Kappa language³⁰, a member of the family of rule-based modeling languages, for building the models. All models were simulated by KaSim3.5 (<http://dev.executableknowledge.org/>). We used Kappa extensions where appropriate, e.g the MetaKappa (<https://github.com/kappamodeler/metakappa>) extension for building the first model to handle the combinatorial explosion caused by three equal binding sites (see Appendix for details). Also, we use the RKappa extension³¹ for sampling the large parameter space, statistical analysis of simulation results, global sensitivity analysis (GSA) and visualization of the Kappa molecular structures as more comprehensible 2D and 3D graphs.

We first investigated the capability of rule-based models to reproduce clathrin cage structures based on random self-assembly processes. For this we assume that clathrin triskelia interact in 3D, in a well-mixed solution and all binding sites of the clathrin triskelia are assumed to be identical. Due to the combinatorial nature of the clathrin molecule association, the size of aggregates is unbounded and limited only by the amount of available substrate.

We started with a reduced model of triskelia monomers similar to Perelson and Goldstein’s equilibrium and continuous model²¹, in which monomers carry three identical equally reactive binding sites. Two variants of this model were implemented in the rule-based Kappa language to investigate the polymerization of

branched complexes from a single class of trivalent agents under 'rings allowed' and 'rings forbidden' conditions similar to that proposed by²⁰ (Model 1).

We then developed a more elaborate model, based on clathrin monomers, that considers triskelia as a predefined complex of three monomers. This model more accurately reproduces the structure of clathrin with distinct legs and binding sites along with specified defined steric and chirality constraints (Model 2). It also contains explicit rules describing formation of penta- and hexagonal rings and demonstrates the dynamics of closed cage structure formation. All the models presented here are kinetic and do not include notions of space. However these could be added by use of existing extensions like SpatialKappa²⁶ or Geometric Kappa³² if required later.

1. Equireactive trivalent agent model

In the first model (Model 1) we simplify the realistic triskelia structure of clathrin to the trivalent agent Cl3 with three identical binding sites. This is effectively a kinetic version of the model described by Perelson and Goldstein in 1985^{16,21} (Figure 1A, Supplementary Data). As clathrin is known to aggregate on the membrane, we assume that with complex growth its ability to diffuse would decrease. Thus, in our configuration complex growth happens preferentially via addition of new monomers rather than merging of existing complexes, in the same way as in Perelson and Goldstein.

The (κ) rule looks as follows:

'proximal binding' $a(A,A,A),a(A) \rightarrow a(A!1,A,A),a(A!1) \# @ 'pbk' (0)$,

where ' pbk ' is the rate of binding.

To ensure stability of the rings in clathrin complexes we make an assumption that molecules with three occupied binding sites cannot dissociate. Thus, dissociation is only possible at the periphery of the complex when at least one binding site is/remains free.

'proximal dissociation' $a(A!1,A),a(A!1) \rightarrow a(A,A),a(A) \# @ 'pdk'$,

where ' pdk ' is the rate of dissociation.

This rule partially contradicts the work of Perelson and Goldstein, where the dissociation is possible only at the monomer level. However, the rule includes the dissociation of terminal monomers as a special case.

We studied the random polymerisation of trivalent monomers under two traditional Flory- Stockmayer assumptions: 'ring forbidden' (Model 1A) and 'rings allowed' (Model 1B).

In the case of Model 1A ('rings forbidden'), the intramolecular bonds between the binding sites of the same polymers are not allowed as the only free agent (with all three sites non-occupied) can bind the polymer. The detailed models for the original Perelson's model and its two Kappa implementations: Model 1A and Model 1B are presented in Supplementary Data.

In the case of Model 1B ('ring allowed') intra-molecular reactions are allowed, so that rings of different sizes may occur. As in^{20,21,33} reactions occur with an equal probability for each of the free binding site to react until the reaction extent $R_{\text{ext}} = 1$, which means that all binding sites are fully occupied. Although cubical structures of clathrin were observed experimentally under special conditions³⁴, the formation of rings of size 4 and less

is not reported under conditions approximating intracellular environments. Hence we set a specific constraint on the polymer chain ability to make intramolecular bonds only when ring size (nring) exceeds 5 bonds in length.

'ring closure' a(A),a(A) -> a(A!1),a(A!1) #@ 'pring' (0.0:'nring')

In the rule above *'pring'* is a rate of ring closure, while *'nring'* refers to the minimal number of bonds in the ring (set to 5 in this case). The constraints enforce limitations on the condition of equal reactivity to be always fulfilled; yet the probability to close a short ring within a large complex is quite small. We also assume the equilibrium constants for initiation, elongation and branching are equal.

2.Triskelia model

To generate a more realistic model we next considered clathrin monomers and their structural properties. Each monomer consists of a proximal region ("P", light green in Figure 1B), which contains a binding domain on its "right", long part ("r") and "left", short part ("l"), and the distal region ("d", dark green in Figure 1B). Domains in the proximal region facilitate the internal binding of monomers to form trimers. The additional binding sites "Pp" and "Pd" in the proximal region allow binding amongst different triskelia. Binding rules presume the 'right' part of one monomer can only bind to the "left" part of another, and so forth to make correct triskelia structures (Figure 1C).

In kappa language this is expressed in the following way:

Cl(l!1,r!2),Cl(r!1,l!3),Cl(r!3,l!2)

"Cl" refers to a single clathrin molecule with proximal right ("r") and left ("l") binding site. All distal parts of the long legs are oriented in one direction, showing a clockwise drift/turn (Figure 1C).

Once assembled, triskelia form the structural unit for the polymerisation process, which is governed by the interaction of domains localised on the right, long leg of each monomer. These are: a proximal (Pp), a distal "receiving" (Pd) and distal "giving" (d) domain. Based on the given clathrin triskelia structure, formation of one bond utilizes four triskelia simultaneously: two monomers bind with their proximal parts, and two form additional bonds with their distal parts (see Supplementary Data for triskelia binding code and a visualization). As was shown by den Otter et al.¹⁷ and Fotin et al.³⁵, the proper orientation of all four legs is vital for formation of closed structures. Initial polymerization steps along with the model rules are presented in detail in Supplementary Data.

In addition to the binding rule, a few specific rules enforce the closure/formation of pentagons and hexagons. Dissociation is implemented as follows. Closed rings cannot be reopened. At least one monomer needs to be unbound for dissociation to happen. Details can be seen in the model code in the Supplementary Data, which shows the rules used in the current model version.

Data Availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

Results

We investigated the ability of rule-based models to reproduce the clathrin cage structures based on a random self-assembly process. Specifically, two traditional Flory- Stockmayer conditions: “rings forbidden” and “rings allowed” were applied separately, similar to²⁰. All models were simulated 5000 times with parameter ranges shown in Table 1.

1. Trivalent model.

In the first model we used a simplified triskelia structure of clathrin with a trivalent agent Cl3 containing three identical binding sites with equal reactivity, similar to the Perelson and Goldstain model in 1985²¹.

The key parameters that have been analyzed are (see also²¹):

$$R_{ext} = \frac{2 * N_{bond}}{3 * amount} \quad (1)$$

$$\alpha = 6 * K * C_t = 6 * \frac{pbk * (N_A * V)}{pdk} = 6 * N_t * \frac{pbk}{pdk} \quad (2)$$

where R_{ext} - reaction extent, α - nondimensional equilibrium constant, N_{bond} - the number of bonds in the polymer, and K – the equilibrium constant. C_t and N_t describe the total concentration and total number of monomers (respectively), *amount* – amount of available triskelia.

We showed that in the “ring forbidden” setup, the distribution of free clathrin with dependence on R_{ext} exactly followed the prediction of Perelson’s theory (Fig 2 A and B). The vast majority of parameter sets in “ring forbidden” are grouped around $R_{ext} = 0.5$, and the dependency between R_{ext} and N_{bond}/N_t is linear. We found that R_{ext} never exceeded the theoretical limit of gel formation (Figure 2B) while in most of the “ring allowed” instances, reactions stopped only when the available binding sites were saturated (Figure 2 B and D).

To explore the types of complexes our simulations produced, we calculated the size of the largest aggregate (W_{max}) and the number of the rings in the system. The latter was estimated as the cyclomatic number of the clathrin graph, which is the number of bonds that need to be removed to form an acyclic graph:

$$C_{rank} = E_g - V_g + C_g \quad (3)$$

with E_g number of edges and V_g number of nodes in the graph. C_g is the number of connected components in the graph. We found that the number of rings in the system (C_{rank}) almost always reached the theoretical

limit (Fig 2D), where the total number of monomers was equal to the size of the largest aggregate (W_{\max}) in agreement with analysis from Falk et al.²⁰

In agreement with²⁰, when intramolecular bonds are allowed (Model 1b) ring formation only starts after gel structure formation (Fig 2D), when the reaction extent reaches the 0.5 threshold. This means that in the simple agent model closed cages would be formed only when 7/8 of the available clathrins form a large single complex.

Further analysis (Supplementary Data) shows that probability of the ring closure grows with the size of the ring. Therefore, the number of short rings (pentagons and hexagons) is quite low even when we set the rate of the ring closure reactions to infinity (Supplementary Data and Figure 3). Therefore we conclude that the simple model is not able to describe the closed cage structures, as the clathrin geometry provides the optimal mutual disposition of the monomers only when 5- and 6-membered rings are formed. To resolve this we developed a more plausible model as follows.

2. Triskelia model.

Model 2 described above corresponds to a more realistic structure of clathrin with distinct regions within the monomer and respective binding sites that reflect the experimental literature^{10,12}. We also introduced a specific rule for orientation of the monomers to ensure that the “right” site of one monomer binds the “left” side of another. This preserves the correct geometry of triskelia and chirality of the monomers. To ensure we obtain realistic clathrin complexes, 5- and 6- ring closure reactions were explicitly specified.

We started with parameter sampling for the model. To ensure comparability between simulations we used the same parameter sets as before by assigning the ring closure rate the value of “pring” to both hexagons and pentagons. Again, the two cases - “ring forbidden” and “ring allowed” were investigated.

The behavior of the “ring forbidden” version of Model 2 is clearly similar to the behavior of the Model 1 and theoretical predictions of Perelson (Figure 4A). The number of free triskelia monotonically decreases towards zero at $R_{ext} = 0.6$. The difference between the theoretical prediction of 0.5 and the observed value is explained by the association rule in the Model 2, which does not prevent associations of clusters and therefore does not follow the monomer attachment mechanisms considered in Perelson²¹. Association between clusters results in a higher numbers of triskelia with all their legs involved in the complex formation, which in turn prevents their dissociation.

The “ring allowed” version of the Model 2 (Figure 4B) follows the same scenario as Model 1 (Figure 2D) and behaves as predicted by theory²⁰. Ring formation starts only after solution to gel transition at $R_{ext} = 0.6$. Contrary to the Model 1, the number of rings does not grow linearly with the size of the complex (Figure 2C). Instead, due to the system not being allowed to form rings of arbitrary size, we obtain many small complexes with few or no rings (Supplementary Figure 2).

For simplicity of simulation and comparison with Model 1 we did not introduce separate kinetic constants for 5- and 6-ring closure. As a result, the vast majority of the rings in our simulations are pentagons. Nevertheless we observed a number of hexagons as well. The relatively high number of octagons observed is a consequence of high number of 5-rings, as hull of two adjacent pentagons can form an octagon (Figure 5C).

To explore the geometry of complexes, which contain 5- and 6-rings we used a set of all possible combinations of pentagons and hexagons as described in²². Table 2 shows that pentagons tend to form adjacent dodecahedron-like structures (see g551, Figure 5A, Supplementary Figure 5), while hexagons are most often surrounded by pentagons as visualized in structure g661 (Figure 5B). We found no clear distinction between ring forming and ring preventing values in parameter sets (Supplementary Figure 3). To further investigate which parameters influence the ring formation the most we performed GSA on Model 2 with the “ring allowed” condition (Supplementary Table 1). We thus concluded that Model 2 is able to produce various structures of different shapes (Figure 5) without the initial constraints, but that they do not all necessarily end up being cage-like structures.

The type of clathrin cage formed *in vivo* is known to depend on the ratio of pentagons and hexagons^{3,22}. Moreover, planar clathrin consists of just hexagons. As an example we tested to see whether our model could be reconciled with the invagination mechanism (e.g. described in Avinoam et. al.⁵). Avinoam’s (2015) mechanism requires the presence of pentagons. To reproduce this we tuned the rate constants for pentagon and hexagon closure and changed the equilibrium of association and dissociation rates for them. First we simulated the model where only 6-rings were allowed by setting the 5-ring closure reaction to 0 to form a planar structure (Figure 6A and Supplementary movie 1). When the reaction extent was close to 1, 5-ring closure was allowed by adjusting rate constant to non-zero value. With a rate of closure for 5- and 6-ring close to each other we observed invaginations, but they never reached the scissing stage so that the completely closed structure never occurred (Supplementary movie 2). At this point we set the rate of closure for 5-rings to infinite and after 10^4 events we obtained the structures shown in Figure 6B and Supplementary Movie 3.

To evaluate the influence of rates of pentagon and hexagon closure/disruption we performed GSA on the model starting with a flat hexagonal mesh (Supplementary Table 2). Here, b and d are the coefficients defining the extent to which pentagon closure is faster than hexagon closure (b), and hexagon compared to pentagon dissociation (d); $rng5$ and $rng6$ are the ratios of ring closure to ring disruption for pentagons and hexagons, respectively. For each parameter the significance level is calculated as described in³⁶. The rate of the pentagon closure did not significantly influence any property of the system, while the rate of hexagon dissociation appeared important for the size of the most frequent complex ($wNmax$) and the presence of hexagon-containing subgraphs (g501, g511, g521, g522, g601, g611, g621, g622, g631, g632, g633, g641, g642, g643 in Supplementary Tables 1 and 2)²². During the course of a simulation we were able to obtain different numbers of closed cages in almost half of the parameter sets, which indicates that the formation of flat structures requires additional constraints, while cage formation happens spontaneously⁴.

Discussion

Computational models describing formation of clathrin-coated vesicles (CCVs)^{2,17,23,37} mostly focus on clathrin self-association or its association with the membrane. However, vesicle recycling is regulated via a large number of signalling processes^{2,38}. Existing computational models struggle to incorporate these regulatory elements either because of high computational cost, which becomes prohibitive in case of incorporation of all involved protein types, or because the structure/type of the model can/does not include the reactions controlled by regulatory systems. For example, the equilibrium model²¹ considered growth of pits as a linear set of reactions, assuming that all three legs of the new triskelia in the pit assemble using the best possible free sites in the net. As shown by simulations in¹⁷ and confirmed in our Models 1A and B this is not the case.

As was proposed in³⁹, these signaling processes can be incorporated into models as a modification of clathrin association/dissociation rates. With these factors in mind we have developed a model capable of describing the formation of CCVs, avoiding the more resource expensive computational algorithms and using a modeling format familiar to the signal transduction modelling community.

Our first version of the model, which described clathrin as a trivalent agent demonstrated that formation of closed structures required an additional manual closure to achieve 5- and 6- rings. With the flexibility of the clathrin molecule and no evidence for energy differences between penta- and hexameric rings we saw no preferences towards either specific ring composition. Weak interactions, which have been proposed to have a major effect on the association of clathrin legs³⁹, and comparatively low bending energy of the clathrin lattice suggest that when on the flat part of the membrane, clathrin will create a flat hexagonal lattice. That process was considered in³, where clathrin was modeled as hexagonal lattice with 5- and 7-sided rings occurring as defects, but the study only considered the equilibrium state, whereas in our analysis we were able to investigate the kinetics of the process. Although the “canonic” mechanism of clathrin pits formation proposes constant curvature growth as a function of clathrin polymerization⁴⁰, the evolution of curved clathrin structures from flat plaque has also some supporting experimental evidence^{5,38}. The recent study of Leyton-Puig et al.⁷ reports the ability of clathrin plaques to act as hubs for CME and proposes actin polymerisation and actin-based adhesion are major regulating factors for their remodeling⁷.

Our model shows that switching pentagon ring formation on/off allows the process to switch between planar patches and closed cages. *In vivo*, this switching could be driven by changes in physical properties of the membrane or by additional regulatory mechanisms^{1,37,41}.

In our model we assume the size and the shape of the clathrin lattice to be controlled by three processes: i) the association/dissociation of triskelia; ii) the 5-ring formation/dissociation and iii) the 6-ring formation/dissociation. Several other factors are known to influence the cage and coat formation and dissociation^{42,43}. For example in⁴⁴, the main difference in pentagon and hexagon closure is attributed to the stiffness of the underlying membrane, while in⁴¹ the rigidity variation of the clathrin net itself is explained by binding to an adaptor protein (AP2, AP3, AP180)^{8,45}. Their influence on clathrin coat formation has been studied in distinct experimental setups and binding to clathrin has been confirmed. Due to their influence on clathrin triskelia structure and hence their ability to influence coat formation it might be debatable if their main

role is in maintaining a flat structure or “forcing”/inducing the formation of vesicles. This mechanism could be easily embedded into the model (see the example in Supplementary Data).

The clathrin light chain is an additional part of the triskelia, which connects to the heavy chain in the region extending from the self-association domain to the knee³⁹. One of the possible conformations can force the knee to bend in a direction that inhibits cage formation. This inhibitory effect is thought to be regulated (inhibited) by interaction with Ca ions or by lowering the pH³⁹. The light chain also influences the rigidity of the clathrin lattice and its ability to bend the lipid membrane at low temperature⁴. The light chain contains 19 serines that are potential kinase targets (GRK2) and phosphorylation of the light chain has been proposed as a discriminator for different cargo inclusion in the vesicle⁴⁶. An example of how the model can be extended to incorporate the above mechanism is presented in Supplementary Data.

The rule-based approach we have used allows us to build and compare kinetic models that describe different possible mechanisms of clathrin cage formation, from direct assembly from monomers at the vesicle budding site to the invagination of flat membrane plaque. More in depth functional details such as the role of N-WASP through Arp2/3⁷ can help to expand models and gain deeper insights. Hence, our implementation is easily extendable allowing the future inclusion of more detailed mechanistic models of CME regulation.

Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 720270 (HBP SGA01).

Competing financial interests

The author(s) declare no competing financial interests.

Authors contribution

AS, OS model design; AS, OS, KFH model simulation, GSA, analysis; AS, KFH, JDA, writing; OS direction and writing

* AS and KFH equally contributed to this work

References

1. McMahon, H. T. & Boucrot, E. Molecular mechanism and physiological functions of clathrin-mediated endocytosis. *Nat. Rev. Mol. Cell Biol.* **12**, 517–533 (2011).
2. Jung, N. & Haucke, V. Clathrin-mediated endocytosis at synapses. *Traffic* **8**, 1129–1136 (2007).
3. Jin, A. J. & Nossal, R. Topological mechanisms involved in the formation of clathrin-coated vesicles. *Biophys. J.* **65**, 1523–1537 (1993).
4. Dannhauser, P. N. *et al.* Effect of clathrin light chains on the stiffness of clathrin lattices and membrane budding. *Traffic* **16**, 519–533 (2015).
5. Avinoam, O., Schorb, M., Beese, C. J., Briggs, J. A. G. & Kaksonen, M. Endocytic sites mature by continuous bending and remodeling of the clathrin coat. *Science* **348**, 1369–1372 (2015).

6. Dannhauser, P. N. & Ungewickell, E. J. Reconstitution of clathrin-coated bud and vesicle formation with minimal components. *Nat. Cell Biol.* **14**, 634–639 (2012).
7. Leyton-Puig, D. *et al.* Flat clathrin lattices are dynamic actin-controlled hubs for clathrin-mediated endocytosis and signalling of specific receptors. *Nat. Commun.* **8**, 16068 (2017).
8. Smith, S. M., Baker, M., Halebian, M. & Smith, C. J. Weak Molecular Interactions in Clathrin-Mediated Endocytosis. *Front Mol Biosci* **4**, 72 (2017).
9. Saheki, Y. & De Camilli, P. Synaptic vesicle endocytosis. *Cold Spring Harb. Perspect. Biol.* **4**, a005645 (2012).
10. Fotin, A. *et al.* Structure of an auxilin-bound clathrin coat and its implications for the mechanism of uncoating. *Nature* **432**, 649–653 (2004).
11. Girard, M., Allaire, P. D., McPherson, P. S. & Blondeau, F. Non-stoichiometric relationship between clathrin heavy and light chains revealed by quantitative comparative proteomics of clathrin-coated vesicles from brain and liver. *Mol. Cell. Proteomics* **4**, 1145–1154 (2005).
12. Kirchhausen, T., Owen, D. & Harrison, S. C. Molecular structure, function, and dynamics of clathrin-mediated membrane traffic. *Cold Spring Harb. Perspect. Biol.* **6**, a016725 (2014).
13. Fotin, A. *et al.* Structure determination of clathrin coats to subnanometer resolution by single particle cryo-electron microscopy. *J. Struct. Biol.* **156**, 453–460 (2006).
14. Pearse, B. M. Coated vesicles from pig brain: purification and biochemical characterization. *J. Mol. Biol.* **97**, 93–98 (1975).
15. Pastan, I. & Willingham, M. C. The pathway of endocytosis. *J. Mol. Biol.* **97**, 1–44 (1985).
16. Pearse, B. M. & Crowther, R. A. Structure and assembly of coated vesicles. *Annu. Rev. Biophys. Biophys. Chem.* **16**, 49–68 (1987).
17. den Otter, W. K., Renes, M. R. & Briels, W. J. Asymmetry as the key to clathrin cage assembly. *Biophys. J.* **99**, 1231–1238 (2010).
18. den Otter, W. K. & Briels, W. J. The generation of curved clathrin coats from flat plaques. *Traffic* **12**, 1407–1416 (2011).
19. Matthews, R. & Likos, C. N. Structures and pathways for clathrin self-assembly in the bulk and on membranes. *Soft Matter* **9**, 5794–5806 (2013).
20. Falk, M. & Thomas, R. E. Molecular size distribution in random polyfunctional condensation with or without ring formation: computer simulation. *Can. J. Chem.* **52**, 3285–3295 (1974).
21. Perelson, A. S. & Goldstein, B. The equilibrium aggregate size distribution of self-associating trivalent molecules. *Macromolecules* **18**, 1588–1597 (1985).

22. Schein, S. & Sands-Kidner, M. A geometric principle may guide self-assembly of fullerene cages from clathrin triskelia and from carbon atoms. *Biophys. J.* **94**, 958–976 (2008).
23. Banerjee, A., Berezhkovskii, A. & Nossal, R. Stochastic model of clathrin-coated pit assembly. *Biophys. J.* **102**, 2725–2730 (2012).
24. Danos, V., Feret, J., Fontana, W., Harmer, R. & Krivine, J. Rule-based modelling and model perturbation. *Transactions on Computational Systems Biology XI* 116–137 (2009).
25. Monine, M. I., Posner, R. G., Savage, P. B., Faeder, J. R. & Hlavacek, W. S. Modeling multivalent ligand-receptor interactions with steric constraints on configurations of cell-surface receptor aggregates. *Biophys. J.* **98**, 48–56 (2010).
26. Sorokina, O., Sorokin, A., Armstrong, J. D. & Danos, V. A simulator for spatially extended kappa models. *Bioinformatics* **29**, 3105 (2013).
27. Danos, V., Feret, J., Fontana, W. & Krivine, J. Scalable simulation of cellular signaling networks. *Computational Methods In Systems Biology, Proceedings* 139–157 (2009).
28. Colvin, J. *et al.* RuleMonkey: software for stochastic simulation of rule- based models. *BMC Bioinformatics* **11**, 404 (2010).
29. Sneddon, M. W., Faeder, J. R. & Emonet, T. Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. *Nat. Methods* **8**, 177–183 (2011).
30. Danos, V., Feret, J., Fontana, W. & Krivine, J. Abstract interpretation of cellular signalling networks. *Verification, Model Checking, and Abstract Interpretation* 83–97 (2008).
31. Sorokin, A., Sorokina, O. & Armstrong, J. D. RKappa: Statistical sampling suite for Kappa models. in *Hybrid Systems Biology* (eds. Maler, O., Halasz, A. & Piazza, C.) 128–142 (Springer, 2015).
32. Danos, V., Honorato-Zimmer, R., Jaramillo-Riveri, S. & Stucki, S. Rigid Geometric Constraints for Kappa Models. *Electron. Notes Theor. Comput. Sci.* **313**, 23–46 (2015).
33. Goldstein, B. & Perelson, A. S. Equilibrium theory for the clustering of bivalent cell surface receptors by trivalent ligands. Application to histamine release from basophils. *Biophys. J.* **45**, 1109–1123 (1984).
34. Sorger, P. K., Crowther, R. A., Finch, J. T. & Pearse, B. M. Clathrin cubes: an extreme variant of the normal cage. *J. Cell Biol.* **103**, 1213–1219 (1986).
35. Fotin, A. *et al.* Molecular model for a complete clathrin lattice from electron cryomicroscopy. *Nature* **432**, 573–579 (2004).
36. Marino, S., Hogue, I. B., Ray, C. J. & Kirschner, D. E. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J. Theor. Biol.* **254**, 178–196 (2008).
37. Muthukumar, M. & Nossal, R. Micellization model for the polymerization of clathrin baskets. *J. Chem.*

- Phys.* **139**, 121928 (2013).
38. Ungewickell, E. J. & Hinrichsen, L. Endocytosis: clathrin-mediated membrane budding. *Curr. Opin. Cell Biol.* **19**, 417–425 (2007).
 39. Wilbur, J. D. *et al.* Conformation switching of clathrin light chain regulates clathrin lattice assembly. *Dev. Cell* **18**, 854–861 (2010).
 40. Lampe, M., Vassilopoulos, S. & Merrifield, C. Clathrin coated pits, plaques and adhesion. *J. Struct. Biol.* **196**, 48–56 (2016).
 41. Nossal, R. Energetics of clathrin basket assembly. *Traffic* **2**, 138–147 (2001).
 42. Böcking, T., Aguet, F., Harrison, S. C. & Kirchhausen, T. Single-molecule analysis of a molecular disassemblase reveals the mechanism of Hsc70-driven clathrin uncoating. *Nat. Struct. Mol. Biol.* **18**, 295–301 (2011).
 43. Doherty, G. J. & McMahon, H. T. Mechanisms of endocytosis. *Annu. Rev. Biochem.* **78**, 857–902 (2009).
 44. Shraiman, B. I. On the role of assembly kinetics in determining the structure of clathrin cages. *Biophys. J.* **72**, 953–957 (1997).
 45. Saleem, M. *et al.* A balance between membrane elasticity and polymerization energy sets the shape of spherical clathrin coats. *Nat. Commun.* **6**, 6249 (2015).
 46. Ferreira, F. *et al.* Endocytosis of G protein-coupled receptors is regulated by clathrin light chain phosphorylation. *Curr. Biol.* **22**, 1361–1370 (2012).
 41. Ferreira, F. *et al.* Endocytosis of G Protein-Coupled Receptors Is Regulated by Clathrin Light Chain Phosphorylation. *Current Biology* **22**, 1361-1370 (2012).

Table 1. Ranges for parameter space exploration.

Structure	Max number
g531	2
g532	4
g541	16
g551	250
g643	2
g651	4
g661	36

Table 2. Number of pent-Rings (g5) and hex-Rings(g6) found in GSA of the “ring allowed” version of the Model 2

Parameter description	Parameter name	min	max
Association rate constant	<i>pbk</i>	10 E-6	1.00
Dissociation rate constant	<i>pdk</i>	10 E-6	1.00
Ring closure rate constant	<i>pring</i>	10 E-6	1.00
Amount of available triskelia	<i>amount</i>	10 E2	10 E4

Figure 1. Structure of agents for Model 1 and Model 2. Three identical binding sites in a simple agent (A) interact with each other to form a lattice (Model1). Monomer (B) of detailed Model 2 has two sites to form the triskelia hub (l,r) and three sites to interact with other triskelia (d, Pd,Pp)(C).

Figure 2. Simulation (5000 parameter sets) of the trivalent model with “ring forbidden” (A) and “ring allowed” (C, D) assumptions. A. The number of free agents N_{free} decreases with R_{ext} and trends to 0 at $R_{ext} = 0.5$ B. The relationship of α and R_{ext} under different experimental conditions: “no ring”, “ring” and “infinite ring” C. The dependency between the size of the largest aggregate and cyclomatic number under “ring allowed” condition D. Relationship between reaction extent and loop structure under “rings allowed”.

Figure 3. Distribution of different cyclic structures obtained from 5000 simulations. A. 5-membered rings B. 6-membered rings C. 7-membered rings D.8-membered rings.

Figure 4. Results of simulation of “ring forbidden” and “ring allowed” models. A. The number of free agents N_{free} decreases with R_{ext} towards 0 at $R_{ext} = 0.6$ for “ring forbidden” model. B. The number of rings (cyclomatic number of the graph) per triskelia in the “ring allowed” model.

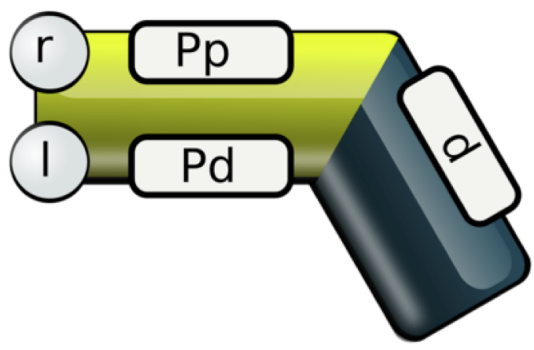
Figure 5. Most populated structures obtained in 5000 simulations of the unconstrained model. A. Most populated pentagon structure. B. Most populated hexagon structure C. An 8-ring formed by two pentagons.

Figure 6. Results of model simulation with different K_d for 5 and 6-membered rings. A. Only hexagons are allowed. B. Rate closure for 5-rings was set to infinite.

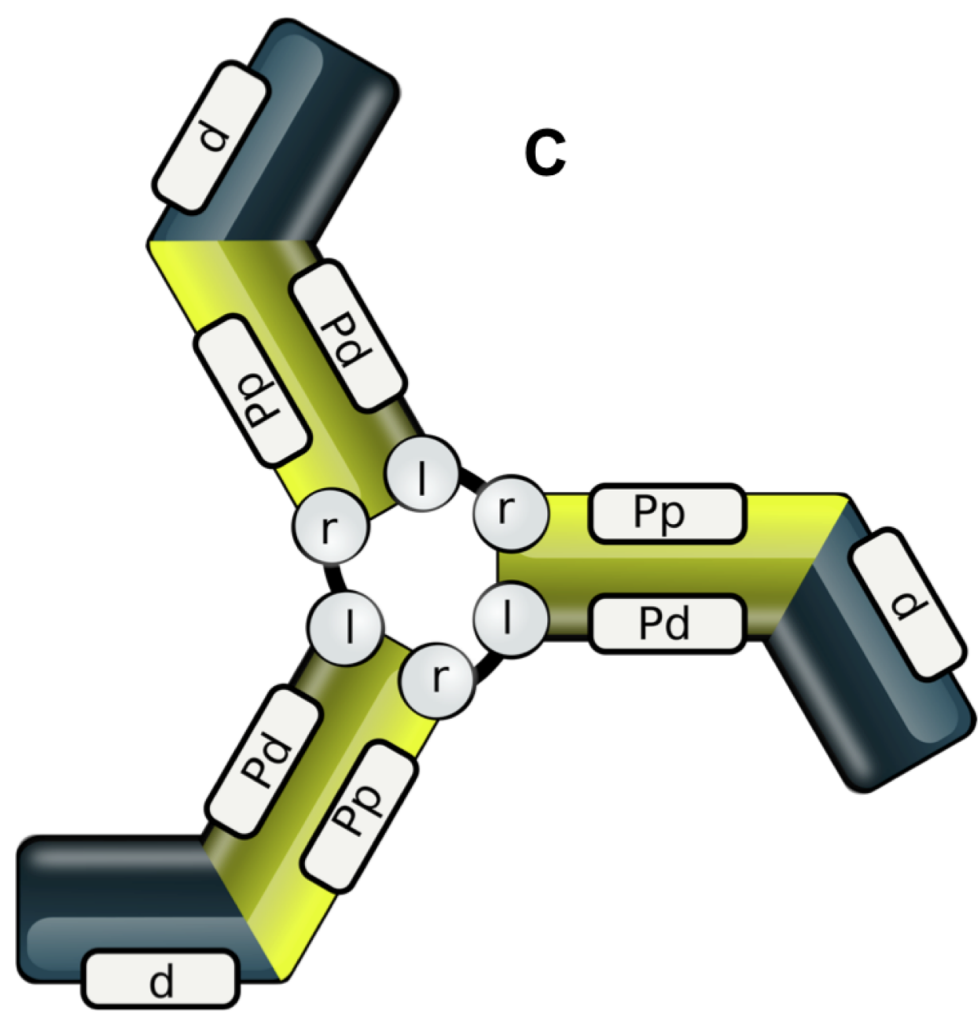
A



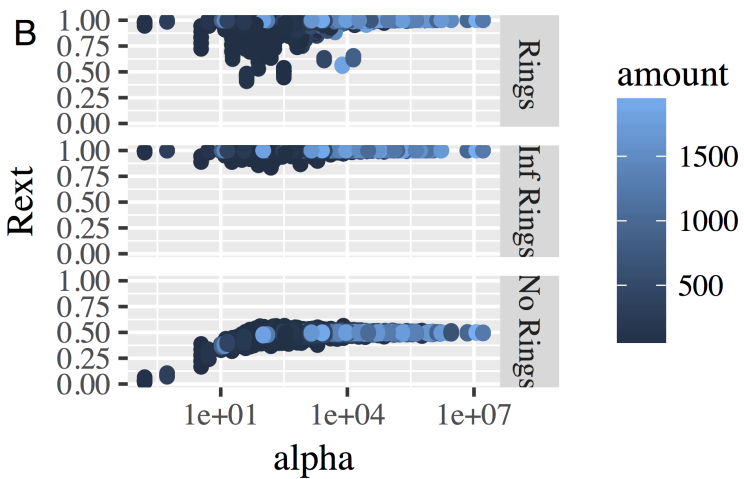
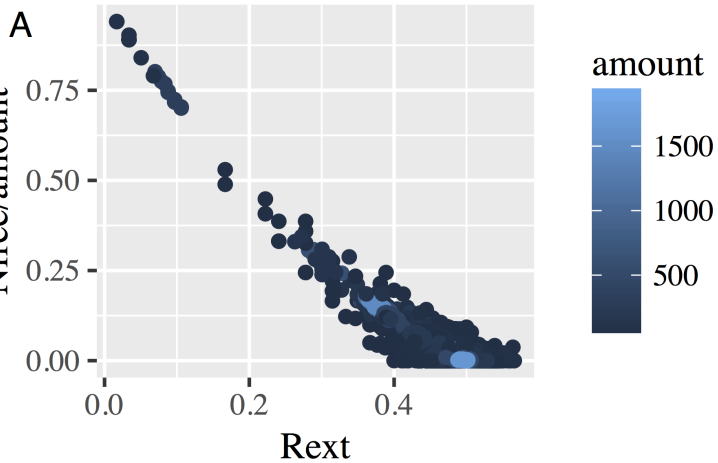
B



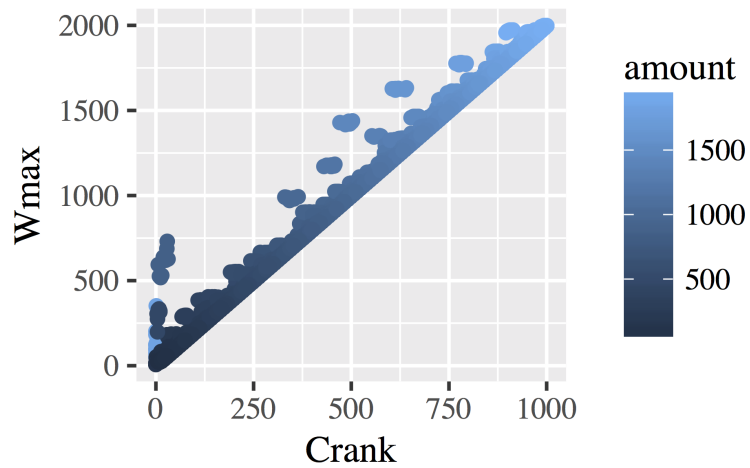
C



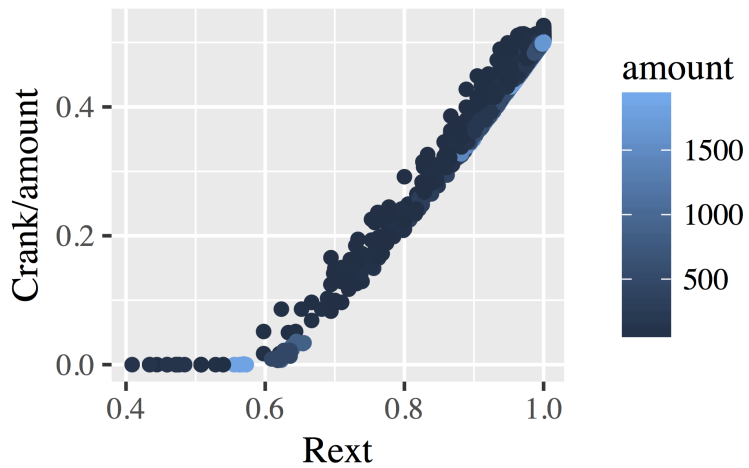
Ringless

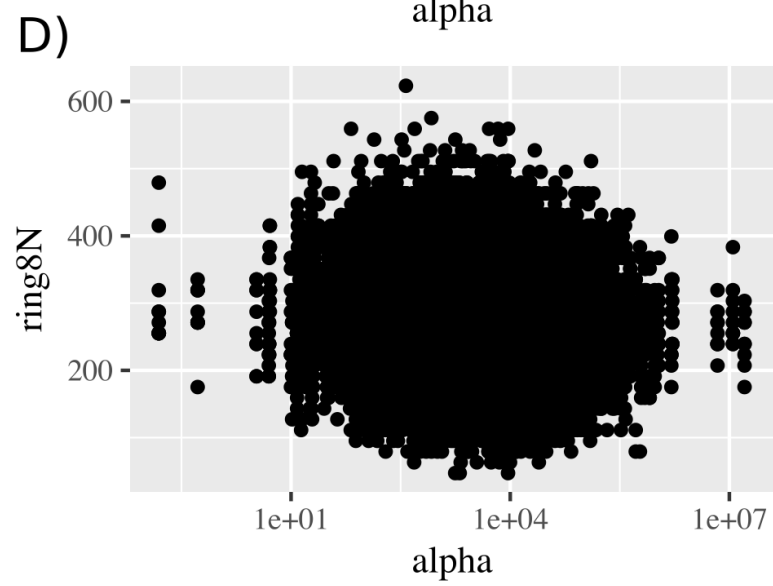
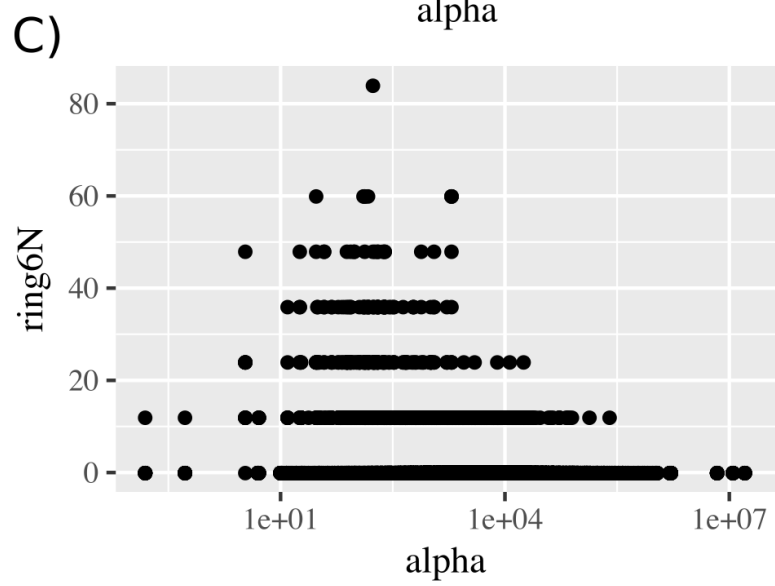
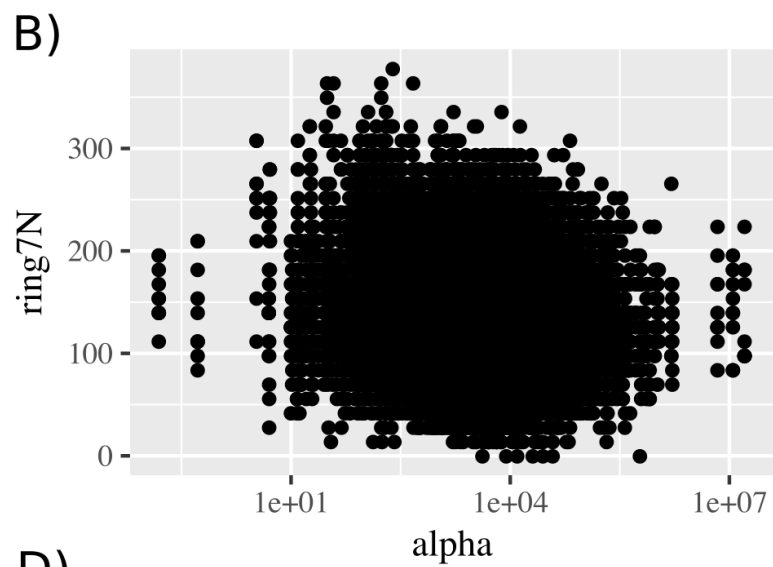
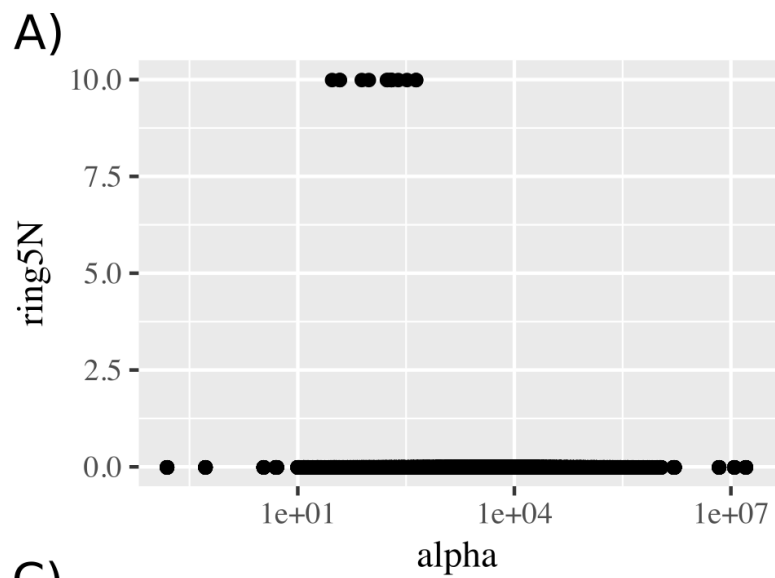


Rings

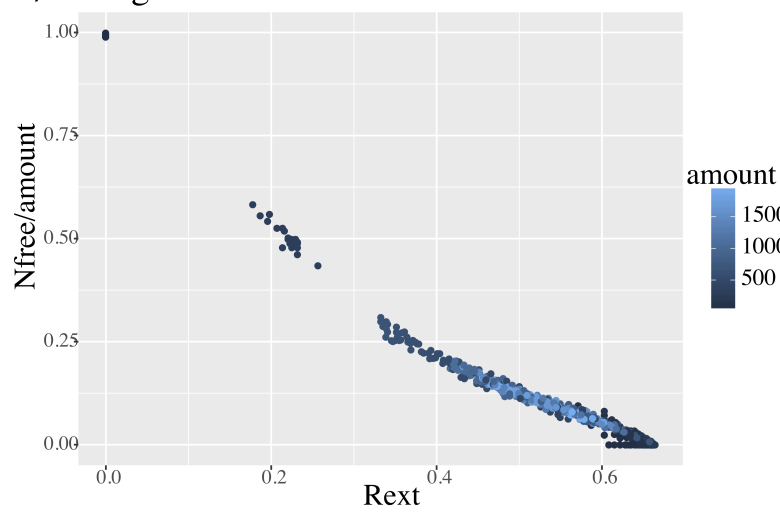


Ring

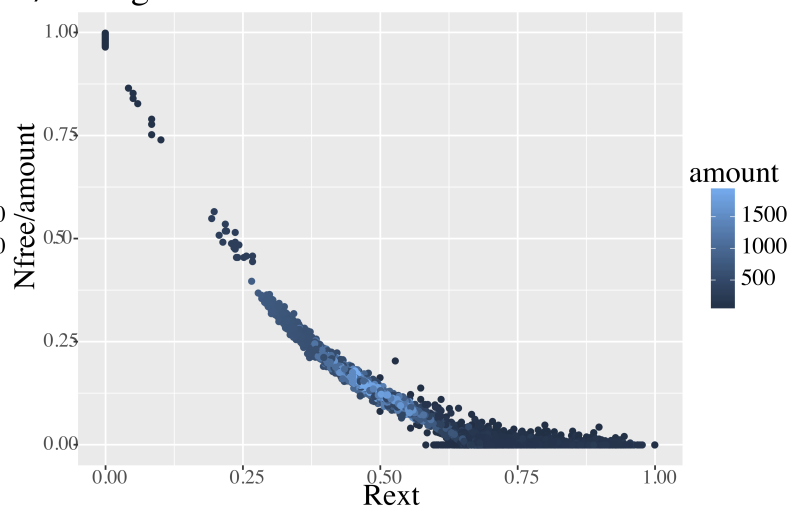




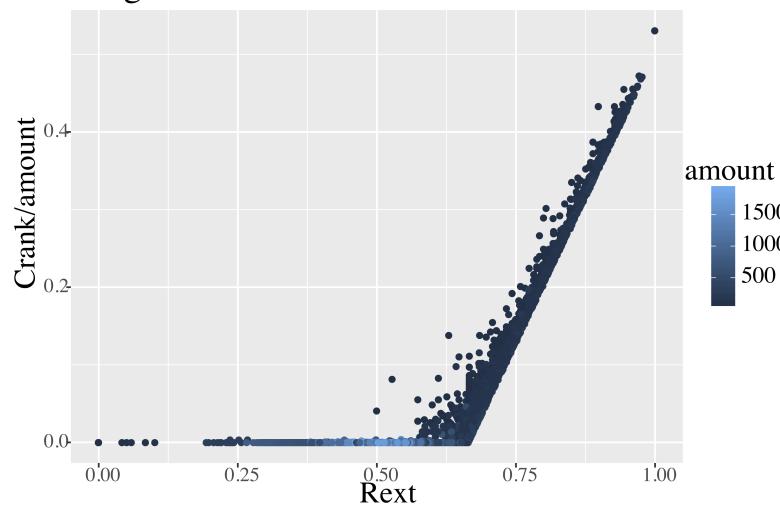
A) Ringless



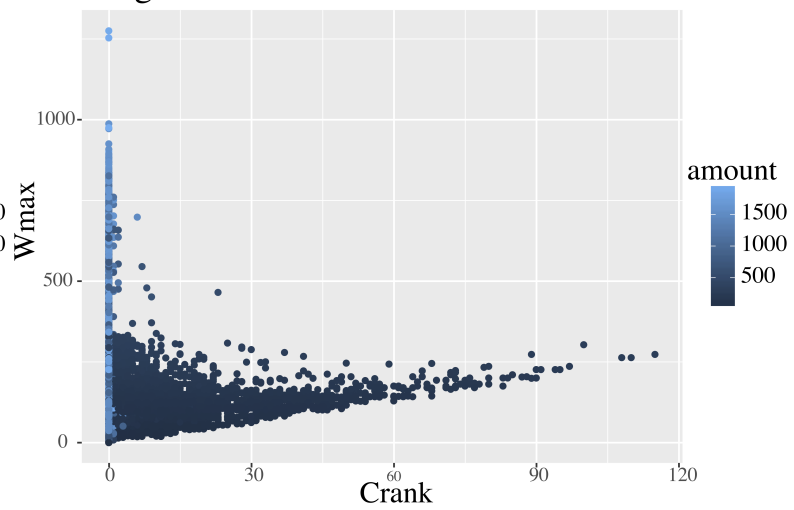
B) Rings



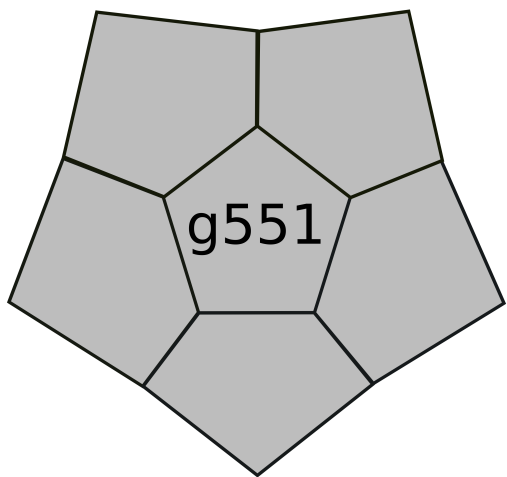
C) Ring



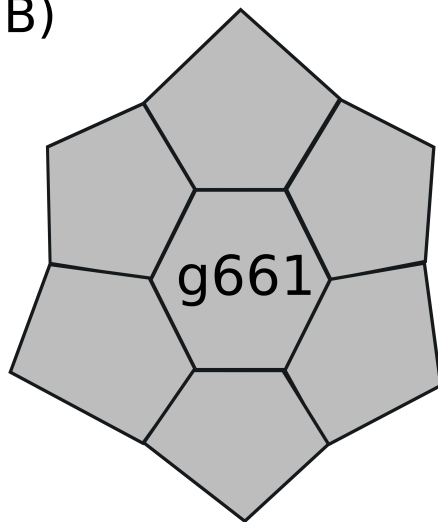
D) Rings



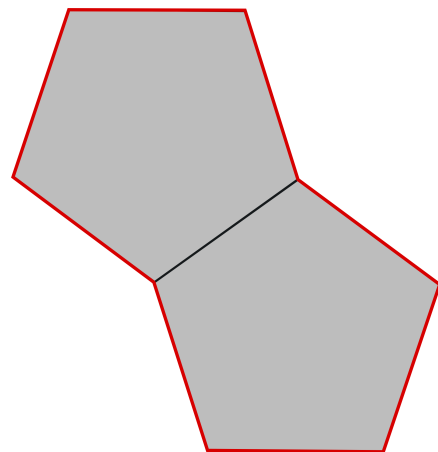
A)



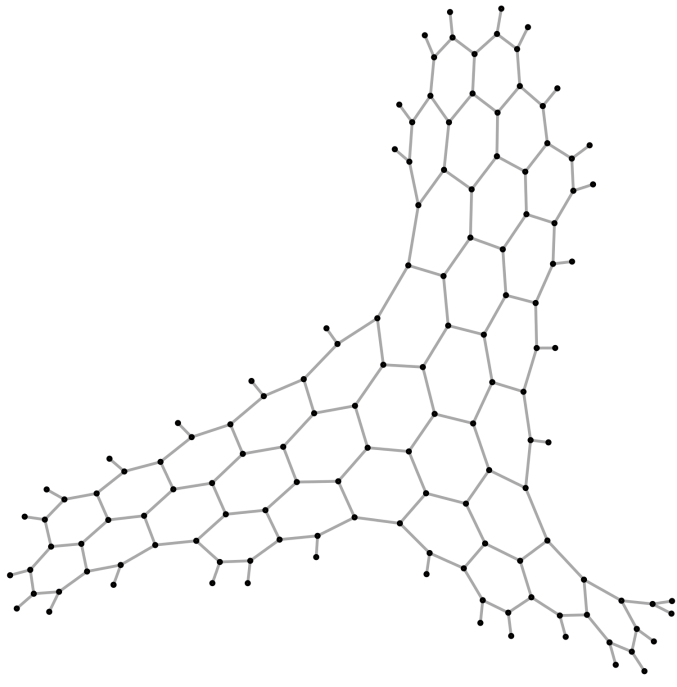
B)



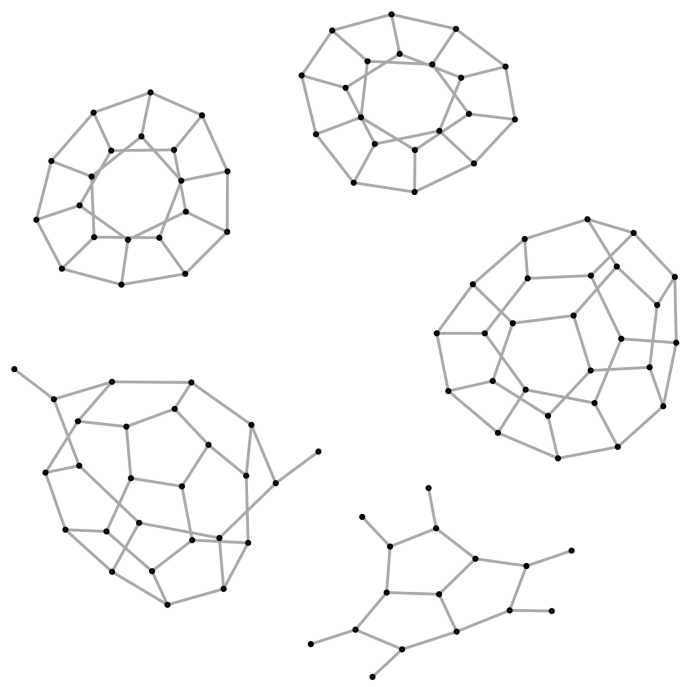
C)



A



B



Appendix H

Disease in Synaptic Models

Katharina F. Heil^{1,2*}, Emilia M. Wysocka^{1*}, Oksana Sorokina¹, Jeanette Hellgren Kotaleski², T. Ian Simpson¹, J. Douglas Armstrong¹ and David C. Sterratt¹



*These authors contributed equally to this work.

¹ - Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh

² - Computational Biology, School of Computer Science and Communication, Royal Institute of Technology (KTH), Stockholm

The Appendix and references of the draft can be found with the digital supplementary material (“Disease-in-Synaptic-Models-appendix.pdf” in folder “synaptic-review”).

Analysis of proteins in computational models of synaptic plasticity

Katharina F. Heil^{1,2}, Emilia M. Wysocka¹, Oksana Sorokina¹, Jeanette Hellgren Kotaleski², T. Ian Simpson¹, J. Douglas Armstrong¹, David C. Sterratt^{1*}

1 School of Informatics, University of Edinburgh, Edinburgh, Scotland, UK

2 Computational Science and Technology, School of Computer Science and Communication, KTH Royal Institute of Technology, Stockholm, Sweden

 These authors contributed equally to this work.

* david.c.sterratt@ed.ac.uk

Abstract

The desire to explain how synaptic plasticity arises from interactions between ions, proteins and other signalling molecules has propelled the development of biophysical models of molecular pathways in hippocampal, striatal and cerebellar synapses. The experimental data underpinning such models is typically obtained from low-throughput, hypothesis-driven experiments. We used high-throughput proteomic data and bioinformatics datasets to assess the coverage of biophysical models.

To determine which molecules have been modelled, we surveyed biophysical models of synaptic plasticity, identifying which proteins are involved in each model. We were able to map 4.2% of previously reported synaptic proteins to entities in biophysical models. Linking the modelled protein list to Gene Ontology terms shows that modelled proteins are focused on functions such as calmodulin binding, cellular responses to glucagon stimulus, G-alpha signalling and DARPP-32 events.

We cross-linked the set of modelled proteins to sets of genes associated with common neurological diseases. We found some examples of disease-associated proteins that are well represented in models, such as voltage-dependent calcium channel family (*CACNA1C*), dopamine D1 receptor, and glutamate ionotropic NMDA type 2A and 2B receptors. Many other disease-associated genes have not been included in models of synaptic plasticity, for example *COMT* and *MAOA*. To determine targets to include in future models, we incorporated pathway enrichment results, and identified *LAMTOR*, a gene uniquely associated with Schizophrenia, which is closely linked to the MAPK pathway found in some models.

Our analysis provides a map of how molecular pathways underpinning neurological diseases relate to synaptic biophysical models which can, in turn, be used to explore how these molecular events might bridge scales into cellular processes and beyond. The map illustrates disease areas where biophysical models have good coverage, as well as domain gaps that require significant further research.

Author summary

The 100 billion neurons in the human brain are connected by a billion trillion structures called synapses. Each synapse contains hundreds of different proteins. Some proteins sense the activity of the neurons connected by the synapse. Depending on what they

sense, the proteins in the synapse are rearranged and new proteins are synthesised. This changes how strongly the synapse influences its target neuron, and underlies learning and memory. Scientists build computational models to reason about the complex interactions between proteins. Here we list the proteins that have been included in computational models to date. For good reasons, models do not always specify proteins precisely, so to make the list we had to translate the names used for proteins in models to gene names, which are used to identify proteins. We found that the list of modelled proteins contains only 4.2% of proteins associated with synapses, suggesting more proteins should be added to models. We used lists of genes associated with neurological diseases to suggest proteins to include in future models.

Introduction

Activity-dependent synaptic plasticity is necessary for learning and memory [1]. Since the discovery of long term potentiation (LTP) and long term depression (LTD) [2, 3], it has been shown that synaptic plasticity can depend strongly on patterns of pre- and post-synaptic firing [4] and neuromodulators [5]. Forms of plasticity vary between types of synapses and brain region [4], which could be explained by the local proteome, i.e. the expressed proteins and their abundances. PSD-95 knock-outs demonstrate the influence of the proteome on synaptic plasticity [6]. Synaptic plasticity underlies behaviour, as evidenced by the effect of antagonising NMDA receptors [1], and synaptic proteins underlie disease [7].

Computational models of synaptic plasticity are important tools for understanding synaptic and neural function. Models at a phenomenological level, such as spike-timing dependent plasticity (STDP) models, link firing patterns in the pre- and postsynaptic neurons to changes in synaptic strength with little or no reference to the underlying molecules [8]. Biophysical models refer to at least some known molecular actors in synaptic plasticity. In 2009 there were at least 117 biophysical postsynaptic signal transduction models [9] and the number is growing [10, 11]. When they include molecular entities and phenomena they can also be used to study dysfunction, and potentially model pharmacological interventions.

Recent advances in tissue and cell extraction techniques and sample processing allow localised proteomes to be determined, e.g. the synapse including the smaller presynaptic or postsynaptic proteomes [12, 13]. Our recent analysis of 37 published synaptic proteomic datasets (in preparation; data from July 2017 in S1 Table) contains 1,867 presynaptic genes, 5,053 postsynaptic genes and 5,862 synaptic genes (with human EntrezID identifiers) respectively. These numbers are large compared to results from individual studies. Nevertheless, data inclusion was highly restrictive and the augmented numbers can be partly explained by higher experimental sensitivity and the broad use of high-throughput techniques.

These synaptic protein lists make it possible to compare systematically proteins contained in computational models of synapses with those proteins likely to be in the synapse. In this paper we: (1) survey a selection of biophysical models of synaptic plasticity, identifying which proteins are involved in each model, and describing the complexity and detail of description of signalling pathways within the models; (2) compare the proteins in models with synaptic protein lists, thus showing what fraction of synaptic proteins have been considered in models; (3) identify the functional classes of proteins in models; and (4) compare the proteins in models with those involved in neurological diseases. Clearly the coverage of synaptic molecules found in the existing ‘model space’ is going to be very incomplete given the intense amount of effort required to develop each model but here we sought to explore systematically molecular coverage to identify significant gaps that might offer new opportunities.

Analysis of proteins in synaptic models

Before outlining our analysis, we first address a fundamental issue we encountered. Computational models contain a diverse cast of players, including proteins, second messengers, reporters, ions and others. Models vary in how precisely they specify proteins; for example Bhalla and Iyengar [14] specify AC1, AC2 and AC8, whereas Castellani et al. [15] and Oliveira et al. [16] specify AC, which could, in principle, map to any of the adenylate cyclases expressed in the synapse. This presents a problem when mapping models to molecular identifiers, which we addressed by developing a mapping from what we refer to as model “entities” to gene families. For example a protein such as Calmodulin 1 can be mapped onto a single gene (*CALM1*), but a family of proteins such as metabotropic glutamate receptors maps onto more than one gene (*GRM1-GRM8*). By definition, second messengers or ions do not map onto gene symbols.

The concept of entities allowed each model’s constituents to be catalogued faithfully and then mapped onto identifiers according to the steps shown in Fig 1: (1) select models to analyse; (2) determine all entities (e.g. proteins, protein multimers or families, ions and second messengers) that are contained in each model; (3) map these entities onto gene identifiers and higher level families; and (4) use the lists of entities in each model and the mappings to undertake comparative analyses. These analyses include: comparison of modelled proteins with pre- and postsynaptic proteomic datasets; identification of properties of modelled genes, in particular cellular pathways, gene ontology terms and disease; and comparison of models with each other.

Selection of models

We selected 30 published computational, biophysical models of plasticity or related pathways in hippocampal, striatal, cerebellar or generic synapses (Table 1). Models that we regarded as phenomenological or descriptive, i.e. models describing a function with no explicit reference to an underlying mechanism, were excluded. For example, models of spike-timing dependent synaptic plasticity are phenomenological, since they contain an empirical function that maps spike times onto changes in plasticity with no reference to proteins.

The process of identifying the model constituents can be time-consuming, especially when machine-readable descriptions are not available. In order to address our questions regarding the molecular coverage of synaptic models, it sufficed to select a set of models that we were reasonably confident gave good coverage of modelled proteins, rather than to identify entities in all published models. We assessed molecular coverage of pre-2010 models from the tables in Manninen et al. [9] and we screened models published between 2010 and December 31st 2015.

Sources of models

A number of the models we selected are written in standardised modelling languages and hosted in large scale repositories such as ModelDB [44], BioModels [45], DOQCS [46] and the CellML repository [47]. ModelDB is a curated database of computational neuroscience models at the molecular and electrophysiological levels, written in a number of languages. BioModels hosts models which focus on biochemical and cellular systems at the physiological and biochemical levels, unrestricted by the biological subject [45,48]. In the curated branch of BioModels, models have to be annotated according to the Minimal Information Requested in the Annotation of Biochemical Models (MIRIAM) standard [49], thus meaning that model constituents are mapped to external identifiers. CellML is both a model format and a repository.

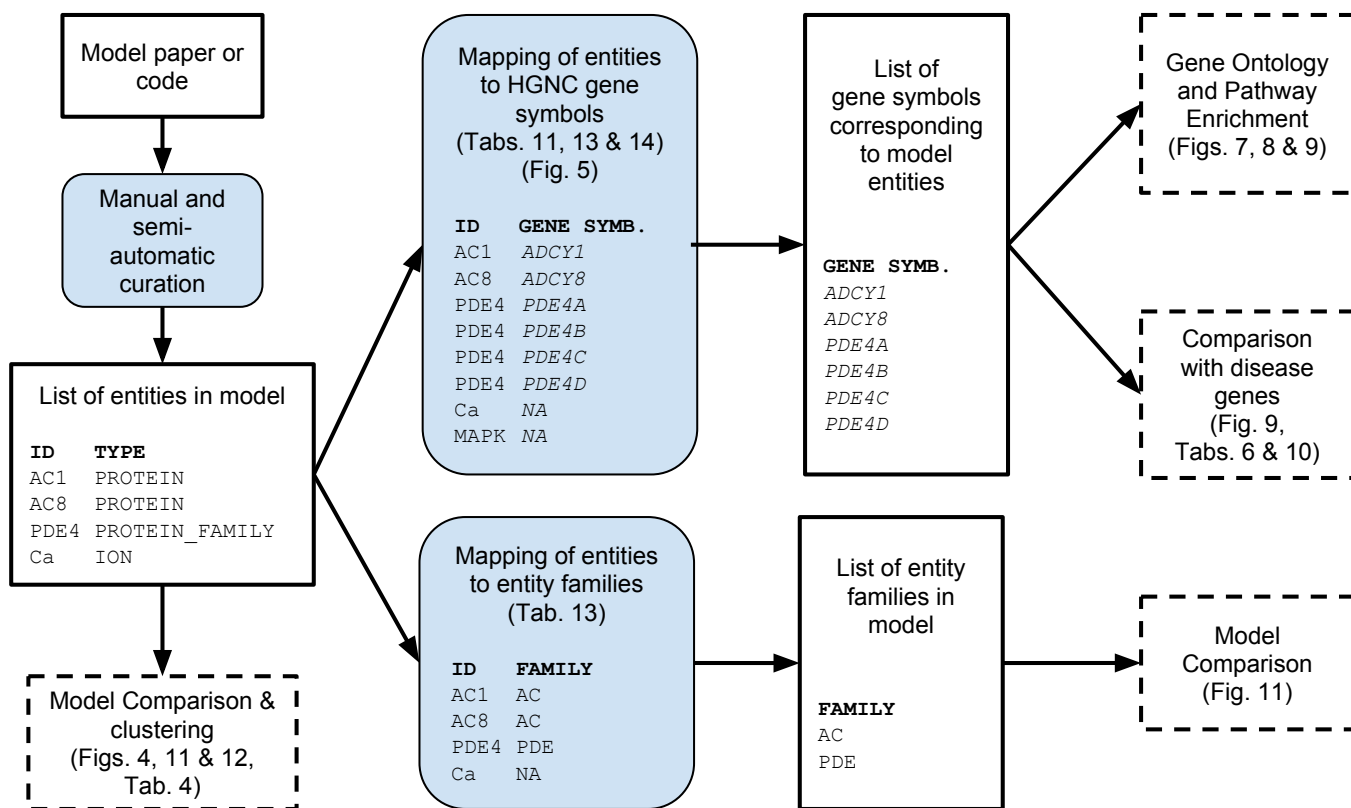


Fig 1. Overview of the modelling paper analysis process. Sets of data are shown in boxes with black rectangular borders. Processes are shown in boxes with blue backgrounds and curved corners. Final analyses are shown in boxes with dashed borders. “ID” refers to the modelled entity. Boldface type refers to column headers.

The CellML repository hosts a wide range of biological models, which have documentation pages generated from the meta-data supplied by model authors. DOQCS (Database of Quantitative Cell Signalling) is a database tailored for storing chemical kinetics and reaction level information [46]. The chemical-level description of each model corresponds to the GENESIS/Kinetikit simulator and reflects reaction diagrams or ordinary differential equation (ODE) equations.

Table 2 summarises the numbers of models we analysed that are stored in repositories and other locations, and the format of the model descriptions. Three of the 7 models deposited in the BioModels database were curated to MIRIAM standards. Around half of all catalogued models (14) had non-machine readable descriptions. Models in this group were generally difficult to explore and extracting information from them proved challenging. There were 18 machine-readable models available from publication attachments, on institute or lab servers and the four public modelling databases; some models were deposited in more than one database. With two exceptions models were not duplicated in ModelDB and BioModels; the Bhalla and Iyengar [14] model was present in all four public modelling databases, and the Nakano

Table 1. Summary of models.

Paper	Vars./comp.	Entities	Vars./comp./ Entities	Region
Antunes and De Schutter (2012) [17]	103	19	5.4	Cereb. Purk.
Antunes et al. (2016) [18]		17		Cereb. Purk.
Bhalla and Iyengar (1999) [14]	100	42	2.4	Hipp. CA1 Pyr.
Byrne et al. (2009) [19]	82	3	27.3	Hipp. CA1 Pyr.
Castellani et al. (2001) [20]	36	5	7.2	Cortex**
Castellani et al. (2005) [15]	33	13	2.5	Ex. glut. syn.**
Graupner and Brunel (2007) [21]	16	5	3.2	Hipp. CA1 Pyr.
Gutierrez-Arenas et al. (2014) [22]	188	34	5.5	Striatial MSPN, D1R expressing
Hernjak et al. (2005) [23]	9	5	1.8	Cereb. Purk.
Khan et al. (2011) [24]	12	1	12.0	Hipp. CA1 Pyr.
Kim et al. (2010) [25]	54	18	3.0	Hipp. CA1 Pyr.
Kim et al. (2011) [26]	16	17	1.0	Hipp. CA1 Pyr.
Kim et al. (2013) [27]	10	18	0.6	Striatial MSPN, mGluR1 expressing
Kötter (1994) [28]		12		striatal MSPN
Kuroda et al. (2001) [29]		20		Cereb. Purk.
Li et al. (2012) [30]	95	8	11.9	Generic excitatory spine
Mattioni and Le Novère (2013) [31]	13	9	1.4	Striatial MSPN
Miller et al. (2005) [32]	58	4	14.5	**
Nair et al. (2015) [33]	80	16	5.0	Striatial MSPN, D1R and D2R expressing*
Nakano et al. (2010) [34]	189	28	6.8	Striatial MSPN, D1R expressing
Oliveira et al. (2010) [16]	31	9	3.4	HEK293 cells
Oliveira et al. (2012) [35]	113	28	4.0	Striatial MSPN
Pepke et al. (2010) [36]	156	3	52.0	**
Qi et al. (2010) [37]	115	13	8.8	Striatial MSPN
Smolen et al. (2006) [38]	23	9	2.6	Hipp. CA1 Pyr.
Smolen et al. (2012) [39]	14	6	2.4	Hipp. CA1 Pyr.
Sorokina et al. (2011) [40]	1,000,000	55	18,181.8	Ext. glut. syn.
Stefan et al. (2008) [41]	49	3	16.3	**
Zeng and Holmes (2010) [42]	14,296,081	6	2,382,680.2	Hipp. DG
Zhabotinsky et al. (2006) [43]	58	11	5.3	Hipp. CA1 Pyr.

“Paper” refers to the analysed model. “Vars/comp.” is the number of molecular variables per compartment, a measure of the complexity of the model; this was not assessed for all papers. “Entities” is the number of entities in the model, and “Vars./Enties” is the ratio between the number of variables per compartment and the number of entities. This roughly corresponds to the level of detail of the model. “Region” refers to the brain region or cell type where the model is situated (** – no cell specified). Abbreviation: Cereb. Purk., cerebellar Purkinje cell; Ex. glut. syn., excitatory glutamatergic synapse; Hipp. CA1 Pyr., hippocampal CA1 pyramidal cells; Hipp. DG, hippocampal dentate gyrus cell; MSPN, medium spiny projection neuron; * – denotes that there is more than one model presented in a study and numbers in this table refer to the one with the larger number of “Entities”.

et al. [34] model was found in ModelDB and BioModels. We did not test the functionality or reproducibility of models; only the availability and relative ease of exploration were examined.

105
106
107

Table 2. Overview of locations of models and their formats.

Type	Location	Format	Fraction	
non-machine-readable	attached to publication	descriptions in appendices, text files, spreadsheets, reaction diagrams, equations	14/30	
	or within publication content			
machine-readable	attached to publication	software-specific	3/30	
	institutional/lab servers		3/30	
	public modelling databases	ModelDB	any of: NEURON, Python, C, C++, GENESIS, Java, Matlab, XPP, etc.	8/30
		BioModels	all of (automatically translated): SBML, CellML, VCML, XPP, SciLab, Octave, BioPAX	7/30
		CellML	CellML	1/30
	DOQCS	GENESIS	2/30	

Fractions refer to the number of models in the category relative to the total of annotated models. Each machine-readable model can be part of several categories. See text for details.

Features of models

We extracted a number of features from each model to highlight their similarities and differences (Table 1). To quantify the model size, we counted the number of entities that appear in the model. We also extracted information on numbers of dynamic variables per compartment (“Vars/comp.”). Variables are values describing quantities that change in the model. A compartment is defined as a spatial subsection within the model. Since the number of compartments varies with the fineness of the spatial mesh used, the number of variables scales with the number of compartments, but the number of variables per compartment will be a constant, independent of the spatial discretisation used to simulate the model. To provide a measure of model complexity, we used the ratio of the number of variables per compartment and the number of entities (“Vars./Comp./Entities”, Table 1).

For example, in a model of calcium binding to a buffer in a single compartment, there are two entities: calcium (an ion) and the buffer (a protein). There are three variables, namely the concentrations of free calcium, free buffer and calcium-buffer complex. To model diffusion of calcium, buffer and calcium-buffer complex, space could be divided into 100 compartments. The number of variables would then be 300, but the number of variables per compartment would be 3. There would still only be two entities in this model – calcium and the buffer – and the variables per compartment per entity ratio would be 1.5.

A high ratio of variables per compartment to entities reflects a detailed description of a small pathway. For example the model of Byrne et al. [19] – whose stochastic model describes binding of calcium, calmodulin (CaM) and calcium/calmodulin dependent kinase II (CaMKII) – has 82 variables per compartment and 3 entities, making a ratio of 27.3. The 82 variables correspond to the combinations of calcium bound to the N and C lobes of calmodulin and whether or not these complexes are bound to CaMKII. Dealing with this complexity in the simulation is achieved by using an agent-based

Table 3. Frequency of entity types found in models.

Type	Frequency	Examples
Ion	2	Magnesium, Calcium
Neurotransmitter	5	Adenosine, Dopamine
Others	2	ATP and PIP2, intermediates in the IP3/DAG pathway
Protein	95	Neurogranin
Protein family	52	calmodulin, which may correspond to one of calmodulin-1, calmodulin-2 or calmodulin-3
Protein multimer	8	AMPA receptor, which comprises a tetramer of GluR1, GluR2, GluR3 and GluR4 proteins.
Reporter	1	AKAR3
Second messenger	8	GTP (Guanosine triphosphate) or cAMP (cyclic AMP).
Total	173	

Gillespie method in which the states of individual molecules rather than populations of molecules are followed through the simulation [50]. Agent-based simulation also allows the more extreme example of Zeng and Holmes [42], who modelled the Ca^{2+} -CaM-CaMKII pathway and the binding of Ca^{2+} -CaM to calcineurin, which for consistency with genetic nomenclature we refer to as PP3 rather than PP2B (see Discussion). Along with calbindin and neurogranin, the model has 6 entities in total and 14,296,081 possible states (i.e. variables), making a ratio of 2,382,680 variables per compartment per entity. In this case the large number of states arises because each of the 6 subunits of CaMKII can be in one of 21 states, which gives rise to 14,296,051 configurations according to the necklace function [51]. Notes on this and other calculations are contained in the “Model classification” spreadsheet in S1 File.

At the other end of the spectrum, a low variable to entity ratio indicates larger pathways with each interaction modelled in less detail. For example, the ODE-based model of Bhalla and Iyengar [14], with 44 entities and approximately 100 variables per compartment, has a ratio of 2.3 variables per compartment per entity.

Identifying entities in models

To identify the entities in each model, the publication describing the model and, if available, an electronic description of the model were examined by one of the authors. For each entity, we recorded the name used in the model publication and our standard entity identifier. Models do not always specify the entities involved precisely. We discussed ambiguous cases together and erred on the side of not imputing the identity of a protein; for example a “Plasticity related protein” [39] was not mapped to an entity identifier.

We identified 178 distinct entities across the 30 catalogued models (see S2 Table for full list). As well as an identifier, each entity has a long name and a type which can be one of: “ion”, “neurotransmitter”, “others”, “protein”, “protein family”, “protein multimer”, “reporter” or “second messenger”. Table 3 shows how many of each type of entity were identified, and gives examples. The most frequent entity type is “protein”, followed by “protein family” and then “protein multimer”.

The rationale for having three protein types – “proteins”, “protein families” and “protein multimers” – was to allow us to record as precisely as possible what was meant in each computational model. A “protein” is a specific protein e.g. neurogranin, encoded by a specific gene (*NRGN*), so it is unambiguous as to which gene is implied by the model. The same gene may produce multiple isoforms due to gene duplicates or alternate splicing. For example *PRKCZ* produces two isoforms, atypical protein kinase

C, ζ (PKC ζ) and autonomously active isoform of atypical protein kinase C, ζ (PKM ζ) [52]. A “protein multimer” is a multiprotein complex, e.g. an α -amino-3-hydroxy-5-methyl-4-isoxalone propionic acid receptor (AMPA), which comprises a tetramer of a selection of GluR1, GluR2, GluR3 and GluR4 proteins. In this example, if the model only specified “AMPA” there would be ambiguity about which of the GluR1–4 subunits are implied by the model. Coding AMPA as a “protein multimer” allows this ambiguity to be recorded and resolved as desired. A “protein family” is a protein from a family of proteins, e.g. calmodulin, which may correspond to one of calmodulin-1, calmodulin-2 or calmodulin-3. Again, it is not clear which protein is implied by the model, though later we will use information about the synaptic proteome to narrow down the possibilities.

“Ions”, “neurotransmitters” and “second messengers” were assigned to individual classes. AKAR3 is the only entity that was classified as a “reporter” [33]. The FLIM-AKAR reporter was included in the model to reflect the experimental setup where it is used to measure PKA dynamics. ATP and PIP2, both intermediates in the IP3/DAG pathway were classified as “other”. ATP itself can produce a second messenger and is often referred to as a precursor or “coenzyme”. Similarly, PIP2 is frequently acting as a precursor of a second messenger [27].

The full catalogue of all model entities for all models is shown as a binary matrix in Fig 2. The models are ordered according to the results of Ward’s 2D hierarchical clustering applied to the matrix (as implemented in R’s *hclust* function with the *Ward.2D* method). This catalogue is the basis for the rest of the analysis.

Mapping entities to gene identifiers

In order to compare synaptic models with the synaptic proteome, we needed to map each protein entity onto the proteins to which it might correspond. The construction of this mapping is shown in Fig 3. Based on common practice in bioinformatics we decided to use HUGO Gene Nomenclature Committee (HGNC) gene symbols and NCBI Entrez Gene IDs to identify proteins/genes. The one-to-one mapping from HGNC gene symbols to NCBI human Entrez Gene IDs [53] allowed this approach.

As presented in Fig 3, entities of type “protein” were mapped directly to HGNC gene symbols. Entities classified as “protein family” and “protein multimer” required an intermediate mapping step. We searched for ontologies that could be used to identify as many of these entities as possible and map them to HGNC gene symbols. After thorough analysis of available bioinformatic resources (see Methods) we decided to use HGNC gene families to map entities of type “protein family” and “protein multimer” to genes. For each such entity, we tried to identify a corresponding HGNC gene family, and used manual NCBI mapping (see Methods) to check if the genes contained in this family seemed likely to be what was meant in the models. For example, we mapped the entity “Dopamine receptors” (DRD) to the HGNC family “Dopamine receptors”, which contains the genes *DRD1*, *DRD2*, *DRD3*, *DRD4* and *DRD5*. Since this seemed a reasonable set, we accepted the mapping.

For some entities no one HGNC family gave a reasonable set of proteins, but the intersection between two or more families did. For example the genes corresponding to SHANK, by which we mean the family of proteins encoded by *SHANK1*, *SHANK2* and *SHANK3*, may be selected from the gene families list by choosing all genes that are in the “Ankyrin repeat domain containing” (ANKRD) and “PDZ domain containing” (PDZ) gene families. When we could not find a corresponding HGNC family or a combination of HGNC families, we constructed our own mapping (see Methods). Since “ions”, “neurotransmitters”, “others”, “reporters” and “second messengers” are not proteins, they were by definition excluded from the mapping to gene names.

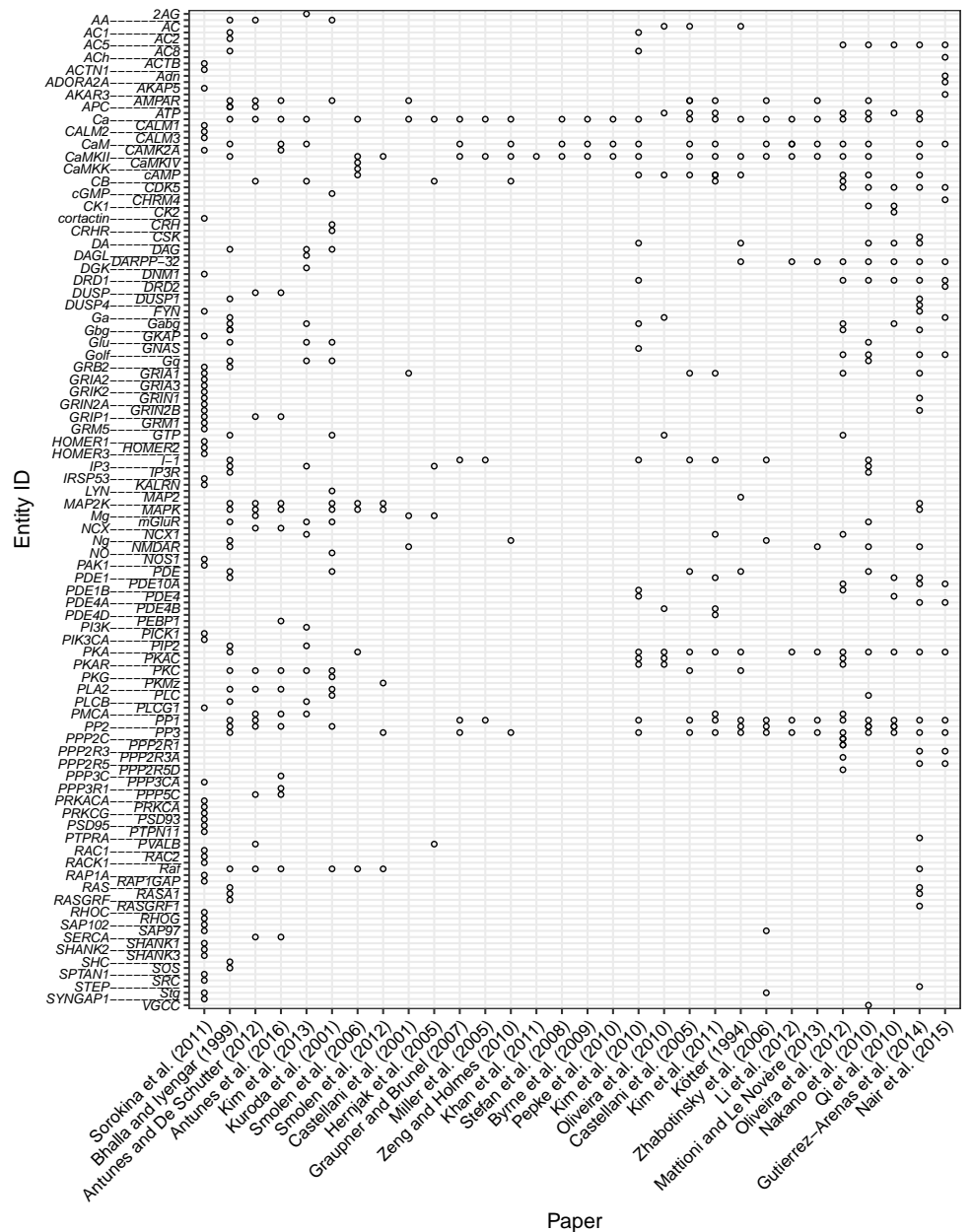


Fig 2. Matrix of entities in models. The occurrence of an entity in a model is indicated by open circles. Entity IDs are staggered for readability.

Once gene families corresponding to 61 "protein families" and "protein multimers" were identified we could map each family or multimer onto a set of genes (S3 Table and S4 Table). 331 unique HGNC gene symbols were identified based on protein families and multimers. The union of this set of symbols with the 96 genes mapped directly from type "protein" forms the "full set of HGNC gene symbols in models" dataset, which contains a total of 386 HGNC gene symbols. A number of "protein families" mapped onto the same genes; for example the families PDE and PDE1 both contain *PDE1A* and *PDE1B*.

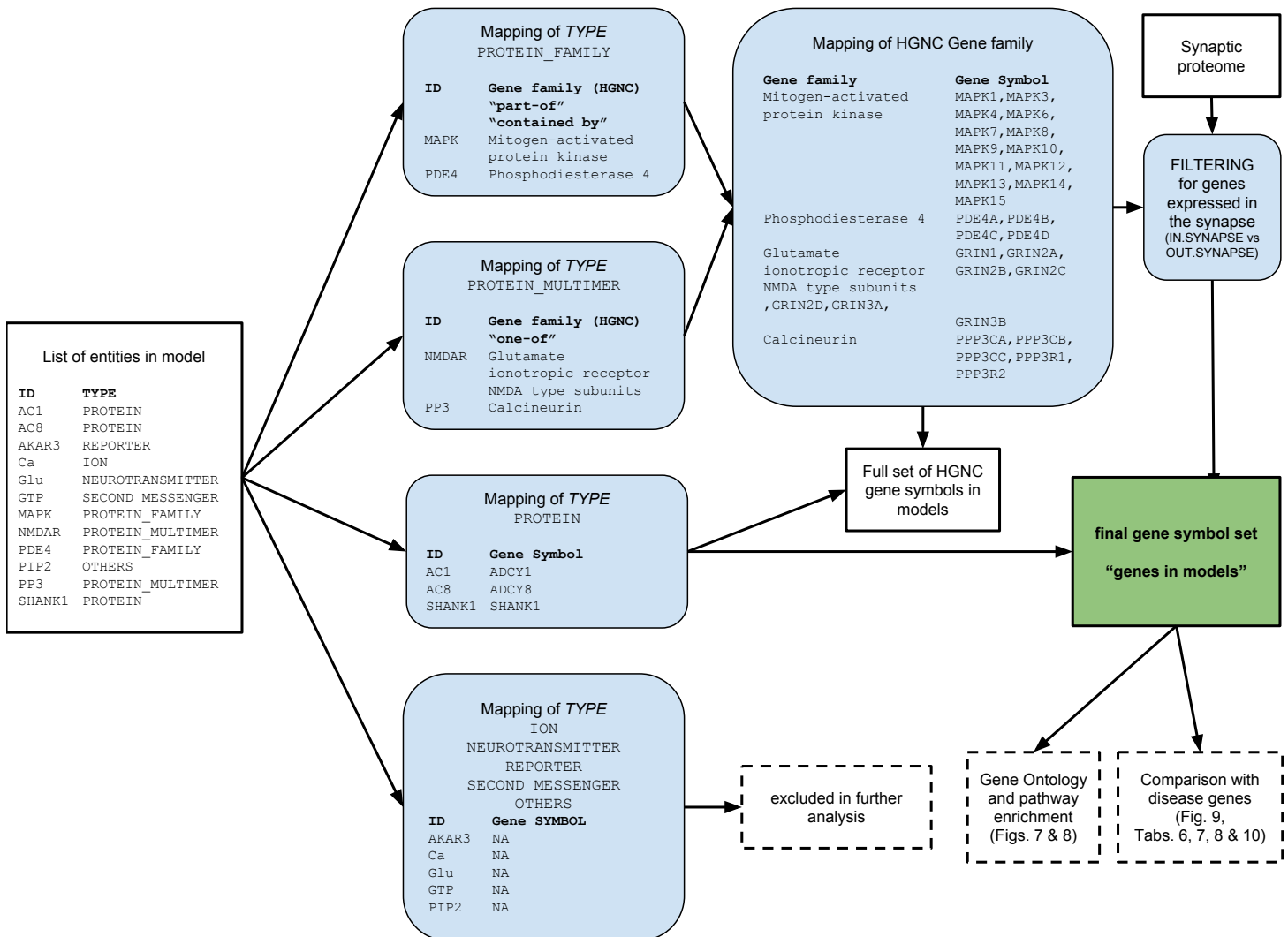


Fig 3. Overview of entity to Gene Symbol mapping process. Sets of data are shown in boxes with black rectangular borders. Mappings are shown in boxes with blue backgrounds and curved corners. Dashed lines indicate additional information, and the key outcome is highlighted in a box with green background. Bold font refers to column headers.

Comparison with proteomic data

HGNC families are general gene classes and do not contain information about tissue specificity or expression patterns. To identify proteins found in the synapse, we used a meta-analysis of published proteomic datasets of the presynapse, postsynapse and synaptosome (in preparation). The individual references, as of July 2017, can be found in S1 Table.

The synaptosome is the largest data subset extracted from brain homogenate. The term synaptosome refers to the complete presynaptic terminal including mitochondria, synaptic vesicles and the postsynaptic membrane together with the postsynaptic density (PSD) [54,55]. The PSD is a tightly connected, dense region of the postsynaptic membrane which hosts a number of different receptors and regulatory units. The presynapse and postsynapse are subsets of the synaptosome, and can be separated

through experimental steps [56, 57].

The union of these three datasets, which we refer to as the “synaptic proteome”, comprises 6,706 genes and is based on data obtained from 37 publications and 39 datasets (data as of July 2017). The extracted proteome was used to filter the “full set of HGNC Gene symbols in models” (see Fig 3 and Identifying entities in models). We found that every “protein family” (S3 Table) and “protein multimer” (S4 Table) in our list contains at least one gene overlapping with the synaptic proteome. Genes not expressed in the synapse (“OUT SYNAPSE” in S3 Table and S4 Table) were excluded from further analysis. This filtering step reduces the 331 genes in families to 239 HGNC gene symbols. Together with directly mapped proteins this leaves 294 unique HGNC gene symbols describing all mapped genes in models, where families and multimers were screened for the presence in the synapse. From now on we refer to this gene set as “genes in models” (see green box, Fig 3).

The overlap between the final set of “genes in models” and the synaptic proteome, as well as its subsets (presynaptic, postsynaptic, and synaptosome), is visualised in the Venn diagram in Fig 4. It can be seen that 46% of “genes in models” (135 genes) are found in all three synaptic proteome datasets. Significantly lower numbers are expressed in individual sub-datasets. These are 3, 14 and 21 genes for the presynapse, postsynapse and synaptosome respectively (representing 1.0%, 4.7% and 7.1% of genes in models). When disregarding “genes in models” present in the intersection of all three datasets, more modelled genes are found in the postsynapse or synaptosome (143 genes) than the presynapse or synaptosome (27 genes). Thus, postsynaptic genes appear to be the most highly modelled subset. However, relative to the total size of the respective proteomes, only 5.1% of postsynaptic genes (258 “genes in models” out of 5,053 postsynaptic genes) versus 7.6% of presynaptic genes (142 “genes in models” out of 1,867 presynaptic genes) are represented in the models.

Nine modelled genes, all of type “protein” are not present in the synaptic proteome datasets (see lower right of the circle in Fig 4). Further investigation uncovered evidence for all of them being expressed in the synapse (Table 4), so these 9 genes remained in the set of “genes in models”. These cases illustrate that, despite the number of proteins found in recent publications, proteomic datasets are still incomplete.

Table 4. Proteins in models and not to be found in synaptic datasets.

Entity ID	Gene	Reason for inclusion
ADORA2A	<i>ADORA2A</i>	Adenosine A2a receptors (A2aR) are expressed with D2R receptors [33]
CALM2	<i>CALM2</i>	Unpublished dataset
CHRM4	<i>CHRM4</i>	Muscarinic cholinergic receptor shown to be expressed in gonadotropin releasing hormone neurons [58]
CRH	<i>CRH</i>	Corticotropin-releasing factor, regulating the release of adrenocorticotropin in synapses [59]
DRD1	<i>DRD1</i>	D1 subtype of the G-protein coupled dopamine receptor – the most abundant in the central nervous system; presence in neurons confirmed [60]
DRD2	<i>DRD2</i>	D2 subtype of the G-protein coupled dopamine receptor; presence in neurons confirmed [60]
DUSP1	<i>DUSP1</i>	Model specifies that DUSP1 feedback loop occurs in the dendritic shaft, the soma and the nucleus [22]
I-1	<i>PPP1R1A</i>	Unpublished dataset
PPP2R3A	<i>PPP2R3A</i>	Preliminary studies suggest PPP2R3A is present in both cytoplasm and nucleus of cells in the striatum [61]. PPP2R3A mediates Ca ²⁺ -dependent dephosphorylation at Thr-75 of DARPP-32 [61].

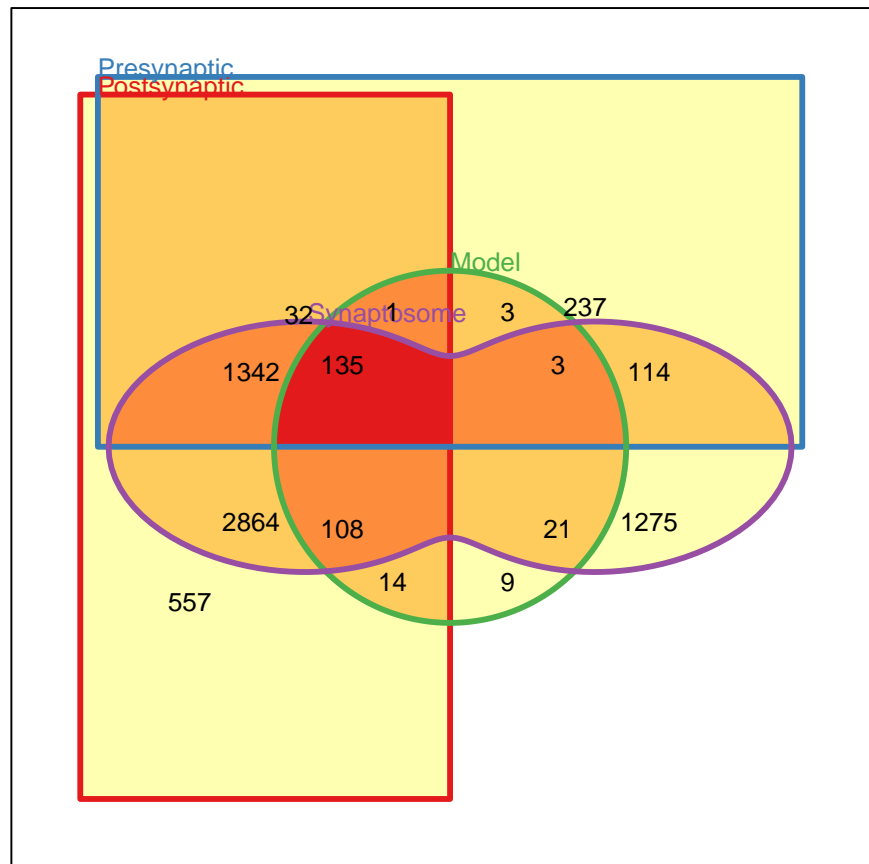


Fig 4. Relationships between the sets of genes in postsynaptic, presynaptic, synaptosome datasets and the sets of genes possibly present in models.

Postsynaptic genes in red, presynaptic in blue, the synaptosome in purple and genes in models in green. Numbers refer to the number of genes in each subset and shading shows how many sets a region belongs to (white – none; red – all four). It can be seen that the number of genes in the proteome but not included in models is an order of magnitude bigger than the number of proteins included in models and the proteomic datasets. There are only 9 genes (listed in Table 4) found in models and none of the proteomic datasets.

Enrichment analysis of modelled genes

After compiling the “genes in models” list, we related it to existing biological knowledge, in the form of gene sets annotated with various biological categories, supplied through a number of databases. Depending on each database’s focus, structured, controlled, and descriptive terms are associated to each gene. In this study, we chose to use the following ontologies: Gene Ontology (GO) [62], REACTOME Pathway Database (REACTOME) [63] and Disease Ontology (DO) [64]. Amongst these GO is the largest and most commonly used ontology, classifying genes within domains including Molecular Function, Biological Process and Cellular Compartment. We also used REACTOME, a free and manually curated database in which genes are tagged with terms representing biochemical reactions and pathways they are involved in. A

271

272

273

274

275

276

277

278

279

280

281

the top enriched pathways. The first two terms are parallel to each other on the pathway hierarchy and have a common parent term of “GPCR downstream signalling”. A comparison of the remaining members of this pathway with the enrichment results shows that they are all significantly enriched in terms of our “genes in models”. The identification of signalling pathways highlights a focus of the analysed models indicating the central role of G-protein signalling.



Fig 6. REACTOME enrichment analysis results for “genes in models”. The synaptic proteome was used as background dataset. The list of significant terms was obtained with the Fisher’s exact test and the *elim* algorithm, followed by Benjamini and Yekutieli multiple testing correction. The terms shown in clouds scored less than 0.01 *p*-value after the correction.

When considering genes annotated with common diseases, Fig 7A shows a significant enrichment of schizophrenia associated genes in the set of “genes in models”, followed by bipolar disorder, Huntington’s disease and Alzheimer’s disease. The order of results is slightly rearranged when considering the whole cell as a background dataset (Fig 7B). For instance, Alzheimer’s disease becomes more prominent, showing the second highest significance for enrichment in our dataset of interest. On the other hand, bipolar disorders drops down the list to the fifth position and autistic disorder appears in the results. This shows how different diseases not only affect specific tissues but can affect a larger number of body regions inducing their effect.

Modelled genes and their overlap with disease genes

Based on the preceding enrichment analyses we wanted to test for specific associations of modelled genes with disease. Since synapses play a crucial role in signal transduction and are affected in many neurological diseases, these were addressed in more detail. We picked seven representative examples of neurological disorders, 6 of which were based on a list published by the Genes 2 Cognition online initiative: Attention Deficit Hyperactivity Disorder (ADHD), Alzheimer’s Disease (AD), Autism, Bipolar Disorder

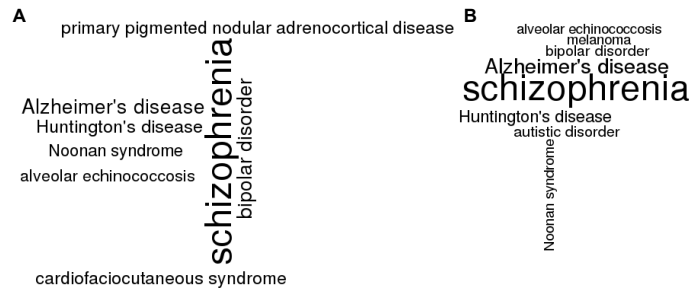


Fig 7. DO enrichment analysis results of “genes in models”. Two background datasets were used: synaptic proteome (A) and all human protein coding genes (B). The list of significant terms was obtained with the Fisher’s exact test and the *elim* algorithm, followed by Benjamini and Yekutieli multiple testing correction. The terms shown in clouds scored less than 0.01 *p*-value after the correction.

(BD), Depression and Schizophrenia. The seventh example was Parkinson’s Disease (PD), motivated by our research interests. The list is a representative rather than exhaustive sample of diseases affecting synapses, including diseases of mental health, developmental disorders, as well as diseases of anatomical entity, such as neurodegenerative diseases. Table 5 gives the DO identifiers and short descriptions of each disease.

331
332
333
334
335
336

Table 5. Diseases of Interest and short descriptions.

Disease	DOID	Description
Alzheimer’s Disease (AD)	DOID:10652	Tauopathy, characterized by memory lapses, emotional instability and progressive loss of mental ability. It results in progressive memory loss, impaired thinking, changes in personality and mood, up to profound decline in cognitive and physical functioning.
Attention Deficit Hyperactivity Disorder (ADHD)	DOID:1094	Specific developmental disorder, characterized by co-existence of attentional problems and hyperactivity.
Autistic Disorder	DOID:12849	An autism spectrum disorder, characterized by symptoms across three symptom domains (communication, social, restricted repetitive interests and behaviors) and delayed language development.
Bipolar Disorder	DOID:3312	A mood disorder that involves alternating periods of mania and depression.
Major Depressive Disorder (MDD)	DOID:1470	An endogenous depression that is characterized by an all-encompassing low mood accompanied by low self-esteem, and by loss of interest or pleasure in normally enjoyable activities.
Parkinson’s Disease (PD)	DOID:14330	Synucleinopathy, based on the degeneration of the central nervous system that often impairs motor skills, speech, and other functions.
Schizophrenia	DOID:5419	Psychotic disorder, characterized by a disintegration of thought processes and of emotional responsiveness.

Onto Suite Miner [70] was used to obtain all genes linked to the DO IDs from the databases supplying gene–disease association information (GeneRIF, OMIM and EnsemblVariation). The various databases have different approaches to disease-gene annotations. EnsemblVariation relies on genetic mutations (mostly Single Nucleotide Polymorphisms, SNPs), whereas OMIM and GeneRIF contain curated text annotations describing disease–gene associations from which data can be extracted using text-mining tools. The different sources were considered individually and jointly. All results refer to the full set of disease associated genes irrespective of the original data source. The

337
338
339
340
341
342
343
344

number of genes linked to each of the diseases can be seen in the “Disease Genes” row in Table 6.

Table 6. Overlap of modelled and disease genes.

Disease	AD	ADHD	Autistic Disorder	Bipolar Disorder	MDD	PD	Schizophrenia
Disease Genes	1511	665	575	1140	616	620	1844
Disease Genes in the Synapse	645 (43%)	233 (35%)	255 (44%)	379 (33%)	202 (33%)	262 (42%)	828 (45%)
Disease Genes in Synapse and in modelled Genes	63 (9.8%)	20 (8.6%)	30 (11.8%)	45 (11.9%)	23 (11.4%)	16 (6.1%)	92 (11.1%)

Overlap of modelled and disease genes and their presence in the synapse and our modelled gene set. Disease information is based on GeneRif, OMIM and EnsemblVariation database data. “AD” stands for Alzheimer’s Disease, “ADHD” for Attention Deficit Hyperactivity Disorder and “PD” for Parkinson’s Disease. Numbers in brackets refer to the percentages. Percentages in the “Disease Genes in the Synapse” column are relative to the total of “Disease Genes” and “Disease Genes in Synapse and in Modelled Genes” is relative to the number of “Disease Genes in Synapse”.

Since not all disease genes are expressed in the synapse, we used the synaptic proteome (see Comparison with proteomic data) to filter the disease associated genes for genes that are expressed in the synapse (see the “Disease genes in the synapse” row, Table 6). Since almost all modelled genes are expressed in the synapse we only present numbers describing the overlap between disease proteins found in the synapse and modelled genes (see the “Disease Genes in Synapse and in Modelled Genes” row, Table 6).

The number of genes associated with diseases varies over a threefold range, from 575 for autistic disorder to 1844 for schizophrenia. However, the proportions of genes associated with a disease and expressed in the synapse range between 33% (Bipolar Disorder and Major Depressive Disorder) and 45% (Schizophrenia). The number of overlapping modelled genes and disease-associated genes (in the synapse) varies between diseases. Schizophrenia has the highest net overlap (92 genes), but also shows the highest number of total associated genes (1844). In total, between 6.1% (Parkinson’s Disease) and 11.8% (Autistic Disorder) of disease genes associated with any of the selected diseases expressed in the synapse appeared in at least one model.

If a gene is associated with many neurodegenerative diseases, its overall function is likely to be generic, leading to a synaptic dysfunction that is not specific to a certain disease. Including such genes in models might explain mechanisms underlying multiple diseases but will not help to model specific diseases. We therefore searched for synaptic genes common to a number of diseases. Table 7 shows the 32 synaptic genes linked to three or more of the diseases included in the analysis. Seven genes are associated to six or all seven tested diseases. The top coverage disease associated genes, found in models annotated, include the protein family voltage-dependent calcium channel family *CACNA1C* and *CACNB2* and dopamine D1 and D2 receptors (*DRD1*, *DRD2*), the inotropic glutamate NMDA receptors, type subunit 2A and 2B (*GRIN2A*, *GRIN2B*) as well as the glutamate metabotropic receptor 5 (*GRM5*). Of the set of modelled genes, 130 (around 50% of the total) are not associated with any of the seven diseases.

In summary, the fraction of genes modelled is relatively small and might indicate that it is challenging to use existing models to make disease predictions. On the other hand the modelled genes can be starting points to extend models to obtain better disease insights, as will be considered in the Discussion (Approaches to including non-modelled disease genes in models).

Table 7. Modelled genes associated with three or more of the selected diseases.

Gene Names	ADHD	AD	Autistic Disorder	Bipolar Disorder	MDD	Schizophrenia	PD
<i>CACNA1C, DRD2, GRIN2A, GRIN2B</i>	1	1	1	1	1	1	1
<i>GRM5</i>	1	1	1	1	0	1	1
<i>CACNB2, DRD1</i>	1	1	1	1	1	1	0
<i>HOMER1</i>	0	1	1	0	1	1	1
<i>CACNA1S, GRM7</i>	1	0	1	1	1	1	0
<i>NOS1</i>	1	1	0	1	0	1	1
<i>GNB3, GRM2</i>	0	1	0	1	1	1	0
<i>GRIA2</i>	0	1	1	0	1	1	0
<i>GNAL</i>	1	0	0	1	1	1	0
<i>PLA2G6</i>	0	0	0	1	0	1	1
<i>ATP2A3, CACNA2D1, GRM3</i>	0	0	0	1	1	1	0
<i>GRIK2, GRM8, GRIP1, PPP1R1B</i>	0	0	1	1	0	1	0
<i>DLG4, NRG1</i>	0	1	0	0	0	1	1
<i>GRIA4</i>	0	1	0	0	1	1	0
<i>FYN, GRIA1, GRIN1, GRM1, GNB2L1</i>	0	1	0	1	0	1	0
<i>SHANK3</i>	1	0	1	0	0	1	0

The number in each cell indicates whether the genes in the Gene Names column are associated (1) or not associated (0) with the diseases indicated in the column headings.

Family trees of entities

Our identification of entities in models makes it possible to query in which models a particular entity is contained. The mapping of entities to genes allows querying models by genes that are, or may be, modelled. It is also desirable to query models by families of molecules, such as PDE4. For example Gutierrez-Arenas et al. [22] and Nair et al. [33] include *PDE4A*, whereas Kim et al. [26] and Oliveira et al. [16] include *PDE4B* in their models, and Kim et al. [25] and Qi et al. [37] specify PDE4.

To enable query by class or family, we determined 29 hierarchical family trees of “proteins”, “protein families” and “protein multimers” implied by the sets of genes corresponding to each (Fig 8). Each “protein family” or “protein multimer” entity is the parent to one or more “proteins” or “protein families”. Each child corresponds to a subset of the proteins in the parent. Tree structures were generated for all “protein multimers” and for “protein families” where a member of that family has been modelled explicitly in at least one of our analysed models. This meant that, for example, PP1 is not represented, since none of its children *PPP1CA*, *PPP1CB* and *PPP1CC* appear in any model explicitly. Individual proteins appear only if they are part of a family or multimer, and they appear in a model – thus, for example, *GRIA4* and *GRIN3* do not appear. Proteins that do not belong to a family, e.g. PSD95 (*DLG4*), are not shown.

Any entity that is part of a family can be mapped to the root node of its tree. Entities that do not belong to a family are implicitly their own root. This mapping of “entities to entity families” (Fig 1) can be applied to the model-entity catalogue (Fig 2) to give the simplified summary mapping of models to 104 family roots shown in Fig 9. This facilitates comparison of entities across models trying to address the differences in model detail between models.

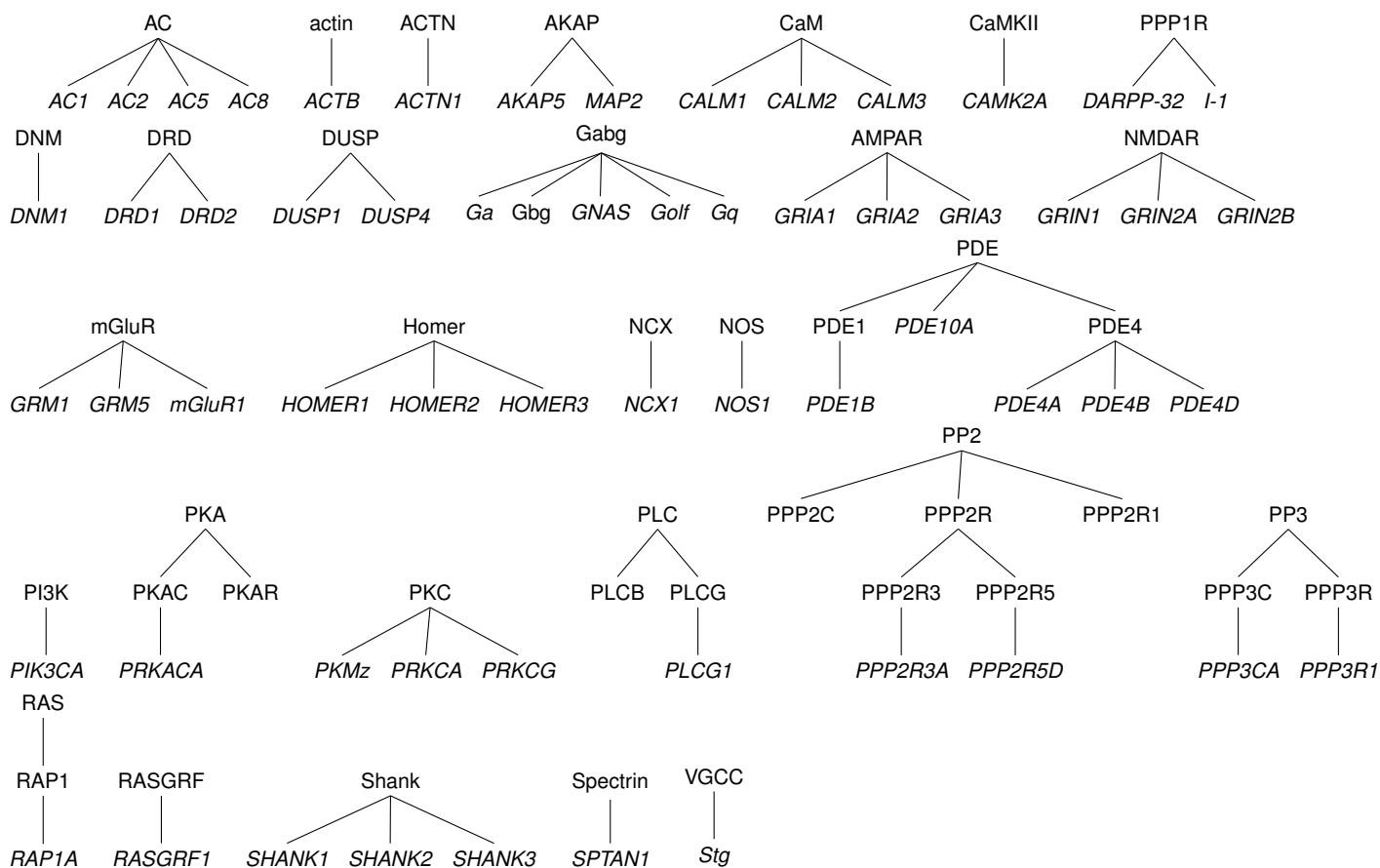


Fig 8. Family trees of “protein families” and “protein multimers”. “Proteins” are shown in italics; “protein families” and “protein multimers” in roman. “Proteins” that do not belong to any family are not shown. Only proteins that are specified in models are shown.

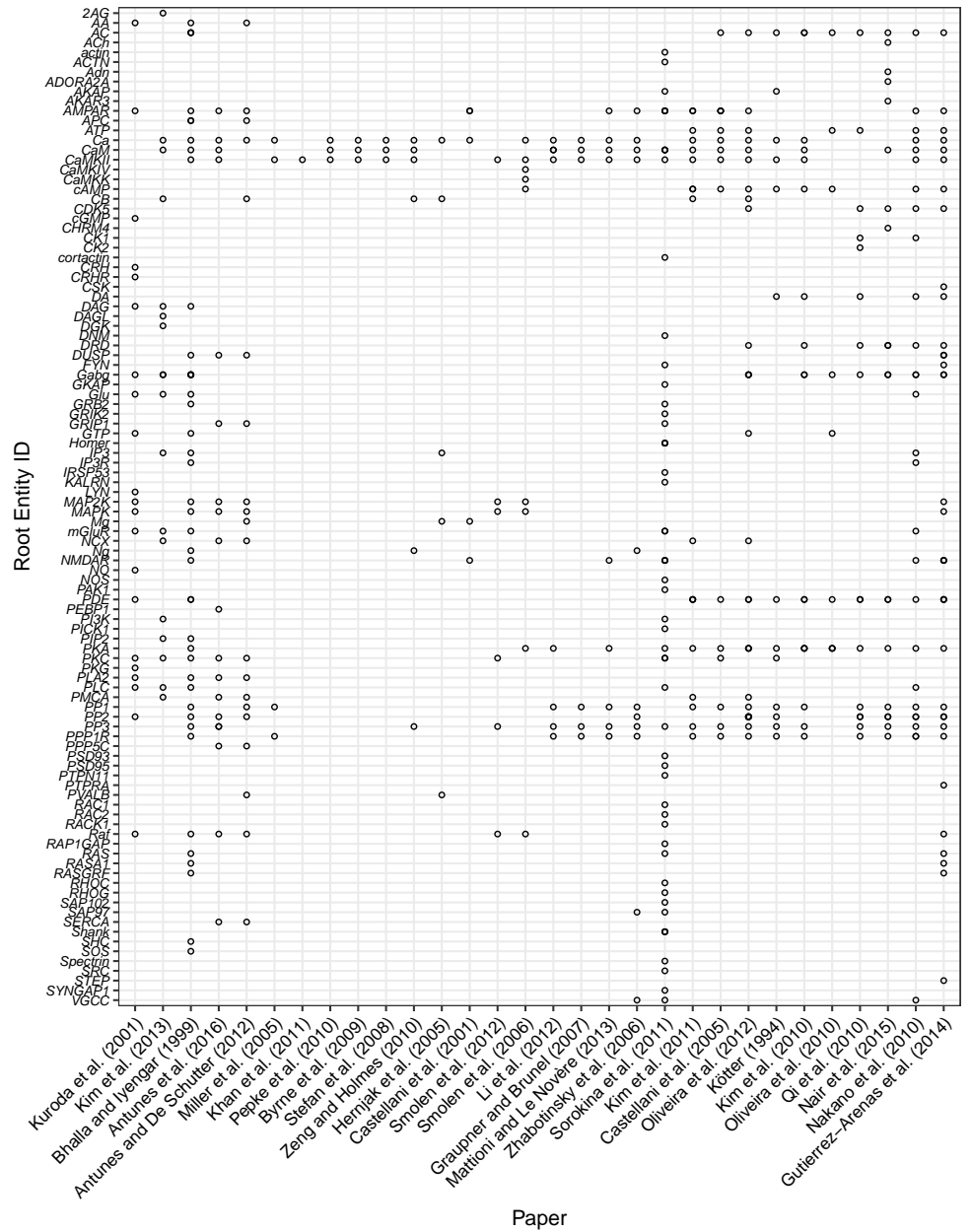


Fig 9. Summary mapping of entities in models. The occurrence of a root entity in a model is indicated by open circles. Lower-level entities are folded into their root entity.

Frequency of modelling

Table 8. Numbers of entities or entity families found in models.

Entity family	Models	Frequency	% Frequency
2AG, actin, ACTN, Adn, AKAP, AKAR3, CaMKIV, CaMKK, cGMP, CHRM4, cortactin, CRH, CRHR, CSK, DAGL, DGK, DNM, GKAP, GRIK2, Homer, IRSP53, KALRN, LYN, NO, NOS, PAK1, PEBP1, PICK1, PSD93, PSD95, PTPN11, PTPRA, RAC1, RAC2, RACK1, RAP1GAP, RHOC, RHOG, SAP102, Shank, SHC, SOS, Spectrin, SRC, STEP, SYNGAP1	1	46	47.4
APC, CK1, FYN, GRB2, IP3R, PI3K, PIP2, PVALB, RASA1, RASGRF, SAP97, SERCA	2	12	12.4
AA, DAG, GRIP1, Mg, Ng, RAS, VGCC	3	7	7.2
CDK5, DUSP, Glu, GTP, IP3, PLA2	4	6	6.2
DA, DRD, mGluR, NCX, PLC, PMCA	5	6	6.2
CB, NMDAR	6	2	2.1
ATP, MAP2K, MAPK, Raf	7	4	4.1
cAMP, Gabg, PKC, PP2	9	4	4.1
AC	10	1	1.0
AMPA, PDE	12	2	2.1
PPP1R	14	1	1.0
PKA	15	1	1.0
PP1	16	1	1.0
PP3	17	1	1.0
CaM	18	1	1.0
CaMKII	22	1	1.0
Ca	23	1	1.0

“Models” is the number of models containing the entity or at least one member of the family. “Frequency” is the number of appearances of the family or entity in the given number of models, and “% Frequency” is the frequency expressed as a percentage.

To give an indication of which are the frequently modelled entities and families of entities, we determined the number of models in which each of the root entities in Fig 10 appears (Table 8). About 50% of root entities appear only in one model. In total, 26 (about 25%) of the entity roots were included in five models or more. The three most frequently modelled entities and families are CaM, CaMKII and Ca, which are included in 18, 22 and 23 out of 30 analysed models respectively. This is due to a number of models focusing specifically on the Ca^{2+} -CaM-CaMKII pathway or including it as a model part, reflecting the central role of phosphorylation of CaMKII by Ca^{2+} -bound CaM in synaptic biology. These top coverage families are followed by families such as calcineurin (PP3) and protein phosphatase 1 (PP1), cAMP-dependent protein kinase (PKA) and PPP1R (the receptor subunit of PP1), which are included in the models that model dephosphorylation of CaMKII via the Ca^{2+} -PP3-I1-PP1 pathway. Receptor related families such as AMPAR appear with lower frequency, reflecting the fact that, while crucial for synaptic physiology, not all models include them as a readout mechanism for LTP and LTD. Even though our coverage of models is not complete, it seems likely that cataloguing further models will not change the order much.

Comparing models based on their entities

Having annotated the models with entities enabled us to compare models with each other by applying a hierarchical clustering approach to the model-entity root mapping (Fig 9). Ward's 2D method, as implemented in R's *hclust* function was used to give the dendrogram shown in Fig 10. The dendrogram splits into 4 clusters, each of which contains a majority of models from one brain region (cerebellum, hippocampus, striatum) or contains only a generic model.

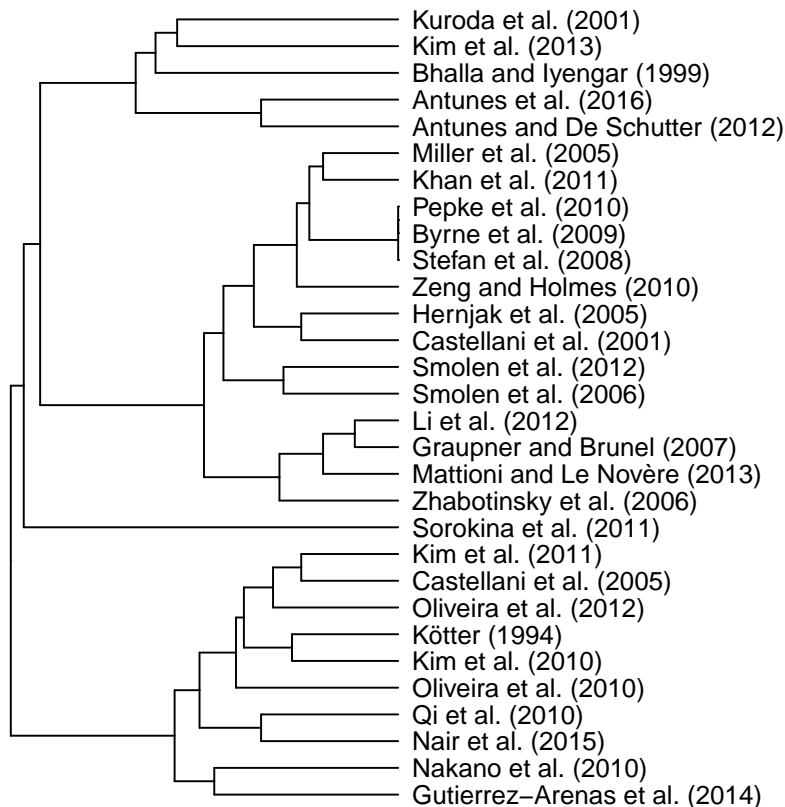


Fig 10. Clustering of the model-entity family root matrix. Clustering of the matrix in Fig 9 as implemented in R's *hclust* function with the *Ward.2D* method. The colour of the citation indicates the brain region modelled: hippocampus (blue), striatum (red), cerebellum (olive), generic (black). Clusters referred to in the text are indicated by the circled numbers.

The cluster labelled 1 is dominated by 3 cerebellar models [17, 18, 29] (olive text), and also contains the hippocampal model (blue text) of Bhalla and Iyengar [14] and the striatal model (red text) of Kim et al. [27]. It can be seen in the matrix of root entities (Fig 9) that distinctive proteins and families in this cluster are PKC (shared by all 5 models), PLA2 (in 4 of 5 models), DAG and PMCA (in 3 of 5 models), and the Raf-MAP2K-MAPK pathway (4 of 5 models).

Most of the 14 models belonging to cluster 2 are hippocampal (7) or generic (5), along with the cerebellar model of Hernjak et al. (2005) [23] and the striatal model of Mattioni and Le Novère [31]. Three models (Byrne et al. [19], Pepke et al. [36] and Stefan et al. [41]) are clustered together as they all contain the identical set of entities: Ca, CaM and CaMKII. The closely related model of Zeng and Holmes [42] includes

calbindin (CB) as well, and the closely related models of Miller et al. [32] and Khan et al. [24] are also centred on CaMKII. The related models of Smolen et al. (2006) [38] and Smolen et al. (2012) [39] feature the MAPK pathway, in addition to CaMKII. The group of models containing Li et al. [30], Graupner and Brunel [21], and Zhabotinsky et al. [43] are all variations on the CaMKII phosphorylation-dephosphorylation circuit, all adding PP1 and PP3 (calcineurin) to the Ca^{2+} -CaM-CaMKII pathway.

The sole member of cluster 3, the model of Sorokina et al. [40], is dissimilar to other models, reflecting the large number of entities, particularly scaffolding proteins, which are contained in this model but not in others.

Cluster 4 mostly contains striatal models [22, 25, 26, 33, 34, 37], with the exceptions of the generic model of Castellani et al. [15] and the hippocampal models of Kim et al. [25, 26]. These models are some of the few non-striatal models to contain the adenylyate cyclase (AC)-cyclic adenosine monophosphate (cAMP)-PKA pathway as well as hydrolysis of cAMP to adenosine monophosphate (AMP) by phosphodiesterase (PDE). Dopamine and G-coupled protein receptors also feature in this cluster.

The bias of each cluster towards a particular brain region indicates that the clustering is meaningful. However, the bias may arise more from choices modellers have made about which pathways to include in models of the various regions. For example, dopamine receptors are included in most striatal models and are only included in a few hippocampal models. Nevertheless, the clustering provides a different view of the landscape of models, and could be used to identify models with similar composition, whose behaviour it might be insightful to compare. We also applied the clustering to the full model-entity matrix (Fig 2), with similar results, though slightly less meaningful groupings.

Approaches to including non-modelled disease genes in models

Knowing which disease associated genes are included in models helps models with high potential to explain disease impact on the synapse to be identified (Modelled genes and their overlap with disease genes). It also allows us to identify disease associated proteins which do not appear in the models we analysed. Of all disease associated genes, 1,248 are found in the synaptic proteome but not in any of the analysed models. Table 9 shows the 32 genes that are associated with 5, 6 or all 7 diseases, and which do not appear in any of the investigated models. Of these, *COMT* and *SLC6A3* are associated with all 7 diseases of interest. Since these genes are associated with all or many studied diseases, they could be of interest when it comes to gaining a better understanding of generic disease dysfunctions.

Supporting the idea that genes implicated in many diseases could be potentially targets for modelling, we identified two genes, *COMT* and *MAOA*, that have been included in metabolic models [71, 72]. Functionally, the catechol O-methyltransferase (*COMT*) degrades catechols, such as dopamine, by catalysing their methylation. This methylation results in one of the major degradative pathways of the catecholamine transmitters [73]. Dopamine is included in a number of analysed models [74, 75], and it could be possible to explore what happens in these models if there is an excess of dopamine due to *COMT* malfunction.

Genes associated with all studied diseases could represent generic disease mechanisms, in which case exploring the role of *COMT* in dopaminergic models would indicate the possible influence of the gene in many diseases. An alternative approach is to consider disease specific genes not appearing in models and associated with only one of the selected diseases. Integrating such proteins into pre-existing models could thus help to gain disease-specific insights. 824 of the disease associated genes are specific to one disease only. To identify genes that can be integrated into existing models, the list

Table 9. Disease associated genes not appearing in any of the annotated models.

Gene Names	ADHD	AD	Autistic Disor- der	Bipolar Disor- der	MDD	Schizo- phre- nia	PD
<i>COMT, SLC6A3</i>	1	1	1	1	1	1	1
<i>GIGYF2</i>	1	0	1	1	1	1	1
<i>GSK3B, ABCB1</i>	1	1	0	1	1	1	1
<i>ANK3, ENO1, KIF5C, MAOA, PRNP, SLC17A6, CSMD1</i>	1	1	1	1	1	0	1
<i>ACE, GAD1</i>	0	1	1	1	1	1	0
<i>DDC, FMR1</i>	1	0	1	1	0	1	1
<i>APAF1, DFNA5, ELAVL2, GRIK1, HINT1, ITIH1, ITIH3, ITIH4, STT3A, LIG4, NDUFB1, NDUFB7, NPY, NTRK3, GATB, SMARCA2, MAD1L1, PRPF3, SH3PXD2A, TRANK1, PPIF, NT5C2, KIF21B, RPRD2, SYNE1, NGEF, TENM4, GNL3, MPP6, MRPS21, RAB39A, CNNM2, OXR1, ANKS1B, VARS2, AS3MT, PALB2, DCTN5, PPP1R21, MTPN, SLC39A12, CHSY3</i>	1	0	1	1	1	0	1
<i>CNR1</i>	1	1	0	0	1	1	1
<i>YWHAZ</i>	1	1	1	0	0	1	1
<i>SNAP25</i>	1	1	1	0	1	0	1
<i>CNTNAP2</i>	1	1	1	1	0	0	1

The number in each cell indicates whether the genes in the Gene Names column are associated (1) or not associated (0) with the diseases indicated in the column headings. The table only lists genes that are associated with four or more diseases.

of non-modelled disease associated genes was compared with genes in pathways enriched in the modelled genes.

For example, all disease genes unique to Schizophrenia were compared with the list of genes in pathways significantly enriched in the modelled genes, giving a list of 8 genes, each of which is found in one or more pathways (Table 10). One of these genes is *LAMTOR2*. The *LAMTOR2:LAMTOR3* complex binds MAPK components [76], together with other members of the mitogen-activated protein kinase (MAP2K) and mitogen-activated protein kinase (MAPK) activation pathway, such as *RAF1*, *MAPK1*, *MAPK3* and *MAP2K2*. In this role it contributes to the activation of the MAPK pathway which has a central role in striatal and cerebellar synapses. Including the influence of *LAMTOR2* on the activity of MAPK in a pre-existing model could hence help to better understand its role in and effects on schizophrenia. Integrating *LAMTOR2* activity in the model could be done mechanistically, or functionally, for example by influencing the MAPK concentration.

Discussion

We have developed a catalogue of genes whose corresponding proteins correspond to entities in computational models of synaptic plasticity. To achieve this we developed a new set of standard identifiers for entities in computational models, and mapped those entities corresponding to proteins and protein families onto genes. Although time and lack of machine-readable model descriptions constrained the number of models we could analyse, by selecting models from three brain regions (hippocampus, striatum and

Table 10. Schizophrenia specific genes not found in models and appearing in pathways that are enriched in annotated models.

Gene Name	Gene Name (long)	REACTOME pathway	Pathway ID
<i>CCK</i>	cholecystokinin	G alpha (q) signalling events	R-HSA-416476
<i>LAMTOR2</i>	late endosomal/lysosomal adaptor, MAPK and MTOR activator 2	MAP2K and MAPK activation, FCERI mediated MAPK activation, VEGFR2 mediated cell proliferation, RAF/MAP kinase cascade	R-HSA-5674135, R-HSA-2871796, R-HSA-5218921, R-HSA-5673001
<i>PSMB1</i>	proteasome subunit beta 1	FCERI mediated MAPK activation, VEGFR2 mediated cell proliferation, RAF/MAP kinase cascade	R-HSA-2871796, R-HSA-5218921, R-HSA-5673001
<i>PSMB4</i>	proteasome subunit beta 4	FCERI mediated MAPK activation, VEGFR2 mediated cell proliferation, RAF/MAP kinase cascade	R-HSA-2871796, R-HSA-5218921, R-HSA-5673001
<i>PSMC1</i>	proteasome 26S subunit and ATPase 1	FCERI mediated MAPK activation, VEGFR2 mediated cell proliferation, RAF/MAP kinase cascade	R-HSA-2871796, R-HSA-5218921, R-HSA-5673001
<i>PSMC4</i>	proteasome 26S subunit and ATPase 4	FCERI mediated MAPK activation, VEGFR2 mediated cell proliferation, RAF/MAP kinase cascade	R-HSA-2871796, R-HSA-5218921, R-HSA-5673001
<i>PSMD2</i>	proteasome 26S subunit and non-ATPase 2 and	FCERI mediated MAPK activation, VEGFR2 mediated cell proliferation, RAF/MAP kinase cascade	R-HSA-2871796, R-HSA-5218921, R-HSA-5673001
<i>TUBB3</i>	tubulin beta 3 class III	Chaperonin-mediated protein folding	R-HSA-390466

cerebellum) we are confident that we have covered the bulk of proteins in models. 511

We were able to identify 294 genes that could be mapped to entities in 512
computational models. This corresponds to 4.2% of the 6,706 known genes in the 513
synaptic proteome. Enrichment analysis showed that, compared to the set of proteins 514
found in the synapse, the genes in models tended to have more signalling functions, 515
which reflects the focus on signalling pathways in such models. This suggests 516
considerable scope for including new molecules in models. However, models of synapses 517
at the molecular level are already complex and are beset by problems of determining 518
parameters. One strategy to prioritise molecules to add to models is to chose those most 519
relevant for disease. Our comparison of the list of genes in models with databases of 520
gene-disease association shows that many disease-associated genes are not currently 521
included in synaptic models, and suggests targets for future modelling. 522

Targeting disease-relevant proteins for modelling 523

The genes in models are more associated with neurological diseases, such as 524
Schizophrenia, Alzheimer's, Huntington's disease and bipolar disorder, than randomly 525
selected genes in the synaptic proteome or the whole genome. Nevertheless, depending 526
on the disease, the number of disease-associated genes included in models range between 527
6% and 12% of the disease-associated genes in the synapse. This suggests that there is 528
considerable potential to include disease-related genes in models. Including these 529
molecules could make these models more useful in helping elucidate disease mechanisms 530
and helping to identify new drug targets. 531

We identified two un-modelled genes associated with 7 neurological diseases, *COMT* 532
and *MAOA* and we found they have close functional links with existing models. By 533
incorporating pathway enrichment results, we identified *LAMTOR*, a gene uniquely 534

associated with Schizophrenia. *LAMTOR* is linked to the MAPK pathway, which features in a number of existing models. This demonstrates the utility of our approach for identifying which proteins to incorporate in existing models so that they can make disease-associated predictions. Further investigation using this approach could indicate other target proteins to add to existing synaptic pathway models to make them more informative about the influence of diseases on the synapse.

A new ontology for computational neuroscience models

The challenge we faced mapping model entities to genes highlighted a gap between bioinformatics, where each gene is well-defined and has a commonly used identifier, and computational neuroscience, where the elements of models are defined at varying levels of precision: for example they may be proteins, protein families or multimers of proteins. Even within the same model, one element may be specified precisely, for example a particular isoform (PKM ζ), and another element may be generic, for example “plasticity related proteins” [39]. From a bioinformatics perspective this may seem offensive, but from the viewpoint of computational neuroscience it is entirely valid: a computational model can be seen as a means to reasoning about a hypothesis; the formulation of the model is the hypothesis and the simulations embody the reasoning that generates the predictions arising from the hypothesis [77]. The modelling process sometimes even requires hypothetical elements, which have no existing identifier. For example, one seminal computational neuroscience model [78] contained hypothetical elements (“gating particles”) that predicted essential features of ion channels function.

The problem of mapping model constituents onto biological entities was noted by the originators of the MIRIAM standard [49]. This standard suggests solving the problem of mapping entities at different levels of abstraction by using a “HasVersion” qualifier to map reactants in models to multiple entities, e.g. to map IP3R to Inositol 1,4,5-triphosphate receptors type 1, 2 and 3. Most of the models we investigated had not been annotated to MIRIAM standards, and we found it more efficient to define our own ontology containing proteins and protein families. We found that existing ontologies such as UniProt, HGNC gene families [79] and Neurolex [80] were not extensive enough to map proteins specified at different levels of precision (e.g. PDE4A, PDE4) to common families (e.g. PDE), though HGNC gene families covered about half of the protein families we identified.

In the absence of a suitable ontology, we used HGNC gene families and curated other family relationships manually to give a full list of entities (see S2 Table) and mappings of proteins to families and multimers in which they occur (see S3 Table and S4 Table). These tables form the kernel of an ontology, and we have demonstrated that it can be used to determine the potential genes underlying the proteins in computational models, and to cross-link these genes with expression data. Furthermore, we have demonstrated that the ontology can be used to compare models, for example using hierarchical clustering, and to summarise of how often various protein families have been modelled. By annotating models with identifiers of brain region or neuron type, the set of possible proteins belonging to a model could be narrowed down according to the genes that are expressed in a given region. The same procedure could be used to link the genetic content of synaptic models with other types of data, for example spatial expression data from the Allen Brain atlas. This would make it possible to check that a particular model was valid in the brain region it is supposed to represent, or, conversely, could be used to find brain regions for which a particular model might be valid.

The number of models analysed in this paper was limited by the time it took us to annotate models we had not constructed. While some repositories, such as the curated branch of BioModels, enforce curation of models to MIRIAM standards [49], it would be desirable for all models to be annotated consistently at the time of publication or

deposition in a repository. Annotation would be a fairly quick process for authors familiar with the models, and the quality of the information would be higher than if annotated by third parties. Three of the 30 models we investigated were annotated to MIRIAM standards. We did not use the MIRIAM annotations of these models, partly so that our annotation of models was consistent and partly because the MIRIAM standard suggests mapping to external identifiers that are often at a finer level of granularity than we needed to compare models to proteomic data. Were more models curated to MIRIAM standards, it would be worthwhile developing a mapping to our identifiers.

As discussed above, some models are of necessity not precise about which protein is specified. To address this, one option would be for the computational neuroscience and bioinformatics communities to adopt an ontology along the lines of the ones we have generated here. If the ontology were stored in the Interlex dynamic lexicon of biomedical terms, a development of Neurolex [80], it would be straightforward for authors to suggest new terms or relationships. The model metadata could be stored by adding fields to existing repository schema, or our data could be converted to a standalone, API-enabled database.

Nomenclature

The nomenclature we have used for entities has been decided by the authors. We have been guided by gene names, and some of our choices might be controversial, for example naming PP2B (calcineurin) PP3. Our rationale for using identifiers related to gene names is so there is more consistency between the names of members in a family. For example, in Fig 8, PP3 is the parent of the catalytic and regulatory subunits PPP3C and PPP3R; having PP2B as a parent would not be equally consistent. It would be desirable for the computational neuroscience and bioinformatics communities to agree a common nomenclature.

New directions in modelling

We have demonstrated the potential of our method of identifying entities in models and mapping them to genes to suggest new, disease-relevant directions for modelling. We believe there is considerable potential for the work to be adopted to suit the needs of the community. Our data and mapping tables and code to reproduce the results in this paper are available (S1 File) and suggestions for additions or amendments are welcome. [We will also be making our files available via github.]

More speculatively, despite the challenge of expanding the number and relevant proteins in models of synaptic plasticity, we believe that the time has come to incrementally increase the number of proteins involved in models, especially those involved in disease mechanisms.

Methods

Identifying entities in models

The question of what entities mean is outlined in Analysis of proteins in synaptic models, subsection Identifying entities in models. The constituent entities of each model were identified by one of the authors (EMW, KFH or DCS) reading the paper, or extracting elements from a machine-readable representation of the model, for example CellML [14] or Kappa [40] descriptions. The name used to identify the entity in the model was then mapped to the standardised list of entities that we built up as we looked through the models. In some cases model entities were not specified enough to allow us to map them

unambiguously onto a model entity – for example “Plasticity Related Protein” [39]. We did not consider a complex as an entity – for example a Ca-CaM-CaMKII complex would give rise to Ca (ion), CaM (“protein”) and CaMKII (“protein multimer”). In naming our standard entities, we have tried to use names commonly used in models, but for entities that have not appeared in many models we have tended to use the newer standard names that appear in the NCBI or UniProt databases.

Mapping entities to a unique gene identifier

To obtain a common identifier for all entities we searched for an ontology that could be used to identify our entities, especially “protein families” and “protein multimers”. We considered a number of potential ontologies:

The Computational Neuroscience Ontology

(<http://bioportal.bioontology.org/ontologies/CNO>) This ontology covers the description of the modelling technique (e.g. Integrate-and-fire neurons) rather than the components of the model.

HGNC Gene families (<http://www.genenames.org/>)

The Human Gene Organisation Gene Nomenclature Committee (HGNC) approves unique symbols and names for human genes, and also places genes in families, based on characteristics such as function, homology, domains and phenotype [79]. Placing genes into families is a manual process, often involving specialists who are expert in that family of genes. Often, but not always, genes in the same family have a common root symbol. The process of defining families is ongoing.

InterPro protein families (<http://www.ebi.ac.uk/interpro>)

The InterPro Consortium is a federation amalgamating protein signature databases (Gene3D, Conserved Domain Database, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY, Structure-Function Linkage Database and TIGRFAMs) [81]. Protein signatures are predictive models build on fragments of amino acid sequences that share local features (e.g. conservation at different positions) known to be associated with a function or structure [82]. There are multiple computational approaches that are detecting such patterns and define types of signatures [83]. The similarity in signature matches between proteins is used to define a hierarchy of families.

Manual NCBI search (www.ncbi.nlm.nih.gov/gene/)

The National Center for Biotechnology Information (NCBI) provides access to biomedical and genomic information. We used their searchable database of genes, which can be queried with a number of different identifiers.

We intended to map out entities using information supplied by one of these ontologies, but no one source proved sufficient. In InterPro, there are a number of families that correspond exactly to proteins, for example Phospholipase A2 (IPR001211) and Phosphoinositide phospholipase C (IPR001192). However, some proteins, including SOS1 and SOS2, belong to very broad families.

In the HGNC database we identified a relatively large number of our entities that correspond to existing HGNC gene families. For example the HGNC Homer family (short for “Homer scaffolding proteins”) comprises the genes *HOMER1*, *HOMER2* and *HOMER3* and the genes *PPP3CA*, *PPP3CB*, *PPP3CC*, *PPP3R1* and *PPP3R2* belong to the HGNC PP3 family. Other entities do not correspond to a single gene family, but can be extracted from the database by selecting multiple families. For example SHANK, by which we mean the family of proteins encoded by *SHANK1*, *SHANK2* and *SHANK3*

may be selected from the gene families list by selecting all genes that are in the “Ankyrin repeat domain containing” (ANKRD) and “PDZ domain containing” (PDZ) gene families. Some of our entities cannot be recovered by searching for families. For example SOS (by which we mean the proteins encoded by *SOS1* and *SOS2*) are in both the “Rho guanine nucleotide exchange factors” and “Pleckstrin homology domain containing” families, but so are 35 other proteins.

We also curated our own mappings by manually querying the NCBI portal by searching for human genes matching a full protein name and a common gene prefix, suffix or infix, if available. For example, Entrez IDs for a “protein family” of Voltage-dependent calcium channel were obtained with the following query: ‘Voltage-dependent calcium channel[All Fields] AND CACN*[All Fields] AND ”Homo sapiens”[Organism]’. The top 20 results were considered and only entries with the closest description and gene summary to the search term were extracted.

Although we were not able to map all our entities by relying on only one ontology, we found that HGNC families covered more of our entities than Interpro, so we used this as a basis for developing an ontology to describe the molecular components of computational neuroscience models. We tried to map all entities of type “protein family” and “protein multimer” to HGNC families. Manual NCBI mappings were used to check and verify that HGNC families represented the modelled group of genes.

In situations where we were unable to find a corresponding HGNC family we (1) suggested some protein groups to be added to the list of HGNC families and await approval of the request; (2) we had no choice but to fall back on our manual NCBI mapping. The combination of the above lead us to our final mappings. S3 Table and S4 Table show identified HGNC families as well as the genes belonging to them. The superscript given with the HGNC family name indicates its origin, the official HGNC mapping vs. custom mapping. The columns “IN SYNAPSE” and “OUT SYNAPSE” are explained in Analysis of proteins in synaptic models, Comparison with proteomic data.

Enrichment Analysis

A commonly used method to find statistically significant commonalities between large gene lists is enrichment analysis, also known as over-representation analysis. Based on information contained in ontological databases, enrichment analysis can show if a set of “genes of interest” contains a significantly high number of genes with the same annotation. This approach allows us to gain a better understanding of underlying common themes in our “genes in models” list.

The underlying principle of such an enrichment analysis is to estimate, for each specific category annotated in the database of interest, if the number of genes in our genes of interest set associated with a certain category is larger than expected by chance. To test this relationship statistically, the hypergeometric distribution or one-tailed Fisher’s exact test is commonly applied. Both are known to be equivalent [84].

The four key numbers required to carry out the statistical calculations are:

1. The number of elements in the full dataset, also considered as the background dataset, N . In our case these are all proteins part of the synaptic proteome.
2. The number of elements n in the subset of the full dataset which is tested for enrichment. This is the number of genes in the “genes in models” list.
3. The number of elements associated to a certain trait in the full dataset, T . It corresponds to the set of genes annotated to any term in one of the databases, e.g. “Schizophrenia”, which describes a disease in the DO database.

-
4. The subset of n shared by the elements found in T , denoted as t . This refers to the number of genes within a category that are also present in our “genes in models” list.

The probability of encountering the exact number of hits t of interest given N , n and T is calculated with the hypergeometric probability $h(t; N, n, T)$:

$$h(t; N, n, T) = \frac{\binom{T}{t} \binom{N-T}{n-t}}{\binom{N}{n}} \quad (1)$$

To describe the probability of finding greater than or equal to the number of items of interest t , we use the cumulative hypergeometric probability:

$$p(t; N, n, T) = \sum_{x=t}^T h(x; N, n, T) = \sum_{x=t}^T \frac{\binom{T}{x} \binom{N-T}{n-x}}{\binom{N}{n}} \quad (2)$$

If this probability is less than a criterion (e.g. $p < 0.01$), the dataset is regarded as enriched [84] for the tested category.

For the analysis, ontology terms for all genes in the background dataset N were obtained. Initially two background sets were considered, containing (1) all genes in the genome and (2) all proteins found in the synapse. Since results were quite similar and the focus of this study is on the synaptic region rather than the whole organism, we only present results obtained with the second dataset as the background set of genes.

We analysed all terms that had at least one gene associated to our “genes in models”. For each such term, the p -value was calculated, indicating potential enrichment, and then corrected for multiple comparison, using the Benjamini and Yekutieli [85] method. Terms with adjusted p -values smaller than 0.01 are presented in the final results.

topONTO and topGO

Ontologies that supply functional annotation information are organised in a hierarchical structure, with the most generic terms at the top, and the most specific ones at the bottom. The higher the term is located in the hierarchy, the more genes are associated with it as it aggregates all genes from its child terms. Hence, a single gene can be found at different levels of annotation specificity. Depending on the purpose of the analysis it is important to be able to choose the level of retrieved terms.

To retrieve the most specific and refined terms among significantly enriched ones, we used an algorithm proposed by Alexa et al. [86] and implemented for the GO database by the R *topGO* package. Since GO is represented as a Directed Acyclic Graph (DAG), the authors incorporated the underlying GO graph topology in the term scoring approach, removing strong correlations commonly occurring between high level terms. This allows the enrichment of a very generic term to be ignored, and less frequent but more specific and potentially more interesting low level ones to be identified.

Assuming that a child term is potentially more interesting than its more generic ancestors, significance of a term is calculated depending on its child terms. Out of multiple versions implementing this idea, we used the *elim* algorithm paired with Fisher’s exact test. The decision was based on the clear number of comparisons conducted by the algorithm. This number was further used to correct for the false discovery rate.

In the *elim* approach [86], enrichment analysis starts at the bottom of the ontology graph. If a child term is significantly enriched amongst the genes of interest, this influences the number of genes annotated with its ancestor terms. All genes associated with the enriched child term are removed from the ancestor terms leaving most specific ones with the minimal indicated significance.

We discovered that the algorithm leads to more refined results than a set-based enrichment analysis that ignores the ontology structure. Therefore, we were interested in applying a same approach to other gene annotation sets. This can be achieved with the *topOnto* R package [69]. It extends the advantage of the Alexa et al. method [86] to any hierarchically structured dataset. Since both REACTOME and DO satisfy this requirement, we were able to apply the same approach to all chosen annotation sets.

Supporting information

S1 File. Data and code. A zip file containing the data tables, and mapping and analysis code that reproduces the results in this paper.

S1 Table. Synaptic Proteome Studies. List of synaptic proteome publications and respective datasets used in this study.

S2 Table. Full list of entities. List of entities containing the ID, name, type and, for proteins, mapping to genes.

S3 Table. Protein family members. List of entities in distinct protein families – “in” and “out” of the synapse.

S4 Table. Protein multimer members. List of entities in distinct protein multimers – “in” and “out” of the synapse.

References

1. Martin SJ, Grimwood PD, Morris RGM. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annu Rev Neurosci.* 2000;23(1):649–711. doi:10.1146/annurev.neuro.23.1.649.
2. Bliss TV, Lømo T. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J Physiol (Lond).* 1973;232(2):331–356.
3. Lynch GS, Dunwiddie T, Gribkoff V. Heterosynaptic depression: a postsynaptic correlate of long-term depression. *Nature.* 1977;266:737–739.
4. Abbott LF, Nelson SB. Synaptic plasticity: taming the beast. *Nat Neurosci.* 2000;3:1178–1183.
5. Nadim F, Bucher D. Neuromodulation of neurons and synapses. *Curr Opin Neurobiol.* 2014;29:48–56. doi:10.1016/j.conb.2014.05.003.
6. Carlisle HJ, Fink AE, Grant SG, O’Dell TJ. Opposing effects of PSD-93 and PSD-95 on long-term potentiation and spike timing-dependent plasticity. *J Physiol (Lond).* 2008;586(Pt 24):5885–5900. doi:10.1113/jphysiol.2008.163469.
7. Pocklington AJ, Cumiskey M, Armstrong JD, Grant SGN. The proteomes of neurotransmitter receptor complexes form modular networks with distributed functionality underlying plasticity and behaviour. *Mol Syst Biol.* 2006;2(1). doi:10.1038/msb4100041.

-
8. Morrison A, Diesmann M, Gerstner W. Phenomenological models of synaptic plasticity based on spike timing. *Biol Cybern.* 2008;98(6):459–478. doi:10.1007/s00422-008-0233-1. 806
807
808
 9. Manninen T, Hituri K, Kotaleski JHH, Blackwell KT, Linne MLL. Postsynaptic signal transduction models for long-term potentiation and depression. *Front Comput Neurosci.* 2010;4. 809
810
811
 10. Nair AG, Gutierrez-Arenas O, Eriksson O, Jauhiainen A, Blackwell KT, Kotaleski JH. Modeling intracellular signaling underlying striatal function in health and disease. *Prog Mol Biol Transl Sci.* 2014;123:277–304. 812
813
814
 11. Blackwell KT, Jedrzejewska-Szmek J. Molecular mechanisms underlying neuronal synaptic plasticity: systems biology meets computational neuroscience in the wilds of synaptic plasticity. *Wires Syst Biol Med.* 2013;5(6):717–731. 815
816
817
 12. Lassek M, Weingarten J, Volkandt W. The synaptic proteome. *Cell Tissue Res.* 2015;359(1):255–65. doi:10.1007/s00441-014-1943-4. 818
819
 13. Bayés À, Collins MO, Croning MD, van de Lagemaat LN, Choudhary JS, Grant SG. Comparative study of human and mouse postsynaptic proteomes finds high compositional conservation and abundance differences for key synaptic proteins. *PLoS ONE.* 2012;7(10):e46683. 820
821
822
823
 14. Bhalla US, Iyengar R. Emergent properties of networks of biological signalling pathways. *Science.* 1999;283:381–387. 824
825
 15. Castellani GC, Quinlan EM, Bersani F, Cooper LN, Shouval HZ. A model of bidirectional synaptic plasticity: from signaling network to channel conductance. *Learn Memory.* 2005;12(4):423–432. 826
827
828
 16. Oliveira RF, Terrin A, Di Benedetto G, Cannon RC, Koh W, Kim M, et al. The role of Type 4 phosphodiesterases in generating microdomains of cAMP: large scale stochastic simulations. *PLoS ONE.* 2010;5(7):e11725. doi:10.1371/journal.pone.0011725. 829
830
831
832
 17. Antunes G, De Schutter E. A stochastic signaling network mediates the probabilistic induction of cerebellar long-term depression. *J Neurosci.* 2012;32(27):9288–300. doi:10.1523/JNEUROSCI.5976-11.2012. 833
834
835
 18. Antunes G, Roque AC, Simoes-de Souza FM. Stochastic induction of long-term potentiation and long-term depression. *Sci Rep.* 2016;6:30899. doi:10.1038/srep30899. 836
837
838
 19. Byrne MJ, Putkey JA, Waxham NN, Kubota Y. Dissecting cooperative calmodulin binding to CaM kinase II: a detailed stochastic model. *J Comput Neurosci.* 2009;27(3):621–638. doi:10.1007/s10827-009-0173-3. 839
840
841
 20. Castellani GC, Quinlan EM, Cooper LN, Shouval HZ. A biophysical model of bidirectional synaptic plasticity: Dependence of AMPA and NMDA receptors. *Proc Natl Acad Sci USA.* 2001;98:12772–12777. 842
843
844
 21. Graupner M, Brunel N. STDP in a bistable synapse model based on CaMKII and associated signaling pathways. *PLoS Comput Biol.* 2007;3(11):e221. 845
846
 22. Gutierrez-Arenas O, Eriksson O, Kotaleski JH. Segregation and crosstalk of D1 receptor-mediated activation of ERK in striatal medium spiny neurons upon acute administration of psychostimulants. *PLoS Comput Biol.* 2014;10(1):e1003445. doi:10.1371/journal.pcbi.1003445. 847
848
849
850
-

-
23. Hernjak N, Slepchenko BM, Fernald K, Fink CC, Fortin D, Moraru II, et al. Modeling and analysis of calcium signaling events leading to long-term depression in cerebellar Purkinje cells. *Biophys J*. 2005;89(6):3790–3806. 851–853
24. Khan S, Zou Y, Amjad A, Gardezi A, Smith CL, Winters C, et al. Sequestration of CaMKII in dendritic spines in silico. *J Comput Neurosci*. 2011;31(3):581–594. 854–855
25. Kim M, Huang T, Abel T, Blackwell KT. Temporal sensitivity of protein kinase A activation in late-phase long term potentiation. *PLoS Comput Biol*. 2010;6(2):1–14. doi:10.1371/journal.pcbi.1000691. 856–858
26. Kim M, Park AJ, Havekes R, Chay A, Guercio LA, Oliveira RF, et al. Colocalization of protein kinase A with adenylyl cyclase enhances protein kinase A activity during induction of long-lasting long-term-potential. *PLoS Comput Biol*. 2011;7(6):e1002084. 859–862
27. Kim B, Hawes SL, Gillani F, Wallace LJ, Blackwell KT. Signaling pathways involved in striatal synaptic plasticity are sensitive to temporal pattern and exhibit spatial specificity. *PLoS Comput Biol*. 2013;9(3):e1002953. doi:10.1371/journal.pcbi.1002953. 863–866
28. Kötter R. Postsynaptic integration of glutamatergic and dopaminergic signals in the striatum. *Prog Neurobiol*. 1994;44(2):163–196. 867–868
29. Kuroda S, Schweighofer N, Kawato M. Exploration of signal transduction pathways in cerebellar long-term depression by kinetic simulation. *J Neurosci*. 2001;21(15):5693–702. 869–871
30. Li L, Stefan MI, Le Novère N. Calcium input frequency, duration and amplitude differentially modulate the relative activation of calcineurin and CaMKII. *PLoS ONE*. 2012;7(9):e43810+. doi:10.1371/journal.pone.0043810. 872–874
31. Mattioni M, Le Novère N. Integration of biochemical and electrical signaling – multiscale model of the medium spiny neuron of the striatum. *PLoS ONE*. 2013;8(7):e66811. doi:10.1371/journal.pone.0066811. 875–877
32. Miller P, Zhabotinsky AM, Lisman JE, Wang XJJ. The stability of a stochastic CaMKII switch: dependence on the number of enzyme molecules and protein turnover. *PLoS Biol*. 2005;3(4):e107+. doi:10.1371/journal.pbio.0030107. 878–880
33. Nair AG, Gutierrez-Arenas O, Eriksson O, Vincent P, Hellgren Kotaleski J. Sensing positive versus negative reward signals through adenylyl cyclase-coupled GPCRs in direct and indirect pathway striatal medium spiny neurons. *J Neurosci*. 2015;35(41):14017–14030. 881–884
34. Nakano T, Doi T, Yoshimoto J, Doya K. A kinetic model of dopamine- and calcium-dependent striatal synaptic plasticity. *PLoS Comput Biol*. 2010;6(2):e1000670. 885–887
35. Oliveira RF, Kim M, Blackwell KT. Subcellular location of PKA controls striatal plasticity: stochastic simulations in spiny dendrites. *PLoS Comput Biol*. 2012;8(2):e1002383. doi:10.1371/journal.pcbi.1002383. 888–890
36. Pepke S, Kinzer-Ursem T, Mihalas S, Kennedy MB. A dynamic model of interactions of Ca^{2+} , calmodulin, and catalytic subunits of Ca^{2+} /calmodulin-dependent protein kinase II. *PLoS Comput Biol*. 2010;6(2):e1000675. doi:10.1371/journal.pcbi.1000675. 891–894
-

-
37. Qi Z, Miller GW, Voit EO. The internal state of medium spiny neurons varies in response to different input signals. *BMC Syst Biol.* 2010;4(1):1–16. doi:10.1186/1752-0509-4-26. 895
896
897
38. Smolen P, Baxter DA, Byrne JH. A model of the roles of essential kinases in the induction and expression of late long-term potentiation. *Biophys J.* 2006;90(8):2760–2775. doi:10.1529/biophysj.105.072470. 898
899
900
39. Smolen P, Baxter DA, Byrne JH. Molecular constraints on synaptic tagging and maintenance of long-term potentiation: a predictive model. *PLoS Comput Biol.* 2012;8(8). 901
902
903
40. Sorokina O, Sorokin A, Armstrong JD. Towards a quantitative model of the post-synaptic proteome. *Mol Biosyst.* 2011;7:2813–2823. 904
905
906
doi:10.1039/C1MB05152K.
41. Stefan MI, Edelstein SJ, Le Novère N. An allosteric model of calmodulin explains differential activation of PP2B and CaMKII. *Proc Natl Acad Sci USA.* 2008;105(31):10768–10773. doi:10.1073/pnas.0804672105. 907
908
909
42. Zeng S, Holmes WR. The effect of noise on CaMKII activation in a dendritic spine during LTP induction. *J Neurophysiol.* 2010;103(4):1798–1808. 910
911
912
doi:10.1152/jn.91235.2008.
43. Zhabotinsky AM, Camp RN, Epstein IR, Lisman JE. Role of the neurogranin concentrated in spines in the induction of long-term potentiation. *J Neurosci.* 2006;26(28):7337–7347. doi:10.1523/jneurosci.0729-06.2006. 913
914
915
44. Hines ML, Morse T, Migliore M, Carnevale NT, Shepherd GM. ModelDB: A database to support computational neuroscience. *J Comput Neurosci.* 2004;17(1):7–11. doi:10.1023/B:JCNS.0000023869.22017.2e. 916
917
918
45. Chelliah V, Juty N, Ajmera I, Ali R, Dumousseau M, Glont M, et al. BioModels: ten-year anniversary. *Nucleic Acids Res.* 2015;43(Database issue):D542–548. 919
920
921
doi:10.1093/nar/gku1181.
46. Sivakumaran S, Hariharaputran S, Mishra J, Bhalla US. The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks. *Bionformatics.* 2003;19(3):408–15. 922
923
924
47. Lloyd CM, Lawson JR, Hunter PJ, Nielsen PF. The CellML model repository. *Bionformatics.* 2008;24(18):2122–3. doi:10.1093/bioinformatics/btn390. 925
926
48. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol.* 2010;4:92. 927
928
929
930
doi:10.1186/1752-0509-4-92.
49. Le Novère N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, et al. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol.* 2005;23(12):1509–15. doi:10.1038/nbt1156. 931
932
933
50. Stefan MI, Bartol TM, Sejnowski TJ, Kennedy MB. Multi-state modeling of biomolecules. *PLoS Comput Biol.* 2014;10(9):e1003844. 934
935
936
doi:10.1371/journal.pcbi.1003844.
51. Weisstein EW. Necklace; 2017. From MathWorld—A Wolfram Web Resource. Available from: <http://mathworld.wolfram.com/Necklace.html>. 937
938
-

-
52. Hernandez AI, Blace N, Crary JF, Serrano PA, Leitges M, Libien JM, et al. Protein kinase M ζ synthesis from a brain mRNA encoding an independent protein kinase C ζ catalytic domain: implications for the molecular mechanism of memory. *J Biol Chem.* 2003;278(41):40305–40316. doi:10.1074/jbc.M307065200. 939–942
53. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 2011;39(Database issue):D52–57. doi:10.1093/nar/gkq1237. 943–945
54. Whittaker V, Michaelson I, Kirkland RJA. The separation of synaptic vesicles from nerve-ending particles (synaptosomes). *Biochem J.* 1964;90(2):293. 946–947
55. Bai F, Weizmann FA. Synaptosome proteomics. In: *Subcellular Proteomics.* Springer; 2007. p. 77–98. 948–949
56. Sokolow S, Henkins KM, Williams IA, Vinters HV, Schmid I, Cole GM, et al. Isolation of synaptic terminals from Alzheimer’s disease cortex. *Cytom Part A.* 2012;81(3):248–254. doi:10.1002/cyto.a.22009. 950–952
57. Dieterich DC, Kreutz MR. Proteomics of the synapse—a quantitative approach to neuronal plasticity. *Mol Cell Proteomics.* 2016;15(2):368–381. doi:10.1074/mcp.R115.051482. 953–955
58. Vastagh C, Rodolosse A, Solymosi N, Liposits Z. Altered expression of genes encoding neurotransmitter receptors in GnRH neurons of proestrous mice. *Front Cell Neurosci.* 2016;10. 956–958
59. Silverman AJ, Hou-Yu A, Chen WP. Corticotropin-releasing factor synapses within the paraventricular nucleus of the hypothalamus. *Neuroendocrinology.* 1989;49(3):291–299. 959–960
60. Mystek P, Tworzydło M, Dziejzicka-Wasylewska M, Polit A. New insights into the model of dopamine D1 receptor and G-proteins interactions. *BBA-Mol Cell Res* 2015;1853(3):594–603. 962–964
61. Ahn JH, Sung JY, McAvoy T, Nishi A, Janssens V, Goris J, et al. The B γ /PR72 subunit mediates Ca $^{2+}$ -dependent dephosphorylation of DARPP-32 by protein phosphatase 2A. *Proc Natl Acad Sci USA.* 2007;104(23):9876–81. doi:10.1073/pnas.0703589104. 965–968
62. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2015;43(D1):D1049. doi:10.1093/nar/gku1179. 969–970
63. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res.* 2016;44(D1):D481. doi:10.1093/nar/gkv1351. 971–973
64. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* 2015;43(D1):D1071. doi:10.1093/nar/gku1011. 974–977
65. Jimeno-Yepes AJ, Sticco JC, Mork JG, Aronson AR. GeneRIF indexing: sentence selection based on machine learning. *BMC Bioinformatics.* 2013;14(1):171. 978–979
66. McKusick VA. Mendelian inheritance in man: a catalog of human genes and genetic disorders. vol. 1. JHU Press; 1998. 980–981

-
67. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's online Mendelian inheritance in man (OMIM®). *Nucleic Acids Res.* 2009;37(suppl 1):D793–D796. 982
983
68. Chen Y, Cunningham F, Rios D, McLaren WM, Smith J, Pritchard B, et al. Ensembl variation resources. *BMC Genomics.* 2010;11(1):293. 984
985
69. He X, Simpson TI. statbio/topOnto: topOnto v1.0; 2017. Available from: 986
<https://doi.org/10.5281/zenodo.819735>. 987
70. He X, Simpson TI. statbio/OntoSuite-Miner: OntoSuite-Miner v1.0; 2017. 988
Available from: <https://doi.org/10.5281/zenodo.819726>. 989
71. Qi Z, Miller GW, Voit EO. Computational systems analysis of dopamine 990
metabolism. *PLoS ONE.* 2008;3(6):e2444. 991
72. Sass MB, Lorenz AN, Green RL, Coleman RA. A pragmatic approach to 992
biochemical systems theory applied to an α -synuclein-based model of Parkinson's 993
disease. *J Neurosci Methods.* 2009;178(2):366–377. 994
73. Harrison PJ, Weinberger DR. Schizophrenia genes, gene expression, and 995
neuropathology: on the matter of their convergence. *Mol Psychiatr.* 996
2005;10(1):40. 997
74. Männistö PT, Kaakkola S. Catechol-O-methyltransferase (COMT): biochemistry, 998
molecular biology, pharmacology, and clinical efficacy of the new selective COMT 999
inhibitors. *Pharmacol Rev.* 1999;51(4):593–628. 1000
75. Weinshilboum RM, Otterness DM, Szumlanski CL. Methylation 1001
pharmacogenetics: catechol O-methyltransferase, thiopurine methyltransferase, 1002
and histamine N-methyltransferase. *Annu Rev Pharmacol.* 1999;39(1):19–52. 1003
76. De Araujo ME, Erhart G, Buck K, Müller-Holzner E, Hubalek M, Fiegl H, et al. 1004
Polymorphisms in the gene regions of the adaptor complex 1005
LAMTOR2/LAMTOR3 and their association with breast cancer risk. *PLoS* 1006
ONE. 2013;8(1):e53768. 1007
77. Sterratt D, Graham B, Gillies A, Willshaw D. Principles of Computational 1008
Modelling in Neuroscience. Cambridge, UK: Cambridge University Press; 2011. 1009
78. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its 1010
application to conduction and excitation in nerve. *J Physiol (Lond).* 1011
1952;117:500–544. 1012
79. Gray KA, Seal RL, Tweedie S, Wright MW, Bruford EA. A review of the new 1013
HGNC gene family resource. *Hum Genomics.* 2016;10:6. 1014
[doi:10.1186/s40246-016-0062-6](https://doi.org/10.1186/s40246-016-0062-6). 1015
80. Larson SD, Martone ME. NeuroLex.org: an online framework for neuroscience 1016
knowledge. *Front Neuroinform.* 2013;7:18. [doi:10.3389/fninf.2013.00018](https://doi.org/10.3389/fninf.2013.00018). 1017
81. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. 1018
The InterPro database, an integrated documentation resource for protein families, 1019
domains and functional sites. *Nucleic Acids Res.* 2001;29(1):37–40. 1020
[doi:10.1093/nar/29.1.37](https://doi.org/10.1093/nar/29.1.37). 1021
82. Sheridan RP, Venkataraghavan R. A systematic search for protein signature 1022
sequences. *Proteins.* 1992;14(1):16–28. [doi:10.1002/prot.340140105](https://doi.org/10.1002/prot.340140105). 1023
-

-
83. Orengo CA, Bateman A, Uversky V, editors. Protein families: relating protein sequence, structure, and function. Wiley; 2014. 1024
1025
84. Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*. 2007;23(4):401–7. 1026
doi:10.1093/bioinformatics/btl633. 1027
1028
85. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4):1165–1188. 1029
doi:10.1214/aos/1013699998. 1030
1031
86. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. 2006;22(13):1600–1607. doi:10.1093/bioinformatics/btl140. 1032
1033
1034

Appendix I

Acronyms

bp	base pairs
CDS	gene coding sequence
CME	Clathrin Mediated Endocytosis
DO	Disease Ontology
DOID	Disease Ontology Identifier
DIP	Database of Interacting Proteins
EnsVar	Ensembl Variation
GeneRIF	Gene Reference into Function
GEO	Gene Expression Omnibus
GO	Gene Ontology
GPCR	G-protein coupled receptor
GWAS	Genome Wide Association Study
HGNC	HUGO Gene Nomenclature Committee
HPRD	Human Protein Reference Database
ID	identifier
NCBI	National Center for Biotechnology Information
OMIM	Online Mendelian Inheritance in Man
PD	Parkinson's Disease
SNP	Single Nucleotide Polymorphism
PPI	Protein-Protein Interaction
PPIN	Protein-Protein-Interaction Network

List of Figures

1.1	The Central Dogma of Molecular Biology. Individual constitutive units are visualized (part A) as well as the union of all the units (part B). Part C shows examples of experimental tools that can be used to study the different levels of information.	3
1.2	The concept of PD map and its visualization (taken from Fujita et al. (2014)).	4
1.3	The tetrapartite synapse of principal neurons, consisting of the pre- and postsynaptic compartment, synaptic cleft, astrocytic endfeet, and extracellular matrix. The tightly regulated protein composition in the different regions can be seen. SV stands for synaptic vesicle (taken from Dieterich and Kreutz (2016)).	8
1.4	Work presented in this thesis. Blue boxes refer to data, orange boxes indicates processes, green boxes highlight analytical steps and outcomes are shown in magenta boxes. . . .	21
2.1	Overview of the mapping approach, showing input data and the obtained output. Dark blue boxes indicate raw data, light blue highlights the columns of interest in the respective files. Green boxes refer to processes and the magenta box highlights the outcome. */** highlight the information that was cross-linked between the files.	26
3.1	Work presented in Chapter 3, focusing on the workflow used to analyse data and generate a combined PD dataset. Dark blue boxes refer to published data, light blue boxes are generated datasets, green boxes describe processes and magenta boxes show outcomes.	36
3.2	Venn Diagram showing the overlap of genes significantly associated with PD (based on Entrez ID count). The different coloured ellipses represent Entrez IDs that have been associated with PD based on a microarray expression study. The four compared studies are: Chandrasekaran and Bonchev (2013) (red), Simunovic et al. (2009) (green), Zhang et al. (2005) (blue) and Moran et al. (2006) (turquoise). Numbers in overlapping regions indicate genes found in one or more studies.	42
3.3	Disease Ontology (DO) graph showing PD with its parent and child terms. Disease Ontology Identifiers (DOIDs) indicated in brackets are the official disease identifiers, used to extract associated genes. PD as well as all subtypes, indicated in boxes with blue borders are used in the analysis.	43
3.4	Venn Diagram showing the overlap of genes significantly associated with PD based on data retrieved with topONTO; PD subtype term results are included.	45

3.5	Single Nucleotide Polymorphism (SNP)-gene association classification of SNPs extracted from the Ensembl Variation (EnsVar) database, based on their position on the genome relative to the gene.	46
3.6	Venn diagram showing genes associated with PD based on a SNP (derived from EnsVar). The different circles indicate the relative position of the SNP to the gene (Figure 3.5). Numbers refer to gene numbers and one gene can be affected by several SNPs.	47
3.7	Schematic illustration of gene components, highlighting the gene coding sequence. . .	48
3.8	Venn diagram summarising all the filtering steps of the data retrieved with topONTO. Gene numbers with a star refer to gene sets that are part of the final PD associated gene set.	49
3.9	Venn diagram showing the overlap of PD associated gene sets retrieved from different sources.	50
4.1	Overview of work presented in Chapter 4. Input databases are shown in dark blue boxes (turquoise represents a special case addressed in the text and excluded in the final PPI set). Light blue stands for newly generated and curated datasets. A yellow box refers to processes, leading to an analytical result (pink boxes). Green boxes represent outputs of this chapters analysis or future results.	58
4.2	PPIs based on a certain “interaction type”. Data from all four source databases, not filtered for direct interactions is visualized (“unique PPIs mapped to human IDs” in Table 4.5). The x-axis shows the interaction type in alphabetical order.	68
4.3	Tree structure of the interaction type ontology branch and respective MI-ID.	69
4.4	PPIs in the dataset based on a certain “interaction type”. Data from all four source databases filtered for direct interactions is visualized. The x-axis shows the interaction type in alphabetical order.	70
4.5	Overlap of PPIs supplied by the different databases.	71
4.6	Overlap of PPIs between the three main databases that were kept to retrieve the final (human, unique, direct PPI dataset).	73
4.7	Human, unique, direct PPIs found in different source databases (based on information in the “source database” column of the mitab25 files). The x-axis shows the source database in alphabetical order.	74
4.8	Human, unique, direct PPI coverage in different source databases (based on “source database” count).	75
4.9	Human, unique, direct PPI coverage in different publications (based on “publication identifier” count).	75
5.1	Overview of used data, analytical processes and outcomes of Chapter 5. Dark blue boxes refer to published data, light blue boxes are generated datasets, yellow boxes refer to analytical steps, green boxes describe processes and magenta boxes show outcomes.	82
5.2	Increase in unique synaptic proteins, identified in different studies over the indicated years. Years without a bar reflect that no new data were added in those years.	86

5.3	Number of proteins found in the regional proteomic studies and their coverage in a specific number of studies (blue bar). The red bar indicates the percentage of proteins identified with the respective coverage relative to the studied dataset.	88
5.4	Coverage of presynaptic proteins, relative to year of first detection. X-axis label contains the year, number of studies published in that year (“studies”) and the number of studies published in the year and all coming years (“total studies”).	89
5.5	Coverage of postsynaptic proteins, relative to year of first detection. X-axis label contains the year, number of studies published in that year (“studies”) and the number of studies published in the year and all coming years (“total studies”).	89
5.6	Coverage of synaptosome proteins, relative to year of first detection. X-axis label contains the year, number of studies published in that year (“studies”) and the number of studies published in the year and all coming years (“total studies”).	90
5.7	Coverage of joint synaptic proteome proteins, relative to year of first detection. X-axis label contains the year, number of studies published in that year (“studies”) and the number of studies published in the year and all coming years (“total studies”).	90
5.8	Overlap of unique, human Entrez IDs of the genes identified in the presynapse, postsynapse and synaptosome proteome. All genes are included (minimum coverage = 1). . .	97
5.9	Overlap of unique human Entrez IDs of the genes identified in the presynapse, postsynapse and synaptosome proteomes (minimum coverage of considered proteins as indicated above).	98
5.10	GO enrichment of the set of genes expressed in all three datasets (presynapse, postsynapse and synaptosome). Enrichment was tested compared to the whole synapse as a background. Results for different GO ontologies are shown. Fisher test, the <code>elim</code> algorithm and Benjamini and Yekutieli multiple testing correction were used. Colour gradient (violet to blue) and size (small to large) reflect significance of the terms. . . .	101
5.11	Overlap of the three regional synaptic proteomes with PD associated genes (minimum coverage = 1).	103
5.12	Overlap of the three regional synaptic proteomes with PD associated genes (adjusted coverage of synaptic proteins in all regional datasets).	104
6.1	Overview of data, processes, and outcomes of Chapter 6. Dark blue boxes refer to published data, light blue boxes are generated datasets, green boxes describe processes and magenta boxes show outcomes.	114
6.2	Detailed overview including network clustering, enrichment, and key-protein as well as community detection processes in Section 6.3.2. Dark blue boxes refer to published data, light blue boxes are generated datasets, yellow boxes refer to analytical tools, green boxes describe processes and magenta boxes show outcomes.	122
6.3	Presynaptic PPINs. Different clustering algorithm results are highlighted. Red coloured nodes represent PD associated genes. Grey “background” shows network edges. . . .	124
6.4	Joint synaptic PPINs. Different clustering algorithm results are highlighted. Red coloured nodes represent PD associated genes. Grey “background” shows network edges. . . .	125

6.5	Clustering highlighting the overlap of PD associated genes in significantly PD enriched communities. The x-axis shows Entrez IDs and the y-axis indicates the dataset, algorithm and community number in which the community was found to be enriched for PD associated genes. “pre”, “post”, “synapt” and “synapse” refer to the presynaptic, postsynaptic, synaptosome and joint synaptic proteome.	130
6.6	Clustering highlighting the overlap of all genes in significantly PD enriched communities (including all community genes). x- and y-axis labelling are as in Figure 6.5. . . .	131
6.7	Coverage of different proteins in the PD enriched communities in the three enriched community clusters. Colours represent clusters. The x-axis indicates the coverage. Genes are associated to the coverage based on the number of PD enriched communities in the respective cluster they appear in.	133
6.8	Genes in Cluster 1 (minimum coverage of two). Opacity represents the coverage. Red squares are PD associated genes.	134
6.9	Genes in Cluster 2 (minimum coverage of two). Opacity represents the coverage. Red squares are PD associated genes.	138
6.10	Genes in Cluster 3 (minimum coverage of two). Opacity represents the coverage. Red squares are PD associated genes.	141
D.1	Postsynaptic PPINs. Different clustering algorithm results are highlighted. Red coloured nodes represent PD associated genes. Grey “background” represents network edges. . .	174
D.2	Synaptosome PPINs. Different clustering algorithm results are highlighted. Red coloured nodes represent PD associated genes. Grey “background” represents network edges. . .	175

List of Tables

1.1	Community clustering algorithms used to divide networks into communities.	15
3.1	Databases which EnsVar retrieves its data from.	39
3.2	Four gene expression microarray studies, used to obtain PD associated differently expressed genes. “Publication” refers to the study, “Brain Region” describes the tissue that was analysed, “Array Type” gives information about the array (all Affymetrix human GeneChips covering the whole human genome). The significance threshold shows p-value and fold-change information applied during original data analysis. “Associated Genes” shows the number of genes identified in the study, “Additional Information” contains further details, “Mapped Genes to Entrez ID” refers to the number of genes which were extracted from the study and successfully mapped to a unique Entrez ID and “Sample Size” refers to the number of samples (PD cases/controls) tested in the study. 41	41
3.3	Number of genes associated with PD based on the <code>topONTO</code> query. Columns refer to the different source databases and “Disease (Subtype)” refers to the different PD subtypes (see Figure 3.3). Numbers in brackets refer to the number of genes associated with only the disease subtype but not PD itself.	44
3.4	PD associated genes referenced in all three sources (ordered numerically by Entrez ID). “Genetic Evidence” can be “E” for EnsVar, “G” for Gene Reference into Function (GeneRIF) or “O” for Online Mendelian Inheritance in Man (OMIM); “Microarray Study” is “C” (Chandrasekaran and Bonchev, 2013), “M” (Moran et al., 2006), “S” (Simunovic et al., 2009) or “Z” (Zhang et al., 2005). “Literature Reference” lists the pubmedID of the paper(s) where the PD link was recorded.	51
4.1	15 standard mitab columns together with their content, including an example (randomly selected, not consistent between different columns).	60
4.2	Five of the most commonly used PPI databases. “Main Identifier” refers to the identifier used for the interactors (given in columns one and two of the psimi25 standard format files). Most recent release refers to the point of writing (May 2017).	62
4.3	Overview of the four databases of choice. All supply data in the mitab25 format. Two different, recent releases, per database are listed together with the data-file names. . . .	63

4.4	Overview of PPIs obtained from four databases (August 2016). Numbers represent PPI counts based on the row count in the file. Some PPIs may occur multiple times and duplicates such as (a-b and b-a are counted separately). “Human” means that both interactors were associated with the human taxid (9606). “Unique” PPIs represents the unique number of PPIs (filtered for mirrored duplicates). Direct interactions were obtained by filtering for direct-only interaction types.	66
4.5	Overview of PPIs obtained from four databases (March 2017). See caption Table 4.4. .	66
4.6	PPI count depending on different filter settings.	72
5.1	Synaptic proteome publications and respective datasets used in this study. “# genes” refers to the number of proteins, mapped to human Entrez IDs identified in the study. Studies are sorted based on presynapse, postsynapse, synaptosome and ascending depending on the year of publication. Studies highlighted with * contain two datasets. More details can be found in the Appendices C.1, C.2, C.3.	84
5.2	Genes detected with top coverage in the presynaptic proteome (ordered by coverage and alphabetically by gene name).	93
5.3	Genes detected with top coverage in the postsynaptic proteome (ordered by coverage and alphabetically by gene name).	94
5.4	Genes detected with top coverage in the synaptosome proteome (ordered by coverage and alphabetically by gene name).	95
5.5	Genes with top coverage, detected in the joint synaptic proteome (ordered by coverage and alphabetically by gene name). “pre”, “post” and “synapt” refer to the presynapse, postsynapse and synaptosome proteomes respectively.	96
5.6	Number of genes in the synaptic regional proteomes, filtered for coverage.	98
5.7	Significantly enriched functional GO terms of the gene sets specifically expressed in only one of the three regional synaptic proteome datasets. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, <i>elim</i> algorithm and Benjamini and Yekutieli multiple testing correction; significance p-value threshold was set to 0.05.	99
5.8	Overlap of PD associated genes with regional synaptic proteomes. “unique” refers to disease genes only overlapping with the indicated regional dataset and “total” refers to all the PD associated genes found in the respective proteome. Hypergeometric testing was carried out considering the full genome as a background (all human protein coding genes, referred to as “genome background”) as well as the synaptic proteome. Grey numbers indicate that the significance threshold of 0.05 was not reached.	105
5.9	Functional GO enrichment of PD associated genes expressed in the synapse and elsewhere. The gene sets of interest were enriched compared to all synaptic genes (“synapse”) as well as all human protein coding genes, apart from the ones expressed in the synapse (“elsewhere”). Results were obtained using the Fisher exact test, <i>elim</i> algorithm and Benjamini and Yekutieli multiple testing correction; significance p-value threshold was set to 0.05 (representation in alphabetical order).	108

6.1	Overview statistics of the PPINs of the presynaptic, postsynaptic, synaptosome and joint synaptic proteome. Number of genes refers to the number of proteins in the proteome (mapped to human Entrez IDs). Nodes are proteins and edges PPIs. “bcc” stands for biggest connected component. “Clustering Coefficient” refers to the global measure. “Density”, “Diameter” and “Power-law Alpha” values are overall network measures. Details can be found in Section 2.4.	115
6.2	Top 10 nodes with maximum degree and highest betweenness score in the four different networks. “deg” refers to degree and “btw” to betweenness. Numbers in parenthesis refer to the rank. Grey scaled numbers are outside the top 10; “-” indicates missing genes in the respective datasets. A PD link is indicated in the last column.	117
6.3	Synaptic PD associated genes with a top 10 degree value and their betweenness scores (together with the overall rank in the respective network). The table is sorted by coverage in the different datasets and based on the first available node degree based on the table columns. “degree” refers to node degree and “btw” to betweenness. Numbers in parenthesis refer to the rank. Grey numbers are outside the top 10; “-” indicates missing genes in the respective datasets.	120
6.4	Results obtained from the clusterings of the networks of the different regional datasets and using different clustering algorithms. Columns 3, 4, 5, 9 and 10 refer to the number of respective communities. “smaller 4” and “larger 200” refers to the number of nodes per community. Remaining columns refer to the number of nodes per community.	126
6.5	PD enriched communities in the different networks based on one of the four datasets and one of the five clustering algorithms (p-value < 0.05, after multiple testing correction). All enriched communities are listed, irrespective of their size. Rows are ordered based on dataset and algorithm. “synapse” refers to the joint synaptic proteome. Grey font highlights communities with less than four genes.	128
6.5	PD enriched communities in the different networks based on one of the four datasets and one of the five clustering algorithms (p-value < 0.05, after multiple testing correction). All enriched communities are listed, irrespective of their size. Rows are ordered based on dataset and algorithm. “synapse” refers to the joint synaptic proteome. Grey font highlights communities with less than four genes.	129
6.6	Three clusters of PD enriched communities. Cluster numbers and colour code as in Figures 6.5, 6.6 and 6.7. “Community Number” refers to the community in the original network; “Genes” refers to the number of genes in the community; “Communities in cluster” refers to the colour coded clusters; columns 8-12 refer to community counts in total and unique for the three clusters.	132
6.7	PD associated genes in Cluster 1. Ordered by coverage and Entrez ID.	135
6.8	GO terms enriched in at least two communities of Cluster 1 (alphabetical order); significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, <code>elim</code> algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values not supplied since different in distinct enriched clusters).	136
6.9	PD associated genes in Cluster 2. Ordered by coverage and Entrez ID.	139

6.10	GO terms enriched in at least two communities of Cluster 2 (alphabetical order); significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, <i>elim</i> algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values not supplied since different in distinct enriched clusters).	140
6.11	PD associated genes in Cluster 3. Ordered by coverage and Entrez ID.	141
6.12	GO terms enriched in at least two communities of Cluster 3; significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, <i>elim</i> algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values not supplied since different in distinct enriched clusters).	142
A.1	Genes manually identified to be linked to PD in reviewed papers (ordered alphabetically by Gene Name short). PMCID shows the reference where the gene-disease association was identified.	163
B.1	MI IDs specifying all direct interaction types, used to filter PPIs (orderd alphabetically based on the description).	165
B.2	MI IDs specifying source databases of PPIs (orderd alphabetically based on description).	167
C.1	Presynaptic proteome publications and respective datasets. "Count" shows the number of proteins, mapped to human Entrez IDs found in the study.	169
C.1	Presynaptic proteome publications and respective datasets. "Count" shows the number of proteins, mapped to human Entrez IDs found in the study.	170
C.2	Postsynaptic proteome publications and respective datasets. "Count" shows the number of proteins, mapped to human Entrez IDs found in the study.	170
C.2	Postsynaptic proteome publications and respective datasets. "Count" shows the number of proteins, mapped to human Entrez IDs found in the study.	171
C.3	Synaptosome proteome datasets and respective publications. "Count" shows the number of proteins, mapped to human Entrez IDs found in the study.	171
F.1	GO terms enriched in at least two communities in Cluster 1 (alphabetical order of GO terms); significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, <i>elim</i> algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values available upon request since different in distinct enriched clusters).	179
F.2	GO terms enriched in at least two communities in Cluster 2 (alphabetical order of GO terms); significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, <i>elim</i> algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values available upon request since different in distinct enriched clusters).	184

- F.3 GO terms enriched in at least two communities in Cluster 3 (alphabetical order of GO terms); significance p-value threshold was set to 0.05. The gene sets of interest were enriched compared to all genes expressed in the synapse. Results were obtained using the Fisher exact test, `elim` algorithm and Benjamini and Yekutieli multiple testing correction. Exact p-values available upon request since different in distinct enriched clusters). 185

Bibliography

- Aasly, J., Shi, M., Sossi, V., Stewart, T., Johansen, K., Wszolek, Z. K., Uitti, R. J., Hasegawa, K., Yokoyama, T., Zabetian, C., et al. (2012). Cerebrospinal fluid amyloid β and tau in *Irrk2* mutation carriers. *Neurology*, 78(1):55–61.
- Aghazadeh, Y. and Papadopoulos, V. (2016). The role of the 14-3-3 protein family in health, disease, and drug development. *Drug discovery today*, 21(2):278–287.
- Ahmed, Z., Timsah, Z., Suen, K. M., Cook, N. P., Lee IV, G. R., Lin, C.-C., Gagea, M., Marti, A. A., and Ladbury, J. E. (2015). Grb2 monomer-dimer equilibrium determines normal versus oncogenic function. *Nature communications*, 6.
- Aibara, S., Katahira, J., Valkov, E., and Stewart, M. (2015). The principal mrna nuclear export factor *nxf1*: *Nxt1* forms a symmetric binding platform that facilitates export of retroviral cte-rna. *Nucleic acids research*, 43(3):1883–1893.
- Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. (2004). Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580.
- Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2016). Hippie v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, page 985.
- Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–1607.
- Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). Mckusick’s online mendelian inheritance in man (omim®). *Nucleic acids research*, 37(suppl 1):D793–D796.
- Archbold, J. K., Whitten, A. E., Hu, S.-H., Collins, B. M., and Martin, J. L. (2014). Snare-ing the structures of *sec1/munc18* proteins. *Current opinion in structural biology*, 29:44–51.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.
- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

- Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752.
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2012). Ncbi geo: archive for functional genomics data sets-update. *Nucleic acids research*, 41(D1):D991–D995.
- Bartholome, O., Van den Ackerveken, P., Gil, J. S., de la Brassinne Bonardeaux, O., Leprince, P., Franzen, R., and Rogister, B. (2017). Puzzling out synaptic vesicle 2 family members functions. *Frontiers in molecular neuroscience*, 10.
- Bayés, À., Collins, M. O., Croning, M. D., van de Lagemaat, L. N., Choudhary, J. S., and Grant, S. G. (2012). Comparative study of human and mouse postsynaptic proteomes finds high compositional conservation and abundance differences for key synaptic proteins. *PLoS one*, 7(10):e46683.
- Bayés, À., Collins, M. O., Galtrey, C. M., Simonnet, C., Roy, M., Croning, M. D., Gou, G., van de Lagemaat, L. N., Milward, D., Whittle, I. R., et al. (2014). Human post-mortem synapse proteome integrity screening for proteomic studies of postsynaptic complexes. *Molecular brain*, 7(1):88.
- Bayés, À., Van De Lagemaat, L. N., Collins, M. O., Croning, M. D., Whittle, I. R., Choudhary, J. S., and Grant, S. G. (2011). Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature neuroscience*, 14(1):19–21.
- Beal, M. F. (1998). Mitochondrial dysfunction in neurodegenerative diseases. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1366(1):211–223.
- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.
- Berggård, T., Linse, S., and James, P. (2007). Methods for the detection and analysis of protein–protein interactions. *Proteomics*, 7(16):2833–2842.
- Biesemann, C., Grønberg, M., Luquet, E., Wichert, S. P., Bernard, V., Bungers, S. R., Cooper, B., Varoqueaux, F., Li, L., Byrne, J. A., et al. (2014). Proteomic screening of glutamatergic mouse brain synaptosomes isolated by fluorescence activated sorting. *The EMBO journal*, page e201386120.
- Blanco-Arias, P., Einholm, A. P., Mamsa, H., Concheiro, C., Gutiérrez-de Terán, H., Romero, J., Toustrup-Jensen, M. S., Carracedo, Á., Jen, J. C., Vilsen, B., et al. (2009). A c-terminal mutation of atp1a3 underscores the crucial role of sodium affinity in the pathophysiology of rapid-onset dystonia-parkinsonism. *Human molecular genetics*, 18(13):2370–2377.
- Bliss, C. A., Danforth, C. M., and Dodds, P. S. (2014). Estimation of global network statistics from incomplete data. *PLoS one*, 9(10):e108471.

- Blom, H., Bernhem, K., and Brismar, H. (2016). Sodium pump organization in dendritic spines. *Neurophotonics*, 3(4):041803–041803.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.
- Bobrovsky, P., Manuvera, V., Polina, N., Podgorny, O., Prusakov, K., Govorun, V., and Lazarev, V. (2016). Recombinant human peptidoglycan recognition proteins reveal antichlamydial activity. *Infection and immunity*, 84(7):2124–2130.
- Boldt, K., Van Reeuwijk, J., Lu, Q., Koutroumpas, K., Nguyen, T.-M. T., Texier, Y., Van Beersum, S. E., Horn, N., Willer, J. R., Mans, D. A., et al. (2016). An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. *Nature communications*, 7.
- Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60.
- Bonifati, V. (2014). Genetics of parkinson's disease—state of the art, 2013. *Parkinsonism & related disorders*, 20:S23–S28.
- Boyken, J., Grønberg, M., Riedel, D., Urlaub, H., Jahn, R., and Chua, J. J. E. (2013). Molecular profiling of synaptic vesicle docking sites reveals novel proteins but few differences between glutamatergic and gabaergic synapses. *Neuron*, 78(2):285–297.
- Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177.
- Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2008). On modularity clustering. *IEEE transactions on knowledge and data engineering*, 20(2):172–188.
- Bredesen, D. E., Rao, R. V., and Mehlen, P. (2006). Cell death in the nervous system. *Nature*, 443(7113):796.
- Breydo, L., Wu, J. W., and Uversky, V. N. (2012). α -synuclein misfolding and parkinson's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1822(2):261–285.
- Brinkmalm, A., Brinkmalm, G., Honer, W. G., Moreno, J. A., Jakobsson, J., Mallucci, G. R., Zetterberg, H., Blennow, K., and Öhrfelt, A. (2014). Targeting synaptic pathology with a novel affinity mass spectrometry approach. *Molecular & Cellular Proteomics*, 13(10):2584–2592.
- Brun, C., Herrmann, C., and Guénoche, A. (2004). Clustering proteins from interaction networks for the prediction of cellular functions. *BMC bioinformatics*, 5(1):95.
- Burré, J., Beckhaus, T., Schägger, H., Corvey, C., Hofmann, S., Karas, M., Zimmermann, H., and Volkhardt, W. (2006). Analysis of the synaptic vesicle proteome using three gel-based protein separation techniques. *Proteomics*, 6(23):6250–6262.

- Calì, T., Ottolini, D., and Brini, M. (2014). Calcium signaling in parkinson's disease. *Cell and tissue research*, 357(2):439–454.
- Caligiore, D., Helmich, R. C., Hallett, M., Moustafa, A. A., Timmermann, L., Toni, I., and Baldassarre, G. (2016). Parkinson's disease as a system-level disorder. *npj Parkinson's Disease*, 2:16025.
- Catterall, W. A. (2011). Voltage-gated calcium channels. *Cold Spring Harbor perspectives in biology*, 3(8):a003947.
- Chandrasekaran, S. and Bonchev, D. (2013). A network view on parkinson's disease. *Computational and structural biotechnology journal*, 7(8):1–18.
- Chang, R. Y. K., Etheridge, N., Nouwens, A. S., and Dodd, P. R. (2015). Swath analysis of the synaptic proteome in alzheimer's disease. *Neurochemistry international*, 87:1–12.
- Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al. (2016). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, page 1102.
- Chen, X., Duan, L.-H., Luo, P.-c., Hu, G., Yu, X., Liu, J., Lu, H., and Liu, B. (2016). Fbxo6-mediated ubiquitination and degradation of erol1 inhibits endoplasmic reticulum stress-induced apoptosis. *Cellular Physiology and Biochemistry*, 39(6):2501–2508.
- Chen, X. and Pan, W. (2014). The treatment strategies for neurodegenerative diseases by integrative medicine. *Integrative Medicine International*, 1(4):223–225.
- Chen, Y., Cunningham, F., Rios, D., McLaren, W. M., Smith, J., Pritchard, B., Spudich, G. M., Brent, S., Kulesha, E., Marin-Garcia, P., et al. (2010). Ensembl variation resources. *BMC genomics*, 11(1):293.
- Cheng, D., Hoogenraad, C. C., Rush, J., Ramm, E., Schlager, M. A., Duong, D. M., Xu, P., Wijayawardana, S. R., Hanfelt, J., Nakagawa, T., et al. (2006). Relative and absolute quantification of postsynaptic density proteome isolated from rat forebrain and cerebellum. *Molecular & cellular proteomics*, 5(6):1158–1170.
- Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6):066111.
- Cohen, L. D., Zuchman, R., Sorokina, O., Müller, A., Dieterich, D. C., Armstrong, J. D., Ziv, T., and Ziv, N. E. (2013). Metabolic turnover of synaptic proteins: kinetics, interdependencies and implications for synaptic maintenance. *PloS one*, 8(5):e63191.
- Collins, M. O., Husi, H., Yu, L., Brandon, J. M., Anderson, C. N., Blackstock, W. P., Choudhary, J. S., and Grant, S. G. (2006). Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *Journal of neurochemistry*, 97(s1):16–23.

- Compta, Y., Ezquerra, M., Muñoz, E., Tolosa, E., Valldeoriola, F., Rios, J., Cámara, A., Fernández, M., Buongiorno, M. T., and Marti, M. J. (2011). High cerebrospinal tau levels are associated with the rs242557 tau gene variant and low cerebrospinal β -amyloid in parkinson disease. *Neuroscience letters*, 487(2):169–173.
- Cook, C., Stetler, C., and Petrucelli, L. (2012). Disruption of protein quality control in parkinson's disease. *Cold Spring Harbor perspectives in medicine*, 2(5):a009423.
- Coultrap, S. J. and Bayer, K. U. (2014). Nitric oxide induces ca^{2+} -independent activity of the ca^{2+} /calmodulin-dependent protein kinase ii (camkii). *Journal of Biological Chemistry*, 289(28):19458–19465.
- Credle, J. J., George, J. L., Wills, J., Duka, V., Shah, K., Lee, Y.-C., Rodriguez, O., Simkins, T., Winter, M., Moechars, D., et al. (2015). Gsk-3 β dysregulation contributes to parkinson's-like pathophysiology with associated region-specific phosphorylation and accumulation of tau and α -synuclein. *Cell death and differentiation*, 22(5):838.
- Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., et al. (2014). The reactome pathway knowledge-base. *Nucleic acids research*, 42(D1):D472–D477.
- Cruceanu, C., Alda, M., Grof, P., Rouleau, G. A., and Turecki, G. (2012). Synapsin ii is involved in the molecular pathway of lithium treatment in bipolar disorder. *PLoS one*, 7(2):e32680.
- Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9.
- Dahlhaus, M., Li, K. W., van der Schors, R. C., Saiepour, M. H., van Nierop, P., Heimel, J. A., Hermans, J. M., Loos, M., Smit, A. B., and Levelt, C. N. (2011). The synaptic proteome during development and plasticity of the mouse visual cortex. *Molecular & Cellular Proteomics*, 10(5):M110–005413.
- Danos, V., Feret, J., Fontana, W., and Krivine, J. (2008). Abstract interpretation of cellular signalling networks. In *International Workshop on Verification, Model Checking, and Abstract Interpretation*, pages 83–97. Springer.
- Daugaard, M., Rohde, M., and Jäättelä, M. (2007). The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions. *FEBS letters*, 581(19):3702–3710.
- Davis, D., Yaveroğlu, Ö. N., Malod-Dognin, N., Stojmirovic, A., and Pržulj, N. (2015). Topology-function conservation in protein–protein interaction networks. *Bioinformatics*, 31(10):1632–1639.
- Dawson, V. L. and Dawson, T. M. (1996). Nitric oxide neurotoxicity. *Journal of chemical neuroanatomy*, 10(3):179–190.
- De Lau, L. M. and Breteler, M. M. (2006). Epidemiology of parkinson's disease. *The Lancet Neurology*, 5(6):525–535.

- Desai, S., Kumar, A., Laskar, S., and Pandey, B. (2014). Differential roles of atf-2 in survival and dna repair contributing to radioresistance induced by autocrine soluble factors in a549 lung cancer cells. *Cellular signalling*, 26(11):2424–2435.
- Dexter, D. T. and Jenner, P. (2013). Parkinson disease: from pathology to molecular disease mechanisms. *Free Radical Biology and Medicine*, 62:132–144.
- Di Maio, V. (2008). Regulation of information passing by synaptic transmission: a short review. *Brain research*, 1225:26–38.
- Dieterich, D. C. and Kreutz, M. R. (2016). Proteomics of the synapse—a quantitative approach to neuronal plasticity. *Molecular & Cellular Proteomics*, 15(2):368–381.
- Distler, U., Schmeisser, M. J., Pelosi, A., Reim, D., Kuharev, J., Weiczner, R., Baumgart, J., Boeckers, T. M., Nitsch, R., Vogt, J., et al. (2014). In-depth protein profiling of the postsynaptic density from mouse hippocampus using data-independent acquisition proteomics. *Proteomics*, 14(21-22):2607–2613.
- Dosemeci, A., Makusky, A. J., Jankowska-Stephens, E., Yang, X., Slotta, D. J., and Markey, S. P. (2007). Composition of the synaptic psd-95 complex. *Molecular & Cellular Proteomics*, 6(10):1749–1760.
- Dosemeci, A., Tao-Cheng, J.-H., Vinade, L., and Jaffe, H. (2006). Preparation of postsynaptic density fraction from hippocampal slices and proteomic analysis. *Biochemical and biophysical research communications*, 339(2):687–694.
- Duce, J. A., Tsatsanis, A., Cater, M. A., James, S. A., Robb, E., Wikhe, K., Leong, S. L., Perez, K., Johanssen, T., Greenough, M. A., et al. (2010). Iron-export ferroxidase activity of β -amyloid precursor protein is inhibited by zinc in alzheimer's disease. *Cell*, 142(6):857–867.
- Dziarski, R. (2004). Peptidoglycan recognition proteins (pgrps). *Molecular immunology*, 40(12):877–886.
- Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210.
- Eggers, C., Schwartz, F., Pedrosa, D. J., Kracht, L., and Timmermann, L. (2014). Parkinson's disease subtypes show a specific link between dopaminergic and glucose metabolism in the striatum. *PloS one*, 9(5):e96629.
- Emmert-Streib, F., Dehmer, M., and Haibe-Kains, B. (2014). Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in cell and developmental biology*, 2.
- Esposito, G., Ana Clara, F., and Verstreken, P. (2012). Synaptic vesicle trafficking and parkinson's disease. *Developmental neurobiology*, 72(1):134–144.
- Exome Variant Server (2012). NHLBI GO Exome Sequencing Project (ESP), Seattle, WA. <http://evs.gs.washington.edu/EVS/>. Accessed: 2017-02-27.

- Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human molecular genetics*, 23(22):5866–5878.
- Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2016). The reactome pathway knowledgebase. *Nucleic acids research*, 44(D1):D481–D487.
- Farr, C. D., Gafken, P. R., Norbeck, A. D., Doneanu, C. E., Stapels, M. D., Barofsky, D. F., Minami, M., and Saugstad, J. A. (2004). Proteomic analysis of native metabotropic glutamate receptor 5 protein complexes reveals novel molecular constituents. *Journal of neurochemistry*, 91(2):438–450.
- Fassio, A., Patry, L., Congia, S., Onofri, F., Piton, A., Gauthier, J., Pozzi, D., Messa, M., Defranchi, E., Fadda, M., et al. (2011). Syn1 loss-of-function mutations in autism and partial epilepsy cause impaired synaptic function. *Human molecular genetics*, 20(12):2297–2307.
- Fernández, E., Collins, M. O., Uren, R. T., Kopanitsa, M. V., Komiyama, N. H., Croning, M. D., Zografos, L., Armstrong, J. D., Choudhary, J. S., and Grant, S. G. (2009). Targeted tandem affinity purification of psd-95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Molecular systems biology*, 5(1):269.
- Filiou, M. D., Bisle, B., Reckow, S., Teplytska, L., Maccarrone, G., and Turck, C. W. (2010). Profiling of mouse synaptosome proteome and phosphoproteome by ief. *Electrophoresis*, 31(8):1294–1301.
- Föcking, M., Dicker, P., Lopez, L. M., Hryniewiecka, M., Wynne, K., English, J. A., Cagney, G., and Cotter, D. R. (2016). Proteomic analysis of the postsynaptic density implicates synaptic function and energy pathways in bipolar disorder. *Translational Psychiatry*, 6(11):e959.
- Folador, E. L., Hassan, S. S., Lemke, N., Barh, D., Silva, A., Ferreira, R. S., and Azevedo, V. (2014). An improved interolog mapping-based computational prediction of protein–protein interactions with increased network coverage. *Integrative Biology*, 6(11):1080–1087.
- Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., et al. (2015). Cosmic: exploring the world’s knowledge of somatic mutations in human cancer. *Nucleic acids research*, 43(D1):D805–D811.
- Franco, I. S. and Shuman, H. A. (2012). A pathogen’s journey in the host cell: Bridges between actin and traffic. *Bioarchitecture*, 2(2):38–42.
- Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

- Fujita, K. A., Ostaszewski, M., Matsuoka, Y., Ghosh, S., Glaab, E., Trefois, C., Crespo, I., Perumal, T. M., Jurkowski, W., Antony, P. M., et al. (2014). Integrating pathways of parkinson's disease in a molecular interaction map. *Molecular neurobiology*, 49(1):88–102.
- Fury, W., Batliwalla, F., Gregersen, P. K., and Li, W. (2006). Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 5531–5534. IEEE.
- Geifman, N., Monsonogo, A., and Rubin, E. (2010). The neural/immune gene ontology: clipping the gene ontology for neurological and immunological systems. *BMC bioinformatics*, 11(1):458.
- Giardine, B., Riemer, C., Hefferon, T., Thomas, D., Hsu, F., Zielenski, J., Sang, Y., Elnitski, L., Cutting, G., Trumbower, H., et al. (2007). Phencode: connecting encode data with mutations and phenotype. *Human mutation*, 28(6):554–562.
- Gingras, A.-C. and Raught, B. (2012). Beyond hairballs: the use of quantitative mass spectrometry data to understand protein–protein interactions. *FEBS letters*, 586(17):2723–2731.
- Glaab, E. and Schneider, R. (2015). Comparative pathway and network analysis of brain transcriptome changes during adult aging and in parkinson's disease. *Neurobiology of disease*, 74:1–13.
- Goodier, J. L., Cheung, L. E., and Kazazian Jr, H. H. (2012). Mov10 rna helicase is a potent inhibitor of retrotransposition in cells. *PLoS genetics*, 8(10):e1002941.
- Gorini, G., Ponomareva, O., Shores, K. S., Person, M. D., Harris, R. A., and Mayfield, R. D. (2010). Dynamin-1 co-associates with native mouse brain bk ca channels: Proteomics analysis of synaptic protein complexes. *FEBS letters*, 584(5):845–851.
- Greenamyre, J. T., Sherer, T. B., Betarbet, R., and Panov, A. V. (2001). Complex i and parkinson's disease. *IUBMB life*, 52(3-5):135–141.
- Grønborg, M., Pavlos, N. J., Brunk, I., Chua, J. J., Münster-Wandowski, A., Riedel, D., Ahnert-Hilger, G., Urlaub, H., and Jahn, R. (2010). Quantitative comparison of glutamatergic and gabaergic synaptic vesicles unveils selectivity for few proteins including mal2, a novel synaptic vesicle protein. *Journal of Neuroscience*, 30(1):2–12.
- Hallett, P. J. and Standaert, D. G. (2004). Rationale for and use of nmda receptor antagonists in parkinson's disease. *Pharmacology & therapeutics*, 102(2):155–174.
- Han, J.-D. J., Bertin, N., Tong, H., Goldberg, D. S., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88.

- Hawrylycz, M., Ng, L., Feng, D., Sunkin, S., Szafer, A., and Dang, C. (2014). The allen brain atlas. In *Springer Handbook of Bio-/Neuroinformatics*, pages 1111–1126. Springer.
- Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., Van De Lagemaat, L. N., Smith, K. A., Ebbert, A., Riley, Z. L., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489(7416):391–399.
- He, X. and Simpson, T. I. (2017a). statbio/ontosuite-miner: Ontosuite-miner v1.0.
- He, X. and Simpson, T. I. (2017b). statbio/toponto: toponto v1.0.
- Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., Von Mering, C., et al. (2004a). The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nature biotechnology*, 22(2):177–183.
- Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., et al. (2004b). Intact: an open source molecular interaction database. *Nucleic acids research*, 32(suppl 1):D452–D455.
- Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. *Database*, 2016:bav096.
- HGNC (1979). HUGO Gene Nomenclature Committee. <http://www.genenames.org/about/overview>. Accessed: 2017-03-09.
- Hirakis, S. P., Boras, B. W., Votapka, L. W., Malmstrom, R. D., McCulloch, A. D., and Amaro, R. E. (2015). Bridging scales through multiscale modeling: a case study on protein kinase a. *Frontiers in physiology*, 6:250.
- Hirsch, E. C., Vyas, S., and Hunot, S. (2012). Neuroinflammation in parkinson's disease. *Parkinsonism & related disorders*, 18:S210–S212.
- Hosaka, M. and Südhof, T. C. (1999). Homo- and heterodimerization of synapsins. *Journal of Biological Chemistry*, 274(24):16747–16753.
- Howard, M. A., Elias, G. M., Elias, L. A., Swat, W., and Nicoll, R. A. (2010). The role of sap97 in synaptic glutamate receptor dynamics. *Proceedings of the National Academy of Sciences*, 107(8):3805–3810.
- Hu, C., Chen, W., Myers, S. J., Yuan, H., and Traynelis, S. F. (2016). Human grin2b variants in neurodevelopmental disorders. *Journal of pharmacological sciences*, 132(2):115–121.
- Hurley, M. J. and Dexter, D. T. (2012). Voltage-gated calcium channels and parkinson's disease. *Pharmacology & therapeutics*, 133(3):324–333.

- Hwang, W., Cho, Y.-R., Zhang, A., and Ramanathan, M. (2006). A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms for Molecular Biology*, 1(1):24.
- Imbrici, P., Camerino, D. C., and Tricarico, D. (2013). Major channels involved in neuropsychiatric disorders and therapeutic perspectives. *Frontiers in genetics*, 4.
- Irwin, D. J., Lee, V. M.-Y., and Trojanowski, J. Q. (2013). Parkinson's disease dementia: convergence of [alpha]-synuclein, tau and amyloid-[beta] pathologies. *Nature Reviews Neuroscience*, 14(9):626–636.
- Jimeno-Yepes, A. J., Sticco, J. C., Mork, J. G., and Aronson, A. R. (2013). Generif indexing: sentence selection based on machine learning. *BMC bioinformatics*, 14(1):171.
- Johnson, K. A., Conn, P. J., and Niswender, C. M. (2009). Glutamate receptors as therapeutic targets for parkinson's disease. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, 8(6):475–491.
- Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate Discrete Distributions, Set*, volume 444. John Wiley & Sons.
- Jordan, B. A., Fernholz, B. D., Boussac, M., Xu, C., Grigorean, G., Ziff, E. B., and Neubert, T. A. (2004). Identification and verification of novel rodent postsynaptic density proteins. *Molecular & Cellular Proteomics*, 3(9):857–871.
- Joshi, S. and Whiteheart, S. W. (2017). The nuts and bolts of the platelet release reaction. *Platelets*, 28(2):129–137.
- Jupp, S., Burdett, T., Leroy, C., and Parkinson, H. E. (2015). A new ontology lookup service at embl-ebi. In *SWAT4LS*, pages 118–119.
- Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., Bader, G. D., Xenarios, I., Wojcik, J., Sherman, D., et al. (2007). Broadening the horizon—level 2.5 of the HUPO-PSI format for molecular interactions. *BMC biology*, 5(1):44.
- Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C. J., Binder, J. X., Malone, J., Vasant, D., et al. (2014). Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research*, page 1011.
- Kitano, H. (2002a). Computational systems biology. *Nature*, 420(6912):206–210.
- Kitano, H. (2002b). Systems biology: a brief overview. *Science*, 295(5560):1662–1664.
- Kleiger, G. and Mayor, T. (2014). Perilous journey: a tour of the ubiquitin–proteasome system. *Trends in cell biology*, 24(6):352–359.

- Klein, C. and Westenberger, A. (2012). Genetics of parkinson's disease. *Cold Spring Harbor perspectives in medicine*, 2(1):a008888.
- Knight, R. and Verkhatsky, A. (2010). Neurodegenerative diseases: failures in brain connectivity? *Cell death and differentiation*, 17(7):1069.
- Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M., et al. (2017). The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1):D865–D876.
- Koschützki, D. and Schreiber, F. (2008). Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene regulation and systems biology*, 2:193.
- Kowal, S. L., Dall, T. M., Chakrabarti, R., Storm, M. V., and Jain, A. (2013). The current and projected economic burden of parkinson's disease in the united states. *Movement Disorders*, 28(3):311–318.
- Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., and Maglott, D. R. (2014). Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(D1):D980–D985.
- Laßek, M., Weingarten, J., and Volknaandt, W. (2015). The synaptic proteome. *Cell and tissue research*, 359(1):255–265.
- Lee, A. S., De Jesús-Cortés, H., Kabir, Z. D., Knobbe, W., Orr, M., Burgdorf, C., Huntington, P., McDaniel, L., Britt, J. K., Hoffmann, F., et al. (2016). The neuropsychiatric disease-associated gene *cacnalc* mediates survival of young hippocampal neurons. *Eneuro*, 3(2):ENEURO–0006.
- Lev, N., Melamed, E., and Offen, D. (2003). Apoptosis and parkinson's disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 27(2):245–250.
- Li, J.-Q., Tan, L., and Yu, J.-T. (2014). The role of the *lrrk2* gene in parkinsonism. *Molecular neurodegeneration*, 9(1):47.
- Lowenstein, E., Daly, R., Batzer, A., Li, W., Margolis, B., Lammers, R., Ullrich, A., Skolnik, E., Bar-Sagi, D., and Schlessinger, J. (1992). The sh2 and sh3 domain-containing protein *grb2* links receptor tyrosine kinases to ras signaling. *Cell*, 70(3):431–442.
- Lu, W., Wan, X., Liu, B., Rong, X., Zhu, L., Li, P., Li, J., Wang, L., Cui, L., and Wang, X. (2014). Specific changes of serum proteins in parkinson's disease patients. *PloS one*, 9(4):e95684.
- Lüscher, C. and Isaac, J. T. (2009). The synapse: center stage for many brain diseases. *The Journal of physiology*, 587(4):727–729.
- Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl 1):D54–D58.

- Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2010). Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 39(suppl_1):D52–D57.
- Magrinelli, F., Picelli, A., Tocco, P., Federico, A., Roncari, L., Smania, N., Zanette, G., and Tamburin, S. (2016). Pathophysiology of motor dysfunction in parkinson's disease as the rationale for drug treatment and rehabilitation. *Parkinson's Disease*, 2016.
- Martins-Branco, D., Esteves, A. R., Santos, D., Arduino, D. M., Swerdlow, R. H., Oliveira, C. R., Januario, C., and Cardoso, S. M. (2012). Ubiquitin proteasome system in parkinson's disease: A keeper or a witness? *Experimental neurology*, 238(2):89–99.
- McCain, J. (2013). The mapk (erk) pathway: investigational combinations for the treatment of braf-mutated metastatic melanoma. *Pharmacy and Therapeutics*, 38(2):96.
- McKusick, V. A. (1998). *Mendelian inheritance in man: a catalog of human genes and genetic disorders*, volume 1. JHU Press.
- Mclean, C., Xin, H., Simpson, I. T., and Armstrong, D. J. (2016). Improved Functional Enrichment Analysis of Biological Networks using Scalable Modularity Based Clustering. *Journal of Proteomics & Bioinformatics*, 9(1):9–18.
- McMahon, H. T. and Boucrot, E. (2011). Molecular mechanism and physiological functions of clathrin-mediated endocytosis. *Nature reviews. Molecular cell biology*, 12(8):517.
- Moran, L. B., Duke, D., Deprez, M., Dexter, D., Pearce, R., and Graeber, M. (2006). Whole genome expression profiling of the medial and lateral substantia nigra in parkinson's disease. *Neurogenetics*, 7(1):1–11.
- Morciano, M., Beckhaus, T., Karas, M., Zimmermann, H., and Volkandt, W. (2009). The proteome of the presynaptic active zone: from docked synaptic vesicles to adhesion molecules and maxi-channels. *Journal of neurochemistry*, 108(3):662–675.
- Morciano, M., Burré, J., Corvey, C., Karas, M., Zimmermann, H., and Volkandt, W. (2005). Immunolocalization of two synaptic vesicle pools from synaptosomes: a proteomics analysis. *Journal of neurochemistry*, 95(6):1732–1745.
- Mortiboys, H., Furnston, R., Bronstad, G., Aasly, J., Elliott, C., and Bandmann, O. (2015). UdecA exerts beneficial effect on mitochondrial dysfunction in lrrk2g2019s carriers and in vivo. *Neurology*, 85(10):846–852.
- Muangpaisan, W., Mathews, A., Hori, H., and Seidel, D. (2011). A systematic review of the worldwide prevalence and incidence of parkinson's disease. *Journal of the Medical Association of Thailand*, 94(6):749.
- Murtagh, J., Eddy, R., Shows, T., Moss, J., and Vaughan, M. (1991). Different forms of go alpha mRNA arise by alternative splicing of transcripts from a single gene on human chromosome 16. *Molecular and cellular biology*, 11(2):1146–1155.

- NCBI, R. C. (2016). Database resources of the national center for biotechnology information. *Nucleic acids research*, 44(D1):D7.
- Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.
- Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133.
- Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104.
- Newman, M. E. (2006b). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.
- Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.
- Oda, K., Matsuoka, Y., Funahashi, A., and Kitano, H. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular systems biology*, 1(1).
- Oliva, C., Escobedo, P., Astorga, C., Molina, C., and Sierralta, J. (2012). Role of the maguk protein family in synapse formation and function. *Developmental neurobiology*, 72(1):57–72.
- Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., et al. (2013). The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, page gkt1115.
- Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F. S., Cesareni, G., et al. (2012). Protein interaction data curation: the international molecular exchange (imex) consortium. *Nature methods*, 9(4):345–350.
- Ostrerova, N., Petrucelli, L., Farrer, M., Mehta, N., Choi, P., Hardy, J., and Wolozin, B. (1999). α -synuclein shares physical and functional homology with 14-3-3 proteins. *Journal of Neuroscience*, 19(14):5782–5791.
- Ozbabacan, S. E. A., Engin, H. B., Gursoy, A., and Keskin, O. (2011). Transient protein–protein interactions. *Protein Engineering Design and Selection*, 24(9):635–648.
- Paliwal, M. and Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert systems with applications*, 36(1):2–17.
- Paoletti, P., Bellone, C., and Zhou, Q. (2013). Nmda receptor subunit diversity: impact on receptor properties, synaptic plasticity and disease. *Nature Reviews. Neuroscience*, 14(6):383.

- Patil, A., Kinoshita, K., and Nakamura, H. (2010). Hub promiscuity in protein-protein interaction networks. *International journal of molecular sciences*, 11(4):1930–1943.
- Paul, M. K. and Mukhopadhyay, A. K. (2004). Tyrosine kinase—role and significance in cancer. *International journal of medical sciences*, 1(2):101.
- Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., and Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData mining*, 4(1):10.
- Peng, J., Kim, M. J., Cheng, D., Duong, D. M., Gygi, S. P., and Sheng, M. (2004). Semi-quantitative proteomic analysis of rat forebrain postsynaptic density fractions by mass spectrometry. *Journal of Biological Chemistry*.
- Perier, C., Bové, J., and Vila, M. (2012). Mitochondria and programmed cell death in parkinson’s disease: apoptosis and beyond. *Antioxidants & redox signaling*, 16(9):883–895.
- Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010). Transient protein-protein interactions: structural, functional, and network properties. *Structure*, 18(10):1233–1243.
- Pinton, P., Giorgi, C., Siviero, R., Zecchini, E., and Rizzuto, R. (2008). Calcium and apoptosis: Er-mitochondria ca^{2+} transfer in the control of apoptosis. *Oncogene*, 27(50):6407.
- Pizzuti, C. and Rombo, S. E. (2014). Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 30(10):1343–1352.
- Prasad, T. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human protein reference database-2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772.
- Priller, C., Bauer, T., Mitteregger, G., Krebs, B., Kretschmar, H. A., and Herms, J. (2006). Synapse formation and function is modulated by the amyloid precursor protein. *Journal of Neuroscience*, 26(27):7212–7221.
- Qureshi, H. Y., Li, T., MacDonald, R., Cho, C. M., Leclerc, N., and Paudel, H. K. (2013). Interaction of 14-3-3 ζ with microtubule-associated protein tau within alzheimer’s disease neurofibrillary tangles. *Biochemistry*, 52(37):6445–6455.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Racanelli, V. and Rehmann, B. (2006). The liver as an immunological organ. *Hepatology*, 43(S1).

- Ran, J., Li, H., Fu, J., Liu, L., Xing, Y., Li, X., Shen, H., Chen, Y., Jiang, X., Li, Y., et al. (2013). Construction and analysis of the protein-protein interaction network related to essential hypertension. *BMC systems biology*, 7(1):32.
- Rao, V. S., Srinivas, K., Sujini, G., and Kumar, G. (2014). Protein-protein interaction detection: methods and analysis. *International journal of proteomics*, 2014.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555.
- Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(1):016110.
- Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007). Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407.
- Robb, G. B. and Rana, T. M. (2007). Rna helicase a interacts with risc in human cells and functions in risc loading. *Molecular cell*, 26(4):523–537.
- Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123.
- Rüenauer, K., Menon, R., Svensson, M., Carlsson, J., Vogel, W., Andrén, O., Nowak, M., and Perner, S. (2014). Prognostic significance of ywhaz expression in localized prostate cancer. *Prostate cancer and prostatic diseases*, 17(4):310.
- Ryu, K.-Y., Maehr, R., Gilchrist, C. A., Long, M. A., Bouley, D. M., Mueller, B., Ploegh, H. L., and Kopito, R. R. (2007). The mouse polyubiquitin gene *ubc* is essential for fetal liver development, cell-cycle progression and stress tolerance. *The EMBO journal*, 26(11):2693–2706.
- Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(suppl 1):D449–D451.
- Satoh, K., Takeuchi, M., Oda, Y., Deguchi-Tawarada, M., Sakamoto, Y., Matsubara, K., Nagasu, T., and Takai, Y. (2002). Identification of activity-regulated proteins in the postsynaptic density fraction. *Genes to Cells*, 7(2):187–197.
- Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. (2012). HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PloS one*, 7(2):e31826.
- Schapira, A. H. (2013). Calcium dysregulation in parkinson's disease. *Brain*, 136(7):2015–2016.
- Schlachetzki, J. C. and Winkler, J. (2015). The innate immune system in parkinson's disease: a novel target promoting endogenous neuroregeneration. *Neural regeneration research*, 10(5):704.

- Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W. A. (2011). Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946.
- Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W. A. (2012). Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946.
- Schwenk, J., Harmel, N., Brechet, A., Zolles, G., Berkefeld, H., Müller, C. S., Bildl, W., Baehrens, D., Hüber, B., Kulik, A., et al. (2012). High-resolution proteomics unravel architecture and molecular diversity of native ampa receptor complexes. *Neuron*, 74(4):621–633.
- Selimi, F., Cristea, I. M., Heller, E., Chait, B. T., and Heintz, N. (2009). Proteomic studies of a single cns synapse type: the parallel fiber/purkinje cell synapse. *PLoS Biol*, 7(4):e1000083.
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology*, 46(1):561–584.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.
- Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311.
- Siddiqui, I. J., Pervaiz, N., and Abbasi, A. A. (2016). The parkinson disease gene snca: Evolutionary and structural insights with pathological implication. *Scientific reports*, 6.
- Simunovic, F., Yi, M., Wang, Y., Macey, L., Brown, L. T., Krichevsky, A. M., Andersen, S. L., Stephens, R. M., Benes, F. M., and Sonntag, K. C. (2009). Gene expression profiling of substantia nigra dopamine neurons: further insights into parkinson's disease pathology. *Brain*, 132(7):1795–1809.
- Soffer, S. N. and Vázquez, A. (2005). Network clustering coefficient without degree-correlation biases. *Physical Review E*, 71(5):057101.
- Sokolow, S., Henkins, K. M., Williams, I. A., Vinters, H. V., Schmid, I., Cole, G. M., and Gyls, K. H. (2012). Isolation of synaptic terminals from alzheimer's disease cortex. *Cytometry Part A*, 81(3):248–254.
- Spatola, M. and Wider, C. (2014). Genetics of parkinson's disease: the yield. *Parkinsonism & related disorders*, 20:S35–S38.
- Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539.

- Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A. D., and Cooper, D. N. (2014). The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, 133(1):1–9.
- Stuehr, D. J. (2004). Enzymes of the l-arginine to nitric oxide pathway. *The Journal of nutrition*, 134(10):2748S–2751S.
- Sun, Q. and Turrigiano, G. G. (2011). Psd-95 and psd-93 play critical but distinct roles in synaptic scaling up and down. *Journal of Neuroscience*, 31(18):6800–6808.
- Swinton, J. (2013). *Vennerable: Venn and Euler area-proportional diagrams*. R package version 3.0/r82.
- Tanaka, Y. (1957). *Report of the international committee on genetic symbols and nomenclature*. Union of International Sci Biol Ser B, Colloquia No. 30.
- Tanner, C. M. (1991). Abnormal liver enzyme-mediated metabolism in parkinson's disease a second look. *Neurology*, 41(5 Suppl 2):89–91.
- Tansey, M. G. and Goldberg, M. S. (2010). Neuroinflammation in parkinson's disease: its role in neuronal death and implications for therapeutic intervention. *Neurobiology of disease*, 37(3):510–518.
- Thenganatt, M. A. and Jankovic, J. (2014). Parkinson disease subtypes. *JAMA neurology*, 71(4):499–504.
- Traag, V. A. and Bruggeman, J. (2009). Community detection in networks with positive and negative links. *Physical Review E*, 80(3):036115.
- Trinidad, J., Thalhammer, A., Specht, C., Schoepfer, R., and Burlingame, A. (2005). Phosphorylation state of postsynaptic density proteins. *Journal of neurochemistry*, 92(6):1306–1316.
- Trinidad, J. C., Thalhammer, A., Specht, C. G., Lynn, A. J., Baker, P. R., Schoepfer, R., and Burlingame, A. L. (2008). Quantitative analysis of synaptic phosphorylation and protein expression. *Molecular & Cellular Proteomics*, 7(4):684–696.
- Tsai, Y.-C., Greco, T. M., Boonmee, A., Miteva, Y., and Cristea, I. M. (2012). Functional proteomics establishes the interaction of sirt7 with chromatin remodeling complexes and expands its role in regulation of rna polymerase i transcription. *Molecular & Cellular Proteomics*, 11(5):60–76.
- Tsirigotis, M., Zhang, M., Chiu, R. K., Wouters, B. G., and Gray, D. A. (2001). Sensitivity of mammalian cells expressing mutant ubiquitin to protein-damaging agents. *Journal of Biological Chemistry*, 276(49):46073–46078.
- Turner, P. R., O'connor, K., Tate, W. P., and Abraham, W. C. (2003). Roles of amyloid precursor protein and its fragments in regulating neural activity, plasticity and memory. *Progress in neurobiology*, 70(1):1–32.

- Uezu, A., Kanak, D. J., Bradshaw, T. W., Soderblom, E. J., Catavero, C. M., Burette, A. C., Weinberg, R. J., and Soderling, S. H. (2016). Identification of an elaborate complex mediating postsynaptic inhibition. *Science*, 353(6304):1123–1129.
- Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419.
- UniProt Consortium et al. (2017). Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169.
- Uversky, V. N. (2008). α -synuclein misfolding and neurodegenerative diseases. *Current Protein and Peptide Science*, 9(5):507–540.
- van Rooden, S. M., Colas, F., Martínez-Martín, P., Visser, M., Verbaan, D., Marinus, J., Chaudhuri, R. K., Kok, J. N., and van Hilten, J. J. (2011). Clinical subtypes of parkinson's disease. *Movement Disorders*, 26(1):51–58.
- van Rossum, G. (1995). *Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI)*. Amsterdam.
- Venderova, K. and Park, D. S. (2012). Programmed cell death in parkinson's disease. *Cold Spring Harbor perspectives in medicine*, 2(8):a009365.
- Vidal, M., Cusick, M. E., and Barabasi, A.-L. (2011). Interactome networks and human disease. *Cell*, 144(6):986–998.
- Voglis, G. and Tavernarakis, N. (2006). The role of synaptic ion channels in synaptic plasticity. *EMBO reports*, 7(11):1104–1110.
- Wakabayashi, K., Tanji, K., Mori, F., and Takahashi, H. (2007). The lewy body in parkinson's disease: Molecules implicated in the formation and degradation of α -synuclein aggregates. *Neuropathology*, 27(5):494–506.
- Walikonis, R. S., Jensen, O. N., Mann, M., Provance, D. W., Mercer, J. A., and Kennedy, M. B. (2000). Identification of proteins in the postsynaptic density fraction by mass spectrometry. *Journal of Neuroscience*, 20(11):4069–4080.
- wan Li, K., Hornshaw, M. P., Van der Schors, R. C., Watson, R., Tate, S., Casetta, B., Jimenez, C. R., Gouwenberg, Y., Gundelfinger, E. D., Smalla, K.-H., et al. (2003). Proteomics analysis of rat brain postsynaptic density: implications of the diverse protein functional groups for the integration of synaptic physiology. *Journal of Biological Chemistry*.
- Wanders, R. J. and Waterham, H. R. (2006). Biochemistry of mammalian peroxisomes revisited. *Annu. Rev. Biochem.*, 75:295–332.
- Wang, J., Li, M., Deng, Y., and Pan, Y. (2010). Recent advances in clustering methods for protein interaction networks. *BMC genomics*, 11(Suppl 3):S10.

- Wang, J. Y. (2014). The capable abl: what is its biological function? *Molecular and cellular biology*, 34(7):1188–1197.
- Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.
- Weingarten, J., Laßek, M., Mueller, B. F., Rohmer, M., Lunger, I., Baeumlisberger, D., Dudek, S., Gogesch, P., Karas, M., and Volkandt, W. (2014). The proteome of the presynaptic active zone from mouse brain. *Molecular and Cellular Neuroscience*, 59:106–118.
- Wetie, N., Armand, G., Sokolowska, I., Woods, A. G., Roy, U., Loo, J. A., and Darie, C. C. (2013). Investigation of stable and transient protein–protein interactions: past, present, and future. *Proteomics*, 13(3-4):538–557.
- Whittaker, V., Michaelson, I., and Kirkland, R. J. A. (1964). The separation of synaptic vesicles from nerve-ending particles (synaptosomes'). *Biochemical Journal*, 90(2):293.
- Wilhelm, B. G., Mandad, S., Truckenbrodt, S., Kröhnert, K., Schäfer, C., Rammner, B., Koo, S. J., Claßen, G. A., Krauss, M., Haucke, V., et al. (2014). Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins. *Science*, 344(6187):1023–1028.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3:160018.
- Wong, K., Grove, J., Grandinetti, A., Curb, J., Yee, M., Blanchette, P., Ross, G., and Rodriguez, B. (2010). Association of fibrinogen with parkinson disease in elderly japanese-american men: a prospective study. *Neuroepidemiology*, 34(1):50–54.
- Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). Dip: the database of interacting proteins. *Nucleic acids research*, 28(1):289–291.
- Xia, J., Benner, M. J., and Hancock, R. E. (2014). Networkanalyst-integrative approaches for protein–protein interaction network analysis and visual exploration. *Nucleic acids research*, 42(W1):W167–W174.
- Yoshimura, Y., Yamauchi, Y., Shinkawa, T., Taoka, M., Donai, H., Takahashi, N., Isobe, T., and Yamauchi, T. (2004). Molecular constituents of the postsynaptic density fraction revealed by proteomic analysis using multidimensional liquid chromatography-tandem mass spectrometry. *Journal of neurochemistry*, 88(3):759–768.
- Yu, X., Wang, C., and Li, Y. (2006). Classification of protein quaternary structure by functional domain composition. *BMC bioinformatics*, 7(1):187.

- Yuste, R. (2015). The discovery of dendritic spines by cajal. *Frontiers in neuroanatomy*, 9.
- Zhang, B., Park, B.-H., Karpinets, T., and Samatova, N. F. (2008). From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics*, 24(7):979–986.
- Zhang, Y., James, M., Middleton, F. A., and Davis, R. L. (2005). Transcriptional analysis of multiple brain regions in parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 137(1):5–16.