October 2018

Newsletter Issue 5

AETIONOMY –

Organising Knowledge about Neurodegenerative Disease Mechanisms for the Improvement of Drug Development and Therapy

IN THIS ISSUE

Project Coordinators:
UCB Pharma, Fraunhofer SCAI

Contact persons:
Dr. Phil Scordis
Prof. Martin Hofmann-Apitius

www.aetionomy.org

imi innovative medicines initiative

AETIONOMY

EDITORIAL

TOWARDS A MECHANISM-BASED TAXONOMY OF ALZHEIMER'S AND PARKINSON'S DISEASE

AETIONOMY is a consortium brought together under the European Innovative Medicines Initiative to tackle the problem of the classification of neurodegenerative diseases. The diagnosis of Alzheimer dementia (AD) is based on a relatively nonspecific phenotype: progressive memory impairment associated with a pathological finding of amyloid plaques and neurofibrillary tangles. Similarly, for Parkinson disease (PD), the phenotype is a progressive movement disorder that is associated with damage to the substantia nigra and the presence of Lewy bodies. In both conditions, single-gene familial forms have highlighted heterogeneous biological pathways resulting in identical disease phenotypes. These pathways are highly likely to be involved in the sporadic forms of the disease, but their exact role and the molecularly driven subgroupings must be identified if we are to make progress in developing new therapies for these forms. AETIONOMY is systematically collecting publicly available and proprietary data. A semantic framework – a formalized representation of the essential knowledge on neurodegenerative diseases that is both computer-readable and understandable by humans – forms the backbone for all data retrieval and annotation. It will be mined to identify new molecularly defined subgroups of AD/PD patients.

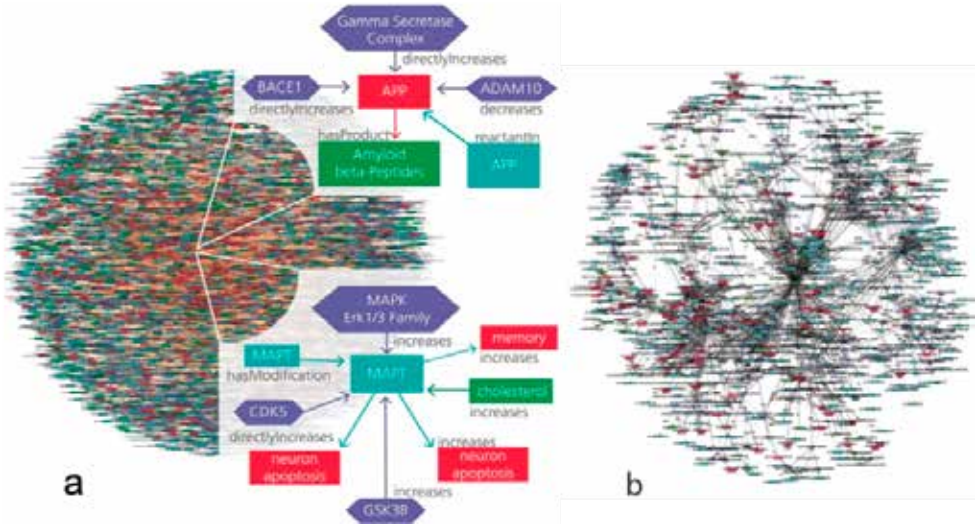
As we do not believe that there is one path or one single modelling approach that is suited to deliver the candidate mechanisms that form the basis for the new taxonomy, we are applying different modelling strategies. A wide spectrum of mining strategies is supported by the AETIONOMY knowledge base, including causal reasoning (based on OpenBEL), graph mining and association mining via pathophysiology graphs. Candidate mechanisms that are causally involved in the aetiology of the disease and potentially useful as classification tools are identified by applying these mining approaches. Finally, AETIONOMY will validate a selection of candidate mechanisms (that bear the potential to establish a taxon in the new mechanism-based taxonomy), supported by our clinical study.

The Consortium is jointly led by Doctor Phil Scordis from the biopharmaceutical company UCB Pharma SPRL, and Professor Martin Hofmann-Apitius, Fraunhofer Institute SCAI.

DISEASE MAPS

COLLECTING KNOWLEDGE FROM LITERATURE AND DATA TO MAP THE KNOWLEDGE ABOUT ALZHEIMER'S AND PARKINSON'S DISEASE

Scientists often describe biological processes in the form of pathways and chains of interactions between molecules. Most of this information is hidden in the unstructured text of scientific publications and need to be organized into computable models. Therefore scientific knowledge of physiological functions and pathological actions in disease were acquired from disease-related articles, reviews, and databases. Fraunhofer's literature mining system 'SCAIView Neuro' was used to retrieve a list of genes, reported to be linked to a pathology, of which the top 100 genes were selected based on their relevancy to the query. Documents tagged for these genes were manually filtered for normal and disease states. Furthermore, In case of Alzheimer's, documents related to top 10 AD related genes were obtained from the AlzGene Database. Additional documents are collected from the databases such as KEGG, Reactome and BioCarta, where the references for the each disease related pathway are used to extract knowledge. Manual curations and extractions of statements and their coding in the Biological Expression Language (BEL) is used to extract information and knowledge. Biological entities (indicated as subjects or objects) and relationships (as causal chains) reported in these scientific articles were encoded and . manually reviewed by experts. As a result of these approach and efforts Fraunhofer generated disease models for AD and PD. The following figures shows the AD BEL disease vs. healthy model:

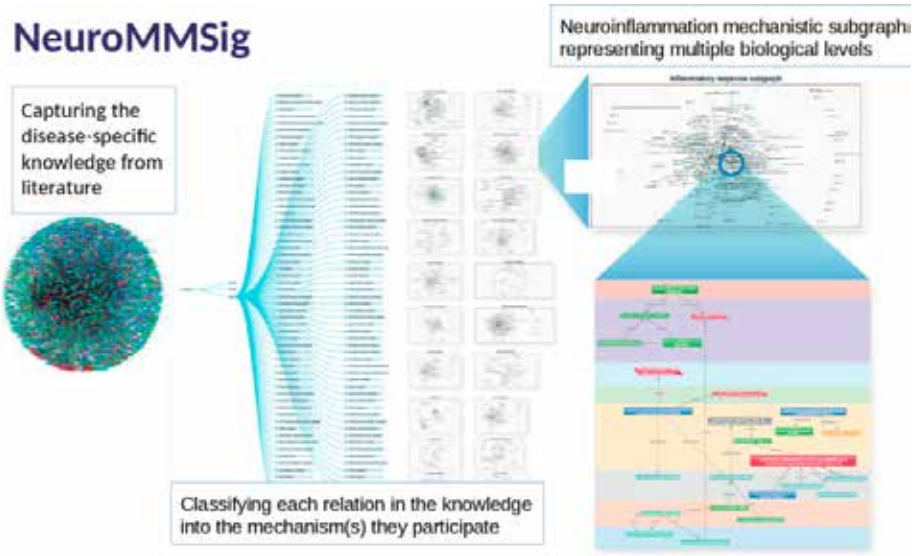


- *The **AD BEL model** generated by Dr. Alpha Tom Kodamullil and team: 35.266 citations and 44.437 BEL statements => 9.645 nodes and 10.251 edges.*
- *The **PD BEL model** generated by Reagon Karki and team: 432 citations and 2.236 BEL statements => 1.424 nodes and 2.690 edges.*

MULTIMODAL MECHANISTIC SIGNATURES

NEUROMMSIG – INVENTORY OF NEURODEGENERATIVE DISEASE MECHANISMS

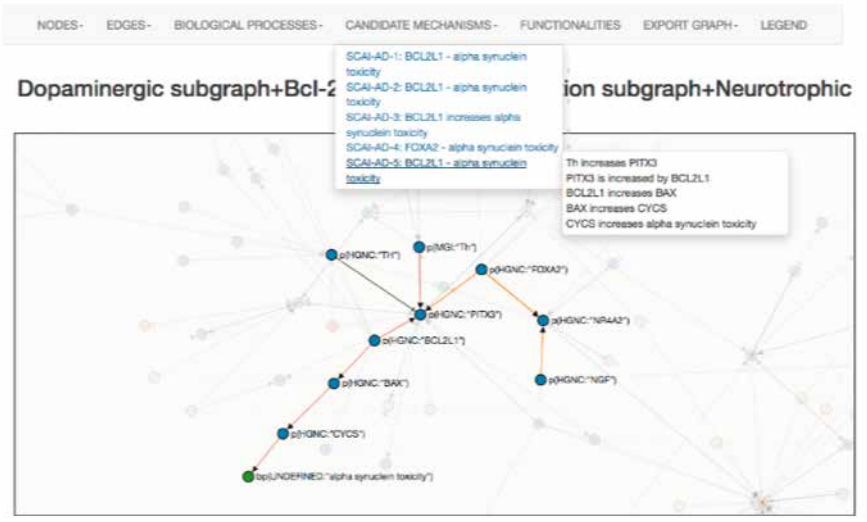
As described published knowledge is organized into computable models. This is the basis for the essential activity in AETIONOMY to generate hypotheses about multiscale mechanisms of neuro-degenerative pathophysiology. Conceptually, we identified and organized disease specific features, at different scales, to perform data-driven analysis. This analysis serves to identify robust combinations of features that correspond to disease subtypes. The mechanisms of neurodegenerative pathophysiology, that distinguish the disease subtypes – referred to as our hypotheses – will be tested, iteratively elaborated and validated through data generated by apposite patient studies. In order to retrieve the main mechanisms involved in these neurodegenerative disorders, a list of pathways and mechanistic knowledge was extracted from PubMed using Fraunhofer's information retrieval system 'SCAIView Neuro'. This list was preprocessed and curated due to the large number of synonyms found in the literature leading to a final inventory of pathways and mechanisms, that served as a guideline for annotating each individual statement (triplet/assertion) in the disease models (Biological Expression Language - BEL). We also emphasized inclusion of all well-known mechanisms (e.g., amyloid cascade, neuro-inflammation, mitochondrial dysfunction, …) as entries to our mechanism repository 'NeuroMMSig'. The next step was to individually annotate and evaluate all of the triplets in the models with their respective candidate mechanisms. During this process, we performed literature and database searches in order to find out, to which candidate mechanism the entities in each BEL statement belonged. This whole process results in a repository of computable disease specific mechanisms (126 for AD and 76 for PD) as shown below:



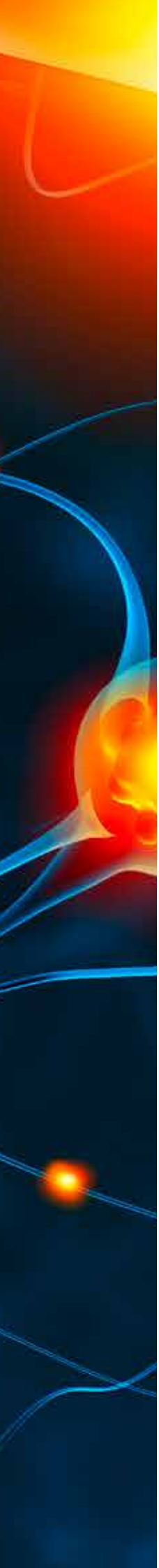
Multimodal data is necessary in order to map biological entities to the clinical studies in neurodegeneration since they contain not only genetic markers, but variables from brain scans to neuro-psychological assessments. Conventional pathway analysis tools such as Gene Set Enrichment Analysis (GSEA)/Molecular signatures (MSIG), are limited to the molecular gene and in particular gene expression layer. In contrast, in our approach NeuroMMSig entries were enriched with imaging features, variant information Single Nucleotide Polymorphism (SNPs), miRNA, clinical studies, and drugs/chemicals, making them essentially multiscale and multimodal representations of candidate mechanisms. The complexity of mechanistic information represented enables NeuroMMSig to accept not only molecular (e.g., gene expression) information. As a consequence, the approach taken with NeuroMMSig is overcoming several of the limitations associated with conventional pathway analysis tools. In summary, NeuroMMSig comprises a candidate mechanism collection from the major neurological disorders, represents a high resolution and curated knowledge base incl. candidate mechanisms converted to computable networks (graphs). The procedure described above to build a repository of disease specific mechanisms was achieved within approximately 1 year of work for the AD model and 6 months of work for the PD due to their size. During this process, database models were created including, for instance, which entities were assigned to candidate mechanisms and other multimodal enrichment data.

NeuroMMSig Biological process example

Users should select candidate nodes based on their interest. From all data-mapped nodes to this selected node, candidate mechanisms (represented as a chain of causation or paths) are displayed in the „Candidate mechanism“ tab. Clicking in each candidate mechanism, allows you to navigate and visualize it in the interface. An example of how candidate mechanism are displayed is shown below:



This example has been generated by submitting a gene set of FOXA2, TH, BCL2L1, NGF in the context of Parkinson's. In the visualization site, „alpha synuclein toxicity“ was selected as a biological process. Here, PITX3 is suggested to be the key player in this particular mechanism.



NeuroMMSig Derived Hypotheses

AETIONOMY disease hypotheses to be tested are represented as networks and stored in NeuroMMSig. The system is developed to enable patient subgroup stratification based on multimodal and multiscale patterns that indicate a perturbation of mechanisms. A small description of some of the proposed mechanisms is depicted in the Table on the right. The process of neuroinflammation and the immune system are involved in the pathology of both, AD and PD (Table 1). In fact, there are currently other IMI projects, such as the PHAGO project, trying to target key players in AD within these biological processes. For that reason, NeuroMMSig has been enriched with mechanistic subgraphs related to these two processes such as chemokine signaling, cytokine signaling, interferon signaling, toll like receptor, inflammatory response, and immune system response subgraphs. These subgraphs contain biomarkers selected from WP5 such as YKL-40, TLR4, and MRP14. Having these biomarkers in the subgraphs will allow testing the generated hypothesis once the clinical studies have been carried out. The clinical measurements can be mapped to nodes in the networks calculating a score for each patient enabling patient subgroup identification. Since NeuroMMSig is inherently multimodal, not only the biomarkers will be mapped but also other indices like imaging features or metabolites.

AD specific hypotheses	PD specific hypotheses
Syndecan-mediated uptake (heparan sulfate proteoglycan (HSPG) – mediated uptake) hypothesis; related to endocytosis processes	LRRK2 (most relevant SNP found on literature)
KANSL1 and the corticotropin-releasing hormone receptor – related "shared mechanism" identified by genetics and imaging analysis on chromosome17	Epigenetics (SNCA methylation in the CNS) PDE4D biomarker
AD-diabetes comorbidity Insulin signaling crosstalk to major AD pathophysiology mechanisms	Mitochondrial dysfunction
Neuro-inflammation	Neuro-inflammation

The use of knowledge graphs representing pathophysiology mechanisms for the stratification of patient subgroups is a non-trivial undertaking. Whereas the clustering of clinical data can identify patterns of clinical readouts, that can be tested in independent clinical data sets for their ability to stratify patients according to the identified pattern, a mechanism candidate needs first to be mapped to variables in clinical data and the significance of the values for the mapped variables needs to be estimated or calculated (e.g., based on thresholds). In the case of discrete variables (e.g., SNPs), the absence or presence of a SNP can be scored. SNPs are likely to be the most frequently used variables to be mapped, as they are routinely measured in research cohorts such as ADNI and PPMI and they are widely used to strategy patient (risk) subgroups. As single SNPs may not be directly "mappable" (because e.g., SNPs linked to mechanisms have not been measured in a study cohort due to different technology platforms for SNP detection), methods for the assignment of SNPs to loci have to be applied. NeuroMMSigDB entries come with a LD-block annotation, which allows for definition of loci and a mapping of SNPs in NeuroMMSigDB mechanisms to SNPs measured in cohorts via LD-blocks.
Some NeuroMMSigDB entries comprise disease stage annotations and such association can be used as a partitioning concept (which, however, does not go beyond the diagnosis of the clinical experts recruiting the patients in the cohort). However, if combined with other modalities (SNPs, imaging readouts), the stage-specific assignment and the mechanistic context may gain an explanatory potential that would trigger more in-depth analysis of that mechanism and its role in stage-specific phenotypes (e.g., certain neuro-psychological assessments; progression patterns; biomarker trajectories). We expect to get more insights in the possible mapping of candidate mechanisms to disease stages during the validation against independent cohort data.

DISEASE HYPOTHESES

CHOSEN CANDIDATE MECHANISMS AND ANALYSES

An essential activity in AETIONOMY is to generate hypotheses about multiscale mechanisms of neurodegenerative pathophysiology. Conceptually, we identify and organize disease specific features, at different scales, to perform data-driven analysis. This analysis serves to identify robust combinations of features that correspond to disease subtypes. The mechanisms of neurodegenerative pathophysiology, that distinguish the disease subtypes – referred to as our hypotheses – are being tested, iteratively elaborated and validated, through data generated by apposite patient studies. On a more strategic and methodological level, the overarching goal is therefore to develop an investigative and procedural blueprint that is generally applicable to disease taxonomy efforts also in other therapeutic areas.

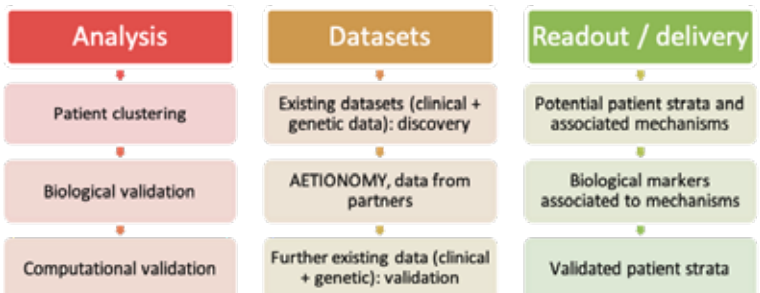
NeuroMMSig characterizes more then 200 candidate mechanisms for both the diseases for Alzheimer's and Parkinson's disease including the AETIONOMY in-silico computed and clinical hypotheses, which are listed in the following enumeration showing also overlaps between AD and PD:

1. Neuroinflammation (in AD, UKB) => MRP14, MRP8
2. Astroglialinflammation (in AD, IDIBAPS/BBRC) => YKL-40 (CHI3L1)
3. Syndecans uptake of aggregate proteins in neurodegenerative diseases
   => AD: GSK3, APOE, DPYSL2, MAP1B, PTP1B, RTN1. STMN1.
   PD: HNRNPK, PARK7
4. Role of insulin pathway in AD and PD progression => STK11, INSR
5. Stress-induced comorbidity of AD and PD (Chromosome 17 locus)
   => CRH, CRHR1, MAPT, KANSL1
6. Cross-talk between mitochondrial dysfunction and neuroinflammation (in PD, ICM)
   => HSD17B10 and TFAM in CSF
7. SNCA methylation and PD progression (in PD, UKB) => SNCA

AD

PD

The analysis of candidate mechanisms follows the following plan:



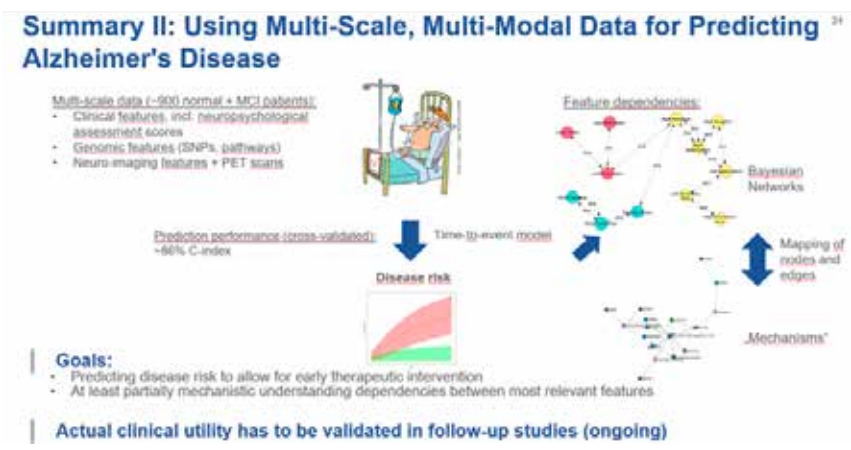
IN SILICO VALIDATION

LINKING MECHANISMS TO DISEASE RISK

We tried to better understand the role of NeuroMMSig mechanistic hypotheses in the context of the transition from cognitively normal or mild impaired stage to Alzheimer's Disease. For this purpose we developed a highly predictive machine learning model for pre-symptomatic patients, which scores the risk for an individual patient to transit to Alzheimer's Disease. In a second step we then linked the predictive factors in the model to NeuroMMSig mechanisms (Figure below).

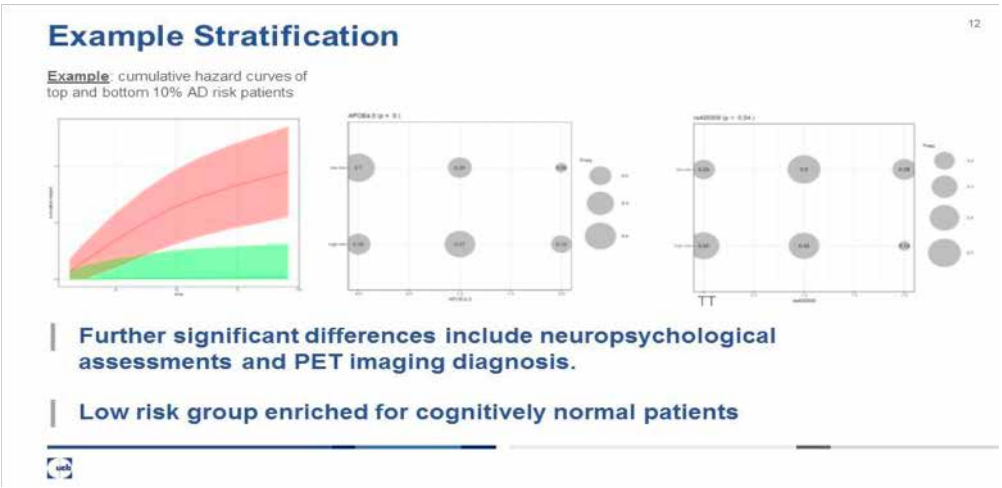
Where do we see the value? Early diagnosis of AD is essential for successful disease management and chance to attenuate symptoms by disease modifying drugs. In the past, a number of cerebrospinal fluid (CSF), plasma and neuro-imaging based biomarkers have been proposed. Still, in current clinical practice, AD diagnosis cannot be made until the patient shows clear signs of cognitive decline, which can partially be attributed to the multi-factorial nature of AD. Having a predictive model, which allows to assess an individual's risk to transit from a pre-symptomatic situation to AD could thus be of high relevance.

Our model integrated rich genotype information (including newly developed SNP functional pathway impact scores), neuro-imaging (volume measurements of brain regions, PET scan results) as well as clinical data from 900 normal and MCI individuals extracted from the Alzheimer' s Disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu/), a large scale observational study started in 2004 to evaluate the use of diverse types of biomarkers in clinical practice. A second aim of this work was to better understand the biological mechanisms driving the conversion of normal/MCI into AD pathology, which may ultimately open the door to novel therapeutic options. To this end, we employed a combination of data driven probabilistic and knowledge driven mechanistic approaches. More specifically, we used Bayesian Networks to uncover the interplay across biological scales between genetic variants, pathways, PET scan results and neuro-imaging related features. Together with manually curated cause-effect chains extracted from the literature, this allowed us to partially reconstruct biological mechanisms that could play a role in the conversion of normal/MCI into AD pathology.



PATIENT STRATIFICATION DERIVED FROM DISEASE RISK MODEL

AETIONOMY disease hypotheses to be tested are represented as networks and stored in NeuroMMSig. The system is developed to enable patient subgroup stratification based on multimodal and multiscale patterns that indicate a perturbation of mechanisms. A small description of some of the proposed mechanisms is depicted in the Table on the right. The process of neuroinflammation and the immune system are involved in the pathology of both, AD and PD (Table 1). In fact, there are currently other IMI projects, such as the PHAGO project, trying to target key players in AD within these biological processes. For that reason, NeuroMMSig has been enriched with mechanistic subgraphs related to these two processes such as chemokine signaling, cytokine signaling, interferon signaling, toll like receptor, inflammatory response, and immune system response subgraphs. These subgraphs contain biomarkers selected from WP5 such as YKL-40, TLR4, and MRP14. Having these biomarkers in the subgraphs will allow testing the generated hypothesis once the clinical studies have been carried out. The clinical measurements can be mapped to nodes in the networks calculating a score for each patient enabling patient subgroup identification. Since NeuroMMSig is inherently multimodal, not only the biomarkers will be mapped but also other indices like imaging features or metabolites.



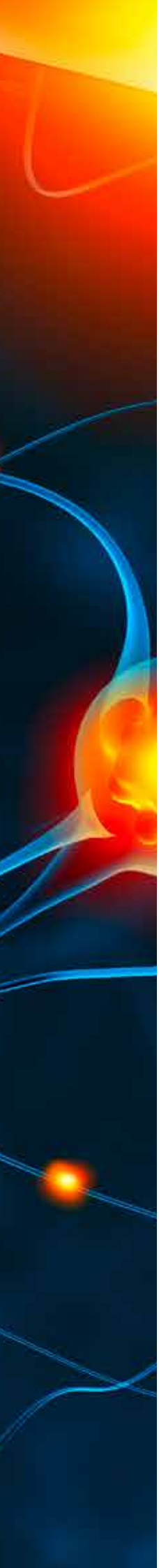
UNSUPERVISED JOINT AD/PD PATIENT CLUSTERING

AETIONOMY aims to establish a molecular disease taxonomy of neurodegenerative diseases. At its core this goal implies the existence of molecularly defined patient sub-groups, which could diverge from the current classification of neurodegenerative diseases. As outlined above, AETIONOMY has taken a knowledge driven approach to define AD and PD disease mechanisms. The question is whether these mechanisms can - possibly in combination - discriminate patient sub-groups and specifically help identifying mixed AD/PD subtypes. The latter would call for a substantial revision of the way, in which neurodegenerative diseases are understood at present.

In order to address these questions, partners UCB and Fraunhofer have established a data mining methodology to group AD and PD patients using SNP based genotypes and shared AD/PD mechanisms derived from BEL encoded knowledge graphs. This mechanism enhanced approach involves a mapping of SNPs to genes encoded in shared molecular mechanisms and dimensionality reduction (e.g. autoencoder networks) followed by clustering with mixture of autoencoders and sparse Non-Negative Matrix Factorization. Our method has been applied to a merged ADNI and PPMI dataset, which contains de novo AD/ PD patients and those, who converted into AD during the course of the study. Identified clusters were well separated, statistically stable and showed (after correction for age, ethnicity and gender effects) statistically significant differences w.r.t. clinical features in AD, such as inter-cranial volume measurements. The validation of the established grouping in comparison to genotypes of healthy controls, patients from the independent ROSMAP cohort (AD) and different PD studies (AETIONOMY PD, ICEBERG PD, DIGPD) is currently ongoing. Additional available omics data from ROSMAP (proteomics, DNA methylation, CHIPseq, gene expression) and AETIONOMY PD (proteomics, DNA methylation) will be used to understand differences between genotype based clusters and to provide a biological contextualization. We expect the work on joint AD / PD patient clustering to go on after the end of the funding for AETIONOMY in 2019.

UNSUPERVISED PD PATIENT CLUSTERING

Following an alternative approach partner ICM is currently developing a Non-Negative Matrix Trifactorization method to cluster PD patients in the DIGPD cohort based on genotype. The approach in particular considers the grouping of SNPs into genes, which fall into the same NeuroMMSig mechanism. Based on the developed method an initial grouping of PD patients has been established and is currently validated using different PD studies (AETIONOMY PD, ICEBERG PD). We expect this work to continue after the end of the funding period for AETIONOMY.



VIRTUAL DEMENTIA COHORT (VDC)

One major limiting factor for the generation of mechanism-based taxonomies is the accessibility and availability of relevant patient-level data. The challenge starts already at the level of study design and patient recruitment. Whereas patient recruitment and biomaterial sample collection is comparably simple and straightforward in the area of systematic autoimmune disorders (taking blood samples is a procedure commonly accepted by patients), the situation is fundamentally different in the case of neurodegenerative disease research. Spinal taps require patients to accept that a needle is inserted into their spine and CSF needs to be collected repeatedly in longitudinal studies aimed at monitoring progression. Furthermore, the fact that an "elevated biomarker level" may indicate an increased risk to develop a neurodegenerative disease without a real chance to treat that disease prevents both, potential patients, as well as healthy controls from enrolling in studies.

This shows already in the comparison between the two projects working on mechanism-based taxonomies: during the runtime of the project, AETIONOMY could only recruit a substantially smaller number of patients and only in one disease area (Parkinsonism), whereas PRECISESADS was able to recruit for all autoimmune diseases and in much larger numbers. The consequences for research on neurodegeneration are dramatic: the number of "referential studies" with sufficient statistical power is very limited. In essence, there are 4 major studies published for Alzheimer´s Disease: ADNI AddNeuroMed, AIBL, and Rosmap, the associated data can be accessed after authorization by the data owners (usually a committee of the consortium that runs the study). In the area of Parkinsonism, the most widely recognized study is the PPMI (Parkinson Progression Marker Initiative) study; the only study that may be comparable to PPMI (by both, size and longitudinal aspects) is the Oxford Parkinson Disease Center (OPDC) discovery cohort.

The strong dependency on ADNI in the Alzheimer area and PPMI in the field of Parkinsonism results in a strong publication bias when it comes to data-driven analyses using patient-level data. The ADNI consortium is co-author on more than 1100 publications; however, there is neither a systematic comparison between the major AD studies, nor is there an independent "validation" data set that has been generated completely independent from ADNI. Whether AddNeuroMed and AIBL could serve as such "independent validation data sets", remains to be shown. At least with respect to ethnicity, all these studies are heavily biased towards Caucasian haplotypes.

To overcome hurdles to access patient data for the in-silico validation of disease mechanisms and ultimately a first validation of a mechanism-based taxonomy of neurodegenerative diseases, we developed the concept of Virtual Dementia Cohorts (VDCs). VDCs are synthetic (artificial) data sets that share features and characteristics of real-world study cohorts in the area of neurodegenerative diseases. VDCs bear the potential to overcome some of the substantial challenges we face in translational neurodegeneration research, namely the:

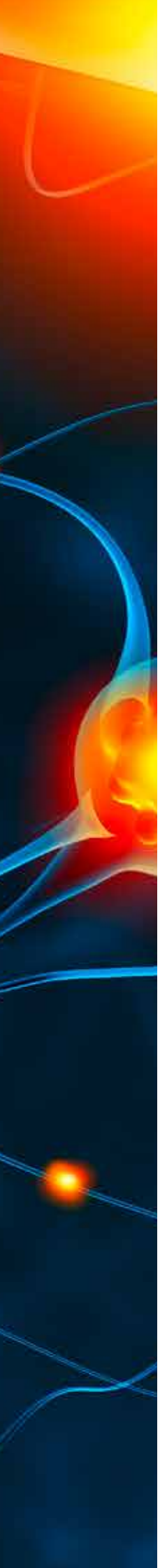
- sharing of patient-level data without compromising patient data privacy
- blending and merging of heterogeneous, complex clinical data sets
- increasing the number of virtual patients to match statistical requirements

- post-hoc enrichment of limited clinical data with additional features
- integration of data and knowledge
- ability to address counterfactual questions
- computing of "what-if" scenarios
- ability to simulate trials in silico
- ability to "play with data" in the sense that modelers and miners can test new methods

Like real-world clinical studies, VDCs may have a time dimension and can thus reflect disease progression. A fully validated VDC will generate patient trajectories, which resemble those observed in real patients. This includes biomarkers, e.g. CSF protein markers, as well as neuro-imaging related features such as volumes of different brain regions. A VDC therefore represents the multi-modal and multi-scale nature of neuro-degenerative diseases. This opens the opportunity to mine VDCs in the future: For example, we may want to use VDCs to cluster patients with respect to their disease progression, and characterize these groups with respect to non-common genetic variants. Due to the possibility to simulate as many virtual patients as desired there are no principal limitations of statistical power. Of course, findings derived from VDCs constitute only hypotheses and will require further validation using data from real patients. However, these validation studies could be much more focused than current approaches. For example, they could concentrate on testing only a handful of SNPs, which have been previously identified from a VDC analysis. One potential way to improve on the representativeness of studies is the "blending" with other, related studies and the "enrichment" with additional information from focused observations. Such enrichment and blending would result in a widening of the variable space describing subjects that may potentially develop signs and symptoms of neurodegenerative diseases over time and the enrichment of these variables with values specific for a wider spectrum of subjects (not only by ethnic background, but also by lifestyle, education, nutrition etc.) would be highly desirable in order to be able to generalize findings from major studies such as ADNI or PPMI.

SIMULATING VIRTUAL BRAIN CONNECTOMES

Research in structural and functional neuroimaging showed altered brain connectivity in AD. In this study, we investigated the whole-brain resting state functional connectivity (FC) and structural connectivity (SC) of the subjects with AD, LMCI, EMCI and NC from the ADNI database.
ADNI is an ongoing, longitudinal, multicentre study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease. Although, ADNI is multi feature and contains different range of MRI sequences, but for most of the patients the dataset is not complete.
In this project, we filled these gaps using personalized large-scale brain network modelling (The Virtual Brain (TVB)) and completed the ADNI database. Statistically, from 244 selected patients after quality check, only 12 of them have complete dataset. Due to the importance of a complete data set comprising anatomical and functional for each individual, we aimed to complete the dataset by simulating the missing parts.



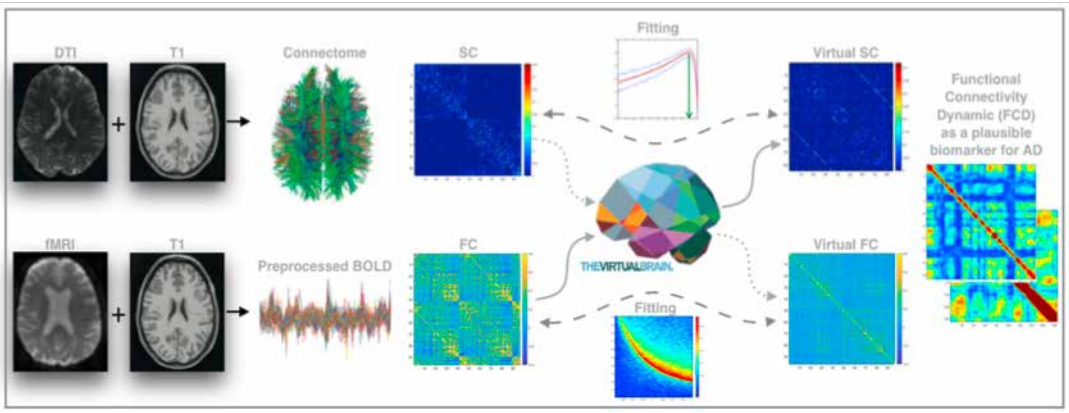
It has been demonstrated that the correlation structure of spontaneous BOLD fluctuations (FC) relates to the underlying anatomical circuitry as obtained by diffusion tensor spectrum imaging (DTI). Nevertheless, how FC relates to the anatomical connectivity and brain dynamics various methods have been proposed and all are effective and complementary. Here, we consider both linear and nonlinear models to understand structure-function relationships and having complete set of features for every subject.
In this study, for completing the imaging data of ADNI like patient cohort, we have used two models. The first model to make data completion is a linear system of stochastic first order differential equations, the connectome-based Ornstein-Uhlenbeck process of TVB. Additionally, for capturing the nonlinearity of the system which may have direct relation with the pathology of Alzheimer's disease, we used a connectome-based mean-field whole brain model based on the Wong-Wang local dynamics.

We proposed a whole-brain computational approach to model the whole-brain structural and functional connectivity of each subject by TVB. All analyses and simulations were carried out using connectivity matrices based on a 96-brain parcellation. In Fig. 1 the TVB pipeline for completing missing data in ADNI has been illustrated.
As a first step, tractography was performed using data from DTI and by using T1 images and implementing the 96-brain parcellation, the SC matrices for each subject were built. For constructing the FC matrices, after preprocessing the fMRI images, the parcellation were implemented in order to reach BOLD signals. Afterwards, by calculating the Pearson correlation coefficient of signals, the FC matrices were built.
At the second step, the simulation and parameter fitting were performed on 12 subjects for Ornstein-Uhlenbeck process and Wong-Wang model. By linear Ornstein-Uhlenbeck process we could simulate SC from FC and when SC were missing and vice versa. As it is illustrated in Table 1, we could virtualize FC of 76 patients from their SC and SC of 156 patients based on their FC.
The Wong-Wang model was used for the simulations of the group of patients, for which their functional data were missing in order to simulated their BOLD signals besides having more sophisticated FC and FCD matrices.

We developed methods to compute the missing data on basis of the available data, and for that one of connectivity matrices, either functional or structural is necessary. Metrics applied in this study are Functional Connectivity Dynamics (FCD) of simulated and empirical time-series.
In this project, using the linear approach of TVB with an Ornstein-Uhlenbeck process, we could systematically complete the missing data in ADNI. This pipeline is built symmetrically, meaning that SC can be reproduce from FC or FC can be predicted from SC. Furthermore, using the nonlinear approach of TVB with Wong-Wang model, we could reach the same target in a different way. The advantage of linear approach is that it is computationally less costly, which shows its importance when working with big data, however, by nonlinear approach we took advantage of its realistic simulation for pathological studies on AD. It worth to mention that in nonlinear approach besides simulating FC from SC, we can simulate the BOLD time-series which is necessary for constructing FCD as a plausible biomarker of AD.

In order to verify the quality of the reconstruction, we use the 12 available subjects, for which both empirical structural and functional information are available. The similarity of simulated

data to empirical had up to 0.45 correlation, which is a confirmation of the level of accuracy of the simulation.
In line with the idea of symmetry in the pipeline, this is desirable for future work to investigate the nonlinear simulation of SC from FC by Effective connectivity. We demonstrated data completion is feasible when one of the structural or functional data is missing and we have solved the problem by filling all the gaps. The ADNI database has been now extended from 12 complete data sets to 156 comprising DTI, MRI and fMRI data.



The TVB pipeline for completing ADNI data.

MULTI-SCALE LONGITUDINAL MODELS FOR VIRTUAL COHORTS

In addition to simulating virtual brain connectomes, UCB and Fraunhofer have recently developed a method to model longitudinal clinical cohorts across biological scales and different biological and clinical modalities. The key idea is to represent a longitudinal clinical cohort as a Bayesian Network model. Since Bayesian Networks are generative models representing a multivariate statistical distribution they can subsequently be used to generate virtual patients. Moreover, Bayesian Networks can be used to make predictions. That means they can also be used for prognosis purposes.
There are a number of non-trivial challenges associated:

- Data in clinical studies often contains missing values, which are not entirely random, but could be correlated with a specific reason (e.g. patient drop out due to symptom worsening)
- Data in referential clinical cohorts, such as ADNI and PPMI is high dimensional, specifically, if genotype information is considered.
- Bayesian Network learning is NP hard. The identification of the true causal network structure is therefore statistically and computationally extremely challenging.

We addressed these challenges via the following approach:

1. Explicit modeling of missing data via auxiliary variables.
2. Splitting of original variables into informative groups and non-linear dimensionality reduction using deep autoencoders within each group. The result is a group-wise score for each individual patient.
3. Constraining Bayesian Network structures via prior knowledge.

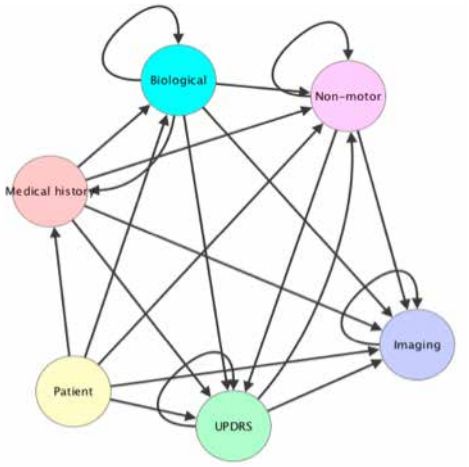


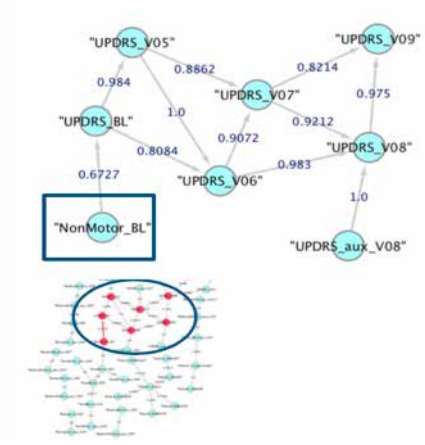
Figure 1: Edges allowed in Bayesian Network (BN). The graph illustrates allowed dependencies between six groups of features (Biological, Non-motor, Imaging, UPDRS, Patient, Medical History). The BN was hence restricted to pick edges only out of the set of depicted dependencies

As an example, Figure 1 shows allowed edges between defined variable groups in PPMI.

Different algorithms for learning the Bayesian Network structure were used and compared with each other via cross-validation. Subsequently, our approach was tested in different ways:

1. Assessing the prediction performance: Is the model able to predict a group level score (e.g. UPDRS) for a patient, which has not been used to train the model?
2. Which edges appear stable, if the Bayesian Network is repeatedly (here: 1000 times) learned, if patients from the training data are re-sampled by replacement?
3. Using the Bayesian Network as a generative model, do virtual patients look reasonable similar to real patients? In particular: Can dissimilar patients be identified? Can a general purpose classifier (e.g. a Random Forest) discriminate between real and virtual patients?

Figure 2 below gives an impression of stable edges, which we identified with our Bayesian Network approach in PPMI. Each edge connects two variable groups. Each variable group aggregates different features (e.g. non-motor symptoms), and within each group the relative impact of original variables can be assessed.



variable	relative_importance
NonMotor_SCAU11_BL::often	1.0000000
NonMotor_PTINBOTH_BL::Patient	0.9927775
NonMotor_PTINBOTH_BL::Patient and Informant	0.9608899
NonMotor_CNTRLSEX_BL::No	0.9022760
NonMotor_CNTRLSEX_BL::Yes	0.8685675
NonMotor_SCAU2_BL::often	0.8267905
NonMotor_BJLOT3_BL::Incorrect	0.7850295
NonMotor_BJLOT3_BL::Correct	0.7833341
NonMotor_TMTRWD_BL::No	0.7586994
NonMotor_TMTRWD_BL::Yes	0.7382647
NonMotor_TMSEX_BL::No	0.6932767
NonMotor_TMSEX_BL::Yes	0.6624058
NonMotor_SCAU10_BL::often	0.6386403
NonMotor_SCAU20_BL::often	0.6276819

Figure 2: Stable Bayesian Network features learned from PPMI.

Figure 3 demonstrates the ability of our model to make disease prognosis. The plot shows the cross-validated accuracy for predicting cognitive impairment of ADNI patients as a function of time.

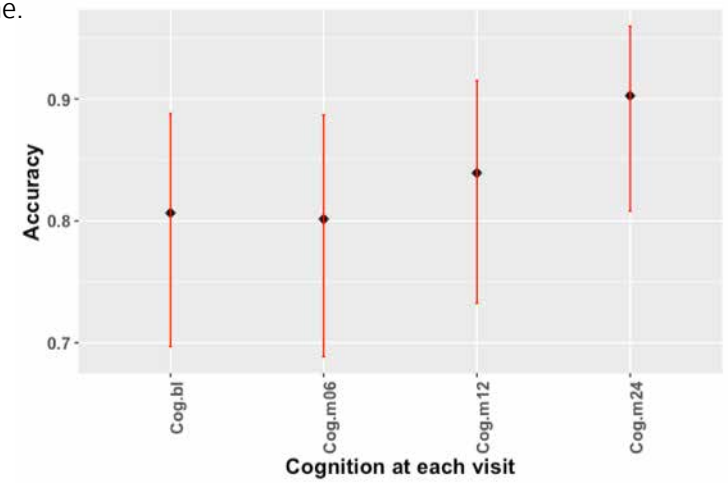


Figure 3: cross-validated accuracy for predicting cognitive impairment of AD patients at different visits. Predictions were always made for each patient in the test set by taking all data for the same patient up to the previous visit as evidence. At baseline all other baseline variables (except for cognitive impairment scores) were used as evidence.

Figure 4 visualizes the distribution of real PPMI and virtual PD patients in a multiple-correspondence analysis plot. No visual discrimination between real and virtual patients is possible, but based on statistical tests a few virtual patients can be identifed as potential outliers compared to the distribution of real patients. Further validations using the approach described above are still ongoing.

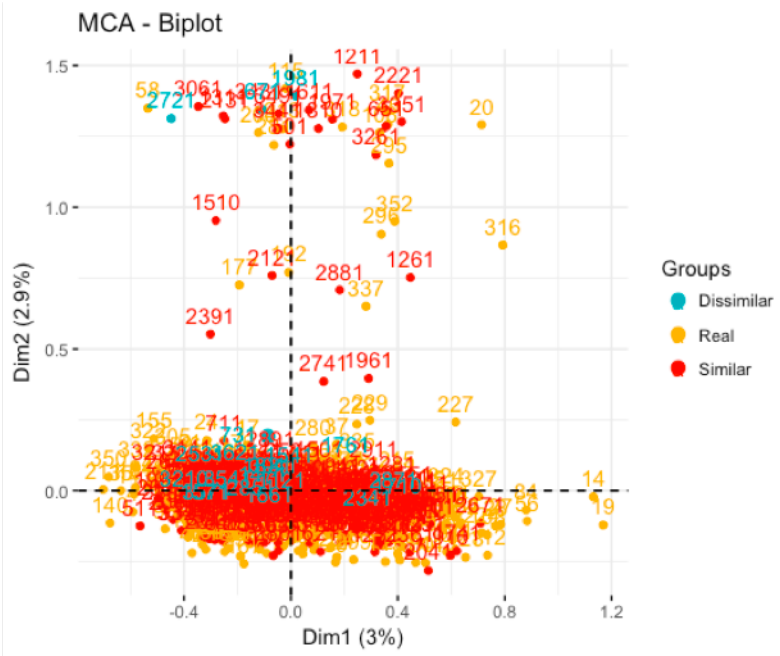


Figure 4: multiple correspondence analysis plot of real and virtual patients. Potential outliers compared to the distribution of real patients are marked in green.

FURTHER INFORMATION

SELECTED PUBLICATIONS

Hofmann-Apitius, M., Alarcón-Riquelme, M. E., Chamberlain, C., & McHale, D. (2015). Towards the taxonomy of human disease. Nature Reviews Drug Discovery, 14(2).

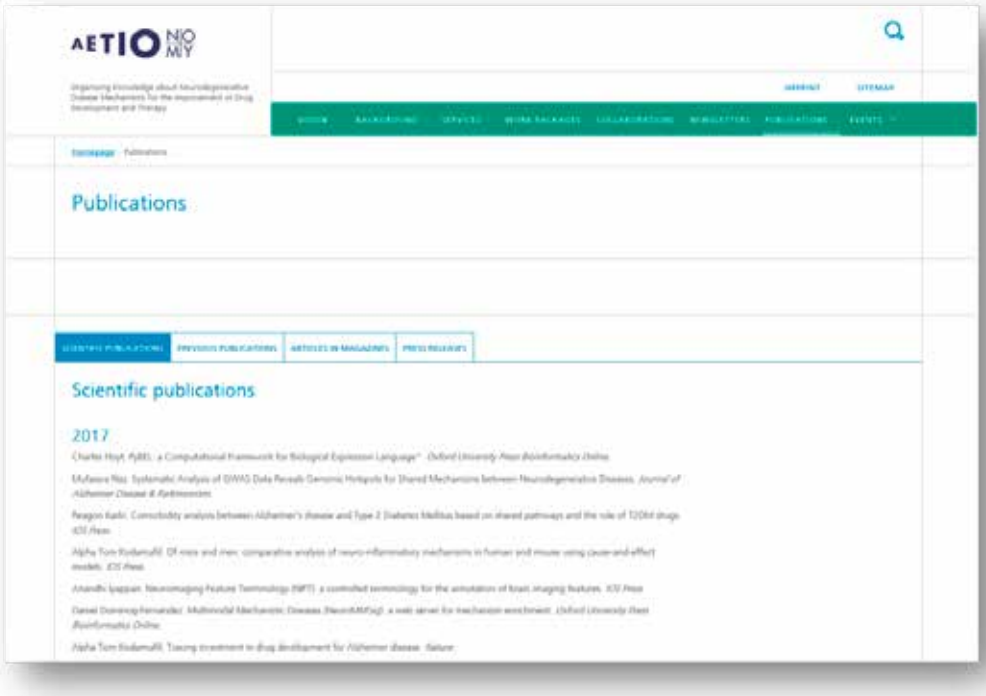
Kodamullil, A. T., Younesi, E., Naz, M., Bagewadi, S., & Hofmann-Apitius, M. (2015). Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. Alzheimer's & Dementia, 11(11), 1329-1339.

Daniel Dominog-Fernandez. Multimodal Mechanistic Diseases (NeuroMMSig): a web server for mechanism enrichment. Oxford University Press Bioinformatics Online.

Shashank Khanna, Daniel Domingo-Fernandez, Anandhi Iyappan, Mohammad Asif Emon, Martin Hofmann-Apitius, Holger Fröhlich (2018), Using Multi-Scale Genetic, Neuroimaging and Clinical Data for Predicting Alzheimer's Disease and Reconstruction of Relevant Biological Mechanisms, Scientific Reports, 8, Article number: 11173

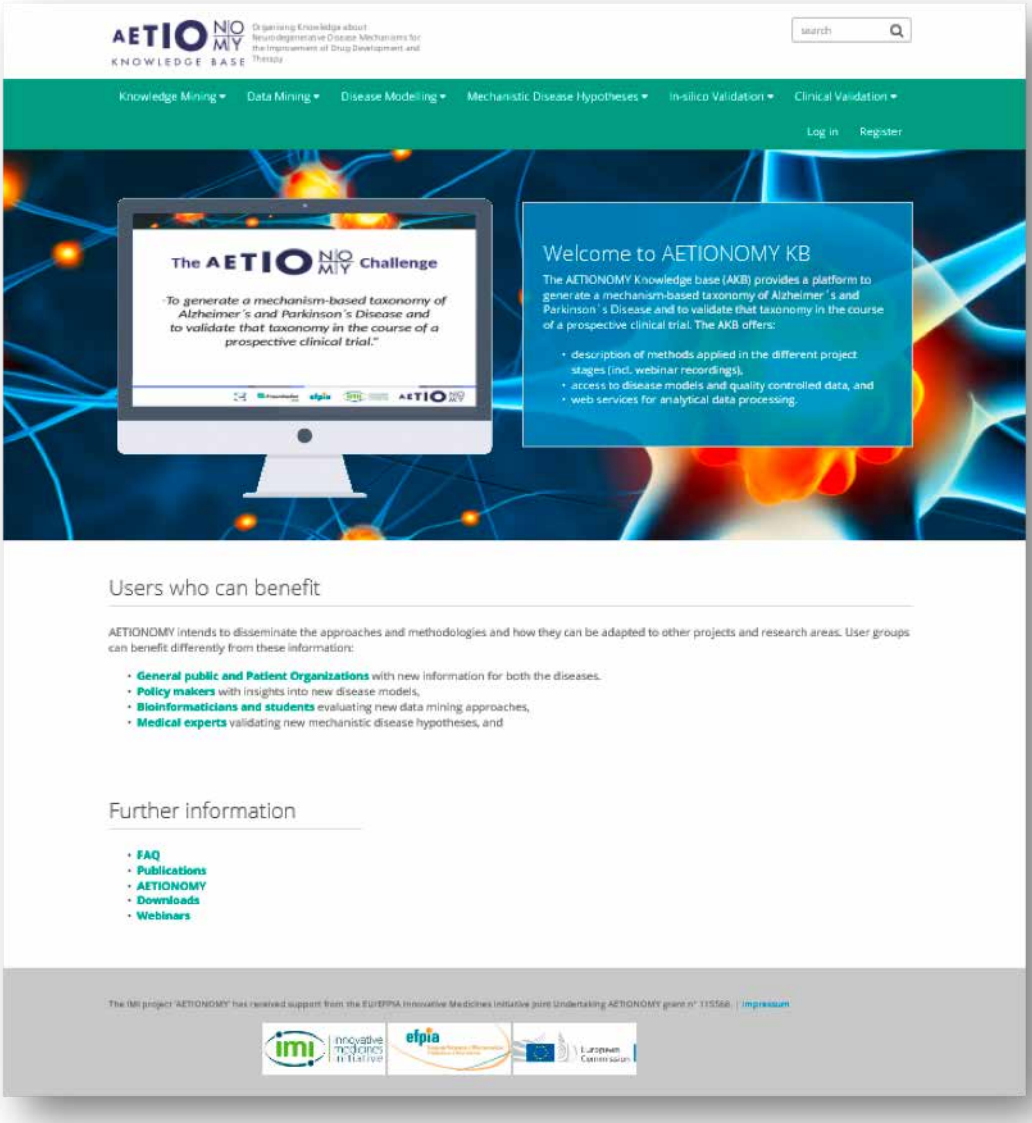
MORE PUBLICATIONS

https://www.aetionomy.eu/en/publications.html



MORE INFORMATION ON OUR APPROACHES:

https://data.aetionomy.scai.fraunhofer.de



The AETIONOMY Project:
The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertalking under grant agreements n° 115568, resources of which are composed of financial contribution from the European Union´s Seventh Framework Programme (FP7/2017-2013) and EFPIA companies´ in kind contribution.