Modeling Neurodegenerative Disease Risk and Progression

Holger Fröhlich

04.04.2019

Inspired by patients.
Driven by science.

universitätbonn

Motivation

Diseases have a time dimension!

We need to understand and model time series data to address important questions in Precision Medicine

- Earlier disease diagnosis / prevention
- Disease progression / prognosis

Case Studies

Early diagnosis: time-to-event models

- Alzheimer‘s Disease risk model

Progression: multivariate time series clustering

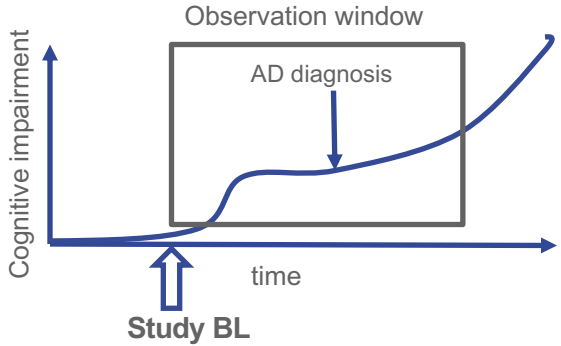
- Alzheimer‘s Disease progression

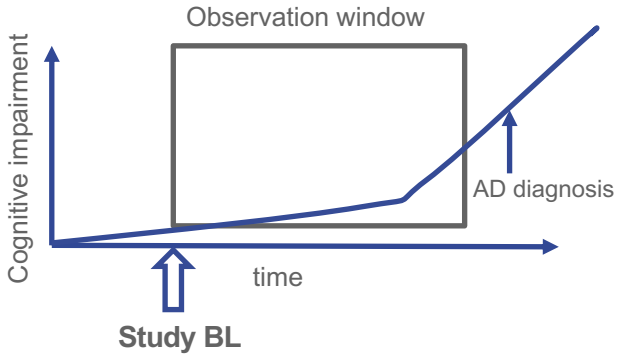
Simulation of multivariate patient trajectories

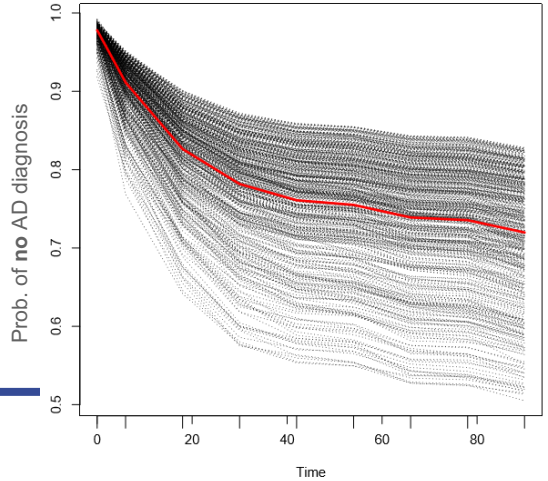
- Talk tomorrow

Early Disease Diagnosis

- **We don‘t know, how long patients have been in a pre-symptomatic phase, before entering the study**
- **Actual AD diagnosis can fall outside the observation window: right censored data**
- **Use well-established theory from survival analysis**







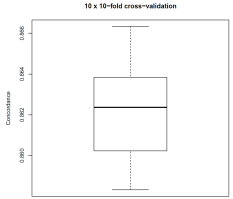
$$h(t|x) = h_0(t)\exp(-f(x))$$

Using Multi-Scale Data for Predicting Alzheimer's Disease and Reconstruction of Biological Mechanisms

Multi-scale baseline data (~900 normal / MCI patients):

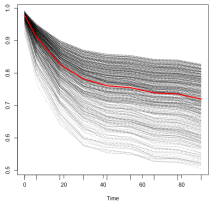
- Clinical features, incl. neuropsychological assessment scores
- Genetic variants (SNPs)
- Neuro-images (MRT)

Prediction performance:
~86% C-index

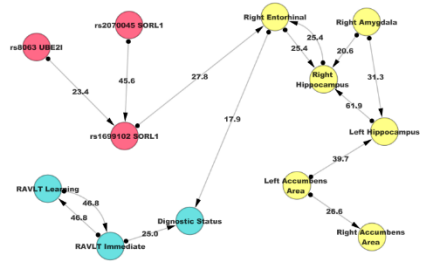


Multi-modal gradient boosted Cox regression

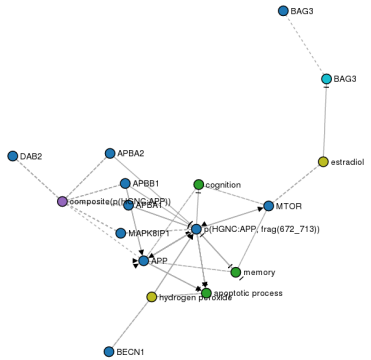
Disease risk



Model interpretation:



Bayesian Networks

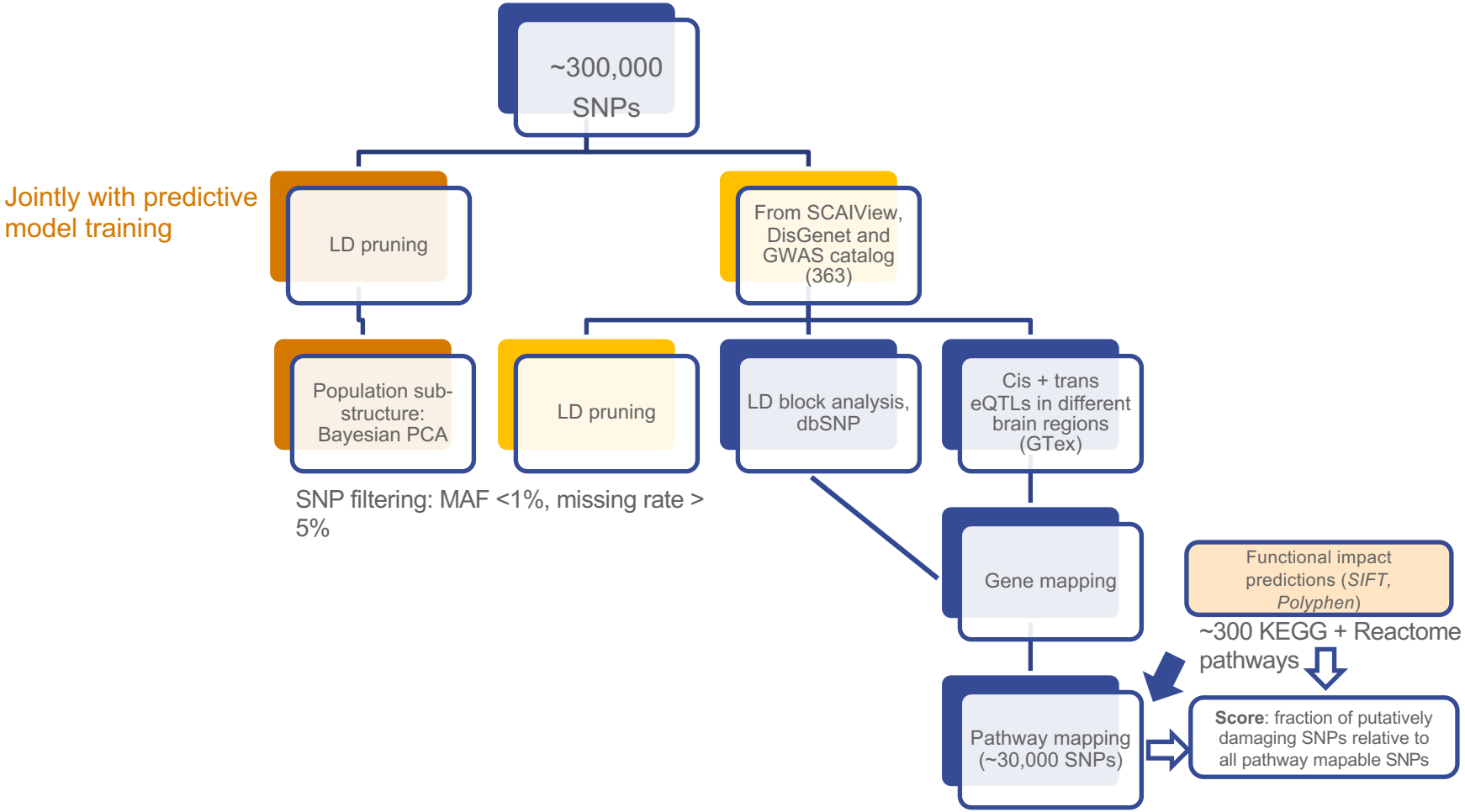


Literature validation

Goals:

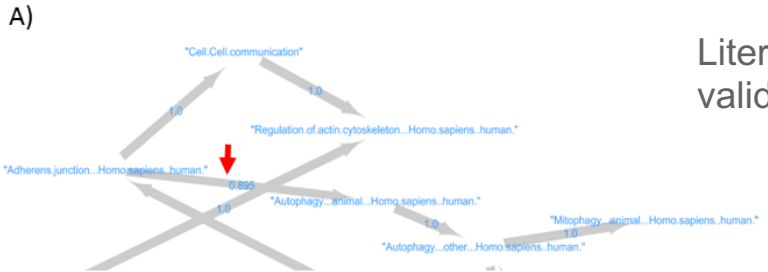
- Predicting disease risk to allow for early therapeutic intervention
- At least partial understanding of transition mechanisms

Khanna, … & Fröhlich, Scientific Reports, 2018


~300,000 SNPs
Jointly with predictive model training
LD pruning
From SCAIView, DisGenet and GWAS catalog (363)
Population sub-structure: Bayesian PCA
LD pruning
LD block analysis, dbSNP
Cis + trans eQTLs in different brain regions (GTex)
SNP filtering: MAF <1%, missing rate > 5%
Gene mapping
Functional impact predictions (SIFT, Polyphen)
~300 KEGG + Reactome pathways
Pathway mapping (~30,000 SNPs)
Score: fraction of putatively damaging SNPs relative to all pathway mapable SNPs

Model interpretation via Bayesian Networks

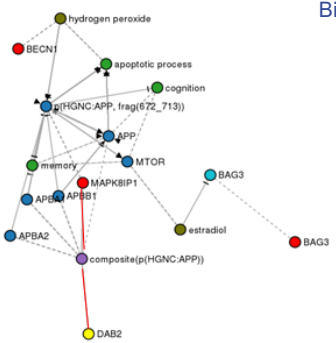
Suggested in Nighot et al., 2016



Literature validation

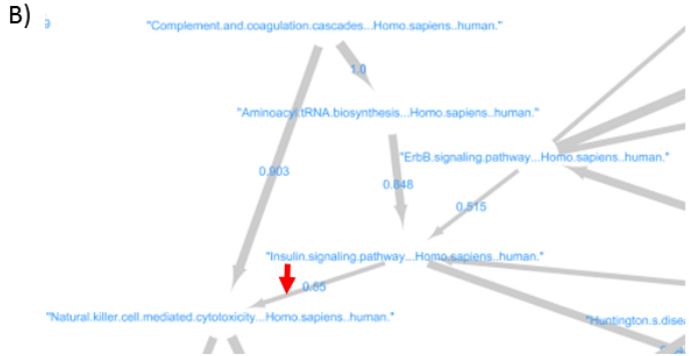
OpenBEL
(Kodamullil et al., 2015)

Adherens junction - autophagy

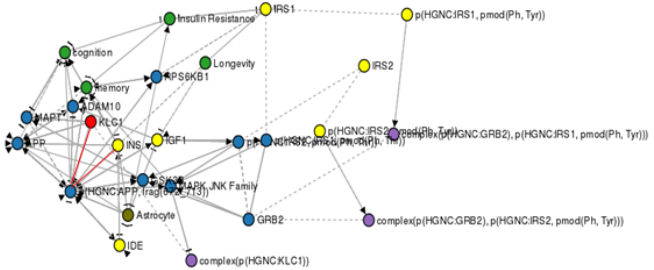


Pathway mapping for OpenBEL: NeuroMMSig (Domingo-Fernandez et al., Bioinformatics, 2017)

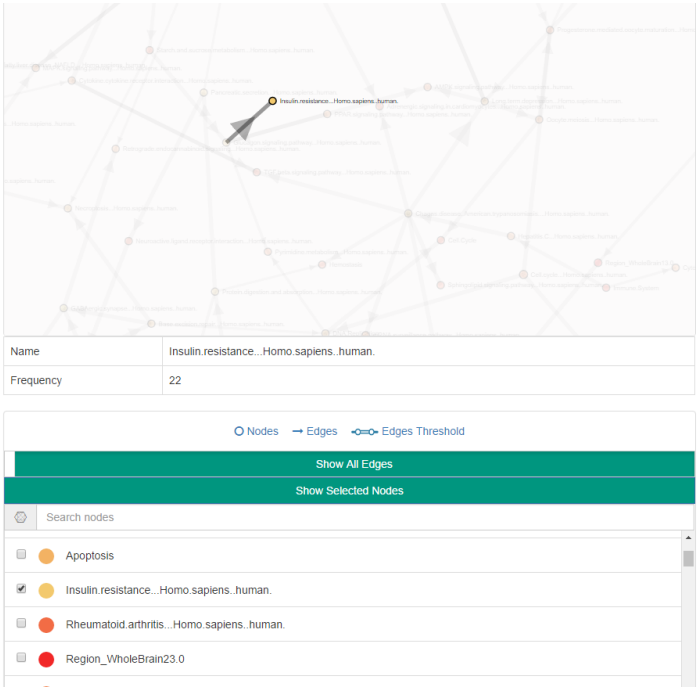
Suggested in Lorini et al., 1994

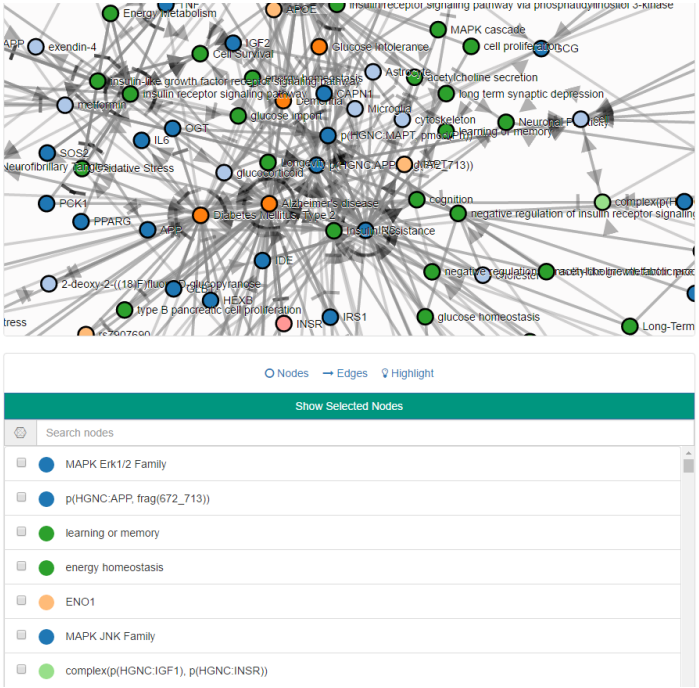


Insulin signaling – NK cell mediated cytotoxicity



An Interactive Web Viewer for Bayesian Networks and Literature Derived Mechanisms

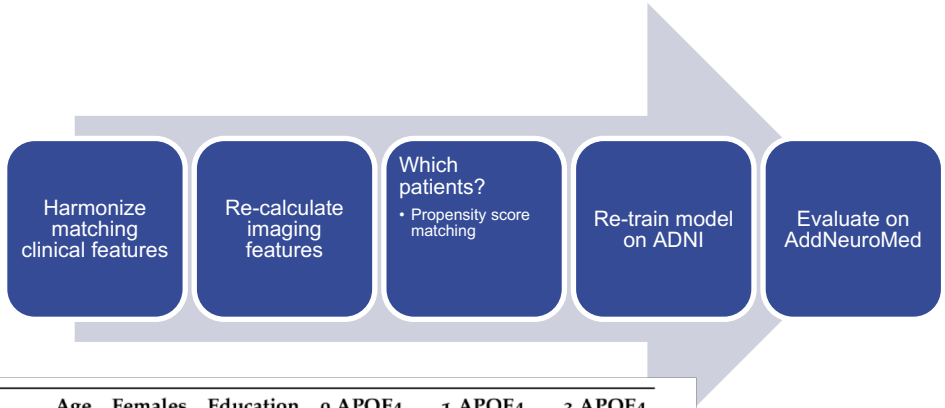




http://neurommsig.scai.fraunhofer.de/bayesian

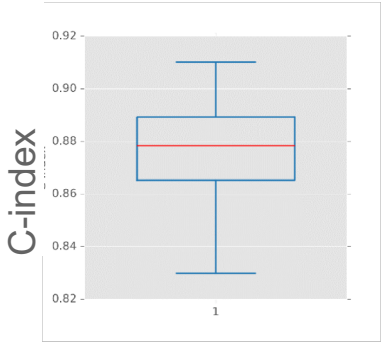
Independent Validation of the AD Risk Model

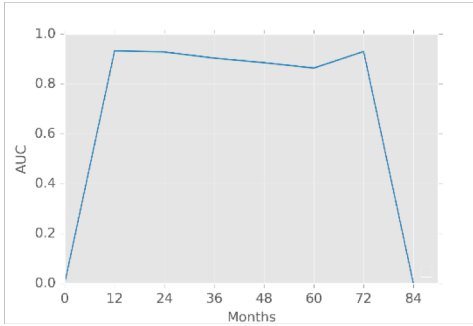
AddNeuroMed cohort: 771 patients

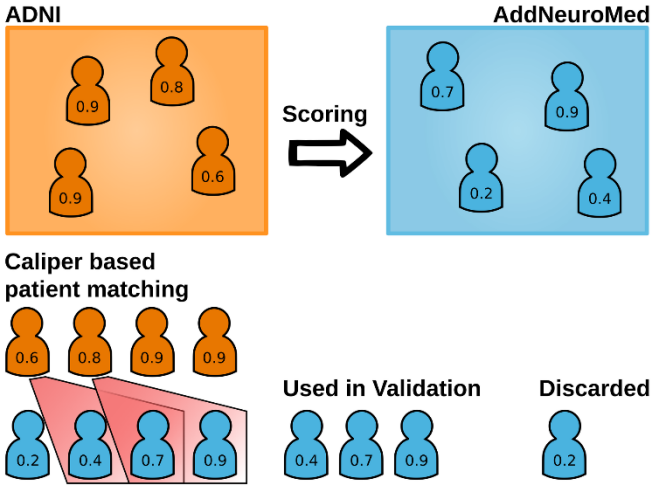


	Age	Females	Education	0 APOE4	1 APOE4	2 APOE4
CTL						
ADNI	74.8	50.1%	16.3	301 (72.5%)	103 (24.8%)	11 (2.7%)
ANM	74.5	59.4%	12.3	380 (74.7%)	118 (23.2%)	11 (2.5%)
MCI						
ADNI	73.0	59.1%	15.9	431 (49.8%)	341 (39.4%)	94 (10.9%)
ANM	76.0	54.7%	10.0	160 (60.4%)	95 (35.8%)	10 (3.8%)
AD						
ADNI	75.0	55.3%	15.2	160 (37.3%)	113 (33.4%)	65 (19.2%)
ANM	78.6	62.9%	9.4	197 (45.7%)	178 (41.3%)	56 (13.0%)

Table 8: Demographic composition of each diagnosis cohort from ADNI and AddNeuroMed. Age and education averages are reported in years. ***0, 1, 2 APOE4:*** *Number of individuals with 0, 1, 2 APOE4 alleles. Percentages give the fraction of the respective data set with the corresponding value.* ***Females:*** *Fraction of female individuals.* ***ANM:*** *AddNeuroMed*







Case Studies

Early diagnosis: time-to-event models

- Alzheimer‘s Disease risk model

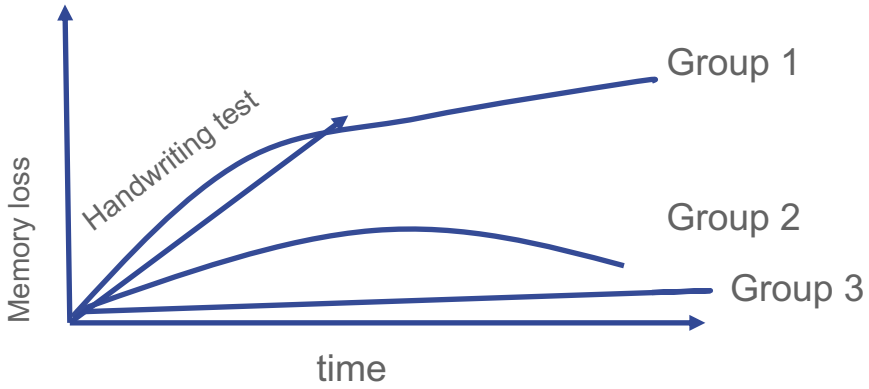
Progression: multivariate time series clustering

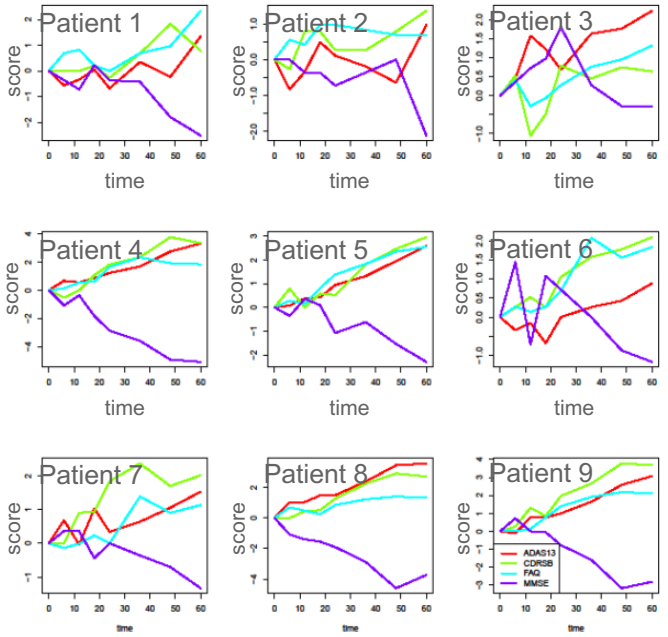
- Alzheimer‘s Disease progression

Simulation of multivariate patient trajectories

- Talk tomorrow

How to find multivariate disease progression clusters?





Disease progression in many diseases is highly heterogenous

- Examples: Parkinson‘s, Alzheimer‘s

Multiple scores can be used to describe different aspects of disease severity

No existing standard technique

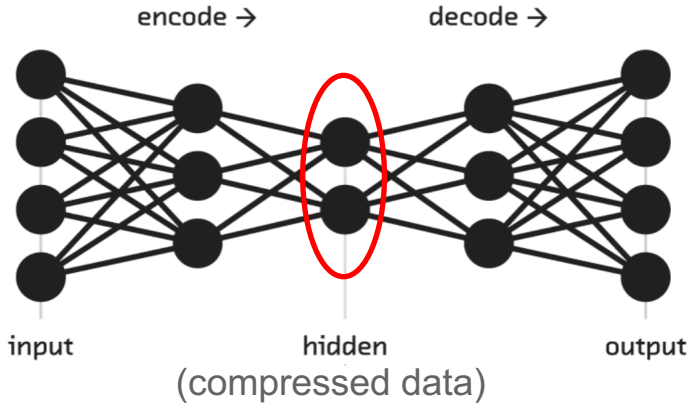
Background: Variational Autoencoders

Idea: Compress data into a lower dimensional representation

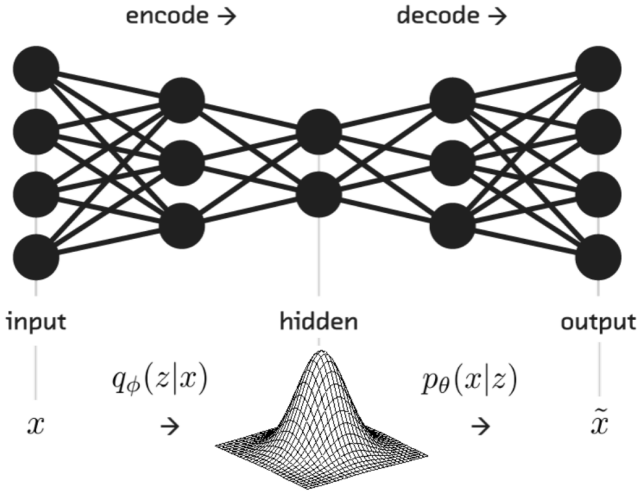
- Objective minimize reconstruction error

Variational autoencoder: make this representation smooth

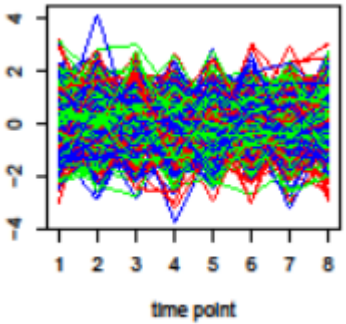
- Map each input data point to a Gaussian



(compressed data)

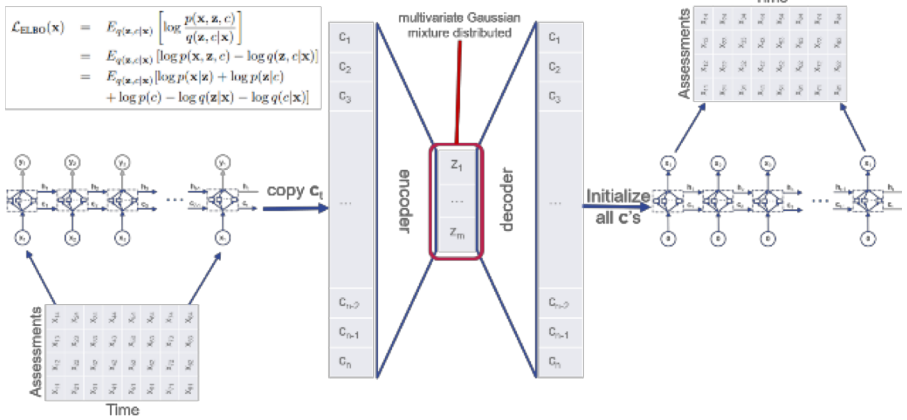


Method: Latent Gaussian Mixtures of Recurrent Variational Autoencoders

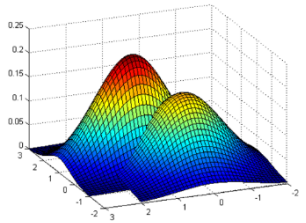


Multivariate time series data

Deep embedding



Latent probabilistic clustering



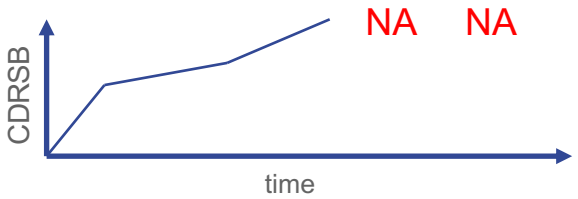
Cluster assignment

- **Challenge: patient drop-out ➔ missing data not at random**
- **Appropriate missing data model is needed**

Jiang et al., Proc IJCAI 2017, arXiv:1611.05148

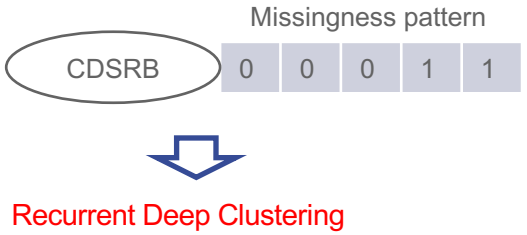
Dealing with Missing Data

Imputation would assume no correlation between missingness and measured features ➔ wrong approach



Our approach:

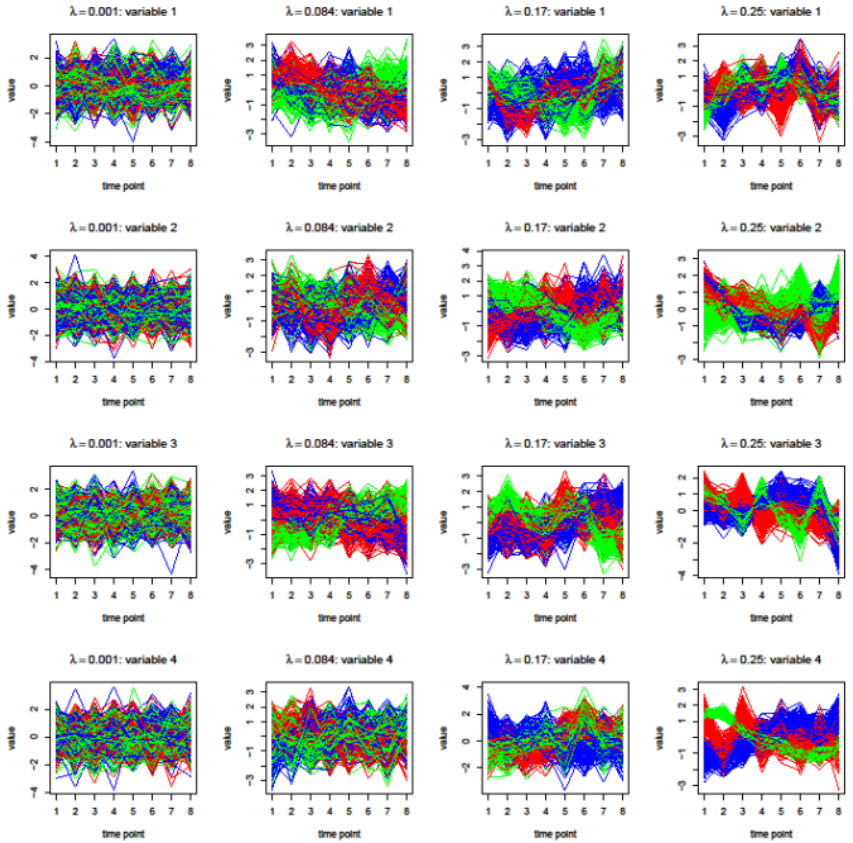
- Indicator variables encode missingness pattern ➔ augment input to neural network ➔ „implicit“ imputation
- Consider only observed data in reconstruction term in ELBO criterion (element-wise multiplication with indicator variable)

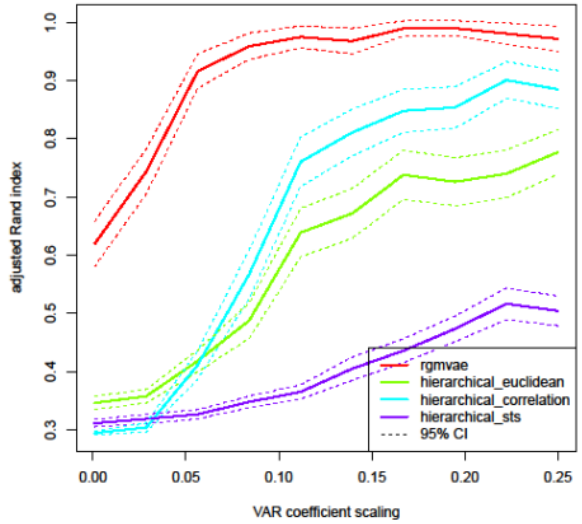


Technical Validation (Simulated Data)

decreasing noise

4 variables

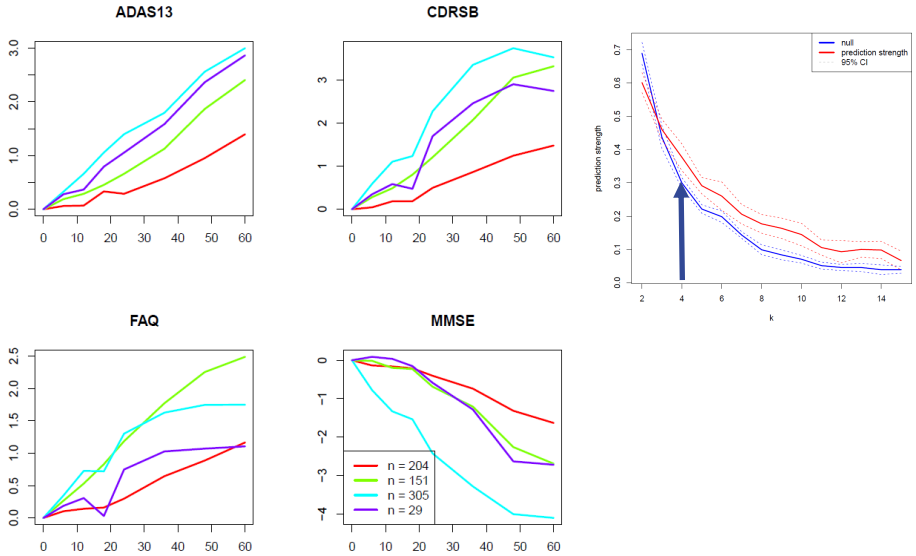




Proposed method outperforms naive approaches

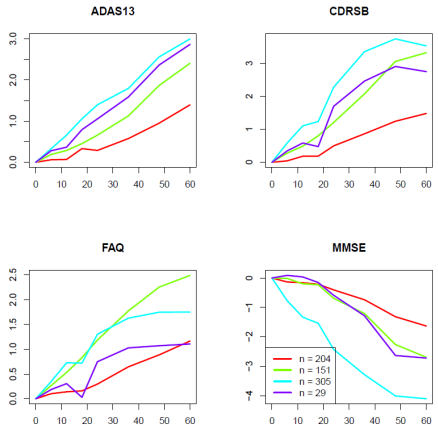
Clustering Results for Alzheimer‘s Disease

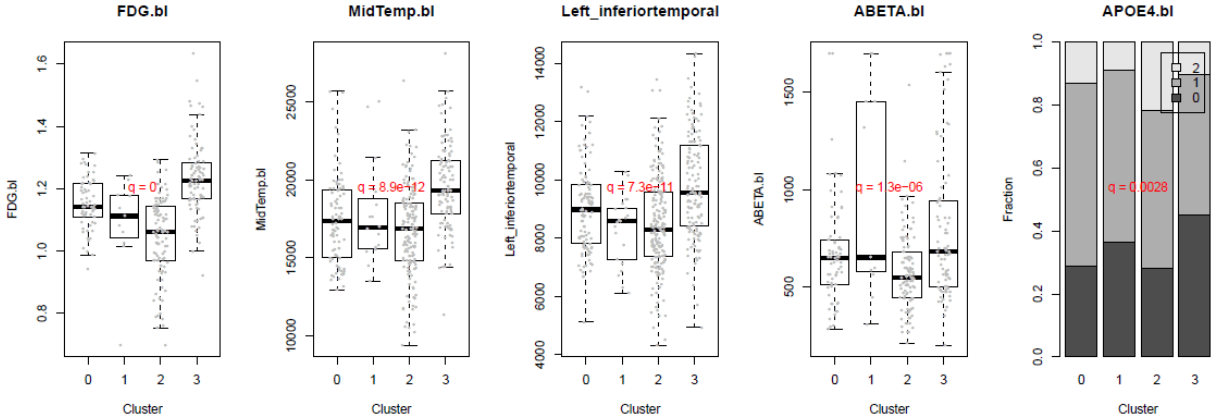
Method yields clearly separated progression clusters for Alzheimer‘s

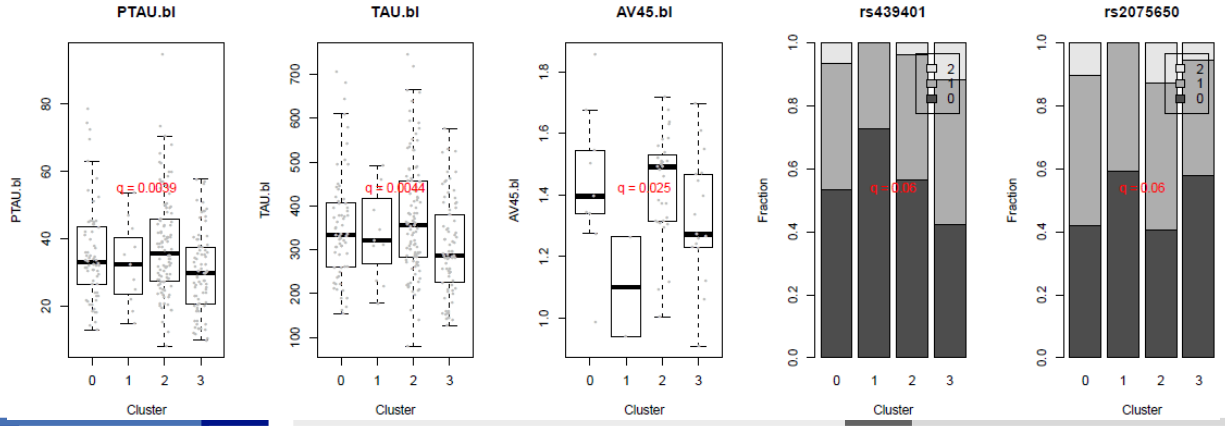


Tibshirani, R. and Walther, G., 2005. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*(3), pp.511-528

Interpretation of Clusters: Alzheimer's Disease



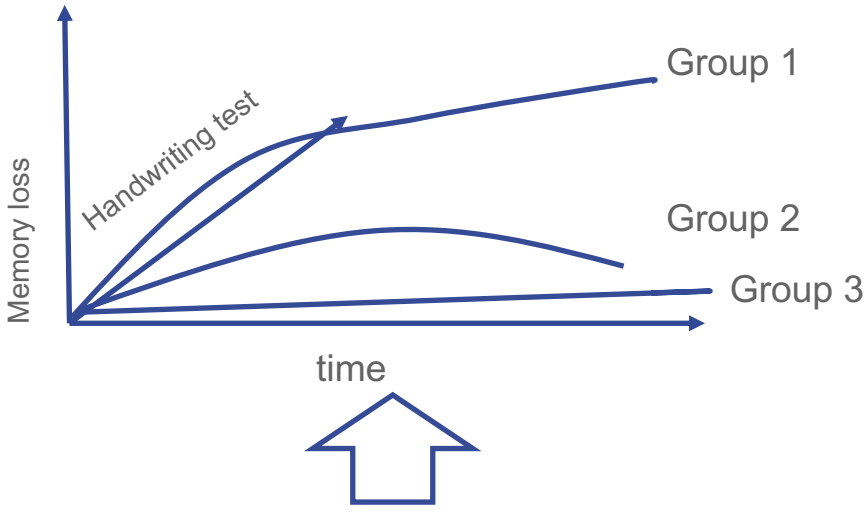




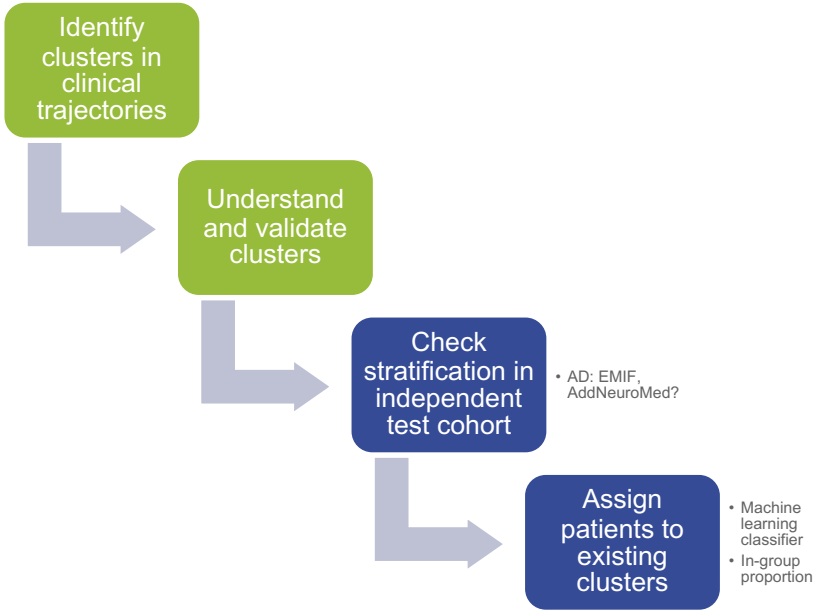
FDG.bl	Glucose uptake marker
MidTemp.bl	Brain region
Left_inferiortemporal	Brain region
ABETA.bl	CSF ABETA levels
APOE4.bl	CSF APOE4 levels
PTAU.bl	CSF PTAU levels
TAU.bl	CSF TAU levels
AV45.bl	ABETA deposits marker
rs439401	APOE4 SNP
rs2075650	TOMM40 SNP

multinomial logistic regression-based likelihood ratio test (corrected for age, gender and education)

Conceptual Idea of Next Steps



Baseline data: genetic, clinical, imaging



Kapp & Tibshirani, Biostatistics 2007

Summary of Case Studies

	Method
Early diagnosis: time-to-event models • Alzheimer‘s Disease risk model	Gradient boosted Cox regression, Bayesian Networks
Progression: multivariate time series clustering • Alzheimer‘s Disease	Recurrent Neural Networks
Simulation of multivariate patient trajectories • Tomorrow‘s talk	Bayesian Networks, autoencoders

Conclusion

Time series analysis is extremely relevant to understand diseases and to answer important questions in Precision Medicine

- Early disease diagnosis
- Disease progression / prognosis

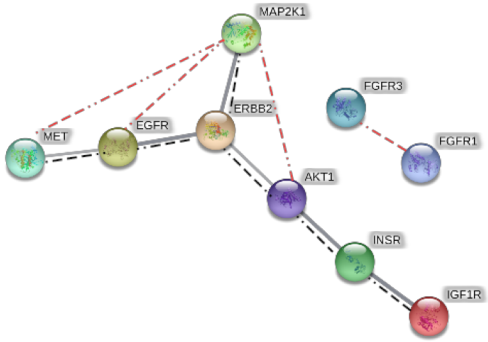
Multitude of methods, right method depends on actual application question

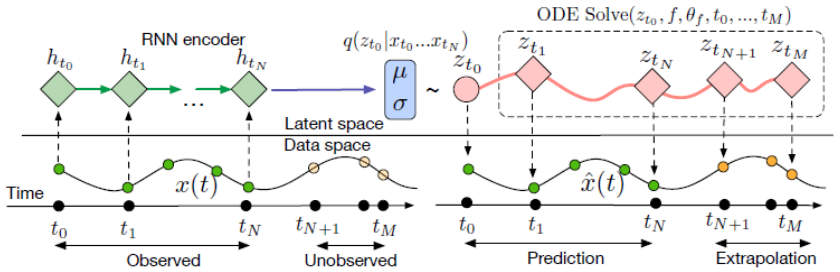
Outlook: Learning the Hidden Dynamics of Disease Development and Progression

Can we learn the mechanism behind disease progression?

- Time series predictions for each individual
- Extrapolation out of domain of training data possible

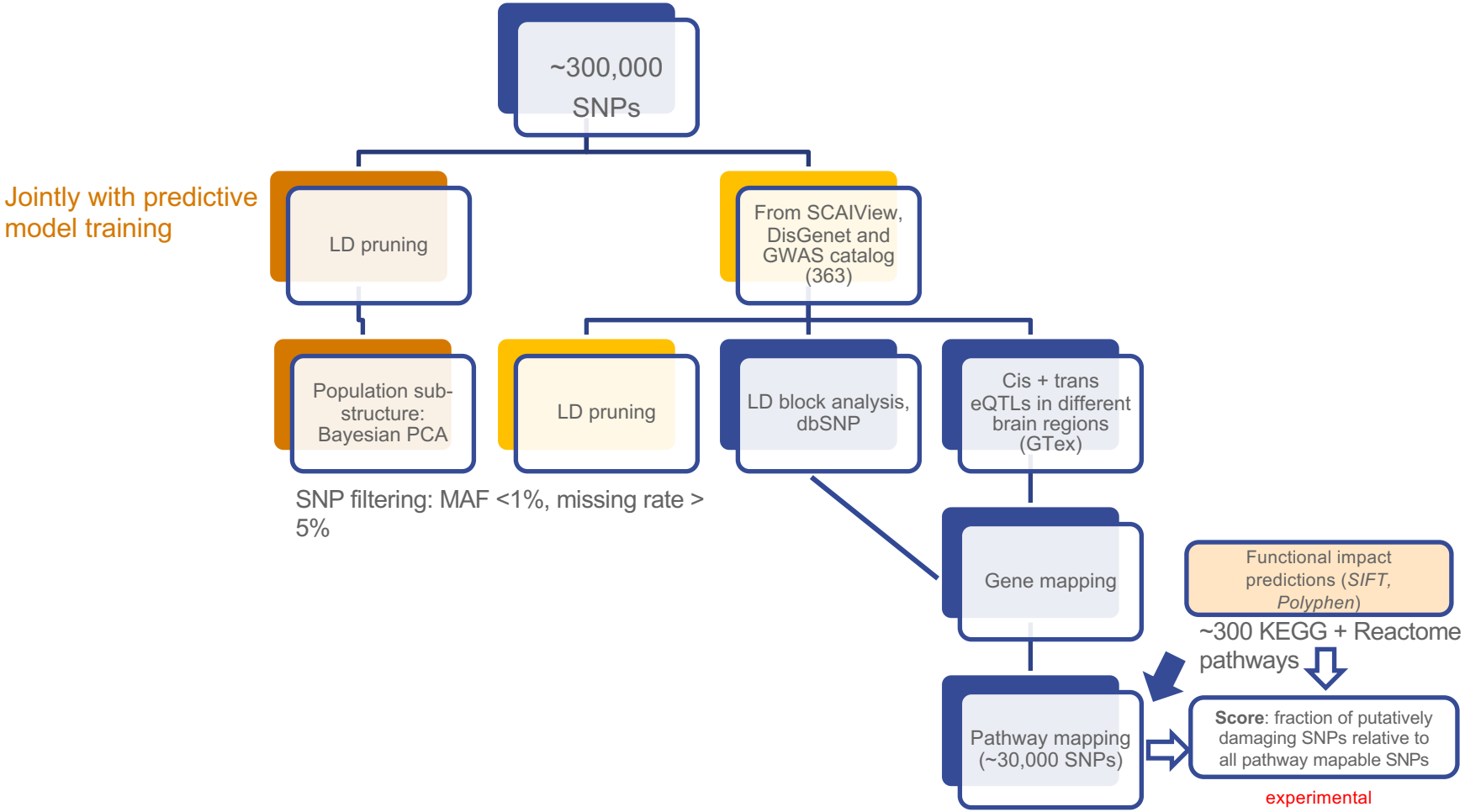
Idea: learn latent ODE (PhD project of Zahra Nasrollah)



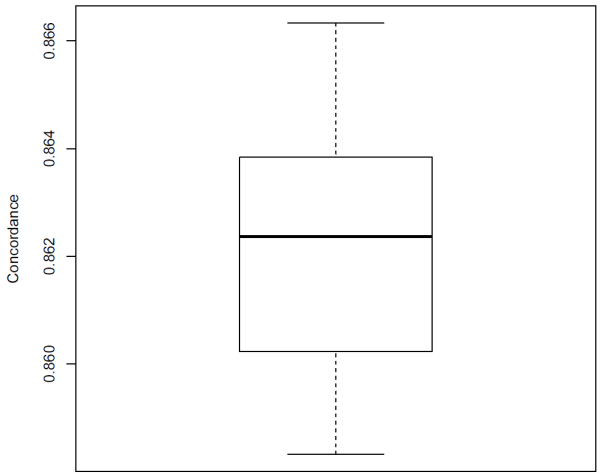


Chen et al., Proc NIPS 2018

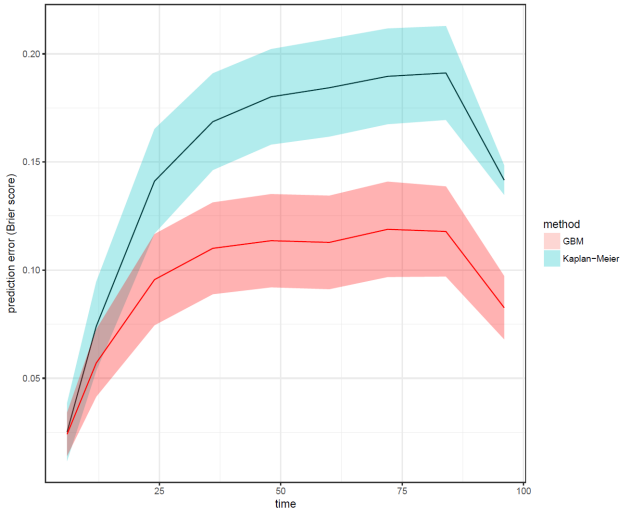
Backup


~300,000 SNPs
Jointly with predictive model training
LD pruning
From SCAIView, DisGenet and GWAS catalog (363)
Population sub-structure: Bayesian PCA
LD pruning
LD block analysis, dbSNP
Cis + trans eQTLs in different brain regions (GTex)
SNP filtering: MAF <1%, missing rate > 5%
Gene mapping
Functional impact predictions (SIFT, Polyphen)
~300 KEGG + Reactome pathways
Pathway mapping (~30,000 SNPs)
Score: fraction of putatively damaging SNPs relative to all pathway mapable SNPs
experimental

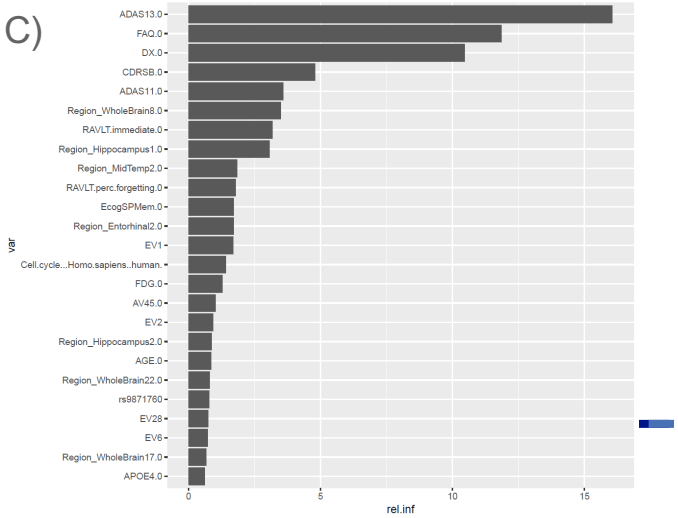
A)


10 x 10−fold cross−validation
Concordance
0.860
0.862
0.864
0.866

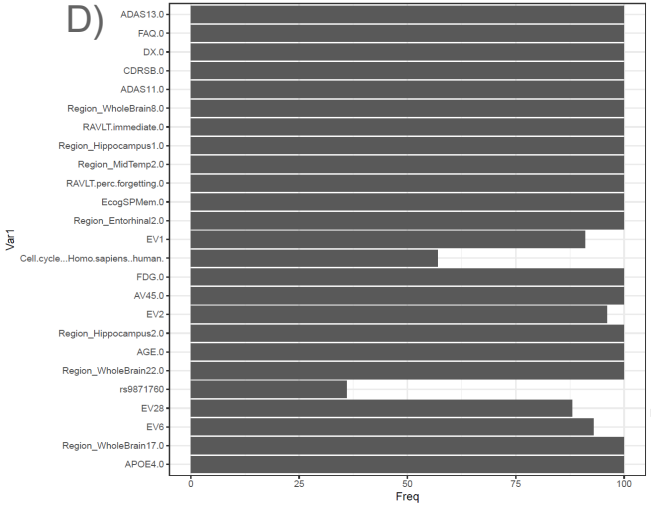
B)


prediction error (Brier score)
0.05
0.10
0.15
0.20
time
25
50
75
100
method
GBM
Kaplan−Meier

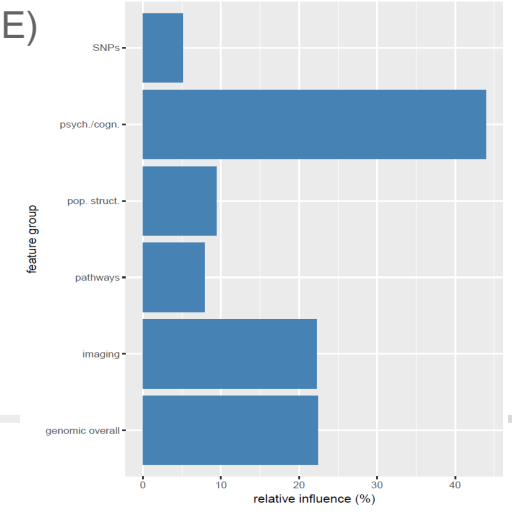
C)


var
ADAS13.0
FAQ.0
DX.0
CDRSB.0
ADAS11.0
Region_WholeBrain8.0
RAVLT.immediate.0
Region_Hippocampus1.0
Region_MidTemp2.0
RAVLT.perc.forgetting.0
EcogSPMem.0
Region_Entorhinal2.0
EV1
Cell.cycle...Homo.sapiens..human.
FDG.0
AV45.0
EV2
Region_Hippocampus2.0
AGE.0
Region_WholeBrain22.0
rs9871760
EV28
EV6
Region_WholeBrain17.0
APOE4.0
0
5
10
15
rel.inf

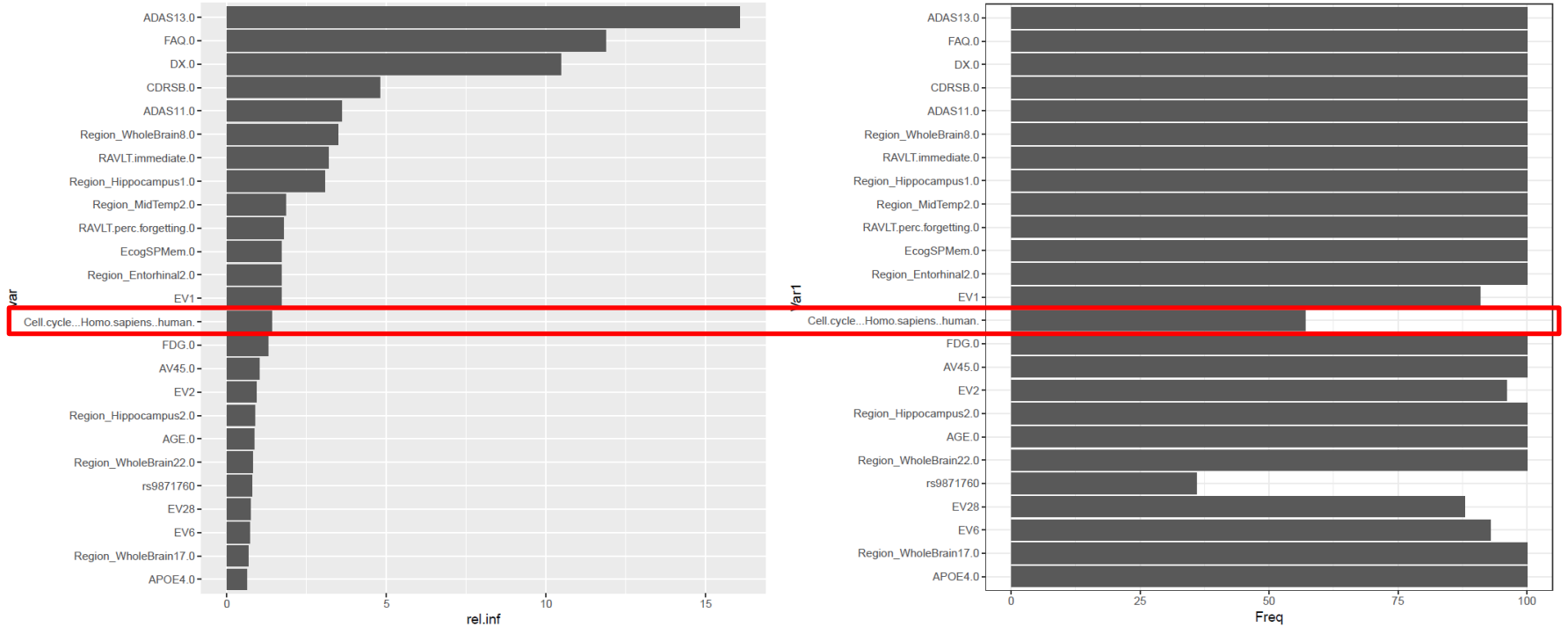
D)


Var1
ADAS13.0
FAQ.0
DX.0
CDRSB.0
ADAS11.0
Region_WholeBrain8.0
RAVLT.immediate.0
Region_Hippocampus1.0
Region_MidTemp2.0
RAVLT.perc.forgetting.0
EcogSPMem.0
Region_Entorhinal2.0
EV1
Cell.cycle...Homo.sapiens..human.
FDG.0
AV45.0
EV2
Region_Hippocampus2.0
AGE.0
Region_WholeBrain22.0
rs9871760
EV28
EV6
Region_WholeBrain17.0
APOE4.0
0
25
50
75
100
Freq

E)


feature group
SNPs
psych./cogn.
pop. struct.
pathways
imaging
genomic overall
0
10
20
30
40
relative influence (%)

Top Relevant Features



> **AD may be caused by aberrant re-entry of different neuronal populations into the cell division cycle (Nagy et al., Neuroscience, 1998)**

Results of Predictive Modeling

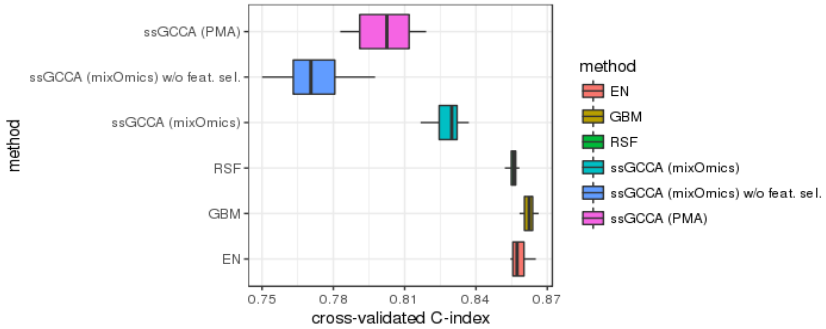
10 times repeated 10-fold cross-validation: ~86% C-index

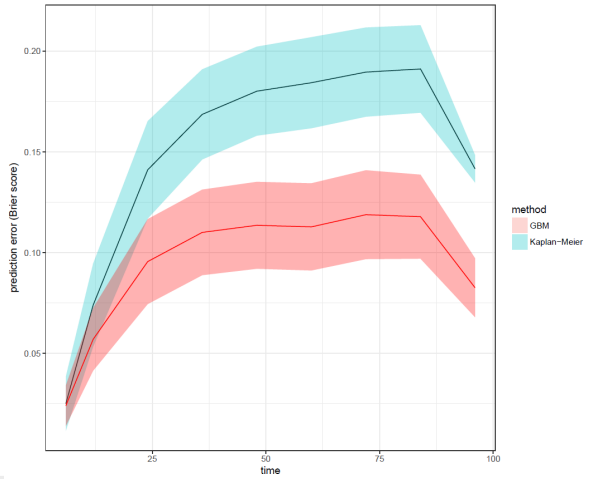
- 50% = chance level
- 100% = perfect

Multi-modal GBM significantly outperforms other methods

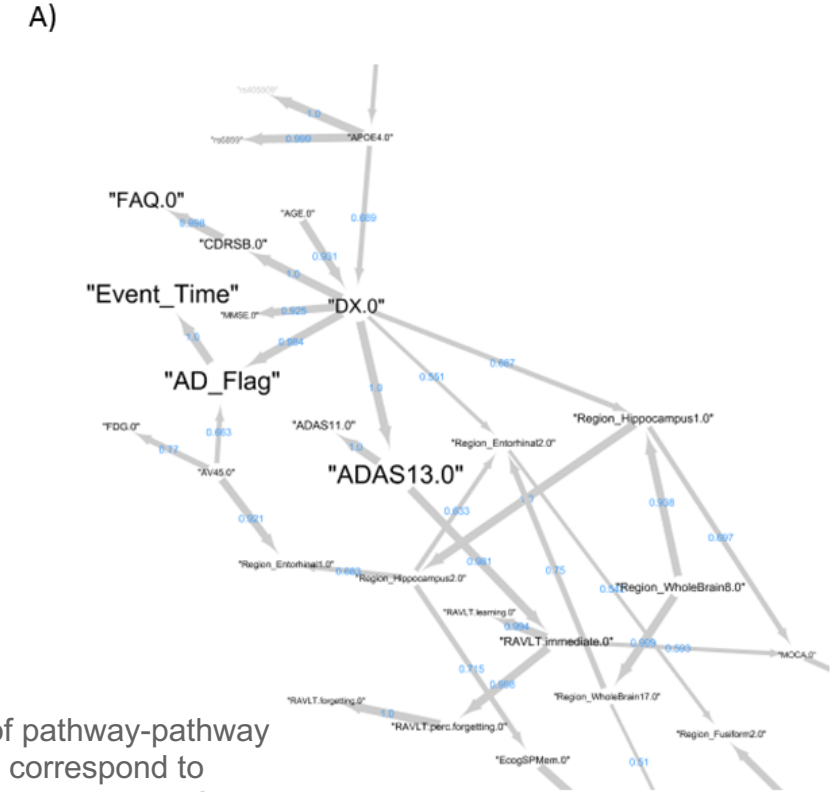
- Similar performance of single GBM, but more features selected

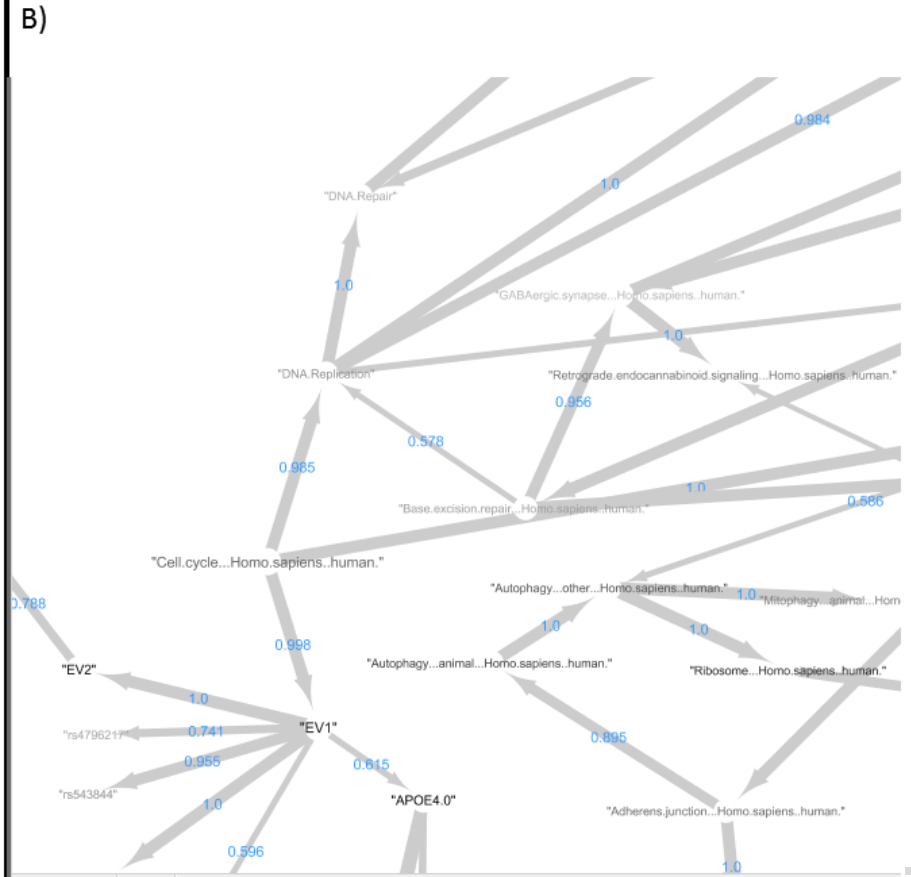
Prediction error over time: gradient boosting outperforms Kaplan-Meier estimator





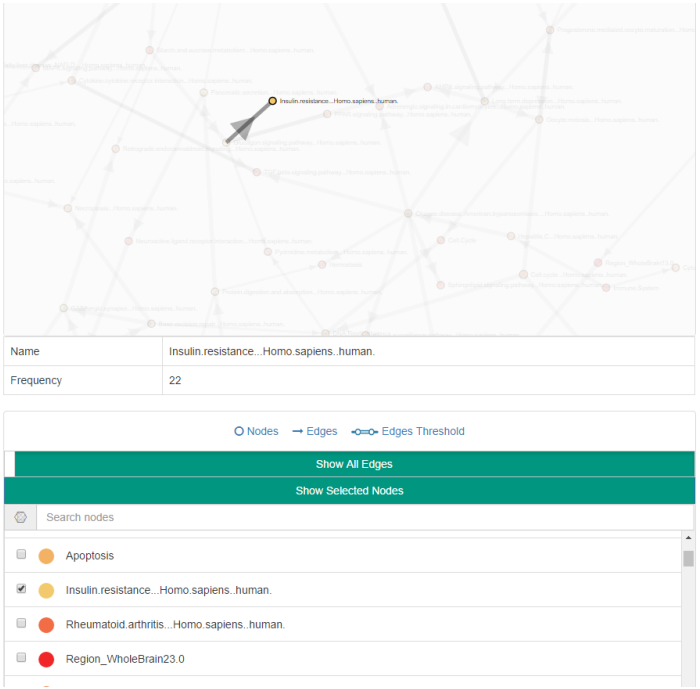
Results of BN Structure Learning (Example Zooms)

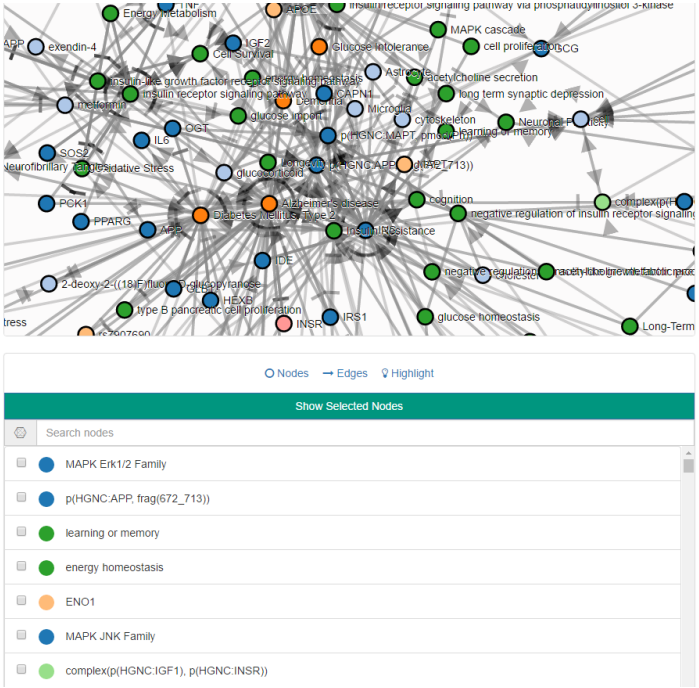




83% of pathway-pathway edges correspond to significant overlap of gene sets (FDR < 5%)

An Interactive Web Viewer for Bayesian Networks and Literature Derived Mechanisms





http://neurommsig.scai.fraunhofer.de/bayesian