# Mechanism-based Stratification of Alzheimer′s and Parkinson′s Disease using Artificial Intelligence

Kumulative Dissertation
zur Erlangung des Doktorgrades (Dr. rer. nat.)
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
**MOHAMMAD ASIF EMRAN KHAN EMON**
aus Chattogram, Bangladesch

Bonn, 2020

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Univ.-Prof. Dr. rer. nat. Martin Hofmann-Apitius
2. Gutachter: Univ.-Prof. Dr. rer. nat. Thomas Schultz
Tag der Promotion: March 10, 2021
Erscheinungsjahr: 2021

# Abstract

The capacity to generate omics and clinical data in biomedical science is growing exponentially over the past decades. Additionally, recent advances in computational power and analyzing capabilities have resulted in overwhelmingly increased interest in the use of big data to solve most problems in biomedical science. Drug discovery and molecular disease taxonomies are two of the most pressing challenges in biomedical science that could be solved by the surge of big data. Hence, there is an urgent need for developing methods that incorporate biomedical data and prior knowledge for drug development and patient stratification in order to achieve the goal of stratified medicine.

In this thesis, we address the aforementioned issues in the context of neurodegenerative diseases. First, we demonstrate a pure knowledge-driven approach for mechanism-based drug repositioning in Alzheimer's disease by curating and analyzing Alzheimer's disease knowledge assembly. Second, we present PS4DR, a drug repositioning workflow that is based on the combination of knowledge- and data-driven approaches. This work combines canonical pathway information and multi-omics data in order to predict drugs that can alter disease etiology. Finally, we showcase a hybrid artificial intelligence-based approach to jointly stratify Alzheimer's and Parkinson's disease patients based on the omics data and prior knowledge of shared molecular mechanisms of the two diseases. The established patient subgroups are reproducible and can be associated with different clinical and molecular disease features.

Finally, this thesis attempts to connect the knowledge- and data-driven strategy for solving two very interesting biomedical problems of drug discovery and patient stratification by using prior knowledge, multi-omics, imaging, and clinical data. Overall, this work is a step towards achieving a better targeted and thus more effective therapy in neurology to reach the ultimate goal of precision medicine concept.

*In loving memory of my Father.*

# Acknowledgment

All that I am today or ever hope to be, I owe to my incredible parents. I would have never made it this far without the unconditional love and supports from my Abbu and Ammu. I would like to thank my two brothers for being my strength during the hard times. I am also grateful to all my friends, who were always there with me when I needed them most.

I would like to thank my supervisor, Prof. Dr. Martin Hofmann-Apitius, for giving me the opportunity to work in this excellent department of SCAI-BIO and putting trust in me. His guidance helped me to navigate through the challenging times in this journey.

I would also like to thank, Prof. Dr. Holger Fröhlich, for supervising me in this thesis. I am glad that I had the opportunity to work with him that gave me the opportunity to improve my research with his insightful guidance and feedback.

I am grateful to all my friendly colleagues and students at SCAI-BIO, for being part of this wonderful journey. Especially, I would like to thank my colleague, Dr. Daniel Domingo-Fernandez, for his supports during the last couple of years. I would also like to convey my gratitude to Stephan Springstubbe, for helping me through different project activities. I am also thankful to my lunch buddy, Mandar Pathare, for being there with me all this time.

Finally, I would like to thank my dear wife, Sabreen, for being incredibly patient and always inspiring me to achieve the final goal.

# Publications

## Thesis publications

- **Mohammad Asif Emon**, Alpha Tom Kodamullil, Reagon Karki, Erfan Younesi, and Martin Hofmann-Apitius. "Using drugs as molecular probes: A computational chemical biology approach in neurodegenerative diseases." *Journal of Alzheimer's Disease* 56, no. 2 (2017): 677-686.

  https://doi.org/10.3233/JAD-160222

- **Mohammad Asif Emon**, Daniel Domingo-Fernández, Charles Tapely Hoyt, and Martin Hofmann-Apitius. *"PS4DR*: A multimodal workflow for identification and prioritization of drugs based on pathway signatures." *BMC Bioinformatics* 21, 231 (2020).

  https://doi.org/10.1186/s12859-020-03568-5

- **Mohammad Asif Emon+**, Ashley Heinson+, Ping Wu+, Daniel Domingo-Fernández, Henri Vrooman, Jean-Christophe Corvol, Phil Scordis, Martin Hofmann-Apitius and Holger Fröhlich. Clustering of Alzheimer's and Parkinson's Disease Based on Genetic Burden of Shared Molecular Mechanisms." *Scientific Reports* 10, 19097 (2020).

  https://doi.org/10.1038/s41598-020-76200-4

## Other publications

- Johann de Jong, **Mohammad Asif Emon**, Ping Wu, Reagon Karki, Meemansa Sood, Patrice Godard, Ashar Ahmad, Henri Vrooman, Martin Hofmann-Apitius, and Holger Fröhlich. "Deep learning for clustering of multivariate clinical patient trajectories with missing values." *GigaScience* 8, no. 11 (2019): giz134.

https://doi.org/10.1093/gigascience/giz134

- Colin Birkenbihl, **Mohammad Asif Emon**, Henri Vrooman, Sarah Westwood, Simon Lovestone on behalf of the AddNeuroMed Consortium, Martin Hofmann-Apitius, Holger Fröhlich, and the Alzheimer's Disease Neuroimaging Initiative. "Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia - lessons for translation into clinical practice." *EPMA Journal* 11, 367–376 (2020).

  https://doi.org/10.1007/s13167-020-00216-z

- Meemansa Sood, Akrishta Sahay, Reagon Karki, **Mohammad Asif Emon**, Henri Vrooman, Martin Hofmann-Apitius and Holger Fröhlich. "Realistic Simulation of Virtual Multi-Scale, Multi-Modal Patient Trajectories using Bayesian Networks and Sparse Autoencoders." *Scientific Reports* 10, 10971 (2020).

  https://doi.org/10.1038/s41598-020-67398-4

- Shashank Khanna, Daniel Domingo-Fernández, Anandhi Iyappan, **Mohammad Asif Emon**, Martin Hofmann-Apitius, and Holger Fröhlich. "Using multi-scale genetic, neuroimaging and clinical data for predicting alzheimer's disease and reconstruction of relevant biological mechanisms." *Scientific reports* 8, no. 1 (2018): 1-13.

  https://doi.org/10.1038/s41598-018-29433-3

- Daniel Domingo-Fernández, Alpha Tom Kodamullil, Anandhi Iyappan, Mufassra Naz, **Mohammad Asif Emon**, Tamara Raschka, Reagon Karki, Stephan Springstubbe, Christian Ebeling, and Martin Hofmann-Apitius. "Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment." *Bioinformatics* 33, no. 22 (2017): 3679-3681.

  https://doi.org/10.1093/bioinformatics/btx399

# Contents

# Acronyms

**AD** Alzheimer's Disease.

**ADNI** Alzheimer's Disease Neuroimaging Initiative.

**AI** Artificial Intelligence.

**AUC** Area Under the Curve.

**BEL** Biological Expression Language.

**BioPAX** Biological Pathways Exchange.

**CNN** Convolutional Neural Network.

**EHR** Electronic Health Record.

**GWAS** Genome-wide ssociation study.

**HGNC** HUGO Gene Nomenclature Committee.

**KEGG** Kyoto Encyclopedia of Genes and Genomes.

**MeSH** Medical Subject Headings.

**ML** Machine Learning.

**NDD** Neurodegenerative Disease.

**PD** Parkinson's Disease.

**PPMI** Parkinson's Progression Markers Initiative.

**SBML** Systems Biology Markup Language.

**SNP** Single Nucleotide Polymorphism.

# 1 Introduction

The term 'precision medicine' refers to the customized treatment of an individual or a group of patients based on their specific disease characteristics [1]. It has a few more popular synonyms that are often interchangeably used, such as personalized medicine, stratified medicine, targeted therapy, or deep phenotyping. This innovative approach tries to identify subpopulation among the patients by using advanced biomedical tools and big data analytics to predict the treatment and prevention strategies that will work best for each patient group. Precision medicine intends to replace the traditional clinical practice of 'one-size-fits-all' treatments, with tailored treatments for each patient group based on their unique biological condition i.e., genetic predisposition, sex, age, ethnicity, lifestyle, environment, etc. While the broader concept of precision medicine has been part of traditional healthcare in many ways, the term itself and its integral concept have become recently very popular due to its great success in the oncology research field [2]. However, such a successful implementation of the precision medicine approach still remains very challenging in complex multifactorial diseases like the neurodegenerative disease (NDD) research field [3].

Meanwhile, the emergence of ever-growing patient-level big data and the increased application of advanced machine learning (ML) techniques in biomedical science bring a new momentum in precision medicine in recent years. While primarily large scale multi-omics and clinical data have been fuelling biomedical research, the recent availability of more patient-level big data like electronic health records (EHRs) and smart device data gives the real opportunity to revolutionize the health care system [4]. ML techniques that can integrate such big data are being successfully incorporated in many sectors of the modern health care system including disease diagnosis and prognosis, patient stratification, drug discovery, clinical

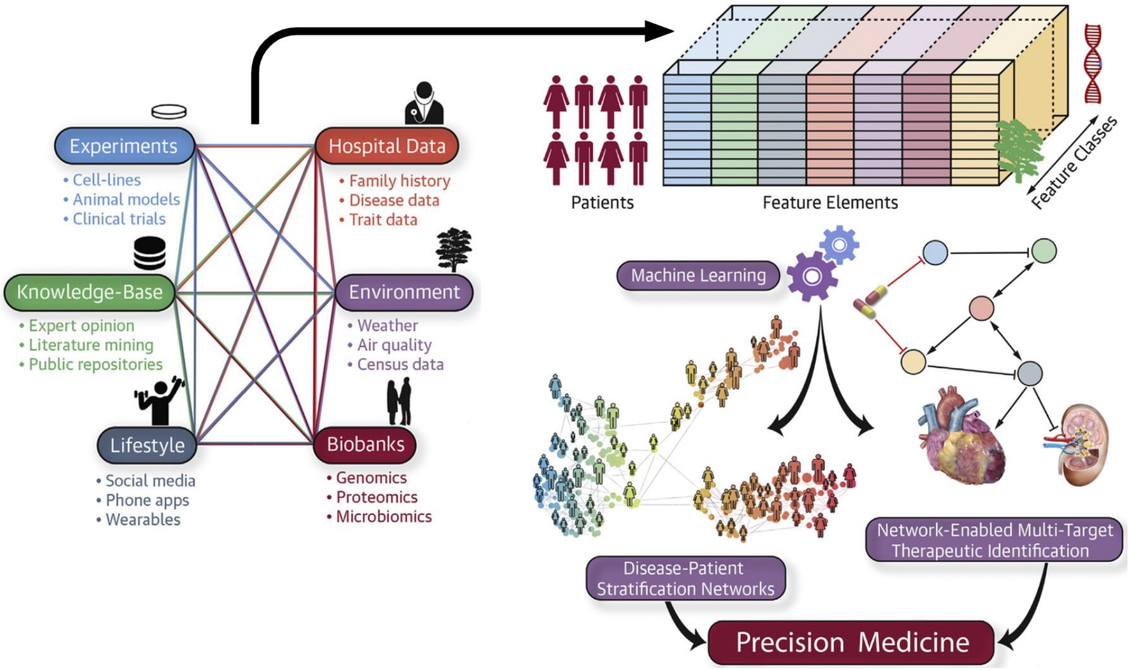


**Figure 1: Machine learning-based Precision Medicine workflow.** Big data in biomedical science enabling us to implement machine learning-based sophisticated workflows in precision medicine. This workflow shows how experimental data, clinical data, lifestyle data, expertly curated knowledge, etc., can be integrated and analyzed by smart machine learning algorithms to stratify patients in order to achieve the precision medicine goal. This figure was modified from [7].

trial designs, etc [5]. However, the immense complexities of multifactorial disease pathologies and thus the inability to decode underlying disease mechanisms limit the hypothesis-driven analyses of these multimodal data. Consolidated knowledge about complex disorders like Alzheimer's disease (AD) and Parkinson's disease (PD) captured from scientific literature and public databases can come to rescue by providing a holistic view of disease biology [6]. Hence, developing advanced machine learning methods to enable better integration of the domain-specific knowledge with these multi-modal data is of critical importance in achieving success in precision medicine. In addition to enabling us to better understand the complex etiology of multifactorial diseases, such integrative approaches can help us to facilitate the research on some burning problems of some sub-domains such as patient stratification, advance clinical trial designs, smart drug discovery programs, etc., that need to be solved first to advance precision medicine research.

## 1.1 Neurodegenerative Diseases

Neurodegenerative disease (NDD) is an umbrella term for a heterogeneous group of conditions that are characterized by the progressive degeneration of the structure and function of the brain or peripheral nervous system. Despite extensive investment and research for unraveling mechanisms of NDDs, the etiology of this group of diseases remains unelucidated due to their complex multifactorial nature [8]. Due to the failure of a deep understanding of the disease pathogenicity, there is no treatment to cure neurodegenerative disorders as of yet. There are only some FDA approved drugs that can partially alleviate the disease symptoms [9]. The increase of elderly people in recent years partially contributes to the upsurge of such age-dependent disorders. The global cost for treatment and care of dementia is estimated to reach 2 trillion US dollars by 2030, which will have an immense socio-economic impact as well [10]. Alzheimer's disease (AD) and Parkinson's disease (PD) are the two most common neurodegenerative diseases among the elderly population [11]. The following sections will focus on introducing these two neurodegenerative diseases AD and PD.

### 1.1.1 Alzheimer's Disease

One of the most common forms of dementia, Alzheimer's disease (AD) is characterized by progressive neurodegeneration of the nervous system that slowly disrupts memory and thinking skills and, eventually, the ability to carry out simple day to day activities [12]. The neuronal damage in AD initially starts in the hippocampus, which plays a major role in learning and memory [13]. Amyloid plaques and neurofibrillary tangles, the two most common hallmarks of AD, are widely believed to be responsible for such neural damage [14]. They prevent neurons from properly functioning, restrict the neuronal communications, and eventually lead neurons to death. Later, the damage reaches the areas in the cerebral cortex, an area responsible for learning language and social behavior [15]. Eventually, it affects other parts of the brain and over time, the patient gradually loses his or her ability to live and function without assistance from a caregiver [15, 16].

Even though there exist a lot of hypotheses for the development and progression of AD, both amyloid plaque accumulation and the presence of neurofibrillary tangles (NFT) remain the two main hallmarks of AD pathology. According to the amyloid hypothesis, abnormal proteolytic processing of the amyloid precursor protein (APP) by beta-secretase (BACE-1) and gamma-secretase results in the production and accumulation of neurotoxic amyloid-beta peptide (A$\beta$42) [17]. However, the tau hypothesis postulates, abnormal phosphorylation of TAU protein (MAPT) contributes to the formation of hyperphosphorylated tau that leads to the formation of the neurofibrillary tangles in AD [18]. In addition to these two leading hypotheses, numerous other physiological mechanisms including neuroinflammation, oxidative stress, mitochondrial dysfunction, insulin signaling, etc., are also linked with AD pathogenesis [19]. While a majority of AD cases can be attributed by the late-onset sporadic form, genetically the disease can be divided into two main subtypes [20, 21]:

- **Familial Alzheimer's disease (FAD)**: This particular AD form is related to the autosomal dominant inheritance of three causative genes, i.e., amyloid precursor protein (APP) gene, presenilin1 (PSEN1) gene, and presenilin 2 (PSEN2) gene. Even though this group accounts for only 5% of cases, it can provide significant insights into the pathogenesis of the sporadic case.
- **Sporadic Alzheimer's disease (SAD):** Unlike familial AD, both genetic and environmental factors may play roles in determining the sporadic AD form. While APOE is considered to be the main responsible gene in sporadic AD, many other genes, like TREM2, ABCA1, ABCA7, MTHFD1, BIN1, etc. were reported to be associated with the disease form. The complete etiology of sporadic AD is not yet well characterized, despite being one of the most investigated diseases.

While many hypotheses have been formulated based on the known disease hallmarks and symptoms, very little could be brought to the light about the actual disease etiology. This knowledge gap about the disease etiology could be attributed to the failure in developing disease-modifying drugs that alter the progression [22]. Despite much research being done, there are only four approved drugs to treat symptoms until now; donepezil, rivastigmine, galantamine, and memantine. Failure in finding the right treatment starting point or identifying the right therapeutic targets or establishing precise and accurate clinical methodologies are some reasons to be blamed for this setback [22]. Hence, there is an urgent need for a greater effort, increased funding and a well-planned strategy to overcome this lacking.

### 1.1.2 Parkinson's Disease

Parkinson's disease (PD) is the most common movement disorder and the second most common neurodegenerative disease after AD, which affects nearly 1% of the population above the age of 60. PD is a neurodegenerative disease with initial clinical manifestations of abnormal motor symptoms that are the result of the loss of midbrain dopaminergic neurons in the substantia nigra. Resting tremor, bradykinesia, rigidity, and postural instability are the most common characteristic motor impairments collectively known as parkinsonism that are seen in PD patients due to the loss of pigmented dopaminergic neurons of the substantia nigra pars compacta (SNpc). Behavioral and cognitive declines become apparent gradually with the progression of neuronal loss in the other area of the brain at an advanced stage. Similar to other neurodegenerative disorders, PD is characterized by the presence of intracellular protein aggregates: Lewy bodies and Lewy neurites. While tau proteins and neurofibrillary tangles present among others, misfolded $\alpha$-synuclein is the key component of the Lewy bodies and neurites [23, 24].

Considering the multifaceted nature of PD with a wide range of clinical symptoms and pathology, the complete etiology and pathogenesis of PD have yet to be established. Several pathways, like oxidative stress, mitochondrial dysfunction, inflammation, and ubiquitin proteasome system (UPS) have been implicated with the disease pathogenicity. However, the exact etiology is largely not understood as of now. Hence, there are no treatments available that can fully treat the clinical syndrome or alter the natural history of PD [25].

While most PD is sporadic in nature, there exists a small number of familial form that consists of up to 15% of the PD cases.

- **Familial Parkinson's Disease:** This PD subtype is caused by a single gene mutation and at least eight genes responsible for this familial form have been discovered so far. They include SNCA, PRKN, LRRK2, DJ-1, UCHL1, NR4A2, GIGYF2 and PINK1. Familial PD patients are quite heterogeneous in nature with either early or late symptomatic onset, slow or rapid progression, and autosomal recessive or dominant modes of inheritance [26].

- **Sporadic Parkinson's Disease:** Sporadic PD is multifactorial in nature and a complex interplay between several familial PD linked genes and the environment play a role to characterize the disease etiology. While this is the most common form of PD among the population, they are often clinically and pathologically quite indistinguishable from the familial cases. Hence, the biochemical pathways underlying both cases might be shared to some extent [27].

### 1.1.3 Crosstalk between Alzheimer's and Parkinson's Disease

While small fractions of both AD and PD patients suffer from the familial variety of the disease with the autosomal dominant pattern of inheritance and high penetrance, both the diseases appear predominantly in a sporadic fashion. The clinical manifestations of AD are characterized by dementia and cognitive impairment, while PD patients mainly exhibit motor deficits [11]. However, despite their diverse clinical and pathological manifestations, AD and PD share a number of common features and mechanisms in terms of molecular pathogenesis [28]. The overlaps between AD and PD are reported at various levels including genetics, cellular mechanisms and biological pathway level based on the evidence obtained from postmortem studies and experimental models [29]. While the accumulation of altered protein aggregates is considered to be the most characteristic hallmarks of AD or PD, other biological events such as mitochondrial dysfunction, oxidative stress, inflammation are believed to be involved in the disease etiology [30]. Moreover, the involvement of several genes including MAPT, SNCA, TREM2, PON1, GSTO, NEDD9, etc., in the disease biology of AD and PD show a clear existence of genetic overlap in these diseases [31, 32].

These extensive crosstalks among biological mechanisms and pathways and the genetic or molecular factors that trigger these disrupted mechanisms can play crucial roles to obtain a deeper understanding of the AD and PD pathogenesis. As a result, it presents a unique opportunity for a joint stratification of AD and PD patients by using the knowledge of such molecular and mechanistic overlaps between two diseases and patient-level big data via advanced machine learning and statistical models. While various omics, clinical and imaging data have been the main fuel for such patient subgrouping in different diseases in the oncology field, availability of additional mechanistic knowledge of diseases will have far-reaching benefits for the successful stratification strategies in multifactorial complex disease fields like NDDs. Such a joint AD-PD patient subgrouping will be beneficial for finding similar disease-modifying and therapeutic strategies that will serve the ultimate goal of precision medicine.

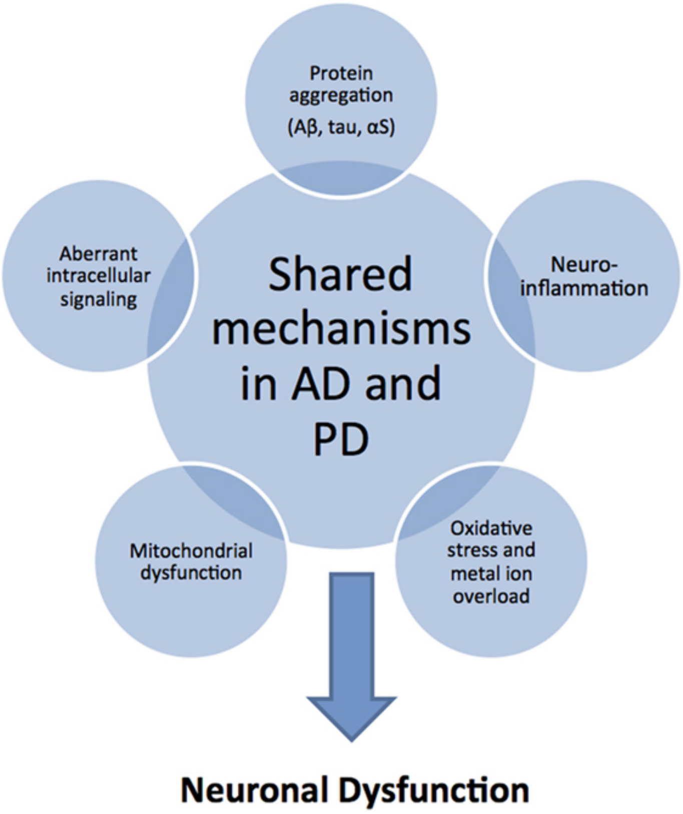

Protein aggregation (Aβ, tau, αS)
Aberrant intracellular signaling
Shared mechanisms in AD and PD
Neuro-inflammation
Mitochondrial dysfunction
Oxidative stress and metal ion overload
Neuronal Dysfunction

**Figure 2: Shared pathological mechanisms between Alzheimer's disease (AD) and Parkinson's disease (PD).** Protein aggregation, neuroinflammation, oxidative stress, mitochondrial dysfunction, and aberrant intracellular signaling are the most prominent biological processes shared by both AD and PD [30].

## 1.2 Knowledge Modeling: an essential tool for translational research

Biomedical data have been exponentially growing both in volumes and varieties due to the advancement in high throughput data generation technologies [33]. While high throughput sequencing of bulk populations is yet not an old concept, brand new techniques like single-cell omics sequencing is making it possible to generate thousands or even millions of measurements concurrently from a single sample. Moreover, with the widespread availability of patient-level data such as electronic health records (EHR) data, smart digital device data, imaging data, etc., biomedical data growth is even expected to outpace Moore's law of computational power increment [34]. Hence, biomedical data fulfills all three requirements of the big data - volume, velocity, and variety. This enormous big data in the biomedical field is presenting us the opportunity to explore the biological systems in such granularity as never seen before. This wave of big data is also generating an enormous amount of information and knowledge. However, such unprecedented growth of knowledge and information poses new challenges for systematic analysis and interpretation of the available resources. Formalization and capturing of the knowledge in a particular domain (i.e., disease) in a computable form, permit the development of tools for understanding the complexity of the domain.

Knowledge modeling in systems biology is such a tool for mapping and representing the existing knowledge about a particular biological domain to enable novel interpretation of biomedical data [35]. With the formalization of available knowledge, they help to explain the relevant biological mechanisms and predict how the system might act when perturbed by therapeutic intervention or other environmental challenges [36]. Hence, they are the bridge that can fill the gaps between biomedical research and the translation of these researches into impactful clinical practices. While many modeling approaches are available to capture biological knowledge, we will focus on the conceptual modeling for the knowledge representation as it goes within the scope of this thesis. Unlike mathematical modeling, the conceptual model captures the salient aspects of a biological system to structure our conceptualizations into the relevant entities and their relationships to organize biological complexity.

Systems Biology Markup Language (SBML), Biological Pathway Exchange Language (BioPAX), and Biological Expression Language (BEL) are the most common conceptual knowledge formats in the biomedical domain. In the following we will briefly discuss these three knowledge formats.

- **Systems Biology Markup Language (SBML).** Systems Biology Markup Language (SBML) is a standard format for representing biochemical reaction networks. This XML-based format is widely used for storing computational models of biological processes and making them interoperable. SBML can be used for the representation of a wide range of biological circumstances, such as metabolic pathways, gene regulatory networks, cell signaling pathways, disease models, etc [37]. SBML can capture quantitative information of molecular species and their concentrations, interactions among these entities, and kinetic laws for these reactions. Hence, considering the dynamics of multiscale interactions of biological systems, SBML models are suitable for simulations of stochastic kinetic models [38]. It is also worth noting that instead of trying to be a universal language for quantitative models, SBML is rather intended for trading the salient features of a sophisticated systems biology model between different software systems and databases [39].

- **Biological Pathways Exchange (BioPAX).** Biological Pathway Exchange (BioPAX) is a standard language to capture and facilitate the exchange of pathway data from heterogeneous information sources. It uses Web Ontology Language (OWL) formats to represent biological pathways at the molecular and cellular level. While the proper utilization of a large amount of data across different pathway databases is hindered due to their incompatible storing formats, BioPAX makes it considerably easier to collect, index, interpret and share pathway data. It can capture and index a broad range of metabolic, signaling, molecular, and gene regulatory networks. BioPAX has been used in different databases to represent millions of interactions from thousands of pathways in various organisms in a computable form. Hence, BioPAX allows information exchange between pathway users, databases and software tools. However, unlike SBML, BioPAX is not able to capture the dynamic and quantitative aspects of the biological process [40].

- **Biological Expression Language (BEL).** Biological Expression Language (BEL) is a high-level systems biology modeling language that captures causal and correlative relationships between different biological entities. BEL enables the assembly of scientific findings in the life sciences in a context-specific manner across multi-scales. It captures the relationships among the entities in a subject-predicate-object (triple) format (Figure 3). As the object of a BEL triple can serve as the subject of one or many other triples, a set of BEL statements can be used to develop a knowledge base or knowledge assembly in the form of a conceptual graph [41]. This computable knowledge

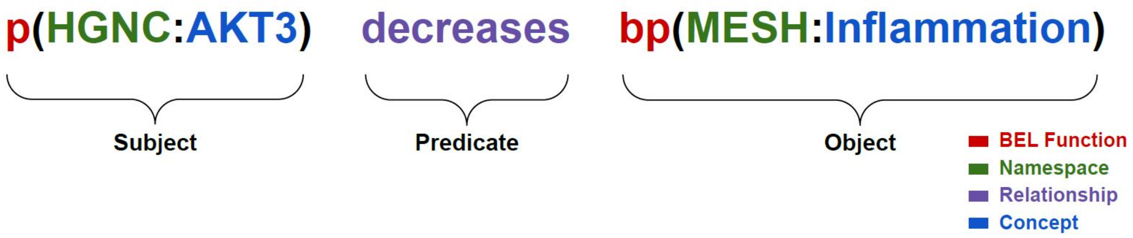


**Figure 3: A BEL Statement / Triple example.** The triple represents that AKT3 kinase protein decreases the inflammation. 'AKT3' is a subject, 'decreases' is a predicate, and 'Inflammation' is an object. The subject and object can either be a molecular entity, like genes, proteins, chemicals, or an abstract concept like biological processes, disorders, biochemical reactions, etc. The predicate represents the type of relationship between subject and object. BEL allows the usage of external namespaces for the formal representation of the concepts, like the HUGO Gene Nomenclature Committee (HGNC) and Medical Subject Headings (MeSH) used in this example.

assembly can be subjected to various graph algorithms for visualization and inference. A simple example of a BEL statement capturing the kinase activity of AKT3 decreases inflammation in the brain is depicted in Figure 3.

While BioPAX has become the de-facto language to integrate different pathways and interaction databases, BEL can serve as a semantic platform for multi-scale knowledge and data integration [42]. The semantic flexibility in BEL allows us to capture, integrate and analyze a wide range of mechanistic details of biological phenomena, i.e., from the molecular to organism scale. Hence, it makes BEL a perfect candidate for modeling sophisticated biological phenomena, like complex disease biology. This detailed capturing of the interactions between different entities in a biological system allows BEL to reason over the previously unknown mechanisms and processes. Moreover, different data-driven network analyses algorithms such as RCR [43] and NPA [44] can be applied to the BEL knowledge assemblies for different clinical applications.

BEL has been demonstrated as a useful format for building and analyzing knowledge models in various complex biological conditions in both diseased and non-diseased conditions [45, 46]. While such knowledge models have helped to understand how disrupted biology in the normal system leads to disease conditions, organizing them into respective knowledge graphs according to their participation in different biological processes is critical for understanding the interplay between multiple mechanisms in multifactorial diseases [6]. However, it remains to be seen if such organized knowledge graphs can be used to answer some of the burning issues in biomedical research such as, (i) could these graphs be used to stratify patients to facilitate precision medicine? and (ii) could these graphs also be targeted by known or novel drugs?

## 1.3 Machine Learning

The term machine learning was coined by famous American computer scientist Arthur Samuel in 1959 [47]. Machine learning (ML) is an application of artificial intelligence (AI) that gives computers the ability to automatically learn and improve from experience. It is the area of computational science that deals with analyzing and interpreting patterns and structures in data with the help of sophisticated statistical and computational algorithms. ML enables the computer to learn, reason, and make the decision without any human interference. Tom M. Mitchell provided a more formal definition of machine learning: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" [48]. ML has been predominantly incorporated into our day to day activities such as internet search, traffic prediction, purchase recommendations, social media services, email spam, and malware filtering, etc. While ML is already proved to be an indivisible part of modern life, it is starting to demonstrate its potential in biomedical research recently.

Biomedical science has become extraordinarily data intensive. The availability of large volumes of high throughput '-omics' data that capture large scale biology such as the genome, proteome, transcriptome, etc., has been a major driving force in the development of precision medicine. A comparison study of four major big data domains including genomics, astronomy, YouTube, and Twitter predicts that genomics alone will equal or surpass the other three domains in data generation and analysis within the next decade [49]. Moreover, the addition of personalized clinical data like EHR, imaging, digital device data, etc., with omics data have dramatically changed the precision medicine research. These enormous multimodal data enabled us to develop and implement state of the art machine learning (ML) algorithms based methods. Concurrent advancement of the biomedical data generation and ML application in this field are starting to demonstrate that more and more ML applications in the health sector are inevitable.

Many of these ML-based methods are equipped for multimodal data integration with existing knowledge which allows capturing complex relationships among the features within and across multiscale biology. Hence, they can provide a holistic view of the biological system by enabling better mechanistic insights of diseases [50]. Altogether, this might lead to the identification of new biomarkers which might be instrumental for early diagnosis, prognosis, or prediction of the disease. While machine learning has the potential to impact various domains of biomedical research, we will discuss some of these fields in this section where ML is already making impacts.

i. **ML in Disease Diagnosis:** Among all application fields of ML in clinical practice, workflows for the disease diagnosis are most likely to have a translational impact in the near future. Many ML-based approaches that can process various data modalities to identify the probable diagnosis are already in action or currently undergoing regulatory steps toward the market. For example, Google has developed various ML-based image analyses workflows to help identify cancerous tumors in different cancer types like lung cancer [51], breast cancer [52], etc. Both studies implemented state-of-the-art deep learning-based artificial intelligence algorithms to detect cancerous cells by extracting features from CT scans and whole slide images. The prediction performances of these algorithms were on-par with the radiologists with very high AUC. Another major example includes Stanford's deep learning algorithm to identify skin cancer that can perform as good as an expert dermatologist [53]. They have implemented a single deep convolutional neural networks based algorithm to classify skin cancer with the AUC of 91% by using pixels from histopathological images and disease labels as inputs. There are lots of successful ML-based studies published recently that are able to diagnose different diseases like, diabetic retinopathy [54], breast cancer [55], glaucoma [56], Parkinson's diseases [57].

ii. **ML in Drug Discovery:** While the drug discovery process is highly complicated, time-consuming, expensive and depends on numerous factors, the application of machine learning algorithms could significantly simplify and shorten this process. ML algorithms and tools could be used in all stages of drug discovery and development, from drug compound screening to clinical trials. Some of the successful applications of ML algorithms in various tasks of the drug discovery domain include novel targets identification [58], disease-target association prediction [59], better disease mechanisms understanding [60], small molecule design and optimization [61], biomarkers development [62], etc. Inspired by many successful implementations of machine learning in this field, many of the major pharmaceutical companies are shifting their focuses on machine learning-based drug discovery projects [63]. A more detailed literature review in this field has been presented later in the drug repositioning section.

iii. **ML in Patient Stratification:** Patient stratification is the most crucial element of precision medicine and yet still quite an open research field for machine learning applications. While we are still dependent on age-old definitions of diseases and disease subtypes, we can move beyond the observational correlation between pathological analysis and clinical outcome for disease classification with the help of current advancement in biomedical data generation and machine learning algorithms implementation. For example, Coudray *et al.* (2018) successfully demonstrated that non-small cell lung cancer (NSLC) patients can be classified based on histopathology images using a deep convolutional neural network [64]. Another similar study by Esteva *et al.* (2017) also successfully classified skin cancer using histological images with their deep learning framework [53]. While these types of phenotype-based disease classification can be still useful for better therapeutic care, the ultimate goal for achieving precision medicine is to subgroup disease based on the molecular characteristics. Ramazzotti and colleagues (2018) presented a novel workflow to identify molecular subtypes of 36 types of cancer by integrating multi-omics data via Multikernel Learning [65]. Another prominent example of such molecular profiling of the disease is the study by Ceccarelli *et al.* (2016), where they clustered Glioma patients with an unsupervised random forest method [66]. A more comprehensive literature review on the application of ML in patient stratification has been presented in the final section of the introduction.

iv. **ML augmented Physician:** The true potential of ML in precision medicine has become more apparent over the past few years by its successful implementation in various biomedical sectors. However, we still need to be skeptical about completely relying on ML when it comes to emergency medical decisions involving patients' life risk. For example, we can train a model to predict the risk of emergency hospital admissions based on the patient's past history including various medical and non-medical factors. However, taking a decision on whether or not to admit the patient solely based on the model's prediction could be fatal. Hence, a collaborative system that can harness the full potential of ML methods to analyze patient data and help the physician to augment their decision making is more realistic in many health care tasks. Moreover, such a system can assist physicians by presenting the results of the diagnostic by using available biomarkers, clinical data, published research, EHRs, smart device data, etc [67]. For example, Google demonstrated a deep convolutional neural network-based framework to accurately detect diabetic retinopathy in fundus photographs [54]. However, instead of replacing clinicians, such a workflow is intended to increase clinician's productivity by reducing the overall cost of diagnosis by automating time-consuming and low cognitive value tasks.

While the above discussions might give the impression that machine learning-powered precision medicine is just a matter of time to become reality, there are several challenges that remained to be solved before benefitting from such tools. Here, we will discuss some of the key challenges in implementing ML-based tools in precision medicine.

i. **Insufficient prediction performance:** The descriptive ability of the data with respect to the clinical endpoint of interest determines the prediction performance of the ML model in biomedical research. However, most of the real-world data are either mostly phenotypic but miss the molecular signatures that are important for successful model training (e.g., EHRs), or mostly having interesting molecular patterns but miss the clinical information for successful interpretation (e.g., genomic variant data). Such data poses big challenges in the separation of true signals from technical noise in big data analysis [68]. As a result, many machine learning models in the precision medicine domain constantly fail to achieve satisfactory predictive power to impress clinicians. Hence, it is essential to identify and integrate the right data modalities for any machine learning models that could extract parts of the relevant signal if not complete. Moreover, some measures like using a sufficiently large patient cohort, utilizing expert-curated knowledge to identify true biological variation from noisy data, careful definitions of clinical outcomes in complicated diseases, avoiding selection biases during sample selection, etc., while designing the study could significantly improve the model's prediction power.

ii. **Insufficient Data:** The success of machine learning models, deep learning models in particular, largely depends on the availability of a huge amount of training data. Although data generation in biomedical science has become a lot easier and cheaper than ever, huge chunks of data are still only available in the oncology and nervous system disease research fields. Moreover, most of these available data are either not of high quality or lack proper annotations. While increasing the amount of usable and correctly annotated data is the key to overcome this challenge, there are few algorithmic adaptations like transfer learning, multiple training runs, surrogate datasets, etc., can be introduced to handle this issue. The basic assumption of transfer learning is that predictive features learned in one application domain can also be applied to a different, but related application domain. Multiple training runs models are implemented by training thousands of models simultaneously with different learning parameters and ultimately select the highest performing model. However, such an implementation is computationally very expensive and time-consuming. Surrogate datasets, on the other hand, represent the noisy versions of the original dataset for the model's performance evaluation in the absence of additional validation data.

iii. **Multiple data modalities:** Advanced data generation technologies have enabled us to produce massive amounts of omics data from various sources like the genome, transcriptome, proteome, metabolome, etc. While traditionally these multimodal data have been treated in isolation by most of the machine learning methods, integrative analysis methods are the true future of precision medicine research. Various integrative ML algorithms are being implemented gradually to solve this challenge including combining various data at the input level, merging extracted features from different data modalities from independent models, aggregated predictions made by multiple ML models, etc [69]. However, implementations of new ML algorithms to enable the integration of new data types remain to be deeply investigated for more robust clinical outcome predictions. Ultimately, integration of new data modalities like EHR, smart digital device data, image-derived features with additional layers of omics data and expert-curated knowledge is essential to get the most benefit from such integrative approaches.

iv. **Interpretability and explanation:** While the predictive performance of ML models is an important criterion to judge how good the model is, the interpretability of such a system's prediction is far more important in clinical problems. It is very crucial to understand and explain how these systems work and make predictions to get clinicians and policymakers on board for the successful integration of such systems into the health care system. While ML algorithms are extraordinarily good in detecting complex patterns from large data to provide accurate predictions, they are not that impressive when it comes to providing a deeper theoretical, mechanistic, or causal understanding of the predicted outcomes. The higher the complexity of a model the less likely for them to be interpretable. Figure 4 demonstrates the classical tradeoff problem between the model complexity and their interoperability. Although method development for interpreting ML models is at a relatively early stage, particularly in biomedical science, ML models are gradually becoming interpretable through the use of different attribution methods. Attribution methods try to find the relevance or contribution of each feature in a model to make the prediction. Using prior knowledge of the biological system such as genes,

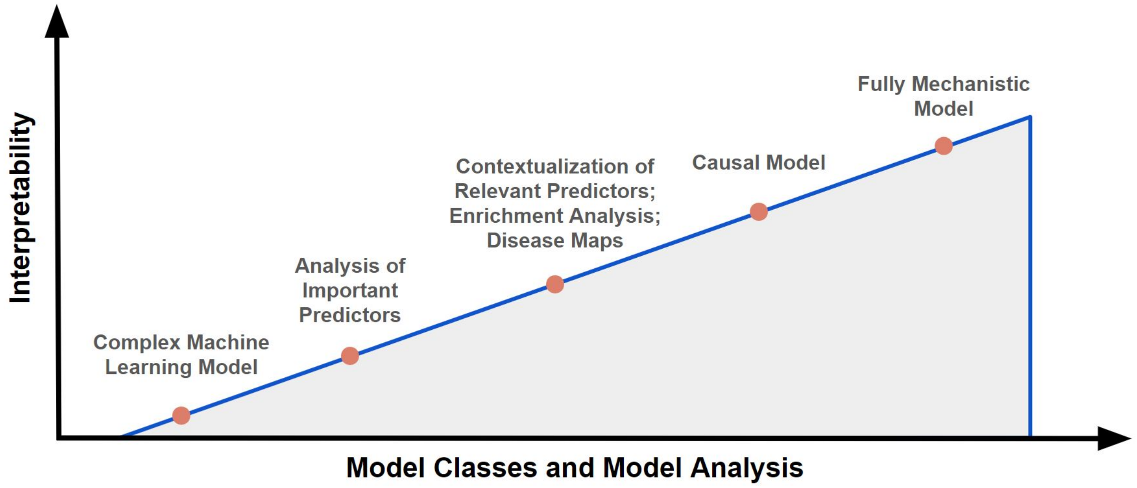

Interpretability
Fully Mechanistic Model
Causal Model
Contextualization of Relevant Predictors; Enrichment Analysis; Disease Maps
Analysis of Important Predictors
Complex Machine Learning Model
Model Classes and Model Analysis

**Figure 4: Interpretability of machine learning models via model analysis.** This chart demonstrates the model's interpretability as a spectrum. Complex ML models are at one side of the spectrum with very low interpretation power. There is a detailed understanding of the exact molecular and pathophysiological mechanisms at the other side of the spectrum which links a model with a defined clinical endpoint. This figure was adapted from [71].

proteins, cells, tissues, biological processes, or disease conditions for guiding the design of ML systems is another approach that can help for the interpretation of a model's prediction. Yu *et al.* (2018) explain the potential of such approaches and label them as "visible approaches" [70]. However, such prior knowledge is not readily available for most of the disease domains and hence, will not be effective unless additional efforts are invested.

## 1.4 Patient Stratification

The term 'patient stratification' represents the most central aspect of precision medicine: distinguishing patients into subgroups or strata based on their inherent features. Hence, such strata represent systematic differences in the patient population. Patient stratification is the first step toward precision medicine, also known as stratified medicine. While broad-scale features like age, sex, ethnicity, diet, etc., have been used for defining patients into different treatment groups from the dawn of modern clinical practices, use of a wide array of individual data like molecular, genetic, clinical and further biomarker information has attracted much attention in this field lately. The rapid advancement in generating high-throughput omics data for the measurements of such molecular biomarkers, like DNA, RNA, protein, metabolites, etc., played critical roles to promote patient stratification in recent clinical studies [72, 73].

Moreover, the emergence of more personalized patient-level 'big data' like EHR, imaging, digital device data, etc., has propelled the patient stratification researches with our increased capacity of rapidly producing, storing and analyzing such data in the last couple of years [74]. One of the main rationales for the patient stratification is to identify the group of patients who will have the most effective responses from a specific treatment. This approach is fundamentally different from traditional clinical practice based on the idea of 'one-size-fits-all' treatments. Technological and analytical advances in biomedical research made it possible to pursue this most anticipated concept by enabling stratified medicine for the patients based on their group profiles. Profiling breast cancer patients based on the presence of human epidermal growth factor receptor (HER)-2 is one of the successful examples [75]. While HER-2 positive patients showed a more aggressive form of the disease, clinical trials showed monoclonal antibody trastuzumab is more effective against the HER-2 positive patients. Another success story is the prevention of late-stage melanoma progression by the BRAF inhibitor vemurafenib. A phase III clinical trial showed patients with BRAF V600E mutation have better responses to vemurafenib treatment [76].

The key motivation for patient stratification is having a much better molecular and mechanistic understanding of disease to improve therapeutic strategies for better patient care. In short, with the application of advanced biomedical tools, such an approach will enable clinicians to predict which treatment and prevention strategies will work best for a given patient group. The benefits of precision medicine in the field of healthcare are immense [77, 78]:

- better understanding of the underlying mechanisms of the diseases within the sub-populations.
- earlier disease detection and possible disease prevention
- ability to predict therapies that will result in higher desirable outcomes
- improved clinical decision making and disease management
- better drug prescription with less predictable side effects

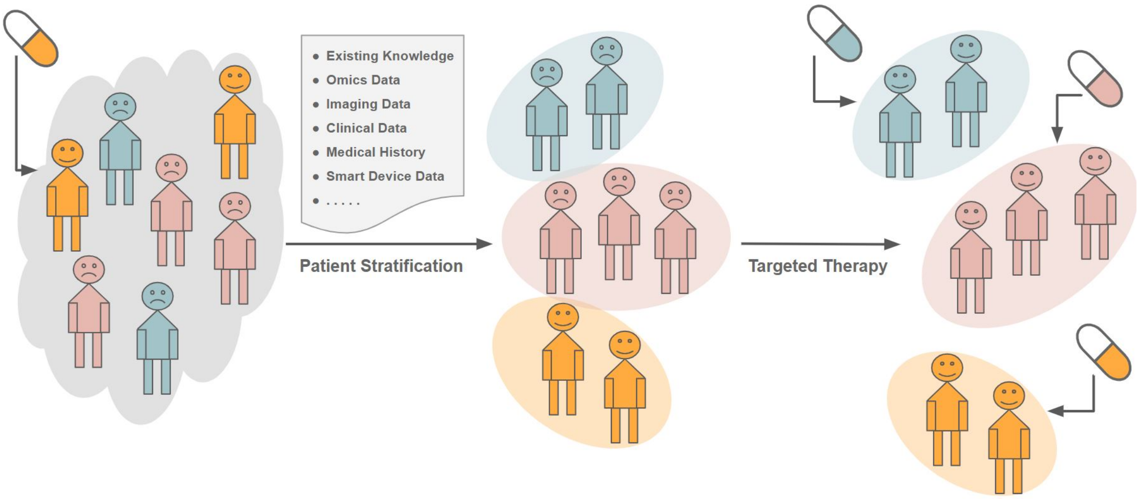


**Figure 5: A schematic representation of the patient stratification concept.** Treatments often fail to achieve the desired outcome in clinical trials due to the presence of unresponsive patient groups in the selected cohort. Statistical or/and machine learning models that can incorporate a wide range of individual features such as omics data, imaging data, clinical data, smart device data, etc., in addition to existing knowledge about the diseases allow clustering patients into separate pools. This stratification enables targeted therapy of those patient groups, more commonly referred to as precision medicine.

- well designed clinical trials with the selection of likely responders at baseline
- reduced time, cost, and failure rate of clinical trials due to better patient cohorts selection

While patient stratification has become a common practice in medical oncology with many successful clinical implementations, it is yet to mark the footsteps in the field of multifactorial complex neurodegenerative diseases like AD and PD. The success stories in oncology could be explained by (i) the presence of extraordinarily large knowledge bases of cancers built on many decades of research, and (ii) increased capacity of rapidly producing, storing and analyzing cancer data [79, 80]. While the new data can reveal novel trends in the disease, the presence of many cancer knowledge bases allow researchers to validate their findings with existing domain knowledge. Unfortunately, both the data and knowledge bases in neurodegenerative disease research are less prominent compared to the cancer field. Therefore, new methods and tools that can utilize both knowledge and data for patient stratification in neurodegenerative diseases are necessary.

## 1.5 Machine Learning and Patient Stratification

Machine Learning is becoming the driving force in both pre-clinical and clinical research over the past few years. It's potential in precision medicine has become more apparent mainly because of the many successful implementations of deep learning algorithms to a variety of biomedical activities including diagnosis, prognosis, disease risk predictions, patient stratifications, etc [5]. The advances of ML approaches in biomedical research made it possible for meaningful utilization of the surge of big biomedical data such as omics, EHR, imaging, digital device data, etc. While proper use of these big data still poses a great challenge to the community, ML is leading the research with the successful implementations of various advanced deep learning algorithms. Patient stratification is one of the most important tasks in the precision medicine field where the successful application of advanced ML approaches has been demonstrated. Hence, we can move beyond the classical disease definition system with the help of these advanced machine learning algorithms which facilitate mechanism-based patient stratifications.

Oncology is the most prolific biomedical research field for the successful demonstrations of machine learning applications for patient stratification. While the application of ML in patient stratification is still at the experimental level in most disease domains, ML-based cancer patient stratifications have demonstrated its successful implementations at the clinical level. For instance, Esteva *et al.* (2017) developed a deep convolutional neural network (CNN) based method which can successfully classify skin cancer using histological images. Their method demonstrated a certified dermatologist level competency in classifying different skin cancer types [53]. Another CNN based framework, developed by Coudray and colleagues (2018), successfully demonstrated that non-small cell lung cancer (NSLC) patients can be classified based on histopathology images [64]. Their method can successfully classify adenocarcinoma, squamous cell carcinoma and normal tissues with the area under the curve (AUC) of 0.97. Sevakula and colleagues (2018) developed a framework for cancer classification using gene expression data and a transfer learning technique. They first used stacked autoencoders to generate common data representations among different cancers and then used those common features to initialize the parameters of each model for particular cancer types [81]. Ramazzotti *et al.* (2018) presented a Multikernel Learning-based novel workflow to identify molecular subtypes of various types of cancer by integrating multi-omics data [65]. They successfully demonstrated the potential of multi-omics data integration and analysis for patient stratification. Ceccarelli *et al.* (2016) implemented an unsupervised random forest method to cluster Glioma patients by using a whole-genome sequencing dataset [66].

On the other hand, successful implementations of patient stratification via machine learning in other disease domains remained mostly challenging either due to the complex natures of the multifactorial diseases or due to the availability of very little data and/or knowledge. Ting *et al.* (2017) published a very successful deep learning-based framework for the identification of diabetic retinopathy and related eye diseases. Their convolutional neural network (CNN) based model was trained with thousands of retinal images from multiethnic populations with diabetes and able to identify different eye diseases with very high sensitivity and specificity [82]. Our own work (De Jong *et al.* 2019), developed a deep learning-based framework that can successfully cluster multivariate time series data with lots of missing values. Their method of a variational autoencoder with recurrence (VaDER) can successfully stratify AD and PD patients into different subgroups with clinically relevant feature associations [83]. In another instance, Lopez *et al.* (2018) implemented an unsupervised machine learning method to find patient clusters based on genetic signatures. They successfully demonstrated their framework on the genome-wide association data of 191 multiple sclerosis patients to cluster them [84]. Tosto *et al.* (2016) applied k-Means clustering algorithms on longitudinal assessments and neuropathological data from 3502 AD patients. They were able to identify distinct subgroups of AD patients with distinct clinical and neuropathological features [85]. Another similar study by Gamberger *et al.* (2016)

used the multi-layer clustering method on both biological and clinical datasets from Alzheimer's disease neuroimaging initiative (ADNI) database. Their method identified three distinct homogenous AD patient clusters with significant dementia problems [86]. Recently, Zeiberg and colleagues (2019) successfully used simple logistic regression to stratify patients with higher risks of acute respiratory distress syndrome (ARDS) by using EHR data [87].

While the above-mentioned studies demonstrate the true potential of data-driven prediction in patient stratification in different disease domains, most of the approaches had very little interpretation capabilities about their predictions. This is due to the inherent nature of the ML model's black box nature. Since most ML models capture complex and non-linear correlations between predictor variables and clinical outcomes, it is often very difficult to explain how they work. However, it is almost impossible to convince clinicians and policymakers to accept such black-box models for their decision making. Hence, it is essential to work on developing more interpretable ML models for the successful adaptation of such systems in the healthcare system. The integration of prior knowledge is a very promising approach to interpret ML models in the context of current biomedical knowledge. Moreover, advances in such integrative ML models can help in the generation of new hypotheses and understanding the mechanisms underlying disease conditions. In the following paragraph, we will discuss some of such integrative studies that incorporated prior knowledge with other data modalities for their predictions.

Cun and Fröhlich (2012) compared performances of several machine learning algorithms for patient stratification by using transcriptomics data from six public breast cancer datasets. They found that the incorporation of prior knowledge from the pathway and protein-protein interaction databases can greatly enhance the interpretability of the model predictions [88]. In another big collaborative study, Costello and colleagues (2014) also compared performances of several ML algorithms to predict therapeutic responses in various breast cancer cell lines using multi-omics data. They observed all methods that integrated pathway knowledge had higher prediction performance than other methods [89]. a set of tumor mutation profiles can be stratified into subtypes that are both biologically and clinically informative. Hofree *et al.* (2013) demonstrated a network-based stratification of tumor mutations in three different cancer datasets which are both biologically and clinically informative. They were able to stratify different tumor subtypes and associate them into distinct network modules by using somatic tumor mutation data in the context of various human interaction networks [90]. Chang *et al.* (2015) proposed another workflow based on a multivariate Cox proportional hazards regression model to predict different risk groups of lung adenocarcinoma patients by using gene expression datasets. Later, they used biological function or pathway analysis to associate potential biological mechanisms involved with each risk group [91]. In a recent study, Manica *et al.* (2019) proposed a pathway-induced multiple kernel learning (PIMKL) algorithm for the patient stratification by integrating prior knowledge and multimodal cancer data. They used multiple interaction-aware kernel functions to integrate molecular interaction networks and annotated pathways as prior knowledge with multi-omics cancer datasets. The proposed framework was not only able to successfully classify breast cancer patients but also provided insights into the molecular mechanisms that underlie the classification [92].

Although integrative machine learning approaches that can incorporate prior knowledge with the data for the patient stratification are becoming widespread in the oncology field as discussed above, such methods are still to be successfully implemented in the field of neurodegenerative diseases. Such patient stratification will not only enable us to identify different treatment response groups in these complex diseases but also help us to identify responsible mechanisms for each of the subgroups. Finally, such stratifications may also explain the role of various common and new biomarkers for both Alzheimer's and Parkinson's disease etiology.

## 1.6 Drug Repositioning

Drug repositioning is the process of using approved drugs for one indication to treat a new indication [93]. This practice has become increasingly attractive since it can bypass many steps of traditional drug discovery as it relies on known drugs. On the other hand, the classical drug discovery process is complicated, time-consuming, and very expensive. Despite the continuous efforts from researchers in academia and pharmaceutical companies, the whole process remains highly failure-prone. Developing a new prescription drug takes approximately 10 to 15 years and 1.5 billion dollars [94]. Drug repositioning strategy can significantly reduce the time and effort of the drug development process by bypassing pre-clinical testing and sometimes even some preliminary phases of clinical trials [95]. Therefore, the drug repositioning process has emerged as a strong alternative to traditional drug discovery research among pharmaceutical companies [96]. While 113 new drugs and biologics were approved in 2017, 36 previously approved drugs were repositioned for new indications [97].

While drug repositioning had mostly relied on serendipitous discoveries in the past, this domain has started to incorporate systematic approaches based on computational analyses over the last decades [98]. However, it is only recently that computational drug repositioning practices have become widespread due to the increased availability of omics and clinical data. Current computational drug repositioning methods have evolved to integrate and utilize various multi-omics data including genetic [99], chemical [100], pharmacological [101], and clinical [102] data. Additionally, easy access to millions of patient-level data like EHRs, clinical imaging, smart device data, etc., through big resources and consortiums like UK biobank [103] and European Medical Information Framework (EMIF) [104] will have an unprecedented impact on the future computational drug repositioning research.

Computational drug repositioning methods focus on two main different orientations defined by the information source, namely (i) drug-oriented, where repositioning strategy begins from the chemical or pharmaceutical perspective, and (ii) disease-oriented, where repositioning strategy begins from the clinical perspective of disease or its pathology. Based on these orientations computational drug repositioning methods can further be categorized into different classes including target-based, expression-based, knowledge-based, chemical structure-based, pathway-based and mechanism of action-based [105]. In the following section, we will discuss different drug repositioning studies around neurodegenerative diseases.

One of the very first examples of successful drug repositioning in NDD is the recommendations of using an epilepsy drug Zonisamide for the treatment of PD [106]. Murata and colleagues (2001) found that zonisamide can also improve the symptoms of PD patients while treating an epilepsy patient with PD. However, recent advances in molecular signature-based drug repositioning techniques that integrate multi-omics data, enabling successful implementation of such methods in NDDs as well. Zhang *et al.* (2016) illustrated a drug repositioning workflow that integrates multi-omics data (i.e., genomics, epigenomics, proteomics, and metabolomics) and prior knowledge of disease pathogenesis (PubMed [107] and OMIM [108]) and drug-target interactions (i.e., DrugBank [109] and Therapeutic Target Database [110]) from publicly available sources [111]. Their systematic data mining based workflow was able to identify 18 druggable protein targets in AD. They further prioritized these anti-AD targets with their ranking algorithms and identified 7 drugs that can prevent the activities of prioritized anti-AD targets. Molecular docking based drug repositioning methods are in practice for a while and hold big potential in NDD drug discovery research. Xie and colleagues (2016) performed a virtual screening of hundreds of FDA-approved drugs on seven known major AD drug targets and shortlisted several drugs that showed extremely high binding free energies with them [112]. Furthermore, they compared the gene expression profiles of those identified drugs against the docking result in the context of associated pathways to predict several FDA-approved drugs including droperidol, glimepiride, risperidone, etc., as potential multi-target candidates for the treatment of AD.

While there is a surge of computational drug repositioning approaches of various kinds in recent years, the introduction of machine learning-based methods has the true potential to revolutionize this research domain. For instance, Jamal *et al.* (2015) presented a machine learning-based approach that analyzes chemical descriptors of thousands of striatal-enriched protein tyrosine phosphatase (STEP) inhibitor compounds for drug candidate predictions in AD via a combination of machine learning, molecular docking, and molecular dynamics (MD) simulation approaches [113]. They first prioritize molecules based on their chemical properties and activities via different machine learning algorithms (i.e., Naive Bayes, Random Forest, and SVM) and later screened the prioritized compounds by using molecular docking and MD simulations for gaining better insights of their binding mechanisms and strength. A study by Romeo-Guitart *et al.* (2018) proposed a systems biology and artificial intelligence-based drug discovery framework, therapeutic performance mapping system (TPMS), to identify neuroprotective agents by using proteomic data from preclinical models and molecular interaction database [114]. First, using literature knowledge they built a protein network on motoneuron root avulsion, and then converted it into topological maps associated with mathematical equations. Next, they combined the output of the machine learning model trained with proteomic data from the preclinical models with the mathematical models. Finally, the mathematical models identified a putative neuroprotective drug combination of two drugs, Acamprosate and Ribavirin, that can promote neuroprotection, nerve regeneration and functional recovery. Another very recent study by Zeng and colleagues (2019) demonstrated a network-based deep learning framework, deepDR, for in silico drug repositioning by integrating 10 different networks [115]. The deepDR pipeline uses random walk-based network representation to capture network structural information from a complicated heterogeneous network that was built with 10 different drug-related networks. Then, they trained a multi-modal deep autoencoder (MDA) model that can learn compact and low-dimensional features of drugs from the heterogeneous network to predict new drug-disease associations. Finally, they used these low dimensional features to train a collective variational autoencoder (cVAE) model to predict potential associations between drugs and diseases. The deepDR pipeline exhibited high prediction performance and prioritized a couple of potential repurposed drugs for AD and PD.

Although there are several promising drug repositioning approaches available in the area of NDDs, we still lack any drugs that can ameliorate NDD diseases (i.e., AD and PD). While multifactorial and complicated disease biologies of AD and PD make it extremely challenging to find the right target(s) for drug development, systematic integration of prior knowledge either in the form of knowledge assemblies or canonical pathways will give better insights into the pathophysiological conditions to identify drug candidates that tagetes underlying mechanisms. Hence, advanced machine learning-based drug repositioning approaches that not only use patient-level multi-omics and clinical data but also incorporate prior knowledge to enable mechanism oriented drug candidate predictions are required to overcome this challenge.

Finally, despite drug repositioning and precision medicine are two distinct fields of biomedical research, they share the common goal of developing better treatment by disentangling underlying disease mechanisms [116]. As precision medicine focuses on enabling better understanding, characterization, and classification of disease, drug repositioning can leverage these deeper disease understandings to find already approved drugs that can alter the activity of identified targets or pathophysiological mechanisms[117]. Moreover, with the availability of ever-expanding patient-level data (i.e., omics and clinical data) and successful demonstration of advanced machine learning approaches in the biomedical research field, the systematic difference between the underlying workflows of precision medicine and drug repositioning is narrowing down dramatically. Therefore, the concepts of precision medicine and drug repurposing could be seen as two sides of the same coin and needed to be used together in order to get their full benefit in the healthcare system.

## 1.7 Problem Statements

This thesis addresses an important aspect of the precision medicine research of stratifying neurodegenerative disease patients based on molecular and mechanistic disease signatures. The key contributions of our work are two folds: i) investigating prior knowledge from AD and PD knowledge assemblies for the mechanism-based patient taxonomy and drug target identifications, and ii) building a deep learning-based hybrid machine learning model for the joint stratification of AD and PD patients. The thesis is structured as follows:

**Chapter 2** introduces the domain-specific knowledge assembly building and its applications in decoding biologically interesting problems in that domain. We have curated the Alzheimer's Disease knowledge assembly [45] to enrich it with drug interaction information from scientific literature and different chemical interaction databases (i.e., CTD, TTD, and STITCH). Using this chemically enriched AD knowledge assembly we identified some crucial mechanisms in AD that could be targeted by existing drugs in other NDDs. We showed such knowledge assembly could not only help us to portray how different chemicals interact in the context of a biological process but also to enable us to identify druggable mechanisms in NDDs.

**Chapter 3** presents *PS4DR*, a customizable workflow to incorporate prior knowledge with multi-omics datasets from various public databases to predict drug repositioning candidates in different diseases. *PS4DR* integrates high-throughput omics data (i.e., genomics and transcriptomics) from disease and drug perturbations with the prior knowledge in the form of a pathway knowledge (i.e., KEGG, Reactome, and Biocarta) to predict approved drugs in new indications. While two omics data modalities in our workflow help to pinpoint molecular determinants underlying different disease conditions, integrating mechanistic knowledge from the pathway databases enable us to identify which biological processes are interrupted in original disease conditions and/or drug perturbed disease conditions. Finally, anticorrelation scores calculated for both drugs and diseases help to prioritize drug candidates for each disease.

**Chapter 4** introduces a hybrid AI approach to cluster AD and PD patients based on their molecular and mechanistic profiles. We established an unsupervised deep learning clustering approach for the joint stratification of AD and PD patients based on prior knowledge in the form of our AD and PD knowledge assemblies. Our workflow demonstrated the use of mechanistic knowledge with omics data not only successfully stratify complex neurodegenerative diseases (i.e., AD and PD) but also can interpret the predicted patient clusters in the context of current biomedical knowledge. Moreover, underlying mechanisms for each of the patient clusters contain separate druggable targets that will enable us to pursue targeted therapy.

Finally, the last chapter of this thesis summarizes the core message of 'mechanism-based patient stratification and drug repositioning' and discusses the limitations and possible future directions of this work.

# 2 Using Drugs as Molecular Probes: A Computational Chemical Biology Approach in Neurodegenerative Diseases

## Introduction

While data, knowledge, and information in biomedical science are increasing exponentially over the past decades, our standings of disease biology in multifactorial and complex non-mendelian diseases remain inexplicable [118]. The knowledge deficit on the causation of diseases even gets bigger in the area of neurodegenerative diseases. Such ill knowledge of complex disease etiologies translates into the lack of effective treatments for most of those diseases despite gigantic investments of the pharmaceutical companies [22]. This work explored the potential of knowledge assembly, a consolidated and computable collection of domain-specific knowledge, in investigating biological phenomena in the context of neurodegenerative disease. It leverages the semi-automatically curated domain knowledge around Alzheimer's disease to predict repositioning candidates by exploring and analyzing disrupted disease biology in AD.

# Chapter 2

Reprinted with permission from "Emon M. A. *et al.*. (2017) Using drugs as molecular probes: A computational chemical biology approach in neurodegenerative diseases." Journal of Alzheimer's Disease 56.2 (2017): 677-686". Copyright © Emon M. A. *et al.*, (2017).

Journal of Alzheimer's Disease 56 (2017) 677–686
DOI 10.3233/JAD-160222
IOS Press

# Using Drugs as Molecular Probes: A Computational Chemical Biology Approach in Neurodegenerative Diseases

Mohammad Asif Emran Khan Emon[a,b], Alpha Tom Kodamullil[a,b], Reagon Karki[a,b], Erfan Younesi[a] and Martin Hofmann-Apitius[a,b,*]
[a] *Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany*
[b] *Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn-Aachen International Center for IT, Bonn, Germany*

Accepted 7 November 2016

**Abstract**. Neurodegenerative diseases including Alzheimer's disease are complex to tackle because of the complexity of the brain, both in structure and function. Such complexity is reflected by the involvement of various brain regions and multiple pathways in the etiology of neurodegenerative diseases that render single drug target approaches ineffective. Particularly in the area of neurodegeneration, attention has been drawn to repurposing existing drugs with proven efficacy and safety profiles. However, there is a lack of systematic analysis of the brain chemical space to predict the feasibility of repurposing strategies. Using a mechanism-based, drug-target interaction modeling approach, we have identified promising drug candidates for repositioning. Mechanistic cause-and-effect models consolidate relevant prior knowledge on drugs, targets, and pathways from the scientific literature and integrate insights derived from experimental data. We demonstrate the power of this approach by predicting two repositioning candidates for Alzheimer's disease and one for amyotrophic lateral sclerosis.

Keywords: Alzheimer disease, amyotrophic lateral sclerosis, biological expression language, disease-drug modeling, drug repositioning, neurodegenerative diseases

## INTRODUCTION

The human brain represents the most complex biological system, both structurally and functionally. Due to the inherent complexity, treating or even alleviating brain diseases, particularly neurodegenerative diseases, is not trivial. Development of drugs against neurodegenerative diseases has turned out to be among the greatest challenges in the pharmaceutical industry, as reflected by the high attrition rates and withdrawal of high profiled pharmaceutical companies from research on relevant indication areas [1]. A recent survey of success rates of drugs in clinical phases between 2003 and 2011 demonstrates that the likelihood of approval for drug candidates in the category of neurodegenerative diseases was only 9.8%, mainly limited by efficacy issues [2]. In fact, older empirical drug discovery methods are ignorant of mechanisms of actions and modern target-based drug discovery strategies follow a reductionist approach that excessively focuses on drug-receptor interactions and pharmacodynamics/pharmacokinetic properties of the candidate molecule. Both approaches do not consider the complex interplay of various biological entities across multiple biological scales and largely ignore the concept of polypharmacology [3].

Although new postgenomic technologies have produced a considerable amount of data at the molecular level, there has been little progress in inferring disease

*Correspondence to: Prof. Dr. Martin Hofmann-Apitius, Head of the Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53754 Sankt Augustin, Germany. Tel.: +49 2241 14 2802; Fax: +49 2241 14 2656; E-mail: martin.hofmann-apitius@scai.fraunhofer.de.

ISSN 1387-2877/17/$35.00

mechanisms from these data. To overcome this hurdle, computational modeling methods, particularly model-driven systems analysis approaches, have opened up the opportunity to interpret biological datasets in a mechanistic context. Most of the relevant published studies, particularly in the data-rich field of cancer research, have used such methods to model the mechanism drug response around one or two signaling pathways using quantitative data and very limited amount of prior knowledge from literature [4, 5]. However, in a data-scarce field of research such as neurodegenerative diseases, aggregation of prior knowledge plays a key role in unraveling the puzzle of mechanisms underlying disease. This strategy leads to a second category of methods that aim at complementary integration of prior knowledge and experimental data to increase the interpretation power. Biological Expression Language (BEL) (http://www.openbel.org) is a comparably new and state-of-the-art mechanistic modeling syntax that offers a method to combine literature-derived 'cause-and-effect' relationships and data-driven results into a consolidated causal network model, which is amenable to further analysis for mechanistic biological interpretation. We have recently demonstrated the benefit of BEL modeling approaches in the area of neurodegenerative diseases by differential analysis of the mechanisms of Alzheimer's disease (AD) [7].

For example, in the context of AD where most likely multiple mechanisms contribute to its pathology, systematic approaches such as a BEL-based mechanism-of-action discovery method instead of conventional drug-target-centric methods are more likely to deliver new, promising drug candidates. BEL-based modeling can help to repurpose other, already approved drugs with polypharmacological properties from other indications, as many of the drug targets are functionally pleiotropic and involved in multiple diseases. Drug repositioning, defined as the process of identifying and developing new indications for existing drugs, is also known as "drug redirecting", "drug repurposing", or "drug reprofiling" [8]. Off label use of Food and Drug Administration (FDA) approved drugs are very popular in many disease treatments; for instance, 50–75% of prescribed drug therapies for cancer are counted for off label uses [9]. One of the well-known examples for drug repositioning is sildenafil (Viagra), which is used in erectile dysfunction, but was initially developed to treat angina [10]. Another benefit of drug repositioning is that it offers very low risk, as repositioning candidates have already passed through several stages of clinical development. Therefore, repositioning can offer a better risk-versus-reward trade-off compared with other strategies in drug development.

Motivated by the capabilities that come with the cause-and-effect modeling approach, we have developed a causal model of drug-target interactions in the context of NDD. We demonstrate how scattered information existing in the scientific literature can be mechanistically linked to support the detection of putative drug action mechanisms in a defined disease context.

## METHODS

### *Construction of drug-target interaction model around NDD*

Using SCAIView (http://academia.scaiview.com/), our literature mining environment [11], we retrieved NDD related drugs based on PubMed abstracts with the query ([MeSH Disease: "Neurodegenerative Diseases"]) AND [Drug Names]. Next, we extracted related mechanistic information of all the drugs that reached clinical trials for NDDs using simple queries with the defined disease context (NDD) and specific drug names. For example, ([MeSH Disease: "Neurodegenerative Diseases"]) AND [Drug Names: "Donepezil"] has been performed to extract all PubMed articles containing information related to the mode-of-action of donepezil in the NDD context. We manually extracted causal information from these articles and coded into a BEL model. Then, we integrated this model with our in-house AD model [7], to achieve greater disease biology context for the analysis. In order to enrich this primary model by additional interactions for NDD related drugs, we extracted all interactions related to these drugs such as drug-drug interactions, target-target interactions, and drug-target interactions that occur within the brain from different drug interactions databases including Comparative Toxicogenomics Database (http://ctdbase.org/), Therapeutic Target Database (http://bidd.nus.edu.sg/group/cjttd/), DrugBank (http://www.drugbank.ca/), and STITCH database (http://stitch.embl.de/). The purpose of this enrichment was to integrate the biology context around drug targets, in particular those causal relationships that can be used to describe the physiological mode-of-action of a drug-target combination.

*Model analysis and visualization*

We used DAVID (http://david.abcc.ncifcrf.gov), a tool widely used for gene set enrichment analysis, to find the most significant pathways and biological processes associated to the drug targets in our model. The Cytoscape software was used for subnetwork extraction and visualization [12].

*Target similarity approach*

Using the query ([MeSH Disease: "Neurodegenerative Diseases"]) AND [Drug names] in SCAIView, PubMed abstracts were searched for all drugs mentioned in the context of neurodegenerative diseases. Targets of all NDD drugs were systematically compared against targets of five approved drugs for AD, and ranked based on the number of shared targets. Hence, NDD drugs having the highest number of shared targets with approved AD drugs were considered for further analysis in the context of AD. Only those targets having direct interactions are considered for this approach to avoid redundancy. Drugs from these lists were then spotted in the AD-specific BEL-based model for the prediction of similarity of their mechanism of action with five approved AD drugs.

## RESULTS

*Analysis of NDD-specific cause-and-effect model*

The retrieved mechanistic information from the text mining tool SCAIView was manually inspected and filtered for relevant information. Both literature and data driven information were encoded into the NDD-specific BEL model and this model contains 9645 nodes and 26,660 edges including 7,215 genes/proteins, 442 biological processes, 101 disease concepts, and 1,081 chemical entities, coded into 34,403 BEL statements (Fig. 1). This model is comprised of several types of interactions such as drug-target interactions, drug-disease interactions, target-target interactions, target-pathway interactions, and drug-pathway interactions represented by mainly 'increases', or 'decreases' types of relationships in BEL.

GSEA for this model resulted in a list of significant pathways, in which Alzheimer's disease pathway was on top of the list, followed by the Amyotrophic Lateral Sclerosis (ALS) pathway (Table 1). Two 'target sets' associated with these two pathways were selected from the model for further analyses with the intention of the identification of potential drug repurposing candidates from the model.

*Model-based mechanistic analysis of drug repositioning candidates*

We systematically analyzed our model in order to detect the mechanism of action of the drugs in the context of their causal relationships with the available targets, pathways, and biological processes in the NDD-specific mechanistic model. This sort of analysis helps to find possible interaction similarities between drugs of one indication to other indications within the disease context. In this study, we were able to predict three candidate drugs for drug repurposing by using our enriched NDD-specific BEL model:

*Donepezil as potential repurposing candidate for ALS*

Functional analysis of genes/proteins in our model revealed the "ALS disease pathway" as the putative shared pathway with AD, suggestive of evidence to explore the possibility of repositioning drugs between these two diseases. Further analysis of the ALS pathway sub-network based on our model led to the identification of the AD approved drug donepezil as a potential candidate for repositioning. Donepezil affects 26 proteins in the ALS mechanistic pathway sub-network in our model.

Mutant SOD1 protein is believed to be a key player in the pathology of ALS, which disturbs the normal physiological conditions and initiates a number of pathways that ultimately lead to the disease condition [13]. Our mechanistic analysis reveals that donepezil can prevent effects of mutant SOD1 by interfering the activities of many proteins that are altered by this mutation under ALS conditions. Mutant SOD1 protein in ALS mainly exerts its effect by three mechanisms that ultimately lead to neuronal cell death. Firstly, mutant SOD1 can exert its effect by stimulating pro-apoptotic proteins BAD and BAX and inhibiting the activity of anti-apoptotic proteins BCL2 and BCL2L1, which leads to an increase in cytochrome C (CYCS) release from the mitochondria [14, 15]. The activation of BAD and BAX can be also be achieved by recruiting TP53 proteins via mutant SOD1 [16]. The released CYCS interacts and forms a complex with APAF1 in the presence of ATP and activates the key player of the cell death CASP9, which subsequently activates CASP3 and initiates cell death [17]. Secondly, mutant SOD1 can initiate

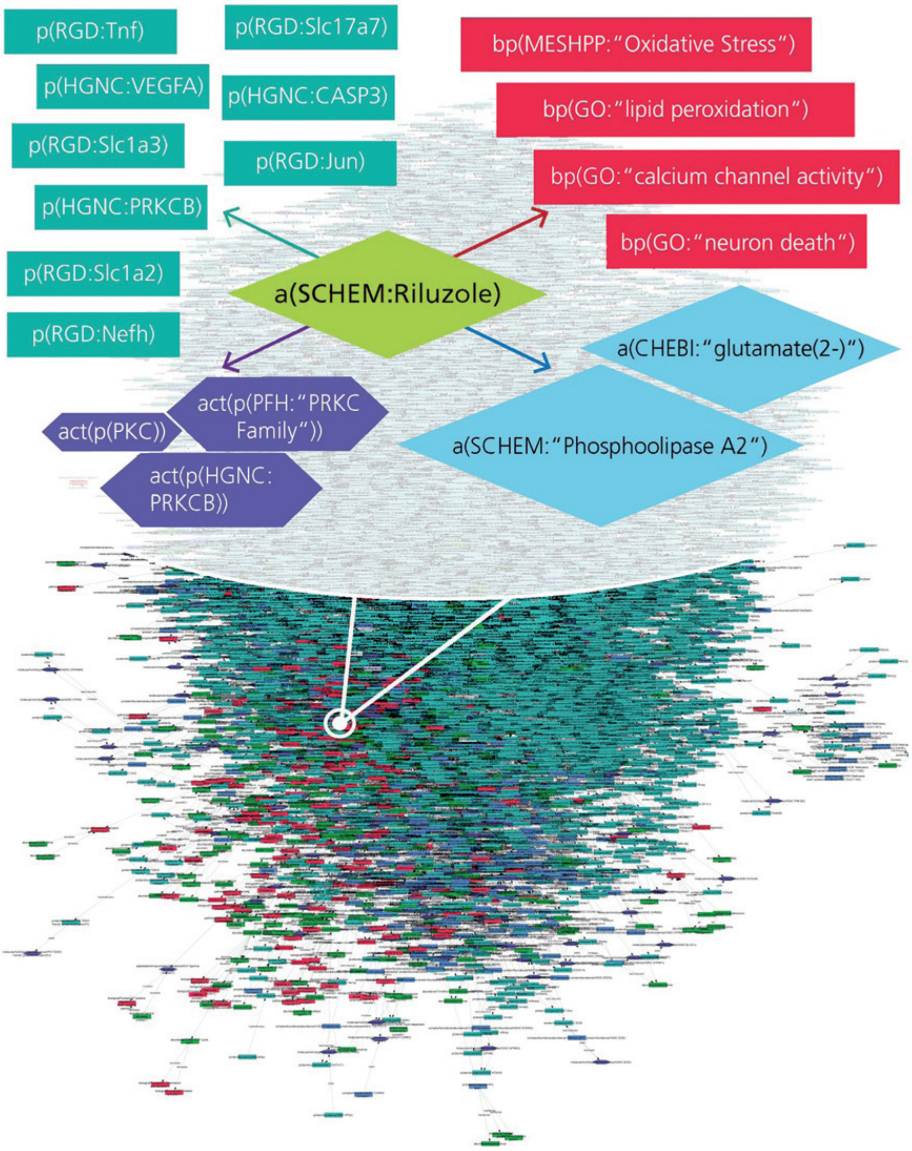


Fig. 1. AD-specific BEL model enriched with drug-target interactions. The extract represents various interaction types for Riluzole such as drug-protein, drug-bioprocess, and drug-pathology interactions encoded into the BEL model.

Table 1
Top Pathways from the gene set enrichment analysis (GESA)

| Pathway Names | Count | *p*-value | Bonferroni | FDR |
|---|---|---|---|---|
| Alzheimer's disease | 30 | 2.02E–29 | 2.43E–27 | 2.32E–26 |
| Amyotrophic lateral sclerosis (ALS) | 24 | 7.85E–18 | 9.41E–16 | 9.01E–15 |
| Pathways in cancer | 50 | 1.98E–14 | 2.37E–12 | 2.27E–11 |
| Prostate cancer | 26 | 4.82E–14 | 5.78E–12 | 5.53E–11 |
| MAPK signaling pathway | 39 | 1.20E–10 | 1.43E–08 | 1.37E–07 |
| Neurotrophin signaling pathway | 26 | 1.46E–10 | 1.75E–08 | 1.67E–07 |
| Bladder cancer | 16 | 1.52E–10 | 1.83E–08 | 1.75E–07 |
| Calcium signaling pathway | 31 | 1.55E–10 | 1.86E–08 | 1.78E–07 |
| Pancreatic cancer | 20 | 2.04E–10 | 2.44E–08 | 2.34E–07 |
| Toll-like receptor signaling pathway | 23 | 4.13E–10 | 4.96E–08 | 4.75E–07 |

oxidative stress via P38 signaling pathway, which in turn inhibits EAAT2, a regulator of the glutamate concentration [18]. The inhibition of EAAT2 produces excess glutamate in synapses, which overstimulate glutamate receptors and initiate high calcium influx in the cytosol and produce reactive oxygen species

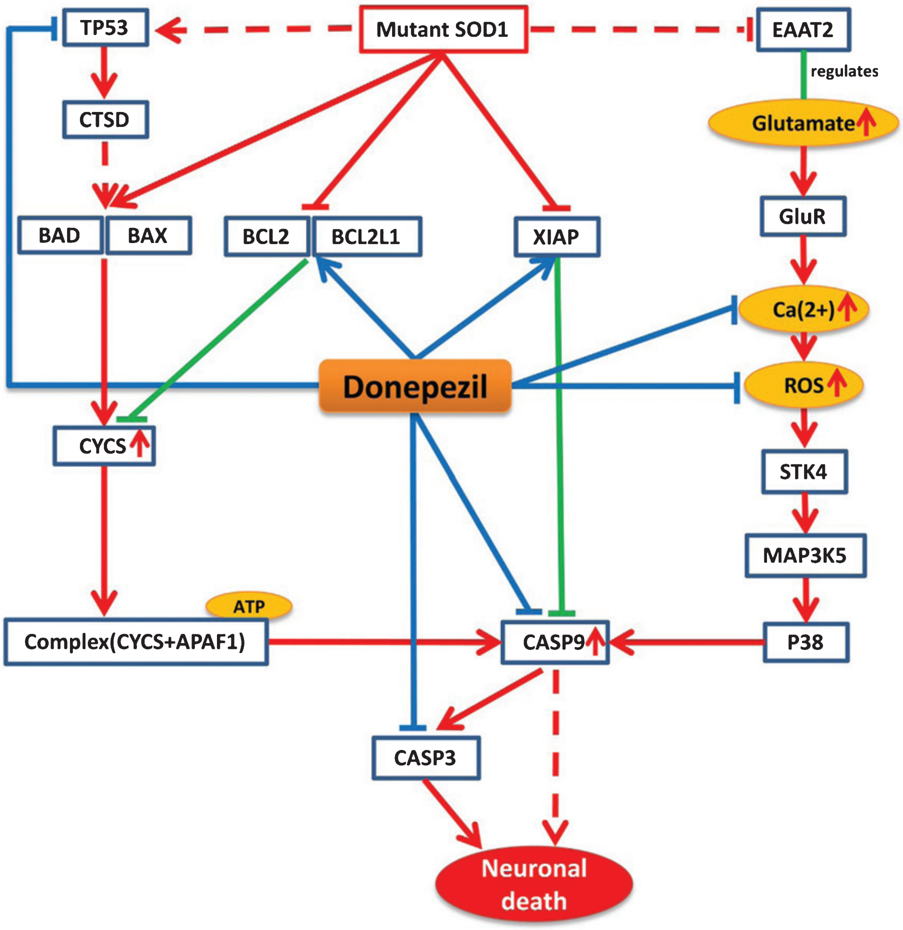


Fig. 2. Model-based prediction of donepezil's mode-of-action in the context of the ALS Pathway. The figure illustrates how donepezil modifies the ALS shared pathway. Red lines represent perturbations in disease condition and green lines indicate normal physiological processes, while blue lines indicate drug effects on targets. Arrows represent increased activities of entities while T lines stand for decreased activities of entities and dotted lines represent intermediate interactions.

(ROS), which will ultimately activate CASP9 [19]. In the third pathway, mutant SOD1 inhibits the activity of XIAP, which regulates CASP9 in normal condition, ultimately leading to activation of CASP9 and CASP3 and initiating neuronal death [20].

According to our model, donepezil can interact with several significant targets in each of these three routes of the ALS pathway. Donepezil can increase the activity of BCL2, BCL2L1, and XIAP, which are inhibited by mutant SOD1 in ALS. It is also able to reduce the level of $Ca^{2+}$ and ROS production in oxidative stress and to inhibit the activation of CASP9 and CASP3 too (Fig. 2).

Epidemiological evidences suggest that early treatment of donepezil in mild cognitive impairment plays a neuroprotective role by preventing neuronal cell death in the hippocampus, hence, reduces the likelihood of disease progression to AD [21, 22]. Interestingly, cognitive dysfunction and inflammation in ALS are broadly associated with morphological changes in the hippocampal region due to excessive neuronal cell death [23–25]. Therefore, based on mechanistic analysis of our model along with the evidences presented; we can hypothesize that donepezil may be a promising repurposing candidate for treating ALS and absolutely worthy of further investigations.

### *Riluzole as potential repurposing candidate for AD*

Being motivated by the donepezil reposition prospect for ALS, we investigated the likelihood of repurposing any ALS drugs for AD with the help of our model. Using our literature-mining environment SCAIView, we found riluzole as the most prominent and effective drug for ALS treatment until now, which helps to prolong the survival of ALS patients. Therefore, we inspected all interactions related to riluzole

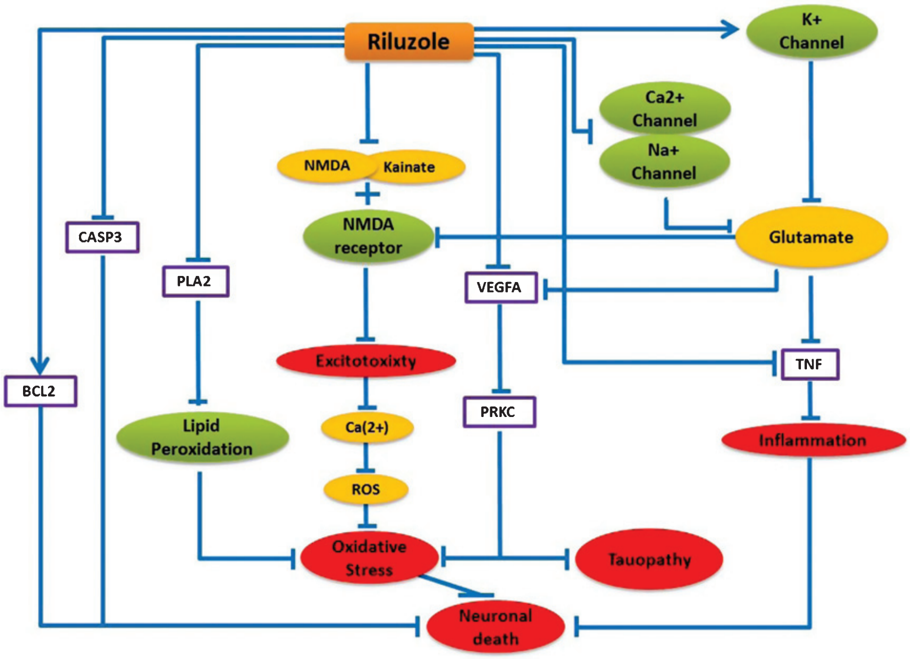


Fig. 3. Schematic representation of Riluzole mechanism of action and its neuroprotective Effect in the context of AD. Blue lines here represent only the alternative effect of riluzole on these pathways. Purple boxes represent the direct protein target and green nodes represent channels and receptors, which can be targeted by riluzole. Yellow nodes represent targeted ions/chemicals and red nodes represent biological processes.

in our model to investigate its potential influences on the AD pathology. Interestingly, our model was able to bring together some mechanisms of actions of riluzole, which interfere with some crucial mechanisms of AD etiology as described below (Fig. 3).

Riluzole is usually known to prolong life longevity of ALS patients by inhibiting $Ca^{2+}$ and $Na^{+}$ channel activities and increasing $K^{+}$ channel activity, which result in decreased glutamate release in the cell [26]. Decreased glutamate concentration contributes to TNF inactivation, which leads to inhibition of inflammatory processes in the brain. The model predicts that riluzole can be beneficial in AD by interfering with a number of mechanistic routes in AD pathology, as follows: in addition to glutamate release inhibition, riluzole can stop neuronal excitotoxicity by inhibiting NMDA and Kainate binding to the NMDA receptor. Riluzole can impede oxidative stress by inhibiting lipid peroxidation via blocking PLA2 activity. Riluzole can also suppress VEGFA and PRKC activities, which are found to be upregulated in AD [27, 28] and believed to be involved in tau phosphorylation [29, 30]. Moreover, riluzole can inhibit pro-apoptotic CASP3 and stimulate anti-apoptotic BCL2 protein to prevent apoptosis [31]. Excessive N-methyl-D-aspartate (NMDA) receptor activation is believed to mediate calcium-dependent glutamate excitotoxicity in different neurodegenerative disorders like AD [32]. Our mechanistic model predicts that glutamate release inhibitor riluzole can provide further therapeutic benefits in AD when used in combination with memantine, the first-in-class approved drug for AD, by modifying excess transmission of synaptic glutamate. Additionally, an ongoing clinical trial NCT01703117, where riluzole is being tested for treating mild stage AD patients, provides further supporting evidence for the mechanism hypothesis we present here. Therefore, we feel encouraged to speculate that riluzole might have therapeutic benefits for AD.

### *Identification of potential drugs for AD by a target similarity approach*

The mechanistic prediction capability of our model inspired us to pursue this slightly different approach for exploring the repositioning potential of drugs present in our model. Analysis for finding common targets between NDD drugs in our model and 5 approved AD drugs identified resveratrol and simvastatin, as drugs that share the highest number of targets with approved AD drugs. Interestingly, these two drugs are already being proposed or investigated for their therapeutics effects in AD (Supplementary

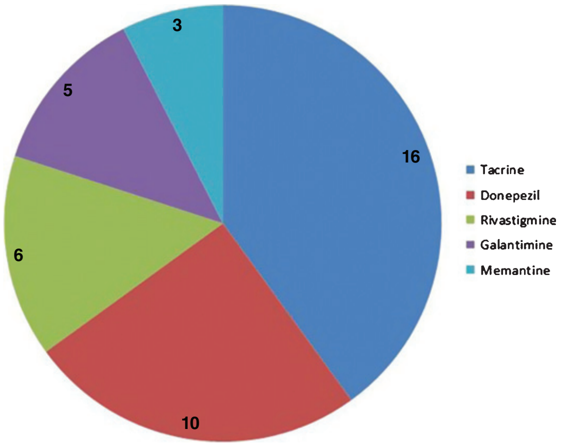


Fig. 4. Distribution of common targets between cyclosporine and five AD approved drugs. The pie chart shows the number of targets that cyclosporine shares with 5 different approved AD drugs.

Table 1), which points to the legitimacy of our approach to find drugs that can be used for similar therapeutic approaches based on their target similarity. To this end, we have selected another top-ranked drug "cyclosporine", from our target similarity list, for further mechanistic analysis.

### *Cyclosporine mode-of-action analysis in the context of AD*

Cyclosporine is a drug against rheumatoid arthritis and has common targets with all five AD approved drugs (Fig. 4), but has not been previously considered for its potential effects in AD.

Instead of being an immunosuppressive agent, the NDD-specific BEL model predicts that cyclosporine can exert neuroprotective effects in various alternative ways. Cyclosporine mainly inhibits immunocompetent lymphocytes (T1 helper cells) in a specific and reversible manner so that T1H cells decrease IL2 and IFNG production and release, which leads to suppression of the immune system response partially (Fig. 5).

According to the model prediction in Fig. 5, cyclosporine can also repress the apoptosis of neuronal cells by inhibiting Cyclophilin D, a member of mitochondrial permeability transition complex (MPTP) [33]. Cyclophilin D inhibition results in regulation of the MPTP complex and decreased Cytochrome C (CYCS) release from mitochondria [34, 35]. This inhibition of CYCS prevents CASP9 and CASP3 mediated apoptotic cell death [20]. The apoptosis inhibition through prevention of CYCS release is also facilitated by the inhibitory effects of cyclosporine on the anti-apoptotic protein BCL2, pro-apoptotic BAD, BAX, and also AKT [36, 37]. Cyclosporine activity inhibits BAX and BAD via stimulation of AKT activity [38], which in turn inhibits GSK3β that phosphorylates and activates BAX [39]. It also inhibits calcineurin, which results in repression of inflammation [40] and down-regulation of ACHE and BCHE, potentially via increasing AKT activity [41, 42], while AKT degeneration leads to increased ACHE and BCHE levels in AD [32]. There is also evidence that cyclosporine decreases ABCB1 and ABCC2 activity [38, 43], which has been reported to increase amyloid-β (Aβ) accumulation in the brain of AD patients [44].

Further support for this hypothesis was provided by a number of patents that explain the putative mechanisms we reconstructed for the potential role of cyclosporine in AD. The claims sections of these patents state clearly, that apart from immunosuppressive activity, cyclosporines could also be effective to improve disease condition by interfering cyclophilins activity and Aβ accumulation. According to US Patents US6583265 and US7538084, cyclosporines can have therapeutic effect in AD by inhibiting the catalytic activity of cyclophilins. A European Patent, EP1893226, recommends the use of cyclosporine to treat AD by preventing Aβ accumulation in the brain in addition to their cyclophilin inhibition activity. Therefore, cyclosporine can be proposed as a multipotent therapeutic agent for AD treatment and this hypothesis bears potential for further clinical investigation.

## DISCUSSION

Structural and functional complexity of the human brain has posed serious challenges to the development of novel therapeutics against neurodegenerative diseases. Capturing this complexity across different molecular entity types and various biological scales can be assisted by computational systems modeling approaches that aim at linking molecular mechanisms to clinical phenotypes. Particularly, in complex diseases like AD, integrating all the entities and bioprocesses involved in the disease into consolidated, cause-and-effect models bears some potential to shed light on interdependent processes and pathways that remain unnoticed in the shadow of disease complexity otherwise. In fact, representing *a priori* relevant knowledge in the form of causal relationship models confers enhanced interpretation power that is well suited to back up experimental data and generate new

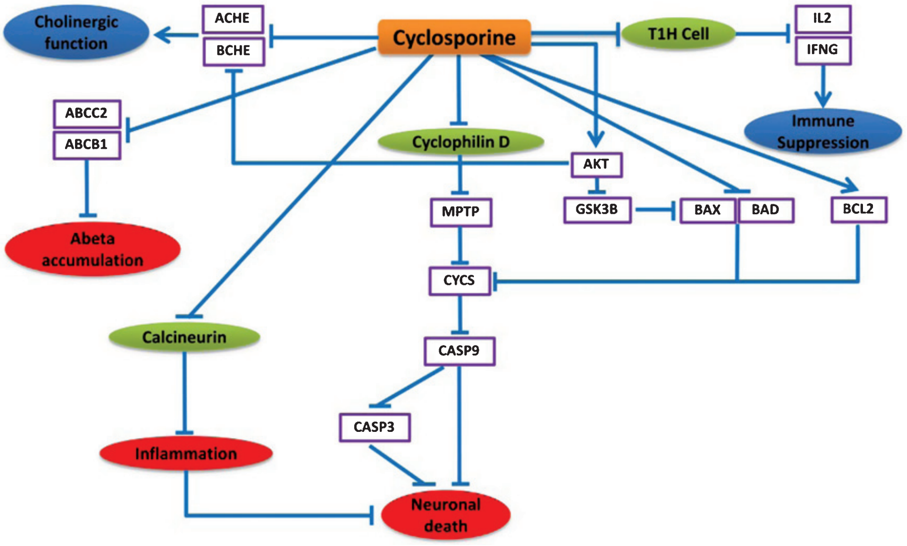


Fig. 5. Neuroprotective effects of cyclosporine in the context of AD. This cartoon demonstrates the mode of action of cyclosporine explicitly. The blue lines here represent the alternative effect of cyclosporine on these different pathways. Cyclosporine mainly inhibits T helper cells to suppress the immune system. Cyclosporine found to affect neuronal cell death by inhibiting cyclophilin D that prevents the cytochrome C release and CASP9, CASP3 activation. Cyclosporine can also down regulate ACHE and BCHE which can provide improved cholinergic function. Moreover, cyclosporine might be useful to prevent amyloid beta accumulation by preventing ABCC2 and ABCB1 proteins.

testable hypotheses. Once such mechanistic, context-sensitive models are available, the molecular space can be enriched for chemical entities to facilitate prediction of mode-of-action for drugs and biomarkers.

As demonstrated in this work, disease-specific mechanistic models that are enriched with chemical entities can be used not only to explain the physiological action mode of approved drugs or candidate drugs, but also to explore the multi-targeting nature of potent compounds and predict the suitability of existing drugs for repurposing in another indication area as well. Based on their role in our cause-and-effect drug-target network, two FDA approved drugs, riluzole and cyclosporine, may be repurposing candidates for AD. Another FDA approved drug, donepezil, could be a potential repurposing candidate for ALS. Although our inferences are based upon the aggregated *a priori* knowledge consolidated in BEL models, further functional or translational validation can be provided by integration of experimental data such as gene expression values. Cross-validation of our models with the signature-based results of Siavelis et al. [14] indicates that rilozule and cyclosporine belong to PKC and GSK3 inhibitor classes of repurposing candidates for AD.

Our approach of using drugs as molecular probes supports the notion that integration of literature-driven information into a formalized model can be instrumental for prediction, analysis, and interpretation of possible biological mechanisms underlying a disease process. Using this approach, we could demonstrate that potential new roles of existing approved drugs can be predicted based on a meaningful functional context. Nevertheless, BEL based mechanistic models, of course, cannot be considered as a replacement for any structure-activity relationship (SAR) model based drug discovery approach. On the contrary, the BEL model presented here merely provides a common platform to put drug-target information into a functional, mechanistic context that focuses on causes and effects and allows for prediction of the repurposing potential of drugs.

It should be noted here, however, that computational models like the ones presented in this study assist hypothesis generation and candidate prioritization. Indeed, these models are merely precursors to clinical and laboratory research findings so that predicted candidates enriched with supporting evidence should be ultimately confirmed by experimental and clinical studies. But such prioritized candidates at this stage can guide future validation efforts in experimental research settings with lower decision-making and investment risk.

## CONCLUSION

Failure of conventional drug discovery and development approaches to deliver new drugs for complex disease like, AD or ALS, has proven that the "one size fits all" paradigm can no longer hold true for chronic and complex idiopathic diseases, particularly in the area of neurodegenerative diseases. This is because of the inherent multitargeting nature of therapeutic agents that modify often unknown pathways with unwanted effects. However, this property can be used positively for repositioning of already approved drugs if the mechanism of action for these drugs can be shown in the context of other diseases. Thus, consolidating the mechanistic information within causal computational models lends support to scientists and decision makers to substantiate their hypotheses based on collective information from both knowledge- and data-driven approaches. It is foreseen that, with the consistent growth of published knowledge and advent of big data, such mechanistic models will play an increasingly important role in the future generation of drug discovery and repurposing pipelines.

## ACKNOWLEDGMENTS

The research leading to these results has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking under AETIONOMY grant agreement n°115568, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and in kind contributions from EFPIA companies.

Authors' disclosures available online (http://j-alz.com/manuscript-disclosures/16-0222r2).

## SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/10.3233/JAD-160222.

## REFERENCES

[1] Wegener G, Rujescu D (2013) The current development of CNS drug research. *Int J Neuropsychopharmacol* **16**, 1687-1693.

[2] Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J (2014) Clinical development success rates for investigational drugs. *Nat Biotechnol* **32**, 40-51.

[3] Enna SJ, Williams M (2009) Challenges in the search for drugs to treat central nervous system disorders. *J Pharmacol Exp Ther* **329**, 404-411.

[4] Wierling C, Kühn A, Hache H, Daskalaki A, Maschke-Dutz E, Peycheva S, Li J, Herwig R, Lehrach H (2012) Prediction in the face of uncertainty: A Monte Carlo-based approach for systems biology of cancer treatment. *Mutat Res* **746**, 163-170.

[5] Chen WW, Schoeberl B, Jasper PJ, Niepel M, Nielsen UB, Lauffenburger DA, Sorger PK (2009) Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol Syst Biol* **5**, 239.

[6] Subramanian A, Subramanian A, Tamayo P, Tamayo P, Mootha VK, Mootha VK, Mukherjee S, Mukherjee S, Ebert BL, Ebert BL, Gillette MA, Gillette MA, Paulovich A, Paulovich A, Pomeroy SL, Pomeroy SL, Golub TR, Golub TR, Lander ES, Lander ES, Mesirov JP, Mesirov JP (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550.

[7] Kodamullil AT, Younesi E, Naz M, Bagewadi S, Hofmann-Apitius M (2015) Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. *Alzheimer's Dement* **11**, 1329-1339.

[8] Ashburn TT, Thor KB (2004) Drug repositioning: Identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* **3**, 673-683.

[9] Pfister DG (2012) Off-label use of oncology drugs: The need for more data and then some. *J Clin Oncol* **30**, 584.

[10] Novac N (2013) Challenges and opportunities of drug repositioning. *Trends Pharmacol Sci* **34**, 267-272.

[11] Friedrich CM, Dach H, Gattermayer T, Engelbrecht G, Benkner S, Hofmann-Apitius M (2008) @neuLink: A service-oriented application for biomedical knowledge discovery. *Stud Health Technol Inform* **138**, 165-172.

[12] Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics* **27**, 431-432.

[13] Andersen PM (2006) Amyotrophic lateral sclerosis associated with mutations in the CuZn superoxide dismutase gene. *Curr Neurol Neurosci Rep* **6**, 37-46.

[14] Vukosavic S, Dubois-Dauphin M, Romero N, Przedborski S (1999) Bax and Bcl-2 interaction in a transgenic mouse model of familial amyotrophic lateral sclerosis. *J Neurochem* **73**, 2460-2468.

[15] Ido A, Fukuyama H, Urushitani M (2011) Protein misdirection inside and outside motor neurons in amyotrophic lateral sclerosis (ALS): A possible clue for therapeutic strategies. *Int J Mol Sci* **12**, 6980-7003.

[16] González de Aguilar JL, Gordon JW, René F, de Tapia M, Lutz-Bucher B, Gaiddon C, Loeffler JP (2000) Alteration of the Bcl-x/Bax ratio in a transgenic mouse model of amyotrophic lateral sclerosis: Evidence for the implication of the p53 signaling pathway. *Neurobiol Dis* **7**, 406-415.

[17] Guégan C, Vila M, Rosoklija G, Hays AP, Przedborski S (2001) Recruitment of the mitochondrial-dependent apoptotic pathway in amyotrophic lateral sclerosis. *J Neurosci* **21**, 6569-6576.

[18] Lee JK, Hwang SG, Shin JH, Shim J, Choi E-J (2014) CIIA prevents SOD1(G93A)-induced cytotoxicity by blocking ASK1-mediated signaling. *Front Cell Neurosci* **8**, 179.

[19] Boillée S, Vande Velde C, Cleveland DW (2006) ALS: A disease of motor neurons and their nonneuronal neighbors. *Neuron* **52**, 39-59.

[20] Inoue H, Tsukita K, Iwasato T, Suzuki Y, Tomioka M, Tateno M, Nagao M, Kawata A, Saido TC, Miura M, Misawa H, Itohara S, Takahashi R (2003) The crucial role of caspase-9

in the disease progression of a transgenic ALS mouse model. *EMBO J* **22**, 6665-6674.

[21] Petersen RC, Thomas RG, Grundman M, Bennett D, Doody R, Ferris S, Galasko D, Jin S, Kaye J, Levey A, Pfeiffer E, Sano M, van Dyck CH, Thal LJ (2005) Vitamin E and donepezil for the treatment of mild cognitive impairment. *N Engl J Med* **352**, 2379-2388.

[22] Min D, Mao X, Wu K, Cao Y, Guo F, Zhu S, Xie N, Wang L, Chen T, Shaw C, Cai J (2012) Donepezil attenuates hippocampal neuronal damage and cognitive deficits after global cerebral ischemia in gerbils. *Neurosci Lett* **510**, 29-33.

[23] Westeneng H-J, Verstraete E, Walhout R, Schmidt R, Hendrikse J, Veldink JH, van den Heuvel MP, van den Berg LH (2015) Subcortical structures in amyotrophic lateral sclerosis. *Neurobiol Aging* **36**, 1075-1082.

[24] Raaphorst J, van Tol MJ, de Visser M, van der Kooi AJ, Majoie CB, van den Berg LH, Schmand B, Veltman DJ (2015) Prose memory impairment in amyotrophic lateral sclerosis patients is related to hippocampus volume. *Eur J Neurol* **22**, 547-554.

[25] Brownell A-L, Kuruppu D, Kil K-E, Jokivarsi K, Poutiainen P, Zhu A, Maxwell M (2015) PET imaging studies show enhanced expression of mGluR5 and inflammatory response during progressive degeneration in ALS mouse model expressing SOD1-G93A gene. *J Neuroinflammation* **12**, 217.

[26] Albo F, Pieri M, Zona C (2004) Modulation of AMPA receptors in spinal motor neurons by the neuroprotective agent riluzole. *J Neurosci Res* **78**, 200-207.

[27] Tang H, Mao X, Xie L, Greenberg DA, Jin K (2013) Expression level of vascular endothelial growth factor in hippocampus is associated with cognitive impairment in patients with Alzheimer's disease. *Neurobiol Aging* **34**, 1412-1415.

[28] Yoo MH, Hyun HJ, Koh JY, Yoon YH (2005) Riluzole inhibits VEGF-induced endothelial cell proliferationin vitro and hyperoxia-induced abnormal vessel formation in vivo. *Investig. Ophthalmol Vis Sci* **46**, 4780-4787.

[29] Hoshi M, Nishida E, Miyata Y, Sakai H, Miyoshi T, Ogawara H, Akiyama T (1987) Protein kinase C phosphorylates tau and induces its functional alterations. *FEBS Lett* **217**, 237-241.

[30] Correas I, Díaz-Nido J, Avila J (1992) Microtubule-associated protein tau is phosphorylated by protein kinase C on its tubulin binding domain. *J Biol Chem* **267**, 15721-15728.

[31] Hassanzadeh K, Roshangar L, Habibi-asl B, Farajnia S, Izadpanah E, Nemati M, Arasteh M, Mohammadi S (2011) Riluzole prevents morphine-induced apoptosis in rat cerebral cortex. *Pharmacol Reports* **63**, 697-707.

[32] Parihar MS, Brewer GJ (2010) Amyloid-β as a modulator of synaptic plasticity. *J Alzheimers Dis* **22**, 741-763.

[33] Kajitani K, Fujihashi M, Kobayashi Y, Shimizu S, Tsujimoto Y, Miki K (2008) Crystal structure and putative function of small Toprim domain-containing protein from Bacillus stearothermophilus. *Proteins* **70**, 311-319.

[34] Xie J-R, Yu L-N (2007) Cardioprotective effects of cyclosporine A in an in vivo model of myocardial ischemia and reperfusion. *Acta Anaesthesiol Scand* **51**, 909-913.

[35] Duan X, Ji B, Yu K, Hei F, Liu J, Long C (2011) Acidic buffer or plus cyclosporine A post-conditioning protects isolated rat hearts against ischemia-reperfusion injury. *Perfusion* **26**, 245-252.

[36] Hage-Sleiman R, Esmerian MO, Kobeissy H, Dbaibo G (2013) p53 and ceramide as collaborators in the stress response. *Int J Mol Sci* **14**, 4982-5012.

[37] Weon JB, Yun BR, Lee J, Eom MR, Ko HJ, Lee HY, Park DS, Chung HC, Chung JY, Ma CJ (2014) Neuroprotective effect of steamed and fermented Codonopsis lanceolata. *Biomol Ther (Seoul)* **2014**, 246-253.

[38] Yamazaki M, Li B, Louie SW, Pudvah NT, Stocco R, Wong W, Abramovitz M, Demartis A, Laufer R, Hochman JH, Prueksaritanont T, Lin JH (2005) Effects of fibrates on human organic anion-transporting polypeptide 1B1-, multidrug resistance protein 2- and P-glycoprotein-mediated transport. *Xenobiotica* **35**, 737-753.

[39] Linseman DA, Butts BD, Precht TA, Phelps RA, Le SS, Laessig TA, Bouchard RJ, Florez-McClure ML, Heidenreich KA (2004) Glycogen synthase kinase-3beta phosphorylates Bax and promotes its mitochondrial localization during neuronal apoptosis. *J Neurosci* **24**, 9993-10002.

[40] Norris CM, Kadish I, Blalock EM, Chen K-C, Thibault V, Porter NM, Landfield PW, Kraner SD (2005) Calcineurin triggers reactive/inflammatory processes in astrocytes and is upregulated in aging and Alzheimer's models. *J Neurosci* **25**, 4649-4658.

[41] Herink J, Krejčová G, Bajgar J, Svoboda Z, Květina J, Živnú P, Palička V (2003) Cyclosporine A inhibits acetylcholinesterase activity in selected parts of the rat brain. *Neurosci Lett* **339**, 251-253.

[42] Jennen DGJ, Magkoufopoulou C, Ketelslegers HB, van Herwijnen MHM, Kleinjans JCS, van Delft JHM (2010) Comparison of HepG2 and HepaRG by whole-genome gene expression analysis for the purpose of chemical hazard identification. *Toxicol Sci* **115**, 66-79.

[43] Luo L, Sun YJ, Yang L, Huang S, Wu YJ (2013) Avermectin induces P-glycoprotein expression in S2 cells via the calcium/calmodulin/NF-??B pathway. *Chem Biol Interact* **203**, 430-439.

[44] Song JS, Chae J-W, Lee K-R, Lee BH, Choi EJ, Ahn SH, Kwon K-I, Bae MA (2011) Pharmacokinetic characterization of decursinol derived from Angelica gigas Nakai in rats. *Xenobiotica* **41**, 895-902.

# Conclusions

This study showed the integration of drug interaction information to the AD knowledge assembly helped to enhance the understanding of the mechanism of actions of those drugs in the context of AD disease biology. This new incorporation enabled AD knowledge assembly to link thousands of drug-target information with various multi-scale AD biomarker knowledge (i.e., from the molecular level to tissue or organ level) and thus made it more comprehensive for the holistic view of disease biology. Hence, it created the opportunity for the researchers to look into those drugs from the disease biology perspective instead of investigating them only at target levels. As a result, it allows us to go beyond the target-based drug repositioning to predict drugs that focus rather on whole disrupted disease biology. Later, the network analysis of this enriched AD knowledge assembly enabled us to identify several disrupted mechanisms in AD and predicted several drugs that could modify the actions of several targets in those identified mechanisms. Looking forward, the semi-automatic curation workflow developed during the course of this study has the potential to serve as a guideline to enrich the knowledge assemblies by using BEL as a medium for knowledge capturing and representation. Finally, this work has demonstrated that a knowledge assembly is a useful tool for the mechanism-based drug repositioning in the context of NDDs, while the benefit of such a workflow could be translated to any other diseases.

# 3 PS4DR: A multimodal workflow for identification and prioritization of drugs based on pathway signatures

## Introduction

Traditional drug development remains a low yielding 'hit and miss' approach, despite the endless efforts from the researchers and colossal investments of pharmaceutical companies. Drug repositioning, on the other hand, has emerged as a more suitable alternative due to its flexibility of bypassing many steps in traditional drug discovery and thus reduced cost [96]. Moreover, increased availability of advanced computational methods and big scale multi-omics data in biology encouraged the widespread adoption of computational drug repositioning strategies. However, there is an urgency for more systematic approaches for drug repositioning that can integrate multi-omics data with mechanistic information from prior knowledge. Such methods will not only be able to predict repositioning candidates based on the patterns observed from omics data but also be able to provide a better understanding of the drug's mechanism of action in the disease context. To address this challenge, we demonstrate PS4DR, a flexible drug repositioning workflow based on the incorporation of pathway information, genomics, and transcriptomics data.

# Chapter 3

Reprinted with permission from "Emon M. A. *et al.*. (2020) *PS4DR*: A multi-modal workflow for identification and prioritization of drugs based on pathway signatures. BMC Bioinformatics, 20:243". Copyright © Emon M. A. *et al.*, (2020).

Emon *et al. BMC Bioinformatics* (2020) 21:231
https://doi.org/10.1186/s12859-020-03568-5

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

# PS4DR: a multimodal workflow for identification and prioritization of drugs based on pathway signatures

Mohammad Asif Emon[1,2*], Daniel Domingo-Fernández[1,2*], Charles Tapley Hoyt[1,2] and Martin Hofmann-Apitius[1,2]

* Correspondence: mohammad.asif.emon@scai.fraunhofer.de; daniel.domingo.fernandez@scai.fraunhofer.de
[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (Fraunhofer SCAI), 53757 Sankt Augustin, Germany
Full list of author information is available at the end of the article

## Abstract

**Background:** During the last decade, there has been a surge towards computational drug repositioning owing to constantly increasing *-omics* data in the biomedical research field. While numerous existing methods focus on the integration of heterogeneous data to propose candidate drugs, it is still challenging to substantiate their results with mechanistic insights of these candidate drugs. Therefore, there is a need for more innovative and efficient methods which can enable better integration of data and knowledge for drug repositioning.

**Results:** Here, we present a customizable workflow (*PS4DR)* which not only integrates high-throughput data such as genome-wide association study (GWAS) data and gene expression signatures from disease and drug perturbations but also takes pathway knowledge into consideration to predict drug candidates for repositioning. We have collected and integrated publicly available GWAS data and gene expression signatures for several diseases and hundreds of FDA-approved drugs or those under clinical trial in this study. Additionally, different pathway databases were used for mechanistic knowledge integration in the workflow. Using this systematic consolidation of data and knowledge, the workflow computes pathway signatures that assist in the prediction of new indications for approved and investigational drugs.

**Conclusion:** We showcase *PS4DR* with applications demonstrating how this tool can be used for repositioning and identifying new drugs as well as proposing drugs that can simulate disease dysregulations. We were able to validate our workflow by demonstrating its capability to predict FDA-approved drugs for their known indications for several diseases. Further, *PS4DR* returned many potential drug candidates for repositioning that were backed up by epidemiological evidence extracted from scientific literature. Source code is freely available at https://github.com/ps4dr/ps4dr.

**Keywords:** Drug repositioning, Drug discovery, Multi-omics, Pathways, Software, Bioinformatics

## Background

De novo drug discovery remains a time-consuming, costly, and failure-prone process, despite advances in high-throughput data generation techniques and analytical approaches. On average, it takes approximately 10 to 15 years and 1.5 billion dollars to bring a drug to market [1]. While traditional drug discovery research is able to propose numerous candidate drugs, the majority of them fail in clinical trials due to lack of efficacy or undesired effects in these trials [2]. Therefore, drug repositioning has emerged as an alternative in drug discovery research [3] that hinges on identifying new indications for investigational or approved drugs in order to reduce the time and cost of preclinical development and primary stages of clinical trials.

Computational drug repositioning methods have recently become popular due to the increased availability of drug-related *-omics* data through sources like CMap (Connectivity Map [4]) and LINCS (Library of Integrated Network-Based Cellular Signatures [5]) (see Tanoli et al. [6] for a review on databases and methods). In recent years, they have evolved to accommodate and utilize novel high-throughput data such as genetic [7], chemical [8], pharmacological [9], and clinical [10]. Computational drug repositioning methods can be categorized as (i) drug-based, where knowledge comes from the chemical or pharmaceutical perspective, or (ii) disease-based, where the strategy focuses on different aspects of the disease, such as symptomatology or pathology [11]. Following, we outline methods from both categories that involve the usage of transcriptomics and GWAS data for drug repositioning purposes.

Transcriptomics data has historically been used to unravel the molecular mechanisms of complex diseases [12–14]. Accordingly, numerous drug repositioning approaches have relied on contrast experiments of transcriptomics readouts such as disease samples, drug perturbed cells and animal models to identify drugs that revert the signature of the disease and eventually its pathogenic phenotype to ultimately predict new indications for existing drugs [4, 15, 16]. To facilitate novel approaches that could systematically exploit this concept, Lamb et al. [4] developed a comprehensive catalog of small molecule perturbed gene expression signatures called CMap. They demonstrated that gene expression signatures can be used to identify drugs with shared mechanisms of action (MoAs), discover unknown MoAs of drugs, and propose potential new therapeutics. Furthermore, a variant of the CMap method was later used by Sirota et al. [16] to compare disease gene signatures against drug-induced gene expression signatures to score each drug-disease pair based on their similarity profile for drug repositioning.

However, the high dimensionality of gene expression signatures has motivated the use of network-based analysis to assist in the interpretation of biological processes which are perturbed by a given drug. Not only are these analyses instrumental in determining relevant molecular signatures as markers of phenotypes but also in garnering novel mechanistic insights into various biological functions and disease. For example, Iorio et al. [15] used Gene Set Enrichment Analysis (GSEA [17]) to build a drug similarity network from the distances of the GSEA scores for each drug pair in order to investigate the biological processes enriched in a set of drug subnetworks to identify compounds with similar MoAs. Suthram et al. [18] integrated disease gene expression signatures with large scale protein-protein interaction networks to identify disease similarities. They discovered a set of common pathways and processes which were dysregulated in most of the investigated diseases and that could be targeted by the drugs indicated for other diseases. Keiser et al. [19] showed that drug-target interaction networks could be used to predict off-targets for known drugs by comparing the similarity of the ligands that bind to the corresponding targets.

Single nucleotide polymorphisms (SNPs) have gained attention in biomedical research due to the impact of genetic variations in numerous complex diseases. Although the majority of SNPs do not have an effect on the phenotypic outcome, some might be directly involved in disease etiology by affecting the associated gene's function depending on their occurrence in the genomic loci. Therefore, identifying disease-associated SNPs via genetic studies (e.g., GWAS) and targeting the corresponding genes has become a common practice for generating hypotheses to investigate molecular mechanisms of disease. Accordingly, new methods are being developed to incorporate GWAS knowledge in the drug repositioning domain. For instance, Sanseau et al. [7] collected disease-associated genes from the GWAS Catalog [20] and evaluated whether these genes were targeted by drugs. In their post hoc analysis, they observed that these genes were more likely to be a drug target than housekeeping genes. They mapped GWAS genes to the genes which were targeted by drugs listed in the pharmaprojects database (http://www.pharmaprojects.com/) and later proposed that drugs with indications different from the GWAS traits could be of potential drug repositioning interest. In another instance, Lencz and Malhotra [21] used the results from large scale GWAS conducted by the Psychiatric Genomics Consortium–Schizophrenia Workgroup (PGC–SCZ) [22] to predict drug repositioning candidates in schizophrenia. First, they identified the overlap between the known drug targets from Rask-Andersen et al. [23] and potential schizophrenia candidate genes from GWAS. Next, they characterized the MoA of drugs targeting the overlapped genes to propose drugs for schizophrenia treatment. Further, Zhang et al. [24] illustrated another strategy to use GWAS data for prioritizing candidate genes from the GWAS identified loci for drug repositioning. They prioritized genes by scoring them with seven criteria such as cis-eQTL, text mining, and functional enrichment to propose new targets for colorectal cancer drug treatments.

While studies have leveraged transcriptomics and genetics data for prioritizing drug repositioning candidates independently, recent approaches have started to utilize them in combination with other data types. So et al. [25] proposed a framework for drug repositioning by combining GWAS-imputed transcriptome signatures and drug-induced changes in gene expression (CMap) in the field of psychiatric disorders. They imputed gene expression signatures from GWAS summary statistics instead of using expression data from microarray or RNA-sequencing studies and compared them with drug-induced expression changes. Zhang et al. [26] demonstrated another drug repositioning workflow by mining *-omics* data such as GWAS, proteomics, and metabolomics from publicly available sources to find diabetic risk proteins and then filtered them to druggable targets. They further analyzed the pathogenicity of these prioritized targets and found several drugs for these targets that have the potential for diabetic treatments. Later, Ferrero and Agarwal [27] presented a systematic approach which integrated GWAS data and gene expression signatures from diseases and drugs perturbation to generate drug repositioning hypotheses. They demonstrated that (i) GWAS-associated genes in disease are more likely to be differentially expressed in the same disease, and (ii) drug perturbed genes in disease are enriched for GWAS-associated genes in the same disease. They eventually proposed statistically significant drug-disease pairs from the latter analysis could be used for drug repositioning.

Above we surveyed the state-of-the-art in silico strategies for drug repositioning by using transcriptomics and GWAS data. However, there is a lack of systematic approaches that can integrate mechanistic knowledge from pathways with data from multiple modalities to ultimately provide a better understanding of the drug's mechanism of action in the disease context. Therefore, we introduce *PS4DR*, a multimodal and integrative workflow that uses multiple data modalities (i.e., GWAS and transcriptomics) together with pathway knowledge to predict approved drugs in new indications. Finally, we show that our workflow is able to identify FDA-approved drugs for their known indications and predict new indications for existing drugs using publicly available datasets.

## Results

We developed *PS4DR*, an automated workflow that enables the integration of multimodal datasets together with pathway information from different canonical pathway databases to predict drug repositioning candidates in different diseases (Fig. 1). We showcase *PS4DR* using real-world gene expression signatures (i.e., Open Targets [28] and LINCS) and GWAS data (i.e., GWASdb [29], GWAS Catalog [20], GRASP [30], and PheWAS [31]). First, the workflow filters disease and drug transcriptomics (i.e., gene expression signatures) with the help of GWAS data. The next step involves calculating pathway signatures for diseases and drugs via pathway enrichment analysis with the filtered dataset. Finally, *PS4DR* performs an anti-correlation analysis by calculating correlation scores between the pathway signatures of drugs and diseases to prioritize drugs for each disease. Below, we show the utility of the workflow with three applications on how this tool can serve to i) identify drug repositioning candidates, ii) prioritize drug combinations, and iii) propose drugs that simulate disease dysregulations.

### Identifying drug repositioning candidates

As a first application, we explored the list of 26 diseases for which our workflow predicted drug repositioning candidates. While our workflow predicted plenty of drug candidates, we considered two criteria to prioritize predicted drugs. First, we prioritized all drugs in each disease based on their negative correlation scores. However, a drug could have a negative correlation score by only reverting a minority of the pathways dysregulated in the disease. Therefore, we also consider the relative number of the dysregulated pathways reverted by a drug for the prioritization process. While this prioritization approach facilitated narrowing down the candidate lists, we are aware that each of the drugs exhibiting negative correlation scores might have the potential to revert the disease condition even if they alter very few dysregulated pathways.

The distribution and Q-Q plots for the majority of the diseases that output drug predictions demonstrate that the correlation scores follow a normal distribution (Additional file: Fig. S1 and Fig. S2). Hence, we applied an arbitrary threshold to the correlation score to prioritize the proposed candidate drugs in each disease. We would like to point out that we used the same threshold for all diseases since we are exploring

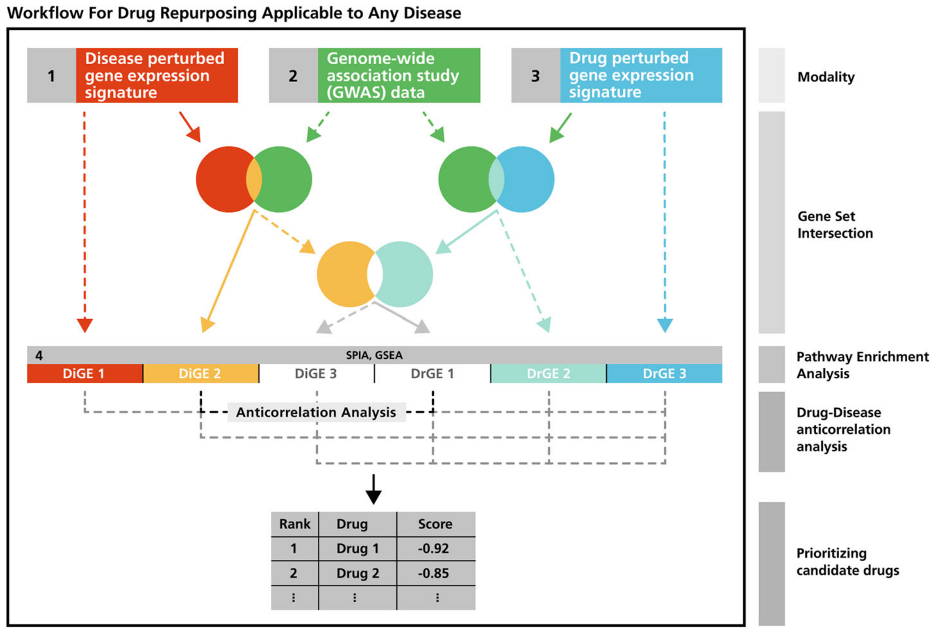

Workflow For Drug Repurposing Applicable to Any Disease
1
Disease perturbed gene expression signature
2
Genome-wide association study (GWAS) data
3
Drug perturbed gene expression signature
Modality
Gene Set Intersection
4
SPIA, GSEA
DiGE 1
DiGE 2
DiGE 3
DrGE 1
DrGE 2
DrGE 3
Pathway Enrichment Analysis
Anticorrelation Analysis
Drug-Disease anticorrelation analysis
Rank
Drug
Score
1
Drug 1
-0.92
2
Drug 2
-0.85
⋮
Prioritizing candidate drugs

**Fig. 1** *An overview of the PS4DR workflow.* The workflow requires three different datasets as inputs, (i) disease perturbed gene expression signatures, (ii) genome-wide association study (GWAS) data, and (iii) drug perturbed gene expression signatures. The first and optional part of the workflow involves different filtering steps based on gene set intersection operations that enable the identification of genes in the gene expression signatures that have also been identified in a GWAS of the studied disease. To retain the maximum flexibility in the workflow, users can decide which of the filtering steps they wish to apply, if any. The next step uses the transcriptomics datasets, filtered or not, to conduct pathway enrichment analysis and evaluate the direction of perturbation for each affected pathway in a particular disease context. While the dotted lines in the figure represent all possible combinations of the filtering steps that can be applied and lead to the pathway enrichment step, solid lines show the option we chose to demonstrate the workflow. Finally, the last step uses the correlation of the pathway scores calculated by the previous step to prioritize drugs that are predicted to invert the pathway signatures observed in a given disease context

multiple indications; however, this threshold could be selected individually for each disease based on their underlying correlation score distributions. The applied threshold discarded drugs with a correlation score greater than − 0.4 or drugs which did not cover more than 50% of the affected pathways in the disease. This filtering step, intended to reduce the number of hits and facilitate the manual investigation of the results, returned a list of predicted drug candidates for 19 diseases (Additional file 1: Table S1). We further investigated the proposed drugs for five conditions to see whether *PS4DR* was able to identify FDA-approved drugs for their known indications and predict new indications for existing drugs in the prioritized list.

First, we focused on the predicted drug list for melanoma. We searched DrugBank [32] and scientific literature to collect evidence for the proposed drugs and summarized our findings in Table 1. Seven of nine predicted drugs are either already being used as cancer drugs or currently being studied in different clinical trials. This motivates further investigation of these drugs as repositioning candidates for the treatment of melanoma.

The topmost drug in our predicted shortlist, Crizotinib, a non-small cell lung cancer (NSCLC) drug, has been reported for its positive effect on melanoma by two studies [33, 34]. While Surriga et al. [33] suggested that Crizotinib could be used in adjuvant therapy for uveal melanoma due to its c-Met activity inhibition, recent research

**Table 1** Drug repositioning candidates for Melanoma. Drugs showing a negative correlation score less than or equal to − 0.40 and affecting more than 50% of the dysregulated pathways in melanoma. The last column outlines the current uses of the given drug in other conditions according to DrugBank and scientific literature

| Drug | DrugBank ID | Correlation Score | Affected Pathways (%) | Description |
|---|---|---|---|---|
| **Crizotinib** | DB08865 | −0.64 | 74.07 | Used for the treatment of locally advanced or metastatic non-small cell lung cancer (NSCLC). |
| **Olmesartan** | DB00275 | −0.85 | 55.56 | Used for the treatment of hypertension. |
| **Sepantronium** | – | −0.21 | 74.07 | Clinical trials in advanced non-small-cell lung cancer. |
| **Bortezomib** | DB00188 | −0.52 | 62.96 | Used for the treatment of multiple myeloma. |
| **Fluspirilene** | DB04842 | −0.5 | 55.56 | Used for the treatment of schizophrenia. |
| **Vistusertib** | DB11925 | −0.44 | 66.67 | Under investigation for the treatment of Advanced Gastric Adenocarcinoma. |
| **Olaparib** | DB09074 | −0.44 | 66.67 | A poly (ADP-ribose) polymerase (PARP) inhibitor indicated for the treatment of Ovarian and Breast Cancer. |
| **Tivozanib** | DB11800 | −0.44 | 66.67 | Used in trials for the treatment of solid tumors, Ovarian Cancer, Glioblastoma, Prostate Cancer among others. |
| **Belinostat** | DB05015 | −0.43 | 55.56 | Used for the treatment of patients with relapsed or refractory peripheral T-cell lymphoma (PTCL). |

reported strong kinase fusion association with different melanoma subtypes [35] and encouraged the testing of kinase fusion inhibitor Crizotinib for melanoma treatment [34]. The third drug, Sepantronium, a selective small-molecule survivin suppressant, was reported to reduce the accumulation of survivin in G2/M mitotic arrest and induce apoptosis in human malignant melanoma cells in combination therapy with docetaxel [36, 37]. The following drug in Table 1, Bortezomib, is an approved drug for multiple myeloma that was suggested as a treatment for melanoma in combination therapy with temozolomide due to its ability to induce apoptosis and autophagic formation in human melanoma tumors [38, 39]. Another FDA approved drug Olaparib (for breast and pancreatic carcinoma), was also found to be effective against melanoma by inhibiting repair of single-strand DNA breaks in different combination therapies [40, 41].

The last two approved drugs in the list (i.e., Tivozanib for renal cell carcinoma and Belinostat for peripheral T-cell lymphoma) have been positively associated with a better response in melanoma [42, 43]. Moreover, another mTOR inhibitor drug, Vistusertib (AZD-2014), currently in phase II clinical trial for meningioma, was reported to have a positive impact by mTORC1/2 inhibition of the resistance to MAPK pathway inhibitors in melanomas with high oxidative phosphorylation [44, 45]. Interestingly, we also have two drugs, Olmesartan, for hypertension, and Fluspirilene, for schizophrenia, from very different therapeutic areas in our shortlist. While no reports of their potential role in melanoma treatment have been found yet, numerous studies have suggested their applicability in different cancer treatments [46–49].

We have found three drugs in breast carcinoma (Additional file 1: Table S1). The first drug, AT-7519, a selective inhibitor of specific Cyclin-Dependent Kinases (CDKs), is under investigation for the treatment of leukemia, lymphoma, myelodysplastic syndrome, and solid tumors [32]. This is in concordance with the study by Yu et al. [50] describing how a subgroup of breast cancer patients benefited from the treatment of

CDK4 kinase inhibitors. The next drug, Omacetaxine Mepesuccinate, used for chronic myeloid leukemia, is in a clinical trial (NCT01844869) for treating advanced solid tumors (i.e., breast, lung, colorectal and melanoma). Finally, Rigosertib has shown potent antitumor activity in various preclinical models such as breast cancer and pancreatic cancer xenografts and is currently under clinical trial [51].

Similarly, we found that six out of eight drugs proposed for pancreatic carcinoma are either already being used in different cancers or have been suggested in the literature, as we discuss below (Additional file 1: Table S1). The first drug, Fenofibrate, an antilipemic agent, was reported to inhibit pancreatic cancer cell proliferation via activation of p53 mediated by upregulation of MEG3 [52]. The next drug, Menadione, was found to induce reactive oxygen species to promote apoptosis via redox cycling in pancreatic cells [53, 54]. Fluoxetine, originally an antidepressant agent, was also reported to work as a chemosensitizer and acts with other cancer drugs to overcome multidrug resistance in cancer cells [55]. An investigational cancer drug, Tosedostat, was found to be well-tolerated and clinically active against pancreatic ductal adenocarcinoma patients in phase I/II clinical trial ([56]; NCT02352831). Another drug, AZD-6482, a selective PI3Kβ inhibitor, could be useful in pancreatic cancer treatment because of its apoptotic effect in cancer cell lines [57]. Praziquantel was reported to inhibit cancer cell growth when used synergistically with paclitaxel via downregulating the expression of X-linked inhibitor of apoptosis protein (XIAP) [58].

While our workflow showed very promising results in cancer, we wanted to explore the results in complex disorders with no available treatments, such as Alzheimer's disease (AD) and multiple sclerosis (MS). In the case of AD, the workflow provided fourteen shortlisted candidates (Table 2). The top drug on the list is Sirolimus (rapamycin), an immunosuppressant, already proposed for the treatment of AD by different studies [59–61]. It has been suggested that the therapeutic effect of this drug is due to the reduction of amyloid-beta levels caused by its inhibition of the mTOR signaling pathway

**Table 2** Drug repositioning candidates for Alzheimer's disease (AD). Drugs showing a negative correlation score less than or equal to −0.40 and affecting more than 50% of the dysregulated pathways in AD

| Drug | DrugBank ID | Correlation Score | Affected Pathways (%) |
|---|---|---|---|
| **Sirolimus (Rapamycin)** | DB00877 | −0.69 | 66.67 |
| **Pevonedistat** | DB11759 | −0.66 | 60.61 |
| **Nilotinib** | DB04868 | −0.64 | 60.61 |
| **Terfenadine** | DB00342 | −0.57 | 57.58 |
| **Doxylamine Succinate** | DB00366 | −0.57 | 54.55 |
| **Halcinonide** | DB06786 | −0.57 | 51.52 |
| **Promazine Hydrochloride** | DB00420 | −0.53 | 66.67 |
| **Mosapride** | DB11675 | −0.45 | 60.61 |
| **Pimozide** | DB01100 | −0.45 | 57.58 |
| **Ritanserin** | DB12693 | −0.45 | 57.58 |
| **Betamethasone** | DB00443 | −0.44 | 66.67 |
| **Cinacalcet Hydrochloride** | DB01012 | −0.43 | 72.73 |
| **Methapyrilene Hydrochloride** | DB04819 | −0.43 | 72.73 |
| **Trametinib** | DB08911 | −0.40 | 60.61 |

[61]. Another compound, Pimozide, an antipsychotic agent, was recently suggested as a potential AD therapeutic which was reported to reduce toxic forms of tau protein by enhanced autophagy activity via AMPK-ULK1 axis stimulation [62]. Interestingly, we have two cancer drugs, Pevonedistat and Nilotinib, which could have potentially positive effects on AD treatment ([63–65]; NCT02947893). Pevonedistat, a neddylation inhibitor, could prevent neuronal damage and ameliorates cognitive deficits by preventing NRF2 protein degradation via inhibiting neddylation [63, 65]. Nilotinib, a tyrosine kinase inhibitor, has also been found to be very promising to delay the progression of AD by enhanced amyloid-beta clearance ([64]; NCT02947893).

Animal studies have demonstrated that the blockade of muscarinic receptors results in increased levels of acetylcholine and improve cognition [66]. Therefore, another proposed drug, Terfenadine which is a muscarinic receptor antagonist and has not yet been linked to AD, could be a potential repositioning candidate. Similarly, several 5-HT6R antagonists have advanced to different phases of clinical trials ([67]; NCT02258152; NCT02580305) as treatments for AD. The results also suggest another drug in the list, Ritanserin, that has not been directly indicated for AD. The high score proposed by our workflow to this serotonin receptor antagonist may be explained by its regulation of the neuronal cholinergic and glutamatergic pathways, both dysregulated in AD. Furthermore, there is increasing evidence showing that neuroinflammation significantly contributes to AD pathogenesis [68, 69]. Hence, it is not surprising to find two anti-inflammatory agents in our list (i.e., Betamethasone and Halcinonide) that could be worth investigating as potential repositioning drugs. Finally, Doxylamine Succinate, a neurotransmitter agent and histamine antagonist, is also a promising candidate since the beneficial effects of histamine antagonists in AD have been reported in multiple studies [70–72].

Finally, we investigated the top ranked drugs proposed by *PS4DR* for multiple sclerosis (MS). Ranked at the top of the list, *PS4DR* successfully recovered methylprednisolone, a corticosteroid with anti-inflammatory action prescribed to treat acute exacerbations in patients with MS [73] (Additional File 1: Table S1).

### Prioritizing drug combinations

Although we have illustrated that our workflow is able to identify candidate compounds for drug repositioning, combining multiple drugs can provide more benefits since the number of affected pathways can be increased by taking advantage of their synergistic effects. Therefore, we applied our workflow to all drug pair combinations in all diseases in order to identify therapies that could have a greater effect than single-drug treatments. For this application, we exclusively considered combinations of two drugs for two reasons: i) application of multiple drugs is usually counterproductive since it increases the number of side effects and ii) calculation time increases exponentially with an increasing number of drugs.

We investigated the predictions of our workflow in breast cancer to verify if we have more drugs with a good negative correlation score and affected pathways (%). While we had three drugs from our single-drug prediction approach, we were able to retrieve 489 drug pairs from the drug combination approach with the same thresholds. To facilitate manual investigation, we increased our threshold of correlation score to less than or equal to − 0.50 and affected pathways greater than or equal to 80% and were still able to retrieve 34 drug pairs (Additional file 1: Table S2). Here, all 19 new drugs in these 34 pairs are partnered with one of the top two drugs, AT-7519 or Omacetaxine Mepesuccinate, from the single-drug approach. Fourteen of the new drugs have partnered with both AT-7519 or Omacetaxine Mepesuccinate. While we have found literature evidence for the beneficial role of seven of these new drugs in the treatment of breast cancer, another six drugs are reported to have positive effects in other solid tumor based cancer treatment as described below. The third drug from the single-drug approach, Rigosertib, which was reported to have antitumor activity in breast cancer cell lines [51], has partnered with both AT-7519 or Omacetaxine Mepesuccinate. BGJ-398, a fibroblast growth factor receptor inhibitor in the list, significantly prevented the outgrowth of tumor organoids in metastatic breast cancers [74]. An approved cancer drug, Erlotinib Hydrochloride, epidermal growth factor receptor inhibitor, has shown a very positive response rate when treated combinedly with Capecitabine and Docetaxel in advanced breast cancer patients [75]. Another drug Selumetinib, a tyrosine kinase inhibitor, is currently being tested in several clinical trials (i.e., NCT03162627; NCT03742102; NCT02503358) for different cancer types, including breast cancer. TAK-715 is a p38 MAP kinase inhibitor in the list that cross-reacts with casein kinase ε (CKIε). Since CKIε mutations have been linked with the proliferation of different breast cancer cell lines, this drug could be explored to repurpose it for breast cancer treatment [76]. Another investigated drug, Tivantinib, has also shown positive effect on breast cancer model by reducing the metastasis via c-MET inhibition [77]. Megestrol Acetate, a progesterone receptor agonist, is under various clinical trials either alone or in combination with other cancer drugs for breast cancer treatment (i.e., NCT03306472 and NCT03024580).

AZD-1775, a drug that inhibits the G2–M cell-cycle checkpoint gatekeeper WEE1 kinase, has been used in multiple trials studying the treatment of lymphoma, ovarian cancer, and adult glioblastoma [32, 78]. Another drug, Axitinib, a selective vascular endothelial growth factor receptor (VEGFR) inhibitor, is under investigation in different clinical trials for various cancer types (i.e., NCT02129647; NCT03494816; NCT03472560). Moreover, four other drugs i,e., BMS-777607, PF-04217903, R-406, and Isotretinoin are reported to have positive effects in different solid tumor cancer types in different studies [32, 79–81].

### Proposing drugs that simulate disease pathway signatures

While we have initially focused on the drugs with the most negative correlation scores, we also anticipated a potential utility for drugs showing positive correlations. Well-characterized drugs with high positive correlation scores can provide information about how pathways or targets could be implicated in the molecular basis of the disease. Hence, as an extended application, the workflow may be used additionally as a prioritization tool to identify drugs that could be potentially employed to generate in-vitro or in-vivo models. By investigating the correlation scores (Fig. 2), researchers can readily identify drugs that could be used for this purpose. Our workflow predicted induction of disease pathway signatures for Pevonedistat in diabetes mellitus, Alvocidib in Crohn’s disease, and Entinostat and panobinostat in systemic lupus erythematosus

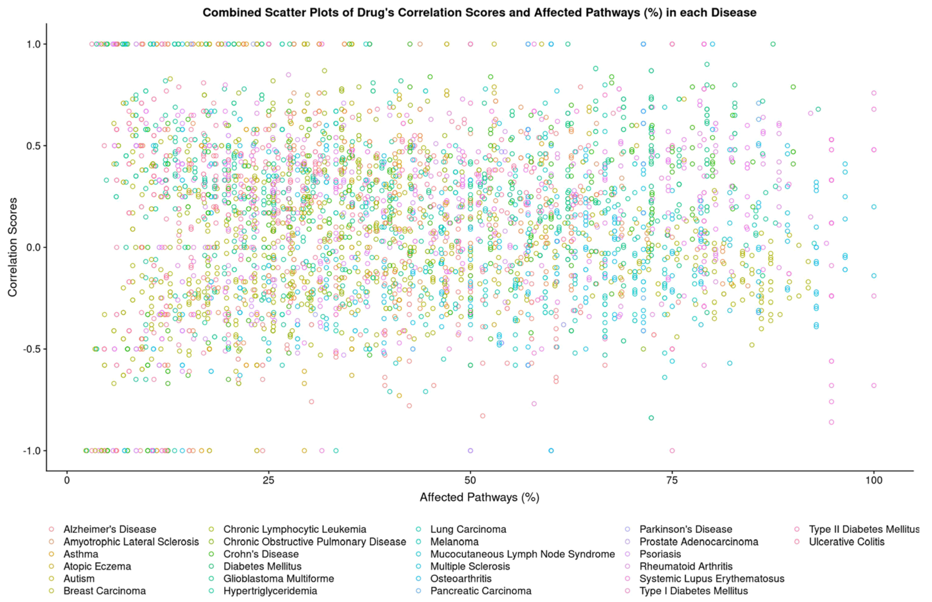


**Fig. 2** Combined scatter plots of the drug's correlation scores against affected pathways (%) in each disease. The relative number of target pathways affected by the drug in the disease context is plotted along the x-axis and correlation scores on the y-axis. Drugs in the top-right corner of the plot might be interesting for developing in vitro disease models since this group of drugs shows positive correlation scores, covering a broad range of the affected pathways. The circles represent drugs and the color coding indicates their respective disease indication, as shown at the bottom

(SLE) through very high positive correlation scores in addition to their broad coverage of affecting disease pathways. We see the need for further investigations of all the drugs with both high positive correlation scores and a high percentage of affected pathways for their use in potential disease model development.

## Discussion

Numerous innovative and interesting methods are constantly being developed to exploit high-throughput biological data in drug discovery research. However, there is still an urgent need for reproducible approaches which could systematically combine mechanistic knowledge with high-throughput data for drug repositioning purposes. In this work, we propose *PS4DR*, a drug repositioning workflow that combines data- and knowledge-driven information for predicting novel indications for prescribed drugs. We demonstrate the workflow using publicly available databases for disease and drug *-omics* data and employing pathway knowledge from various canonical pathway databases. The results show how *PS4DR* provides a comprehensive overview of the targeted pathways by drug or drug combinations and how this information can be useful to identify drug repositioning candidates. Finally, we validated the results of the workflow with epidemiological evidence extracted from the scientific literature to demonstrate that the workflow also prioritizes already approved drugs for numerous conditions.

However, our work is not without limitations, which we plan to address in future research. The connection between drug perturbed gene expression signatures, GWAS data, and disease-specific gene expression signatures is based on statistics derived from gene overlap. While the two latter datasets are disease-specific, drug-derived information is not contextualized. The linkage across the datasets could be more informative if there would be datasets available with drug perturbed gene expression signatures from disease models. Moreover, using advanced techniques such as deep learning [82] or network-based [83] methods to bridge different data modalities by inferring the association between heterogeneous features (i.e., genes, diseases) could also be viable alternative approaches to contextualize the data. Additionally, our workflow is limited to the availability of summarized disease- and drug- perturbed gene expression signatures. Finally, we would like to mention that the drug combination strategy approach is agnostic to other important processes such as kinetics, whether target genes are expressed in the tissue and whether the proposed drugs can be delivered to the tissue.

Although we applied the workflow to 43 diseases and 547 FDA approved and 126 investigational drugs (clinical trial phase I-III), the flexible design of the workflow allows for it to be run using any disease or drug for which GWAS and transcriptomics data is available. Similarly, other pathway databases could be used in the pathway enrichment step instead of the ones we are proposing. Therefore, we plan to use other datasets in the future such as DSigDB for drug-induced gene expression [84] as well as other pathway databases such as WikiPathways [85]. We also anticipate that incorporating new data modalities such as proteomics and eQTLs could be another prospect for enhancement of the workflow. While we have not considered drug side effects in our current work, integrating side effect information in a future extension could lead to better predictions. Moreover, we purposely restricted our analysis to exclusively approved drugs and those under clinical trial since our study was focused on finding repositioning drug candidates. However, the presented workflow could be applied to all LINCS drug perturbed gene expression signatures for drug discovery purposes. Running the workflow with novel datasets not only will provide new insights on candidate drugs but also allow to evaluate the reproducibility of the findings presented in this work.

## Conclusions

Here, we have presented *PS4DR*, a reproducible drug repositioning workflow that exploits multimodal datasets to predict drug candidates with the help of pathway knowledge. We have demonstrated how integrating pathway knowledge with transcriptomics and GWAS data can elucidate a drug’s mode of action in a disease condition as well as identify potential new applications for a drug. Our workflow predicted numerous drug candidates for several diseases which were validated with epidemiological evidence extracted from the literature and clinical trials. In addition, the modular design of the workflow enables investigators to choose any dataset from proprietary or public databases which suit their experimental needs. While the increased amount and dimensionality of personalized health data are improving health care, we hope our systematic approach to integrate contextual knowledge with data will pave the way towards mechanism-based drug repositioning in precision medicine research.

## Methods

Previous work from Ferrero and Agarwal [27] demonstrated that genes associated with a disease have a tendency to be differentially expressed both in a disease and drug context. Following their hypothesis, we propose a new workflow, *PS4DR*, that can exploit transcriptomics and GWAS data together with pathway knowledge to predict the drugs that best revert the pathway dysregulations observed in a given pathophysiological context. We compared the results generated using the PS4DR workflow with the drug-disease associations presented by Ferrero and Agarwal [27]. These results can be found in Additional file 1: Text Section 3.

In the following subsections, we describe our modular and flexible workflow (Fig. 1). We begin by introducing the different data modalities (e.g., GWAS, gene expression signatures, etc.) and the resources used in the workflow in the application scenario, followed by the data preprocessing steps. Finally, we discuss in detail the different components of the workflow, its implementation, and how it can be adapted to other software tools.

### Data modalities

*PS4DR* uses two different data modalities: GWAS and transcriptomics data. This section describes the datasets used for each modality for the case scenario. While we used various publicly available datasets as described below, users can use any other public or proprietary datasets of their preference in the workflow.

#### *GWAS data*

We have collected genetic association data from different publicly available GWAS datasets (i.e., GWASdb, GWAS catalog, GRASP, and PheWAS). We integrated these datasets by using the Systematic Target OPportunity assessment by Genetic Association Predictions (STOPGAP) [86] analysis pipeline that enables merging different GWAS datasets and calculating their linkage disequilibrium (LD) to capture a wider spectrum of relevant genetic signals. While STOPGAP offers already processed datasets, we have used the pipeline in our workflow to process the most recent datasets from the above-mentioned sources. All the data processed with STOPGAP were downloaded on 2nd March 2019.

#### *Gene expression data*

We have used two different sources i.e., (i) LINCS and (ii) Open Targets to collect gene expression datasets for drug perturbations and diseases in our workflow, respectively. The LINCS dataset is a collection of gene expression signatures obtained by exposing cells to a wide variety of known and novel perturbing agents following the L1000 assay. This dataset was retrieved from the Harmonizome database [87] since it provides an already processed version of the original datasets with more convenient attribute tables that define significant associations between genes and attributes such as cell lines, drugs, and dose information. Furthermore, we made use of Open Targets, a platform that brings together multiple data types by comprehensive and robust data integration from many public databases. It has been widely used for investigations on target identification and prioritization. We have retrieved gene expression signatures data for different diseases using the Open Target's RESTful API on the 5th of March, 2019. Finally, to demonstrate the scalability of PS4DR, we provide the source code to run the workflow with CREEDS [88], an analogous dataset to the two used as case scenarios in the manuscript (https://github.com/ps4dr/ps4dr/tree/master/data/creeds).

### Data preprocessing

Since the workflow utilizes a large number of datasets coming from multiple resources in the two data modalities (i.e., genome-wide association data and gene expression signatures) used in the workflow, a series of preprocessing steps were required to harmonize the data to make them interoperable (Fig. 3).

We harmonized Medical Subject Headings (MeSH) [89] concepts used in GWAS studies to facilitate interoperability with the DEG data from Open Targets that exclusively uses the Experimental Factor Ontology (EFO [90]) to catalog disorders. Similarly, we used Ensembl identifiers as the overarching nomenclature that harmonizes all different gene identifiers (e.g., HGNC, Entrez Gene, etc.) in the multiple datasets. The mappings from MeSH to EFO terms were performed using the EFO ontology (version: 2.105). The conversion from different gene identifiers to Ensembl IDs was conducted with the Ensembl release 97 with the biomaRt R package [91]. Finally, LINCS compound identifiers were mapped to PubChem compound identifiers using the mapping table provided by the Ma'ayan Laboratory (http://amp.pharm.mssm.edu/static/hdfs/harmonizome/data/lincscmapchemical/gene_attribute_edges.txt.gz) and then from PubChem compound identifiers to ChEMBL identifiers using UniChem's RESTful API [92].

These preprocessing steps enabled us to retrieve a total of 174,648 associations between 17,959 genes in 613 diseases from GWAS data. We have used EFO identifiers of these 616 diseases to retrieve their corresponding gene expression signatures in Open Targets using its API. Finally, DEG signatures were fetched for 183 diseases with 259,594 associations between 23,998 genes. Moreover, we also retrieved 17,074 associations between 1060 diseases and 2103 drugs from Open Targets which were at least in clinical trial phase I. Finally, we obtained 1,427,757 associations between 8107 genes and 2700 perturbing agents from the LINCS dataset.

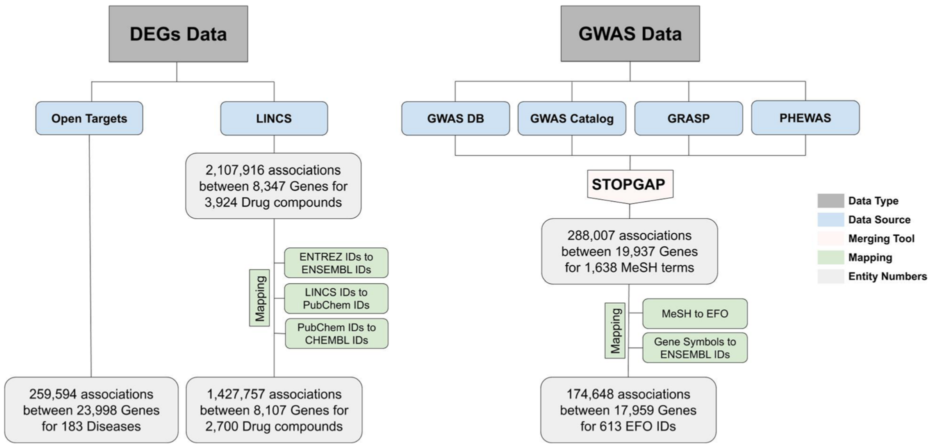

DEGs Data
GWAS Data
Open Targets
LINCS
GWAS DB
GWAS Catalog
GRASP
PHEWAS
2,107,916 associations between 8,347 Genes for 3,924 Drug compounds
STOPGAP
288,007 associations between 19,937 Genes for 1,638 MeSH terms
Data Type
Data Source
Merging Tool
Mapping
Entity Numbers
Mapping
ENTREZ IDs to ENSEMBL IDs
LINCS IDs to PubChem IDs
PubChem IDs to CHEMBL IDs
Mapping
MeSH to EFO
Gene Symbols to ENSEMBL IDs
259,594 associations between 23,998 Genes for 183 Diseases
1,427,757 associations between 8,107 Genes for 2,700 Drug compounds
174,648 associations between 17,959 Genes for 613 EFO IDs

**Fig. 3** Data preprocessing workflow. This workflow describes the preprocessing of gene expression signatures (left side) and GWAS data (right side) to make them interoperable, as well as the primary and final outcome after the preprocessing. Preprocessing steps include multiple intermediary mappings to get common identifiers for Genes (ENSEMBL identifiers), chemicals (ChEMBL identifiers) and diseases (EFO identifiers)

### Filtering via gene set enrichment

The *PS4DR* workflow contains a series of optional filtering steps that enable identifying the genes in the transcriptomics data that have also been reported in GWAS for the same disease. While this step adds the disease context [27] to the gene expression signatures, we leave the possibility for users to omit this step and directly proceed to the pathway enrichment analysis step. Following, we describe each of the filtering steps that are based on calculating the significance of the overlap between the gene sets of the transcriptomics and GWAS data using Fisher's Exact test.

#### *Disease gene expression signatures and GWAS data*

This filtering step is based on calculating the significance of the overlap between gene sets from disease gene expression signatures and GWAS data for each disease pair using Fisher's Exact test. To adjust for multiple testing, $p$-values were corrected with the Benjamini-Hochberg correction [93], and gene sets with a corrected $p$-value above 0.05 were removed. We obtained 26,214 significantly overlapped disease pair gene sets among all the diseases, while 43 of these gene sets originated from the same diseases. These are the 'disease-specific gene sets' from 43 diseases, which are both genetically associated and differentially expressed in the same disease. As previously reported by Ferrero and Agarwal [27], we also observed gene sets from GWAS and transcriptomics data of the same disease are more likely to show a significant overlap compared to gene sets from different diseases (Additional file 1: Fig. S3).

#### *Drug gene expression signatures and GWAS data*

Using the same strategy as the previous step, we filtered drug perturbed gene expression signatures using GWAS data to retain significantly overlapped gene sets. Here, a more stringently adjusted $p$-value threshold of less than or equal to $1e^{-10}$ was used to limit the false positive associations since the drug perturbed data do not have any direct disease context. However, we used additional drug-disease associations retrieved from Open Targets to give disease context, to an extent, to the drug perturbed gene expression signatures. Finally, we obtained 22,551 significantly overlapped gene sets which are genetically associated with a particular disease and also differentially expressed by drug perturbations in the same disease context.

#### *Disease gene expression signatures, drug gene expression signatures, and GWAS data*

The final step involves further filtering of the resulting gene sets of the two previous filtering steps by applying the same strategy. The aim of this final filtering step is to retrieve drug perturbed differentially expressed gene sets in a disease which are also genetically associated with that same disease. In our case scenario, we obtained 14,631 unique drug-disease pairs with significant gene sets ($q$-value $> 0.05$) from all possible drug-disease pairs (total number of pairs). These two gene sets (i.e., disease-specific and drug-specific gene sets) will be used in the next step for each disease to identify the drugs that revert the signatures observed in the disease condition.

### Pathway enrichment analysis

We next use pathway enrichment analysis in each disease to calculate the sign of pathway dysregulation (i.e., up- or down-regulation) in both of the input datasets (i.e., disease-specific gene sets and drug-specific gene sets) using one or multiple pathway databases of reference. By running pathway enrichment analysis, we obtain two vectors, one for each input dataset, indicating the sign of dysregulation for each pathway (i.e., up- or down-regulated and no change). Here, it is important to note that pathway enrichment acts as a dimensionality reduction technique by narrowing down the genetic space (on the scale of thousands) to the pathway space (on the scale of hundreds) (Additional file 1: Text Section 4). Although numerous pathway enrichment methods can be applied to the workflow (e.g., GSEA, Signaling Pathway Impact Analysis (SPIA) [94]), the method applied must ultimately provide the sign of pathway dysregulation since this information will be used in the following step for drug prioritization.

Here, we demonstrate the workflow using one of the most popular topology-based enrichment methods, SPIA, on three pathway databases (i.e., KEGG [95]; Reactome [96]; and Biocarta [97]). Since SPIA requires the pathway input files in a specific binary matrix format, we have used two different tools to prepare pathway datasets for SPIA input. The SPIA package already provides a function to prepare the pathway input file for KEGG's KGML files. Therefore, we have downloaded the latest KGML files from KEGG's ftp site on 27 June 2019 and used the SPIA function 'makeSPIAdata' to convert them to the SPIA required input format. However, this function only works with the KGML file format, which is a modified XML used by KEGG. Therefore, we used *graphite* (v 1.30.0 - release 2019-04-17) [98] to create additional pathway input files for SPIA calculations. First, we retrieved the Reactome and Biocarta pathway files by using the graphite function 'pathways' and then we prepared SPIA input files of these two databases by using another function, 'prepareSPIA'. Both these data sets were time-stamped with 2019-04-17. However, as previously mentioned, the workflow could be adapted to employ other pathway enrichment analysis methods such as GSEA (Additional file 1: Text Section 2). First, we performed SPIA on 43 'disease-specific gene sets' in order to evaluate signed pathway dysregulation in a disease context. Next, we conducted SPIA for 'drug-specific gene sets in disease' which gives signed pathway dysregulation for all available approved drugs and those under clinical trial in each of 43 diseases. Moreover, to evaluate whether SPIA results can be statistically significant, we performed SPIA with the simulated pathways created using the genes from KEGG, Reactome, and Biocarta. The results of SPIA from these randomly simulated pathway constructs rarely yielded significant up- or down-regulated pathways for any of the diseases we tested; thus, this confirms that true pathways are biologically meaningful (Fig. 4).

### Drug prioritization: correlation score

The final part of the workflow uses the results of pathway enrichment methods to prioritize drugs based on how well they can counteract the overall pathway signatures on each disease. First, only the statistically significant pathways ($q$-value $< 0.05$) which are up- or down-regulated in drug and diseases contexts are considered. Next, to facilitate calculating the correlation scores, each affected pathway is assigned with + 1 or − 1 depending on whether it is up- or down-regulated, respectively. Finally, Pearson's

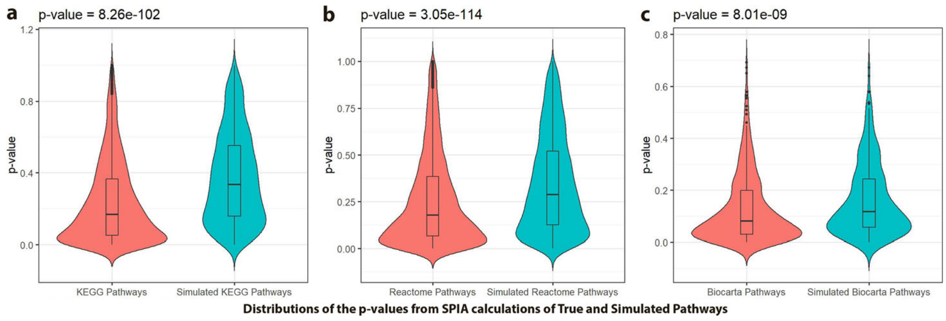


**Fig. 4** Distributions of the *p*-values resulting from SPIA true and simulated pathways represented as violin plots for a) KEGG, b) Reactome, and c) Biocarta pathway databases. Mann-Whitney U test confirmed that the distributions are significantly different for all three pathway databases (KEGG: *p*-value = 8.26e$^{-102,}$ Reactome: *p*-value = 3.05e$^{-114}$, Biocarta: *p*-value = 8.01e$^{-09}$). These results demonstrate that while true pathways yield meaningful results (i.e., lower *p*-values), simulated pathways are rarely significantly enriched

correlation coefficient is calculated using the drug pathway signature vectors against the disease pathway signature vectors. This step results in a list of 26 diseases, while some of the diseases did not have any drugs with a correlation score as the standard deviation was zero for both vectors. Alternatively, Levenshtein distance [99] was also used to calculate the dissimilarity score between the drug and disease pathway signature vectors. We selected arbitrary thresholds for correlation scores (i.e., less than or equal to − 0.4) and affected pathways (i.e., greater than or equal to 50%) to reduce the

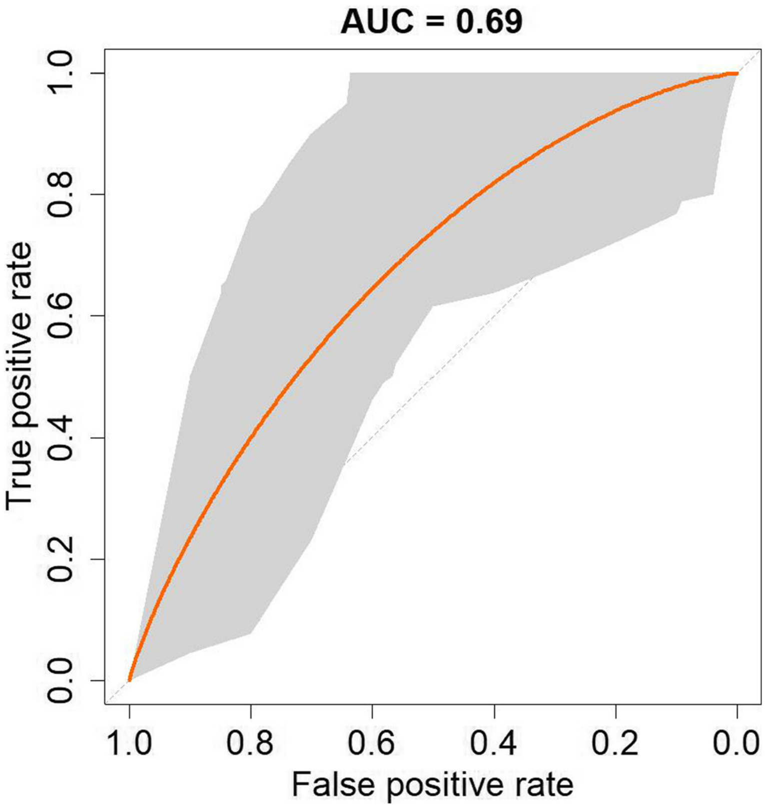


**Fig. 5** ROC curve of *PS4DR* predicted drugs. ROC curve with 95% confidence interval obtained using existing clinical trials for predicted drugs as positive labels and correlation scores as the ranking metric

number of drug candidates in each disease for further manual investigation. However, users can decide the threshold according to their preferences. As a validation step, we generated the ROC curve (Fig. 5) for the predicted drug-disease associations by using the correlation scores as predictors and their available clinical trial evidence as labels. The resulting AUC of 0.69 demonstrates that *PS4DR* can prioritize several drugs for given diseases that are already on clinical trials. While we achieved a slightly higher AUC-ROC than Ferrero et al. (AUC-ROC = 0.64), we must note some subtle methodological differences. First, we used a dataset that is 2 years newer than Ferrero et al. (2019 versus 2017). Second, we used anti-correlation scores as the predictor instead of adjusted *p*-values from Fisher's test for significantly overlapped genesets. Third, we used the same methodology to calculate the AUC, but because of our prioritization, had a smaller number of drug-disease pairs. This was reflected in our wider confidence intervals (0.59–0.82).

### Software and code

R 3.5.1 was used for all data processing and analysis. All code is publicly available at https://github.com/ps4dr/ps4dr under the Apache 2.0 License. Dependencies of the modules used by the workflow and their specific versions are outlined in the repository. Furthermore, we packaged the workflow into a single shell script that can run all the steps with a single command, thus, enabling the reproducibility of the results in the future. Finally, the README file includes an introduction and a tutorial on how to use *PS4DR* and how to add or modify modules within the workflow.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s12859-020-03568-5.

**Additional file 1.** This text file contains all supplementary text, tables and figures referenced in the manuscript.

### Abbreviations

AMPK: AMP-activated protein kinase; CKIε: Casein Kinase ε; CMap: Connectivity Map; CREEDS: CRowd Extracted Expression of Differential Signatures; EFO: Experimental Factor Ontology; FDA: Food and Drug Administration; GSEA: Gene Set Enrichment Analysis; GWAS: Genome-Wide Association Study; LD: Linkage Disequilibrium; LINCS: Library of Integrated Network-Based Cellular Signatures; MeSH: Medical Subject Headings; MoA: Mechanisms of Action; PGC–SCZ: Psychiatric Genomics Consortium–Schizophrenia Workgroup; SLE: Systemic Lupus Erythematosus; SNP: Single nucleotide polymorphism; SPIA: Signaling Pathway Impact Analysis; VEGFR: Vascular Endothelial Growth Factor Receptor; XIAP: X-linked inhibitor of apoptosis protein

### Acknowledgments

We would like to thank Dr. Ludwig Geistlinger for his helpful suggestions regarding the application of SPIA in our workflow and the reviewers of the manuscript for their helpful comments.

### Availability and requirements

Project name: *PS4DR*.
Project home page: https://github.com/ps4dr
Operating system(s): Linux.
Programming language: R.
Other Requirements: R 3.5.1.
License: Apache License 2.0.
Any restrictions to use by non-academics: None.

### Authors' contributions

MAE and DDF conceived and designed the study. MAE implemented the scripts and conducted the application scenario. MAE, DDF, and CTH wrote the paper. MHA supervised the work and acquired the funding. All authors have read and approved the final manuscript.

### Funding

This work was supported by the Fraunhofer Society under the MAVO project, the Human Brain Pharmacome (https://pharmacome.scai.fraunhofer.de). This work was also supported by the EU/EFPIA Innovative Medicines Initiative Joint Undertaking under AETIONOMY [grant number 115568], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies in-kind contribution. This funding body did not play a role in the design of the study and collection, analysis, and interpretation of data, or in writing the manuscript.

### Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the *PS4DR*'s GitHub repository, [https://github.com/ps4dr/ps4dr]. The datasets generated and/or analyzed during the current study are publicly available at https://github.com/ps4dr/results.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (Fraunhofer SCAI), 53757 Sankt Augustin, Germany. [2]Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, 53117 Bonn, Germany.

Received: 30 July 2019 Accepted: 28 May 2020
Published online: 05 June 2020

### References

1. Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. Nat Rev Drug Discov. 2004;3(5):417. https://doi.org/10.1038/nrd1382.
2. Waring MJ, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov. 2015;14(7):475. https://doi.org/10.1038/nrd4609.
3. Li J, et al. A survey of current trends in computational drug repositioning. Briefings Bioinformatics. 2015;17(1):2–12. https://doi.org/10.1093/bib/bbv020.
4. Lamb J, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;*313*(5795):1929–35. https://doi.org/10.1126/science.1132939.
5. Duan Q, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res. 2014;42(W1):W449–60. https://doi.org/10.1093/nar/gku476.
6. Tanoli Z. Exploration of databases and methods supporting drug repurposing: a comprehensive survey. Briefings Bioinformatics. 2020;bbaa003. https://doi.org/10.1093/bib/bbaa003.
7. Sanseau P, et al. Use of genome-wide association studies for drug repositioning. Nat Biotechnol. 2012;30(4):317. https://doi.org/10.1038/nbt.2151.
8. Luo H, et al. DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical–protein interactome. Nucleic Acids Res. 2011;39(suppl_2):W492–8. https://doi.org/10.1093/nar/gkr299.
9. Lee HS, et al. Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. BMC Syst Biol. 2012;6(1):80. https://doi.org/10.1186/1752-0509-6-80.
10. Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. PloS One. 2011;6(12):e28025. https://doi.org/10.1371/journal.pone.0028025.
11. Dudley JT, et al. Exploiting drug–disease relationships for computational drug repositioning. Briefings Bioinformatics. 2011;12(4):303–11. https://doi.org/10.1093/bib/bbr013.
12. Cookson W, et al. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009;10(3):184. https://doi.org/10.1038/448645a.
13. Emilsson V, et al. Genetics of gene expression and its effect on disease. Nature. 2008;452(7186):423. https://doi.org/10.1038/nature06758.
14. Schadt EE, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet. 2005;37(7):710. https://doi.org/10.1038/ng1589.
15. Iorio F, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Nat Acad Sci. 2010;107(33):14621–6. https://doi.org/10.1073/pnas.1000138107.
16. Sirota M, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Translational Med. 2011;3(96):96ra77. https://doi.org/10.1126/scitranslmed.3001318.
17. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Nat Acad Sci. 2005;*102*(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
18. Suthram S, et al. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comp Biol. 2010;6(2):e1000662. https://doi.org/10.1371/journal.pcbi.1000662.
19. Keiser MJ, et al. Predicting new molecular targets for known drugs. Nature. 2009;462(7270):175. https://doi.org/10.1038/nature08506.
20. Welter D, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2013;42(D1):D1001–6. https://doi.org/10.1093/nar/gkt1229.

21. Lencz T, Malhotra AK. Targeting the schizophrenia genome: a fast track strategy from GWAS to clinic. Mol Psychiatry. 2015;20(7):820. https://doi.org/10.1038/mp.2015.28.
22. Ripke S, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511(7510):421. https://doi.org/10.1038/nature13595.
23. Rask-Andersen M, et al. The druggable genome: evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. Ann Rev Pharmacol Toxicol. 2014;54:9–26. https://doi.org/10.1146/annurev-pharmtox-011613-135943.
24. Zhang J, et al. Use of genome-wide association studies for cancer research and drug repositioning. *PloS one*. 2015a; 10(3):e0116477. https://doi.org/10.1371/journal.pone.0116477.
25. So HC, et al. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. Nature Neuroscience. 2017;20(10):1342. https://doi.org/10.1038/nn.4618.
26. Zhang M, et al. Drug repositioning for diabetes based on omics' data mining. PloS One. 2015b;10(5):e0126082. https://doi.org/10.1371/journal.pone.0126082.
27. Ferrero E, Agarwal P. Connecting genetics and gene expression data for target prioritisation and drug repositioning. Biodata Mining. 2018;11(1):7. https://doi.org/10.1186/s13040-018-0171-y.
28. Koscielny G, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 2016; 45(D1):D985–94. https://doi.org/10.1093/nar/gkw1055.
29. Li MJ, et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res. 2011;40(D1):D1047–54. https://doi.org/10.1093/nar/gkr1182.
30. Leslie R, et al. GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics. 2014;30(12):i185–94. https://doi.org/10.1093/bioinformatics/btu273.
31. Denny JC, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnology. 2013;*31*(*12*):1102 10.1038%2Fnbt.2749.
32. Wishart DS, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017;46(D1): D1074–82. https://doi.org/10.1093/nar/gkx1037.
33. Surriga O, et al. Crizotinib, a c-Met inhibitor, prevents metastasis in a metastatic uveal melanoma model. Mol Cancer Ther. 2013;12(12):2817–26. https://doi.org/10.1158/1535-7163.MCT-13-0499.
34. Wiesner T, et al. Kinase fusions are frequent in Spitz tumours and spitzoid melanomas. Nature Communications. 2014;5: 3116. https://doi.org/10.1038/ncomms4116.
35. Turner J, et al. Kinase gene fusions in defined subsets of melanoma. Pigment Cell Melanoma Res. 2017;30(1):53–62. https://doi.org/10.1111/pcmr.12560.
36. Lewis KD, et al. A multi-center phase II evaluation of the small molecule survivin suppressor YM155 in patients with unresectable stage III or IV melanoma. Investigational new drugs. 2011;29(1):161–6 https://doi.org/10.1007/s10637-009-9333-6.
37. Yamanaka K, et al. Antitumor activity of YM155, a selective small-molecule survivin suppressant, alone and in combination with docetaxel in human malignant melanoma models. Clinical Cancer Res. 2011;17(16):5423–31. https://doi.org/10.1158/1078-0432.CCR-10-3410.
38. Amiri KI, et al. Augmenting chemosensitivity of malignant melanoma tumors via proteasome inhibition: implication for bortezomib (VELCADE, PS-341) as a therapeutic agent for malignant melanoma. Cancer Res. 2004;64(14):4912–8. https://doi.org/10.1158/0008-5472.CAN-04-0673.
39. Selimovic D, et al. Bortezomib/proteasome inhibitor triggers both apoptosis and autophagy-dependent pathways in melanoma cells. Cellular Signalling. 2013;25(1):308–18. https://doi.org/10.1016/j.cellsig.2012.10.004.
40. Czyż M, et al. PARP1 inhibitor olaparib (Lynparza) exerts synthetic lethal effect against ligase 4-deficient melanomas. Oncotarget. 2016;7(46):75551 10.18632%2Foncotarget.12270.
41. McNeil EM, et al. The toxicity of nitrofuran compounds on melanoma and neuroblastoma cells is enhanced by Olaparib and ameliorated by melanin pigment. DNA Repair. 2013;12(11):1000–6. https://doi.org/10.1016/j.dnarep.2013.08.017.
42. Friedman AA, et al. Landscape of targeted anti-cancer drug synergies in melanoma identifies a novel BRAF-VEGFR/PDGFR combination treatment. PloS one. 2015;10(10):e0140310. https://doi.org/10.1371/journal.pone.0140310.
43. Gimsing P. Belinostat: a new broad acting antineoplastic histone deacetylase inhibitor. Expert Opin Invest Drugs. 2009; 18(4):501–8. https://doi.org/10.1517/13543780902852560.
44. Gopal YV, et al. Inhibition of mTORC1/2 overcomes resistance to MAPK pathway inhibitors mediated by PGC1α and oxidative phosphorylation in Schmid, P., Forster, M. D., Summers, Y. J., Good, J., Sarker, S. J., Lim, L., … & Middleton, G. W. (2017). A study of vistusertib in combination with selumetinib in patients with advanced cancers: TORCMEK phase Ib results.melanoma. Cancer Res. 2014;74(23):7037–47. https://doi.org/10.1158/0008-5472.CAN-14-1392.
45. Schmid P, et al. A study of vistusertib in combination with selumetinib in patients with advanced cancers: TORCMEK phase Ib results; 2017. https://doi.org/10.1200/JCO.2017.35.15_suppl.2548.
46. Abd-Alhaseeb MM, et al. Olmesartan potentiates the anti-angiogenic effect of sorafenib in mice bearing Ehrlich's ascites carcinoma: role of angiotensin (1–7). PLoS One. 2014;9(1):e85891. https://doi.org/10.1371/journal.pone.0085891.
47. Masamune A, et al. The angiotensin II type I receptor blocker olmesartan inhibits the growth of pancreatic cancer by targeting stellate cell activities in mice. Scand J Gastroenterology. 2013;48(5):602–9. https://doi.org/10.3109/00365521.2013.777776.
48. Patil SP, et al. Identification of antipsychotic drug fluspirilene as a potential p53-MDM2 inhibitor: a combined computational and experimental study. J Computer-Aided Mol Design. 2015;29(2):155–63. https://doi.org/10.1007/s10822-014-9811-6.
49. Shi XN, et al. In silico identification and in vitro and in vivo validation of anti-psychotic drug fluspirilene as a potential CDK2 inhibitor and a candidate anti-cancer drug. PloS One. 2015;10(7):e0132072. https://doi.org/10.1371/journal.pone.0132072.
50. Yu Q, et al. Requirement for CDK4 kinase function in breast cancer. Cancer Cell. 2006;9(1):23–32. https://doi.org/10.1016/j.ccr.2005.12.012.
51. Nuthalapati S, et al. Preclinical pharmacokinetic and pharmacodynamic evaluation of novel anticancer agents, ON01910. Na (Rigosertib, Estybon™) and ON013105, for brain tumor chemotherapy. Pharmaceutical Res. 2012;29(9):2499–511. https://doi.org/10.1007/s11095-012-0780-y.

52. Hu D, et al. Fenofibrate inhibited pancreatic cancer cells proliferation via activation of p53 mediated by upregulation of LncRNA MEG3. Biochem Biophys Res Commun. 2016;471(2):290–5. https://doi.org/10.1016/j.bbrc.2016.01.169.
53. Criddle DN, et al. Menadione-induced reactive oxygen species generation via redox cycling promotes apoptosis of murine pancreatic acinar cells. J Biol Chemi. 2006;281(52):40485–92. https://doi.org/10.1074/jbc.M607704200.
54. Osada S, et al. The utility of vitamin K3 (menadione) against pancreatic cancer. Anticancer Res. 2008;28(1A):45–50.
55. Zhou T, et al. Fluoxetine synergys with anticancer drugs to overcome multidrug resistance in breast cancer cells. Tumor Biology. 2012;33(5):1299–306. https://doi.org/10.1007/s13277-012-0377-4.
56. Wang-Gillam A, et al. A phase I/II study combining tosedostat with capecitabine in patients with metastatic pancreatic ductal adenocarcinoma (PDAC); 2017. https://doi.org/10.1200/JCO.2017.35.4_suppl.410.
57. Xu PF, et al. PI3Kβ inhibitor AZD6482 exerts antiproliferative activity and induces apoptosis in human glioblastoma cells. Oncology reports. 2019;41(1):125–32 https://doi.org/10.3892/or.2018.6845.
58. Wu ZH, et al. Praziquantel synergistically enhances paclitaxel efficacy to inhibit cancer cell growth. Plos One. 2012;7(12): e51721. https://doi.org/10.1371/journal.pone.0051721.
59. Bové J, et al. Fighting neurodegeneration with rapamycin: mechanistic insights. Nat Rev Neurosci. 2011;12(8):437. https://doi.org/10.1038/nrn3068.
60. Cai Z, Yan LJ. Rapamycin, autophagy, and Alzheimer's disease. J Biochem Pharmacol Res. 2013;1(2):84.
61. Spilman P, et al. Inhibition of mTOR by rapamycin abolishes cognitive deficits and reduces amyloid-β levels in a mouse model of Alzheimer's disease. PloS One. 2010;5(4):e9979. https://doi.org/10.1371/journal.pone.0009979.
62. Kim YD, et al. Pimozide reduces toxic forms of tau in TauC3 mice via 5′ adenosine monophosphate-activated protein kinase-mediated autophagy. J Neurochemistry. 2017;142(5):734–46. https://doi.org/10.1111/jnc.14109.
63. Andérica-Romero AC, et al. The MLN4924 inhibitor exerts a neuroprotective effect against oxidative stress injury via Nrf2 protein accumulation. Redox Biol. 2016;8:341–7. https://doi.org/10.1016/j.redox.2016.02.008.
64. Lonskaya I, et al. Nilotinib-induced autophagic changes increase endogenous parkin level and ubiquitination, leading to amyloid clearance. J Mol Med. 2014;92(4):373–86. https://doi.org/10.1007/s00109-013-1112-3.
65. Scudder SL, Patrick GN. Synaptic structure and function are altered by the neddylation inhibitor MLN4924. Mol Cell Neuroscience. 2015;65:52–7 10.1016%2Fj.mcn.2015.02.010.
66. Clader JW, Wang Y. Muscarinic receptor agonists and antagonists in the treatment of Alzheimer's disease. Curr Pharmaceutical Design. 2005;11(26):3353–61. https://doi.org/10.2174/138161205774370762.
67. Benhamú B, et al. Serotonin 5-HT6 receptor antagonists for the treatment of cognitive deficiency in Alzheimer's disease. J Med Chem. 2014;57(17):7160–81. https://doi.org/10.1021/jm5003952.
68. Lee YJ, et al. Inflammation and Alzheimer's disease. Arch Pharmacal Res. 2010;33(10):1539–56. https://doi.org/10.1007/s12272-010-1006-7.
69. Rubio-Perez JM, Morillas-Ruiz JM. A review: inflammatory process in Alzheimer's disease, role of cytokines. Scientific World J. 2012;2012. https://doi.org/10.1100/2012/756357.
70. Nuutinen, S., and Panula, P. (2010). Histamine in neurotransmission and brain diseases. In Histamine in Inflammation (pp. 95–107). Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-8056-4_10.
71. Passani MB, Blandina P. Histamine receptors in the CNS as targets for therapeutic intervention. Trends Pharmacol Sci. 2011;32(4):242–9. https://doi.org/10.1016/j.tips.2011.01.003.
72. Vohora D, Bhowmik M. Histamine H3 receptor antagonists/inverse agonists on cognitive and motor processes: relevance to Alzheimer's disease, ADHD, schizophrenia, and drug abuse. Front Syst Neuroscience. 2012;6:72. https://doi.org/10.3389/fnsys.2012.00072.
73. Sloka JS, Stefanelli M. The mechanism of action of methylprednisolone in the treatment of multiple sclerosis. Multiple Sclerosis J. 2005;11(4):425–32. https://doi.org/10.1191/1352458505ms1190oa.
74. Wendt MK, et al. Fibroblast growth factor receptor splice variants are stable markers of oncogenic transforming growth factor β1 signaling in metastatic breast cancers. Breast Cancer Res. 2014;16(2):R24. https://doi.org/10.1186/bcr3623.
75. Twelves C, et al. Erlotinib in combination with capecitabine and docetaxel in patients with metastatic breast cancer: a dose-escalation study. Eur J Cancer. 2008;44(3):419–26. https://doi.org/10.1016/j.ejca.2007.12.011.
76. Verkaar F, et al. Inhibition of Wnt/β-catenin signaling by p38 MAP kinase inhibitors is explained by cross-reactivity with casein kinase Iδ/ɛ. Chem Biol. 2011;18(4):485–94. https://doi.org/10.1016/j.chembiol.2011.01.015.
77. Previdi S, et al. Breast Cancer–Derived bone metastasis can be effectively reduced through specific c-MET inhibitor tivantinib (ARQ 197) and shRNA c-MET knockdown. Mol Cancer Ther. 2012;11(1):214–23. https://doi.org/10.1158/1535-7163.MCT-11-0277.
78. Matheson CJ, et al. Targeting WEE1 kinase in cancer. Trends Pharmacol Sci. 2016;37(10):872–81. https://doi.org/10.1016/j.tips.2016.06.006.
79. Ghotra VP, et al. SYK is a candidate kinase target for the treatment of advanced prostate cancer. Cancer Res. 2015;75(1): 230–40. https://doi.org/10.1158/0008-5472.CAN-14-0629.
80. Hong WK, et al. Prevention of second primary tumors with isotretinoin in squamous-cell carcinoma of the head and neck. New Engl J Med. 1990;323(12):795–801. https://doi.org/10.1056/NEJM199009203231205.
81. Zou HY, et al. Sensitivity of selected human tumor models to PF-04217903, a novel selective c-Met kinase inhibitor. Mol Cancer Ther. 2012;11(4):1036–47. https://doi.org/10.1158/1535-7163.MCT-11-0839.
82. Liang M, et al. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM Transactions Computational Biology Bioinformatics. 2014;12(4):928–37. https://doi.org/10.1109/TCBB.2014.2377729.
83. Lan W, et al. Predicting microRNA-disease associations based on improved microRNA and disease similarities. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2018;15(6):1774–82. https://doi.org/10.1109/TCBB.2016.2586190.
84. Yoo M, et al. DSigDB: drug signatures database for gene set analysis. Bioinformatics. 2015;*31*(18):3069–71. https://doi.org/10.1093/bioinformatics/btv313.
85. Slenter DN, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2017;*46*(D1):D661–7. https://doi.org/10.1093/nar/gkx1064.
86. Shen J, et al. STOPGAP: a database for systematic target opportunity assessment by genetic association predictions. Bioinformatics. 2017;33(17):2784–6. https://doi.org/10.1093/bioinformatics/btx274.

87. Rouillard AD, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database. 2016;2016. https://doi.org/10.1093/database/baw100.
88. Wang Z, et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. Nature Commun. 2016;7:12846. https://doi.org/10.1038/ncomms12846.
89. Lipscomb CE. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*. 2000;88(3):265.
90. Malone J, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010;26(8):1112–8. https://doi.org/10.1093/bioinformatics/btq099.
91. Durinck S, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21:3439–40. https://doi.org/10.1093/bioinformatics/bti525.
92. Chambers J, et al. UniChem: a unified chemical structure cross-referencing and identifier tracking system. J Cheminformatics. 2013;5(1):3 https://doi.org/10.1186/1758-2946-5-3.
93. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Series B (Methodological). 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
94. Tarca AL, et al. A novel signaling pathway impact analysis. Bioinformatics. 2008;25(1):75–82. https://doi.org/10.1093/bioinformatics/btn577.
95. Kanehisa M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016;*45*(D1): D353–61. https://doi.org/10.1093/nar/gkw1092.
96. Fabregat A, et al. The Reactome pathway Knowledgebase. Nucleic Acids Res. 2018;*46*(D1):D649–55. https://doi.org/10.1093/nar/gkx1132.
97. Nishimura D. BioCarta.. Biotech Software & Internet Report. Computer Software J Scient. 2001;2(3):117–20. https://doi.org/10.1089/152791601750294344.
98. Sales G, et al. Graphite-a Bioconductor package to convert pathway topology to gene network. BMC Bioinformatics. 2012;13(1):20. https://doi.org/10.1186/1471-2105-13-20.
99. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet Phys Doklady. 1966;10(8): 707–10.

## Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Chapter 3**

# Conclusions

We have presented PS4DR, a flexible workflow that explores the feasibility of multi-omics data integration with prior knowledge for mechanism-based drug repositioning. PS4DR was shown to be capable of incorporating pathway knowledge different canonical pathway databases (i.e., KEGG, Reactome, and Biocarta) with genomics and transcriptomics data from several diseases and drug perturbations to predict approved drugs for several new indications. The workflow demonstrated how omics data and pathway knowledge can complement each other to illustrate a drug's mode of action on the altered biological process of the disease and prioritize the drug candidates based on their impact on combined disease etiology. While PS4DR showcased its ability in mechanism-based drug repositioning with publicly available omics datasets, it can serve as a genomic and transcriptome data harmonization workflow to make them interoperable with each other and enable users to work with any new disease or drug datasets of their interests. Finally, the success of this approach has emphasized the importance of prior knowledge in deducting the data-driven anomaly in the area of biomedical research.

# 4 Clustering of Alzheimer's and Parkinson's Disease Based on Genetic Burden of Shared Molecular Mechanisms

## Introduction

Precision medicine intends to disengage the disease into distinct molecular subgroups that could be targeted individually by separate treatment. Despite the unprecedented growth in biomedical big data, the vision of precision medicine remains highly ambitious in the field of neurodegenerative diseases. While poor disease understanding due to the multifaceted complexity of NDDs leads to the failure of numerous clinical trials and a waste of billions of dollars, a systematic approach to characterize these diseases on a molecular basis could increase the chances of successful treatments. Hence, there is an unmet need for developing tools for mechanism-based stratification of NDD patients based on their shared disease mechanisms. Looking forward to the future, such methods could enable the customized treatments for each patient subgroup based on their unique biological phenomena (i.e., genetic architecture, transcription profile, proteomic measurements, molecular mechanisms, etc). We introduce an artificial intelligence-based workflow to integrate prior mechanistic knowledge of diseases with patient-level data for the joint stratification of AD and PD patients.

# Chapter 4

Reprinted with permission from "Emon M. A. *et al.*. (2020) Clustering of Alzheimer's and Parkinson's Disease Based on Genetic Burden of Shared Molecular Mechanisms. Scientific Reports 10, 19097 (2020)". Copyright © Emon M. A. *et al.*, (2020).

scientific reports

OPEN

# Clustering of Alzheimer's and Parkinson's disease based on genetic burden of shared molecular mechanisms

Mohammad Asif Emon[1,3,7], Ashley Heinson[2,7], Ping Wu[2,7], Daniel Domingo-Fernández[1,3], Meemansa Sood[1,3], Henri Vrooman[4], Jean-Christophe Corvol[5], Phil Scordis[2], Martin Hofmann-Apitius[1,3] & Holger Fröhlich[1,3,6]✉

One of the visions of precision medicine has been to re-define disease taxonomies based on molecular characteristics rather than on phenotypic evidence. However, achieving this goal is highly challenging, specifically in neurology. Our contribution is a machine-learning based joint molecular subtyping of Alzheimer's (AD) and Parkinson's Disease (PD), based on the genetic burden of 15 molecular mechanisms comprising 27 proteins (e.g. APOE) that have been described in both diseases. We demonstrate that our joint AD/PD clustering using a combination of sparse autoencoders and sparse non-negative matrix factorization is reproducible and can be associated with significant differences of AD and PD patient subgroups on a clinical, pathophysiological and molecular level. Hence, clusters are disease-associated. To our knowledge this work is the first demonstration of a mechanism based stratification in the field of neurodegenerative diseases. Overall, we thus see this work as an important step towards a molecular mechanism-based taxonomy of neurological disorders, which could help in developing better targeted therapies in the future by going beyond classical phenotype based disease definitions.

## Abbreviations

| | |
|---|---|
| A-$\beta$ | Amyloid beta |
| AChE | Acetylcholinesterase |
| AD | Alzheimer's disease |
| ADAS | Alzheimer's disease assessment scale |
| ADNI | Alzheimer's disease neuroimaging initiative |
| AUC | Area under the receiver operating characteristic curve |
| CDRSB | Clinical dementia rating sum of boxes |
| CSF | Cerebrospinal fluid |
| ESS | Epworth sleepiness scale |
| eQTL | Expression quantitative trait loci |
| FAQ | Functional activities questionnaire |
| GO | Gene ontology |
| GWAS | Genome wide association study |
| HADS | Hospital anxiety and depression scale |
| IGP | In-group proportion |
| KEGG | Kyoto encyclopedia of genes and genomes |
| MCI | Mild cognitive impairment |

[1]Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53754 Sankt Augustin, Germany. [2]UCB Pharma (UCB Celltech Ltd.), 208 Bath Road, Slough SL1 3WE, Berkshire, UK. [3]Bonn-Aachen International Center for IT, University of Bonn, Endenicher Allee 19c, 53115 Bonn, Germany. [4]Department of Radiology and Nuclear Medicine, Department of Medical Informatics, Erasmus MC, University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands. [5]ICM - Hôpital Pitié Salpêtrière, 47, bd de l'hôpital, 75013 Paris, France. [6]UCB Pharma (UCB Biosciences GmbH), Alfred-Nobel-Str. 10, 40789 Monheim, Germany. [7]These authors contributed equally: Mohammad Asif Emon, Ashley Heinson and Ping Wu. ✉email: holger.froehlich@scai.fraunhofer.de

| | |
|---|---|
| MMSE | Mini-mental state examination |
| MOCA | Montreal Cognitive Assessment |
| MRI | Magnetic resonance imaging |
| NMF | Non-negative matrix factorization |
| PPMI | Parkinson's progression markers initiative |
| PET | Positron emission tomography |
| PD | Parkinson's disease |
| RBD | Rapid eye movement sleep behavior disorder |
| SNP | Single nucleotide polymorphism |
| UPDRS | Unified Parkinson's disease rating scale |

Many neurological disorders are highly multifaceted, heterogeneous and difficult to treat. The high percentages of clinical trial failures in Alzheimer's Disease (AD) exemplify the unmet clinical need in the field: While Open Targets today lists 140 compounds that have been tested in clinical studies so far[1], there are currently only 4 approved ones for symptomatic treatment on the market[2]. The majority of clinical trial failures in neurology (like in other disease areas) can be attributed to a lack of efficacy, and one of the contributing factors is the selection of the wrong target population[3].

Precision medicine brings the hope of disentangling diseases into separate molecular subgroups, which could be therapeutically targeted more specifically, hence increasing the chances of successful treatment. Moreover, these molecular subtypes may be associated with particular mechanisms, which might allow the identification of novel treatment opportunities. The far-reaching vision is an entirely molecular defined taxonomy of neurological disorders, which should be seen in contrast to the traditional and purely phenotypic way, in which neurological diseases have been defined since the nineteenth century[4,5].

The AETIONOMY project funded within the Innovative Medicines Initiative (IMI) of the European Union has taken a step into this direction (www.aetionomy.eu). While focusing on Alzheimer's and Parkinson's Disease (PD) as important examples, the goal of AETIONOMY was to identify and validate molecular characteristics that could help to stratify AD and PD into more homogeneous patient subgroups. Both neurodegenerative diseases share common properties, such as neuroinflammation[6], aberrant miRNA expression[7], and protein misfolding[8]. Accordingly, it has been suspected for a long time that, despite largely non-overlapping causal genetic variants in genome-wide association studies (GWAS), similarities may be expected at the functional or molecular mechanism level[9–11]. Hence, some authors have suggested to focus the analysis on functional categories rather than on individual genetic variants[10].

The existence of commonly impaired biological processes or mechanisms is also potentially attractive from a therapeutic point of view, since it might open the perspective for a more causal disease treatment. However, the question arises, how homogenous AD and PD patient groups might be with respect to those shared mechanisms, i.e. whether there exist subgroups.

In this work, we explored the genetic burden by single nucleotide polymorphisms (SNPs) on genes that in the literature have been described to play a role in both diseases. We found that, based on aggregate SNP burden scores of common molecular mechanisms in AD and PD, unsupervised machine learning methods can identify distinct and reproducible joint subgroups. We show that these clusters can be associated with distinct clinical, pathophysiological and molecular features on a biological-processes and pathway level, and we investigate the potential clinical utility of these differences by prioritizing drug targets for specific patient subgroups. Altogether, this work shows the possibility of effectively using knowledge about disease mechanisms in combination with modern machine learning techniques to unravel molecular subtypes of AD and PD, which may in the future aid the development of better targeted therapies by contributing to a molecular mechanism-based definition of neurodegenerative diseases.

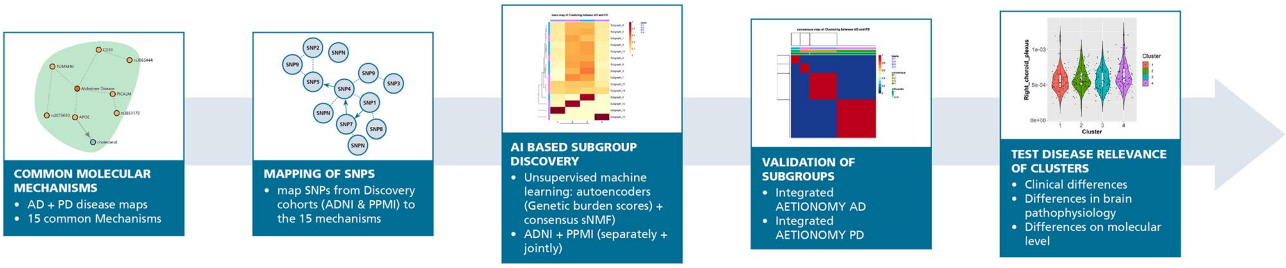
## Results

### Strategy for identifying mechanism based AD/PD subtypes.

Before going into more detail, we briefly outline our general approach for identifying subtypes of sporadic AD and PD idiopathic patients (Fig. 1): Following Tan et al.[12] it is largely driven by the idea of a genetic sub-classification followed by a clinical, imaging based and biological characterization of patients in each cluster to test disease relevance.

Genetic commonalities between AD and PD can only be expected at the biological function level. Hence, the starting point of our work was a comprehensive mapping of the molecular disease landscape of AD and PD based on the scientific literature (see "Methods" section). The result was a set of 15 molecular mechanisms comprising 27 proteins that have been implicated in both diseases (Fig. 2). We mapped 148 SNPs to these genes based on proximity as well as eQTL analysis, see details in Supplements. Using ADNI and PPMI as discovery cohorts (see descriptions in "Methods" section), we calculated for each of the 15 molecular mechanism an aggregate burden score via sparse autoencoders and then used sparse non-negative matrix factorization to identify 4 distinct patient subgroups in AD and PD, see "Methods" section. These subgroups were found independently in both diseases as well as in a merger of ADNI and PPMI patients (Fig. 3A–C).

As a next step, we validated the existence of the identified mixed AD/PD subgroups with the help of disease patients in our integrated AETIONOMY AD and PD cohorts (see description in "Methods" section and Fig. 3D–F). Finally, we tested the disease relevance of the patient subgroups by statistically analyzing the differences of clinical and brain imaging related features as well as transcriptome and methylome profiles in AD and PD patients. Following this high-level overview about our strategy, we will now describe each of the main analysis steps in more detail.



**COMMON MOLECULAR MECHANISMS**

- AD + PD disease maps
- 15 common Mechanisms

**MAPPING OF SNPS**

- map SNPs from Discovery cohorts (ADNI & PPMI) to the 15 mechanisms

**AI BASED SUBGROUP DISCOVERY**

- Unsupervised machine learning: autoencoders (Genetic burden scores) + consensus sNMF)
- ADNI + PPMI (separately + jointly)

**VALIDATION OF SUBGROUPS**

- Integrated AETIONOMY AD
- Integrated AETIONOMY PD

**TEST DISEASE RELEVANCE OF CLUSTERS**

- Clinical differences
- Differences in brain pathophysiology
- Differences on molecular level

**Figure 1.** Strategy for identifying AD + PD subtypes.

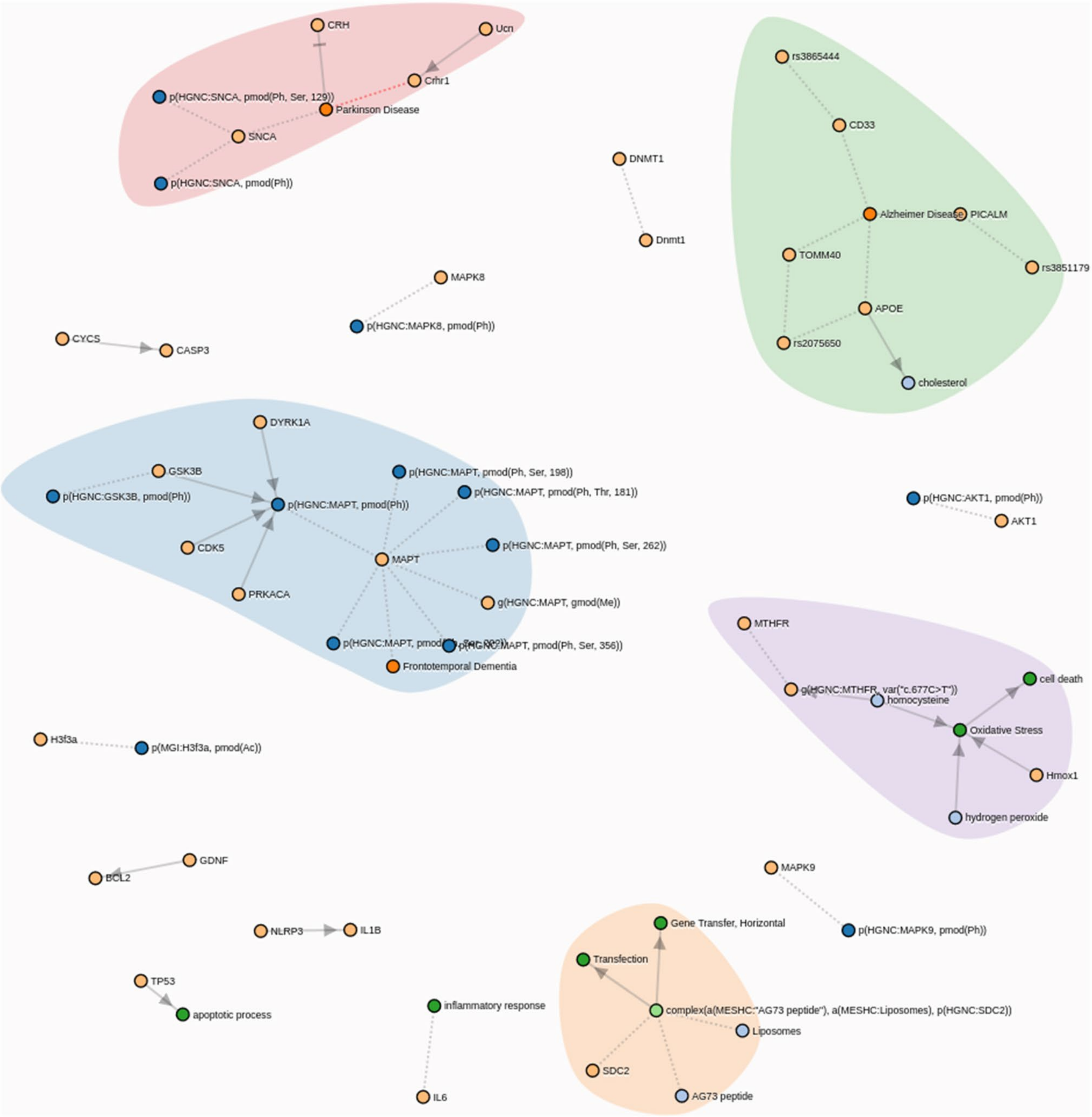


**Figure 2.** Common AD/PD disease mechanisms, see also https://clus2bio.scai.fraunhofer.de/mechanisms.

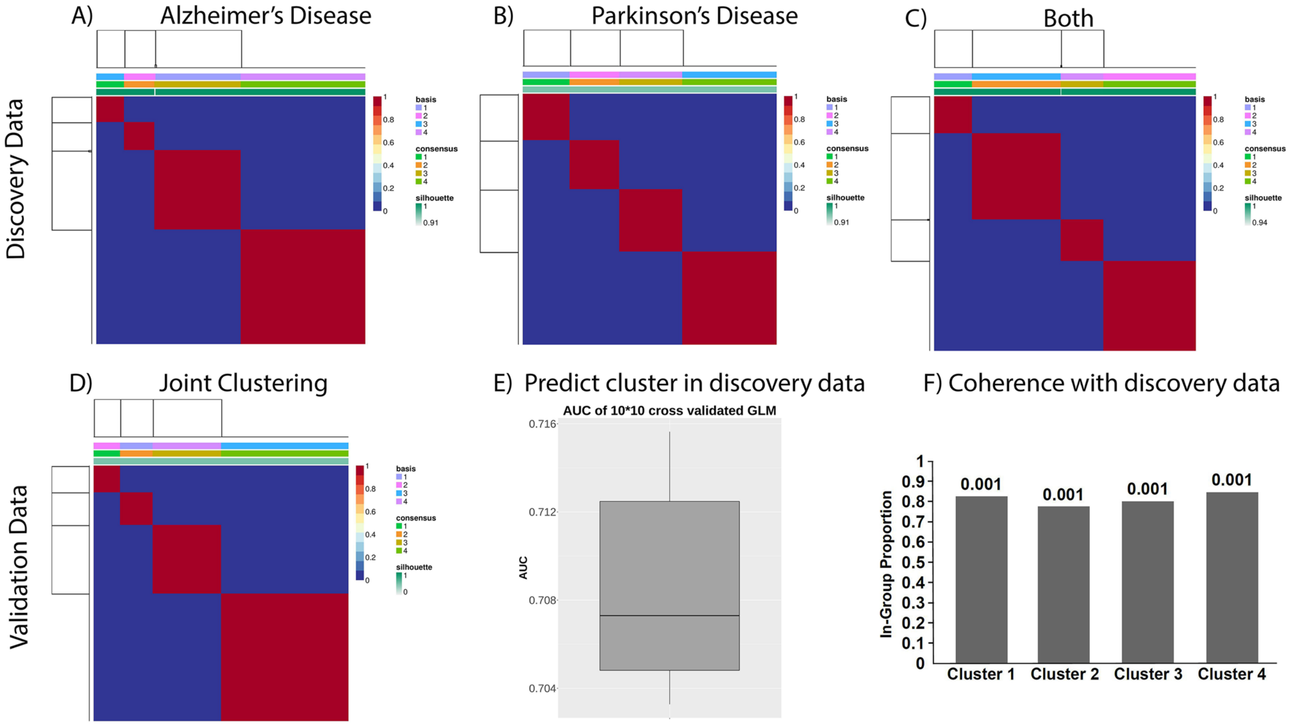


**Figure 3.** Identification of mechanism-induced subtypes in AD and PD: (**A**) Consensus clustering of ADNI AD patients (consensus matrix). (**B**) Consensus clustering of PPMI PD patients (consensus matrix). (**C**) Consensus clustering of merged ADNI + PPMI (consensus matrix). (**D**) Consensus clustering of merged validation data (integrated AETIONOMY AD + PD, consensus matrix). (**E**) Prediction performance of a classifier that allows assigning each patient in a validation cohort to a cluster in the discovery cohort. (**F**) Coherence of joint AD + PD clustering with validation cohort: Shown is the in-group proportion measure and its p-value according to a permutation test.

**Mechanism burden scores allow for reproducible subtyping of AD and PD patients.** Using the data of 148 SNPs mapping to 15 common AD/PD disease mechanisms in 486 AD and 358 PD patients within our discovery cohorts, we developed an unsupervised machine learning approach to discover subgroups (see details in "Methods" section and Supplementary Text p. 13). This approach consisted of two basic steps: (i) sparse autoencoding of the SNPs mapping to each of the 15 mechanisms, resulting into a profile of genetic burden scores; (ii) consensus sparse non-negative matrix factorization to cluster patients and for identifying most discriminative mechanisms. Our method resulted in 4 subgroups in ADNI, PPMI as well as in a merger of ADNI and PPMI patients that were statistically stable and better discriminated than expected by pure chance (Fig. 3A-C, Tables S2–S4); details are described in the "Methods" section and in the Supplementary Text (p. 28). Interestingly, clusters found in the merged AD/PD cohort were all composed of a mixture of AD and PD patients (Figure S22). They were not identical to the ones identified in each disease individually, but showed a highly significant overlap in both cases ($p < 1E{-}16$, $\chi 2$-test). That means our clustering suggests the existence of certain commonalities between AD and PD patients on the level of SNP burden on specific mechanisms. We will discuss the question of disease relevance later.

Due to the particular properties of our employed clustering approach, each of the clusters can be linked back to a particular set of disease mechanisms (Figure S21, Table S1 and https://clus2bio.scai.fraunhofer.de/mechanisms for an interactive view): Cluster 1 reflects the genetic burden on AKT1. AKT1 phosphorylation regulates multiple signaling cascades that are of relevance in both AD and PD[13–15].

Cluster 2 is—among other features—strongly associated with the genetic burden on IL1B, NLRP3, TP53[16–19]. Activation of IL1B by NLRP3 and TP53 play a role in the response of the immune system. Neuroinflammation is a common feature of AD and PD[6].

One of the features of cluster 3 is the genetic burden on MTHFR, which is implicated in hydrogen peroxide and homocysteine regulation as well as cell death and oxidative stress[20], Genetic variants may contribute to the risk of PD[21] and late-onset of AD[22,23].

Cluster 4 reflects the genetic burden on MAPK9, which is implicated in multiple signaling cascades in both diseases[24,25].

Again, these are only examples of representative mechanisms for each cluster. A complete overview can be found in Table S1 and under https://clus2bio.scai.fraunhofer.de/mechanisms.

Our next steps particularly focused on the validation of the existence of the joint AD/PD subgroups. For this purpose, we made use of a merger of our integrated AETIONOMY AD and PD validation cohorts, and we asked two essential questions:

| Cohort | Age | Gender (m/f) | Education | MMSE | CDRSB | ADAS11 | ADAS13 | RAVLT immediate | RAVLT learning |
|---|---|---|---|---|---|---|---|---|---|
| ADNI | 75.55 (9.6) | 288/198 | 16 (5) | 24(3) | 4.5 (2) | 17 (8.33) | 27.67 (10) | 23 (9) | 2 (2) |
| ROSMAP | 87.98 (6.19) | 52/142 | 16 (5) | NA | NA | NA | NA | NA | NA |
| IDIBAPS | 62.42 (9.75) | 15/28 | 12 (9) | 18 (7.5) | NA | NA | NA | NA | NA |

**Table 1.** Demographic and clinical variable summary of AD discovery (ADNI) and validation (ROSMAP and IDIBAPS) cohorts. The data shows only clinically diagnosed sporadic AD cases. The Table shows the median of each variable and the inter-quartile range (IQR) in brackets.

| Cohort | Age | Gender (m/f) | UPDRS1 | UPDRS2 | UPDRS3 off | UPDRS3 on | MOCA | HADS anxiety | Schwab-England |
|---|---|---|---|---|---|---|---|---|---|
| PPMI | 62.5 (14.14) | 238/120 | 5 (5) | 5 (6) | 19 (11.75) | NA | 28 (3) | NA | NA |
| AETIONOMY PD | 64 (11.25) | 59/29 | 7 (6.75) | 7 (8.75) | 28 (17) | 0 (0) | 27 (4) | 5 (5) | 2 (0) |
| DIGPD | 61 (12) | 100/73 | 7 (6) | 5 (5) | 21 (11) | 0 (0) | NA | 6 (4) | 2 (1) |
| ICEBERG | 67 (16.5) | 30/12 | 10 (4) | 7 (6) | 29.5 (11.5) | 0 (0) | 27 (3) | 7 (3) | 2 (0) |

**Table 2.** Demographic and clinical variable summary of PD discovery (PPMI) and validation (AETIONOMY PD, DIGIPD, ICEBERG) cohorts. The data shows only de novo diagnosed idiopathic PD cases. The Table shows the median of each variable and the inter-quartile range (IQR) in brackets.

1. Does an independent clustering of patients in the validation data re-suggest the same number of clusters?
2. Given the panel of 148 SNPs, can we put patients from our validation cohorts into the same clusters that we had previously identified based on our discovery cohorts, and is the correspondingly induced stratification of patients in the validation cohorts coherent with the clustering of patients in the discovery data?

To answer the first question, we re-ran our developed unsupervised machine learning approach (consisting of sparse autoencoding of each of the 15 molecular mechanisms followed by consensus sparse non-negative matrix factorization), which again supported the existence of 4 clusters composed of mixture of AD and PD patients in the merged validation data (Fig. 3D, Table S5, Figure S23).

To answer question two, we first developed a predictive machine learning algorithm, which allowed us to assign any patient in a validation cohort to one of the established clusters (see "Methods" section). Cross-validation based evaluation of the prediction performance of this classifier was conducted and indicated a decent area under receiver operator characteristic curve (AUC) of ~ 70% that was significantly higher than chance level (Fig. 3E), i.e. clusters were predictable.

Secondly, we measured the coherence of the predicted stratification of patients in our validation cohorts with the one identified in our discovery cohorts. This was done by counting the fraction of patients in the validation cohort whose closest patient in the discovery cohort had the same label, yielding the In-Group Proportion (IGP) measure suggested by Kapp and Tibshirani[26], see "Methods" section for details. Accordingly, we could verify a high and statistically significant agreement of clusters predicted for patients in the validation data with those in the merged discovery cohort (Fig. 3F). Overall, we thus concluded that our discovered joint stratification of AD and PD patients was reproducible.

**Comparison of clinical outcome measures between clusters.** Our next steps focused on the question whether our identified patient clusters were disease associated or just reflecting general genetic differences in the population. For this purpose, we used clinical, imaging, transcriptome and methylome data.

We first investigated differences in clinical outcome measures of AD and PD patients across clusters. This was done separately on the basis of each of the individual study used in this paper (AD: ADNI, ROSMAP; PD: PPMI, AETIONOMY PD, ICEBERG, DIGPD), because available clinical data differs between studies (Tables 1, 2), and differences in inclusion/exclusion criteria may bias a combined analysis: Despite the fact that all patients had a time till initial diagnosis of at most 2 years there were significant differences of baseline UPDRS scores between PD studies ($p < 1E\text{-}9$ for MDS-UPDRS I, $p = 0.02$ for MDS-UPDRS II, $p < 1E\text{-}5$ for MDS-UPDRS III off treatment score; Kruskal–Wallis test), and in all cases UPDRS total (sum of MDS-UPDRS I + II + III off treatment scores) in PPMI and DIGPD were lower than in AETIONOMY PD and ICEBERG (median UPDRS total in PPMI: 30, DIGPD: 33, AETIONOMY PD: 42, ICEBERG: 47). Similarly, AD cohorts differed significantly by age ($p < 2E-16$, one-way ANOVA), level of education ($p < 0.01$, Kruskal–Wallis test) and MMSE baseline scores ($p < 1E-10$, Kruskal–Wallis test).

Based on these observations we focused on a statistical analysis within each of the AD and PD cohorts separately. Notably, IDIBAPS was excluded at this point due to the very small sample size (only 29 cases). Summary statistics of major demographic and clinical baseline variables of all clusters in AD and PD can be found in Tables S7 and S8. Within ADNI we compared multiple cognitive assessment scores (CDRSB, ADAS11, ADAS13, MMSE, MOCA, FAQ, RAVLT, and LDELTOTAL) at the visit of first dementia diagnosis (n = 486 patients) across

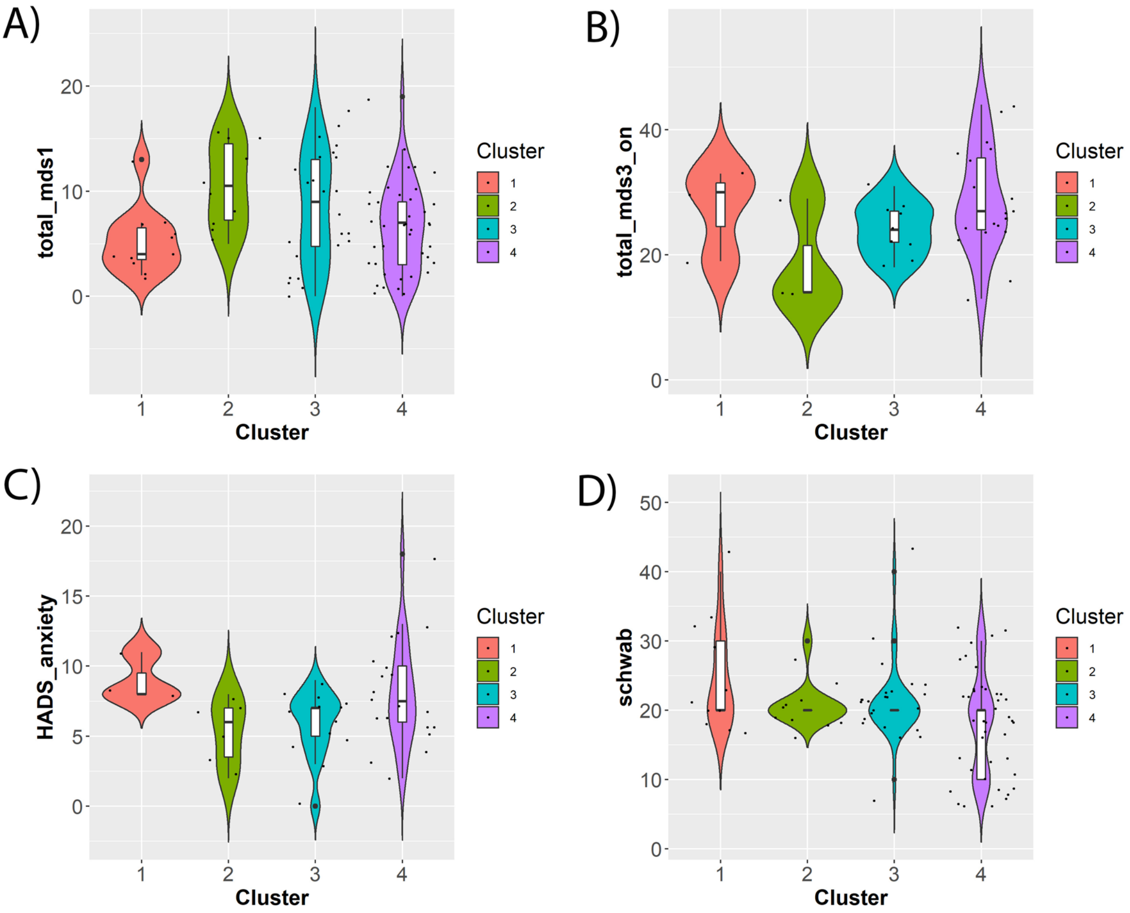


**Figure 4.** Examples of significant differences between clusters with respect to clinical baseline features in PD patients after correction for confounding effects (see "Methods" section). (**A**) MDS-UPDRS I score (AETIONOMY PD); (**B**) MDS-UPDRS III on treatment score (ICEBERG); (**C**) HADS anxiety score (ICEBERG); (**D**) Schwab-England Scale in % (AETIONOMY PD). The Figures shows statistical distributions as violin plots (i.e. boxplots plus kernel density estimates), and individual data points are shown as superimposed dots.

clusters. The provided cognitive tests cover different aspects, such as global cognitive impairment (ADAS11, ADAS13, MMSE, MOCA), logical memory (LDELTOTAL), verbal episodic memory (RAVLT) and activities of daily living (FAQ). For more detailed information about the composition of individual cognition scores we refer to the literature[27–32]. Notably, cluster labels were based on the clustering of the merged ADNI + PPMI and ROSMAP + AETIONOMY PD + ICEBERG + DIGPD cohorts, respectively. Statistical significances were corrected for multiple confounding factors, such as age, gender, time until diagnosis, ethnicity and the use of L-DOPA (the latter for PD patients). Multiple testing correction was applied via the method by Benjamini and Hochberg[33]. Details about the statistical analysis are described in the "Methods" section part of this paper.

According to our analysis, no statistically significant differences of cognitive assessment scores could be found between clusters in AD patients at study baseline (although we notably did observe weakly significant results for working memory cognition assessments in ROSMAP patients). However, as indicated in Fig. 4A–D, PD patients in AETIONOMY PD and ICEBERG demonstrated significant pairwise differences between clusters with respect to several clinical baseline scores, namely MDS-UPDRS I (non-motor aspects of daily living; ICEBERG, AETIONOMY PD), HADS anxiety score (ICEBERG), MDS-UPDRS III (motor examination) on treatment scores (ICEBERG) and Schwab-England Scale (difficulties with activities of daily living; AETIONOMY PD). No significant results were found in PPMI and DIGPD.

In addition to this analysis of baseline variables we also conducted an analysis of follow-up longitudinal data, which was available in ADNI (AD) and PPMI (PD) cohorts. This analysis showed significant differences of the progression of MDS-UPDRS III (motor examination) scores across patient subtypes in PPMI. In ADNI we found significant differences in the progression of global cognitive impairment (ADAS11, ADAS13, CDRSB, MMSE) and verbal episodic memory (RAVLT; see Tables S12, S13).

In summary, clusters are associated with significant differences of clinical disease symptoms and symptom progression of AD and PD patients.

**Association with brain imaging derived features in AD and PD.** In ADNI, AD patients demonstrated highly significant pairwise differences when comparing 193 intracranial volume normalized subcortical

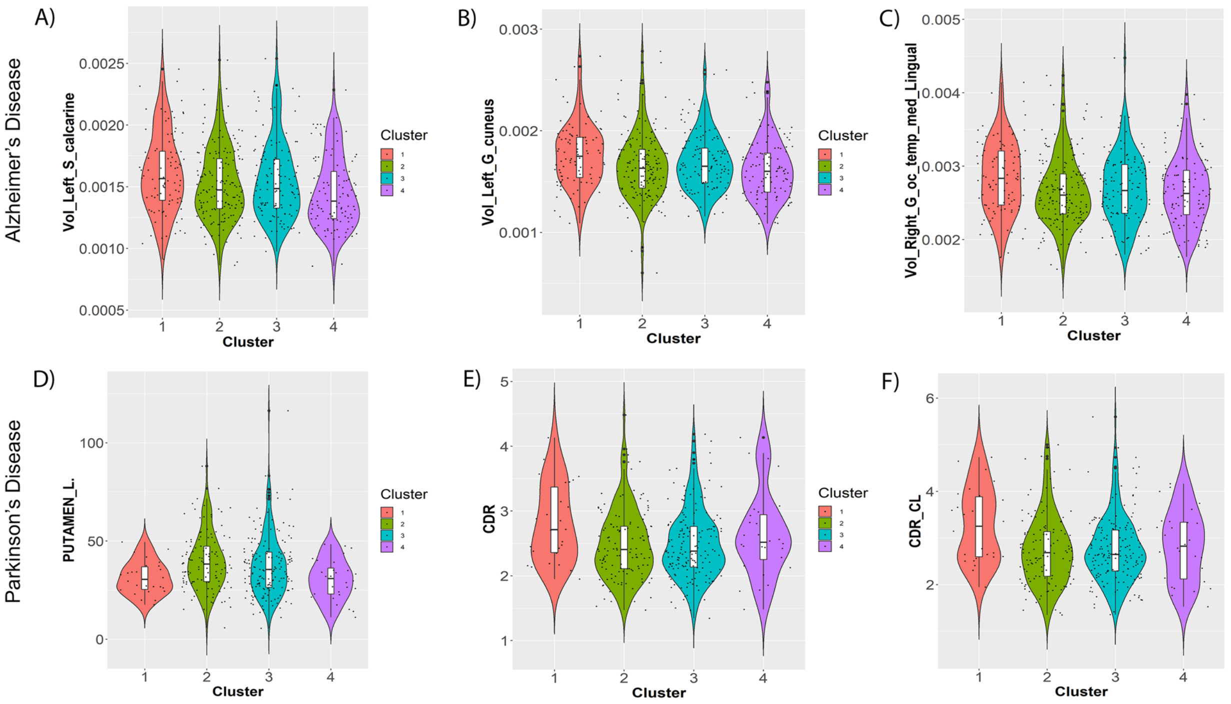


**Figure 5.** Example of significant differences between clusters with respect to brain imaging derived features at study baseline/time of first disease diagnosis (see "Methods" section). (**A**) left calcarine sulcus in AD patients; (**B**) left cuneus gyrus in AD patients; (**C**) volume of right medial occipitotemporal gyrus in AD patients; (**D**) DaTSCAN left Putamen—ratio to age expected value in healthy controls; (**E**) DaTSCAN Count Density Ratio: Caudate/Putamen; (**F**) DaTSCAN Count Density Ratio (CL): Caudate contralateral/Putamen contralateral. The Figures shows statistical distributions as violin plots (i.e. boxplots plus kernel density estimates), and individual data points are shown as superimposed dots.

brain structures of those patients which had a recent AD diagnosis at study baseline (n = 209) and correcting statistical differences for the confounding effects of age and sex. We found significant differences in several brain regions, such as the calcarine sulcus, the cuneus gyrus and the medial occipitotemporal gyrus (Table S14, Fig. 5A–C).

In PPMI, pairwise differences between the clusters were significant for in presynaptic dopaminergic imaging (DaTSCAN) were identified in caudate and putamen (Table S16, Fig. 5D–F). Also, the dopamine receptor density ratio of caudate versus putamen differed significantly between clusters.

Altogether, we concluded that our genetically derived clusters are associated with significant pathophysiological differences in the brain.

### Association with A-$\beta$, transcriptome and methylome changes.

Interestingly, the CSF protein A-$\beta$ showed significant pairwise concentration differences between all clusters in PPMI PD patients (Table S16), but not in ADNI AD subjects. However, there was only weakly significant difference in MOCA cognitive assessment scores across clusters ($p = 0.1$) and no correlation of A-$\beta$ levels with MOCA ($p = 0.53$, Kendall's tau: 0.03). This is in agreement with Melzer et al.[34], who reported no association of amyloid-beta deposits with cognitive decline in PD patients.

We further explored changes in transcriptome and methylome of ROSMAP AD patients on the level of Gene Ontology (GO) terms[35] and KEGG pathways[36] via Gene Set Enrichment Analysis (GSEA)[37]. This analysis was chosen due to the low sample size, and it can only reveal broad trends in the data, namely statistical enrichment of GO terms and pathway at the beginning or end of a fold change ranked list of genes. We here report findings of GO terms and KEGG pathways that were statistically enriched within one particular patient subtype, but not in others compared to cognitively normal controls. Enrichment maps[38] were used to provide a condensed view on biological processes and pathways that were particularly altered within one specific cluster (Figure S25–S41). Enrichment maps represent semantic similarities between GO terms (shown as nodes) via edges, and group GO terms together based on the hierarchical relationship between them. More results (including comparisons of one specific cluster to all others) can be found under https://clus2bio.scai.fraunhofer.de).

According to the highly condensed view of enrichment maps, for example, cluster 1 in AD specifically shows changes in the meiotic cycle compared to healthy donors (Figure S26). In fact, aberrant re-entry of neurons into the cell cycle has long been seen as one of the hallmarks of AD[39,40]. Cluster 2 shows transcriptome changes in microtubule-based processes (Figure S27). Indeed, the tau protein, which under healthy conditions stabilizes microtubule, in AD patients aggregates into insoluble filaments in the brain that represent one of the hallmarks of the disease[41]. Specific features of cluster 3 are gene expression changes of processes related to the termination of protein translation (Figure S28). Reduced global translation rates (and RNA levels) have been observed previously in AD patients[42]. Alteration of apoptosis related pathways is one of the features specific for cluster 4 (Figure S29), which is well known in the context of AD[43]. In addition, patients in this cluster show DNA methylation changes in growth factor beta receptors (Figure S35), which has been reported to promote AD pathology[44]. More results can be found in the Supplements.

PPMI transcriptome and methylome data has a larger sample size, but the main limitation is the fact that measurements have been derived from blood and thus only indirectly mirror the pathological processes in the brain. Accordingly, we here again decided to only focus on GSEA results comparing PD patients in each of the clusters against healthy controls (S30–S32; S37–S41). For example, cluster 1 shows specific methylome changes in the JAK-STAT signaling pathway. Inhibition of this pathway has been suggested as a therapy against PD[45]. Cluster 2 shows methylome changes of microtubule cytoskeleton organization. Tau deposition and filament assembly is one of the hallmarks of PD[46]. Assembly of misfolded proteins in PD yields activation of adaptive immune response[47]. According, transcriptional changes can be observed in cluster 2 as well. Cluster 3 demonstrates methylome changes of lipoprotein metabolism, which has recently been found altered in PD[48]. Cluster 4 shows transcriptional changes in protein ubiquitination, which has been suggested to also play a role in idiopathic forms of PD[49,50]. In addition, methylome changes of several metabolic processes were observed, which is in agreement with recent findings that view PD as a disorder of the cell metabolism[51]. Again, more results (including enrichment maps for GO terms) can be found in the Supplements and under https://clus2bio.scai.fraunhofer.de.

Altogether, our examples suggest that—despite the obvious limitations of the employed molecular data—each of the four clusters can be associated with biological processes that are solely enriched in one cluster and that are well known in the context of both diseases. Epigenetic changes were observed to a much higher extent in PD than in AD.

**Molecular differences between clusters can be linked to known disease mechanisms.** We next explored GO terms (biological processes) and KEGG pathways that were enriched in the *difference* between one cluster to all others. In other words, we looked into differential expression and differential methylation between cluster 1 and all others, cluster 2 and all others, and so on. For each of these comparisons a larger number of biological processes and pathways could be identified in both AD and PD (Tables S18, S20, S22, S24). In agreement to the findings in the last Section, significant differences between clusters in methylation could only be found in PD patients, but not in AD. Transcriptome differences between clusters were observed in both diseases.

We further explored the link between differences at the transcriptome and methylome level among clusters and known disease mechanisms in AD and PD. More specifically, we mapped our initially identified 15 common AD/PD disease mechanisms to disease specific mechanisms defined in the NeuroMMSig database[52]. That means, each of the common AD/PD mechanisms used in our clustering was identified with a certain NeuroMMSig gene set, if it was contained in that gene set. We found at least one NeuroMMSig gene set for each of the 15 mechanisms. Since each NeuroMMSig gene set equals a subgraph in one of our literature derived AD and PD disease maps (see "Methods" section), we could then systematically conduct graph mining. More specifically, we looked for shortest paths linking NeuroMMSig gene sets with biological processes and pathways identified in our omics data analysis. Shortest path calculations considered the causal direction of edges (marking e.g. a phosphorylation event) whenever possible. Due to the large number of results (over 600), we decided to implement an interactive web application for exploration (https://clus2bio.scai.fraunhofer.de/biomarkers). The web application also provides pointers to the scientific literature supporting each of the edges.

In the following, we highlight only selected examples (Fig. 6): As explained previously, cluster 1 is strongly associated with the genetic burden on AKT signaling. At the transcriptional level we observed significant downregulation of genes in the cell cycle process in AD patients (adj. $p$ value 0.03). Both can be linked together, as shown in Fig. 6A. AKT signaling influences acetylcholinesterase (AChE), which is thought to play a role in apoptotic processes[53] and amyloid-beta formation[54]. Amyloid-beta increases NAE1 via APP[55] and influences the entire cell cycle process[56].

In cluster 2, for PD patients we observed differential methylation of genes involved in processes related to microtubule cytoskeleton organization (adj. $p$ value < 0.001). Cluster 2 is—among others—associated with the genetic burden on TP53. As shown in Fig. 6B there is indeed a causal chain between TP53 and microtubule cytoskeleton organization. Elevated TP53 levels have been found to induce apoptosis and inflammation in PD[57]. Apoptotic processes yield a translocation of UTRN from the cytosol to mitochondria and subsequently increases cytochrome C[58] and alpha-syn[59], which itself is involved in microtubule cytoskeleton organization[60].

In cluster 3, for AD patients we observed significant transcriptional downregulation of genes involved in "long term synaptic depression" (adj. $p$ value 0.02). Cluster 3 is at the same time associated to the genetic burden on APOE. The connection between both is highlighted in Fig. 6C. For example, APOE has been suggested to increase insulin resistance[61], which yields synaptic depression of neurons and thus suggests the perception of AD as a "type 3 diabetes"[62].

Once again, these are only examples and further results can be explored via our web application.

**Potential implications for drug development.** Our previous results indicate that our AD/PD clustering can be associated with molecular and pathophysiological differences between patient subgroups. To better understand the potential utility of these patient subgroups for improving future AD and PD therapy, we conducted a target prioritization of all 27 genes involved into the 15 mechanisms that we had previously used in conjunction with SNP data to identify cluster patients. Target prioritization was done via Open Targets[1], which uses genetic evidence as well as literature mining to assign a confidence score to each protein as a potential drug target. In addition, tractability by small molecules and antibodies was considered. Figures S41, S42 highlight that

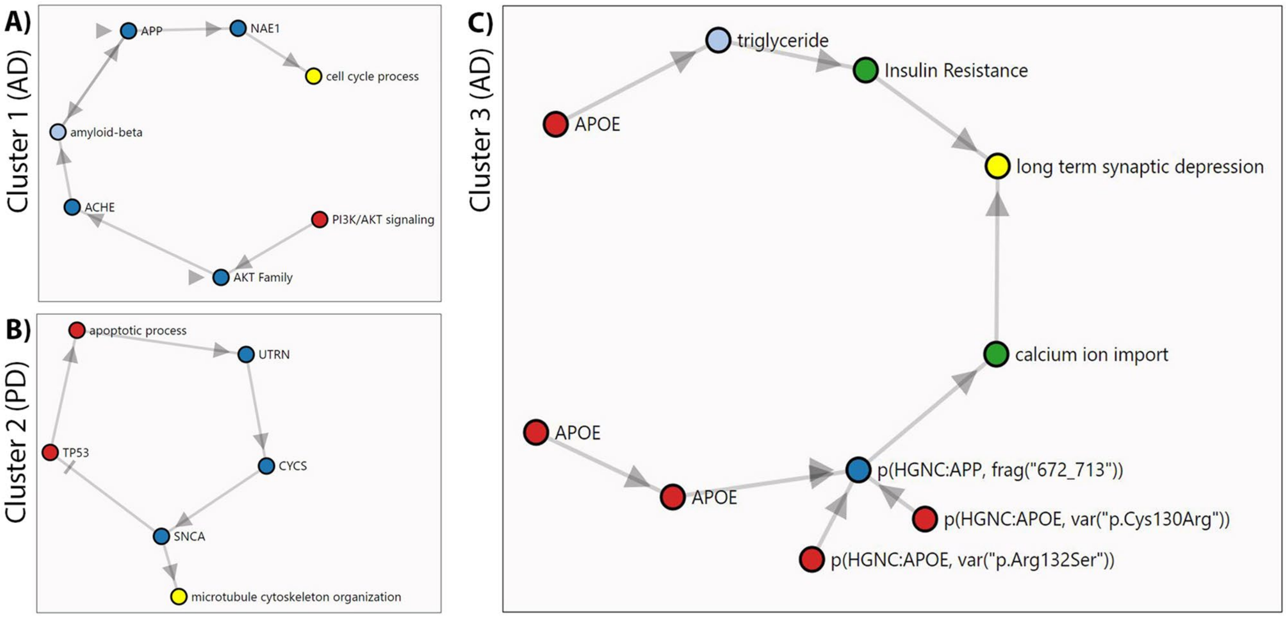


**Figure 6.** Examples of GO terms (biological processes) found significantly enriched in gene expression and/or methylome changes of one cluster compared to all others (yellow) together with their connections to genes playing a role in common AD/PD disease mechanisms (red): (**A**) Connection between AKT signaling (feature of cluster 1) and cell cycle in AD; (**B**) connection between TP53 (features of cluster 2) and microtubule cytoskeleton organization in PD; (**C**) connection between APOE4 (feature of cluster 3) and long term synaptic depression in AD. Intermediate signaling proteins are shown in blue. For more examples visit https://clus2bio.scai.fraunhofer.de/biomarkers.

in both diseases several potential targets could be identified via Open Targets. In addition, some of these targets could be clearly associated to one specific cluster (Table S1): In AD genes CDK5, GSK3B are strongly associated to cluster 2 (Table S1). APOE, PICALM, TOMM40, MTHFR and CD33 are linked to cluster 3. Further potential targets include SNCA, IL6 and CYCS, which are more strongly associated with clusters 2 and 3 than to the rest.

In PD, only SNCA, MAPT and APOE were identified as potential targets (Figure S42). MAPT is strongly associated with cluster 2 and APOE to cluster 3 (Table S1).

Altogether this analysis shows that our patient subtypes might be used to inform better targeted therapeutic strategies in AD and PD in the future.

## Conclusion

Precision medicine offers the hope of delivering the right treatment to the right patient, based on individual characteristics rather than population averages for these characteristics. Precision medicine is only an emerging reality at this moment, and moving closer to this vision will require non-trivial efforts in data mining and machine learning based on the entirety of available patient data[63]. Specifically, in neurology, this is extremely challenging, because on the one hand diseases are often highly multifaceted and on the other hand deep molecular multi-omics data (as frequently employed in cancer research) are difficult or even impossible to obtain for obvious reasons. Accordingly, in this work we started with an intensive literature mining effort, which mapped out the current mechanistic understanding of AD and PD pathologies and allowed us to identify shared molecular mechanisms. These shared molecular mechanisms were used as a starting point for developing a joint molecular subgrouping of AD and PD. More specifically, we used state-of-the-art unsupervised machine learning techniques to identify four mixed AD + PD patient clusters based on SNP burden scores of common AD/PD mechanisms. Importantly, the resultant disease subtypes manifest as mixtures of different mechanisms rather than being instances of single ones.

We validated the existence of patient clusters based on combined genotypes of 561 patients from AETIONOMY PD, ICEBERG, DIGPD, ROSMAP and IDIBAPS studies. Moreover, we conducted an in-depth analysis of clinical, imaging and molecular differences between patient clusters in both diseases. Our work demonstrated that SNP burden on mechanism level can be used to subdivide AD as well as PD patients jointly, and that clusters are associated with clinical, pathophysiological (specifically visible in brain imaging) and molecular differences between patients. We investigated the potential clinical utility of these differences by prioritizing drug targets for specific patient subgroups.

Overall, one should see our approach as complementary to the multitude of existing work that focuses on separate subgrouping of AD and PD based on polygenic risk scores[64], CSF, blood and imaging biomarkers[65–67], or based on clinical outcome measures[68,69]. We see the main distinction of our approach in a better understanding of the stratification potential of common AD and PD disease mechanisms, including the implications for future drug development.

Of course, our work is not without limitations: These can largely be associated to the limited availability of transcriptome and methylome data (with matched SNP genotypes from the same patient) in only two studies (ROSMAP and PPMI) and with relatively low sample sizes in ROSMAP. Moreover, clinical differences between cohorts imposed non-trivial challenges for reaching coherent conclusions regarding the clinical differences between patient subgroups. We thus see a need to more systematically replicate observational clinical studies in the neurology field. At the same time, such studies should preferably be longitudinal and collect multi-omics data from the same patient in a more systematic way than currently done in ROSMAP and PPMI. Such data should then be used to re-validate the findings presented here, specifically in terms of molecular differences between patient subgroups.

Altogether, we see our work as a step towards realizing the far-reaching vision of a completely molecular based definition of human disease, as formulated by Kola and Bell[4] and Strafella et al.[5]. As pointed out before, we see the potential impact of such an effort in the development of better targeted and thus hopefully more efficacious therapies in AD and PD in the future.

## Materials and methods

**Overview about used data.** *Studies used for discovery.* ADNI. Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.adni.loni.usc.edu). The longitudinal observation study includes—among others—486 subjects, which were diagnosed with mild sporadic AD during the study. 206 patients had a recent clinical AD diagnosis at study baseline. Data from ADNI subjects includes SNP based genotype (two different Illumina chip platforms), APOE4 status, CSF biomarkers, volume measurements of seven brain regions as well as different clinical and neuropsychological test results. In addition to the 7 brain volume measurements provided in the original ADNIMERGE dataset we calculated 193 subcortical brain region volumes from raw images, using the parcellation by Destrieux et al.[70], see details in Supplements. An overview about key demographic and clinical features of this study can be found in Table 1.

PPMI. The Parkinson's Progression Markers Initiative (PPMI) (www.ppmi-info.org/data) consists of multiple cohorts from a network of clinical sites with the aim to identify and verify progression markers in PD. It is a longitudinal observation study with data collected using standardized protocols[71]. PPMI comprises of eight cohorts with different clinical and genetic characteristics. Here we used data of 358 de novo diagnosed idiopathic PD patients and 198 healthy controls. All PD patients were initially untreated and diagnosed with the disease for two years or less. The dataset contains information about patient demographics, patient PD history, DaTSCAN imaging, non-motor symptoms, CSF biomarkers (A-$\beta$, $\alpha$-synuclein, dopamine, phospho-tau, total tau) and UPDRS scores. Genotype was available via whole genome sequencing data. In addition, whole blood transcriptome and methylome data was available for n = 306 and n = 277 of the same patients, respectively. The number of healthy controls with available gene expression and DNA methylation data was n = 151 and n = 112. An overview about key demographic and clinical features of this study can be found in Table 2.

*Studies used for validation.* Integrated AETIONOMY AD. Validation data comprised 237 clinically diagnosed sporadic AD cases with available genotype from ROSMAP[72] (n = 194) and 21 additional cases with available genotype from IDIBAPS taken from the AETIONOMY biomarker verification study[73]. We call the union of these 258 AD patients *integrated AETIONOMY AD* in the following. The data included:

- clinical characteristics: e.g. post-mortem diagnosis, age at death, gender
- genome-wide transcriptome (n = 56 AD cases with jointly available genotype and n = 50 cognitively normal controls) and methylome data (n = 53 AD cases with jointly available genotype and n = 34 cognitively normal controls) from post-mortem brain tissue (ROSMAP)

An overview about key demographic and clinical features of this study can be found in Table 1.

Integrated AETIONOMY PD. Validation data comprised idiopathic PD cases with available genotype that were diagnosed with PD for 2 years or less (in agreement to PPMI). 173 out of the 303 cases stem from DIGPD (NCT01564992)[74], 42 from ICEBERG (NCT02305147) and 88 were taken from study that is henceforth referred to as AETIONOMY PD[75]. We call the union of the 303 idiopathic PD patients *integrated AETIONOMY PD* in the following. The datasets are cross-sectional and include typical clinical outcome variables, such as MDS-UPDRS, Hoehn and Yahr stage, cognitive assessment scores (MMSE, MOCA), Epworth sleepiness scale (ESS), REM sleep behavior disorder (RBD), Hospital anxiety, and depression scale (HADS). An overview about key demographic and clinical features of this study can be found in Table 2.

**Identification of common molecular mechanisms.** Common molecular mechanisms between AD and PD were identified with the help of a systematic literature mining approach with post-hoc manual curation. More specifically, the text mining engine SCAIView[76] was used to construct cause-effect relationships between molecules, pathways, biological processes and imaging features in both, AD and PD, see Domingo-Fernandez et al. and Kodamullil et al. for details[52,77] for details. After manual curation, two computable disease maps, one for AD and one PD were created. Finally, we have also made them interactively usable via a dedicated web application (https://neurommsig.scai.fraunhofer.de/).

Calculation of the intersection of cause-effect relationships described in the AD and PD disease maps resulted into 27 genes grouped into 15 cause-effect relationship sub-graphs, called *mechanisms* from now on (see Fig. 2

and https://clus2bio.scai.fraunhofer.de/mechanisms for an interactive view). While some of these mechanisms describe only posttranslational modifications of a single protein, others reflect more complex protein–protein interactions and signaling cascades (Table S1). Key proteins described in both diseases include e.g. APOE, TAU, SNCA and TOMM40. These proteins are involved into several known disease relevant processes that we have made computationally accessible via our earlier developed NeuroMMSig database[52].

We mapped 148 genetic variants (SNPs) measured in ADNI1, ADNI2/GO as well as PPMI to the 27 common AD/PD disease genes via a combination of two strategies: a) proximity (using a 10 kbp window size); and b) eQTL mapping, see details in Supplements on page 2.

**Calculation of SNP burden on molecular mechanism level.** SNP data is inherently extremely sparse, i.e. even "common" genetic variants are comparably seldom in the data. This imposes a major challenge for any clustering algorithm, because the distance between two arbitrary SNP profiles based on the usual 0, 1, 2 encoding then becomes almost identical. That means clustering of raw SNP profiles is prone to become statistically unstable and noisy. To account for this fact, we embedded the 148 SNP profiles of AD and PD patients into a lower dimensional latent space while taking into account the grouping of SNPs according to 15 molecular cause-effect relationship subgraphs (aka molecular mechanisms) defined in the last section. We aimed for making this embedding non-linear to capture SNP-SNP interactions. Very recently, autoencoder networks (a specific deep learning technique) have been proposed for that purpose[78,79]. Autoencoders allow for learning a non-linear and low dimensional representation of SNP data for each patient, i.e. in essence a SNP burden score per mechanism (see Supplements for details). Based on the SNP burden scores a grouping of AD and PD patients can be established via clustering. Details will be explained later.

To maximize the chance for a later interpretation of the clustering and to avoid an imbalance due to differences in the number of mapped SNPs, we learned (sparse) autoencoder based SNP burden scores for each of the 15 mechanisms. That means we ended up with a 15 dimensional vector of genetic burden scores for each patient. Each of these 15 scores can be interpreted in terms of the relative contribution of each SNP to the overall score learned by the autoencoder network (Figures S2–S16). Details about the training procedure for our sparse autoencoder networks are described in the Supplementary Material on page 13.

**Unsupervised machine learning for patient (Bi-)clustering.** Based on the 15 dimensional SNP burden profile of each patient derived from SNP data we clustered patients. We here relied on sparse non-negative matrix factorization (sNMF). Briefly, sNMF factorizes a patients times mechanisms matrix $X$ into a product of two non-negative matrices $W$ and $H$, where $W$ represents a sparse mapping of mechanisms to clusters and $H$ a soft assignment of patients to clusters[80,81]. That means, for each patient cluster it is possible to identify the most influencing mechanisms (see Supplements for details). Hence, sNMF effectively yields a bi-clustering. The entire (bi-)clustering procedure is in practice is an iterative process that is dependent on the initialization of both matrices and should thus be repeated a number of times (here: 50) to yield a consensus. This consensus was used for further analysis.

The number of clusters $k$ corresponds to the number of columns of matrix $W$ and the number of rows of matrix $H$. We chose $k$ based on inspection of three statistical criteria (proportion of ambiguously clustered pairs, silhouette index, cophenetic correlation) and in comparison to a randomly permuted cluster assignment[82–84]. We then decided for the minimal number of clusters $k$ yielding the most stable clustering solution (lowest proportion of ambiguously clustered pairs) that was at the same time exhibiting a significantly larger silhouette index and cophenetic correlation than expected by chance. Details are explained and shown in Supplements on page 28.

**Validation of patient subtypes via independent studies.** Figure 7 gives an overview of our overall validation strategy, which consists of two parts: In the first part we re-clustered patients in our merged AD/PD validation cohort using the same workflow that we had established for our discovery cohort, which re-confirmed the possible existence of 4 clusters in AD and PD (Table S5).

In the second part of our validation, we followed the idea of assigning patients in an independent study to the pre-existing clusters discovered in ADNI and PPMI and then measuring the degree of coherence between the cluster assignments and originally discovered groups. For this purpose, we here adopted an approach proposed in Kapp and Tibshirani[26]: Following that approach we first developed a supervised machine learning classifier on the basis of the SNP data of patients in ADNI and PPMI. This allowed us to predict for any patient in an independent validation cohort the membership to a cluster in the discovery cohort based on the 148 SNP panel described above. We used an $l_1$ penalized logistic regression (i.e. LASSO) as a classification algorithm, and we evaluated the prediction performance of this classifier via a conventional 10 times repeated tenfold cross-validation procedure. That means we subsequently left out 1/10 of our discovery data for testing the classifier and only trained on the remaining 9/10 of the data. Autoencoder training and $l_1$ penalty hyper-parameter optimization was done within the cross-validation loop to prevent overoptimism. The corresponding cross-validated multi-class AUC of 70% is shown in Fig. 3E.

After classifier development, we were able to assign patients from independent studies to the clusters discovered in our discovery cohort (ADNI + PPMI). The in-group proportion measure (IGP) proposed by Kapp and Tibshirani then measured the proportion of patients in the validation study, whose nearest neighbors in the discovery cohort had the same cluster label. An IGP closer to 1 indicates a stronger coherence of the statistical distribution of data in the validation cohort with the clustering of the discovery cohort. An IGP closer to zero indicates disagreement.

To further assess the statistical significance of observed IGP values we performed a permutation test, in which we randomly permuted the cluster assignment of patients and re-calculated the IGP. This was done for 1000

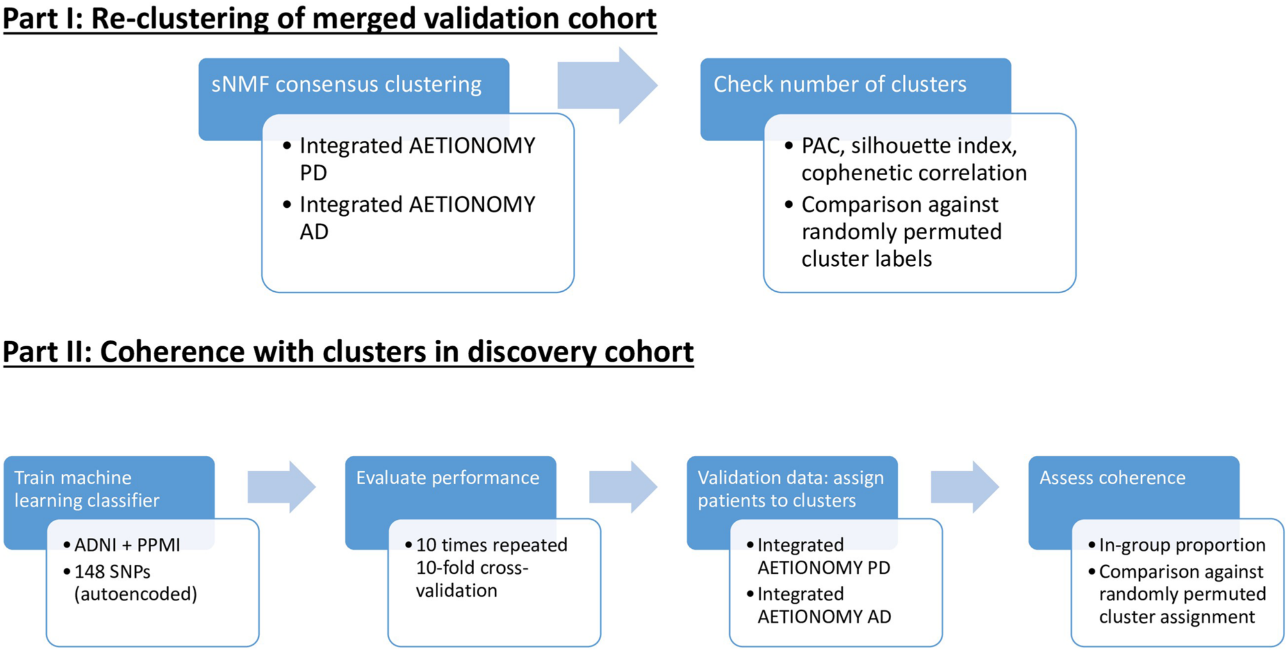

Part I: Re-clustering of merged validation cohort
sNMF consensus clustering
• Integrated AETIONOMY PD
• Integrated AETIONOMY AD
Check number of clusters
• PAC, silhouette index, cophenetic correlation
• Comparison against randomly permuted cluster labels
Part II: Coherence with clusters in discovery cohort
Train machine learning classifier
• ADNI + PPMI
• 148 SNPs (autoencoded)
Evaluate performance
• 10 times repeated 10-fold cross-validation
Validation data: assign patients to clusters
• Integrated AETIONOMY PD
• Integrated AETIONOMY AD
Assess coherence
• In-group proportion
• Comparison against randomly permuted cluster assignment

**Figure 7.** Strategy for validating genetically defined patient clusters.

times. None of the randomly permuted cluster assignments exceeded the IGP of the original clustering, i.e. our obtained results were highly significant.

It is worthwhile mentioning that we also re-calculated the IGP for our integrated AETIONOMY AD and PD cohorts separately to exclude that the observed high coherence was only true for one of the two diseases. Figures S23, S24 clearly demonstrate that no corresponding biases could be observed, i.e. IGP values were in a comparable range.

**Statistical analysis of clusters.** *Clinical data.* Clinically observed differences between patient subgroups might be impaired by multiple confounding factors. To identify these confounders, we initially performed a stepwise multinomial logistic regression (R-package “nnet”) with the cluster indicator as response and several potential confounders as predictors. This approach was chosen to account for the fact that many clinical variables show a highly skewed distribution. Considered confounders included:

- baseline diagnosis (ADNI),
- age (all),
- gender (all),
- marriage status (ADNI),
- education level (ADNI, ROSMAP, AETIONOMY PD),
- sub-study (ADNI1, ADNI2, ADNIGO, ADNI3; ROS, MAP),
- duration of the disease since the first diagnosis (PPMI, AETIONOMY PD, DIGPD, ICEBERG, ROSMAP),
- smoking history (AETIONOMY PD)
- coffee and alcohol consumption (AETIONOMY PD)
- ethnicity, including Spanish origin (PPMI, AETIONOMY PD, ICEBERG, DIGPD, ROSMAP), and
- prior neurological drug treatment (ROSMAP, PPMI, AETIONOMY PD, ICEBERG, DIGPD), including L-DOPA for PD (PPMI, AETIONOMY PD, ICEBERG, DIGPD).

The Akaike Information Criterion (AIC) was used for model selection, resulting in an “optimal” confounder set. It is worthwhile mentioning that none of the considered confounders demonstrated univariately significant association to cluster membership in any dataset according to a likelihood ratio test against the null model.

To determine the influence of clinical outcome measures (e.g. UPDRS3) in a second step we fitted a multinomial logistic regression model that included in addition to the selected confounders exactly one of the clinical variables of interest. In other words, there was a separate multinomial logistic regression for each clinical outcome measure. We then used a likelihood ratio test (Analysis of Deviance/type III ANOVA) to estimate the significance of the influence of the clinical variable of interest while correcting for confounders. In case of nominal significance ($p < 0.05$) we conducted a post-hoc analysis of pairwise differences between clusters using a Wald test. Due to multiple pairwise comparisons and the existence of several clinical variables of interest within each study, we jointly corrected P-values resulting from all statistical tests for multiple testing. This was done via Benjamini and Hochberg’s method[33]. Corresponding results are shown in Tables S9–S11.

Statistical analysis of longitudinal clinical data from ADNI and PPMI studies was performed via a generalized linear mixed model (R-package "lme4") between each pair of clusters. For this purpose, we subtracted from each clinical outcome score its baseline value and divided by the standard deviation of the outcome score at baseline, resulting into a patient specific progression score. Two alternative approaches to model time were considered, namely either as a numerical value or as a categorical factor. Model selection via the AIC was used to choose among these alternatives. Notably, we also included an interaction effect between cluster and time to model potentially existing time inhomogeneous effects (none of them being significant, though). Furthermore, we included a random intercept for each patient. Akin to the situation for the baseline data we performed a stepwise regression to initially select an "optimal" confounder set. Afterwards, a type III ANOVA was conducted to estimate the significance of a given clinical outcome. P-values of pairwise differences between clusters were corrected for multiple testing correction in the same way as described before. Results of the clinical time series analysis are shown in Tables S12–S13.

*Brain imaging.* Statistical analysis of features derived from MRI imaging in ADNI and DaTSCAN in PPMI in principle followed the same approach as those derived from for clinical data at baseline. The only difference was that for analysis of MRI imaging derive features we always used age and sex as confounders, and no further confounders were considered. Results of the statistical analysis are shown in Tables S14–S15. For ADNI we used two types of imaging data:

1. 7 precalculated brain volume measurements available in the ADNIMERGE package. We used always data from that visit, at which the first dementia diagnosis had been given.
2. 193 subcortical brain volumes calculated from Distrieux's parcellation approach, see details in Supplements.

All brain volume measurements were divided by intercranial volumes for normalization purposes before statistical analysis.

*CSF biomarkers.* CSF biomarkers were analyzed in the same way as clinical variables. Results are shown in Table S16. For ADNI AD patients we used always data from that visit, at which the first dementia diagnosis had been given.

*Omics data.* Analysis of transcriptomics (ROSMAP, PPMI), methylomics (ROSMAP, PPMI) data followed common practice in bioinformatics. Details are explained in the Supplements of this paper. Confounder analysis was done akin to clinical data. Accordingly, no confounders were identified in ROSMAP. However, initial quality control suggested a batch effect between the ROS and MAP sub-studies in DNA methylation data, which we corrected via ComBat[85]. In PPMI we used gender (RNAseq) and age (DNA methylation) as confounders. Complete analysis results are available under https://clus2bio.scai.fraunhofer.de/.

*Analysis tools.* We used bcftools (version 1.8) and PLINK (version 1.90b4.1) for SNP recoding from whole genome sequencing and genotyping data respectively.

R 3.5.1 was used for all data analyses purpose. R-package h2o (cluster version 3.26.0.2) was used for the autoencoder model training and logistic regression. We have used R-package NMF (version 0.21.0) for clustering. R-package biomart (version 2.36.1) was used for gene annotation. R-package SNPlocs.Hsapiens.dbSNP.20120608 (version 0.99.11) was used to map genomic coordinates to rsIDs. We have used R-package ggplot2 (version 3.0.0) for producing all the analysis plots in addition to inbuilt tool provided by R-package NMF package (version 0.21.0) for clustering and visualizations.

Received: 12 February 2020; Accepted: 23 October 2020
Published online: 05 November 2020

## References

1. Koscielny, G. *et al.* Open targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).
2. Mehta, D., Jackson, R., Paul, G., Shi, J. & Sabbagh, M. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010–2015. *Expert. Opin. Invest. Drugs* **26**, 735–739 (2017).
3. Fogel, D. B. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp. Clin. Trials Commun.* **11**, 156–164 (2018).
4. Kola, I. & Bell, J. A call to reform the taxonomy of human disease. *Nat. Rev. Drug. Discov.* **10**, 641–642 (2011).
5. Strafella, C. *et al.* Application of precision medicine in neurodegenerative diseases. *Front. Neurol.* **9**, 701 (2018).
6. McKenzie, J. A. *et al.* Neuroinflammation as a common mechanism associated with the modifiable risk factors for Alzheimer's and Parkinson's diseases. *Curr. Aging Sci.* https://doi.org/10.2174/1874609810666170315113244 (2017).
7. Kamal, M. A., Mushtaq, G. & Greig, N. H. Current update on synopsis of miRNA dysregulation in neurological disorders. *CNS Neurol. Disord. Drug Targets* **14**, 492–501 (2015).
8. Soto, C. Protein misfolding and disease; protein refolding and therapy. *FEBS Lett.* **498**, 204–207 (2001).
9. Ahmad, K. *et al.* Commonalities in biological pathways, genetics, and cellular mechanism between Alzheimer disease and other neurodegenerative diseases: an in silico-updated overview. *Curr. Alzheimer Res.* **14**, 1190–1197 (2017).
10. Guio-Vega, G. P. & Forero, D. A. Functional genomics of candidate genes derived from genome-wide association studies for five common neurological diseases. *Int. J. Neurosci.* **127**, 118–123 (2017).
11. Xie, A., Gao, J., Xu, L. & Meng, D. Shared mechanisms of neurodegeneration in Alzheimer's disease and Parkinson's disease. *Biomed. Res. Int.* **2014**, 648740 (2014).

12. Tan, L., Jiang, T., Tan, L. & Yu, J.-T. Toward precision medicine in neurological diseases. *Ann. Transl. Med.* **4**, 104 (2016).
13. Greene, L. A., Levy, O. & Malagelada, C. Akt as a victim, villain and potential hero in Parkinson's disease pathophysiology and treatment. *Cell. Mol. Neurobiol.* **31**, 969–978 (2011).
14. Jha, S. K., Jha, N. K., Kar, R., Ambasta, R. K. & Kumar, P. p38 MAPK and PI3K/AKT signalling cascades in Parkinson's disease. *Int. J. Mol. Cell Med.* **4**, 67–86 (2015).
15. Ahmad, F. *et al.* Reactive oxygen species-mediated loss of synaptic Akt1 signaling leads to deficient activity-dependent protein translation early in Alzheimer's disease. *Antioxid. Redox Signal.* **27**, 1269–1280 (2017).
16. Heneka, M. T. *et al.* NLRP3 is activated in Alzheimer's disease and contributes to pathology in APP/PS1 mice. *Nature* **493**, 674–678 (2013).
17. Leal, M. C., Casabona, J. C., Puntel, M. & Pitossi, F. J. Interleukin-1β and tumor necrosis factor-α: reliable targets for protective therapies in Parkinson's disease?. *Front. Cell Neurosci.* **7**, 53–53 (2013).
18. Lemere, C. A. A beneficial role for IL-1 beta in Alzheimer disease?. *J. Clin. Invest.* **117**, 1483–1485 (2007).
19. Wang, S., Yuan, Y.-H., Chen, N.-H. & Wang, H.-B. The mechanisms of NLRP3 inflammasome/pyroptosis activation and their role in Parkinson's disease. *Int. Immunopharmacol.* **67**, 458–464 (2019).
20. Rozycka, A., Jagodzinski, P. P., Kozubski, W., Lianeri, M. & Dorszewska, J. Homocysteine level and mechanisms of injury in Parkinson's disease as related to MTHFR, MTR, and MTHFD1 genes polymorphisms and L-Dopa treatment. *Curr. Genom.* **14**, 534–542 (2013).
21. Liu, L. *et al.* MTHFR C677T and A1298C polymorphisms may contribute to the risk of Parkinson's disease: a meta-analysis of 19 studies. *Neurosci. Lett.* **662**, 339–345 (2018).
22. Román, G. C. MTHFR gene mutations: a potential marker of late-onset Alzheimer's disease?. *J. Alzheimers Dis.* **47**, 323–327 (2015).
23. Wang, B. *et al.* Association of MTHFR gene polymorphism C677T with susceptibility to late-onset Alzheimer's disease. *J. Mol. Neurosci.* **27**, 23–27 (2005).
24. Bohush, A., Niewiadomska, G. & Filipek, A. Role of mitogen activated protein kinase signaling in Parkinson's disease. *Int. J. Mol. Sci.* **19**, 2973 (2018).
25. Kheiri, G., Dolatshahi, M., Rahmani, F. & Rezaei, N. Role of p38/MAPKs in Alzheimer's disease: implications for amyloid beta toxicity targeted therapy. *Rev. Neurosci.* **30**, 9–30 (2018).
26. Kapp, A. V. & Tibshirani, R. Are clusters found in one dataset present in another dataset?. *Biostatistics* **8**, 9–31 (2007).
27. Kueper, J. K., Speechley, M. & Montero-Odasso, M. The Alzheimer's disease assessment scale-cognitive subscale (ADAS-Cog): modifications and responsiveness in pre-dementia populations. A narrative review. *J. Alzheimers Dis.* **63**, 423–444 (2018).
28. O'Bryant, S. E. *et al.* Staging dementia using clinical dementia rating scale sum of boxes scores. *Arch. Neurol.* **65**, 1091–1095 (2008).
29. Folstein, M. F., Folstein, S. E. & McHugh, P. R. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **12**, 189–198 (1975).
30. Nasreddine, Z. S. *et al.* The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).
31. Teng, E. *et al.* Utility of the functional activities questionnaire for distinguishing mild cognitive impairment from very mild Alzheimer's disease. *Alzheimer Dis. Assoc. Disord.* **24**, 348–353 (2010).
32. Chelune, G. J., Bornstein, R. A. & Prifitera, A. The Wechsler Memory Scale—revised. In *Advances in Psychological Assessment* Vol. 7 (eds McReynolds, P. *et al.*) 65–99 (Springer, US, New York, 1990). https://doi.org/10.1007/978-1-4613-0555-2_3.
33. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
34. Melzer, T. R. *et al.* Beta amyloid deposition is not associated with cognitive impairment in Parkinson's disease. *Front. Neurol.* **10**, 391 (2019).
35. The Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
36. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, 480–484 (2008).
37. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
38. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).
39. Moh, C. *et al.* Cell cycle deregulation in the neurons of Alzheimer's disease. *Results Probl. Cell Differ.* **53**, 565–576 (2011).
40. Raina, A. K., Monteiro, M. J., Mcshea, A. & Smith, M. A. The role of cell cycle-mediated events in Alzheimer's disease. *Int. J. Exp. Pathol.* **80**, 71–76 (1999).
41. Medeiros, R., Baglietto-Vargas, D. & LaFerla, F. M. The role of tau in Alzheimer's disease and related disorders. *CNS Neurosci. Ther.* **17**, 514–524 (2011).
42. Kapur, M., Monaghan, C. E. & Ackerman, S. L. Regulation of mRNA translation in neurons–a matter of life and death. *Neuron* **96**, 616–637 (2017).
43. Obulesu, M. & Lakshmi, M. J. Apoptosis in Alzheimer's disease: an understanding of the physiology, pathology and therapeutic avenues. *Neurochem. Res.* **39**, 2301–2312 (2014).
44. Tesseur, I. *et al.* Deficiency in neuronal TGF-beta signaling promotes neurodegeneration and Alzheimer's pathology. *J. Clin. Invest.* **116**, 3060–3069 (2006).
45. Qin, H. *et al.* Inhibition of the JAK/STAT pathway protects against α-synuclein-induced neuroinflammation and dopaminergic neurodegeneration. *J. Neurosci.* **36**, 5144–5159 (2016).
46. Zhang, X. *et al.* Tau pathology in Parkinson's disease. *Front. Neurol.* **9**, 809 (2018).
47. Mosley, R. L., Hutter-Saunders, J. A., Stone, D. K. & Gendelman, H. E. Inflammation and adaptive immunity in Parkinson's disease. *Cold Spring Harb. Perspect. Med.* **2**, a009381 (2012).
48. Alecu, I. & Bennett, S. A. L. Dysregulated lipid metabolism and its role in α-synucleinopathy in Parkinson's disease. *Front. Neurosci.* **13**, 328 (2019).
49. Chin, L.-S. & Li, L. Ubiquitin phosphorylation in Parkinson's disease: implications for pathogenesis and treatment. *Transl. Neurodegen.* **5**, 1 (2016).
50. Lim, K.-L. & Tan, J. M. Role of the ubiquitin proteasome system in Parkinson's disease. *BMC Biochem.* **8**, S13 (2007).
51. Anandhan, A. *et al.* Metabolic dysfunction in Parkinson's disease: bioenergetics, redox homeostasis and central carbon metabolism. *Brain Res. Bull.* **133**, 12–30 (2017).
52. Domingo-Fernandez, D. *et al.* Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics* **33**, 3679–3681 (2017).
53. Jiang, H., Zhang, J., Zhu, H., Li, H. & Zhang, X. Nerve growth factor prevents the apoptosis-associated increase in acetylcholinesterase activity after hydrogen peroxide treatment by activating Akt. *Acta Biochim. Biophys. Sin. (Shanghai)* **39**, 46–56 (2007).
54. Inestrosa, N. C., Dinamarca, M. C. & Alvarez, A. Amyloid-cholinesterase interactions. implications for Alzheimer's disease. *FEBS J.* **275**, 625–632 (2008).
55. Chen, Y., Liu, W., McPhie, D. L., Hassinger, L. & Neve, R. L. APP-BP1 mediates APP-induced apoptosis and DNA synthesis and is increased in Alzheimer's disease brain. *J. Cell Biol.* **163**, 27–33 (2003).

56. Nizzari, M. *et al.* Neurodegeneration in Alzheimer disease: role of amyloid precursor protein and presenilin 1 intracellular signaling. *J. Toxicol.* **2012**, 187297 (2012).
57. Mogi, M., Kondo, T., Mizuno, Y. & Nagatsu, T. p53 protein, interferon-gamma, and NF-kappaB levels are elevated in the parkinsonian brain. *Neurosci. Lett.* **414**, 94–97 (2007).
58. Frank, S. *et al.* The role of dynamin-related protein 1, a mediator of mitochondrial fission, in apoptosis. *Dev. Cell* **1**, 515–525 (2001).
59. Seo, J.-H. *et al.* Alpha-synuclein regulates neuronal survival via Bcl-2 family expression and PI3/Akt kinase pathway. *FASEB J.* **16**, 1826–1828 (2002).
60. Sekigawa, A. *et al.* Diversity of mitochondrial pathology in a mouse model of axonal degeneration in synucleinopathies. *Oxid. Med. Cell Longev.* **2013**, 817807 (2013).
61. Elbein, S. C. & Hasstedt, S. J. Quantitative trait linkage analysis of lipid-related traits in familial type 2 diabetes: evidence for linkage of triglyceride levels to chromosome 19q. *Diabetes* **51**, 528–535 (2002).
62. de la Monte, S. M. & Wands, J. R. Alzheimer's disease is type 3 diabetes-evidence reviewed. *J. Diabetes Sci. Technol.* **2**, 1101–1113 (2008).
63. Fröhlich, H. *et al.* From hype to reality: data science enabling personalized medicine. *BMC Med.* **16**, 150 (2018).
64. Mukherjee, S. *et al.* Genetic data and cognitively defined late-onset Alzheimer's disease subgroups. *Mol. Psychiatry* https://doi.org/10.1038/s41380-018-0298-8 (2018).
65. Toschi, N. *et al.* Biomarker-guided clustering of Alzheimer's disease clinical syndromes. *Neurobiol. Aging* https://doi.org/10.1016/j.neurobiolaging.2019.08.032 (2019).
66. Martí-Juan, G., Sanroma, G. & Piella, G. Revealing heterogeneity of brain imaging phenotypes in Alzheimer's disease based on unsupervised clustering of blood marker profiles. *PLoS ONE* **14**, e0211121 (2019).
67. Young, A. L. *et al.* Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat. Commun.* **9**, 4273 (2018).
68. Mu, J. *et al.* Parkinson's disease subtypes identified from cluster analysis of motor and non-motor symptoms. *Front. Aging Neurosci.* **9**, 301 (2017).
69. Peter, J. *et al.* Subgroups of Alzheimer's disease: stability of empirical clusters over time. *J. Alzheimers Dis.* **42**, 651–661 (2014).
70. Destrieux, C., Fischl, B., Dale, A. & Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* **53**, 1–15 (2010).
71. Initiative, P. P. M. The Parkinson progression marker initiative (PPMI). *Prog. Neurobiol.* **95**, 629–635 (2011).
72. De Jager, P. L. *et al.* A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* **5**, 180142 (2018).
73. Brosseron, F. *et al.* Multicenter Alzheimer's and Parkinson's disease immune biomarker verification study. *Alzheimers Dement.* https://doi.org/10.1016/j.jalz.2019.07.018 (2019).
74. Corvol, J. C. *et al.* Longitudinal analysis of impulse control disorders in Parkinson disease. *Neurology* **91**, e189–e201 (2018).
75. Corvol, J.-C. *et al.* AETIONOMY, a Cross-Sectional Study Aimed at validating a new taxonomy of Neurodegenerative Diseases: Study design and subject characteristics. *medRxiv* https://doi.org/10.1101/19004804 (2019).
76. Younesi, E. *et al.* Mining biomarker information in biomedical literature. *BMC Med. Inform. Decis. Mak.* **12**, 148 (2012).
77. Kodamullil, A. T., Younesi, E., Naz, M., Bagewadi, S. & Hofmann-Apitius, M. Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. *Alzheimers Dement.* **11**, 1329–1339 (2015).
78. Montañez, C. A. C., Fergus, P., Chalmers, C. & Hind, J. Analysis of extremely obese individuals using deep learning stacked autoencoders and genome-wide genetic data. arXiv:1804.06262 [cs, q-bio] (2018).
79. Xie, R., Wen, J., Quitadamo, A., Cheng, J. & Shi, X. A deep auto-encoder model for gene expression prediction. *BMC Genom.* **18**, 845 (2017).
80. Kim, H. & Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **23**, 1495–1502 (2007).
81. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
82. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comp. Appl. Math.* **20**, 53–65 (1987).
83. Şenbabaoğlu, Y., Michailidis, G. & Li, J. Z. Critical limitations of consensus clustering in class discovery. *Sci. Rep.* **4**, 6207 (2014).
84. Sokal, R. R. & Rohlf, F. J. The comparison of dendrograms by objective methods. *Taxon* **11**, 33–40 (1962).
85. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat* **8**, 118–127 (2007).

## Acknowledgements

We thank Stephanie Carvalho, Aishwarya Alex and Reinhard Schneider for helping us to access the data from IDIBAPS, ICEBERG, DIGPD and AETIONOMY PD studies. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research \& Development, LLC.; Johnson \& Johnson Pharmaceutical Research \& Development LLC.; Lumosity; Lundbeck; Merck \& Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners. A list of names of all of the PPMI funding partners can be found at www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors/.

## Author contributions

H.F. designed the study; M.A.E., A.H., P.W. and D.D.F. implemented workflows and performed the data analyses; M.A.E., A.H. and P.W. processed the discovery and validation data sets; H.V. performed image analysis; H.F., P.S., M.H.A. and J.C.C. advised and supervised the work. All authors were involved in drafting and approving the final manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL. The research leading to these results has received partial support from the Innovative Medicines Initiative Joint Undertaking under grant agreement #115568, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. The AETIONOMY cohort was part of the AETIONOMY project. The design and the patient characteristics are described elsewhere (medRxiv 19004804; https://doi.org/10.1101/19004804). The complete study group can be found in the supplements of this paper. DIGPD was funded by a grant from the French Ministry of Health, Programme Hospitalier de Recherche Clinique AOR0810. The complete study group can be found in the supplements of this paper. ICEBERG was funded by a grant from the program „Investissements d'Avenir" ANR-10-IAIHU-06, and by a grant from the EDF foundation. The complete study group can be found in the supplements of this paper.

## Competing interests

AH, PW, PS and HF received salaries from UCB Biosciences GmbH and UCB Celltech Ltd, respectively (both branches of UCB Pharma S.A.). UCB Pharma had no influence on the content of this work. All remaining authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41598-020-76200-4.

**Correspondence** and requests for materials should be addressed to H.F.

Chapter 4

# Conclusions

In this work, we developed a novel unsupervised machine learning-based approach that could integrate the mechanistic information from knowledge assemblies with patient-level omics data for the molecular mechanism-based stratification of AD and PD patients. While the establishment of such patient stratification remains one of the biggest challenges in the field of NDD pathophysiology, our work demonstrates a robust patient stratification strategies that can also be reproducible in completely independent study cohorts. Moreover, significant associations between different clinical and molecular disease biomarkers (i.e., clinical, imaging, transcriptomics, and methylomics) and the predicted patient subgroups show the clinical relevance of our clustering strategy. To our knowledge, this work is the first successful demonstration of mechanism-based patient stratification in the area of neurodegenerative disease. Further, in developing the clustering workflow, we have tackled a critical bioinformatic problem of reducing the genomic data dimensionality by aggregating the SNPs information into the singular mechanism scores. Looking ahead in the future, this work would facilitate the practice of stratified medicine in the NDD field by enabling us to target distinct pathophysiological mechanisms underlying various patient subgroups.

# 5 Conclusion and outlook

Given the advances in big data technologies in biology and ever-increasing computational power, the biomedical research field has been constantly growing at an unprecedented rate compared to any other research domain over the past decades [33, 119]. While a surge in multi-modal and multi-scale data (e.g., omics, imaging, EHRs, smart device data) has granted access to explore many uncharted territories in biology, the successful integration of artificial intelligence in biomedical science will serve as a powerful tool to achieve that goal. These significant advancements have raised provocative questions like, do we still need hypothesis-driven scientific inquiry, or can we engage in scientific inquiry without any prior beliefs, and be free from established ways of thinking or doing to explore a limitless number of possibilities [120]? However, in practice, all this big data represents highly selected phenomena of the real-world due to sampling bias and tends to exclude the majority of biological work [121]. As a result, data-driven anomalies can be understood only by contrast or the presence of prior assumptions. Hence, both knowledge- and data-driven approaches are fundamental parts of an interactive cycle of knowledge acquisition and complement each other in the new scientific quest [122].

The present work is a systematic attempt to bridge prior knowledge in the NDD domain (i.e., AD and PD) with patient-level omics and clinical data in order to seek answers for two pressing issues in the biomedical science: i) can we leverage prior knowledge to stratify AD and PD patients by integrating omics and clinical data, and if yes, ii) can we target the mechanisms underlying the patient subgroups for stratified medicine development? However, the practice of patient stratification and stratified medicine in the context of NDD is relatively new and more challenging compared to the oncology domain due to their multifactorial nature and poor understanding of the disease etiology [18]. Hence, establishing a successful patient stratification strategy in the field of NDD could end the long-standing drought in drug development in this domain by enabling us to specifically target responsible mechanisms that define each of the subgroups.

First, this thesis addressed one of the fundamental topics in bioinformatics by demonstrating the necessity of having organized domain knowledge in the field of NDDs in order to increase the understanding and interpretation capability of biological systems. The enrichment of AD knowledge assembly with drug interaction information showed organized and computable knowledge can indeed explain various biological phenomena in complex disorders like AD and that explanatory power can further be used to propose mechanism-based drug repositioning candidates. We have addressed another essential problem of data integration with knowledge for explaining the anomalies or trends seen in biological data. While resources like the connectivity map and LINCS have been available to researchers for some time, there have been a lack of drug repositioning workflows that could systematically substantiate their results with mechanistic insights on these candidate drugs. The PS4DR workflow has shown how domain-specific pathway knowledge can play a crucial role in the interpretation of large biological datasets (i.e., genome and transcriptome data) and the prediction of drugs based on how they counteract disrupted biological processes in different disorders. Finally, while lack of subtype and tissue-specific data is a massive problem in most NDDs due to the high degree of cellular heterogeneity and post-mortem sample collection difficulties [123], our AD and PD clustering workflow demonstrated that having prior knowledge can help to effectively interpret non-tissue-specific data. The workflow successfully integrated mechanistic information from knowledge assemblies with patient-level omics data for mechanism-based patient stratification in the area of neurodegenerative diseases.

In addressing all the challenges described, several new algorithmic approaches had to be developed in the course of this thesis. While developing PS4DR, we have established a robust and flexible workflow to harmonize genome-wide association study (GWAS) and transcriptome data in order to make them interoperable with each other. This workflow can enable users to use any new public or proprietary genetic and transcriptomic datasets of their interest to analyze for drug repositioning in any diseases without requiring any substantial data integration efforts. During the AD and PD patient clustering pipeline development, we have solved a very important issue of reducing data dimensionality of large scale, patient-level omics data. We have developed a novel workflow to reduce hundreds to thousands of SNPs into a single score (i.e., mechanism scores) via combining prior knowledge with deep sparse autoencoder algorithms. While we have showcased the workflow with SNPs data, it has the ability to reduce data dimensionality with any type of omics (i.e., genomics, transcriptomics, methylomics, proteomics, and metabolomics) data without requiring any changes in the workflow. In addition to its ability to integrate new data dimensionalities, the workflow is also well equipped to deal with data in massive amounts by enabling parallel computing for multiple model training in computer clusters. Most importantly, in the spirit of open and reproducible science, all the resources, scripts and pipelines that have been developed in the course of this thesis are made publicly available via online web tools and GitHub repositories in order to enable other researchers to reproduce our works and conduct their own investigations.

However, despite its successful implementation and promising results, our work is not without limitations. The knowledge assemblies might sometimes represent highly selected information due to a publication bias in the field of biomedical research [124]. For instance, information on the 'role of insulin signaling in AD' in our knowledge assembly will not be as rich as the information of amyloid or tau hypotheses due to the large publication turnout gap between these topics. This could lead to a critical problem in hypothesis-driven analyses of the data due to knowledge imbalances among different subdomains while interpreting data anomalies. Another pitfall is the lack of availability of different data modalities across different study cohorts. While we used transcriptome and methylome data for the validation of our AD-PD patient stratification work, we could find these two data modalities in only two out of our five validation cohorts. We also had similar issues with regard to a lack of common data modalities (i.e., genetic and transcriptomic data) for the same diseases. As a consequence, we were able to implement the PS4DR workflow for only 43 diseases that had both these data modalities.

While this thesis represents only the work that has been published in peer-reviewed journals, the impact of the work goes far beyond that. The chemically enriched AD knowledge assembly has contributed significantly to pave the way for the Human Brain Pharmacome project [1] and is now being enhanced with quantitative chemical information in order to propose repurposing candidates based on chemical determinants (e.g., molecular weight, structure, IC50) in addition to considering their mechanistic interpretations. Furthermore, the prioritized mechanisms and proposed drugs will be tested experimentally in-situ and in-vitro within the project context. PS4DR is a flexible drug repositioning workflow that presents a great opportunity to integrate any new disease, drug, or pathway dataset. One immediate future effort would be replacing canonical pathway information with

[1] https://pharmacome.scai.fraunhofer.de/

our NeuroMMSig mechanism enrichment server [6] to get pathophysiological mechanistic insights of various NDDs. Such integration will improve mechanistic drug repositioning by enabling data interpretation with respect to disease-specific biology over canonical pathways.

Finally, this thesis work has successfully established mechanism-based AD and PD stratification and demonstrated the implication of such taxonomy in therapeutic target identification to some extent. While our work represents one of the main outcomes of the AETIONOMY project [2], more systematic effort is required to enable us to take full advantage of stratified medicine. One future direction would be using drug perturbed gene expression data (i.e., LINCS dataset) to train a separate autoencoder model to integrate with our knowledge assemblies. Combining this new drug response model with the existing SNPs data model will not only allow us to systematically investigate all mechanisms responsible for those patient groups, but will also help us to identify drugs that work best on different subgroups. This work can also serve as a good starting point for further omics (i.e., transcriptome, methylome, proteome, and metabolome) and clinical (i.e., imaging, EHRs, and smart device data) data integration for the development of predictive machine learning models for more robust patient stratification. Other deep learning algorithms, like the convolutional neural network (CNN), deep belief network (DBN), Restricted Boltzmann machine (RBM), etc., are also worth exploring to identify the top-performing model. Looking ahead in the future, translation of this work into clinical practice would set the stage for bringing the precision medicine concept into reality.

[2] https://www.aetionomy.eu/

# Bibliography

[1] National Research Council, Division on Earth and Life Studies, Board on Life Sciences, Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. en. National Academies Press, **Jan. 2012**.

[2] P. Krzyszczyk, A. Acevedo, E. J. Davidoff, L. M. Timmins, I. Marrero-Berrios, M. Patel, C. White, C. Lowe, J. J. Sherba, C. Hartmanshenn, K. M. O'Neill, M. L. Balter, Z. R. Fritz, I. P. Androulakis, R. S. Schloss, M. L. Yarmush. The growing role of precision and personalized medicine for cancer treatment. en. *Technology* **Sept. 2018**, 6 (3-4), 79–100.

[3] L. Tan, T. Jiang, L. Tan, J.-T. Yu. Toward precision medicine in neurological diseases. en. *Annals of Translational Medicine* **Mar. 2016**, 4 (6).

[4] D. Cirillo, A. Valencia. Big data analytics for personalized medicine. en. *Curr. Opin. Biotechnol.* **Aug. 2019**, 58, 161–167.

[5] F. Azuaje. Artificial intelligence for precision oncology: beyond patient stratification. en. *NPJ Precis Oncol* **Feb. 2019**, 3, 6.

[6] D. Domingo-Fernández, A. T. Kodamullil, A. Iyappan, M. Naz, M. A. Emon, T. Raschka, R. Karki, S. Springstubbe, C. Ebeling, M. Hofmann-Apitius. Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. en. *Bioinformatics* **Nov. 2017**, 33 (22), 3679–3681.

[7] K. W. Johnson, K. Shameer, B. S. Glicksberg, B. Readhead, P. P. Sengupta, J. L. M. Björkegren, J. C. Kovacic, J. T. Dudley. Enabling Precision Cardiology Through Multiscale Biology and Systems Medicine. en. *JACC Basic Transl Sci* **June 2017**, 2 (3), 311–327.

[8] T. Berman, A. Bayati. *What are Neurodegenerative Diseases and How Do They Affect the Brain?* **2018**.

[9] X. Chen, W. Pan. *The Treatment Strategies for Neurodegenerative Diseases by Integrative Medicine*. **2015**.

[10] Y. H. El-Hayek, R. E. Wiley, C. P. Khoury, R. P. Daya, C. Ballard, A. R. Evans, M. Karran, J. L. Molinuevo, M. Norton, A. Atri. *Tip of the Iceberg: Assessing the Global Socioeconomic Costs of Alzheimer's Disease and Related Dementias and Strategic Implications for Stakeholders*. **2019**.

[11] R. L. Nussbaum, C. E. Ellis. *Alzheimer's Disease and Parkinson's Disease*. **2003**.

[12] *What Is Alzheimer's Disease?* `https://www.nia.nih.gov/health/what-alzheimers-disease`. Accessed: 2020-3-24.

[13] K. S. Anand, V. Dhikav. Hippocampus in health and disease: An overview. en. *Ann. Indian Acad. Neurol.* **Oct. 2012**, 15 (4), 239–246.

[14] A. Serrano-Pozo, M. P. Frosch, E. Masliah, B. T. Hyman. Neuropathological Alterations in Alzheimer Disease. en. *Cold Spring Harb. Perspect. Med.* **Sept. 2011**, 1 (1).

[15] C. A. Gold, A. E. Budson. Memory loss in Alzheimer's disease: implications for development of therapeutics. en. *Expert Rev. Neurother.* **Dec. 2008**, 8 (12), 1879–1891.

[16] T. McLaughlin, H. Feldman, H. Fillit, M. Sano, F. Schmitt, P. Aisen, C. Leibman, L. Mucha, J. M. Ryan, S. D. Sullivan, D. E. Spackman, P. J. Neumann, J. Cohen, Y. Stern. Dependence as a unifying construct in defining Alzheimer's disease severity. en. *Alzheimers. Dement.* **Nov. 2010**, 6 (6), 482–493.

[17] D. J. Selkoe, J. Hardy. The amyloid hypothesis of Alzheimer's disease at 25 years. en. *EMBO Mol. Med.* **June 2016**, 8 (6), 595–608.

[18] R. B. Maccioni, G. Farías, I. Morales, L. Navarrete. *The Revitalized Tau Hypothesis on Alzheimer's Disease*. **2010**.

[19] E Mohandas, V Rajmohan, B Raghunath. Neurobiology of Alzheimer's disease. en. *Indian J. Psychiatry* **Jan. 2009**, 51 (1), 55–61.

[20] I. Piaceri. *Genetics of familial and sporadic Alzheimer s disease*. **2013**.

[21] J. Dorszewska, M. Prendecki, A. Oczkowska, M. Dezor, W. Kozubski. Molecular Basis of Familial and Sporadic Alzheimer's Disease. en. *Curr. Alzheimer Res.* **2016**, 13 (9), 952–963.

[22] D. Mehta, R. Jackson, G. Paul, J. Shi, M. Sabbagh. *Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015*. **2017**.

[23] G. E. Alexander. Biology of Parkinson's disease: pathogenesis and pathophysiology of a multisystem neurodegenerative disorder. en. *Dialogues Clin. Neurosci.* **Sept. 2004**, 6 (3), 259–280.

[24] O.-B. Tysnes, A. Storstein. Epidemiology of Parkinson's disease. en. *J. Neural Transm.* **Aug. 2017**, 124 (8), 901–905.

[25] T. R. Mhyre, J. T. Boyd, R. W. Hamill, K. A. Maguire-Zeiss. *Parkinson's Disease*. **2012**.

[26] M. J. Ma, M Joe Ma. *Biopsy Pathology of Neurodegenerative Disorders in Adults*. **2010**.

[27] C. Chai, K.-L. Lim. *Genetic Insights into Sporadic Parkinson's Disease Pathogenesis*. **2014**.

[28] G. Ganguly, S. Chakrabarti, U. Chatterjee, L. Saso. *Proteinopathy, oxidative stress and mitochondrial dysfunction: cross talk in Alzheimer's disease and Parkinson's disease*. **2017**.

[29] J. A. Santiago, V. Bottero, J. A. Potashkin. Dissecting the Molecular Mechanisms of Neurodegenerative Diseases through Network Biology. en. *Front. Aging Neurosci.* **May 2017**, 9, 166.

[30] M. Caruana, R. Cauchi, N. Vassallo. *Putative Role of Red Wine Polyphenols against Brain Pathology in Alzheimer's and Parkinson's Disease*. **2016**.

[31] R. S. Desikan, A. J. Schork, Y Wang, A Witoelar, M Sharma, L. K. McEvoy, D Holland, J. B. Brewer, C.-H. Chen, W. K. Thompson, D Harold, J Williams, M. J. Owen, M. C. O'Donovan, M. A. Pericak-Vance, R Mayeux, J. L. Haines, L. A. Farrer, G. D. Schellenberg, P Heutink, A. B. Singleton, A Brice, N. W. Wood, J Hardy, M Martinez, S. H. Choi, A DeStefano, M. A. Ikram, J. C. Bis, A Smith, A. L. Fitzpatrick, L Launer, C van Duijn, S Seshadri, I. D. Ulstein, D Aarsland, T Fladby, S Djurovic, B. T. Hyman, J Snaedal, H Stefansson, K Stefansson, T Gasser, O. A. Andreassen, A. M. Dale, ADNI, ADGC, GERAD, CHARGE and IPDGC Investigators. Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus. en. *Mol. Psychiatry* **Dec. 2015**, 20 (12), 1588–1595.

[32] A. Xie, J. Gao, L. Xu, D. Meng. *Shared Mechanisms of Neurodegeneration in Alzheimer's Disease and Parkinson's Disease*. **2014**.

[33] J. Luo, M. Wu, D. Gopukumar, Y. Zhao. Big Data Application in Biomedical Research and Health Care: A Literature Review. en. *Biomed. Inform. Insights* **Jan. 2016**, 8, 1–10.

[34] A. W. Toga, I. D. Dinov. Sharing big biomedical data. en. *J Big Data* **June 2015**, 2.

[35] T. Slater. Recent advances in modeling languages for pathway maps and computable biological networks. en. *Drug Discov. Today* **Feb. 2014**, 19 (2), 193–198.

[36] M. Heiner, D. Gilbert. BioModel engineering for multiscale Systems Biology. en. *Prog. Biophys. Mol. Biol.* **Apr. 2013**, 111 (2-3), 119–128.

[37] M Hucka, A Finney, H. M. Sauro, H Bolouri, J. C. Doyle, H Kitano, A. P. Arkin, B. J. Bornstein, D Bray, A Cornish-Bowden, A. A. Cuellar, S Dronov, E. D. Gilles, M Ginkel, V Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A Kremling, U Kummer, N Le Novère, L. M. Loew, D Lucio, P Mendes, E Minch, E. D. Mjolsness, Y Nakayama, M. R. Nelson, P. F. Nielsen, T Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J Stelling, K Takahashi, M Tomita, J Wagner, J Wang, SBML Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. en. *Bioinformatics* **Mar. 2003**, 19 (4), 524–531.

[38] H. Resat, L. Petzold, M. F. Pettigrew. Kinetic modeling of biological systems. en. *Methods Mol. Biol.* **2009**, 541, 311–335.

[39] O. Ruebenacker, I. I. Moraru, J. C. Schaff, M. L. Blinov. Kinetic Modeling using BioPAX ontology. en. *Proceedings* **Nov. 2007**, 2007, 339–348.

[40] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, J. Luciano, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, K. Kandasamy, A. C. Lopez-Fuentes, H. Mi, E. Pichler, I. Rodchenkov, A. Splendiani, S. Tkachev, J. Zucker, G. Gopinath, H. Rajasimha, R. Ramakrishnan, I. Shah, M. Syed, N. Anwar, O. Babur, M. Blinov, E. Brauner, D. Corwin, S. Donaldson, F. Gibbons, R. Goldberg, P. Hornbeck, A. Luna, P. Murray-Rust, E. Neumann, O. Ruebenacker, M. Samwald, M. van Iersel, S. Wimalaratne, K. Allen, B. Braun, M. Whirl-Carrillo, K.-H. Cheung, K. Dahlquist, A. Finney, M. Gillespie, E. Glass, L. Gong, R. Haw, M. Honig, O. Hubaut, D. Kane, S. Krupa, M. Kutmon, J. Leonard,

D. Marks, D. Merberg, V. Petri, A. Pico, D. Ravenscroft, L. Ren, N. Shah, M. Sunshine, R. Tang, R. Whaley, S. Letovksy, K. H. Buetow, A. Rzhetsky, V. Schachter, B. S. Sobral, U. Dogrusoz, S. McWeeney, M. Aladjem, E. Birney, J. Collado-Vides, S. Goto, M. Hucka, N. Le Novère, N. Maltsev, A. Pandey, P. Thomas, E. Wingender, P. D. Karp, C. Sander, G. D. Bader. The BioPAX community standard for pathway data sharing. en. *Nat. Biotechnol.* **Sept. 2010**, 28 (9), 935–942.

[41] C. A. Lieu, K. O. Elliston. Applying a causal framework to system modeling. en. *Ernst Schering Res. Found. Workshop* **2007**, (61), 139–152.

[42] C. T. Hoyt, D. Domingo-Fernández, S. Mubeen, J. M. Llaó, A. Konotopez, C. Ebeling, C. Birkenbihl, Ö. Muslu, B. English, S. Müller, M. P. de Lacerda, M. Ali, S. Colby, D. Türei, N. Palacio-Escat, M. Hofmann-Apitius. *Integration of Structured Biological Data Sources using Biological Expression Language*.

[43] N. L. Catlett, A. J. Bargnesi, S. Ungerer, T. Seagaran, W. Ladd, K. O. Elliston, D. Pratt. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. en. *BMC Bioinformatics* **Nov. 2013**, 14, 340.

[44] F. Martin, T. M. Thomson, A. Sewer, D. A. Drubin, C. Mathis, D. Weisensee, D. Pratt, J. Hoeng, M. C. Peitsch. Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. en. *BMC Syst. Biol.* **May 2012**, 6, 54.

[45] A. T. Kodamullil, E. Younesi, M. Naz, S. Bagewadi, M. Hofmann-Apitius. Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. en. *Alzheimers. Dement.* **Nov. 2015**, 11 (11), 1329–1339.

[46] S. Boué, M. Talikka, J. W. Westra, W. Hayes, A. Di Fabio, J. Park, W. K. Schlage, A. Sewer, B. Fields, S. Ansari, F. Martin, E. Veljkovic, R. Kenney, M. C. Peitsch, J. Hoeng. Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. en. *Database* **Apr. 2015**, 2015, bav030.

[47] A. L. Samuel. *Some Studies in Machine Learning Using the Game of Checkers*. **1959**.

[48] T. Mitchell. *Machine learning meets natural language (Abstract)*. **1997**.

[49] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, G. E. Robinson. Big Data: Astronomical or Genomical? en. *PLoS Biol.* **July 2015**, 13 (7), e1002195.

[50] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, P. Ping. Machine Learning and Integrative Analysis of Biomedical Big Data. en. *Genes* **Jan. 2019**, 10 (2).

[51] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, S. Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. en. *Nat. Med.* **June 2019**, 25 (6), 954–961.

[52] Y. Liu, T. Kohlberger, M. Norouzi, G. E. Dahl, J. L. Smith, A. Mohtashamian, N. Olson, L. H. Peng, J. D. Hipp, M. C. Stumpe. Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. en. *Arch. Pathol. Lab. Med.* **July 2019**, 143 (7), 859–868.

[53] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. en. *Nature* **Feb. 2017**, 542 (7639), 115–118.

[54] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, D. R. Webster. *Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs*. **2016**.

[55] H. Asri, H. Mousannif, H. Al Moatassime, T. Noel. *Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis*. **2016**.

[56] S. J. Kim, K. J. Cho, S. Oh. Development of machine learning models for diagnosis of glaucoma. en. *PLoS One* **May 2017**, 12 (5), e0177726.

[57] E. Abdulhay, N Arunkumar, K. Narasimhan, E. Vellaiappan, V Venkatraman. *Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease*. **2018**.

[58] J. Jeon, S. Nim, J. Teyra, A. Datti, J. L. Wrana, S. S. Sidhu, J. Moffat, P. M. Kim. *A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening*. **2014**.

[59] E. Ferrero, I. Dunham, P. Sanseau. *In silico prediction of novel therapeutic targets using gene–disease association data*. **2017**.

[60] R Das, F Ou, C Washburn, F Innocenti, A Nixon, H Lenz, C Blanke, D Niedzwiecki, I Khalil, B Harms, A Venook. *Bayesian machine learning on CALGB/SWOG 80405 (Alliance) and PEAK data identify a heterogeneous landscape of clinical predictors of overall survival (OS) in different populations of metastatic colorectal cancer (mCRC)*. **2019**.

[61] S. Riniker, Y. Wang, J. L. Jenkins, G. A. Landrum. *Using Information from Historical High-Throughput Screens to Predict Active Compounds*. **2014**.

[62] P. Mamoshina, M. Volosnikova, I. V. Ozerov, E. Putin, E. Skibina, F. Cortese, A. Zhavoronkov. Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. en. *Front. Genet.* **July 2018**, 9, 242.

[63] S. Smith. *43 Pharma Companies Using Artificial Intelligence in Drug Discovery*. `https://blog.benchsci.com/pharma-companies-using-artificial-intelligence-in-drug-discovery`. Accessed: 2020-2-27.

[64] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, A. Tsirigos. *Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning*. **2018**.

[65] D. Ramazzotti, A. Lal, B. Wang, S. Batzoglou, A. Sidow. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. en. *Nat. Commun.* **Oct. 2018**, 9 (1), 4453.

[66] M. Ceccarelli, F. P. Barthel, T. M. Malta, T. S. Sabedot, S. R. Salama, B. A. Murray, O. Morozova, Y. Newton, A. Radenbaugh, S. M. Pagnotta, S. Anjum, J. Wang, G. Manyam, P. Zoppoli, S. Ling, A. A. Rao, M. Grifford, A. D. Cherniack, H. Zhang, L. Poisson, C. G. Carlotti Jr, D. P. d. C. Tirapelli, A. Rao, T. Mikkelsen, C. C. Lau, W. K. A. Yung, R. Rabadan, J. Huse, D. J. Brat, N. L. Lehman, J. S. Barnholtz-Sloan, S. Zheng, K. Hess, G. Rao, M. Meyerson, R. Beroukhim, L. Cooper, R. Akbani, M. Wrensch, D. Haussler, K. D. Aldape, P. W. Laird, D. H. Gutmann, TCGA Research Network, H. Noushmehr, A. Iavarone, R. G. W. Verhaak. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. en. *Cell* **Jan. 2016**, 164 (3), 550–563.

[67] C Krittanawong. The rise of artificial intelligence and the uncertain future for physicians. en. *Eur. J. Intern. Med.* **Feb. 2018**, 48, e13–e14.

[68] J. Fan, F. Han, H. Liu. *Challenges of Big Data analysis*. **2014**.

[69] Y. Li, F.-X. Wu, A. Ngom. *A review on machine learning principles for multi-view biological data integration*. **2016**.

[70] M. K. Yu, J. Ma, J. Fisher, J. F. Kreisberg, B. J. Raphael, T. Ideker. *Visible Machine Learning for Biomedicine*. **2018**.

[71] H. Fröhlich, R. Balling, N. Beerenwinkel, O. Kohlbacher, S. Kumar, T. Lengauer, M. H. Maathuis, Y. Moreau, S. A. Murphy, T. M. Przytycka, M. Rebhan, H. Röst, A. Schuppert, M. Schwab, R. Spang, D. Stekhoven, J. Sun, A. Weber, D. Ziemek, B. Zupan. From hype to reality: data science enabling personalized medicine. en. *BMC Med.* **Aug. 2018**, 16 (1), 1–15.

[72] Y. Feinstein, J. C. Walker, M. J. Peters, S. Nadel, N. Pathan, N. Edmonds, J. Herberg, M. Kaforou, V. Wright, M. Levin, P. Ramnarayan. Cohort profile of the Biomarkers of Acute Serious Illness in Children (BASIC) study: a prospective multicentre cohort study in critically ill children. en. *BMJ Open* **Nov. 2018**, 8 (11), e024729.

[73] I. Lohse, K. Statz-Geary, S. P. Brothers, C. Wahlestedt. Precision medicine in the treatment stratification of AML patients: challenges and progress. en. *Oncotarget* **Dec. 2018**, 9 (102), 37790–37797.

[74] T. Hulsen, S. S. Jamuar, A. R. Moody, J. H. Karnes, O. Varga, S. Hedensted, R. Spreafico, D. A. Hafler, E. F. McKinney. From Big Data to Precision Medicine. en. *Front. Med.* **2019**, 6.

[75] A. D. Hingorani, D. A. v. d. Windt, R. D. Riley, K. Abrams, K. G. M. Moons, E. W. Steyerberg, S. Schroter, W. Sauerbrei, D. G. Altman, H. Hemingway, PROGRESS Group. Prognosis research strategy (PROGRESS) 4: stratified medicine research. en. *BMJ* **Feb. 2013**, 346, e5793.

[76] P. B. Chapman, A. Hauschild, C. Robert, J. B. Haanen, P. Ascierto, J. Larkin, R. Dummer, C. Garbe, A. Testori, M. Maio, D. Hogg, P. Lorigan, C. Lebbe, T. Jouary, D. Schadendorf, A. Ribas, S. J. O'Day, J. A. Sosman, J. M. Kirkwood, A. M. M. Eggermont, B. Dreno, K. Nolop, J. Li, B. Nelson, J. Hou, R. J. Lee, K. T. Flaherty, G. A. McArthur, BRIM-3 Study Group. Improved survival with vemurafenib in

melanoma with BRAF V600E mutation. en. *N. Engl. J. Med.* **June 2011**, 364 (26), 2507–2516.

[77] S. T. Rosen. *Why Precision Medicine Continues to Be the Future of Health Care*. **2017**.

[78] A. Ahmad. Dissecting Patient Heterogeneity Via Statistical Modeling Based on Multi-modal Omics Data. en. **2019**.

[79] J. Xu, P. Yang, S. Xue, B. Sharma, M. Sanchez-Martin, F. Wang, K. A. Beaty, E. Dehan, B. Parikh. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. en. *Hum. Genet.* **Feb. 2019**, 138 (2), 109–124.

[80] E. I. Dumbrava, F. Meric-Bernstam. *Personalized cancer therapy—leveraging a knowledge base for clinical decision-making*. **2018**.

[81] R. K. Sevakula, V. Singh, N. K. Verma, C. Kumar, Y. Cui. Transfer Learning for Molecular Cancer Classification Using Deep Neural Networks. en. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **Nov. 2019**, 16 (6), 2089–2100.

[82] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. A. Finkelstein, E. L. Lamoureux, I. Y. Wong, N. M. Bressler, S. Sivaprasad, R. Varma, J. B. Jonas, M. G. He, C.-Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee, T. Y. Wong. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. en. *JAMA* **Dec. 2017**, 318 (22), 2211–2223.

[83] J. de Jong, M. A. Emon, P. Wu, R. Karki, M. Sood, P. Godard, A. Ahmad, H. Vrooman, M. Hofmann-Apitius, H. Fröhlich. Deep learning for clustering of multivariate clinical patient trajectories with missing values. en. *Gigascience* **Nov. 2019**, 8 (11).

[84] C. Lopez, S. Tucker, T. Salameh, C. Tucker. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. en. *J. Biomed. Inform.* **Sept. 2018**, 85, 30–39.

[85] G. Tosto, S. E. Monsell, S. E. Hawes, G. Bruno, R. Mayeux. *Progression of Extrapyramidal Signs in Alzheimer's Disease: Clinical and Neuropathological Correlates*. **2015**.

[86] D. Gamberger, The Alzheimer's Disease Neuroimaging Initiative, B. Ženko, A. Mitelpunkt, N. Lavrač. *Homogeneous clusters of Alzheimer's disease patient population*. **2016**.

[87] D. Zeiberg, T. Prahlad, B. K. Nallamothu, T. J. Iwashyna, J. Wiens, M. W. Sjoding. Machine learning for patient risk stratification for acute respiratory distress syndrome. en. *PLoS One* **Mar. 2019**, 14 (3), e0214465.

[88] Y. Cun, H. Fröhlich. *Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions*. **2012**.

[89] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-ud din, P. Hintsanen, S. A. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, NCI DREAM Community, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J. W. Gray, G. Stolovitzky. A community effort to assess and improve drug sensitivity prediction algorithms. en. *Nat. Biotechnol.* **Dec. 2014**, 32 (12), 1202–1212.

[90] M. Hofree, J. P. Shen, H. Carter, A. Gross, T. Ideker. Network-based stratification of tumor mutations. en. *Nat. Methods* **Nov. 2013**, 10 (11), 1108–1115.

[91] Y.-H. Chang, C.-M. Chen, H.-Y. Chen, P.-C. Yang. Pathway-based gene signatures predicting clinical outcome of lung adenocarcinoma. en. *Sci. Rep.* **June 2015**, 5, 10979.

[92] M. Manica, J. Cadow, R. Mathis, M. Rodríguez Martínez. PIMKL: Pathway-Induced Multiple Kernel Learning. en. *NPJ Syst Biol Appl* **Mar. 2019**, 5, 8.

[93] S. H. Sleigh, C. L. Barton. *Repurposing Strategies for Therapeutics*. **2010**.

[94] M. Dickson, J. P. Gagnon. The cost of new drug discovery and development. en. *Discov. Med.* **June 2004**, 4 (22), 172–179.

[95] T. T. Ashburn, K. B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. en. *Nat. Rev. Drug Discov.* **Aug. 2004**, 3 (8), 673–683.

[96] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, Z. Lu. A survey of current trends in computational drug repositioning. *Brief. Bioinform.* **Jan. 2016**, 17 (1), 2–12.

[97] A. I. Graul, P Pina, M Stringer. *The years new drugs biologics 2017: Part I*. **2018**.

[98] P. Sanseau, J. Koehler. Editorial: Computational methods for drug repurposing. *Brief. Bioinform.* **July 2011**, 12 (4), 301–302.

[99] P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J Brent Richards, L. R. Cardon, V. Mooser. Use of genome-wide association studies for drug repositioning. en. *Nat. Biotechnol.* **Apr. 2012**, 30 (4), 317–320.

[100] H. Luo, J. Chen, L. Shi, M. Mikailov, H. Zhu, K. Wang, L. He, L. Yang. DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical–protein interactome. *Nucleic Acids Res.* **July 2011**, 39 (suppl_2), W492–W498.

[101] H. S. Lee, T. Bae, J.-H. Lee, D. G. Kim, Y. S. Oh, Y. Jang, J.-T. Kim, J.-J. Lee, A. Innocenti, C. T. Supuran, L. Chen, K. Rho, S. Kim. Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. en. *BMC Syst. Biol.* **July 2012**, 6 (1), 1–10.

[102] L. Yang, P. Agarwal. Systematic Drug Repositioning Based on Clinical Side-Effects. *PLoS One* **Dec. 2011**, 6 (12), e28025.

[103] *UK Biobank*. `https://www.ukbiobank.ac.uk/`. Accessed: 2020-3-21.

[104] *EMIF*. `http://www.emif.eu/`. Accessed: 2020-3-21.

[105] J. T. Dudley, T. Deshpande, A. J. Butte. Exploiting drug–disease relationships for computational drug repositioning. *Brief. Bioinform.* **July 2011**, 12 (4), 303–311.

[106] Zonisamide has beneficial effects on Parkinson's disease patients. *Neurosci. Res.* **Dec. 2001**, 41 (4), 397–399.

[107] pubmeddev. *Home - PubMed - NCBI*. `https://www.ncbi.nlm.nih.gov/pubmed/`. Accessed: 2020-3-22.

[108] *OMIM - Online Mendelian Inheritance in Man*. `https://omim.org/`. Accessed: 2020-3-22.

[109] *DrugBank*. `https://www.drugbank.ca/`. Accessed: 2020-3-22.

[110] *Therapeutic Target Database (TTD)*. `http://db.idrblab.net/ttd/`. Accessed: 2020-3-22.

[111] M. Zhang, G. Schmitt-Ulms, C. Sato, Z. Xi, Y. Zhang, Y. Zhou, P. St George-Hyslop, E. Rogaeva. Drug Repositioning for Alzheimer's Disease Based on Systematic 'omics' Data Mining. *PLoS One* **Dec. 2016**, 11 (12), e0168812.

[112] H. Xie, H. Wen, M. Qin, J. Xia, D. Zhang, L. Liu, B. Liu, Q. Liu, Q. Jin, X. Chen. *In silico drug repositioning for the treatment of Alzheimer's disease using molecular docking and gene expression data*. **2016**.

[113] S. Jamal, S. Goyal, A. Shanker, A. Grover. Checking the STEP-Associated Trafficking and Internalization of Glutamate Receptors for Reduced Cognitive Deficits: A Machine Learning Approach-Based Cheminformatics Study and Its Application for Drug Repurposing. en. *PLoS One* **June 2015**, 10 (6), e0129370.

[114] D. Romeo-Guitart, J. Forés, M. Herrando-Grabulosa, R. Valls, T. Leiva-Rodríguez, E. Galea, F. González-Pérez, X. Navarro, V. Petegnief, A. Bosch, M. Coma, J. M. Mas, C. Casas. Neuroprotective Drug for Nerve Trauma Revealed Using Artificial Intelligence. en. *Sci. Rep.* **Jan. 2018**, 8 (1), 1879.

[115] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, F. Cheng. deepDR: a network-based deep learning approach to in silico drug repositioning. en. *Bioinformatics* **Dec. 2019**, 35 (24), 5191–5198.

[116] Y. Y. Li, S. J. M. Jones. *Drug repositioning for personalized medicine*. **2012**.

[117] A. Talevi, C. L. Bellera. Challenges and opportunities with drug repurposing: finding strategies to find alternative uses of therapeutics. en. *Expert Opin. Drug Discov.* **Apr. 2020**, 15 (4), 397–401.

[118] *Complex Diseases: Research and Applications | Learn Science at Scitable*. https://www.nature.com/scitable/topicpage/complex-diseases-research-and-applications-748/. Accessed: 2020-3-24.

[119] *SJR Subject Bubble Chart*. https://www.scimagojr.com/mapgen.php?maptype=bc&country=US. Accessed: 2020-3-22.

[120] WIRED Staff. *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. https://www.wired.com/2008/06/pb-theory/. Accessed: 2020-3-23. **June 2008**.

[121] S. Leonelli. What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. en. *Big Data Soc* **June 2014**, 1 (1).

[122] F. Mazzocchi. Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. en. *EMBO Rep.* **Oct. 2015**, 16 (10), 1250–1255.

[123] D. G. Hernandez, M. A. Nalls, M. Moore, S. Chong, A. Dillman, D. Trabzuni, J. R. Gibbs, M. Ryten, S. Arepalli, M. E. Weale, A. B. Zonderman, J. Troncoso, R. O'Brien, R. Walker, C. Smith, S. Bandinelli, B. J. Traynor, J. Hardy, A. B. Singleton, M. R. Cookson. Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. en. *Neurobiol. Dis.* **July 2012**, 47 (1), 20–28.

[124] T. Wilholt. *Bias and values in scientific research*. **2009**.