Cross-Modal Knowledge Distillation for PET-Free Amyloid-Beta Detection from MRI

Francesco Chiumento[1] Julia Dietlmeier[1,2] Ronan P. Killeen[3,4]
Kathleen M. Curran[2,4] Noel E. O'Connor[1,2] Mingming Liu[1,2]
[1]Dublin City University, Ireland [2]Insight Research Ireland Centre for Data Analytics, Ireland
[3]St. Vincent's University Hospital, Ireland [4]University College Dublin, Ireland
francesco.chiumento2@mail.dcu.ie

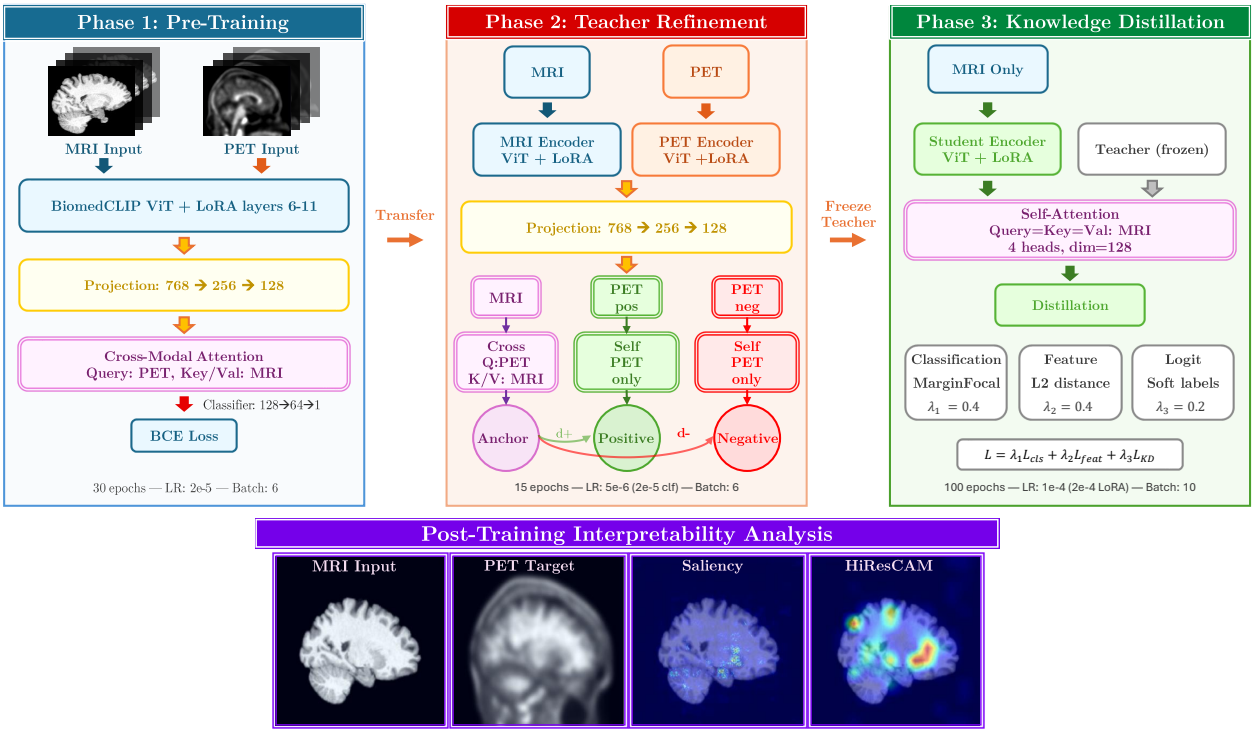


Figure 1. Overview of our PET-guided knowledge distillation framework. A teacher model learns cross-modal alignment between PET and MRI, guided by Centiloid-aware triplet mining. The teacher's knowledge is then distilled to an MRI-only student for PET-free inference.

Abstract

Detecting amyloid-β (Aβ) positivity is crucial for early diagnosis of Alzheimer's disease but typically requires PET imaging, which is costly, invasive, and not widely accessible, limiting its use for population-level screening. We address this gap by proposing a PET-guided knowledge distillation framework that enables Aβ prediction from MRI alone, without requiring non-imaging clinical covariates or PET at inference. Our approach employs a BiomedCLIP-based teacher model that learns PET–MRI alignment via cross-modal attention and triplet contrastive learning with PET-informed (Centiloid-aware) online negative sampling. An MRI-only student then mimics the teacher via feature-level and logit-level distillation. Evaluated across four MRI contrasts (T1w, T2w, FLAIR, T2) and two independent datasets, our approach demonstrates effective knowledge transfer (best AUC: 0.74 on OASIS-3, 0.68 on ADNI) while maintaining interpretability and eliminating the need for clinical variables. Saliency analysis confirms that predictions focus on anatomically relevant cortical regions, supporting the clinical viability of PET-free Aβ screening. Code is available at github.com/FrancescoChiumento/pet-guided-mri-amyloid-detection.*

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that represents the predominant form of dementia. 55.2 million individuals are currently affected worldwide, with projections reaching 78 million by 2030 and costs projected to reach up to $2.8 trillion [36, 57]. The pathological hallmarks of AD are neurofibrillary tangles (NFTs), formed by hyperphosphorylated tau protein within neurons, and extracellular plaques of accumulated Aβ peptide accompanied by neuroinflammation, synaptic dysfunction, mitochondrial and vascular abnormalities [45, 59]. Currently, there is no effective cure for AD, and once it manifests, it follows an irreversible progression [17, 59].

An early event in AD pathophysiology is the deposition of Aβ plaques in the brain, which can occur $\sim$20 years before the onset of symptoms, offering a potential window for secondary prevention [8]. Typically, Aβ is assessed through invasive procedures such as positron emission tomography (PET) or cerebrospinal fluid (CSF) assays. PET uses radiotracers targeting amyloid or tau pathology, creating a spatial map of Aβ distribution in the brain and revealing the extent of AD pathology [4]. Although amyloid levels can be measured in individuals using PET imaging with amyloid-sensitive ligands such as Pittsburgh Compound B (PiB; ^{11}C-PiB) or florbetapir (^{18}F-AV-45) [7, 58], amyloid-PET is expensive, not widely available, and invasive, exposing the patient to ionizing radiation. For this reason, amyloid-PET is not considered cost-effective for diagnostic use at the MCI stage [28], yet demand is projected to increase up to $\sim$20-fold following approval of anti-Aβ therapies (lecanemab, donanemab) that require confirmation of amyloid pathology prior to treatment initiation [52–54].

MRI, conversely, is non-invasive, widely accessible, and provides detailed structural brain information without radioactive tracers [56]. While MRI cannot directly visualize amyloid, Aβ accumulation leads to widespread brain cell loss, which manifests as atrophy in T1-weighted (T1w) sequences [9]. Several studies classify Aβ positivity (Aβ+ vs. Aβ-) from PET images using deep learning models. A smaller subset uses MRI, but typically includes additional non-imaging covariates such as APOEε4 genotype, cognitive scores, hippocampal volumetry, or clinical diagnosis [12, 24, 30, 56]; however, these covariates exhibit high missingness in practice: routine APOE testing is not recommended by ACMG guidelines [16] and is missing in 26% of NACC participants [39]; cognitive measures are absent in 89% of EHR dementia records [35]; and only 23% of European centres perform volumetry analysis [55]. Few studies, on the other hand, have attempted to predict amyloid status using only MRI data without other clinical variables [23].

In this work, we address binary Aβ+ prediction using only MRI as the single modality. Specifically, our research question is how to effectively transfer the knowledge from a multimodal contrastive PET and MRI *teacher* to an MRI-only *student*. With this in mind, the key novelty of our work is that we propose a PET-guided *knowledge distillation* framework based on BiomedCLIP, a ViT pre-trained on 15M biomedical image-text pairs [60] which, to the best of our knowledge, has not been applied in this specific context. The cross-modal *teacher* integrates PET (query) and MRI (key) via multi-head attention (MHA) and is supervised by a combination of classification and *triplet loss*; the MRI-only *student* learns from the *teacher* through feature-level and logit-level distillation with temperature annealing. At inference time, only MRI is required. Fig. 1 illustrates our three-phase distillation approach. The main contributions of our work are outlined below.

- **PET-to-MRI knowledge distillation framework:** We propose a novel cross-modal distillation approach for amyloid prediction, where an MRI-only student learns from a PET+MRI teacher via ℓ_2-normalized feature matching and temperature-scaled logit distillation, enabling PET-free inference without non-imaging covariates. Unlike prior work using direct supervision from binary PET labels or CSF values, our teacher learns from PET's spatial standardized uptake value ratio (SUVR) distribution [25] through cross-modal attention, transferring spatially-informed knowledge to the student model.
- **Contrastive teacher training:** We propose a new adaptation of the triplet loss with Centiloid (CL, a standardized PET amyloid quantification scale [20, 25])-aware negative mining, where online hard-negative selection ensures biochemically distinct triplets, while cross-modal attention and self-attention synthesize patient-level embeddings for triplet learning.
- **Multi-contrast evaluation and cross-dataset transfer:** We evaluate T1w, T2w, FLAIR, and T2* sequences and their combinations on OASIS-3 and ADNI, achieving AUCs up to 0.74 and improving T1w MRI-only results by +19.7% (0.73 vs. 0.61) over Kim et al. [23], and validate cross-dataset knowledge distillation (CDKD) [1, 27].

2. Related Work

2.1. MRI-based Amyloid Prediction

MRI-only methods for predicting Aβ+ remain limited and achieve only moderate discriminative performance. Most studies focus on binary classification of amyloid-β status (Aβ+ vs. Aβ-) in AD research, often using PET or CSF labels within deep learning models [12]. A smaller number of works have investigated MRI-only prediction. Kim et al. [23] trained a 3D EfficientNet on 4,056 exams (ADNI, OASIS-3, and A4), reporting AUCs of 0.61 for T1w and 0.67 for T1w+FLAIR, without clinical variables. In contrast, when structured variables are added, performance increases: Lew et al. [30] combined MRI with demographic,

APOE, and cognitive information to predict PET-derived Amyloid/Tau/Neurodegeneration (ATN) status, obtaining an amyloid AUC of 0.79 using SUVR thresholded via Gaussian mixture modeling. Most prior work either relies on direct supervision or clinical variables, with few truly MRI-only pipelines and no exploitation of PET–MRI complementarity during training [23]. We address this by supervising MRI features with a PET-guided teacher through cross-modal distillation.

Although structured covariates can improve prediction, their use reduces deployability in practice, since key variables (e.g., APOE genotype and detailed cognitive scores) are not consistently available across imaging workflows and cohorts [35, 39, 41, 42, 55]. Also, beyond T1w, integration of additional contrasts such as FLAIR or DTI has been explored in a few studies, but evidence for T2w or T2*-GRE/SWI sequences for Aβ classification remains limited, as these modalities are more commonly used for amyloid-related imaging abnormalities (ARIA) or vascular assessment [44]. Conversely, MRI scans are routinely acquired and recommended as *first-line* neuroimaging, making them a pragmatic and scalable input for automated screening models [40]. This supports multimodal supervision with single-modality inference.

2.2. Multimodal Fusion and Its Limitations

Multimodal approaches (structural, functional, and diffusion MRI) have shown promising results [12] but require all modalities to be available at both training and inference time, limiting real-world applicability. These limitations have motivated alternative strategies that exploit multimodal information during training while enabling single-modality inference. Among these, *knowledge distillation* has emerged as a framework for cross-modal transfer [31].

2.3. Knowledge Distillation in Medical Imaging

Knowledge distillation (KD) is a learning paradigm in which a compact "student" network learns from a high-capacity "teacher" network, improving efficiency or enabling cross-modal transfer. In medical imaging, KD has been applied to tasks such as segmentation, classification, and reconstruction. To the best of our knowledge, PET-to-MRI distillation for amyloid prediction has not been extensively explored [31]. Building on this overview, we identify key limitations in the current literature:

- No use of PET–MRI cross-modal supervision during training;
- Limited evaluation of multi-contrast MRI beyond T1w+FLAIR;
- Limited assessment of model generalization across independent cohorts.

We adopt CL to define amyloid status because CL provides a tracer- and cohort-independent scale [20, 26], enabling cross-dataset transfer between OASIS-3 and ADNI. In contrast, some prior work uses SUVR-based thresholding (including data-driven Gaussian mixture cutoffs) [30] or CSF-based definitions [12], which are internally consistent but less comparable across sites, tracers and cohorts [20]. Since label definitions and targets differ, absolute AUCs are indicative rather than strictly interchangeable.

3. Datasets and Cohorts

Public cohorts We evaluate two large public cohorts for aging and AD (Table 1), with subsets detailed in Table 2.

- **OASIS-3** [27] is a longitudinal dataset spanning ∼30 years, with 1,379 subjects, 2,842 MRI sessions, and 2,157 PET sessions. PET data include SUVR and CL quantification via the PET Unified Pipeline (PUP) [49].
- **ADNI** [1] is a widely used multicenter longitudinal study with over 5,100 participants (3,262 with MRI, 1,089 with PET) sharing imaging and structured data under the ADNI Data Use Agreement and publication policy.

Table 1. **Public cohorts overview.** Amyloid PET tracers used.

Cohort	Subjects	PET Tracers	Notes
OASIS-3 [27]	1,379	PiB, AV-45	Multimodal; PUP
ADNI [1]	5,111	PiB, AV-45	Multicenter

Data usage statement Data are obtained from the longitudinal databases OASIS-3 [27] and the Alzheimer's Disease Neuroimaging Initiative (ADNI) [1]. We construct modality-consistent MRI cohorts using the same procedure across both datasets:

- ($T1w \cap T2w$): all MRI sessions where both T1w and T2w are available for the same subject/timepoint;
- ($T2* \cap FLAIR$): all MRI sessions where both T2* and FLAIR are available for the same subject/timepoint.

Each cohort is partitioned with identical modality constraints; sessions missing contrasts are excluded (no imputation). We retain all timepoints matching the modality combination, considering only subjects with MRI and PET acquired within 365 days. Data splits are performed at the subject level to avoid subject overlap and prevent data leakage across folds. When a subject has multiple PET tracers (PiB, AV-45), all scans are assigned to the same fold.

Table 2. **Cohorts used.** Splits are subject-wise.

Cohort (MRI)	Sessions			Subjects			%Pos (Train)
	Train	Val	Test	Train	Val	Test	
OASIS-3 (T1w ∩ T2w)	518	187	166	346	118	127	21.8
OASIS-3 (FLAIR ∩ T2*)	288	96	89	287	93	88	34.4
ADNI (T1w ∩ T2w)	303	102	90	212	72	68	44.9
ADNI (FLAIR ∩ T2*)	608	193	190	405	132	132	50.2

4. Methodology

4.1. Overview and Framework Architecture

Our methodology addresses Aβ detection without PET images through a three-phase distillation framework (Figs. 2–4). We train a teacher that captures shared PET–MRI patterns, then distill this knowledge into a student operating on MRI-only (single- or multi-contrast: T1-weighted (T1w), T2-weighted (T2w), fluid-attenuated inversion recovery (FLAIR), and T2*-weighted (T2*)). The framework consists of:

1. **Reference labels**: binary Aβ status from PET-derived CL scores (positive if CL > 20.6).
2. **Preprocessing and registration**: N4, HD-BET, ANTs registration to MNI, and slice-level PET–MRI pairing.
3. **Architecture**: BiomedCLIP ViT with LoRA; cross-modal attention (teacher), self-attention (student).
4. **Training**: Phase 1 pre-training with classification (Sec. 4.5), Phase 2 contrastive learning with *CL-aware* online negative sampling for triplets (selecting negatives with $|\text{CL}_{\text{anchor}} - \text{CL}_{\text{neg}}| \geq 5.0$) (Sec. 4.6), and Phase 3 feature and logit distillation (Sec. 4.7, Figs. 2–4).

At test time, only MRI is required; PET images and CL values are not needed.

4.2. Ground Truth: PET-Based Amyloid Quantification

We use PET-derived Centiloid (CL) values as the ground truth for binary amyloid status, which provide a tracer-harmonized scale across AV-45 and PiB acquisitions [20, 25]. OASIS-3 uses the PET Unified Pipeline (PUP) [49] for cortical SUVR normalized to cerebellum. ADNI uses Berkeley's `CENTILOIDS` field (`UCBERKELEY_AMY_6MM`) for AV-45 quantification [43]. We apply a fixed threshold **CL** > 20.6 uniformly across cohorts consistent with established practice [2, 25, 29, 37], without demographic or clinical variables.

The use of CL values serves two distinct purposes: **(i) training**—binary label and *CL-aware negative sampling* in triplet mining (teacher accesses PET spatial distribution); **(ii) inference**—MRI-only, model outputs $p(\text{A}\beta+)$ thresholded at $\theta = 0.5$ (fixed) or θ^* (validation-optimized). Table 2 shows the resulting Aβ prevalence across cohorts; OASIS-3 exhibits class imbalance.

4.3. Data Organization and Preprocessing

MRI Preprocessing MRI and PET scans are temporally paired within $|\Delta t_{\text{MRI–PET}}| \leq 365$ days, consistent with low annualized CL change [6]. For each MRI contrast (T1w, T2w, FLAIR, T2*), we apply: (i) N4 bias field correction with Otsu thresholding using 200 histogram bins to generate a binary foreground mask, excluding background voxels from correction [50]; (ii) HD-BET brain extraction [21], a deep learning-based method trained on multi-site data. HD-BET skull stripping (mask + brain volume) employs test-time augmentation for improved robustness; (iii) robust intensity normalization: clip 0.5–99.5%, min–max to $[0, 1]$, then gamma correction $I_{\text{norm}} = I^{0.9}$ for gray–white matter contrast [10].

PET–MRI Registration PET volumes are smoothed with a Gaussian kernel (FWHM 8 mm) following the PET Unified Pipeline protocol [49] and rigidly aligned to their corresponding skull-stripped MRI using ANTs (`type_of_transform='Rigid'`) [3], then both are registered to the MNI ICBM152 symmetric template using rigid transformation only [14, 15]. Finally, the MRI-to-MNI transform is applied to the PET volume previously aligned to MRI via `ants.apply_transforms` with linear interpolation, yielding standardized files. We use *only rigid transformations* (6 degrees of freedom: 3 translations + 3 rotations) to avoid introducing artificial deformations that could misrepresent amyloid distribution patterns.

Slice Extraction and Pairing After MNI registration ($1\times1\times1$ mm^3 isotropic), from each volume we extract $S_{\text{full}} = 40$ sagittal slices (resized to 224×224 for ViT input) uniformly from the central 40–60% along the left–right axis. During training, we subsample $S_{\text{train}} = 25$ uniformly spaced slices balancing coverage, batch diversity, and efficiency, leveraging BiomedCLIP 2D pretraining with multi-head attention pooling. PET–MRI pairs use matching slice indices for anatomical correspondence. Slices with mean intensity < 0.01 (background) are discarded. We create one fixed subject-level split using `StratifiedGroupKFold` (K=5, seed=42): 3 folds train, 1 val, 1 test; all sessions/tracers per subject remain within one fold to prevent leakage. Stratification is based on the binary amyloid label to maintain class balance. We also check tracer proportions across splits to avoid distribution drift.

Training Strategies During Phases 1–2, we apply synchronized spatial and intensity augmentations (affine transforms, color jitter, blur/noise, random erasing) with shared random seeds for PET–MRI pairs to preserve anatomical correspondence (see supplementary). To address class imbalance (Table 2), we use `WeightedRandomSampler` with inverse frequency weights ($w_i = 1/n_{y_i}$), ensuring balanced mini-batches.

4.4. Network Architecture

We adapt a BiomedCLIP Vision Transformer (ViT) encoder [60] using Low-Rank Adaptation (LoRA) [19] applied to the last six transformer blocks (layers 6–11, 0-indexed; rank $r = 32$) on attention projections (*q/k/v/out*), while keeping earlier blocks frozen. LoRA introduces trainable low-rank

matrices that reparameterize frozen pretrained weights, enabling efficient fine-tuning (the formula and initialization can be found in the supplementary).

Each modality (MRI or PET) is processed as a stack of S slices resized to 224×224. Per-slice CLS tokens (768D) are projected to 128D through a two-layer nonlinear head with LayerNorm, GELU, and dropout. Slice embeddings are then processed by an MHA layer with 4 heads and pooled via learned attention weights $\alpha_s = \mathrm{softmax}(w_s/\tau)$ with temperature τ=2.0 to obtain patient-level representations $\mathbf{e} = \sum_s \alpha_s \mathbf{A}_s$ (see supplementary).

Attention Mechanisms **Teacher:** extracts MRI patterns predictive of PET-derived amyloid pathology through cross-modal attention:

$$\mathbf{A}^{(T)} = \mathrm{CrossAttn}(\mathbf{Q} = \mathbf{E}_{\mathrm{PET}}, \mathbf{K} = \mathbf{V} = \mathbf{E}_{\mathrm{MRI}}), \quad (1)$$

where $\mathbf{E}_{\mathrm{PET}}, \mathbf{E}_{\mathrm{MRI}} \in \mathbb{R}^{B\times S\times 128}$ are batched slice embeddings, then attention-pooled to obtain $\mathbf{e}_T = \sum_s \alpha_s^{(T)} \mathbf{A}_s^{(T)}$. **Student:** aggregates MRI-only features via self-attention:

$$\mathbf{A}^{(S)} = \mathrm{SelfAttn}(\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{E}_{\mathrm{MRI}}), \quad (2)$$

followed by the same pooling to produce $\mathbf{e}_S = \sum_s \alpha_s^{(S)} \mathbf{A}_s^{(S)}$. We match ℓ_2-normalized embeddings $\hat{\mathbf{e}}_S$ to teacher embeddings $\hat{\mathbf{e}}_T$ via feature distillation (Sec. 4.8).

Classifier Both models use a two-layer MLP with ReLU and dropout ($128 \to 64 \to 1$) to produce the logit z. Dropout rates are higher during teacher training (Phases 1–2) and reduced during distillation.

4.5. Phase 1: Teacher Pre-Training

We pre-train the teacher on PET→MRI embeddings (Fig. 2): $\mathbf{e}_T = \mathrm{AttnPool}(\mathbf{E}_{\mathrm{PET}}, \mathbf{E}_{\mathrm{MRI}})$, $z_T = \mathrm{MLP}(\mathbf{e}_T)$, using AdamW [32] (lr 2×10^{-5}, weight decay (wd) 10^{-3}, batch size 6) with binary cross-entropy with logits (BCE) for 30 epochs with gradient clipping and warm-up (see supplementary).

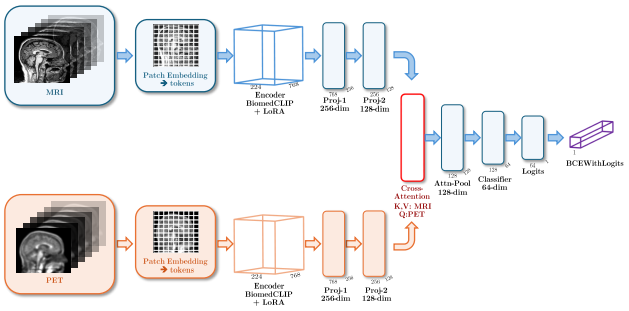


Figure 2. **Phase 1 — Teacher pre-training** with binary classification loss (BCE).

4.6. Phase 2: Contrastive Teacher Refinement

CL-Aware Online Negative Sampling We perform *online* hard-negative mining using PET-derived CL values during training. Triplets (anchor, positive, negative) are formed at patient level: anchor and positive from the same subject/session (PET$^+$) to enforce intra-session consistency, while negatives are cross-subject with CL gap control $|\mathrm{CL}_a - \mathrm{CL}_n| \geq \Delta_{\min} = 5.0$, ensuring negatives are biochemically distinct while remaining challenging to discriminate. If no candidate meets training-set criteria, we apply multi-stage fallback: (i) progressively relax threshold to $\{2.5, 1.0\}$ and select the patient with maximum CL difference; (ii) if unsuccessful, sample uniformly from different subjects, preventing identity collisions (different `subject_id`) and handling missing values. Negative PET slices are preprocessed identically to positives (resize 224×224, CLIP normalization) and embedded by the same teacher. Anchor embeddings use *cross-attention* (PET→MRI), $\mathbf{e}^{\mathrm{anchor}} = \mathrm{AttnPool}(\mathbf{E}^+_{\mathrm{PET}}, \mathbf{E}_{\mathrm{MRI}})$, to inject MRI context; positives and negatives use *self-attention* on PET$^+$ and PET$^-$ respectively. This yields clinically-structured hard negatives, improving separability and encouraging fine-grained amyloid discrimination.

Loss Function In Phase 2, the training objective combines triplet loss with binary cross-entropy and regularization *(see Fig. 3)*:

$$\mathcal{L}_{\mathrm{Phase2}} = \mathcal{L}_{\mathrm{triplet}} + \lambda_{\mathrm{cls}} \cdot \mathrm{BCE}(z_T, y) + \mathcal{L}_{\mathrm{reg}}, \quad (3)$$

where $\mathcal{L}_{\mathrm{triplet}}$ is the triplet loss in Eq. (4), z_T is the teacher's classification logit, $y \in \{0, 1\}$ is the binary amyloid label, and we use binary cross-entropy with logits (BCE) with $\lambda_{\mathrm{cls}} = 1.0$, and $\mathcal{L}_{\mathrm{reg}}$ includes ℓ_2 penalty on embedding norms and anti-collapse terms. Mini-batches are balanced via `WeightedRandomSampler`; we set `pos_weight`=1.0 to avoid redundant reweighting. This objective enables the teacher to learn discriminative embeddings via metric learning and accurate classification.

Triplet Loss We use triplet loss with margin $m = 1.0$:

$$\mathcal{L}_{\mathrm{triplet}} = \frac{1}{N} \sum_{i=1}^{N} \max\{0,\ \|a_i - p_i\|_2 - \|a_i - n_i\|_2 + m\}, \quad (4)$$

where anchor $a = \mathrm{CrossAttn}(\mathbf{E}^+_{\mathrm{PET}}, \mathbf{E}_{\mathrm{MRI}})$, positive $p = \mathrm{SelfAttn}(\mathbf{E}^+_{\mathrm{PET}})$, and negative $n = \mathrm{SelfAttn}(\mathbf{E}^-_{\mathrm{PET}})$ are patient-level embeddings after attention pooling. Regularization $\mathcal{L}_{\mathrm{reg}}$ includes ℓ_2 norm penalty ($\lambda = 0.01$) and adaptive anti-collapse terms. We train for 15 epochs using AdamW with component-specific learning rates: 5×10^{-6} (encoder/attention), 2×10^{-5} (classifier), and CosineAnnealingWarmRestarts scheduler (full details in supplementary).

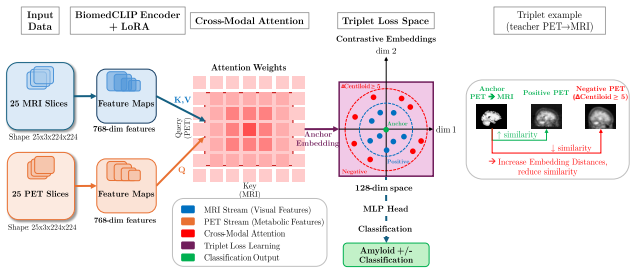


Figure 3. **Phase 2 — Teacher refinement** with online CL-aware triplet mining.

4.7. Phase 3: Knowledge Distillation

MarginFocal Loss We use margin-augmented focal loss with margin-shifted logits $\tilde{z}_i = z_i + m(1 - 2\hat{y}_i)$, where $\hat{y}_i = \mathbf{1}[y'_i > 1/2]$ and y'_i are the (optionally smoothed) targets:

$$\mathcal{L}_{\text{MF}} = \frac{1}{N}\sum_{i=1}^{N}(1-p_{t_i})^{\gamma}\,\text{BCE}_w(\tilde{z}_i, y'_i) + \lambda_{\text{gap}}\,[\,m - (\bar{z}_+ - \bar{z}_-)\,]_+, \tag{5}$$

where γ=2.0, $p_i = \sigma(\tilde{z}_i)$, $p_{t_i} = y'_i p_i + (1-y'_i)(1-p_i)$, and $\bar{z}_\pm$ are batch means of raw logits over positives/negatives. We anneal $m \in [0.3, 1.2]$ over 20 epochs and schedule $\lambda_{\text{gap}} = 0.1$ (epochs 1–10) then 0.3; gap deficit is internally scaled by 0.1 (see supplementary).

Feature Distillation We align the student's MRI-only embeddings $\hat{\mathbf{e}}_S$ with the teacher's cross-modal embeddings $\hat{\mathbf{e}}_T$ via ℓ_2-normalized mean squared error:

$$\mathcal{L}_{\text{feat}} = \|\hat{\mathbf{e}}_S - \hat{\mathbf{e}}_T\|_2^2, \quad \hat{\mathbf{e}} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2}. \tag{6}$$

Normalization focuses on directional alignment rather than magnitude; teacher embeddings are detached to ensure unidirectional knowledge transfer.

Logit Distillation To transfer the teacher's classification calibration and decision boundaries *(see Fig. 4)*, we apply temperature-scaled logit distillation using binary cross-entropy between the student's scaled logits and the teacher's soft targets:

$$\mathcal{L}_{\text{logit}} = T^2 \cdot \text{BCE}\left(\frac{z_S}{T}, \sigma\left(\frac{z_T}{T}\right)\right), \tag{7}$$

where T is the distillation temperature, $\sigma(\cdot)$ is the sigmoid function, and the T^2 factor compensates for the gradient magnitude reduction caused by temperature scaling. Higher temperatures ($T > 1$) produce softer probability distributions enabling the student model to learn from the teacher's uncertainty and relative confidences rather than only from hard binary labels.

Training Configuration Loss weights warm up linearly over 10 epochs: λ_{cls}: 0.3 → 0.4, λ_{feat}: 0.5 → 0.4, λ_{logit}: 0.2, then fixed at $(0.4, 0.4, 0.2)$. Temperature anneals $T = 2.5 \rightarrow 1.0$ over 20 epochs. Batch sizes: 6 (Phases 1–2), 10 (Phase 3); gradient accumulation to 30. In Phase 3 we train for 100 epochs using AdamW with component-specific learning rates: LoRA adapters at 2×10^{-4} (no wd), projection at 1×10^{-4} (wd 10^{-4}), attention/classifier at 1×10^{-4} (wd 10^{-3}). Scheduler: ReduceLROnPlateau (monitoring val. F1; see supplementary).

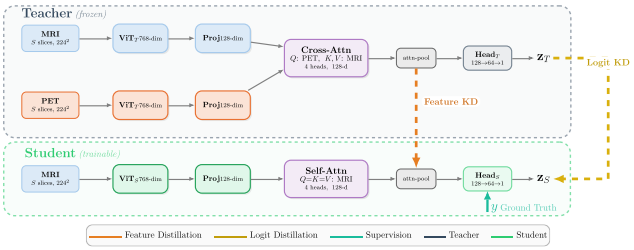


Figure 4. **Phase 3 — Knowledge distillation**. MRI-only student is trained via ℓ_2 feature matching and temperature-scaled logit distillation; PET/CL are used only for teacher supervision.

4.8. Distillation Loss Function

Our training combines three losses (Eqs. (5), (6), (7)):

$$\begin{aligned}\mathcal{L}_{\text{total}} =& \lambda_{\text{cls}}\mathcal{L}_{\text{MF}} \\ &+ \lambda_{\text{feat}}\|\hat{\mathbf{e}}_S - \hat{\mathbf{e}}_T\|_2^2 \\ &+ \lambda_{\text{logit}}\,T^2\,\text{BCE}\left(\frac{z_S}{T}, \sigma\left(\frac{z_T}{T}\right)\right),\end{aligned} \tag{8}$$

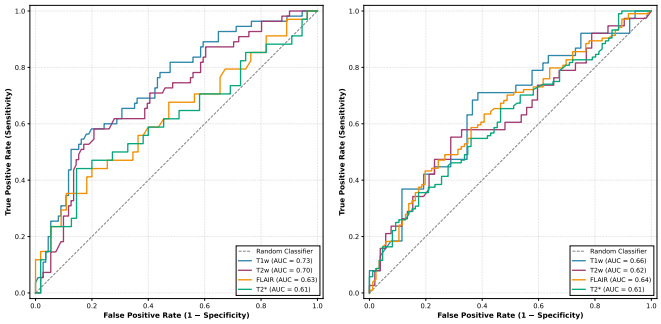
where $\hat{\mathbf{e}} = \text{normalize}(\mathbf{e})$ denotes the ℓ_2-normalized embeddings after attention pooling, z represents the logit output, and T is the distillation temperature (the T^2 factor preserves gradient scale under temperature in binary distillation). Loss weights are annealed as described in Sec. 4.7 [18].

5. Experiments

We evaluate our framework on OASIS-3 and ADNI across four widely available clinical MRI contrasts (T1w, T2w, FLAIR, T2*) capturing complementary tissue properties (anatomy, edema, white-matter lesions, microbleeds) [38, 61]. We first compare with prior MRI-based amyloid prediction methods and present single-modality and multi-contrast results (Sec. 5.1, Table 3), then CDKD (Sec. 5.3), and ablations (Sec. 5.4). Saliency maps confirm spatially plausible focus patterns in predictions (Sec. 5.2). Checkpoints are selected by validation F1 (*Sel.=F1*) or teacher–student embedding similarity (*Sel.=Sim*). Performance is reported at fixed ($\theta = 0.5$) and validation-optimized (θ^*, maximizing F1) thresholds using standard metrics: F1, accuracy (Acc), precision (Prec), recall (Rec), AUC, and negative predictive value (NPV).

5.1. Performance on OASIS-3 and ADNI

Table 3 presents results across both cohorts. T1w achieves the strongest single-sequence performance on OASIS-3 (AUC 0.73 [0.66–0.82]) with balanced metrics at θ^* (F1 0.59, Acc 0.71, NPV 0.81). T2w attains slightly lower AUC (0.70) but comparable F1 (0.55). ADNI results confirm consistency (T1w: AUC 0.66; T2w: 0.62). PET-guided fusion improves MRI-only inference: T1w+T2w→T1w reaches AUC 0.74 on OASIS-3 while preserving ADNI performance (F1 0.61, Rec 0.82 at θ^*). *Sel.=F1* generally outperforms *Sel.=Sim* for single-contrast students, but *Sel.=Sim* benefits multi-contrast models, confirming PET-aligned embeddings stabilize cross-contrast integration. Fig. 5 reveals consistent ranking: T1w>T2w/FLAIR>T2*.



(a) OASIS-3 (b) ADNI

Figure 5. **ROC curves per sequence.** (a) OASIS-3; (b) ADNI.

FLAIR and T2* show moderate discrimination, likely due to smaller sample sizes and weaker structural correlates, yet remain useful at θ^* for high-recall screening.

Comparison with Prior Work Table 4 compares with prior work. Relative to Kim et al. [23], a large MRI-only baseline, our T1w improves AUC from 0.61 to 0.73 (+19.7%) and multi-contrast fusion (T1w+T2w: AUC=0.74 vs. 0.67 for T1w+FLAIR, +10.4%, without non-imaging covariates as in Wang et al. [56]). Key advances include: (1) multi-contrast fusion, (2) CDKD (ADNI↔OASIS-3), and (3) interpretability via saliency and anatomical bias control (Sec. 5.2). Compared to Chiumento et al. [10] (T1w: F1=0.48), PET-guided Centiloid-aware triplet distillation achieves F1=0.59 (+22.9%) extending to CL-standardized multi-contrast evaluation.

Table 4. **Comparison with MRI-based amyloid prediction**. MRI-only inference prioritized; covariate-based method noted.

Method	MRI Input	Ground Truth	Clin.	Training	N	AUC [95% CI]
Kim et al. (2025) [23] (T1w)	**T1w**	**PET (CL)**	**No**	Direct	779	0.61 [0.59–0.63]
Kim et al. (2025) [23] (T1w+FLAIR)	T1w+FLAIR	**PET (CL)**	**No**	Direct	779	0.67 [0.65–0.70]
Wang et al. (2025) [56]	T1w	PET (SUVR>1.11)	Yes†	Transfer	241	0.74
Dolci et al. (2025) [12]	3-modal	CSF (Aβ42)‡	**No**	Direct	–	–
Chiumento et al. (2025) [10]	**T1w**	**PET (CL)**	**No**	Direct	–	–
Ours T1w	**T1w**	**PET (CL)**	**No**	**KD+Trip**	166	**0.73 [0.66–0.82]**
Ours T2w	**T2w**	**PET (CL)**	**No**	**KD+Trip**	166	**0.70 [0.61–0.78]**
Ours T1w+T2w	**T1w+T2w**	**PET (CL)**	**No**	**KD+Trip**	166	**0.74 [0.67–0.82]**

† Requires non-imaging covariates (age, sex, APOE genotype).
– not reported. Dolci: Acc=0.76; Chiumento: F1=0.48 (θ^*).
‡ CSF-based ground truth; not directly comparable to PET-CL.
N = test set; CL = Centiloid; CI = conf. interval (bootstrap 1,000×).

5.2. Interpretability Analysis

Fig. 6 shows saliency/HiResCAM maps across training (epochs 1→25) on OASIS-3 for all MRI contrasts [13, 33, 47]. At epoch 1, attention is diffuse; by epoch 25, the model focuses on regions consistent across contrasts, with spatial alignment between MRI and PET evident in HiResCAM.

Table 3. **Test results on OASIS-3 and ADNI.** *Sel.:* checkpoint by val. F1 or embedding similarity. Metrics: $\theta = 0.5$ vs. optimized θ^*.

	OASIS-3											ADNI										
Modality	Sel.	@0.5				@θ^*				AUC	NPV	Sel.	@0.5				@θ^*				AUC	NPV
		F1	Acc	Prec	Rec	F1	Acc	Prec	Rec				F1	Acc	Prec	Rec	F1	Acc	Prec	Rec		
T1w & T2w available																						
T1w	F1	0.53	0.71	0.57	0.49	0.59	**0.71**	0.56	0.64	0.73	0.81	F1	0.58	0.56	0.48	0.74	**0.61**	0.50	0.46	**0.92**	**0.66**	0.77
T1w	Sim	0.46	0.69	0.54	0.40	0.53	0.49	0.38	**0.87**	0.70	0.83	Sim	0.57	0.54	0.47	0.71	0.59	0.51	0.46	0.82	0.64	0.68
T2w	F1	0.54	0.54	0.40	0.80	0.55	0.63	0.46	0.69	0.70	0.80	F1	0.51	0.63	0.59	0.45	0.57	0.51	0.45	0.76	0.62	0.65
T2w	Sim	0.52	0.69	0.54	0.51	0.54	0.52	0.40	0.84	0.71	0.82	Sim	0.61	0.50	0.46	0.92	**0.61**	0.46	0.44	1.00	0.63	1.00
T1w+T2w→T1w	F1	0.54	0.54	0.40	0.84	**0.61**	**0.71**	0.54	0.71	**0.74**	0.83	F1	0.45	0.64	0.65	0.34	**0.61**	0.56	0.48	0.82	**0.66**	0.73
T1w+T2w→T1w	Sim	0.24	0.70	0.73	0.15	0.54	0.60	0.44	0.73	0.69	0.80	Sim	0.14	0.59	0.60	0.08	0.58	**0.59**	0.51	0.68	0.65	0.69
T1w+T2w→T2w	F1	0.53	0.48	0.38	0.88	0.54	0.58	0.42	0.75	0.70	**0.84**	F1	0.48	0.67	0.70	0.37	0.55	0.49	0.44	0.74	0.61	0.62
T1w+T2w→T2w	Sim	0.47	0.68	0.52	0.42	0.54	0.57	0.42	0.78	0.69	0.81	Sim	0.51	0.62	0.56	0.47	0.53	0.50	0.44	0.66	0.60	0.61
FLAIR & T2* available																						
FLAIR	F1	0.55	0.51	0.42	0.80	0.48	0.58	0.46	0.50	0.63	0.67	F1	0.60	0.60	0.66	0.55	0.66	0.60	0.61	0.71	0.64	0.57
FLAIR	Sim	0.51	0.65	0.55	0.47	0.53	**0.66**	0.57	0.50	0.63	0.71	Sim	0.60	0.57	0.61	0.58	0.66	0.57	0.58	0.77	0.59	0.55
T2*	F1	0.50	0.69	0.64	0.41	0.51	0.52	0.42	0.65	0.61	0.67	F1	0.43	0.55	0.70	0.31	0.68	0.57	0.58	0.81	0.61	0.56
T2*	Sim	0.50	0.45	0.38	0.71	0.52	0.53	0.43	0.68	0.58	0.69	Sim	0.51	0.57	0.68	0.40	**0.74**	0.63	0.60	**0.97**	0.65	**0.86**
FLAIR+T2*→FLAIR	F1	0.49	0.70	0.68	0.38	0.53	0.52	0.42	0.71	0.63	0.69	F1	0.68	0.58	0.58	0.81	0.67	0.56	0.57	0.83	0.61	0.53
FLAIR+T2*→FLAIR	Sim	0.46	0.66	0.59	0.38	0.52	0.63	0.51	0.53	**0.64**	0.70	Sim	0.69	0.61	0.61	0.79	0.73	**0.64**	0.62	0.89	0.67	0.71
FLAIR+T2*→T2*	F1	0.47	0.67	0.62	0.38	0.56	0.54	0.44	0.77	0.63	0.82	F1	0.71	0.56	0.56	0.96	0.71	0.56	0.56	0.95	0.59	0.64
FLAIR+T2*→T2*	Sim	0.54	0.48	0.41	0.79	**0.57**	0.43	0.40	1.00	0.63	1.00	Sim	0.70	0.63	0.62	0.80	0.71	0.56	0.56	0.96	**0.68**	0.64

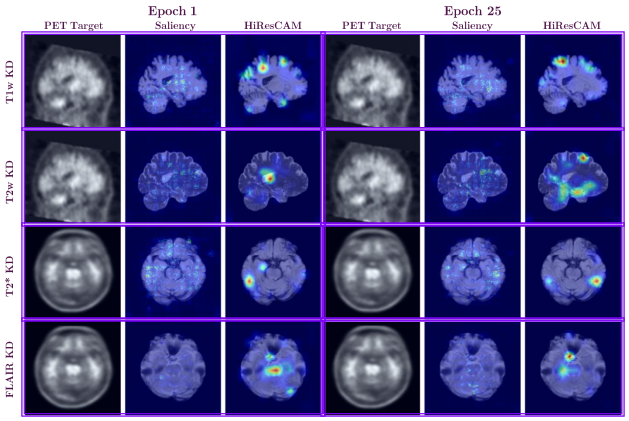


Figure 6. **Saliency/HiResCAM.** OASIS-3, epochs 1 → 25.

5.3. Cross-Dataset Transfer

Table 5 evaluates cross-dataset knowledge distillation (source teacher distills target MRI-only student) [5, 34]. **ADNI→OASIS-3:** ADNI teachers transfer effectively (AUC 0.72–0.73). T2w transfer outperforms native training (T2w: 0.72 vs. 0.70; T1w+T2w→T2w: 0.73 vs. 0.70, Sel.=F1). T1w is stable (ΔAUC $<$ 0.01), while T1w+T2w→T1w reaches F1 0.61 and NPV 0.83 at θ^*. **OASIS-3→ADNI:** T1w transfers best (AUC 0.66, F1 0.60 at θ^*). Fusion preserves or improves NPV under θ^*; *Sel.=F1* balances metrics while *Sel.=Sim* attains higher recall. This reflects target-dataset adaptation in realistic clinical settings with limited paired PET–MRI data.

Table 5. **Cross-dataset knowledge distillation (CDKD).** Source teacher distills target MRI-only student. Metrics at $\theta = 0.5, \theta^*$.

	@0.5				@θ^*				Overall	
Modality	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	AUC	NPV
ADNI (Teacher) KD→ OASIS-3 (Student) — *Sel.=F1*										
T1w	0.27	0.68	0.53	0.18	0.58	**0.70**	0.54	0.64	**0.73**	0.80
T2w	0.55	0.70	0.54	0.56	0.54	0.69	0.54	0.55	0.72	0.77
T1w+T2w→T1w	0.55	0.58	0.42	0.76	**0.61**	0.70	0.53	0.71	0.72	**0.83**
T1w+T2w→T2w	0.39	0.72	0.68	0.27	0.57	0.61	0.45	**0.76**	**0.73**	0.82
OASIS-3 (Teacher) → ADNI (Student) — *Sel.=F1*										
T1w	0.54	0.63	0.58	0.50	**0.60**	**0.61**	0.53	0.68	**0.66**	**0.71**
T2w	0.49	0.60	0.53	0.45	0.57	0.52	0.46	**0.74**	0.61	0.66
T1w+T2w→T1w	0.46	0.60	0.54	0.40	0.56	0.53	0.47	0.71	0.62	0.66
T1w+T2w→T2w	0.10	0.60	1.00	0.05	0.55	0.51	0.45	0.71	0.62	0.63

5.4. Ablation Study

Table 6 shows T1w/T2w ablations. **Phase 1 (Pre-training):** Omitting pre-training lowers OASIS-3 T1w F1 from 0.59 to 0.53 at θ^*, with Prec./Rec. imbalance (0.46/0.75 at $\theta = 0.5$), indicating miscalibration. Larger drop on OASIS-3 (0.59 → 0.53) than ADNI (0.61 → 0.59), likely due to class imbalance. **Phase 2 (Triplet):** Removing triplet mining drops OASIS-3 T1w F1 to 0.51 (ΔF1 = −0.08), showing CL-aware hard-negative separation. **Phase 3A:** Removing embedding alignment yields the largest drop (OASIS-3 T1w: 0.51; T2w: 0.52 at θ^*), lowering T2w NPV to 0.78 (from 0.80–0.82), confirming PET–MRI alignment is critical. **Phase 3B:** Removing logit distillation yields intermediate performance (T1w: 0.55; T2w: 0.52), worsening calibration. **Ranking:** 3A (feature) yields largest F1 gains; 1 (pre-training) helps imbalanced data; 2 (hard negatives) improves discrimination; 3B (logit) refines calibration.

Table 6. **Ablation Study.** T1w/T2w phase impact ($\theta = 0.5, \theta^*$).

		@0.5				@θ^*				Overall	
Dataset	Modality	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	AUC	NPV
No Pre-training (Phase 1)											
OASIS-3	T1w	0.57	0.62	0.46	0.75	0.53	0.71	0.56	0.51	0.74	0.77
OASIS-3	T2w	0.16	0.68	0.56	0.09	0.55	0.62	0.45	0.71	0.67	0.80
ADNI	T1w	0.57	0.57	0.49	0.68	0.59	0.57	0.49	0.73	0.63	0.70
ADNI	T2w	0.51	0.57	0.49	0.53	0.55	0.50	0.44	0.74	0.61	0.63
No Triplet Loss (Phase 2)											
OASIS-3	T1w	0.39	0.72	0.68	0.27	0.51	0.70	0.55	0.47	0.71	0.76
OASIS-3	T2w	0.55	0.61	0.44	0.71	0.56	0.60	0.44	0.78	0.70	0.82
ADNI	T1w	0.62	0.67	0.60	0.63	0.60	0.59	0.51	0.74	0.68	0.71
ADNI	T2w	0.47	0.64	0.64	0.37	0.57	0.52	0.46	0.76	0.60	0.67
No Feature Distillation (Phase 3A)											
OASIS-3	T1w	0.43	0.70	0.58	0.35	0.51	0.68	0.52	0.51	0.71	0.76
OASIS-3	T2w	0.53	0.66	0.49	0.58	0.52	0.56	0.41	0.73	0.71	0.78
ADNI	T1w	0.60	0.56	0.48	0.79	0.61	0.50	0.46	0.92	0.65	0.77
ADNI	T2w	0.50	0.58	0.50	0.50	0.56	0.50	0.45	0.76	0.61	0.64
No Logit Distillation (Phase 3B)											
OASIS-3	T1w	0.61	0.66	0.49	0.78	0.55	0.71	0.57	0.53	0.74	0.77
OASIS-3	T2w	0.54	0.63	0.46	0.66	0.52	0.55	0.40	0.73	0.69	0.78
ADNI	T1w	0.63	0.62	0.54	0.76	0.60	0.53	0.47	0.84	0.67	0.72
ADNI	T2w	0.49	0.58	0.50	0.47	0.57	0.51	0.45	0.76	0.60	0.65

6. Conclusion

We introduce a PET-guided knowledge distillation framework for MRI-only amyloid-β prediction, with no clinical or demographic variables at inference. Unlike prior approaches dependent on genetic or cognitive covariates, our method achieves competitive performance using standard MRI alone, enhancing deployability in settings with high data missingness. A BiomedCLIP teacher, trained via Centiloid-aware triplet mining, transfers knowledge to an MRI-only student network. Evaluations on OASIS-3 and ADNI across four MRI contrasts show strong cross-dataset transfer with CDKD. Ablations confirm that both feature distillation and triplet learning are critical, with pre-training providing additional gains under data imbalance. Saliency analyses reveal anatomically plausible regions contributing to prediction. Overall, this framework offers a promising path toward cost-effective, scalable Alzheimer's biomarker screening using routine MRI. Future work will explore multi-contrast fusion beyond tested pairs, advanced MRI (diffusion, perfusion), and prognostic prediction of amyloid conversion using these sequences.

Acknowledgments

This work was funded by Taighde Éireann – Research Ireland through the Research Ireland Centre for Research Training in Machine Learning (18/CRT/6183). This work was also supported by Taighde Éireann – Research Ireland under Grant number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund through the Research Ireland Insight Centre for Data Analytics at Dublin City University.

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research &Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

Data were provided in part by OASIS-3 (Longitudinal Multimodal Neuroimaging; Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352). AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

References

[1] Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's disease neuroimaging initiative. Public dataset, https://adni.loni.usc.edu/, 2025. 2, 3, 13

[2] Sanka Amadoru, Vincent Doré, Catriona A. McLean, Fairlie Hinton, Claire E. Shepherd, Glenda M. Halliday, Cristian E. Leyton, Paul A. Yates, John R. Hodges, Colin L. Masters, Victor L. Villemagne, and Christopher C. Rowe. Comparison of amyloid PET measured in Centiloid units with neuropathological findings in Alzheimer's disease. *Alzheimer's Research & Therapy*, 12(1):22, 2020. 4

[3] Brian B. Avants, Charles L. Epstein, Murray Grossman, and James C. Gee. Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain. *Medical Image Analysis*, 12(1):26–41, 2008. 4

[4] Weiqi Bao, Fang Xie, Chuantao Zuo, Yihui Guan, and Yiyun Henry Huang. PET Neuroimaging of Alzheimer's Disease: Radiotracers and Their Utility in Clinical Research. *Frontiers in Aging Neuroscience*, 13:624330, 2021. 2

[5] Mariana Bento, Irene Fantini, Justin Park, Leticia Rittner, and Richard Frayne. Deep Learning in Large and Multi-Site Structural Brain MR Imaging Datasets. *Frontiers in Neuroinformatics*, 15:805669, 2022. 8

[6] Ariane Bollack, Lyduine E. Collij, David Vállez García, Mahnaz Shekari, Daniele Altomare, Pierre Payoux, Bruno Dubois, Oriol Grau-Rivera, Mercè Boada, Marta Marquié, Agneta Nordberg, Zuzana Walker, Philip Scheltens, Michael Schöll, Robin Wolz, Jonathan M. Schott, Rossella Gismondi, Andrew Stephens, Christopher Buckley, Giovanni B. Frisoni, Bernard Hanseeuw, Pieter Jelle Visser, Rik Vandenberghe, Alexander Drzezga, Maqsood Yaqub, Ronald Boellaard, Juan Domingo Gispert, Pawel Markiewicz, David M. Cash, Gill Farrar, Frederik Barkhof, and AMYPAD consortium. Investigating reliable amyloid accumulation in Centiloids: Results from the AMYPAD Prognostic and Natural History Study. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 20(5):3429–3441, 2024. 4

[7] Vincent Camus, Pierre Payoux, Louisa Barré, Béatrice Desgranges, Thierry Voisin, Clovis Tauber, Renaud La Joie, Mathieu Tafani, Caroline Hommet, Gaël Chételat, Karl Mondon, Vincent de La Sayette, Jean-Philippe Cottier, Emilie Beaufils, Maria-Joao Santiago Ribeiro, Valérie Gissot, Emilie Vierron, Johnny Vercouillie, Bruno Vellas, Francis Eustache, and Denis Guilloteau. Using PET with 18F-AV-45 (florbetapir) to quantify brain amyloid load in a clinical environment. *European Journal of Nuclear Medicine and Molecular Imaging*, 39(4):621–631, 2012. 2

[8] Richard J. Caselli, Blake T. Langlais, Amylou C. Dueck, Yinghua Chen, Yi Su, Dona E.C. Locke, Bryan K. Woodruff, and Eric M. Reiman. Neuropsychological decline up to 20 years before incident mild cognitive impairment. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 16(3):512–523, 2020. 2

[9] Tamoghna Chattopadhyay, Saket S. Ozarkar, Ketaki Buwa, Neha Ann Joshy, Dheeraj Komandur, Jayati Naik, Sophia I. Thomopoulos, Greg Ver Steeg, Jose Luis Ambite, and Paul M. Thompson. Comparison of deep learning architectures for predicting amyloid positivity in Alzheimer's disease, mild cognitive impairment, and healthy aging, from T1-weighted brain structural MRI. *Frontiers in Neuroscience*, 18:1387196, 2024. 2

[10] Francesco Chiumento, Julia Dietlmeier, Ronan P. Killeen, Kathleen M. Curran, Noel E. O'Connor, and Mingming Liu. Detecting Beta-Amyloid via Cross-Modal Knowledge Distillation from PET to MRI. In *2025 Medical Image Understanding and Analysis Conference (MIUA)*, Leeds, UK, 2025. Frontiers. 4, 7

[11] Lyduine E. Collij, Ariane Bollack, Renaud La Joie, Mahnaz Shekari, Santiago Bullich, Núria Roé-Vellvé, Norman

Koglin, Aleksandar Jovalekic, David Valléz Garciá, Alexander Drzezga, Valentina Garibotto, Andrew W. Stephens, Mark Battle, Christopher Buckley, Frederik Barkhof, Gill Farrar, Juan Domingo Gispert, and Amypad Consortium. Centiloid recommendations for clinical context-of-use from the AMYPAD consortium. *Alzheimer's & Dementia*, 20(12): 9037–9048, 2024. 16

[12] Giorgio Dolci, Charles A. Ellis, Federica Cruciani, Lorenza Brusini, Anees Abrol, Ilaria Boscolo Galazzo, Gloria Menegaz, and Vince D. Calhoun. Multimodal MRI accurately identifies amyloid status in unbalanced cohorts in Alzheimer's disease continuum. *Network Neuroscience*, 9 (1):259–279, 2025. 2, 3, 7

[13] Rachel Lea Draelos and Lawrence Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020. 7, 13

[14] Vladimir S. Fonov, Alan C. Evans, Robert C. McKinstry, C. Robert Almli, and D. Louis Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, 2009. 4, 13

[15] Vladimir S. Fonov, Alan C. Evans, Kelly Botteron, C. Robert Almli, Robert C. McKinstry, D. Louis Collins, and Brain Development Cooperative Group. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1): 313–327, 2011. 4

[16] Jill S. Goldman, Susan E. Hahn, Jennifer Williamson Catania, Susan LaRusse-Eckert, Melissa Barber Butson, Malia Rumbaugh, Michelle N. Strecker, J. Scott Roberts, Wylie Burke, Richard Mayeux, Thomas Bird, and American College of Medical Genetics and the National Society of Genetic Counselors. Genetic counseling and testing for Alzheimer disease: Joint practice guidelines of the American College of Medical Genetics and the National Society of Genetic Counselors. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 13(6):597–605, 2011. 2

[17] Serafettin Gunes, Yumi Aizawa, Takuma Sugashi, Masahiro Sugimoto, and Pedro Pereira Rodrigues. Biomarkers for Alzheimer's Disease in the Current State: A Narrative Review. *International Journal of Molecular Sciences*, 23(9): 4962, 2022. 2

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015. 6

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. 4, 13, 14

[20] Leonardo Iaccarino, Samantha C. Burnham, Ilke Tunali, Jian Wang, Michael Navitsky, Anupa K. Arora, and Michael J. Pontecorvo. A practical overview of the use of amyloid-PET Centiloid values in clinical trials and research. *NeuroImage: Clinical*, 46:103765, 2025. 2, 3, 4

[21] Fabian Isensee, Marianne Schell, Irada Tursunova, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, Martin Bendszus, Klaus Hermann Maier-Hein, and Philipp Kickingereder. Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, 40(17):4952–4964, 2019. 4, 13

[22] Clifford R. Jack, Arvin Arani, Bret J. Borowski, Dave M. Cash, Karen Crawford, Sandhitsu R. Das, Charles DeCarli, Evan Fletcher, Nick C. Fox, Jeffrey L. Gunter, Ranjit Ittyerah, Danielle J. Harvey, Neda Jahanshad, Pauline Maillard, Ian B. Malone, Talia M. Nir, Robert I. Reid, Denise A. Reyes, Christopher G. Schwarz, Matthew L. Senjem, David L. Thomas, Paul M. Thompson, Duygu Tosun, Paul A. Yushkevich, Chadwick P. Ward, and Michael W. Weiner. Overview of ADNI MRI. *Alzheimer's & Dementia*, 20(10):7350–7360, 2024. 16

[23] Donghoon Kim, Jon André Ottesen, Ashwin Kumar, Brandon C. Ho, Elsa Bismuth, Christina B. Young, Elizabeth Mormino, Greg Zaharchuk, and Alzheimer's Disease Neuroimaging Initiative (ADNI). Deep Learning-Based Prediction of PET Amyloid Status Using MRI. *AJNR. American Journal of Neuroradiology*, 46(12):2590–2598, 2025. 2, 3, 7

[24] Jun Pyo Kim, Jonghoon Kim, Hyemin Jang, Jaeho Kim, Sung Hoon Kang, Ji Sun Kim, Jongmin Lee, Duk L. Na, Hee Jin Kim, Sang Won Seo, and Hyunjin Park. Predicting amyloid positivity in patients with mild cognitive impairment using a radiomics approach. *Scientific Reports*, 11: 6954, 2021. 2

[25] William E. Klunk, Robert A. Koeppe, Julie C. Price, Tammie Benzinger, Michael D. Devous, William Jagust, Keith Johnson, Chester A. Mathis, Davneet Minhas, Michael J. Pontecorvo, Christopher C. Rowe, Daniel Skovronsky, and Mark Mintun. The Centiloid Project: Standardizing Quantitative Amyloid Plaque Estimation by PET. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 11 (1):1–15.e4, 2015. 2, 4

[26] Sayantan Kumar, Tom Earnest, Braden Yang, Deydeep Kothapalli, Andrew J. Aschenbrenner, Jason Hassenstab, Chengie Xiong, Beau Ances, John Morris, Tammie L. S. Benzinger, Brian A. Gordon, Philip Payne, Aristeidis Sotiras, and Alzheimer's Disease Neuroimaging Initiative (ADNI). Analyzing heterogeneity in Alzheimer disease using multimodal normative modeling on imaging-based ATN biomarkers. *Alzheimer's & Dementia*, 21(4):e70143, 2025. 3

[27] Pamela J. LaMontagne, Tammie LS Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *medRxiv : the preprint server for health sciences*, 2019. 2, 3, 13

[28] Young-Sil Lee, HyunChul Youn, Hyun-Ghang Jeong, Tae-Jin Lee, Ji Won Han, Joon Hyuk Park, and Ki Woong Kim. Cost-effectiveness of using amyloid positron emission tomography in individuals with mild cognitive impairment. *Cost Effectiveness and Resource Allocation*, 19(1):50, 2021. 2

[29] Sylvain Lehmann, Audrey Gabelle, Marie Duchiron, Germain Busto, Mehdi Morchikh, Constance Delaby, Christophe Hirtz, Etienne Mondesert, Jean-Paul Cristol, Genevieve Barnier-Figue, Florence Perrein, Cédric Turpinat, Snejana Jurici, Karim Bennys, and Alzheimer's Disease Neuroimaging Initiative (ADNI). Comparative performance of plasma pTau181/Aβ42, pTau217/Aβ42 ratios, and individual measurements in detecting brain amyloidosis. *eBioMedicine*, 117:105805, 2025. 4

[30] Christopher O. Lew, Longfei Zhou, Maciej A. Mazurowski, P. Murali Doraiswamy, and Jeffrey R. Petrella. MRI-based Deep Learning Assessment of Amyloid, Tau, and Neurodegeneration Biomarker Status across the Alzheimer Disease Spectrum. *Radiology*, 309(1):e222441, 2023. 2, 3

[31] Xiang Li, Like Li, Minglei Li, Pengfei Yan, Ting Feng, Hao Luo, Yong Zhao, and Shen Yin. Knowledge distillation and teacher–student learning in medical imaging: Comprehensive overview, pivotal role, and future directions. *Medical Image Analysis*, 107:103819, 2026. 3

[32] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. 5, 13

[33] Tanjim Mahmud, Koushick Barua, Sultana Umme Habiba, Nahed Sharmen, Mohammad Shahadat Hossain, and Karl Andersson. An Explainable AI Paradigm for Alzheimer's Diagnosis Using Deep Transfer Learning. *Diagnostics*, 14 (3):345, 2024. 7

[34] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, John-Paul Taylor, Jakub Hort, Jón Snædal, Jaime Kulisevsky, Frederic Blanc, Angelo Antonini, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kłoszewska, Hilkka Soininen, Simon Lovestone, Andrew Simmons, Dag Aarsland, and Eric Westman. The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study. *Medical Image Analysis*, 66:101714, 2020. 8

[35] Nancy Maserejian, Henry Krzywy, Susan Eaton, and James E. Galvin. Cognitive measures lacking in EHR prior to dementia or Alzheimer's disease diagnosis. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 17 (7):1231–1243, 2021. 2, 3

[36] Ana R. Monteiro, Daniel J. Barbosa, Fernando Remião, and Renata Silva. Alzheimer's disease: Insights and new prospects in disease pathophysiology, biomarkers and disease-modifying drugs. *Biochemical Pharmacology*, 211: 115522, 2023. 2

[37] OASIS-3 Imaging Core. OASIS-3: Imaging Methods & Data Dictionary (Version 2.3, Data Release 2.0). Technical report, Washington University in St. Louis, Knight ADRC, 2022. 4

[38] Wiesje Pelkmans, Ellen Dicks, Frederik Barkhof, Hugo Vrenken, Philip Scheltens, Wiesje M. van der Flier, and Betty M. Tijms. Gray matter T1-w/T2-w ratios are higher in Alzheimer's disease. *Human Brain Mapping*, 40(13):3900–3909, 2019. 6

[39] Elizabeth Pirraglia, Ricardo S. Osorio, Lidia Glodzik, Yibeltal Ashebir, and Yongzhao Shao. Subtypes of multiple-etiology dementias and the heterogeneous impact of APOE variants. *Alzheimer's & Dementia*, 21(11):e70872, 2025. 2, 3

[40] Cyrus A. Raji and Tammie L. S. Benzinger. The Value of Neuroimaging in Dementia Diagnosis. *Continuum (Minneapolis, Minn.)*, 28(3):800–821, 2022. 3

[41] Wenhui Ren, Zheng Liu, Yanqiu Wu, Zhilong Zhang, Shenda Hong, and Huixin Liu. Moving Beyond Medical Statistics: A Systematic Review on Missing Data Handling in Electronic Health Records. *Health Data Science*, 4:0176, 2024. 3

[42] Marina Ritchie, Seyed Ahmad Sajjadi, and Joshua D. Grill. Apolipoprotein E Genetic Testing in a New Age of Alzheimer Disease Clinical Practice. *Neurology: Clinical Practice*, 14(2):e200230, 2024. 3

[43] Sarah K. Royse, Davneet S. Minhas, Brian J. Lopresti, Alice Murphy, Tyler Ward, Robert A. Koeppe, Santiago Bullich, Susan DeSanti, William J. Jagust, and Susan M. Landau. Validation of amyloid PET positivity thresholds in centiloids: A multisite PET study approach. *Alzheimer's Research & Therapy*, 13:99, 2021. 4

[44] Michelle Roytman, Faizullah Mashriqi, Khaled Al-Tawil, Paul E. Schulz, Greg Zaharchuk, Tammie L. S. Benzinger, and Ana M. Franceschi. Amyloid-Related Imaging Abnormalities: An Update. *American Journal of Roentgenology*, 220(4):562–574, 2023. 3

[45] Saeid Safiri, Amir Ghaffari Jolfayi, Asra Fazlollahi, Soroush Morsali, Aila Sarkesh, Amin Daei Sorkhabi, Behnam Golabi, Reza Aletaha, Kimia Motlagh Asghari, Sana Hamidi, Seyed Ehsan Mousavi, Sepehr Jamalkhani, Nahid Karamzad, Ali Shamekh, Reza Mohammadinasab, Mark J. M. Sullman, Fikrettin Şahin, and Ali-Asghar Kolahi. Alzheimer's disease: A comprehensive review of epidemiology, risk factors, symptoms diagnosis, management, caregiving, advanced treatments and associated challenges. *Frontiers in Medicine*, 11:1474043, 2024. 2

[46] Jorge Samper-González, Ninon Burgos, Simona Bottani, Sabrina Fontanella, Pascal Lu, Arnaud Marcoux, Alexandre Routier, Jérémy Guillon, Michael Bacci, Junhao Wen, Anne Bertrand, Hugo Bertin, Marie-Odile Habert, Stanley Durrleman, Theodoros Evgeniou, and Olivier Colliot. Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. *NeuroImage*, 183:504–521, 2018. 16

[47] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 7

[48] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 13

[49] Yi Su. Ysu001/PUP. GitHub repository, https://github.com/ysu001/PUP, 2025. 3, 4, 13

[50] Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010. 4, 13

[51] Nicholas J. Tustison, Philip A. Cook, Andrew J. Holbrook, Hans J. Johnson, John Muschelli, Gabriel A. Devenyi, Jeffrey T. Duda, Sandhitsu R. Das, Nicholas C. Cullen, Daniel L. Gillen, Michael A. Yassa, James R. Stone, James C. Gee, and Brian B. Avants. The ANTsX ecosystem for quantitative biological and medical imaging. *Scientific Reports*, 11(1):9068, 2021. 13

[52] U.S. Food and Drug Administration. Leqembi (lecanemab-irmb) injection, for intravenous use: Prescribing Information, 2023. 2

[53] U.S. Food and Drug Administration. KISUNLA (donanemab-azbt) injection, for intravenous use: Prescribing Information, 2024.

[54] Antoine Verger, Igor Yakushev, Nathalie L. Albert, Bart van Berckel, Matthias Brendel, Diego Cecchin, Pablo Aguiar Fernandez, Francesco Fraioli, Eric Guedj, Silvia Morbelli, Nelleke Tolboom, Tatjana Traub-Weidinger, Donatienne Van Weehaeghe, and Henryk Barthel. FDA approval of lecanemab: The real start of widespread amyloid PET use? - the EANM Neuroimaging Committee perspective. *European Journal of Nuclear Medicine and Molecular Imaging*, 50(6): 1553–1555, 2023. 2

[55] Meike W. Vernooij, Francesca Benedetta Pizzini, Reinhold Schmidt, Marion Smits, Tarek A. Yousry, Núria Bargalló, Giovanni Battista Frisoni, Sven Haller, and Frederik Barkhof. Dementia imaging in clinical practice: A European-wide survey of 193 centres and conclusions by the ESNR working group. *Neuroradiology*, 61(6):633–642, 2019. 2, 3

[56] Chenxi Wang, Weiwei Zhang, Ming Ni, Qiong Wang, Chang Liu, Linbin Dai, Mengguo Zhang, Yong Shen, and Feng Gao. Deep-learning based multi-modal models for brain age, cognition and amyloid pathology prediction. *Alzheimer's Research & Therapy*, 17(1):126, 2025. 2, 7

[57] WHO. *A Blueprint for Dementia Research*. World Health Organization, 2022. 2

[58] Ghiam Yamin and David B. Teplow. Pittsburgh Compound-B (PiB) binds amyloid β-protein protofibrils. *Journal of Neurochemistry*, 140(2):210–215, 2017. 2, 13

[59] Jifa Zhang, Yinglu Zhang, Jiaxing Wang, Yilin Xia, Jiaxian Zhang, and Lei Chen. Recent advances in Alzheimer's disease: Mechanisms, clinical trials and new drug development strategies. *Signal Transduction and Targeted Therapy*, 9(1): 211, 2024. 2

[60] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI*, 2 (1):AIoa2400640, 2025. 2, 4

[61] Milica Živanović, Aleksandra Aracki Trenkić, Vuk Milošević, Dragan Stojanov, Miroslav Mišić, Milica Radovanović, and Vukota Radovanović. The role of magnetic resonance imaging in the diagnosis and prognosis of dementia. *Biomolecules and Biomedicine*, 23(2):209–224, 2023. 6

Cross-Modal Knowledge Distillation for PET-Free Amyloid-Beta Detection from MRI

Supplementary Material

A. List of Abbreviations

Table 7 summarizes the main abbreviations used throughout the main paper and supplementary material.

Table 7. **Abbreviations used throughout the main paper and supplementary material.**

Abbr.	Full Form	Abbr.	Full Form
Medical & Imaging			
Aβ	Amyloid-β	OASIS-3	Open Access Series of Imaging Studies 3 [27]
AD	Alzheimer's disease	PET	Positron Emission Tomography
ADNI	Alzheimer's Disease Neuroimaging Initiative [1]	PiB	Pittsburgh Compound B [58]
CL	Centiloid	PUP	PET Unified Pipeline [49]
CSF	Cerebrospinal Fluid	SUVR	Standardized Uptake Value Ratio
FLAIR	Fluid-Attenuated Inversion Recovery	T1w	T1-weighted
T2w	T2-weighted	T2*	T2*-weighted
FWHM	Full Width at Half Maximum	NFTs	Neurofibrillary Tangles
MRI	Magnetic Resonance Imaging	APOE4	Apolipoprotein E $\varepsilon 4$ allele
APOE	Apolipoprotein E	AV-45	Florbetapir (^{18}F-AV-45)
ATN	Amyloid–Tau–Neurodegeneration framework	SWI	Susceptibility-Weighted Imaging
ARIA	Amyloid-Related Imaging Abnormalities	ICBM152	MNI ICBM152 brain template [14]
DTI	Diffusion Tensor Imaging	EHR	Electronic Health Record
GRE	Gradient-Recalled Echo		
Deep Learning & Architecture			
BCE	Binary Cross-Entropy	MLP	Multi-Layer Perceptron
CLS	Classification Token	ReLU	Rectified Linear Unit
GELU	Gaussian Error Linear Unit	ViT	Vision Transformer
KD	Knowledge Distillation	MHA	Multi-Head Attention
LoRA	Low-Rank Adaptation [19]	CDKD	Cross-Dataset Knowledge Distillation
CLIP	Contrastive Language–Image Pre-training	BiomedCLIP	Biomedical CLIP vision–language model
Preprocessing & Tools			
ANTs	Advanced Normalization Tools [51]	MNI	Montreal Neurological Institute [14]
HD-BET	HD Brain Extraction Tool [21]	N4	N4 Bias Field Correction [50]
Evaluation Metrics			
Acc	Accuracy	F1	F1 score
AUC	Area Under the ROC Curve	Prec	Precision
Rec	Recall	NPV	Negative Predictive Value
CI	Confidence Interval	ROC	Receiver Operating Characteristic
Training			
AdamW	Adam with Weight Decay [32]	lr	Learning Rate
FP16	16-bit Floating Point	wd	Weight Decay

B. Supplementary Material Overview

In this section, we provide technical details and additional experiments that support the main paper. In Sec. C, we present further implementation details, including augmentation strategies, hyperparameters for all three training phases, and architecture-specific formulas referenced in the main text. In Sec. D, we report additional qualitative and quantitative results: radar charts to visualize performance trade-offs across metrics and to compare single-sequence models with multi-sequence distilled models (Fig. 7); ablation studies evaluating different Centiloid thresholds for negative patient selection during triplet mining (Sec. D.2, Table 9); ROC curve evolution from the first to the last epoch (Sec. D.3, Fig. 8); and interpretability analyses (Sec. D.4) for both single- and multi-sequence models via gradient-based saliency [48] and HiResCAM [13], which confirm that the model's attention focuses on anatomically plausible regions (Figs. 9–10).

C. Implementation Details

C.1. Data Augmentation

To improve generalization while preserving PET–MRI spatial correspondence and to facilitate reproducibility, we apply the following synchronized augmentations during Phases 1–2 using shared random seeds for each PET–MRI pair:

- **Spatial transformations:** random affine (rotation $\pm 7°$), translation $\pm 5\%$, and isotropic scaling in $[0.95, 1.05]$.
- **Intensity modulations:** color jitter (brightness/contrast $\pm 10\%$), Gaussian blur (kernel size 3, $p = 0.3$), gamma correction $\gamma \in [0.9, 1.1]$ ($p = 0.5$), and Gaussian noise $\sigma \in [0.01, 0.03]$ ($p = 0.5$).
- **Random erasing:** $p = 0.25$, erase scale in $[0.05, 0.12]$.

The same augmentations are applied during Phase 3, with PET–MRI synchronization preserved (shared random seeds). At test time, only resizing and normalization are used.

Framework & hardware. We use PyTorch 2.x with CUDA 12.x on NVIDIA GeForce RTX 4090/5090 GPUs (24–32 GB VRAM).

C.2. Training Configuration

We use PyTorch with seed $= 42$, mixed-precision training (FP16), and gradient accumulation to increase the effective batch size. Hyperparameters are tuned on the validation set and then kept fixed for all experiments. Early stopping is based on validation F1 (Phases 1 and 3) and on a combined score (triplet separation + F1) in Phase 2.

Class Balancing and Sampling To mitigate label imbalance, we use per-class inverse-frequency weights and a weighted sampler. Let n_0, n_1 be the counts of negative/positive samples in the training set; we set per-class weights $w_c = 1/n_c$ and assign each sample the weight of its class. A `WeightedRandomSampler` (replacement) is used to draw mini-batches with balanced label proportions. We do not apply additional positive reweighting in the BCE (i.e., `pos_weight` $= 1.0$) to avoid double-counting the imbalance already handled by sampling.

Reproducibility. Fixed seed (42) for all stochastic operations (data loading, augmentation, model initialization, dropout); CuDNN in deterministic mode with benchmarking disabled.

Training: Mixed precision with `GradScaler` and gradient accumulation; batch sizes 6 (Phases 1–2), 10 (Phase 3).

Early Stopping:
- Phase 1 (Pre-training): validation F1, patience 5
- Phase 2 (Contrastive): validation combined score (triplet separation + F1), patience 3
- Phase 3 (Distillation): validation F1, patience 25

Data loading & checkpointing: 4 DataLoader workers with pin memory enabled. Phase 1 (pre-training): best validation F1. Phase 2 (teacher): best *combined score* (triplet separation + classification F1, weight 1:0.5). Phase 3 (student): best teacher–student similarity and best validation F1 (saved separately).

C.3. Architecture Implementation Details

C.3.1. LoRA Parameterization

Low-Rank Adaptation (LoRA) [19] reparameterizes each projection with pretrained weights $\mathbf{W}_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ as:

$$\mathbf{W}' = \mathbf{W}_0 + s\,\Delta\mathbf{W}, \quad \Delta\mathbf{W} = \mathbf{B}\mathbf{A}, \quad s = \alpha/r, \tag{9}$$

where $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$ are the only trainable parameters, with rank $r = 32$ and scaling factor $\alpha = 32$ (so $s = \alpha/r = 1.0$), while $\mathbf{W}_0$ remains frozen. We rely on the standard LoRA initialization from the PEFT library: the up-projection $\mathbf{B}$ is initialized to zero so that $s\,\mathbf{B}\mathbf{A} = \mathbf{0}$ at the beginning of training, and $\mathbf{W}' = \mathbf{W}_0$. LoRA adapters are applied to attention projections (query, key, value, output) in transformer blocks 6–11 (0-indexed).

C.3.2. Projection Head Architecture

ViT CLS tokens (768D) are projected to 128D using the dropout rates specified in Table 8:

$$\begin{aligned} \mathbf{h}^{(1)} &= \text{Dropout}_1\big(\text{GELU}(\text{LN}_1(\mathbf{W}_1\mathbf{h}))\big), \\ \mathbf{e} &= \text{Dropout}_2\big(\text{LN}_2(\mathbf{W}_2\mathbf{h}^{(1)})\big). \end{aligned} \tag{10}$$

where $\mathbf{W}_1 : 768 \to 256$ and $\mathbf{W}_2 : 256 \to 128$.

C.3.3. Attention Pooling Formula

Given slice embeddings $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_S] \in \mathbb{R}^{S \times 128}$ after MHA (4 heads, $d_{\text{head}} = 32$), with S slices, we compute:

$$w_s = \mathbf{W}_{\text{pool}}\mathbf{A}_s + b, \quad \alpha_s = \text{softmax}(w_s/\tau), \tag{11}$$

where $\mathbf{W}_{\text{pool}} \in \mathbb{R}^{1 \times 128}$, $b \in \mathbb{R}$ is a scalar bias term, and $\tau = 2.0$. The patient representation is $\mathbf{e} = \sum_{s=1}^{S} \alpha_s\,\mathbf{A}_s$.

C.3.4. Classification Head Architecture

Both teacher and student use:

$$z = \mathbf{W}_2(\text{Dropout}(\text{ReLU}(\mathbf{W}_1\mathbf{e}))), \tag{12}$$

with $\mathbf{W}_1 : 128 \to 64$, $\mathbf{W}_2 : 64 \to 1$.

Weight initialization: the attention pooling weight uses Xavier-uniform (gain = 1.0) with bias 0.0; other linear layers use PyTorch defaults. Before student distillation, the classifier output layer (64→1) is re-initialized with Xavier-uniform (gain = 0.5) and bias 0.0.

Table 8. **Dropout schedules.** Dropout rates for projection and classification heads in teacher (Phases 1–2) and student (Phase 3).

	Phases 1–2 (Teacher)	Phase 3 (Student)
Projection head (p_1, p_2)	$(0.5,\ 0.4)$	$(0.3,\ 0.2)$
Classification head p	0.6	0.4

Phase 1 - Complete Hyperparameters
- Epochs: 30
- Batch size: 6
- Optimizer: AdamW with lr $= 2 \times 10^{-5}$, wd $= 1 \times 10^{-3}$ (uniform)
- Slices per subject: 25 (uniformly spaced)
- Loss: BCE (no label smoothing)
- Gradient clipping: max_norm dynamically adjusted (1.0 for epochs 1–2, 2.0 for epochs 3–5, 5.0 thereafter)

Phase 2 - Complete Hyperparameters
- Epochs: 15
- Batch size: 6
- Optimizer: AdamW with component-specific rates:
  - Vision backbone, projection, attention: lr $= 5 \times 10^{-6}$, wd $= 10^{-2}$
  - Classification head: lr $= 2 \times 10^{-5}$, wd $= 5 \times 10^{-3}$
- Scheduler: CosineAnnealingWarmRestarts with $T_0 = 5$, $T_{\text{mult}} = 2$, $\eta_{\min} = 10^{-7}$

Phase 2 Regularization $\mathcal{L}_{\text{reg}}$ includes three components: (i) ℓ_2 penalty on anchor, positive, and negative embedding norms with coefficient 0.01; (ii) inter-anchor similarity penalty, penalizing mean pairwise cosine similarity above 0.5; (iii) anchor-negative similarity penalty (weight 0.5), penalizing mean similarity above -0.1. Components (ii–iii) use progressive epoch-dependent scaling: $s = 0.1$ for epochs 1–3, then linearly increasing to 1.0 by epoch 13.

MarginFocal Loss Details (Phase 3) Complete hyperparameters: $\gamma = 2.0$ (focal parameter), $w = 1.0$ (pos_weight for balanced batches), label smoothing $= 0$, $\varepsilon = 10^{-8}$ (numerical stability floor).

The positive-weighted BCE is: $\text{BCE}_w(\tilde{z}, y') = -wy' \log \sigma(\tilde{z}) - (1 - y') \log(1 - \sigma(\tilde{z}))$.

Margin annealing schedule:
- Epochs 1–6: $m = 0.3$
- Epochs 7–20: m is linearly increased from 0.3 towards the final value
- Epochs 21+: $m = 1.2$ (fixed)

Gap deficit scaling: The term $[\,m - (\bar{z}_+ - \bar{z}_-)\,]_+$ is multiplied by 0.1 internally before adding to the loss, yielding effective $\lambda_{\text{gap}} \approx 0.01$ (epochs 1–10) and 0.03 (epochs 11+).

Phase 3 - Complete Hyperparameters

- Epochs: 100
- Batch size: 10
- Optimizer: AdamW with component-specific rates:
  - LoRA adapters: lr $= 2 \times 10^{-4}$, wd $= 0$ (preserve low-rank structure)
  - Projection modules: lr $= 1 \times 10^{-4}$, wd $= 1 \times 10^{-4}$
  - Attention modules: lr $= 1 \times 10^{-4}$, wd $= 1 \times 10^{-3}$
  - Classification head: lr $= 1 \times 10^{-4}$, wd $= 1 \times 10^{-3}$
- Scheduler: ReduceLROnPlateau (mode=max, factor=0.7, patience=15, threshold=0.005, `min_lr` $= 10^{-5}$, cooldown=2; monitoring validation F1)
- Temperature annealing: $T = 2.5$ (epochs 1–6), linearly decreased (epochs 7–20) towards $T = 1.0$, and fixed at $T = 1.0$ (epochs 21+)

Knowledge Distillation Loss Weights Warm-up During Phase 3, we linearly warm up the loss weights over the first 10 epochs from $(\lambda_{\text{cls}}, \lambda_{\text{feat}}, \lambda_{\text{logit}}) = (0.3, 0.5, 0.2)$ to $(0.4, 0.4, 0.2)$. After epoch 10, the weights are kept fixed at $(\lambda_{\text{cls}}, \lambda_{\text{feat}}, \lambda_{\text{logit}}) = (0.4, 0.4, 0.2)$.

D. Additional Experimental Results

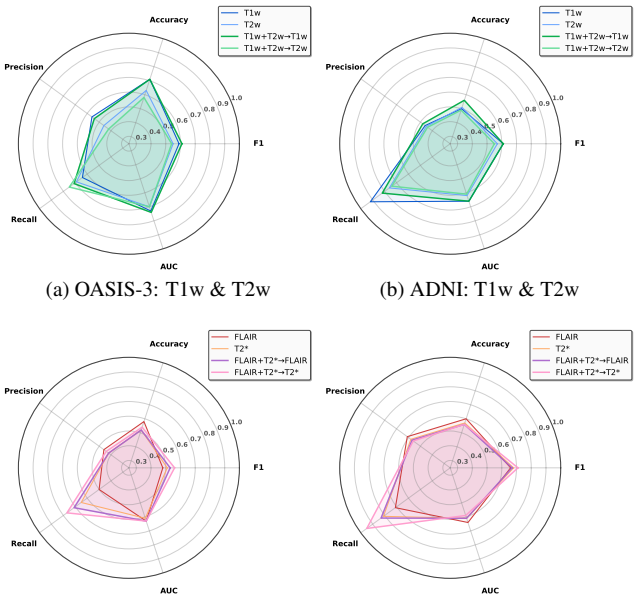
D.1. Multi-Contrast Performance Analysis

Fig. 7 compares models that use a single sequence with models that use multiple sequences and are then tested on a single sequence using five metrics (F1, Accuracy, Precision, Recall, AUC). Multi-sequence models consistently improve recall (e.g., T1w on OASIS-3: 0.64 → 0.71, +10.9%), while also increasing AUC on OASIS-3 (0.73 → 0.74) and accuracy for T1w on ADNI (0.50 → 0.56), and maintaining comparable precision and AUC in both cohorts. On OASIS-3 (Fig. 7c), FLAIR+T2* → T2* achieves the largest F1 improvement (0.51 → 0.56, +9.8%) with substantial recall gains (0.65 → 0.77). In ADNI, the recall gain is larger; in particular, FLAIR+T2* → T2* reaches F1 0.71 with recall 0.95. Across all four charts, multi-sequence distillation yields higher recall while maintaining or improving the other metrics, suggesting that using multiple sequences provides the model with richer context when predicting from a single sequence and is therefore preferable in settings where minimizing false negatives is critical.

D.2. Ablation Study: Impact of Centiloid-Based Negative Mining

To assess the contribution of the Centiloid-guided negative mining strategy used in Phase 2, we vary the minimum amyloid burden difference threshold $\Delta_{\text{CL}}^{\min}$ required between anchor and negative samples (Table 9) and evaluate three configurations for both T1w and T2w sequences:

- **Uniform ($\Delta_{\text{CL}}^{\min} = 0$):** Negatives sampled uniformly, representing a baseline where triplet learning relies on visual similarity.



(a) OASIS-3: T1w & T2w (b) ADNI: T1w & T2w

(c) OASIS-3: FLAIR & T2* (d) ADNI: FLAIR & T2*

Figure 7. **Single vs Multi-Sequence Performance Comparison.** Spider charts comparing single-sequence and multi-sequence distilled models across metrics (F1, Accuracy, Precision, Recall, AUC). Top: T1w+T2w; bottom: FLAIR+T2*. Left: OASIS-3; right: ADNI. Multi-sequence distillation improves recall.

- **Moderate ($\Delta_{\text{CL}}^{\min} = 5.0$):** configuration used in the main experiments, requiring negatives to differ by at least 5 CL units from the anchor.
- **Strict ($\Delta_{\text{CL}}^{\min} = 10.0$):** more restrictive threshold enforcing larger amyloid burden differences, yielding harder negatives but reducing the number of eligible triplets.

Analysis of Results Table 9 shows that using a specific Centiloid gap constraint value ($\Delta_{\text{CL}}^{\min} > 0$) is beneficial compared to using a uniformly selected negative patient. The choice $\Delta_{\text{CL}}^{\min} = 5.0$ provides the best trade-off between sensitivity and specificity across datasets, with better calibration of the metrics. For T1w on OASIS-3, the threshold of 5.0 improves the F1 score from 0.53 to 0.59 (+11.3%) and increases NPV from 0.77 to 0.81 compared to ($\Delta_{\text{CL}}^{\min} = 0$), reducing the number of missed amyloid-positive cases. The recall improvement (0.56→0.64) implies a relative reduction of approximately 18% in the number of false negatives in the test set. The **Strict** strategy ($\Delta_{\text{CL}}^{\min} = 10.0$) underperforms on OASIS-3 T1w (F1 = 0.52, AUC = 0.68), suggesting that using too high a Centiloid value prevents the network from learning smaller differences between patients that are not too dissimilar, as obtained with $\Delta_{\text{CL}}^{\min} = 5.0$. For T2w on OASIS-3, all three strategies lead to similar F1 scores, but using a moderate $\Delta_{\text{CL}}^{\min}$ achieves an AUC of 0.70 (similar to the strict config-

Table 9. **Ablation study: impact of Centiloid-based negative mining.** Test performance on OASIS-3 and ADNI for T1w and T2w models when varying the minimum Centiloid gap $\Delta_{\text{CL}}^{\min}$ between anchor and negative samples. We report metrics both at a fixed decision threshold of 0.5 and at the validation-optimized threshold θ^*.

	OASIS-3											ADNI										
Sequence	$\Delta_{\text{CL}}^{\min}$	@0.5				@θ^*				AUC	NPV	$\Delta_{\text{CL}}^{\min}$	@0.5				@θ^*				AUC	NPV
		F1	Acc	Prec	Rec	F1	Acc	Prec	Rec				F1	Acc	Prec	Rec	F1	Acc	Prec	Rec		
T1-weighted MRI																						
T1w	0.0	0.53	0.49	0.38	0.87	0.53	0.68	0.51	0.56	0.72	0.77	0.0	0.60	0.63	0.56	0.66	0.60	0.60	0.52	0.71	0.65	0.71
T1w	5.0	0.53	0.71	0.57	0.49	0.59	0.71	0.56	0.64	0.73	0.81	5.0	0.58	0.56	0.48	0.74	0.61	0.50	0.46	0.92	0.66	0.77
T1w	10.0	0.53	0.65	0.47	0.60	0.52	0.63	0.46	0.60	0.68	0.77	10.0	0.59	0.60	0.52	0.68	0.58	0.53	0.47	0.76	0.65	0.68
T2-weighted MRI																						
T2w	0.0	0.47	0.66	0.49	0.46	0.55	0.60	0.44	0.75	0.69	0.81	0.0	0.55	0.53	0.46	0.68	0.56	0.51	0.45	0.74	0.59	0.64
T2w	5.0	0.54	0.54	0.40	0.80	0.55	0.63	0.46	0.69	0.70	0.80	5.0	0.51	0.63	0.59	0.45	0.57	0.51	0.45	0.76	0.62	0.65
T2w	10.0	0.55	0.61	0.44	0.71	0.55	0.61	0.44	0.71	0.70	0.80	10.0	0.52	0.59	0.51	0.53	0.58	0.51	0.46	0.79	0.60	0.67

uration and higher than the uniform baseline at 0.69), indicating better calibration.

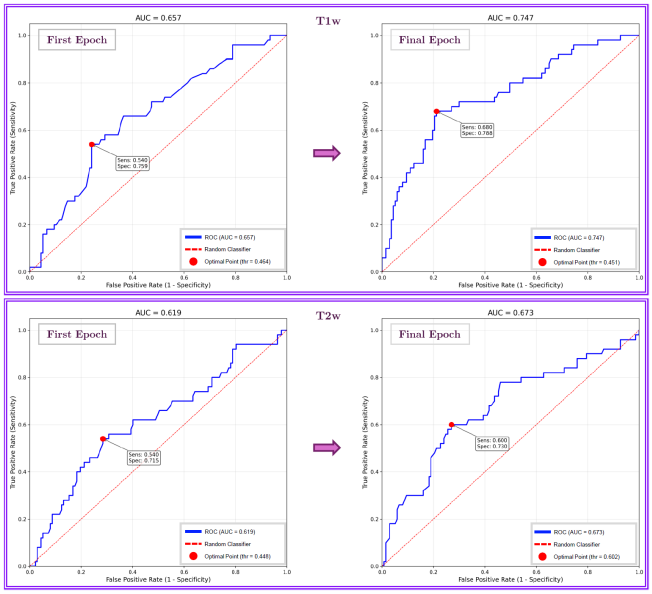
On ADNI, the trends are similar but less pronounced. For T1w, using the Moderate configuration (F1 = 0.61, AUC = 0.66) marginally outperforms both Uniform (F1 = 0.60, AUC = 0.65) and Strict (F1 = 0.58, AUC = 0.65). A similar pattern is observed for T2w, with Moderate achieving F1 = 0.57 and AUC = 0.62. We also note that the Uniform baseline remains competitive on ADNI (e.g., T1w: F1 = 0.60), likely due to its more balanced class distribution, which reduces the risk of trivial negatives. The smaller performance gap on ADNI suggests that CL-aware mining provides greater benefits in imbalanced settings. These results validate our choice of $\Delta_{\text{CL}}^{\min} = 5.0$ as the optimal balance: it enforces biochemically meaningful separation by requiring negatives to differ by at least 5 Centiloid units from the anchor, a threshold that exceeds both the test-retest measurement error (2.5–3.5 CL) and the reliable annual amyloid accumulation rate (3–5 CL/year) [11], while maintaining sufficient triplet diversity for effective contrastive learning. Across both datasets and contrasts, the Moderate setting either matches or outperforms the Uniform and Strict strategies, with the largest absolute gains observed on OASIS-3.

D.3. Training Convergence Analysis

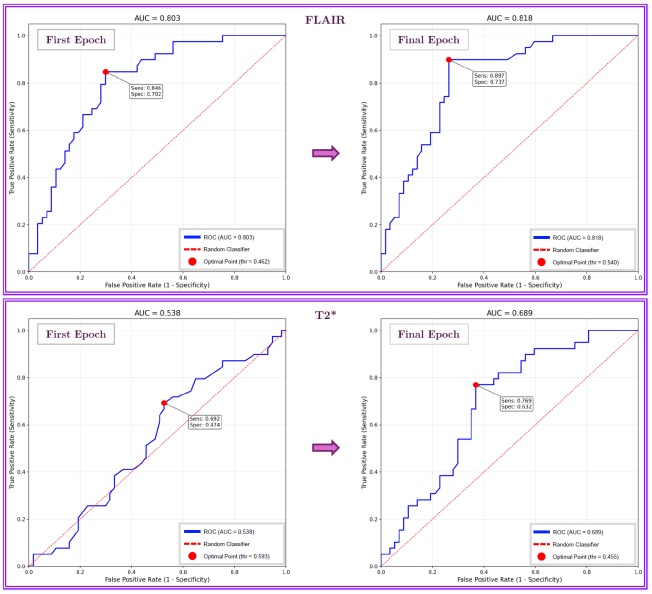
Fig. 8 visualizes the evolution of the ROC curves from the first epoch (Epoch 1) to the last epoch (Final Epoch) on the validation set. We plot the final training epoch rather than the early-stopping checkpoint. All models show positive ΔAUC improvements from the first to the final epoch, confirming effective knowledge distillation. OASIS-3 achieves superior results in terms of AUCs compared to ADNI, possibly due to differences in cohort composition, image acquisition protocols, or amyloid distribution across datasets [22, 46].

D.4. Interpretability

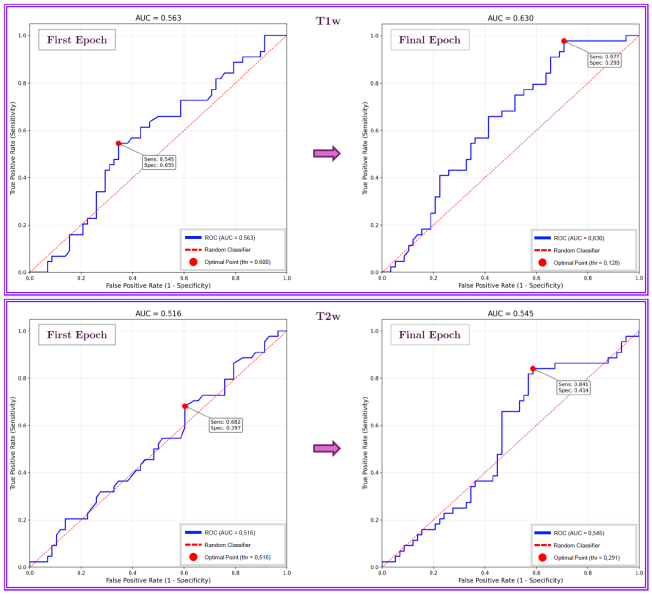
Figs. 9 and 10 show how the model's spatial attention evolves during training. We visualize three epochs (1, 8, and 25) for both single-sequence models (T1w, T2w, FLAIR, T2*) and multi-sequence models (T1w+T2w and FLAIR+T2*), each tested on an individual contrast. For each dataset (OASIS-3 and ADNI) and configuration, we display the target PET image with gradient-based saliency maps and HiResCAM explanations. At epoch 1, both saliency and HiResCAM are relatively diffuse across the brain volume, indicating that the networks initially rely on non-specific global patterns. By epoch 8, the model's attention becomes more structured and begins to concentrate on regions that more closely correspond to the reference PET signal. By epoch 25, the maps are more focal, highlighting neuroanatomically plausible regions. Qualitatively, the attention patterns are consistent between OASIS-3 and ADNI, suggesting that the learned features capture generalizable amyloid-related patterns rather than dataset-specific artifacts.



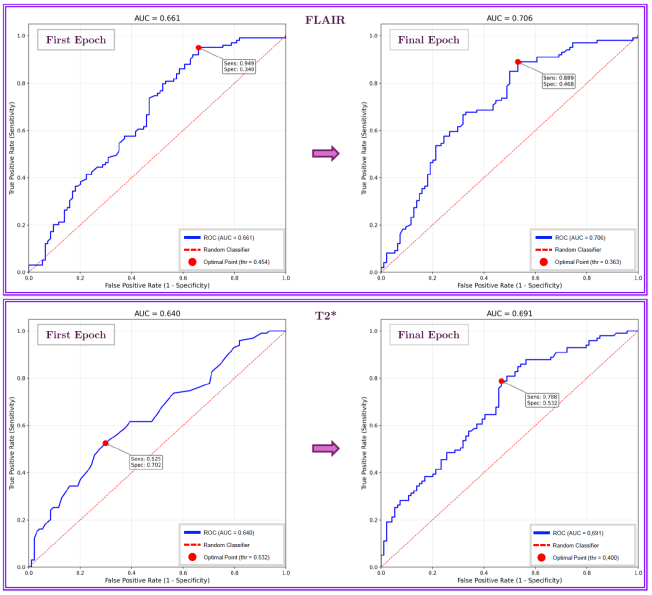
(a) OASIS-3: T1w and T2w



(b) OASIS-3: FLAIR and T2*

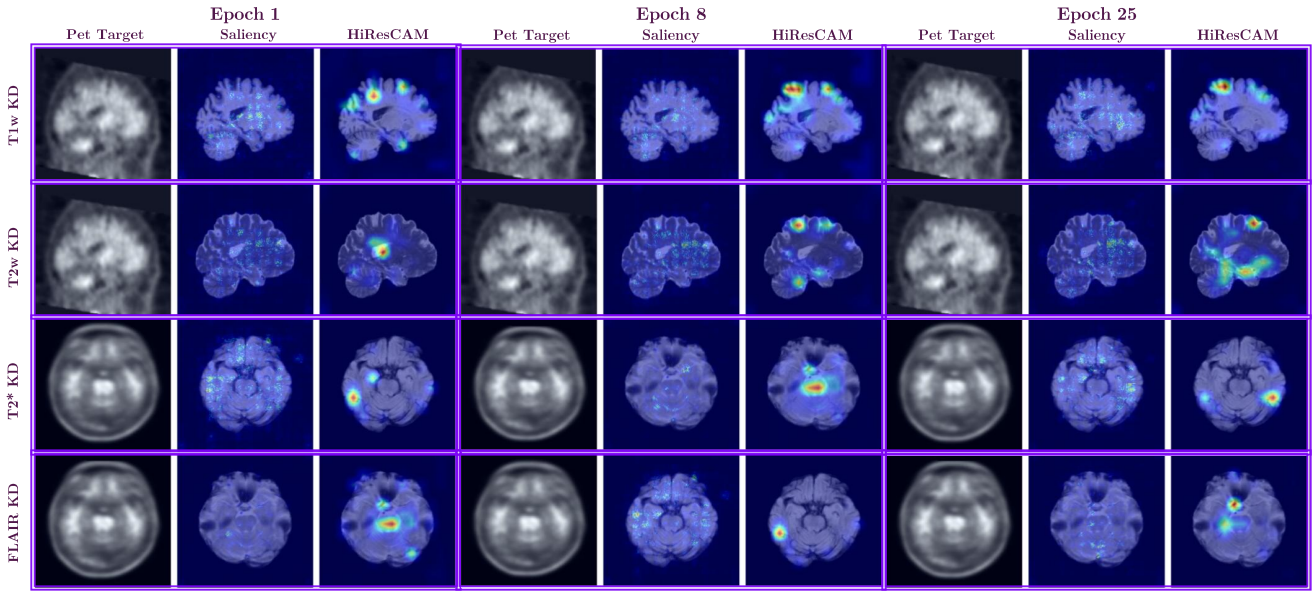


(c) ADNI: T1w and T2w

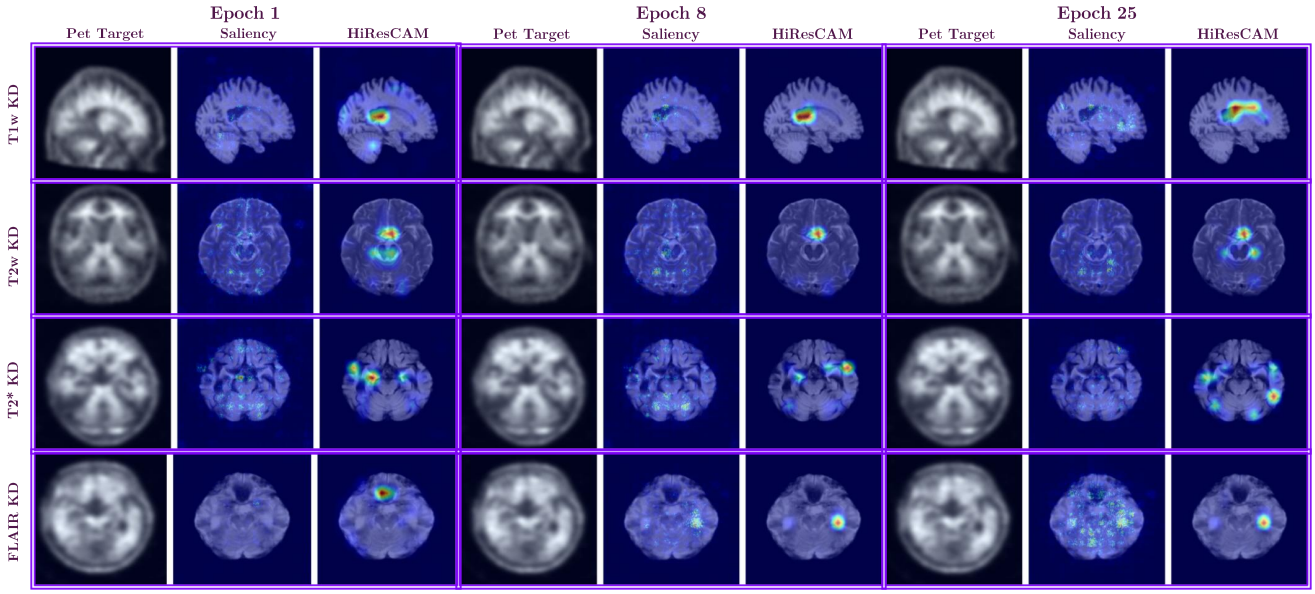


(d) ADNI: FLAIR and T2*

Figure 8. **ROC curve evolution across datasets and sequences.** Validation performance at Epoch 1 (initialization) vs. Final Epoch (convergence). Top row: OASIS-3 dataset for T1w+T2w training (left) and FLAIR+T2* training (right). Bottom row: ADNI dataset with same training configurations. All models show substantial AUC improvements demonstrating effective knowledge distillation.

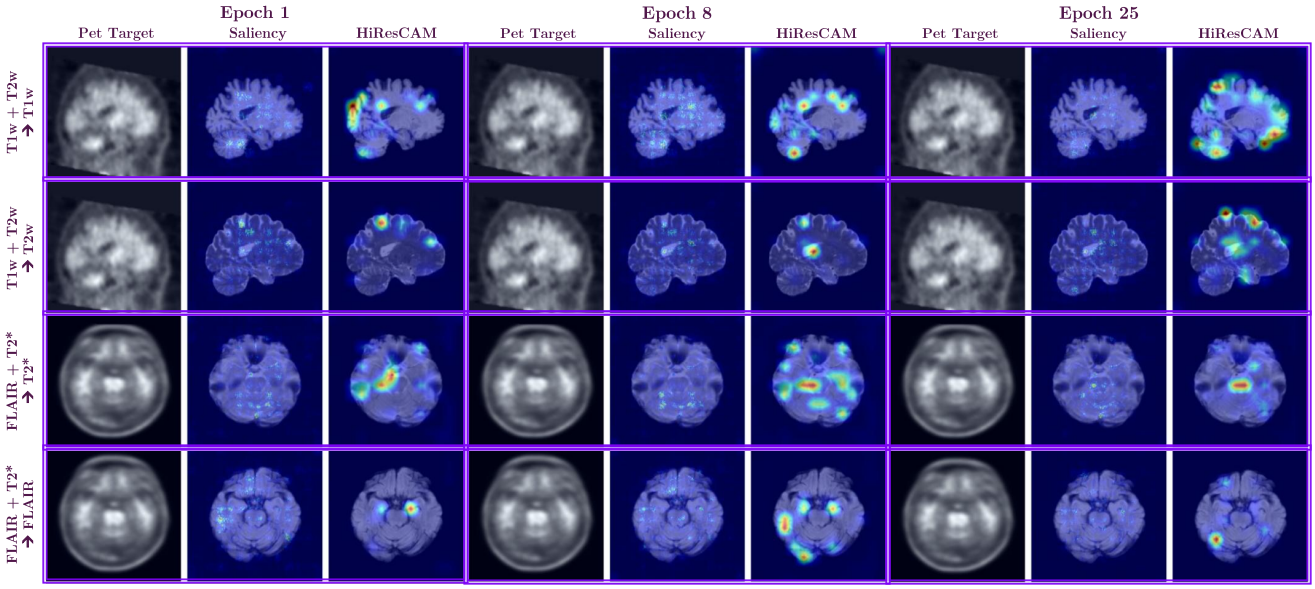


(a) OASIS-3 dataset

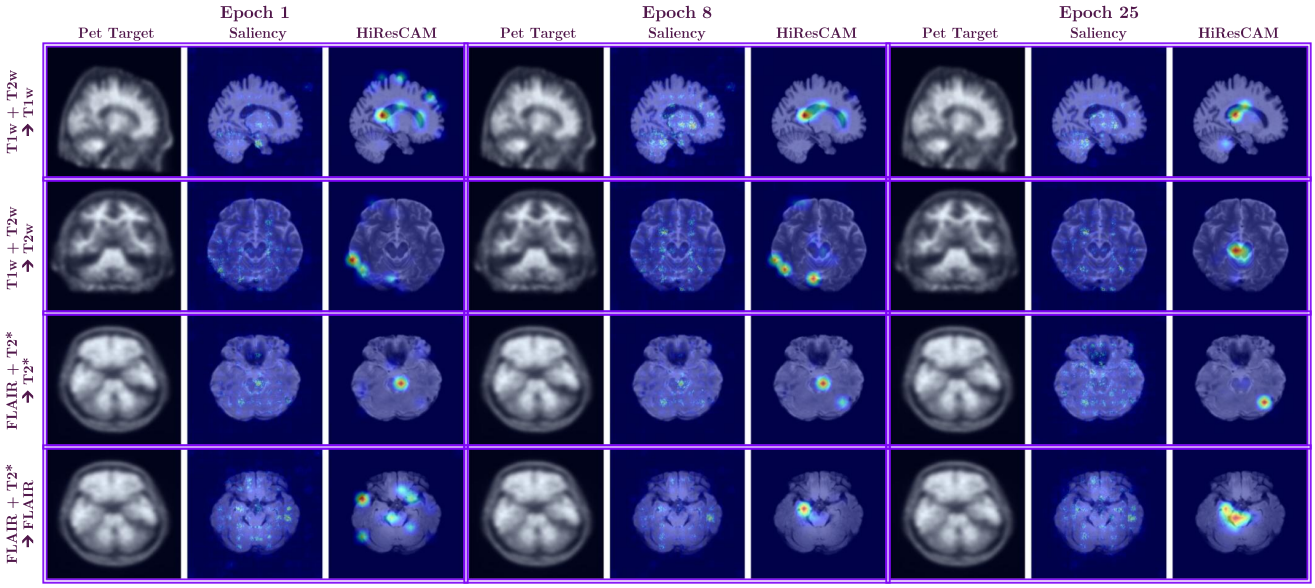


(b) ADNI dataset

Figure 9. **Single-sequence training: attention evolution across datasets.** Training progression (epochs 1, 8, 25) for models trained on individual MRI contrasts. Each row displays PET reference, target MRI, gradient saliency, and HiResCAM maps. The network progressively focuses on anatomically relevant brain structures, with consistent patterns across OASIS-3 and ADNI datasets demonstrating robust generalization.



(a) OASIS-3 dataset



(b) ADNI dataset

Figure 10. **Multi-sequence training with single-sequence inference.** Saliency/HiResCAM evolution (epochs 1, 8, 25) for models trained on paired sequences (T1w+T2w, FLAIR+T2*) and tested on individual contrasts, showing consistent spatial attention patterns across modalities.