AI reveals insights into link between CD33 and cognitive impairment in Alzheimer's Disease

Tamara Raschka[1,2,3,*], Meemansa Sood[1,2], Bruce Schultz[1], Aybuge Altay[1,2,4], Christian Ebeling[1], and Holger Fröhlich[1,2,*]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany
[2]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Bonn, Germany
[3]Fraunhofer Center for Machine Learning, Germany
[4]Current Address: Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany
[*]Corresponding authors

E-Mail: tamara.raschka@scai.fraunhofer.de (TR), holger.froehlich@scai.fraunhofer.de (HF)

Abstract

Modeling biological mechanisms is a key for disease understanding and drug-target identification. However, formulating quantitative models in the field of Alzheimer's Disease is challenged by a lack of detailed knowledge of relevant biochemical processes. Additionally, fitting differential equation systems usually requires time resolved data and the possibility to perform intervention experiments, which is difficult in neurological disorders. This work addresses these challenges by employing the recently published Variational Autoencoder Modular Bayesian Networks (VAMBN) method, which we here trained on combined clinical and patient level gene expression data while incorporating a disease focused knowledge graph. Our approach, called iVAMBN, resulted in a quantitative model that allowed us to simulate a down-expression of the putative drug target CD33, including potential impact on cognitive impairment and brain pathophysiology. Experimental validation demonstrated a high overlap of molecular mechanism predicted to be altered by CD33 perturbation with cell line data. Altogether, our modeling approach may help to select promising drug targets.

Author Summary

For the last 20 years the field of Alzheimer's Disease (AD) is marked by a series of continuous failures to deliver demonstrably effective medications to patients. This is also highlighted by the highly controversial recent approval of Aduhelm (Biogen) by the FDA, which is now investigated internally due to the lack of clear efficacy.

One of the reasons for the continuous failure of trials in AD is the choice of the wrong target mechanism. In essence there is a lack of understanding, how targeting a certain molecule would affect cognitive impairment in human. One way to address this issue is the development of quantitative system level models connecting the molecular level with the phenotype. However, formulating such models in the field of Alzheimer's Disease is challenged by a lack of detailed knowledge of relevant biochemical processes and the connection of molecular mechanisms to cognitive impairment. Additionally, fitting of differential equation systems, which are often used in systems biology, requires time resolved data and the possibility to perform intervention experiments, which is difficult in neurological disorders due to the lack of realistic model systems.

Our work addresses these challenges by employing a novel hybrid Artificial Intelligence (AI) approach combining variational autoencoders with Bayesian Networks. Our proposed approach, named Integrative Variational Autoencoder Modular Bayesian Networks (iVAMBN), was trained on combined clinical and patient level gene expression data while incorporating a disease focused knowledge graph. Our method resulted in an interpretable, quantitative model. It showed connections between various biological mechanisms playing a role in AD. Furthermore, iVAMBN directly connected the molecular level to the disease phenotype. Our model allowed us to simulate a down-expression of the putative drug target CD33. Results showed a significantly increased cognition and predicted perturbation of a number of biological mechanisms. We experimentally validated these predictions using gene expression data from a knock-out THP-1 monocyte cell line. This experiment confirmed our model predictions up to a very high extend. To our knowledge we thus developed the first experimentally validated, quantitative, multi-scale model connecting molecular mechanisms with clinical outcomes in the AD field.

Introduction

Alzheimer's Disease (AD) is a neurodegenerative disorder affecting about 50 million people worldwide, resulting in the inability to perform necessary, daily activities before leading to an often early death [1]. Despite decades of research and more than 2000 clinical studies listed on ClinicalTrials.gov, to date there is no cure, and all existing treatments are purely symptomatic [1]. New disease modifying treatments are urgently needed, but require a better mechanistic understanding of the disease.

A common starting point in this context is to map out the existing knowledge landscape about the disease. In the past few decades, a large number of databases have been developed in the bioinformatics community, such as databases for biological pathways (like KEGG [2], PathwayCommons [3], WikiPathways [4], Reactome [5]), drug-target interactions (like OpenTargets [6], Therapeutic Targets Database [7]), disease-gene associations (like DisGeNET [8]) or protein-protein interactions (like STRING [9], IntAct [10]). All these databases simplify the usage of the respective knowledge for algorithms and models, especially in the field of drug target identification. Moreover, none of these databases have been compiled in a disease focused manner. The Biological Expression Language (BEL) provides this opportunity and can be used to represent literature-derived, disease focused knowledge in the form of attributed graphs in a precise manner. For AD a knowledge graph has been published in [11] and represents the manually curated, disease focused mechanistic interplay between genetic variants, proteins, biological processes and pathways described in the literature, enabling the user to computationally query and integrate knowledge graphs into drug target identification algorithms.

One of the interesting molecules in the AD field is CD33, a transmembrane receptor protein expressed primarily in myeloid lineage cells. It has been associated with decreased risk of AD in GWAS studies [12–18] and discussed as a potential therapeutic target, for example via immunotherapy [14]. In an AD mouse model, a knockout of CD33 mitigated amyloid-β clearance and improved cognition [13, 17, 18]. Similarly, a positive effect on amyloid-β phagocytosis could be observed in CD33 knock-out THP-1 macrophages [16]. In humans a correlation between CD33, cognition and amyloid clearance is known, however, the concrete underlying mechanisms are still not well understood. There is an ongoing clinical trial that is testing the effects of a CD33

inhibitor in patients with mild to moderate AD (NCT03822208). Along those lines, the EU-wide PHAGO project (https://www.phago.eu) funded via the Innovative Medicines Initiatives aimed to develop tools and methods to study the functioning of CD33 and related pathways in AD in order to facilitate decisions about potential drug development programs.

While graphs are useful for describing the disease focused knowledge landscape about AD, the principal incompleteness of disease focused biological knowledge may result in disagreements to observed data. Moreover, graphs do not allow to produce quantitative insights and predictions. For this purpose ordinary (ODEs) and partial differential equations (PDEs) are frequently used in systems biology and systems medicine, as they are able to describe biological mechanisms in a quantitative way. However, their formulation requires a detailed understanding of biochemical reactions, which in the AD field is only available for specific processes, like for example amyloid-β aggregation [19, 20]. Moreover, fitting differential equations usually requires time resolved data and the possibility to perform intervention experiments (as knock-downs or stimulation), which is challenged by the fact that cell lines and mouse models in the AD field can most likely only mimic specific aspects of the human disease [21–23].

A principle alternative to differential equation systems are probabilistic graphical models and in particular Bayesian Networks (BNs), which are quantitative as well. However, standard BN implementations require normally or multinomially distributed data, which is not the case in many applications. Furthermore, structure learning of BNs is an NP hard problem, where the number of possible network structures grows super-exponentially with the number of nodes in the network [24]. Hence, modeling higher dimensional data with a BN raises severe concerns regarding structure identifiability.

Altogether, these challenges lead to the fact that the AD field lacks a comprehensive quantitative model of the interplay between relevant molecules and biological processes, including the role of CD33, up to the phenotype level.

In this work, we developed a - to our knowledge - first quantitative, multi-scale model focused on the multitude of mechanisms governing the CD33 molecule. Our model spans a variety of modalities, including gene expression, brain pathophysiology, demographic information and cognition scores. To address the challenges mentioned before, we started with a disease focused knowledge graph reconstruction, which we clustered into modules to significantly reduce dimen-

sionality. In the following we use the term "module" to denote a set of objects grouped together. Subsequently, we relied on our recently published Variational Autoencoder Modular Bayesian Network (VAMBN) algorithm [25], which is a hybrid Artificial Intelligence (AI) approach combining variational autoencoders [26] with modular Bayesian Networks [27], which is able to model arbitrary statistical distributions. We trained VAMBN on joint clinical and patient level gene expression data while employing a clustered knowledge graph reflecting incomplete prior knowledge about disease mechanisms and their interplay. A simulated knock-down of CD33 and predicted downstream effects could be experimentally validated with gene expression data from a cell line. Overall, we believe that our work helps to move closer towards a systemic and quantitative understanding of the disease, which is the prerequisite for finding urgently needed novel therapeutic options.

Results

In this work, we relied on AD patient data from the Religious Orders Study and Memory and Aging Project (ROSMAP) [28–30] for model training and the Mayo RNAseq Study (Mayo) [31] for external validation. An overview about clinical characteristics of the patients in these studies can be found in Table 1.

Overview about Modeling Strategy

Figure 1 shows an overview about our modeling strategy, which we call integrative VAMBN (iVAMBN), combining clinical and patient-level gene expression data with disease focused knowledge graphs. The first step of our workflow compiles an AD focused knowledge graph describing cause and effect relationships between biological processes, genes and pathologies. The generated graph consisted of 383 nodes and 607 edges. The graph was subsequently clustered into modules with the help of the Markov Clustering algorithm [32] to significantly reduce the number of variables for subsequent modeling steps. Genes within modules were annotated with AD disease mechanisms coming from the NeuroMMSig gene set collection [33].

Using patient-level clinical and gene expression data from post-mortem cerebral cortex tissues, in a second step the VAMBN algorithm was employed to quantitatively model relation-

Table 1: **Patient statistics.** Shown are the number of patients, their age in years (with mean and standard deviation), sex, APOE genotype (binary encoding for at least one present E4 allele), MMSE score (with mean and sd) and Braak stage.

	ROSMAP	Mayo
no. patients	221	82
age	87.95 ± 3.38	82.66 ± 7.61
sex		
male	68	33
female	153	49
APOE		
0	138	39
1	83	43
MMSE	13.16 ± 8.38	-
Braak		
1	7	-
2	6	-
3	42	-
4	71	6
5	88	35
6	7	41

ships between gene modules as well as phenotype related scores (Mini-Mental State Examination (MMSE), Braak staging) and demographic features based on ROSMAP data. ROSMAP was chosen for training of the algorithm, because of the comparably large number of patients (more than 200) and available MMSE plus Braak scores. VAMBN takes as input patient-level data hierarchically organized into pre-defined modules (here: either gene modules or a phenotype related module including i.e. MMSE plus Braak stages), original features (here: demographic and clinical variables like age, sex, APOE genotype, and brain region) and prior knowledge regarding their possible connections. The output is a probabilistic graphical model describing connections between modules and original features. There is a per-patient score for each module, and each of these scores can be further decoded into feature-level gene expression and phenotype data, respectively.

In the third step of our strategy we evaluated, whether our iVAMBN model could also explain gene expression data from the Mayo study. Notably, at this step we only considered the Braak stage in the phenotype module, because the Mayo study does not report MMSE scores. For that purpose we first re-trained our iVAMBN model on ROSMAP while leaving out MMSE scores and then assessed the marginal log-likelihood of the modified model on the Mayo dataset. We

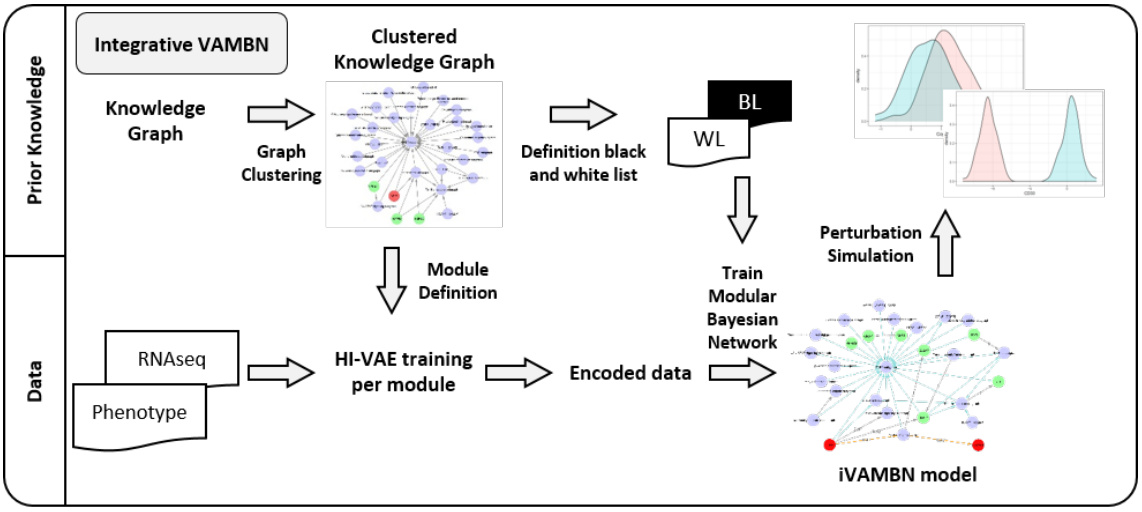

Figure 1: **The Integrative VAMBN (iVAMBN) approach.** The iVAMBN approach integrates gene expression data, clinical and patho-physiological (phenotype) measures (bottom left) into a joint quantitative, probabilistic graphical model. The method initially uses a knowledge graph (top left) for defining modules and for informing about potential connections between them. In a second step, a representation of each module using a Heterogeneous Incomplete Variational Autoencoder (HI-VAE) is learned. In a third step a modular Bayesian Network between autoencoded modules is learned while taking into account the information derived from the knowledge graph. Finally, the iVAMBN model is used to simulate gene perturbation (top right).

then tested the marginal log-likelihood of the true model against randomly permuted versions of the learned probabilistic graph. This allowed us to assess, in how far the model learned on ROSMAP could explain Mayo data better than expected by pure chance.

For the last step, we used our iVAMBN model trained on ROSMAP to simulate several therapeutic interventions, including a CD33 inhibition. Based on available data, we were able to experimentally validate the predicted effects of a CD33 inhibition using CD33 knockout gene expression data from a THP-1 monocyte cell line. More details about the entire iVAMBN approach can be found in the Methods section of this paper.

In the following we elaborate on the results obtained in each of these different steps, while technical details are provided in the Methods part of this article.

Knowledge Graph Compilation

As outlined in the previous section, our modeling approach started with the compilation and Markov clustering of a knowledge graph. The Markov clustering resulted in 32 modules, including 4 single gene modules, namely CD33, HSPB2, HSPB3, and MIR101-1. Most of the non-single gene modules comprised only two genes, while others had multiple combinations, like the GABA subgraph module with 289 genes. The exact number of genes clustered together as well as the result of a statistical over-representation analysis (hypergeometric test) using the AD focused gene set collection NeuroMMSig [33] can be found in Supplementary Table S1. A complete list of molecules within each module can be found in Supplementary Table S2. The modules were considered as nodes of a graph between them, where an edge was set between modules M_1, M_2, if in the original knowledge graph there was at least one gene in M_1 and one in M_2 that was connected via a directed path. The resulting (acyclic) module graph is shown in Figure S1.

Integrative Variational Autoencoder Modular Bayesian Network Model

Integrative VAMBN combines the advantages of Bayesian Networks with the capabilities of variational autoencoders, more specifically Heterogeneous Incomplete Variational Autoencoders (HI-VAEs) [34]. Briefly, the idea is to initially learn a low dimensional Gaussian representation of features mapping to each of the defined modules. HI-VAEs differ from classical variational autoencoders in the sense that they can be applied to heterogeneous input data of different numerical scales, potentially containing missing values. In a second step a Bayesian Network structure is then learned over the low dimensional representations of modules, resulting in a modular Bayesian Network. More details are presented in the Methods part of this paper and in [25].

We here trained an iVAMBN model using the identified modules (i.e. feature groups in the original data) as - potentially multivariate - nodes of a probabilistic graphical model. Noteworthy exceptions are described in detail in Supplementary Note S1. In cases where multiple features mapto one and the same module (i.e. the corresponding node / random variable in the probabilistic graphical model is multivariate), our method initially learns a low dimensional representation using a HI-VAE. Second, we learned the Bayesian Network structure connecting

these modules. At this stage it is possible to provide information about possible connections between modules given in the knowledge derived module graph (Supplementary Figure S1). We tested three different strategies to incorporate the information provided in the module graph:

- *completely data driven:* the entire Bayesian Network was only learned from data,
- *knowledge informed:* the module graph was either used to only initialize Bayesian Network structure learning, to enforce / white list the existence of specific edges, or used for a combination of both, and
- *completely knowledge driven:* strictly constrain edges between modules to those provided via the module graph, and additionally learned ones are only allowed to connect cognition scores, patho-physiological stages, and demographic features. All other possible edges are black listed, i.e. not allowed.

A systematic comparison of these strategies via a cross-validation yielded a better performance of the second strategy (knowledge informed), in which we used the module graph to white list edges and to initialize a greedy hill climbing based structure learning, see details in Methods Section and Supplementary Note S2. That means, iVAMBN was allowed to add additional edges, if the data provided according evidence.

We repeated the knowledge informed modular Bayesian Network learning 1000 times on random bootstrap sub-samples of the data drawn with replacement, hence allowing to quantify the statistical confidence of each inferred edge. The results of this analysis can be found in Supplementary Table S3.

In the following we only focus on the 130 edges that were found in at least 40% of the 1000 modular Bayesian Network reconstructions (Figure 2). Notably, this threshold was only chosen for better visualization purposes and to limit the subsequent discussion. Edges with lower bootstrap probability might also exist in reality despite lower statistical confidence. Nodes corresponding to sex, APOE status, and brain region were not connected to any other nodes with sufficient statistical confidence, meaning that these features might have no impact on the rest of the network. Nodes with only outgoing edges in the network (i.e. source nodes) were: the years of education, the age, and the single gene NAV3. The GABA subgraph (containing

more than 280 genes) and the phenotype module were leaf nodes, meaning they had no outgoing edges. Only patient age had a direct influence on CD33. CD33 had eight directly influenced molecular mechanisms: the GABA subgraph, the Amyloidogenic subgraph (containing genes SRC and APBA2), the Acetylcholine signaling subgraph (containing genes ACHE and PRNP), the Prostaglandin subgraph, and the Chaperone subgraph (containing genes HSPB6, CXCL8, and CCR2). Also, the single gene module, TRAF1, was a child of CD33. Altogether, CD33 had a predicted causal influence on every node, except for the source nodes.

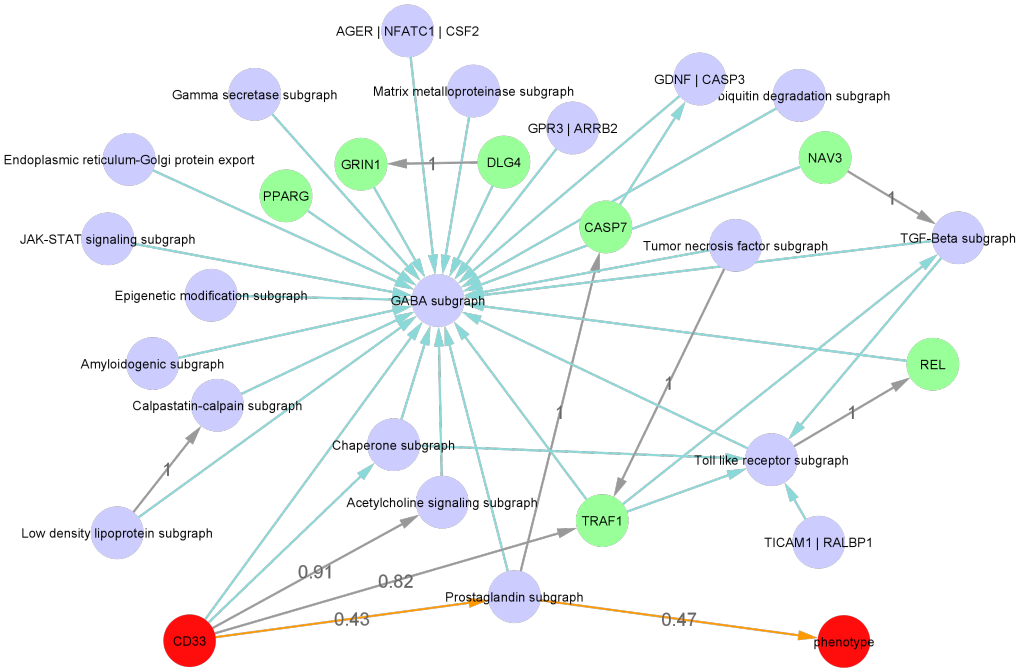

Figure 2: **Network representation of iVAMBN model for ROSMAP data.** Shown are the learned (grey) and knowledge-derived (green) edges between gene modules (purple nodes), single gene modules (green) and CD33 and phenotype module (red). All these edges appeared with bootstrap frequency > 0.4. The newly inferred shortest path between CD33 and phenotype is displayed in orange. Other edges with bootstrap frequency > 0.4 have been removed for visualization purposes, except for those six edges which were trained with a bootstrap confidence of 1.

Model reveals path between CD33 and disease phenotype

As shown in Figure 2 the shortest path between CD33 and the disease phenotype was observed through the Prostaglandin subgraph. All the edges from this connection were newly learned

from data, meaning that they had not previously been identified in the knowledge graph. Nevertheless, these correlations have been previously reported in the literature: Prostaglandines are eicosanoides, which were found to play a role in memory learning and neuroinflammation [35,36]. A major producer is activated microglia, which itself is activated through amyloid-β and produces inflammatory cytokines [37]. Currently, microglia and their effects on AD is a major focus in the field of research [38, 39]. Also, PGD2, a prostaglandin mainly synthesized in neurons, was previously found to be upregulated in AD patients [40]. Pairwise correlation plots between the genes of the prostaglandin pathway and CD33 or phenotype can be found in Supplementary Figure S3.

In total, 130 of the 162 edges of the bootstrapped iVAMBN model were newly learned from the data and had not been previously identified within the literature derived knowledge graph. Out of these 130 edges, six edges had a bootstrap confidence of 100%, meaning that they were learned consistently from 1000 random sub-samples of the data. A list of these edges can be found in Table 2.

Table 2: **Consistently newly learned edges in iVAMBN model.** All edges were found in each of 1000 network reconstructions from randomly subsampled data.

from	to
DLG4	GRIN1
Tumor necrosis factor subgraph	TRAF1
Toll like receptor subgraph	REL
Low density lipoprotein subgraph	Calpastatin-calpain subgraph
Prostaglandin subgraph	CASP7
NAV3	TGF-Beta subgraph

These high confidence edges demonstrated strong pairwise correlations between connected modules. NAV3, for example, had a strong negative correlation with MAVS, a member of the TGF-Beta subgraph module (Figure 3 left). In contrast to that SRSF10 and CREB1, members of the Low density lipoprotein subgraph and Calpastatin-calpain subgraph modules, were strongly positive correlated (Figure 3 right).

Although no direct correlation between NAV3 and MAVS is known, their effects are both linked to AD. NAV3, which is predominantly expressed in the nervous system, is increased in AD patients [41], while MAVS encodes a gene that is needed for the expression of beta interferon and thus contributes to antiviral innate immunity and may protect the cells from apoptosis [42].

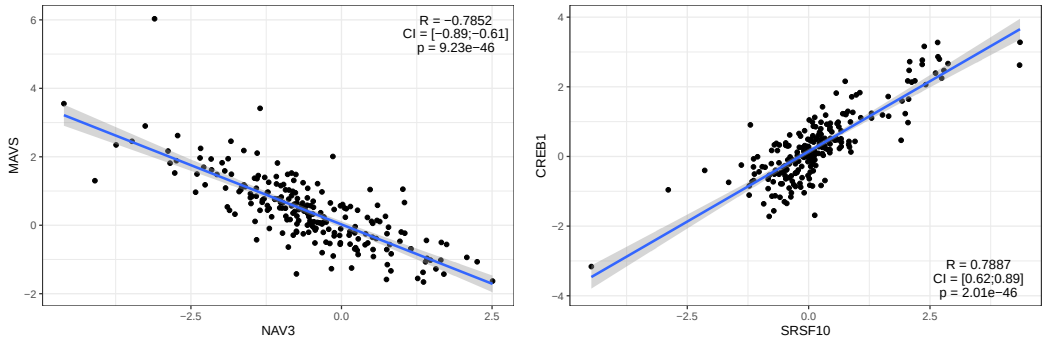

Figure 3: **Quantitative relationships learned by iVAMBN.** Each correlation (R) is shown along with its confidence interval (CI) and multiple testing adjusted p-value. Left: Correlation of NAV3 with TGF-Beta subgraph module member MAVS. Right: Correlation of Low density lipoprotein subgraph module member SRSF10 with CREB1, a member of the Calpastatin-calpain subgraph module. Further plots can be found in Supplementary Figures S3 and S4.

Together with the strong negative correlation seen in the data, one can hypothesize that the increased level of NAV3 in AD leads to a decreased level of MAVS, which elevates apoptosis of the cells.

The strong positive correlation between SRSF10 and CREB1 linked the Low density lipoprotein (LDL) and Calpastatin-calpain subgraphs. LDL is a major APOE receptor, which is the strongest genetic factor for AD, where different alleles are either risk or protective alleles [43]. APOE is also linked to amyloid-β, whose production is increased with elevated activity of calpain due to the decreased levels of calpastatin. Calpastatin is also linked to synaptic dysfunction and to the tau pathology of AD [44, 45]. Tau is another protein that accumulates in the brains of AD patients. The exact underlying mechanisms here are still unknown, but regulatory mechanisms of calpain are highly influenced by Calcium (Ca^{2+}) influx and increased intracellular calcium levels are a main reason for the loss of neuronal function in AD [44–46]. Changes in the Calpastatin-calpain mechanism may therefore also lead to reduced amyloid-β deposition.

External Validation of iVAMBN model

We assessed the ability of the model to explain normalized gene expression data from an independent study, Mayo. Notably, all gene expression data was from the same brain region, namely the cerebral cortex. However, Mayo does not contain MMSE scores. Therefore, we first trained

a modified version of our iVAMBN model on ROSMAP, which only contained the Braak score in the phenotype module, but otherwise had the edges shown in Figure 2. The full list of edges of this model together with their corresponding bootstrap confidences can be found in Supplementary Table S3. We then explored the marginal log-likelihood $\log p(data \mid graph)$ of the model on the Mayo dataset and subtracted the marginal log-likelihood obtained by 1000 random permutations of the network (Figure 4), resulting in an empirical p-value. This showed that our model could explain Mayo gene expression data significantly better than randomly permuted networks ($p = 0.035$) despite the clinical differences between patients in both studies shown in Table 1.

In addition, we trained a separate iVAMBN model on MSBB data and explored the overlap with the ROSMAP model at different thresholds of the bootstrap confidence (Supplementary Figure S5). At the previously chosen 40% threshold the overlap of the newly learned edges contained in the iVAMBN models trained on ROSMAP and MSBB was statistically significant, even if edge directions were considered (hypergeometric test, $p < 1e - 38$).

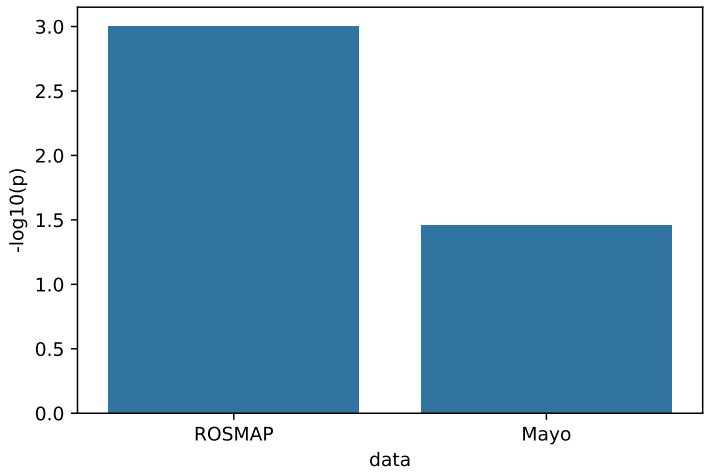

Figure 4: **External model validation.** Statistical significance $-log_{10}(p)$ value of the marginal log-likelihood of the model when evaluated on the training data (ROSMAP) and external validation data (Mayo).

CD33 Down-expression Simulation

To understand the potential systemic consequences of a therapeutic intervention into CD33 we simulated its down-expression. This was achieved by a counterfactual down-expression (here: 9-fold) of CD33 in every patient (Figure 5 (top left)). Due to the fact that iVAMBN is a quantitative model, associated downstream consequences on biological mechanisms and phenotype could be predicted in every patient (see examples in Figure 5). CD33 down-expression simulation (left) results in higher scores of the prostaglandin pathway module (right).

In addition, iVAMBN predicted a significant increase of MMSE scores ($p < 0.001$, Figure 6 (left)), and also a significant decrease of Braak stages ($p < 0.001$, Figure 6 (right)). That means patients are not only predicted to improve the specific cognitive abilities tested by MMSE, but are also predicted to improve brain pathophysiology.

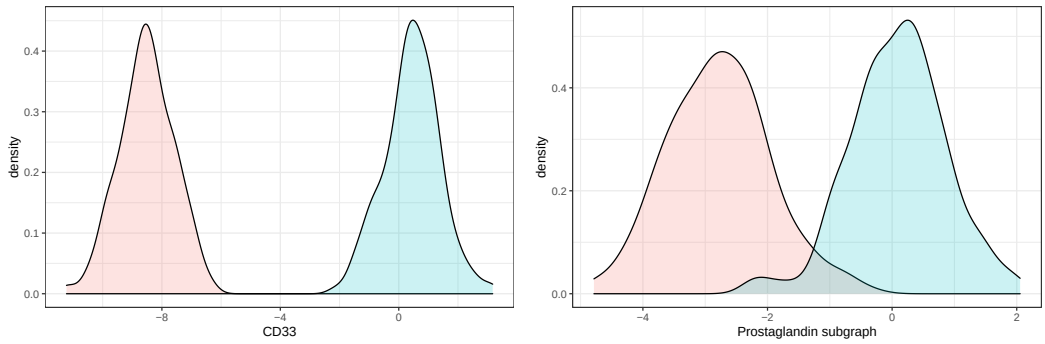

Figure 5: **Module distributions in original and simulated CD33 down-expression.** The blue curve describes the original distribution, while the red one describes the CD33 down-expression scenario. CD33 down-expression simulation (left) results in lower scores of the prostaglandin pathway module (right).

CD33 down-expression reveals significant changes in many mechanisms

Our iVAMBN model predicted significant effects on gene expression of 28 mechanisms and individual genes, respectively (Table 3). Significant changes were, for example, predicted for the genes CASP7 and TRAF7, and the prostaglandin and calpastatin-calpain mechanisms. But also the amyloidogenic mechanism is significantly differential expressed in a CD33 knock-down scenario.

Decreased expression of the amyloidogenic mechanism will thus result in patients with less

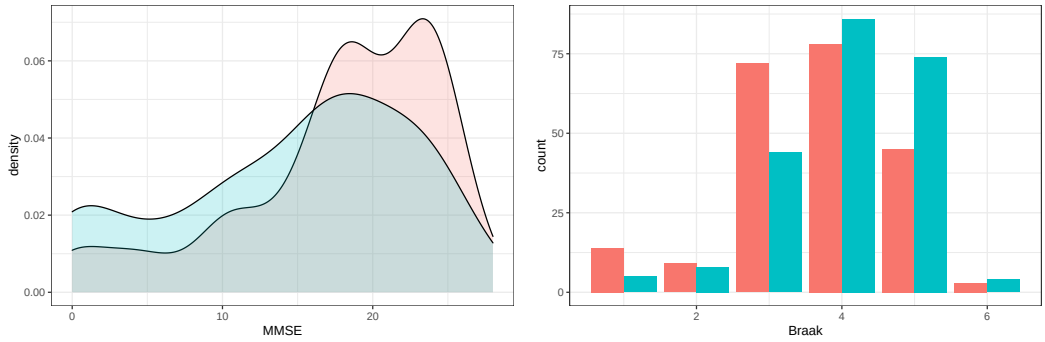

Figure 6: **Predicted changes on phenotype (MMSE and Braak stages) as a consequence of CD33 down-expression.** Distribution of MMSE and Braak stages in CD33 original (blue) and down-expressed (red) patients shows a significant improvement of scores and thus cognition as well as brain pathophysiology.

amyloid-β deposition. While this connection of the amyloidogenic mechanism and AD is clear, others need to be further explored.

The link between Calpastatin-calpain mechanism and AD was already described earlier. The key aspect is its negative influence on amyloid-β deposition.

PGD2, a prostaglandin mainly synthesized in neurons, was previously found to be upregulated in AD patients [40]. Prostaglandines are eicosanoide, which were found to play a role in memory learning and neuroinflammation [35, 36]. A major producer is activated microglia, which itself is activated through amyloid-β and produces inflammatory cytokines [37]. Currently, microglia and their effects on AD is a major focus in the field of research [38, 39]. Again, down-expression of the prostaglandin may result in reduced amyloid-β deposition. Altogether, the vast majority of significantly differential expressed gene sets was highly linked to AD through the amyloid-β cascade.

Experimental validation with cell line data

We checked whether our iVAMBN based predictions experimentally agreed with cell line gene expression data, specifically reflecting wild type (WT) and CD33 knock-out (KO). Our analysis (see details in Methods part) revealed significant changes of 23 AD associated mechanisms and genes in KO versus WT. Interestingly, 19 out of these 23 mechanisms overlapped with those predicted by iVAMBN (Table 3). Likewise, iVAMBN predicted significant changes of 22 genes

Table 3: **Statistical significance of gene modules.** The table shows results of a global test [47], assessing the differential gene set expression of each gene module between WT and down-expression/KO of CD33. P-values of the test within simulated scenario, as well as, p-values from cell line KO are reported and corrected for multiple testing using the Benjamini-Hochberg method. The agreement of both tests is described in the last column, meaning if both tests are either significant or non-significant (+) or if they don't show same direction of significance (-). For GRIN1 no p-value could be computed, as that gene is not present in the cell line data.

Gene module	p-value simulated KD	p-value cell line KO	agreement significance
GABA subgraph	2.75e-04	3.60e-15	+
Toll like receptor subgraph	1.05e-26	1.05e-13	+
Prostaglandin subgraph	6.99e-109	1.02e-09	+
TGF-Beta subgraph	0.592	8.79e-11	-
Calpastatin-calpain subgraph	3.14e-91	5.41e-09	+
JAK-STAT signaling subgraph	0.454	2.91e-11	-
AGER / NFATC1 / CSF2	5.78e-41	0.0129	+
Chaperone subgraph	2.84e-75	2.02e-09	+
REL	4.45e-18	9.96e-11	+
Ubiquitin degradation subgraph	5.15e-20	1.06e-06	+
GRIN1	1.92e-132	NA	
PPARG	2.20e-04	1.78e-03	+
GDNF / CASP3	1.06e-17	2.98e-11	+
Gamma secretase subgraph	4.36e-10	1.93e-03	+
Epigenetic modification subgraph	6.90e-58	7.64e-03	+
TICAM1 / RALBP1	1.46e-16	0.0561	-
Amyloidogenic subgraph	4.54e-69	9.11e-10	+
Tumor necrosis factor subgraph	0.0997	0.769	+
Acetylcholine signaling subgraph	6.74e-04	0.337	-
Matrix metalloproteinase subgraph	0.0708	2.74e-10	-
NAV3	0.176	3.81e-07	-
TRAF1	1.66e-95	2.26e-08	+
CASP7	1.75e-138	0.151	-
GPR3 / ARRB2	4.87e-04	8.02e-04	+
Endoplasmic reticulum-Golgi protein export	5.19e-29	1.78e-11	+
Low density lipoprotein subgraph	0.891	8.11e-06	-
DLG4	5.85e-93	3.44e-07	+
CD33	3.33e-307	8.06e-08	+

and gene sets, respectively, out of which only 3 were false positives at a false discovery rate threshold of 5%. Notably one of the false positive predictions (TICAM1 / RALBP1) had an adjusted p-value of 5.6% in the experimental data.

Overall, we thus observed a high degree of overlap between the dysregulated mechanisms and those predicted by the iVAMBN model, indicating that our model aligns well with the cell line data.

Simulation of the perturbation of other candidate targets

For comparison reasons, we further simulated the effect on the phenotype of a 9-fold up- or down-regulation of all other genes in our model, which showed a directed path to the phenotype module. Genes belonging to modules which were not an ancestor of the phenotype module were excluded, because they could not have any effect on the phenotype according to our model. We simulated for each candidate target an up- as well as a down-regulation.

The simulated dys-regulations showed that none of the candidate targets had a predicted effect on the phenotype stronger than CD33 (Figure S5). Only TRAF6 and TGFB3 down-regulation as well as up-regulation of APBA2, TRAF5 and SALL1 were predicted to increase the mean MMSE score by more than two points, compared to a predicted increase by almost five points via CD33 perturbation.

APBA2 is known to interact with APP and therefore plays a role in the amyloidogenic pathway [48, 49]. TRAF6 was identified in multiple experiments as target of miR-146a which is a key regulator of innate immunity that is up regulated in AD pathology affected brain regions and might also has an effect on amyloid-β metabolism [50]. It was found that treatment with a miR-146a agomir inhibits TRAF6 expression and reduced the cognitive impairment in AD mice [51].

Discussion

The here presented work is the first to demonstrate, to our knowledge, that one can integrate gene expression and clinical data together with qualitative knowledge about cause-and-effect relationships into a quantitative, system medical model of AD. This was achieved via an AI based method, which we combined with a knowledge graph representation of AD. We could show that a simulated CD33 down-expression agrees well with experimental gene expression KO data from a THP-1 cell line. Overall, our model could thus help to understand and quantify intervention effects on a multi-scale biological system level and thus aid the identification of novel therapeutic targets, which are urgently needed in the AD field.

Our model predicted that CD33 down-regulation would yield a significant effect on cognition (MMSE) and brain pathophysiology (Braak scores) through the prostaglandin pathway.

Although the role of prostaglandins is known to play a role in memory, learning and neuroinflammation [35, 36], the exact mechanism by which cognition is affected remains unknown, but seems to be coupled to amyloid-β deposition through microglia. In AD mice, a knockout of CD33 mitigated amyloid-β clearance and improved cognition [17, 18]. A positive effect on amyloid-β phagocytosis could also be observed in CD33 KO THP-1 macrophages [16].

Despite the evidence for a positive effect on cognition, we should mention that CD33 as a possible drug target has possible caveats that have been discussed in the literature [14]: i) It is not clear whether the genetic association of CD33 to AD is causal or just due to linkage disequilibrium with the true causal variant. ii) It is so far not entirely clear, how to therapeutically manipulate the expression level of CD33 in an optimal manner. iii) There might be safety issues due to the fact that CD33 is important for inhibiting immune responses and mediating self-tolerance. Systemic CD33 inhibition could potentially induce inflammatory autoimmune diseases. We therefore see the investigation of CD33 conducted in this paper more as a showcase for our iVAMBN approach rather than making any specific recommendation regarding the therapeutic value of CD33. Integrating known side effects of approved drugs targeting specific proteins in our model's graph structure could provide hints on possible side effects and is an interesting point for further research.

Altogether we see the impact of our work two-fold: first, we have introduced a novel multi-scale, quantitative modeling approach (iVAMBN), which is widely applicable in systems medicine, specifically in situations, where only a partial mechanistic understanding of biological phenomena is given. Secondly, our developed model can be further explored by the AD field and could aid a better understanding of the disease as well as identification of novel therapeutic options.

Methods

AD Knowledge Graph

A major part of this study is a BEL (`https://bel.bio/`) encoded, knowledge graph, which was initially compiled via text mining and later on manually curated via literature. In general, the BEL language helps to build a computer-process-able cause-and-effect relationship model. Each

BEL statement consists of a subject and an object, connected through a relation. Subjects and objects could be many different entities, like genes, proteins or RNA, but also biological processes, pathologies or even chemicals. Therefore, the relations have many different facets, as well. These could be relations like *increases*, *decreases* or *association*, describing the interaction between subject and object. But there are also relationships describing something like a membership of subject and object, for example *hasComponent* and *isA*. The BEL model used here, is an enriched version of the AD cause-and-effect relationship model published in [11] and can be found in the github repository. The enrichment was done around the two genes CD33 and TREM2, such that detailed knowledge about these two genes was gathered in the context of AD.

A filtering step was necessary, in order to get only entities measured in the gene expression data. In this case only gene and protein entities from the knowledge graph can be used. Additionally, the knowledge graph was filtered for only causal interactions, such as *increases*, *decreases*, or *regulates*, resulting in a network with 431 nodes and 673 edges. From that we only took the largest connected component to reduce the dimensionality. Hence, the used graph during our study consisted of 383 nodes and 607 edges, in which any two nodes were connected through some path.

Clustering of Filtered Knowledge Graph One of the key aspects of iVAMBN is grouping of input features (genes, pathophysiological and clinical features) into modules in order to allow for a statistically stable identification of a Bayesian Network structure in a subsequent step. For identifying modules of genes we clustered the knowledge graph with the help of different graph clustering algorithms:

- the Markov Cluster algorithm [32, 52] implemented in the *MCL* package in R [53].
- edge betweenness [54] community detection implemented in the R package *igraph* [55]
- infomap [56] community finding method implemented in the R package *igraph* [55]

After clustering, genes being part of a single cluster were assigned to a corresponding module. Genes being not clustered but only connected to one cluster, were merged into that cluster. Genes being connected to multiple clusters were kept as single gene modules (modules consisting of a single feature) for further analysis. We selected the best clustering algorithm according

to multiple metrics described in [57] including internal density, number of edges inside clusters, average degree, expansion, cut ratio, conductance, and norm cut. Based on these metrics the average ranking of each graph clustering algorithm was computed with the rational in mind, that each cluster should have an high internal density and sparse connections across clusters. This resulted in choosing the markov clustering algorithm for further analyses. The metrics for each clustering algorithm can be found in Supplementary Table S4.

Annotation of Modules with AD Disease Mechanisms For each module, an over-representation analysis for AD associated disease mechanisms was conducted. AD associated mechanisms were retrieved from the NeuroMMSig database [33]. For that purpose, the *enricher* function from the *clusterProfiler* package in R was used, which allows to use user-defined gene set annotations for a hypergeometric test [58]. We annotated each module with the most significant NeuroMMSig gene set after multiple testing correction via control of false discovery rate (Benjamini-Hochberg method).

Gene Expression Data Analysis

RNAseq data from several observational clinical studies, as well as RNAseq data from a cell line knockout experiment, were used in this work. The patient data were from i) the Religious Orders Study and Memory and Aging Project (ROSMAP) [28–30], and ii) the Mayo RNAseq Study (Mayo) [31]. The last one contains two separate datasets referring to separate brain regions, namely cerebellum (CBE) and temporal cortex (TCX). Both studies were accessed through the AMP-AD Knowledge Portal at Synapse using the data deposited in the RNAseq Harmonization Study. The used data are gene counts provided as gene count matrices that had been generated using STAR [59]. Gene counts were normalized to log counts per Million (logCPMs) and counts from AD patients were scaled against the healthy control data within each study. That means for each AD sample and gene the corresponding mean expression value of the same gene in cognitively normal subjects was subtracted. Subsequently we divided the values by the standard deviation of the gene in healthy controls. That means raw expression values were converted into abnormality scores.

After that, the datasets were filtered for AD patients only, resulting in 221 samples for

ROSMAP, and 62 samples in each of the two Mayo studies. Further filtering was done based on the brain region the samples were taken from. While all brain regions in ROSMAP could be mapped to the cerebral cortex via the Uber-anatomy ontology (UBERON) [60], that could only be done for the temporal cortex part of the Mayo study, meaning that the cerebellum samples were excluded. For making the expression data across studies comparable, a batch correction with ComBat [61] was applied to the scaled AD data. This normalized, scaled, and batch corrected data was then used for further analysis steps.

The cell line RNAseq data used during this study is from a THP-1 monocyte cell line with two different genetic backgrounds and two treatments. It can be found under GEO accession GSE155567. A sample could have either wild-type CD33 or a knocked out CD33 gene, plus either a control vector or a SHP-1 knock-down vector, resulting in four different conditions: i) *wild-type with control*, ii) *wild-type with SHP-1 knock-down vector*, iii) *CD33 knockout with control vector*, and iv) *CD33 knockout with SHP-1 knock-down vector*. There were 6 biological replicates per condition. Within the here presented study, only samples containing the control vector were used, resulting in twelve used samples. Therefore samples from condition 1 were called as *wild-type (WT)* samples and samples from condition 3 as *knockout (KO)* samples. Reads were aligned with STAR and gene counts were generated via the *featureCounts* function of the *Rsubread* package [62]. More details about the data can be found in [16] and under GEO accession GSE155567.

Variational Autoencoders (VAE)

Variational autoencoders [26] are one of the most frequently used type of unsupervised neural network techniques. They can be interpreted as a special type of probabilistic graphical model, which has the form $Z \to X$, where Z is a latent, usually multivariate standard Gaussian, and X a multivariate random variable describing the input data. Moreover, for any sample (x, z), we have $p(x \mid z) = N(\mu(z), \sigma(z))$. One of the key ideas behind VAEs is to variationally approximate

$$\log q(z|x) = \log N(z \mid \mu(x), \sigma(x)) \tag{1}$$

This means that $\mu(x)$ and $\sigma(x)$ are the multivariate mean and standard deviation of the approximate posterior $q(z \mid x)$ and are outputs of a multi-layer perceptron neural network (the encoder) that is trained to minimize for each data point x the criterion

$$\log(x) \geq \frac{1}{2} \sum_{j=1}^{D} \left(1 + \log \sigma_j(x)^2 - \mu_j(x)^2 - \sigma_j(x)^2\right) + \frac{1}{L} \sum_{l} \log p(x|z^{(l)}) \tag{2}$$

Here the index j runs over the D dimensions of the input x, and $z = \mu(x) + \sigma(x) \odot \epsilon^{(l)}$ with $\epsilon^{(l)} \sim N(0, I)$ being the lth random sample drawn from a standard multivariate Gaussian, and $\odot$ denotes an element-wise multiplication. Notably, the right summand corresponds to the reconstruction error of data point x by the model, whereas the first term imposes a regularization. We refer to [26] for more details.

Heterogeneous Incomplete Variational Autoencoders (HI-VAE)

Variational autoencoders were originally developed for homogeneous, continuous data. However, in our case variables grouped into the phenotype module do not fulfill this assumption, because Braak stages and MMSE scores are discrete ordinal. In agreement to our earlier work [25] we thus employed the HI-VAE [34], which is an extension of variational autoencoders and allows for various heterogeneous data types, even within the same module. More specifically, the authors suggest to parameterize the decoder distribution as

$$p(x_j \mid z) = p(x_j|\gamma_j = h_j(z)) \tag{3}$$

where $h_j(\cdot)$ is a function learned by the neural network, and γ_j accordingly models data modality specific parameters. For example, for real-valued data we have $\gamma_j = (\mu(z), \sigma_j(z)^2))$, while for ordinal discrete data we use a thermometer encoding, where the probability of each ordinal category can be computed as

$$p(x_j = r \mid \gamma_j) = p(x_j \leq r \mid \gamma_j) - p(x_j \leq r - 1 \mid \gamma_j) \tag{4}$$

with

$$p(x_j \leq r \mid z) = \frac{1}{1 + \exp(-(\theta_j(z) - h_j(z)))} \tag{5}$$

The thresholds $\theta_j(z)$ divide the real line into R regions, and $h_j(z)$ indicates, in which region z falls. The data modality specific parameters are thus $\gamma_j = \{h_j(z), \theta_1(z), ..., \theta_{R-1}(z)\}$ and are modeled as output of a feed forward neural network.

According to [34] we use batch normalization to account for differences in numerical ranges between different data modalities.

For multi-modal data and in particular discrete data a single Gaussian distribution may not be a sufficiently rich representation in latent space. Hence, the authors propose to replace the standard Gaussian prior distribution imposed for z in VAEs by a Gaussian mixture prior with K components:

$$s \sim Categorical(\pi) \tag{6}$$

$$z \mid s \sim N(\mu(s), I_K) \tag{7}$$

where $\pi_k = \frac{1}{K}$ for $k = 1, 2, ..., K$ and s is a one-hot vector encoding of the mixture component. We evaluated different choices of K using a 3-fold cross-validation, while employing the reconstruction error $\frac{1}{L}\sum_l \log p(x|z^{(l)})$ as an objective. In conclusion it turned out that $K = 1$ component was an optimal choice for all modules in our iVAMBN model.

Modular Bayesian Networks

Let $X = (X_v)_{v \in V}$ be a set of random variables indexed by nodes V in a directed acyclic graph (DAG) $G = (V, E)$. In our case each of these nodes corresponds either to lower dimensional embedding of a group of variables (i.e. module) in the original data, or to an original features (e.g. biological sex) in the dataset. According to the definition of a Bayesian Network (BN), the joint distribution $p(X_1, X_2, \ldots, X_n)$ factorizes according to

$$p(X_1, X_2, \ldots, X_n) = \prod_{v \in V} p(X_v \mid X_{pa(v)}) \tag{8}$$

where $pa(v)$ denotes the parent set of node v [27]. In our case random variables follow either a Gaussian or a multinomial distribution, i.e. the BN is hybrid. Notably, no discrete random variable was allowed to be a child of a Gaussian one.

Since the BN in our case is defined over low dimensional representations of groups of variables, we call the structure Modular Bayesian Network (MBN). Notably, a MBN is a special instance of a hierarchical BN over a structured input domain [63–66].

A typical assumption in (M)BNs is that the set of parameters $(\theta_v)_{v \in V}$ associated to nodes V are statistically independent. For a Gaussian node v parameters can thus be estimated by fitting a linear regression function with parents of v being predictor variables [27]. Similarly, for a discrete node $\tilde{v}$ having only discrete parents, parameters can be estimated by counting relative frequencies of variable configurations, resulting into a conditional probability table.

Quantitative Modeling Across Biological Scales via iVAMBN

Model Training

The here presented *Integrative Variational Autoencoder Modular Bayesian Network (iVAMBN)* approach (Figure 1), integrates different biological scales together with a knowledge graph into the previously published Variational Autoencoder Modular Bayesian Network (VAMBN) approach [25]. More precisely, there are four steps to build an iVAMBN model: i) Definition of modules of variables, ii) Training of a HI-VAE for each module, iii) Definition of logical constraints for possible edges in the MBN, and iv) Structure and parameter learning of the MBN using encoded values for each module. These four steps result from the fact that HI-VAEs (as well as any other variants of variational autoencoders) themselves can be interpreted as specific types of BNs and thus the overall log-likelihood of an iVAMBN model can be decomposed accordingly. That means the overall iVAMBN model can be interpreted as a special type of Bayesian Network, see [25] for details.

The four model building steps were followed in the application of the iVAMBN approach in this work as well. The modules of variables were mainly defined through the previously explained Markov clustering of the knowledge graph, plus an additional module summarizing MMSE (Mini–Mental State Examination) and Braak stages into one *phenotype* module. MMSE

measures cognitive impairment by testing the orientation in time and space, recall, language, and attention, while Braak stages refer to the degree of biological brain pathology [67]. Some non-assigned genes, were directly treated as nodes in the MBN construction and thus also called gene modules. The same was done for demographic features, like sex, age, years of education and the APOE genotype.

For training the HI-VAEs for each module a hyperparameter optimization (grid search) was implemented over learning rate (learning rate $\in \{0.001, 0.01\}$) and minibatch size (minibatch size $\in \{16, 32\}$) as in [25]. Each parameter combination was evaluated with the reconstruction loss as objective function in a 3-fold cross-validation scenario.

In general the number of possible MBN DAG structures for n nodes grows super-exponentially with n [24], making identification of the true graph structure highly challenging. Therefore, our aim was to restrict the set of possible DAGs a priori as much as possible via knowledge based logical constraints. More specifically we imposed the following causal restrictions:

- Nodes defined by demographic or clinical features (like age, gender, APOE genotype, and brain region) can only have outgoing edges.
- The phenotype module (= clinical outcome measures) can only have incoming edges.
- Genes and gene modules can not influence demographic or clinical features, except the age.

To additionally integrate prior knowledge defined through the knowledge graph, we tested three different strategies while building a MBN:

1) **Completely data driven**: The knowledge graph is completely ignored for structure learning.
2) **Knowledge informed**: The knowledge graph is used in the greedy hill climbing algorithm for structure learning i) as starting point, ii) as white list (intending that those edges were defined as pre-existing), or iii) as both.
3) **Completely knowledge driven**: The knowledge graph provides the structure of the MBN and additional connections are only allowed for demographics or the phenotype module.

Structure learning of the MBN was always performed via a greedy hill climber using the Bayesian Information Criterion for model selection. We employed the implementation provided in R-package *bnlearn* [68].

Evaluating the Model Fit

To evaluate the fit of the overall iVAMBN model we employed the generative nature of our model: Following a topological sorting of the nodes of the DAG of the MBN we first sampled from the distribution of each node conditional on its parent. Notably, for MBN nodes representing modules this amounted to sample from the posterior of the according HI-VAE, which in practice can be realized via injection of normally distributed noise, see Section Variational Autoencoders, Eq. (2). Subsequently, the random sample was then decoded via the HI-VAE. Altogether we thus generated as many synthetic subjects as real ones. We then compared the marginal distribution of each variable based on the synthetic and the real data. Results, including summary statistics and Kullback-Leibler divergences are shown in the supplementary material. Furthermore, we compared the correlation matrices of synthetic and real data.

CD33 down-expression simulation and analysis

To be able to simulate a down-expression of CD33, we first shifted the distribution of CD33 such that it reflects a 9-fold down-expression of CD33. In agreement to the theory of Bayesian Networks this operation made CD33 conditionally independent of its parents in the MBN, which amounts to deleting any of its incoming edges and resulted into a mutilated MBN. Afterwards we exploited the fact that iVAMBN is a generative model. That means we first drew samples from the conditional densities of the mutilated MBN. Practically this amounted to first topologically sort the nodes in the MBN, hence exploiting the fact that the underlying graph structure cannot have cycles. Subsequently, samples were drawn from the statistical distribution of each node while conditioning on the value of its parents. The result was a per-sample module activity scores, which we then decoded through our HI-VAE models into single gene scores.

Differences between the wild-type and simulated down-expression samples were investigated afterwards via multiple statistical hypothesis tests: First, a linear regression was used to model

the down-expression effect on gene expression and on the different phenotype scores. Second, the *globaltest* package in R was used to test the differential expression of specific gene sets between the wild-type and simulated down-expression group [47]. Those tested gene sets were here defined through the modules' genes used in the MBN, meaning that we tested for differential expression of MBN's gene modules. P-values were adjusted for multiple test scenario with the help of the *subsets* option of *globaltest* and via calculating the false discovery rate. The globaltest for gene sets, as well as the fold change analysis, was also applied to the cell line WT and KO data to be able to validate the results.

Effects of the perturbation of other candidate targets were simulated similarly as the CD33 knock-down. Again, the distribution of the respective target was shifted such that it reflected a 9-fold down- or up-regulation. The module was identified to which the candidate target had been assigned, and all variables (including the perturbed target) mapping to that module were encoded via the previously trained HI-VAE for the module. Subsequently, the effects on the phenotype could be predicted in the same way as described for CD33.

Acknowledgements

This project has received partial funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115976. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

This work was partially developed in the Fraunhofer Cluster of Excellence "Cognitive Internet Technologies" and partially supported via the Fraunhofer Center for Machine Learning.

The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Data generation was supported by the following NIH grants: P30AG10161, P30AG72975, R01AG15819, R01AG17917, R01AG036836, U01AG46152, U01AG61356, U01AG046139, P50 AG016574, R01 AG032990, U01AG046139, R01AG018023, U01AG006576, U01AG006786, R01AG025711, R01AG017216, R01AG003949, R01NS080820, U24NS072026, P30AG19610, U01AG046170, RF1AG057440, and U24AG061340, and the Cure PSP, Mayo and Michael J Fox foundations, Arizona Department of Health Services and the Arizona Biomedical Research Commission. We thank the participants of the Reli-

gious Order Study and Memory and Aging projects for the generous donation, the Sun Health Research Institute Brain and Body Donation Program, the Mayo Clinic Brain Bank, and the Mount Sinai/JJ Peters VA Medical Center NIH Brain and Tissue Repository. Data and analysis contributing investigators include Nilüfer Ertekin-Taner, Steven Younkin (Mayo Clinic, Jacksonville, FL), Todd Golde (University of Florida), Nathan Price (Institute for Systems Biology), David Bennett, Christopher Gaiteri (Rush University), Philip De Jager (Columbia University), Bin Zhang, Eric Schadt, Michelle Ehrlich, Vahram Haroutunian, Sam Gandy (Icahn School of Medicine at Mount Sinai), Koichi Iijima (National Center for Geriatrics and Gerontology, Japan), Scott Noggle (New York Stem Cell Foundation), Lara Mangravite (Sage Bionetworks).

References

[1] Hodson R. Alzheimer's disease. Nature. 2018 7;559:S1-1.

[2] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 2017 1;45.

[3] Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Research. 2011 1;39.

[4] Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research. 2018 1;46.

[5] Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Research. 2018 1;46.

[6] Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N, et al. Open Targets Platform: new developments and updates two years on. Nucleic Acids Research. 2019 1;47.

[7] Wang Y, Zhang S, Li F, Zhou Y, Zhang Y, Wang Z, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Research. 2019 11.

[8] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Research. 2019 11.

[9] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research. 2019 1;47.

[10] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research. 2014 1;42.

[11] Kodamullil AT, 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. Alzheimer's and Dementia. 2015 11;11:1329-39.

[12] Estus S, Shaw BC, Devanney N, Katsumata Y, Press EE, Fardo DW. Evaluation of CD33 as a genetic risk factor for Alzheimer's disease. Acta Neuropathologica. 2019 8;138.

[13] Jiang T, Yu JT, Hu N, Tan MS, Zhu XC, Tan L. CD33 in Alzheimer's Disease. Molecular Neurobiology. 2014 2;49.

[14] Zhao L. CD33 in Alzheimer's Disease – Biology, Pathogenesis, and Therapeutics: A Mini-Review. Gerontology. 2019;65.

[15] Siddiqui SS, Springer SA, Verhagen A, Sundaramurthy V, Alisson-Silva F, Jiang W, et al. The Alzheimer's disease–protective CD33 splice variant mediates adaptive loss of function via diversion to an intracellular pool. Journal of Biological Chemistry. 2017 9;292.

[16] Wißfeld J, Nozaki I, Mathews M, Raschka T, Ebeling C, Hornung V, et al. Deletion of Alzheimer's disease-associated CD33 results in an inflammatory human microglia phenotype. GLIA. 2021.

[17] Griciuc A, Patel S, Federico AN, Choi SH, Innes BJ, Oram MK, et al. TREM2 Acts Downstream of CD33 in Modulating Microglial Pathology in Alzheimer's Disease. Neuron. 2019 9;103.

[18] Griciuc A, Serrano-Pozo A, Parrado A, Lesinski A, Asselin C, Mullin K, et al. Alzheimer's Disease Risk Gene CD33 Inhibits Microglial Uptake of Amyloid Beta. Neuron. 2013 5;78.

[19] Steckmann T, Awan Z, Gerstman BS, Chapagain PP. Kinetics of peptide secondary structure conversion during amyloid B-protein fibrillogenesis. Journal of Theoretical Biology. 2012 5;301:95-102.

[20] Proctor CJ, Pienaar IS, Elson JL, Kirkwood TB. Aggregation, impaired degradation and immunization targeting of amyloid-beta dimers in Alzheimers disease: A stochastic modelling approach. Molecular Neurodegeneration. 2012;7:32.

[21] Oblak AL, Forner S, Territo PR, Sasner M, Carter GW, Howell GR, et al. Model organism development and evaluation for late-onset Alzheimer's disease: MODEL-AD. Alzheimer's & Dementia: Translational Research & Clinical Interventions. 2020 1;6.

[22] Jankowsky JL, Zheng H. Practical considerations for choosing a mouse model of Alzheimer's disease. Molecular Neurodegeneration. 2017 12;12:1-22.

[23] Arber C, Lovejoy C, Wray S. Stem cell models of Alzheimer's disease: Progress and challenges. Alzheimer's Research and Therapy. 2017 6;9:1-17.

[24] Chickering DM, Meek C, Heckerman D. Large-Sample Learning of Bayesian Networks is NP-Hard. 2012 10.

[25] Gootjes-Dreesbach L, Sood M, Sahay A, Hofmann-Apitius M, Fröhlich H. Variational Autoencoder Modular Bayesian Networks for Simulation of Heterogeneous Clinical Study Data. Frontiers in Big Data. 2020 5;3:16.

[26] Kingma DP, Welling M. Auto-Encoding Variational Bayes. 2013 12.

[27] Heckerman D. In: Holmes DE, Jain LC, editors. A Tutorial on Learning with Bayesian Networks. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 33-82.

[28] Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. Nature Neuroscience. 2018 6;21:811-9.

692 [29] Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious
693 Orders Study and Rush Memory and Aging Project. Journal of Alzheimer's Disease. 2018
694 6;64.

695 [30] Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and Findings from the
696 Religious Orders Study. Current Alzheimer Research. 2012 6;9.

697 [31] Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, et al. Human
698 whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative
699 diseases. Scientific Data. 2016 12;3:160089.

700 [32] Dongen SV. Graph Clustering Via a Discrete Uncoupling Process. SIAM Journal on Matrix
701 Analysis and Applications. 2008 1;30.

702 [33] Domingo-Fernández D, Kodamullil AT, Iyappan A, Naz M, Emon MA, Raschka T, et al.
703 Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): A web
704 server for mechanism enrichment. Bioinformatics. 2017;33.

705 [34] Nazábal A, Olmos PM, Ghahramani Z, Valera I. Handling incomplete heterogeneous data
706 using VAEs. Pattern Recognition. 2020 11;107:107501.

707 [35] Bsci DT, Msc GK, Weisinger RS, Sinclair AJ. The role of eicosanoids in the brain. Asia
708 Pac J Clin Nutr. 2008;17:220-8.

709 [36] Ardura-Fabregat A, Boddeke EWGM, Boza-Serrano A, Brioschi S, Castro-Gomez S,
710 Ceyzériat K, et al. Targeting Neuroinflammation to Treat Alzheimer's Disease. CNS Drugs.
711 2017 12;31:1057-82.

712 [37] Biringer RG. The role of eicosanoids in alzheimer's disease. International Journal of Envi-
713 ronmental Research and Public Health. 2019 7;16.

714 [38] Candlish M, Hefendehl JK. Microglia Phenotypes Converge in Aging and Neurodegenerative
715 Disease. Frontiers in Neurology. 2021 5;12.

716 [39] Schwabe T, Srinivasan K, Rhinn H. Shifting paradigms: The central role of microglia in
717 Alzheimer's disease. Neurobiology of Disease. 2020 sep;143:104962.

718 [40] Iwamoto N, Kobayashi K, Kosaka K. The formation of prostaglandins in the postmortem
719 cerebral cortex of Alzheimer-type dementia patients. Journal of Neurology. 1989 2;236:80-4.

720 [41] Shioya M, Obayashi S, Tabunoki H, Arima K, Saito Y, Ishida T, et al. Aberrant microRNA
721 expression in the brains of neurodegenerative diseases: miR-29a decreased in Alzheimer
722 disease brains targets neurone navigator 3. Neuropathology and Applied Neurobiology.
723 2010 feb;36(4):320-30.

724 [42] Seth RB, Sun L, Ea CK, Chen ZJ. Identification and Characterization of MAVS, a Mitochon-
725 drial Antiviral Signaling Protein that Activates NF-κB and IRF3. Cell. 2005 sep;122(5):669-
726 82.

727 [43] Serrano-Pozo A, Das S, Hyman BT. APOE and Alzheimer's disease: advances in genetics,
728 pathophysiology, and therapeutic approaches. The Lancet Neurology. 2021 jan;20(1):68-80.

729 [44] Ferreira A. Calpain Dysregulation in Alzheimer's Disease. ISRN Biochemistry. 2012
730 10;2012:1-12.

731 [45] Vosler PS, Brennan CS, Chen J. Calpain-mediated signaling mechanisms in neuronal injury
732 and neurodegeneration. Molecular Neurobiology. 2008;38:78-100.

733 [46] McCarty MF, Dinicolantonio JJ, Lerner A. A fundamental role for oxidants and intracellular
734 calcium signals in Alzheimer's pathogenesis—and how a comprehensive antioxidant strategy
735 may aid prevention of this disorder. International Journal of Molecular Sciences. 2021 2;22:1-
736 27.

737 [47] Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of
738 genes: testing association with a clinical outcome. Bioinformatics. 2004 1;20.

739 [48] Abou-Fadel J, Vasquez M, Grajeda B, Ellis C, Zhang J. Systems-wide analysis unravels the
740 new roles of CCM signal complex (CSC). Heliyon. 2019 12;5:e02899.

741 [49] Jensen TMT, Albertsen L, Bartling CRO, Haugaard-Kedström LM, Strømgaard K. Prob-
742 ing the Mint2 Protein-Protein Interaction Network Relevant to the Pathophysiology of
743 Alzheimer's Disease. Chembiochem : a European journal of chemical biology. 2018 3.

[50] Goodall EF, Heath PR, Bandmann O, Kirby J, Shaw PJ. Neuronal dark matter: the emerging role of microRNAs in neurodegeneration. Frontiers in cellular neuroscience. 2013 10;7:178.

[51] Mai H, Fan W, Wang Y, Cai Y, Li X, Chen F, et al. Intranasal Administration of miR-146a Agomir Rescued the Pathological Process and Cognitive Impairment in an AD Mouse Model. Molecular therapy Nucleic acids. 2019 12;18:681-95.

[52] Dongen SV. Graph clustering by flow simulation. Proefschrift Universiteit Utrecht; 2000.

[53] Jäger ML. MCL: Markov Cluster Algorithm.. https://CRAN.R-project.org/package=MCL; 2015.

[54] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E. 2004 2;69:026113.

[55] Csardi G, Nepusz T, et al. The igraph software package for complex network research. InterJournal, complex systems. 2006;1695(5):1-9.

[56] Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences. 2008 1;105:1118-23.

[57] Arratia A, Mirambell MR. Clustering assessment in weighted networks. PeerJ Computer Science. 2021;7:1-27.

[58] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS: A Journal of Integrative Biology. 2012 5;16.

[59] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013 1;29:15-21.

[60] Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biology. 2012;13.

[61] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 1;8.

[62] Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Research. 2019 5;47.

[63] Parviainen P, Kaski S. Bayesian Networks for Variable Groups; 2016.

[64] Parviainen P, Kaski S. Learning structures of Bayesian networks for variable groups. International Journal of Approximate Reasoning. 2017 9;88:110-27.

[65] Becker AK, Dörr M, Felix SB, Frost F, Grabe HJ, Lerch MM, et al. From heterogeneous healthcare data to disease-specific biomarker networks: A hierarchical Bayesian network approach. PLOS Computational Biology. 2021 2;17:e1008735.

[66] Gyftodimos E, Flach PA. Hierarchical Bayesian Networks: An Approach to Classification and Learning for Structured Data. In: Vouros GA, Panayiotopoulos T, editors. Methods and Applications of Artificial Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 291-300.

[67] Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathologica. 1991 9;82:239-59.

[68] Scutari M. Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software. 2010;35.

Author contributions

HF designed and supervised the project. TR analysed the data. MS and AA helped with the analysis of the data. BS and CE generated the knowledge graph. TR and HF wrote the manuscript. All authors have read and approved the manuscript.

Supporting information

Supplementary Table S1: Module enrichment analysis

Supplementary Table S2: Module assignment

Supplementary Figure S1: Clustered knowledge graph.

Supplementary Note S1: iVAMBNs Module Definition

Supplementary Note S2: iVAMBNs Knowledge Integration

Supplementary Figure S2: Log likelihood loss (i.e. negative log-likelihood) of different knowledge integration strategies.

Supplementary Figure S3: Quantitative effect between modules of shortest path.

Supplementary Figure S4: Quantitative effect between modules of newly trained edges with confidence 1

Supplementary Table S3: Bootstrap confidence results

Supplementary Figure S5: Overlap of ROSMAP and Mayo network structures

Supplementary Figure S6: Effects on phenotype scores of up- and down-regulation simulations.

Supplementary Table S4: Graph Clustering Metrics

Supplementary Note S3: Evaluating the Model Fit

Supplementary Figure S7: Distribution of single feature in real data versus simulated data.

Supplementary Figure S8: Correlation matrix of real and drawn data.

Supplementary Appendix S1: Github repository of the code for this analysis `https://github.com/traschka/iVAMBN`

Additional information

Data Availability Statement

- All code can be found at: `https://github.com/traschka/iVAMBN`
- ROSMAP and Mayo: `https://adknowledgeportal.synapse.org/Explore/Studies/DetailsPage?Study=syn21241740`
- CD33 KO cell line: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155567`