TabPFN-CBM: A Concept-Interpretable Tabular Foundation Model Framework for Multimodal Alzheimer's Diagnosis

Thomas Yu*, Shu Yang*, Zixuan Wen, Boning Tong, Tianyu Han, Christos Davatzikos, Li Shen†, and for the Alzheimer's Disease Neuroimaging Initiative‡

University of Pennsylvania,
Philadelphia, PA 19104, USA
**Equal contribution*
†*E-mail: li.shen@pennmedicine.upenn.edu*

Foundation models have demonstrated powerful predictive capabilities in biomedical tasks, including the diagnosis of Alzheimer's disease (AD). Tabular foundation models, such as TabPFN, are especially well-suited for clinical applications where tabular data are ubiquitous. However, these models generally do not reveal biomedically meaningful decision pathways. This black-box nature limits their clinical adoption where interpretability is critical. Here, we propose TabPFN-CBM, the first concept bottleneck framework with a tabular foundation model backbone leveraging TabPFN to provide both high predictive performance and interpretability through concept-level explanations. We evaluate our model on multimodal tabular data from the Alzheimer's Disease Neuroimaging Initiative to predict diagnostic classes through an intermediate layer of clinically grounded AD biomarker concepts. Our model achieves competitive performance compared to baseline models while bringing concept interpretability of predicted diagnoses and enabling potential clinician-guided interventions. TabPFN-CBM is publicly available at `github.com/thomasyu28/tabpfn-cbm`.

Keywords: Concept bottleneck models; TabPFN; tabular foundation models; Alzheimer's disease diagnosis; QT-PAD biomarkers; multimodal imaging and genetic data.

1. Introduction

Alzheimer's disease is a progressive neurodegenerative disorder with a complex, multifactorial etiology and a major cause of morbidity and mortality worldwide. Approximately 57 million people are living with dementia globally, with AD accounting for an estimated 60–70% of these cases.[a] The prevalence of AD and related dementias is projected to exceed 100 million cases by 2050.[1,2] As AD remains largely incurable and current treatments show limited ability to alter the long-term disease course, early detection is critical to enable timely interventions and

‡Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

[a]World Health Organization, "Dementia", `https://www.who.int/news-room/fact-sheets/detail/dementia`, updated July 3, 2026.

inform treatment planning aimed at delaying disease progression. Diverse biomarkers including neuroimaging measurements, genetic variants, cognitive assessments, and fluid-based markers provide complementary information for the early identification and characterization of AD.[3,4] Realizing the diagnostic potential of these heterogeneous biomarkers requires models that can integrate multimodal measurements while retaining clinically meaningful interpretability.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) offers a widely used longitudinal resource that integrates different data modalities and clinical diagnoses, thereby enabling multimodal investigation. Although raw neuroimaging scans contain rich voxel-level information, direct modeling is computationally demanding and statistically challenging in the modest cohorts typical of Alzheimer's studies. Instead, Region-of-Interest (ROI) summaries (e.g., regional amyloid burden, cortical thickness) reduce dimensionality, retain anatomical meaning, and can be easily integrated with genetic and other subject-level variables in a standard table format, as demonstrated in prior AD multimodal studies.[5–8] Complementing these lower-level measurements, the Quantitative Templates for the Progression of Alzheimer's Disease (QT-PAD) framework[b] organizes a compact set of cognitive, imaging, and fluid biomarkers that characterize clinically relevant aspects of AD pathology and have supported previous tabular few-shot diagnosis.[9]

Most existing AI approaches to AD diagnosis and multimodal analysis with tabular data rely on conventional machine-learning algorithms or task-specific deep neural networks.[10–13] Tree-based ensembles and lighter-weight logistic regression remain strong baselines for biomedical tables because they perform reliably with limited samples and heterogeneous feature distributions. By contrast, standard deep neural networks often require larger training sets and careful architectural and optimization choices to learn competitive tabular representations.[14,15] Nevertheless, both conventional machine-learning models and task-specific neural networks primarily learn from the available study cohort. They do not natively incorporate external knowledge and often provide limited clinically meaningful explanations for their predictions. These limitations motivate the development of pretrained foundation models that can transfer knowledge across datasets and adapt to new predictive tasks using few labeled examples.

Tabular foundation models (TFMs) have recently emerged as a promising alternative for low-data structured prediction. TabPFN introduced the prior-data fitted network paradigm for tabular classification, in which a bi-dimensional transformer is pretrained on synthetic tables sampled from structural-causal priors to approximate posterior predictive inference. At test time, the labeled training table is supplied as context, enabling predictions for new samples via in-context learning without task-specific gradient updates.[16] The initial model was restricted to small numerical classification datasets, prompting a series of notable extensions. TuneTables learns a compact continuous context for a frozen TabPFN, enabling parameter-efficient adaptation to larger datasets,[17] whereas LoCalPFN combines nearest-neighbor retrieval with task-specific fine-tuning to construct locally adapted contexts for individual queries.[18] TabPFN v2 substantially broadened the core framework to support both classification and regression, numerical and categorical variables, outliers, and irrelevant features. After pretraining on ap-

[b] `https://www.pi4cs.org/qt-pad-challenge`

proximately 130 million synthetic datasets, it demonstrated strong out-of-the-box performance on tables containing up to 10,000 samples and 500 features.[19] Subsequent work confirmed its strong generalization across hundreds of datasets but also identified persistent limitations in high-dimensional, many-class, and large-scale settings, motivating divide-and-conquer inference and lightweight encoder-adaptation strategies.[20,21] More recent generations[22–24] have further expanded the applicable data regime: for example, TabPFN v2.5 targets datasets with up to 50,000 samples and 2,000 features and introduces deployment-oriented model distillation.[22] These properties make TabPFN particularly attractive for small, heterogeneous biomedical cohorts such as ADNI.

Despite these advances, however, the TabPFN models primarily transfer generic statistical and causal regularities learned from synthetic prediction tasks. They do not provide an explicit, clinically grounded decision pathway, making it difficult to trace a diagnosis to established AD constructs. While restricted post-hoc feature importance analyses such as SHAP are possible,[25,26] these methods generally suffer from a lack of interpretability, limiting their deployment into clinical applications. Concept bottleneck models (CBMs) instead impose intrinsic interpretability by decomposing prediction into an input-to-concept model and a concept-to-label model, thereby constraining the downstream decision to pass through a designated layer of human-understandable concepts.[27–29] In addition, this concept layer can be inspected and corrected through test-time intervention (TTI), in which an expert replaces selected concept predictions with measured values before the final decision.[27,30]

In this study, we present TabPFN-CBM: a novel framework that integrates the state-of-the-art TFM, TabPFN, with an intrinsically interpretable concept bottleneck architecture, utilizing multimodal data and clinically meaningful QT-PAD concepts for AD diagnosis. We investigate two key research questions: **(1) how well can the model predict clinically grounded concepts, and how do those concepts contribute to the diagnosis of AD? (2) how would the predicted diagnosis change if a clinician corrected one or more concept values?** The model first predicts non-overlapping QT-PAD biomarker concepts from ADNI tabular inputs comprising neuroimaging and genetic features. It then performs disease classification to predict AD, mild cognitive impairment (MCI), or cognitively normal (CN) through a clinically grounded concept bottleneck. Our key contributions are two-fold:

(i) To the best of our knowledge, this is the first study to introduce a CBM framework on TFMs, showing that pretrained TFMs like TabPFN can serve as an effective concept predictor in a small-sample, high-dimensional, multimodal biomedical setting. TabPFN-CBM achieves comparable performance to a direct black-box TabPFN, while outperforming both sequential and jointly trained baseline neural network-based CBMs. It thus offers intrinsic interpretability without compromising much predictive power.

(ii) Moreover, we demonstrate actionable concept-level interpretability through simulated clinician-guided test-time intervention. We quantify the effects of individual and sequential interventions, examine their consistency across demographic subgroups, and illustrate how TabPFN-CBM can identify high-value measurements and prioritize concept corrections.

2. Methods

2.1. *Dataset*

The genotyping, demographic and imaging data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu).[31–33] The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Up-to-date information about the ADNI is available at `www.adni-info.org`.

Three input modalities were used in this study. Structural MRI features were extracted using FreeSurfer and included 14 subcortical volume measurements, 66 cortical thickness measurements, and a FreeSurfer version indicator, totaling 81 features. Amyloid PET data consisted of 68 standardized uptake value ratio (SUVR) measurements of regional amyloid accumulation derived from AV45 PET imaging. Genetic data consisted of APOE-ϵ genotype and single-nucleotide polymorphisms (SNPs) filtered at a genome-wide significance threshold of $p < 5 \times 10^{-14}$, resulting in 810 SNPs per subject.

Concept labels were derived from the QT-PAD biomarker set, which originally contains 16 clinical measurements: five cognitive assessments (ADAS-Cog 13, CDR Sum of Boxes, RAVLT learning, MMSE, FAQ), two PET imaging measurements (FDG-PET, AV45-PET), three CSF biomarkers (CSF amyloid-β, CSF total tau, CSF p-tau), and six FreeSurfer morphometric measurements (Whole brain volume, Hippocampus, Entorhinal, Ventricles, Middle temporal, Fusiform). The outcome of interest was the diagnosis of cognitively normal, mild cognitive impairment, or Alzheimer's disease.

As subjects have multiple visits with differing data availability, we constructed a single cross-sectional entry per subject by aligning measurements across imaging modalities to the closest visit within a 12-month window. SNP measurements were treated as time-invariant. Subjects with missing modalities, concepts, or diagnosis were excluded, resulting in 505 total subjects for analysis.

Summary statistics for demographics and QT-PAD biomarkers by diagnostic group are reported in Table 1. For continuous data, mean and standard deviation are shown. For categorical data, counts are shown. Group differences were assessed using Kruskal–Wallis tests for continuous variables and chi-square tests for categorical variables, with Benjamini–Hochberg FDR correction applied across all tests.[34,35]

2.2. *TabPFN-CBM*

We propose TabPFN-CBM, a concept bottleneck model consisting of two stages: **concept prediction** using input modalities and **diagnosis prediction** using predicted concepts (Fig. 1). The bottleneck framework requires all information to pass through a concept layer, allowing for greater interpretability and intervention at test-time. Our model training mirrors the sequential training paradigm introduced in the original concept bottleneck model paper, where stage 1 maps the input to the concept space, and stage 2 is trained on the predicted concepts from stage 1 to learn a mapping from concept to prediction.[27]

Table 1: Subject demographics and QT-PAD biomarker summary by diagnostic group. Values are mean ± SD unless otherwise noted. Significance of group differences: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Characteristic	**CN**	**MCI**	**AD**	**Total**
Demographics				
Participants (n)	192	218	95	505
Age (years)**	73.6 ± 6.0	71.5 ± 7.2	74.0 ± 8.2	72.8 ± 7.1
Sex (M/F)	91/101	122/96	54/41	267/238
Education (years)**	16.6 ± 2.5	16.4 ± 2.6	15.6 ± 2.6	16.3 ± 2.6
Cognitive assessments				
ADAS-Cog 13***	9.0 ± 4.3	15.5 ± 7.2	31.3 ± 8.3	16.0 ± 10.2
CDR Sum of Boxes***	0.0 ± 0.2	1.5 ± 0.9	4.5 ± 1.6	1.5 ± 1.8
RAVLT learning***	6.0 ± 2.3	4.4 ± 2.6	1.7 ± 1.6	4.5 ± 2.8
MMSE***	28.9 ± 1.3	28.1 ± 1.8	23.1 ± 2.2	27.5 ± 2.7
FAQ***	0.3 ± 0.9	3.0 ± 3.8	13.9 ± 7.2	4.0 ± 6.4
PET imaging				
FDG-PET***	1.30 ± 0.13	1.23 ± 0.13	1.03 ± 0.16	1.22 ± 0.16
AV45-PET***	1.13 ± 0.20	1.22 ± 0.24	1.40 ± 0.23	1.22 ± 0.24
CSF biomarkers				
CSF amyloid-β***	1256.0 ± 435.6	1016.9 ± 424.6	696.2 ± 314.5	1047.5 ± 456.3
CSF total tau***	243.6 ± 87.4	279.7 ± 134.3	386.6 ± 158.9	286.1 ± 134.1
CSF p-tau***	22.3 ± 8.9	26.7 ± 14.7	38.3 ± 16.6	27.2 ± 14.4
FreeSurfer morphometry (cm^3)				
Whole brain volume*	1049.1 ± 101.0	1045.8 ± 98.5	1011.8 ± 118.6	1041.0 ± 103.9
Hippocampus***	7.43 ± 0.88	6.98 ± 1.15	5.91 ± 0.98	6.96 ± 1.15
Entorhinal***	3.85 ± 0.59	3.63 ± 0.75	2.93 ± 0.65	3.61 ± 0.74
Ventricles***	32.8 ± 16.2	36.3 ± 20.2	46.4 ± 21.3	36.7 ± 19.5
Middle temporal***	20.5 ± 2.4	20.2 ± 2.6	17.8 ± 3.4	19.9 ± 2.8
Fusiform***	18.6 ± 2.3	18.3 ± 2.7	16.5 ± 2.4	18.1 ± 2.6

Stage 1: Concept prediction. The first stage maps the input features $x \in \mathbb{R}^d$ (comprising d multimodal MRI, PET, SNP features) to the QT-PAD concepts using pretrained TabPFN v2.5.[22] We define the predicted concept vector $\hat{c} \in \mathbb{R}^k$, where k represents the number of the selected set of QT-PAD biomarkers:

$$\hat{c} = f_{\text{TabPFN}}(x) \tag{1}$$

To avoid overlap with the input modalities, we exclude amyloid PET (1) and morphometric measurements (6) from the QT-PAD concept set. Nine QT-PAD biomarkers (ADAS-Cog 13, CDR Sum of Boxes, RAVLT learning, MMSE, FAQ, FDG-PET, CSF amyloid-β, CSF total tau, and CSF phosphorylated tau) were used. Concept prediction was evaluated using the R^2 metric, which measures the proportion of variance in the concept explained by the model's predictions.

Stage 2: Diagnosis prediction. The second stage uses the predicted QT-PAD concept vector $\hat{c}$ from stage 1 as input to predict the final diagnosis $\hat{y} \in \{\text{CN}, \text{MCI}, \text{AD}\}$:

$$\hat{y} = g(\hat{c}) \tag{2}$$

Both logistic regression and TabPFN (*distinct from stage 1*) were assessed for mapping concepts to diagnosis $g(\cdot)$, denoted as **TabPFN-CBM$_{\mathbf{LR}}$** and **TabPFN-CBM$_{\mathbf{TFM}}$**, respectively.

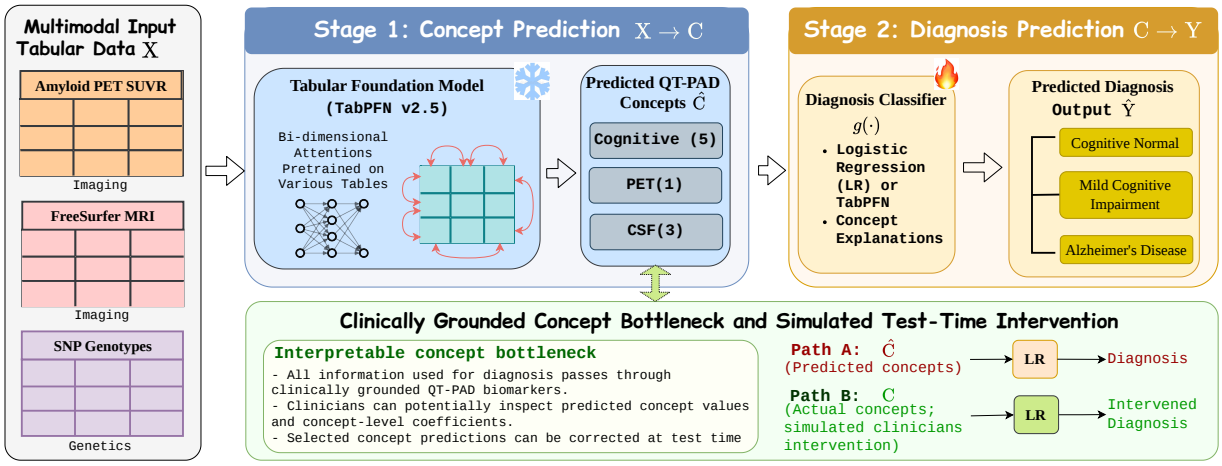


Fig. 1: Overview of the architecture of our sequential TabPFN-CBM framework. Clinically grounded concepts are first predicted from multimodal data and subsequently used for diagnosis classification.

2.3. *Baseline models*

Several baseline models were trained to compare against the performance of our model. Table 2 summarizes the model architectures of TabPFN-CBM and the baseline models.

(1) **Direct models** (X→Y) predict diagnosis from input data without a concept layer, assessing the tradeoff between interpretability and predictive performance that arises from the imposed bottleneck. Two direct models were trained on the same cohort as TabPFN-CBM: logistic regression denoted **Direct$_{\mathbf{LR}}$** and TabPFN denoted **Direct$_{\mathbf{TFM}}$**.

(2) **Concept models** (C→Y) predict diagnosis directly from true concept values. These models allow us to assess the predictive performance in the scenario where concepts are available at inference time. We define a different set of QT-PAD biomarkers used as concepts for concept models. The concepts comprise the following 11 non-cognitive QT-PAD biomarkers: FDG-PET, AV45-PET, CSF amyloid-β, CSF total tau, CSF p-tau, Whole brain volume, Hippocampus, Entorhinal, Ventricles, Middle temporal, and Fusiform. Since AD diagnosis is based partially on cognitive QT-PAD biomarkers, we use all non-cognitive QT-PAD concepts to avoid label leakage.

As subjects are filtered for concept completeness, the cohort for concept models is reduced from 505 subjects to 386 subjects with the complete non-cognitive concept set. Two concept models were trained on the modified cohort: **Concept$_{\mathbf{LR}}$** using logistic regression and **Concept$_{\mathbf{TFM}}$** using TabPFN. These models serve as an upper bound baseline for diagnostic performance based on true biomarker values.

(3) **ANN-based concept bottleneck models** (X→C→Y) predict patient diagnoses from

multimodal tabular input data through a concept layer. These baseline models allow us to isolate the performance increase from replacing the neural network traditionally used in CBM models with a pretrained foundation model. Both sequential training and joint training are assessed, denoted **ANN-CBM (Sequential)** and **ANN-CBM (Joint)** respectively. Sequential training follows the same structure as described for TabPFN-CBM, while joint training trains stage 1 and stage 2 simultaneously, with a hyperparameter λ chosen to weigh concept loss against label loss.[27] ANN-CBM (Sequential) uses an ANN to predict concepts from inputs and logistic regression to predict diagnosis from concepts, while ANN-CBM (Joint) uses an ANN to predict diagnosis from inputs through a concept layer. ANN hyperparameters were selected by validation-set grid search, and the best-performing configuration was used for each variant.

2.4. *Model evaluation*

Data were split into 70% training, 15% validation, and 15% test sets, stratified by diagnosis. To assess robustness and improve reliability of results, we randomly generated five independent data splits using different seeds. Hyperparameter selection was performed using the validation set from the first seed and the selected hyperparameters were evaluated across all five random seeds. Classification performance for CN/MCI/AD diagnosis was assessed using area under the receiver operating characteristic curve (AUC), balanced accuracy, and F1 score, reported as mean ± standard deviation across the five seeds.

Table 2: Model architectures for Alzheimer's disease diagnosis.

Model	**X (input)**	**X→C**	**C (concepts)**	**X/C→Y**
Direct models				
$\text{Direct}_{\text{LR}}$	MRI + PET + SNP	–	–	Logistic regression
$\text{Direct}_{\text{TFM}}$	MRI + PET + SNP	–	–	TabPFN
Concept models				
$\text{Concept}_{\text{LR}}$	–	–	True (11)	Logistic regression
$\text{Concept}_{\text{TFM}}$	–	–	True (11)	TabPFN
Concept bottleneck models				
ANN-CBM (Joint)	MRI + PET + SNP	ANN	Predicted (9)	ANN
ANN-CBM (Sequential)	MRI + PET + SNP	ANN	Predicted (9)	Logistic regression
TabPFN-CBM$_{\text{LR}}$	MRI + PET + SNP	TabPFN	Predicted (9)	Logistic regression
TabPFN-CBM$_{\text{TFM}}$	MRI + PET + SNP	TabPFN	Predicted (9)	TabPFN

Note: X→C denotes the model mapping inputs to concepts. X/C→Y denotes the model mapping inputs to diagnosis for direct models or the model mapping concepts to diagnosis for all other models. The 11 true QT-PAD concepts used by concept models are FDG-PET, AV45-PET, CSF amyloid-β, CSF total tau, CSF p-tau, whole brain volume, hippocampus, entorhinal, ventricles, middle temporal, and fusiform. The 9 predicted concepts used by concept bottleneck models are ADAS-Cog 13, CDR Sum of Boxes, RAVLT learning, MMSE, FAQ, FDG-PET, CSF amyloid-β, CSF total tau, and CSF phosphorylated tau.

2.5. *Concept Intervention*

A hallmark of concept bottleneck models is the ability for experts to directly intervene at the concept level to improve model predictions.[27] TabPFN-CBM allows the predicted concepts to be replaced with their true values to improve AD diagnosis. This enables clinicians to selectively choose the concepts on which to intervene, balancing clinical workload with diagnostic precision.

We measure the effect of test-time intervention in the TabPFN-CBM$_{\text{LR}}$ model. To simulate TTI, we replace predicted concepts with known ground-truth concepts. For an intervention set $\mathcal{I}$ and a given concept j, the intervened value $\tilde{c}_j$ is defined as:

$$\tilde{c}_j = \begin{cases} c_j, & j \in \mathcal{I}, \\ \hat{c}_j, & j \notin \mathcal{I}, \end{cases} \tag{3}$$

where c_j is the true value. We then re-evaluate the stage 2 logistic regression model g with weights unchanged using the updated concept vector $\tilde{c}$:

$$\hat{y}^{(\mathcal{I})} = g(\tilde{c}) \tag{4}$$

We evaluate the effect of concept intervention both individually and sequentially. For individual interventions, we replace each concept with its ground-truth value and measure the change in test AUC, ΔAUC, defined as

$$\Delta\text{AUC} = \text{AUC}_{\text{after intervention}} - \text{AUC}_{\text{no intervention}}. \tag{5}$$

For sequential intervention, concepts were ranked by their individual improvement in validation AUC, from highest to lowest, and progressively replaced with true values in that order. Validation and test split orderings were consistent with each other. All results are reported mean ± standard deviation across five random seeds.

2.6. *Demographic TTI Analysis*

To assess whether the effect of test-time intervention differs across patient subgroups, we stratify the TTI analysis by sex (male/female) and age (split at the median age of 72.8). Within each demographic subgroup, we compute the individual concept ΔAUC, holding the stage 2 classifier fixed. All results are reported as mean ± standard deviation across five random seeds.

3. Results

3.1. *Model Performance*

Table 3 summarizes diagnostic performance on the test set for TabPFN-CBM and baseline models. TabPFN-CBM$_{\text{LR}}$ achieves an AUC of 0.764±0.023, comparable to Direct$_{\text{TFM}}$ with an AUC of 0.766±0.031. Overlapping standard deviations between these two models indicate that enforcing a concept bottleneck layer does not meaningfully sacrifice predictive performance for interpretability. TabPFN-CBM$_{\text{LR}}$ outperforms both ANN-CBM variants (AUC of 0.709 and 0.704 for Joint and Sequential, respectively), demonstrating the impact in performance of a tabular foundation model over traditional neural network backbones in a concept bottleneck model. The weaker performance of TabPFN-CBM$_{\text{TFM}}$ relative to TabPFN-CBM$_{\text{LR}}$ is likely

because TabPFN's in-context mechanism is poorly suited to stage 2, where only nine noisy, model-predicted concepts are available as context.

$\text{Concept}_{\text{TFM}}$, which predicts diagnosis directly from true concept values, achieves the highest performance across all models (AUC 0.802), indicating that the QT-PAD biomarkers selected as concepts are highly informative for AD diagnosis. However, since the complete set of concepts is not routinely collected for all patients,[36] predicting them from available modalities, such as MRI, PET, and genetic data, offers a practical benefit without requiring direct measurement of all QT-PAD biomarkers.

Table 3: Model performance for AD multi-class diagnosis. Highest performance within each model category (direct, concept, concept bottleneck) is in **bold**. Bal. Acc: balanced accuracy.

Model	AUC	Bal. Acc	F1
Direct models			
$\text{Direct}_{\text{LR}}$	0.672±0.035	0.497±0.041	0.485±0.039
$\text{Direct}_{\text{TFM}}$	**0.766±0.031**	**0.598±0.054**	**0.590±0.054**
Concept models			
$\text{Concept}_{\text{LR}}$	0.779±0.031	0.596±0.063	0.596±0.056
$\text{Concept}_{\text{TFM}}$	**0.802±0.026**	**0.607±0.035**	**0.597±0.032**
Concept bottleneck models			
ANN-CBM (Joint)	0.709±0.025	0.549±0.027	0.518±0.022
ANN-CBM (Sequential)	0.704±0.019	0.518±0.011	0.526±0.012
$\text{TabPFN-CBM}_{\text{LR}}$	**0.764±0.023**	**0.591±0.037**	**0.594±0.036**
$\text{TabPFN-CBM}_{\text{TFM}}$	0.749±0.033	0.557±0.042	0.556±0.044

Fig. 2 shows the predictive performance of TabPFN in predicting QT-PAD concepts (Stage 1). CSF amyloid-β was the most accurately predicted concept (R^2=0.63), consistent with the established relationship between amyloid PET and CSF amyloid pathology.[37] Four cognitive assessments (ADAS-Cog 13, MMSE, CDR Sum of Boxes, and FAQ) and FDG-PET were predicted with moderate accuracy (R^2=0.37–0.53), while CSF total tau, CSF p-tau, and RAVLT learning were predicted least accurately (R^2=0.22–0.30). This pattern suggests that biomarkers with stronger physiological links to the input modalities are more easily predicted compared to biomarkers reflecting more distal or behavioral aspects of AD progression.

3.2. *Concept Intervention*

To assess the effect of test-time intervention on the $\text{TabPFN-CBM}_{\text{LR}}$ model, we substitute predicted concepts with their corresponding ground-truth values while keeping the stage 2 classifier fixed. Fig. 3a shows the effect of intervening on each concept individually, ordered from largest to smallest improvement in AUC. Among all concepts, intervening on CDR Sum of Boxes produces the largest improvement in diagnostic performance (ΔAUC=0.178), followed by other cognitive measurements MMSE (ΔAUC=0.042) and FAQ (ΔAUC=0.015). In contrast, FDG-PET and RAVLT learning decrease performance when intervened on individually (ΔAUC=-0.040 and -0.067, respectively).

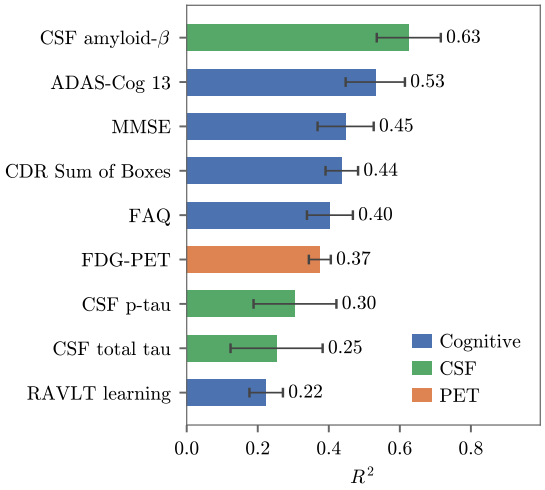


Fig. 2: R^2 for TabPFN's prediction of each QT-PAD concept from SNP, PET, and MRI data, sorted in descending order. Bars are colored by concept category (cognitive, CSF, PET); error bars show standard deviation across five seeds.

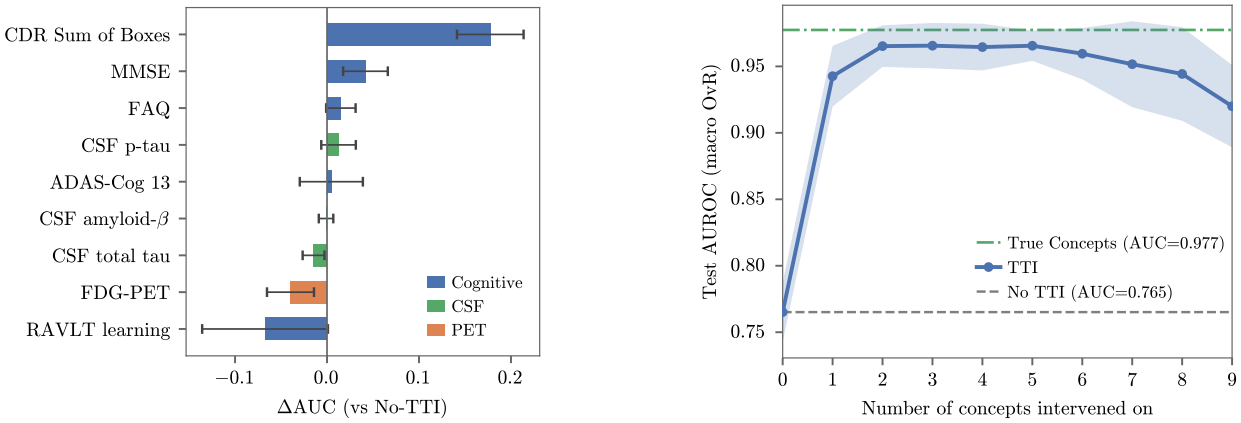
Fig. 3b shows the effect of sequentially intervening on concepts in order of their individual improvement in validation AUC, consistent with the ranking shown in Fig. 3a. A single intervention on CDR Sum of Boxes increased AUC from 0.765 to 0.943, approaching the upper bound achievable with all nine concepts as inputs (True Concepts, AUC=0.977). Performance increased over the first five interventions, peaking at an AUC of 0.966 after five concepts were replaced. With additional interventions, performance gradually declined, reaching an AUC of 0.920 when all concepts were replaced.

The large increase in performance after adding CDR is consistent with its established role in diagnosing Alzheimer's disease progression.[38] Our framework makes this relationship explicit and actionable through intervention, allowing clinicians to potentially identify which individual concepts or sequence of concepts offer the greatest diagnostic value and prioritize obtaining those measurements.

The decline in performance for certain concept interventions likely reflects a distribution shift between predicted and true concept values.[27] The Stage 2 classifier is trained on predicted concepts, whose errors and correlations differ from the ground-truth concept distribution. While replacing highly informative concepts such as CDR Sum of Boxes improves prediction, substituting less important concepts may move samples away from the distribution on which the classifier was trained, resulting in reduced performance despite using more accurate concept values.

3.3. *Demographic TTI Analysis*

Fig. 4 shows individual-concept intervention effects stratified by sex and age. The overall order of concept effects was consistent across all four subgroups and the overall cohort, with CDR Sum of Boxes producing the largest improvement and RAVLT learning and FDG-PET producing small decreases in every stratum. Sex-stratified effects were similar in magnitude



(a) (b)

Fig. 3: Test-time intervention results on the TabPFN-CBM$_{\text{LR}}$ model. (a) Change in AUROC when each QT-PAD concept is individually replaced with its ground-truth value, relative to the no-intervention baseline. Error bars: standard deviation across five seeds. (b) AUROC as concepts are replaced with true values in succession, following the order in (a). True Concepts indicates the upper bound from a logistic regression model trained on the true values of all nine concepts, while No TTI baseline indicates the performance of TabPFN-CBM with no intervention. Shaded region shows standard deviation across five seeds.

between male and female subgroups. Age-stratified effects showed differences between effect size for younger and older subjects. Intervening on CDR Sum of Boxes improved AUC more for patients below the median age of 72.8 years (ΔAUC$=0.206$) than for patients at or above the median (ΔAUC$=0.147$). In this cohort, patients below the median age have a higher proportion of MCI (48.4% vs. 37.8%) and a lower proportion of AD (13.8% vs. 23.9%) relative to older patients. This discrepancy likely reflects the progressive nature of atrophy signals on structural MRI, resulting in a weaker predictive signal for the model in younger, earlier-stage patients.[39]

4. Discussion

In this work, we introduce TabPFN-CBM, a novel framework that pairs the predictive capabilities encoded in a pretrained TFM with an interpretable concept layer clinically grounded for AD diagnosis. It represents the first study to use a TFM as the concept-prediction backbone. In ADNI, its logistic regression variant TabPFN-CBM$_{\text{LR}}$ achieved diagnostic performance comparable to a black-box TabPFN classifier trained directly on input modalities. TabPFN-CBM improved predictive performance over concept bottleneck models with ANN backbones, reflecting the sample-efficiency of TabPFN in the lower-data regime.[15] TabPFN-CBM additionally allows for test-time intervention, transforming the framework into a tool for human collaboration where clinicians can correct predicted concepts to improve predictions. Our simulated TTI results show that while more accurate concept values do not uniformly improve a classifier trained on predicted concepts, intervention on a single concept (CDR Sum of Boxes) achieves near-perfect AUC for multi-class AD diagnosis and such gain applies across patient

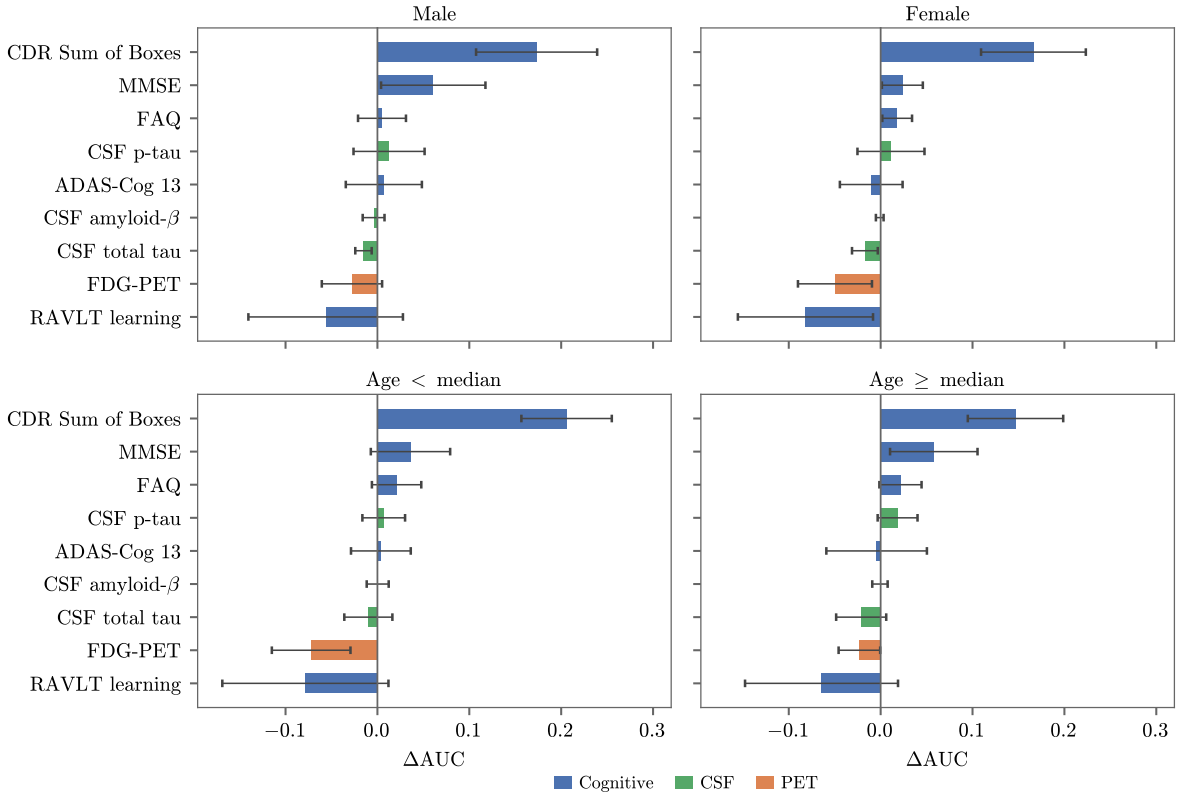


Fig. 4: Demographic subgroup test-time intervention analysis. Subjects are stratified by sex and by age (split at the median age). Bars show change in AUROC when each QT-PAD concept is individually replaced with its ground-truth value, relative to the no-intervention baseline. Error bars indicate standard deviation across five seeds.

subgroups of sex and age.

The gap in performance between TabPFN-CBM and ANN-CBM is consistent with prior literature showing that neural networks require larger sample sizes than tabular foundation models to learn reliable representations.[15] For our sample size of n=505, TabPFN's in-context, prior-fitted approach outperforms models that learn representations from scratch. Many clinical tasks are limited by smaller sample sizes, including rare disease cohorts, studies requiring invasive or costly medical tests, or longitudinal cohorts. Our approach may extend to a range of settings where interpretability is crucial but sample size requirements limit the use of models without pretraining.

Several limitations remain: Currently, the framework is evaluated solely on the single, cross-sectional, and complete-case ADNI cohort. This single timepoint design may not reflect realistic clinical workflows, where cognitive assessments or preliminary diagnoses are often available before imaging or genetic testing. Future work should prioritize external or prospective clinical validation and incorporate longitudinal data to better forecast progressive disease trajectories. In the longitudinal setting, the model should reflect the temporal relationship of the diagnostic process to expand the clinical utility of our framework. Model performance

can be further enhanced by integrating non-tabular modalities, such as clinical notes via language models and raw images via imaging foundation models,[40–43] while automated concept discovery[44–46] could identify novel biomarkers in an unsupervised fashion and extend the model's utility to settings lacking predefined concept labels. Additionally, clinical deployment would require robust handling of potential missing data and providing uncertainty estimates at both the concept and diagnosis levels.

Acknowledgements: This work is supported in part by NIH R01 AG071470, U01 AG066833, U01 AG068057, and U19 AG074879; NSF Grant 2500343; and seed funding from the PSOM AI2D Center at Penn. Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (adni.loni.usc.edu), which was funded by NIH U01 AG024904.

References

1. M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro and C. P. Ferri, The global prevalence of dementia: A systematic review and metaanalysis, *Alzheimer's & Dementia* **9**, 63 (2013).
2. E. Nichols *et al.*, Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019, *The Lancet Public Health* **7**, e105 (2022).
3. C. R. Jack, D. A. Bennett *et al.*, Nia-aa research framework: Toward a biological definition of alzheimer's disease, *Alzheimer's and Dementia* **14**, 535 (2018).
4. C. R. Jack, D. A. Bennett *et al.*, A/t/n: An unbiased descriptive classification scheme for alzheimer disease biomarkers, *Neurology* **87**, 539 (2016).
5. J. Sha, J. Bao, K. Liu, S. Yang, Z. Wen, J. Wen, Y. Cui, B. Tong, J. H. Moore, A. J. Saykin *et al.*, Preference matrix guided sparse canonical correlation analysis for mining brain imaging genetic associations in alzheimer's disease, *Methods* **218**, 27 (2023).
6. S. Yang, A. Wang, J. Bao, S. Mu, Y. Feng, Z. Wen, J. Y. Baik, J. Wen, B. Hou, R. Wang *et al.*, 'diet'deep canonical correlation analysis for high-dimensional genetics study of brain imaging phenotypes in alzheimer's disease, *Alzheimer's & Dementia* **20**, p. e092330 (2024).
7. Z. Yang, J. Wen, A. Abdulkadir, Y. Cui, G. Erus, E. Mamourian, R. Melhem, D. Srinivasan, S. T. Govindarajan, J. Chen *et al.*, Gene-sgan: discovering disease subtypes with imaging and genetic signatures via multi-view weakly-supervised deep clustering, *Nature communications* **15**, p. 354 (2024).
8. J. Bao, J. Wen, C. Chang, S. Mu, J. Chen, M. Shivakumar, Y. Cui, G. Erus, Z. Yang, S. Yang *et al.*, A genetically informed brain atlas for enhancing brain imaging genomics, *Nature communications* **16**, p. 3524 (2025).
9. S. Kearney, S. Yang, Z. Wen, B. Hou, D. Duong-Tran, T. Chen, J. Moore, M. Ritchie and L. Shen, Enabling few-shot alzheimer's disease diagnosis on biomarker data with tabular llms, in *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2025.
10. L. Shen and P. M. Thompson, Brain imaging genomics: Integrated analysis and machine learning, *Proceedings of the IEEE* **108**, 125 (2020).
11. J. Bao, B. N. Lee, J. Wen, M. Kim, S. Mu, S. Yang, C. Davatzikos, Q. Long, M. D. Ritchie and L. Shen, Employing informatics strategies in alzheimer disease research: A review from genetics, multiomics, and biomarkers to clinical outcomes, *Annual Review of Biomedical Data Science* (2024).
12. E. Lin, C.-H. Lin and H.-Y. Lane, Deep learning with neuroimaging and genomics in alzheimer's disease, *International journal of molecular sciences* **22**, p. 7911 (2021).

13. X. Cen, W. Dong, W. Lv, Y. Zhao, F. Dubee, A.-F. A. Mentis, D. Jovic, H. Yang and Y. Li, Towards interpretable imaging genomics analysis: methodological developments and applications, *Information Fusion* **102**, p. 102032 (2024).
14. L. Grinsztajn, E. Oyallon and G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, *Advances in neural information processing systems* **35**, 507 (2022).
15. D. McElfresh, S. Khandagale, J. Valverde, V. Prasad C, G. Ramakrishnan, M. Goldblum and C. White, When do neural nets outperform boosted trees on tabular data?, *Advances in Neural Information Processing Systems* **36**, 76336 (2024).
16. N. Hollmann, S. Müller, K. Eggensperger and F. Hutter, TabPFN: A transformer that solves small tabular classification problems in a second, in *The Eleventh International Conference on Learning Representations*, 2023.
17. B. Feuer, R. T. Schirrmeister, V. Cherepanova, C. Hegde, F. Hutter, M. Goldblum, N. Cohen and C. White, TuneTables: Context optimization for scalable prior-data fitted networks, in *Advances in Neural Information Processing Systems*, 2024.
18. V. Thomas, J. Ma, R. Hosseinzadeh, K. Golestan, G. Yu, M. Volkovs and A. L. Caterini, Retrieval & fine-tuning for in-context tabular models, in *Advances in Neural Information Processing Systems*, 2024.
19. N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister and F. Hutter, Accurate predictions on small data with a tabular foundation model, *Nature* **637**, 319 (2025).
20. H.-J. Ye, S.-Y. Liu and W.-L. Chao, A closer look at TabPFN v2: Understanding its strengths and extending its capabilities, in *Advances in Neural Information Processing Systems*, 2025.
21. S. Liu and H.-J. Ye, TabPFN unleashed: A scalable and effective solution to tabular classification problems, in *Proceedings of the 42nd International Conference on Machine Learning*, 2025.
22. L. Grinsztajn, K. Flöge, O. Key, F. Birkel, P. Jund, B. Roof, B. Jäger, D. Safaric, S. Alessi, A. Hayler, M. Manium, R. Yu, F. Jablonski, S. B. Hoo, A. Garg, J. Robertson, M. Bühler, V. Moroshan, L. Purucker, C. Cornu, L. C. Wehrhahn, A. Bonetto, B. Schölkopf, S. Gambhir, N. Hollmann and F. Hutter, TabPFN-2.5: Advancing the state of the art in tabular foundation models (2025).
23. L. Grinsztajn, K. Flöge, O. Key, F. Birkel, P. Jund, B. Roof, M. Manium, S. B. Hoo, M. Bühler, A. Garg, D. Safaric, J. Robertson, B. Jäger, S. Alessi, A. Hayler, V. Moroshan, L. Purucker, P. Singer, A. Arazi, J. Siems, J. H. Metzen, G. Grab, N. Erickson, S. Guo, E. Kalfon, S. Bing, D. Salinas, C. Cornu, L. C. Wehrhahn, D. Kriuchkova, K. Kaya, L. Sidhoum, M. Salmon, J. Chen, M. Hulsebos, Y. LeCun, S. Müller, B. Schölkopf, S. Gambhir, N. Hollmann and F. Hutter, TabPFN-3: Technical report (2026).
24. C. Kolberg, J. Kreuer, J. Huurdeman, S. Ouaari, K. Eggensperger and N. Pfeifer, TabPFN-wide: Continued pre-training for extreme feature counts (2025).
25. M. Muschalik, H. Baniecki, F. Fumagalli, P. Kolpaczki, B. Hammer and E. Hüllermeier, shapiq: Shapley interactions for machine learning, in *Advances in Neural Information Processing Systems*, eds. A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak and C. Zhang (Curran Associates, Inc., 2024).
26. D. Rundel, J. Kobialka, C. von Crailsheim, M. Feurer, T. Nagler and D. Rügamer, Interpretable machine learning for tabpfn, in *World Conference on Explainable Artificial Intelligence*, 2024.
27. P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim and P. Liang, Concept bottleneck models, in *Proceedings of the 37th International Conference on Machine Learning*, 2020.
28. M. E. Zarlenga, Z. Shams, M. E. Nelson, B. Kim and M. Jamnik, Tabcbm: Concept-based interpretable neural networks for tabular data, *Transactions on Machine Learning Research*

(2023).
29. T. Han, R. Wu, Y. Tian, F. Khader, L. C. Adams, K. K. Bressem, C. Davatzikos, J. N. Kather, L. Shen, D. A. Mankoff *et al.*, Clear: an auditable foundation model for radiology grounded in clinical concepts, *Nature Biomedical Engineering* (2026).
30. S. Shin, Y. Jo, S. Ahn and N. Lee, A closer look at the intervention procedure of concept bottleneck models, in *Proceedings of the 40th International Conference on Machine Learning*, 2023.
31. A. J. Saykin, L. Shen, X. Yao, S. Kim, K. Nho, S. L. Risacher, V. K. Ramanan, T. M. Foroud, K. M. Faber, N. Sarwar, L. M. Munsie, X. Hu, H. D. Soares, S. G. Potkin, P. M. Thompson, J. S. Kauwe, R. Kaddurah-Daouk, R. C. Green, A. W. Toga, M. W. Weiner and Alzheimer's Disease Neuroimaging Initiative, Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans, *Alzheimers Dement* **11**, 792 (2015).
32. L. Shen, P. M. Thompson, S. G. Potkin, L. Bertram, L. A. Farrer, T. M. Foroud, R. C. Green, X. Hu, M. J. Huentelman, S. Kim, J. S. Kauwe, Q. Li, E. Liu, F. Macciardi, J. H. Moore, L. Munsie, K. Nho, V. K. Ramanan, S. L. Risacher, D. J. Stone, S. Swaminathan, A. W. Toga, M. W. Weiner, A. J. Saykin and Alzheimer's Disease Neuroimaging Initiative, Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers, *Brain Imaging Behav* **8**, 183 (2014).
33. M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, J. C. Morris *et al.*, Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials, *Alzheimer's & Dementia* **13**, e1 (2017).
34. W. H. Kruskal and W. A. Wallis, Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* **47**, 583 (1952).
35. Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289 (01 1995).
36. R. Y. Lo, W. J. Jagust, P. Aisen, C. R. Jack Jr, A. W. Toga, L. Beckett, A. Gamst, H. Soares, R. C. Green, T. Montine *et al.*, Predicting missing biomarker data in a longitudinal study of alzheimer disease, *Neurology* **78**, 1376 (2012).
37. D. Alcolea, J. Pegueroles, L. Munoz, V. Camacho, D. López-Mora, A. Fernández-León, N. Le Bastard, E. Huyck, A. Nadal, V. Olmedo *et al.*, Agreement of amyloid pet and csf biomarkers for alzheimer's disease on lumipulse, *Annals of clinical and translational neurology* **6**, 1815 (2019).
38. S. E. O'Bryant, S. C. Waring, C. M. Cullum, J. Hall, L. Lacritz, P. J. Massman, P. J. Lupo, J. S. Reisch, R. Doody, T. A. R. Consortium *et al.*, Staging dementia using clinical dementia rating scale sum of boxes scores: a texas alzheimer's research consortium study, *Archives of neurology* **65**, 1091 (2008).
39. S. M. Tripathi, P. Chutia and A. D. Murray, Neuroimaging biomarkers in alzheimer's disease, *Journal of Dementia and Alzheimer's Disease* **2** (2025).
40. X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian and Y. Wu, Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records (2022).
41. K. Huang, J. Altosaar and R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission (2020).
42. G. Yang, K. Du, Z. Yang, Y. Du, E. Y. W. Cheung, Y. Zheng, M. Yang, Z. Kourtzi, C.-B. Schönlieb, S. Wang and A. D. N. Initiative, Adfound: A foundation model for diagnosis and prognosis of alzheimer's disease, *IEEE Journal of Biomedical and Health Informatics* **29**, 8395 (2025).

43. D. Tak, B. A. Garomsa, A. Zapaishchykova, T. L. Chaunzwa, J. C. Climent Pardo, Z. Ye, J. Zielke, Y. Ravipati, S. Pai, S. Vajapeyam, M. Mahootiha, M. Parker, L. R. G. Pike, C. Smith, A. M. Familiar, K. X. Liu, S. Prabhu, O. Arnaout, P. Bandopadhayay, A. Nabavizadeh, S. Mueller, H. J. W. L. Aerts, R. Y. Huang, T. Y. Poussaint and B. H. Kann, A generalizable foundation model for analysis of human brain mri, *Nature Neuroscience* **29**, 945 (April 2026).
44. T. Oikarinen, S. Das, L. M. Nguyen and T.-W. Weng, Label-free concept bottleneck models, in *International Conference on Learning Representations*, 2023.
45. M. Yuksekgonul, M. Wang and J. Zou, Post-hoc concept bottleneck models, in *International Conference on Learning Representations*, 2023.
46. S. Rao, S. Mahajan, M. Böhle and B. Schiele, Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery, in *European Conference on Computer Vision*, 2024.