GeoSAE: Geometric Prior-Guided Layer-Wise Sparse Autoencoder Annotation of Brain MRI Foundation Models

Favour Nerrise Lucy Yin Mohammad H. Abbasi Kilian M. Pohl Ehsan Adeli†
Stanford University, CA
{fnerrise, lucyyin, mabbasi, kpohl, eadeli}@stanford.edu

Abstract

*Brain MRI foundation models learn rich representations of anatomy, but interpreting what clinical information they encode remains an open problem. Standard sparse autoencoders (SAEs) suffer from severe feature collapse in deep transformer layers, and in Alzheimer's disease (AD) research, aging confounds nearly every clinical variable, making naïve annotation unreliable. We propose **GeoSAE**, a geometry-guided SAE framework that uses the foundation model's learned manifold structure to prevent feature collapse and annotates each surviving feature via age-deconfounded partial correlations. Applied to ∼14k T1-weighted MRI scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Australian Imaging Biomarkers and Lifestyle (AIBL) datasets, **GeoSAE** identifies a compact, fully interpretable feature set that predicts mild cognitive impairment (MCI)-to-AD conversion (AUC 0.746) using only 2% of the embedding dimensions, while comorbidity-annotated features achieve only chance-level performance. The identified features replicate across cohorts without retraining (r=0.97) and localize to neuroanatomically distinct regions consistent with Braak staging. This shows that geometry-guided SAEs can extract interpretable, biomarkers from frozen brain MRI foundation models.*

1. Introduction

Brain MRI foundation models (FMs) pretrained on tens of thousands of scans via self-supervised learning achieve strong downstream transfer for disease classification, age prediction, and brain-age estimation [4, 24, 41]. However, these models produce high-dimensional, entangled representations whose clinical content is difficult to interpret. In Alzheimer's disease (AD) research, where aging is both the primary risk factor and a pervasive confound, this opacity poses a direct clinical risk. Features that appear to predict AD progression may instead reflect normal aging, comorbidity burden, or scanner artifacts [34]. Reliable deployment of brain MRI FMs therefore requires methods that can decompose their representations into discrete, interpretable features and that can control for age when assigning clinical meaning [8, 13].

Existing interpretation methods neither decompose FM representations into discrete features nor control for confounds. Post-hoc attribution methods such as GradCAM [37] and integrated gradients [40] produce spatial saliency maps that highlight which input voxels influence a prediction, but they do not identify what high-level concepts the model has learned to represent internally. Linear probing [3] and concept activation vectors [25] test whether specific clinical variables are linearly decodable from intermediate representations, but they treat the latent space as a monolithic vector and cannot determine which dimensions encode which information. Disentanglement-based methods [18, 23] and concept-based approaches [15] can isolate individual factors of variation, but they require predefined labels during training and architectural modifications to the model itself. Brain MRI FMs are pretrained at scale via self-supervision, so a post-hoc method that can decompose their learned representations into discrete, identifiable features without modifying the model is essential.

Sparse autoencoders (SAEs) provide such a method. An SAE trains a sparse, overcomplete dictionary on a model's internal activations so that each representation is reconstructed from a small set of active dictionary elements, where each element ideally corresponds to a single interpretable concept [6, 9]. Researchers have used SAEs to reveal monosemantic features in large language models [14, 42], to identify interpretable directions in vision transformers [31, 32], and to discover biologically meaningful structure in protein language models [16, 38]. Recent extensions to medical imaging FMs have shown promise in pathology [26], radiology [2], and hematology [10]. However, all of these applications operate on models where feature mortality is modest and domain-specific confounds do not arise. Applying SAEs to neuroimaging FMs introduces three distinct challenges: *(a)* standard SAEs suffer severe feature mortality in deep transformer layers, losing >98% of dictionary elements be-

†Corresponding author.

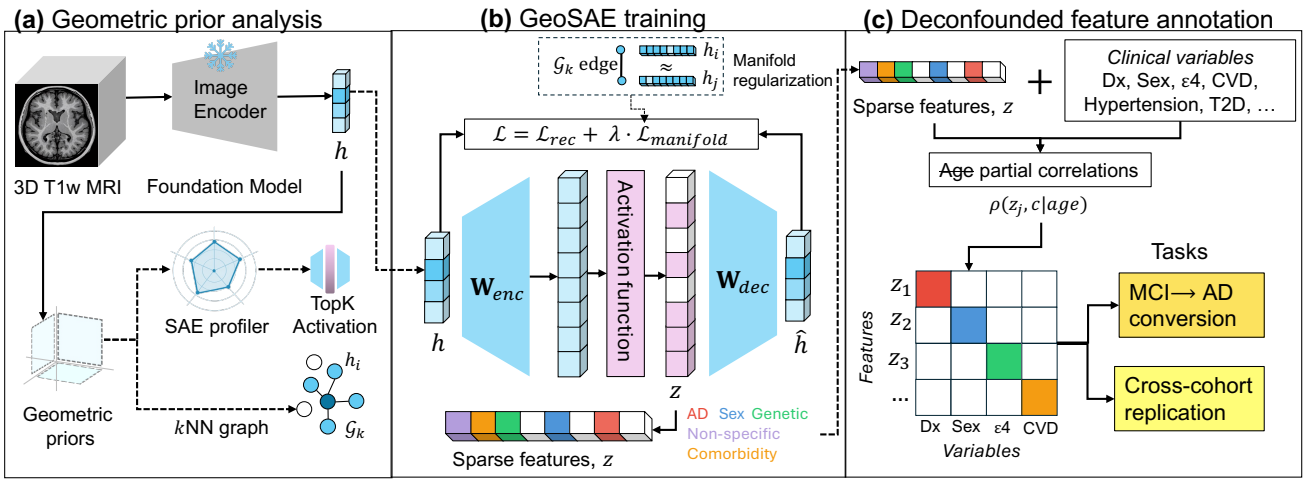

Figure 1. **Overview of GeoSAE.** (a) Geometric prior analysis of a frozen brain MRI FM selects the SAE activation function and constructs a k-NN manifold graph. (b) GeoSAE training uses manifold regularization to prevent feature collapse. (c) Age-deconfounded feature annotation assigns each alive feature to a clinical category for downstream tasks.

cause no geometric prior guides regularization; *(b)* age confounds nearly every clinical variable, so raw feature-outcome correlations conflate neurodegeneration with normal aging; and *(c)* comorbidities such as cardiovascular disease and diabetes share risk factors with AD, so feature annotations must disentangle disease-specific signal from comorbidities.

Our approach. We address all three challenges with **GeoSAE** (**Geo**metry-guided **S**parse **A**uto**E**ncoder), a framework that we apply to a 3D brain MRI ViT [41] (Fig. 1). **GeoSAE** first analyzes the geometry of each transformer layer's representations to select an appropriate SAE activation function and to construct a k-NN manifold graph. To address feature mortality *(a)*, we introduce a manifold regularization term that penalizes dissimilar pre-activations for samples that are neighbors on the FM's learned manifold. This regularization supplies gradients to all dictionary elements, including those zeroed out by the sparsity gate, and thereby prevents feature collapse (Sec. 3.2). To address age confounding *(b)* and comorbidity confounding *(c)*, we annotate each surviving feature via age-deconfounded partial correlations with hierarchical FDR correction and validate the annotations against 40+ secondary clinical variables.

We evaluate **GeoSAE** on approximately 14,500 T1-weighted MRI scans from ADNI [20] and AIBL [12]. At the optimal layer (layer 9), **GeoSAE** produces 161 alive features, 7× more than a standard SAE. The top 16 of these features achieve 100% clinical annotation and predict MCI-to-AD conversion with AUC 0.749, outperforming the full 768-dimensional CLS representation (0.732) while providing clinically interpretable features that replicate across cohorts (r=0.97). Cross-cohort activation consistency reaches r=0.97, a 32% improvement over standard SAEs (r=0.74). Attention-based saliency maps localize each feature to brain regions consistent with Braak staging [43].

Our contributions are threefold: (1) We propose **GeoSAE**, which uses geometric analysis of FM representations to guide SAE activation function selection and introduces a manifold regularization on pre-activations that prevents feature collapse. (2) We design an age-deconfounded annotation pipeline using partial correlations with hierarchical FDR correction to assign each surviving feature to a clinical category. (3) We apply **GeoSAE** across all 12 transformer layers of a brain MRI FM and show that 16 interpretable features at the peak layer predict AD conversion better than the full representation, replicate across independent cohorts, and localize to neuroanatomically plausible regions.*

2. Related Works

Interpreting neural network representations. The problem of understanding what a neural network has learned has motivated a broad family of methods. Gradient-weighted class activation mapping (GradCAM) [37] computes class-discriminative saliency by weighting feature maps with gradients, while integrated gradients [40] accumulate gradients along a path from a baseline input. SHAP [29] unifies several attribution approaches under a game-theoretic framework. All of these methods produce per-voxel importance scores and are effective for localizing predictions, but they do not decompose the model's internal representation into discrete, nameable concepts. Linear probing [3] addresses this by training linear classifiers on intermediate representations to test whether specific variables are encoded, and concept activation vectors (TCAVs) [25] quantify a model's sensitivity to user-defined concept directions. Both approaches,

*Code: https://github.com/favour-nerrise/GeoSAE

however, treat the latent space as a single vector and cannot identify which dimensions or subspaces encode which information. Disentanglement methods such as β-VAE [18] and supervised disentanglement [23, 45] can separate individual factors of variation, but they require labels during training and architectural modifications. Concept bottleneck models [15] similarly require concept supervision. None of these methods can be applied post-hoc to a frozen, self-supervised FM. Our work uses SAEs, which require no labels and no model modification, to decompose frozen FM representations into interpretable features.

Sparse autoencoders for mechanistic interpretability. SAEs were first applied to language models, where Bricken et al. [6] demonstrated that a one-layer MLP in a small transformer could be decomposed into monosemantic dictionary elements. Templeton et al. [42] scaled this approach to Claude 3 Sonnet and showed that SAE features correspond to human-interpretable concepts at the level of individual neurons. Gao et al. [14] systematically evaluated SAE scaling laws and training recipes. In vision, Olson et al. [31] applied SAEs to CLIP and DINOv2 and found that features align with object parts and textures, Pach et al. [32] showed that SAE features in vision-language models are monosemantic, and Stevens et al. [39] proposed evaluation criteria for scientifically rigorous SAE interpretation. Lim et al. [27] introduced PatchSAE, which extracts interpretable concepts with patch-wise spatial attributions from CLIP and reveals how adaptation techniques remap visual concepts. In the biomedical domain, Le et al. [26] used SAEs on a pathology FM and annotated features via enrichment testing against gene expression programs. Abdulaal et al. [2] applied SAEs to a radiology FM for report generation, and Dasdelen et al. [10] decomposed a hematology FM into cell-type-specific features. Simon and Zou [38] applied SAEs to protein language models and annotated features against Gene Ontology terms. A common finding across these studies is that SAE features are more interpretable and more specific than individual neurons. However, none of these works address feature mortality in deep layers or control for confounds during annotation. We address both with geometric regularization and age-deconfounded partial correlations.

Brain MRI foundation models. Self-supervised pretraining on large-scale structural MRI has produced FMs with strong transfer to clinical tasks. BrainIAC [41] trains a ViT on over 85,000 brain MRI scans and achieves state-of-the-art performance on age prediction and disease classification. Barbano et al. [4] pretrain anatomical FMs that capture regional brain structure, Kaczmarek et al. [24] build a SimCLR-based FM for neurological disease diagnosis from 3D MRI, and Rui et al. [36] propose multi-view pre-training that jointly learns from multiple brain imaging modalities. Wald et al. [44] revisit MAE pre-training for 3D medical image segmentation, demonstrating the effectiveness of self-supervised ViTs for volumetric data. These models learn rich representations of brain anatomy, but prior work has shown that neuroimaging embeddings frequently encode scanner manufacturer and acquisition site alongside clinical signal [28], that aging and neurodegeneration produce overlapping structural changes [34], and that comorbidities such as cardiovascular disease and diabetes further confound disease-specific signatures [8, 13]. No prior work has attempted to decompose a brain MRI FM's representations into interpretable features while controlling for these confounds. Our work is the first to apply SAEs to brain MRI FMs and to use the FM's own geometric structure to regularize SAE training.

3. Methods

GeoSAE is a three-stage pipeline (Fig. 1): geometric prior analysis selects an activation function and constructs a manifold graph (Sec. 3.1); a manifold-regularized SAE uses this graph to prevent feature collapse (Sec. 3.2); and age-deconfounded annotation assigns each surviving feature to a clinical category (Sec. 3.3).

3.1. Sparse Autoencoders for Brain MRI Foundation Models

Background. Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{c}_i)\}_{i=1}^{N}$ denote N T1-weighted brain MRI scans with clinical covariates $\mathbf{c}_i = (\text{age}, \text{sex}, \text{APOE4}, \text{diagnosis}, \mathbf{m})$, where $\mathbf{m} \in \{0,1\}^M$ are M comorbidity indicators. A frozen, self-supervised FM produces per-layer [CLS] token representations $\mathbf{h}_i^{(\ell)} = f_{\text{FM}}^{(\ell)}(\mathbf{x}_i) \in \mathbb{R}^d$ for layers $\ell = 1, \ldots, L$. A sparse autoencoder (SAE) [9] maps each dense $\mathbf{h} \in \mathbb{R}^d$ to a sparse code $\mathbf{z} \in \mathbb{R}^{d_{\text{SAE}}}$ ($d_{\text{SAE}} = d \times E$, expansion factor E). The encoder computes pre-activations $\mathbf{a} = \mathbf{W}_{\text{enc}}(\mathbf{h} - \mathbf{b}_{\text{pre}}) + \mathbf{b}_{\text{enc}}$ and applies a sparsity-inducing activation $\mathbf{z} = \sigma(\mathbf{a})$. The decoder reconstructs $\hat{\mathbf{h}} = \mathbf{W}_{\text{dec}}\,\mathbf{z} + \mathbf{b}_{\text{pre}}$ with unit-norm columns, minimizing $\mathcal{L}_{\text{SAE}} = \|\mathbf{h} - \hat{\mathbf{h}}\|_2^2$. Each active dimension of $\mathbf{z}$ ideally corresponds to one interpretable concept.

Feature mortality in deep layers. SAEs have been successfully applied to language models [14, 42] and vision transformers [31], but brain MRI FMs pose a distinct challenge. A feature j is *alive* if $\exists\, i : z_{ij} > 0$ and *dead* otherwise. As transformer representations become increasingly structured with depth, most dictionary elements receive zero gradient through the sparsity gate and are never updated. Dead neuron resampling [6] partially addresses this but destabilizes training for larger dictionaries. We propose instead to use the FM's own geometric structure to regularize SAE training.

Geometric prior analysis. We select the SAE activation function by evaluating four candidates (ReLU [6], JumpReLU [35], TopK [30], SpaDE [19]) against five geometric properties of each layer's representations, including angular vs. radial class structure, dimensionality homogeneity, and sparsity uniformity. For BrainIAC, all layers exhibit

angular-dominant structure (η^2=0.002) with significant negative activations; four of five properties favor TopK.

Manifold graph construction. Beyond activation selection, the geometric analysis yields a k-nearest-neighbor manifold graph $\mathcal{N}_k$ on the FM embeddings. We compute pairwise distances among all N representations $\{\mathbf{h}_i\}$ and connect samples to k nearest neighbors with Gaussian kernel weights:

$$w_{ij} = \exp\Big(- \frac{\|\mathbf{h}_i - \mathbf{h}_j\|^2}{2\sigma^2} \Big) \tag{1}$$

where σ is the median neighbor distance. Samples that the FM considers similar (nearby in representation space) receive high edge weights. The graph serves as the manifold prior for **GeoSAE** training and provides the neighborhood structure that guides regularization in Sec. 3.2.

3.2. Manifold-Regularized Sparse Autoencoder

TopK activation. In our dataset, geometric prior analysis recommends TopK [30] as the activation function across all 12 layers. We train $f_{\text{SAE}}^{(\ell)}$ independently for each layer ℓ. TopK activation retains only the k largest positive pre-activation entries:

$$z_j = \begin{cases} \text{ReLU}(a_j) & \text{if } j \in \text{top-}k(\mathbf{a}) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

TopK enforces exact sparsity by activating precisely k features per input. This simplifies comparison across layers and avoids sensitivity to the ℓ_1 penalty weight. Features outside the top-k set receive exactly zero gradient and remain dead.

Manifold regularization. GeoSAE addresses feature mortality by adding a manifold smoothness penalty on *pre*-activations $\mathbf{a}$ (before the sparsity gate) using the k-NN graph $\mathcal{N}_k$ from Sec. 3.1:

$$\mathcal{L} = \mathcal{L}_{\text{SAE}} + \lambda \sum_{(i,j)\in\mathcal{N}_k} w_{ij} \, \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 \tag{3}$$

This penalty operates on pre-activations, not post-activation codes. Every encoder-based SAE computes pre-activations before applying a sparsity gate, so the manifold term provides gradients to *all* dictionary elements, including those zeroed out by TopK. Samples that are neighbors on the FM's learned manifold are encouraged to have similar pre-activation patterns. This distributes gradient signal across the full dictionary and prevents collapse to a small active set.

Properties. The manifold term operates on pre-activations and is therefore activation-agnostic; it can be combined with any sparsity gate. Setting λ=0 recovers the standard SAE and provides a natural ablation baseline. Subsequent analyses are restricted to alive features ($\exists\, i : z_{ij} > 0$).

Comparison with dead neuron resampling. The standard remedy for feature mortality is dead neuron resampling [6], which periodically reinitializes dead dictionary elements. In our setting, resampling has two drawbacks. First, it causes training instability when combined with manifold regularization because resampled features are initialized far from the manifold-consistent landscape. Second, resampling alone produces features that converge to the dominant signal rather than diverse clinical categories. Manifold regularization avoids both issues because it provides continuous gradient flow to all dictionary elements. The k-NN graph is precomputed once ($\mathcal{O}(N^2 d)$); the per-batch manifold loss adds $\mathcal{O}(Bk\, d_{\text{SAE}})$ operations.

3.3. Deconfounded Feature Annotation

Age confounding. We annotate each alive feature by its association with $3 + M$ clinical variables: age, sex, APOE4, diagnosis, and M comorbidities. Prior SAE annotation methods [10, 26, 38] do not adjust for covariates. In neuroimaging, this confounds aging with neurodegeneration. Without correction, nearly every feature appears "AD-related" because both the feature and diagnosis correlate with age.

Partial correlations. We compute partial Spearman correlations controlling for age. For variable c and feature j, let r_{jc}, $r_{j,a}$, and $r_{c,a}$ denote Spearman correlations between the respective rank-transformed quantities and age a:

$$\rho_{jc\cdot a} = \frac{r_{jc} - r_{j,a}\, r_{c,a}}{\sqrt{1 - r_{j,a}^2}\,\sqrt{1 - r_{c,a}^2}} \tag{4}$$

This removes the variance shared with age from both the feature and the clinical variable before computing their association. Testing all feature×variable pairs simultaneously inflates false positives, so we apply false discovery rate (FDR) correction [5] to p-values from the partial-correlation matrix.

Category assignment. Let $\mathcal{C}$ denote the $3 + M$ non-age variables, grouped into categories (AD-related, sex-related, genetic, comorbidity). Each alive feature j is assigned to the category of the variable with the strongest significant partial correlation:

$$\text{cat}(j) = \arg\max_{c\in\mathcal{C}} \; |\rho_{jc\cdot a}| \quad \text{s.t. } \; p_{jc}^{\text{FDR}} < 0.05 \tag{5}$$

We label features with no significant partial correlation but significant raw age correlation ($p_{\text{age}}^{\text{FDR}} < 0.05$) as "aging" and the remainder as "non-specific." This ensures that each feature receives a single, interpretable clinical label while respecting the dominant confound in neuroimaging data.

4. Experiments

4.1. Dataset and Setup

Datasets. For training and primary evaluation, we use T1-weighted MRI from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [20], consisting of 13,218 scans across 1,974 participants (34.6% cognitively normal (CN), 53.1% MCI, 12.4% AD; mean age 74.6±7.4; 46.9% female).

We also use the Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) dataset [12], which includes 1,266 scans across 690 participants (67.9% CN (including 7.6% subjective memory complaints), 13.5% MCI, 11.0% AD; mean age 73.7±6.7; 53.1% female), for external validation. We preprocessed all volumes [1] with skull-stripping, bias-correction, registration to MNI space, and trilinear resampling to 96^3. Within ADNI, we derive MCI-to-AD conversion labels from longitudinal diagnostic records (926 MCI subjects: 343 converters, 583 stable); AIBL is used for feature annotation replication only, as its MCI converter count (n=11) is insufficient for conversion evaluation. Clinical covariates include M=5 comorbidity indicators in ADNI (hypertension, hyperlipidemia, depression, type 2 diabetes, cardiovascular disease) and M=4 in AIBL (hypertension unavailable). ADNI additionally provides 31 secondary variables spanning demographics, vitals, psychiatric history, cardiovascular indicators, metabolic markers, lifestyle factors, medications, and scanner metadata.

Evaluation tasks. We validate the annotated features through three tasks. *(1) Enrichment testing* (ADNI): we test each category against the 31 secondary variables using per-variable FDR correction. AD-related features should be enriched for AD-adjacent variables (e.g. cognitive scores, AD medications); we characterize features with no clinical association by activation frequency and image-quality statistics to assess whether they encode data-quality artifacts. *(2) Selective prediction* (ADNI): we evaluate category-restricted feature subsets for MCI-to-AD conversion prediction via logistic regression (5-fold stratified CV). The key test is that AD-related features should predict conversion comparably to all features, while comorbidity features should not. *(3) Cross-cohort replication* (AIBL): we apply the ADNI-trained SAE to AIBL without retraining and recompute all clinical correlations. We quantify replicability by annotation agreement (Pearson r between feature-level correlations across cohorts) and activation consistency (Spearman r between per-feature activation magnitudes).

Implementation. The FM is BrainIAC [41], a 12-layer ViT (L=12, d=768). For cross-layer analysis, we train L **GeoSAE** models (λ=0.1, k_{NN}=15, one per layer) with E=2, k=16 (d_{SAE}=1,536). At the peak layer (layer 9), we also train a standard SAE (λ=0) for comparison: it produces 23 alive features versus 161 for **GeoSAE** (7×). We train all SAEs for 100 epochs with Adam (lr 10^{-3}, batch 256) on a 90/10 subject-level split, achieving >99% explained variance. For external validation, we apply the same layer-9 **GeoSAE** to AIBL without retraining. For single-scan prediction, we use the latest available scan per subject to ensure one observation per subject. All conversion experiments use 5-fold stratified cross-validation with subject-level splits (all scans from a subject appear in the same fold) to prevent data leakage from longitudinal scans.

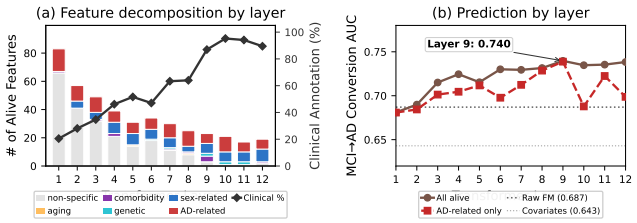

Figure 2. Cross-layer analysis of **GeoSAE** across 12 BrainIAC layers. (a) Stacked bars show alive features by clinical category; the line shows clinical annotation rate. Features consolidate with depth while clinical specificity increases. (b) Conversion AUC peaks at layer 9, then declines as scanner features dominate.

4.2. Cross-Layer Decomposition

Geometric prior results. Geometric prior analysis of the BrainIAC representations identifies angular-dominant structure (η^2=0.002), homogeneous dimensionality, uniform sparsity, and significant negative activations (48.8%); four of five geometric properties favor TopK, and the analysis rules out simplex-based methods. These properties are consistent across all 12 layers, so we use TopK as the activation function for all **GeoSAE** models.

Layer-wise trends. Training **GeoSAE** independently at each of the 12 transformer layers reveals two trends (Fig. 2). First, alive features peak in early layers (447 at layer 2) and consolidate with depth (147 at layer 10), a pattern that mirrors the FM's progressive abstraction from low-level texture to high-level semantic content. Second, conversion AUC increases with depth and peaks at layer 9 (0.731±0.011), then declines as scanner-related features dominate layers 10–12.

Clinical annotation across layers. Clinical annotation rates range from 14% to 29% across layers. At layer 9, the top-16 features by activation frequency achieve 100% annotation (Fig. 3): 5 AD-related, 6 sex-related, and 5 genetic features, with zero non-specific. Non-specific features dominate early layers (80% at layer 1), while AD, sex, and genetic features account for 87% at layer 9. This progressive specialization indicates that deeper layers encode increasingly clinical, rather than low-level, information.

4.3. Selective Prediction

Feature characterization. At layer 9, **GeoSAE** training (λ=0.1) produces 161 alive features (7× more than standard) with 17× lower inter-feature redundancy (mean pairwise $|r|$=0.009 vs. 0.156). The top-16 features by activation frequency achieve 100% clinical annotation (5 AD, 6 sex, 5 genetic; zero non-specific) and AUC 0.746.

Single-scan prediction. Table 1 presents single-scan MCI-to-AD conversion results. **GeoSAE** (16 features) achieves the highest AUC (0.746), surpassing both a supervised 3D ResNet-18 [17] baseline (MONAI [7]) trained from scratch on the same MRI volumes (0.729) and the raw 768-dim

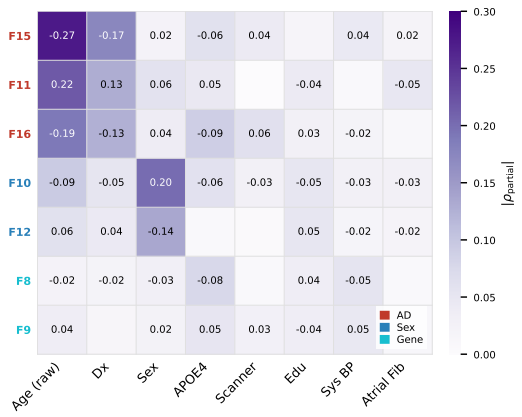

Figure 3. **GeoSAE** annotation (layer 9). $|\rho_{jc\cdot a}|$ for 3 strongest per category (AD, Sex, Genetic); values where FDR $p<0.05$.

Table 1. Single-scan MCI-to-AD conversion prediction (5-fold stratified CV, 926 MCI subjects: 343 converters, 583 stable). d: input dimensionality. Sens/Spec at threshold 0.5. **Bold**: best; underline: second best.

Model	d	AUC	Sens (%)	Spec (%)
Non-imaging (covariates)	3	.643 ± .003	24.1 ± 0.4	85.8 ± 0.5
3D ResNet-18 [17]	96[3]	.729 ± .035	**66.0** ± 20.0	61.4 ± 18.4
Raw FM [41]	768	.687 ± .009	54.0 ± 2.3	71.6 ± 0.7
PCA [22]	16	.734 ± .004	47.0 ± 1.1	82.9 ± 0.6
β-VAE [18]	16	.740 ± .002	48.8 ± 0.9	83.7 ± 0.6
ReLU SAE [6]	16	.728 ± .003	45.9 ± 1.3	81.9 ± 0.7
TopK SAE [30]	16	.737 ± .003	47.9 ± 0.8	83.9 ± 0.6
GeoSAE-rand	16	.574 ± .064	11.0 ± 12.8	**94.1** ± 6.1
GeoSAE (ours)	16	**.746** ± .002	49.2 ± 0.6	82.1 ± 0.7

FM embedding (0.687), while using only 16 dimensions (48× fewer than the raw embedding), the same reduction as PCA and β-VAE but with fully interpretable, cross-cohort-stable features (r=0.97 vs. 0.77; Sec. 4.6). We also include **GeoSAE**-rand, a control that samples 16 alive **GeoSAE** features uniformly at random rather than selecting the top-16 by activation frequency; its near-chance performance (0.574) shows that the gain does not come from arbitrary feature subsampling. PCA and β-VAE [18] at the same dimensionality (d=16) achieve 0.734 and 0.740 respectively, but the β-VAE shows substantially lower cross-cohort replicability (r=0.77 vs. 0.98; Sec. 4.6). The standard SAE (23 alive features, no manifold regularization) achieves 0.737; **GeoSAE** surpasses it with fewer features, which demonstrates the benefit of manifold-guided feature selection. All imaging methods outperform non-imaging covariates (0.643), which confirms that the FM encodes conversion-relevant structure beyond age, sex, and APOE4.

Table 2. Ablation study. Each row modifies one component of **GeoSAE** (E=2, k=16, λ=0.1). Alive: features with >0 activation. **Bold**/underline: best/second-best among ablation variants; category-only rows are negative controls, not competing methods.

Variant	Alive	d	AUC	Sens (%)	Spec (%)
GeoSAE (ours)	161	16	**.746**	**49.2**	82.1
Manifold regularization (λ)					
w/o manifold (λ=0)	23	16	.737	47.9	**83.9**
λ=1.0	121	16	.740	47.3	83.7
λ=10.0	14	16	.722	44.9	82.4
Architecture (E, k)					
E=2, k=8	156	8	.743	45.7	82.3
E=2, k=32	86	32	.729	48.5	81.1
E=4, k=16	383	16	.734	46.3	82.9
E=8, k=16	1404	16	.735	45.9	82.8
Category-only features					
AD-related only	7	7	.720	43.8	81.2
Comorbidity only	13	13	.507	0.4	99.7

4.4. Ablation Studies

Component ablations. Table 2 ablates GeoSAE's key components. Removing manifold regularization (λ=0) reduces alive features from 161 to 23 and annotation from 100% to 87%, while over-regularization (λ=10) degrades AUC. AUC is stable across $\lambda \in [0, 5]$ (ΔAUC < 0.002 except λ=0.5) and similarly robust to the manifold graph neighborhood size ($k_{\text{NN}} \in \{5, 10, 15, 20, 30\}$, ΔAUC < 0.006). Sweeping dictionary size and sparsity ($E \in \{2, 4, 8\}$, $k \in \{8, 16, 32\}$) confirms that E=2, k=16 is optimal; larger dictionaries fragment features and reduce AUC. Without age deconfounding, zero features are assigned to AD-related because $|\rho_{\text{age}}| > |\rho_{\text{diagnosis}}|$ for every feature.

Category ablation. Category-level feature removal tests whether annotations reflect genuine predictive contributions. Dropping AD-related features causes the largest AUC decrease (−0.017), while dropping comorbidity features has negligible effect (−0.003), confirming that AD features are both sufficient (0.720 alone) and *necessary* for conversion prediction. Dropping the 122 non-specific features *improves* AUC to 0.749, which indicates they add noise.

4.5. Brain Region Localization

Attention rollout. Clinical annotation identifies *what* each SAE feature encodes; we next ask *where* in the brain it attends. We compute attention rollout through the first 9 ViT layers, aggregate over the 50 maximally activating samples per feature, and map results to the Harvard-Oxford atlas [21].

Anatomical sub-patterns. The top-4 features ranked by individual MCI-to-AD conversion AUC localize to anatomically distinct sub-patterns (Fig. 4). Features F1 and F2 attend primarily to medial temporal lobe structures (hippocampus, parahippocampal gyrus), consistent with early Braak stages

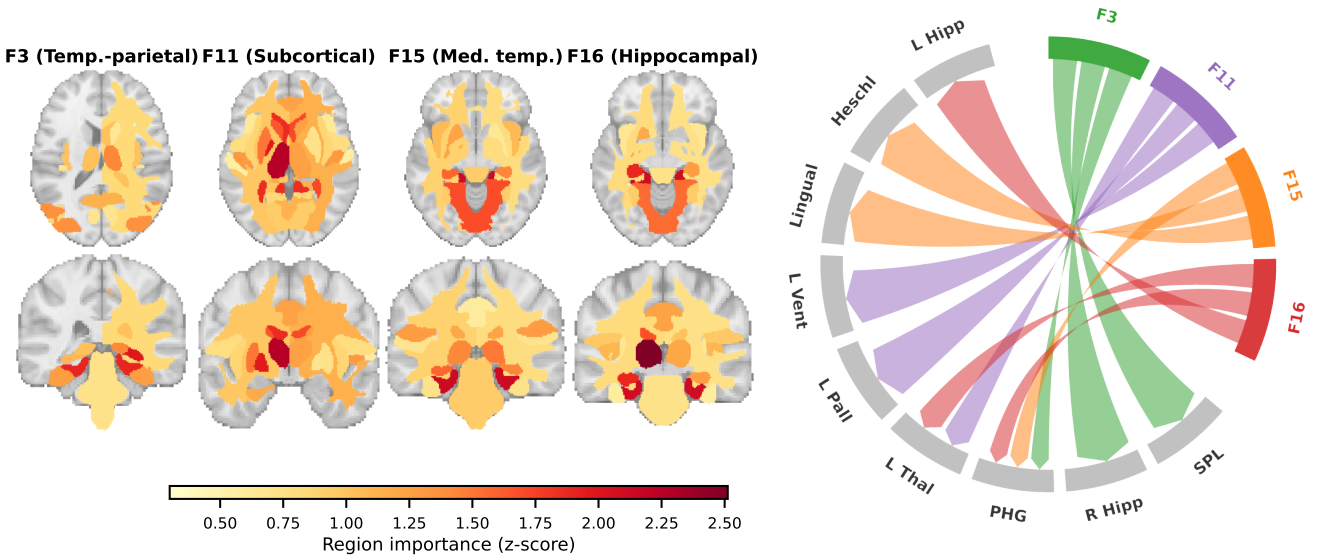

Figure 4. Brain region localization of the top-4 conversion-predictive **GeoSAE** features (F3, F11, F15, F16: top-4 SAE features ranked by individual MCI-to-AD conversion AUC) via attention rollout (layers 1–9) mapped to the Harvard-Oxford atlas. Left: axial (top) and coronal (bottom) slices; color intensity indicates z-scored region importance. Right: chord diagram showing top-3 regions per feature. PHG: parahippocampal gyrus; SPL: superior parietal lobule; L/R Hipp: hippocampus; L Thal/Pall/Vent: thalamus, pallidum, lateral ventricle.

where tau pathology first appears [43]. Feature F3 targets subcortical nuclei (thalamus, pallidum, ventricles), which reflect global atrophy and ventricular enlargement characteristic of advancing neurodegeneration [33]. Feature F4 activates a temporo-parietal network (superior parietal lobule, amygdala, fusiform cortex), consistent with later Braak stages where pathology spreads to association cortices [43].

Visualization. Fig. 4 shows axial and coronal slices with z-scored region importance overlaid on a template brain. The spatial patterns are visually distinct across features: F1–F2 highlight bilateral medial temporal regions, F3 shows a midline subcortical pattern, and F4 produces a distributed lateral pattern. The chord diagram (right panel) confirms that the top-3 regions per feature do not overlap, which indicates that **GeoSAE** decomposes the FM's conversion prediction into neuroanatomically separable components.

4.6. Cross-Cohort Replication and Enrichment

Annotation agreement. When we apply the ADNI-trained layer-9 **GeoSAE** to AIBL without retraining, it shows strong annotation agreement (Fig. 5): r=0.89 for age-partial diagnosis correlations and r=0.84 for age, with consistent category distributions (7 vs. 6 AD-related) despite AIBL's different demographics (75% CN vs. 35% in ADNI, higher cardiovascular disease and type 2 diabetes prevalence).

Activation consistency. The top-16 features achieve activation magnitude r=0.97 and diagnosis-specific pattern r=0.98 (CN) / 0.99 (AD) between ADNI and AIBL. The standard SAE achieves r=0.74 / 0.75 / 0.97 on the same metrics, a 32% lower cross-cohort reproducibility. All 16

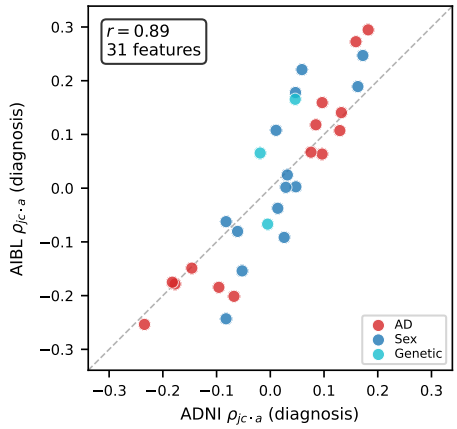

Figure 5. Cross-cohort replication of **GeoSAE** feature annotations. Each point is an SAE feature alive in both ADNI and AIBL, colored by clinical category. Age-partial diagnosis correlations replicate strongly (r=0.89, $p<10^{-10}$) despite different cohort variables.

GeoSAE features replicate in AIBL (100% replication rate). This strong transfer without retraining demonstrates that the manifold-regularized features capture stable biological signal rather than cohort-specific artifacts.

Enrichment testing. Enrichment testing against 31 secondary variables identifies AD medication use as the most prominently encoded secondary signal: memantine (26 of 161 features), donepezil (17), and the composite AD-medication indicator (20). This is consistent with known cholinesterase-inhibitor effects on brain structure [11], where

medicated patients show different atrophy patterns than unmedicated patients. Scanner manufacturer is associated with 11–29 features per layer, which confirms that acquisition confounds are encoded in the FM's representations. Scanner features concentrate in layers 10–12, which further supports the selection of layer 9 as optimal.

5. Conclusion

We present **GeoSAE**, a geometric prior-guided sparse autoencoder framework for interpreting brain MRI foundation models. **GeoSAE** addresses feature mortality in deep transformer layers through manifold regularization on pre-activations and produces 7× more alive features. A core set of 16 features with 100% clinical annotation predicts MCI-to-AD conversion (AUC 0.746), while comorbidity-annotated features achieve only chance-level performance (AUC 0.506); this contrast confirms that the framework separates disease from comorbidity signal. These features replicate across cohorts without retraining (r=0.97) and localize to neuroanatomically distinct sub-patterns aligned with Braak staging. We evaluate on a single FM (BrainIAC) and a single clinical domain (AD), so whether these results extend to other architectures and diseases remains to be established; the k-NN graph construction also scales quadratically, though in practice it is a one-time cost (<1 minute for 13k scans). More broadly, **GeoSAE** demonstrates that the geometric structure of foundation model representations can guide the extraction of clinically transparent, mechanistically interpretable features from black-box embeddings.

Acknowledgements

This work was supported in part by National Institutes of Health grants AG089169, AG084471, AG073356, DA057567, and AA021697, and the Stanford Institute for Human-Centered AI (HAI) Hoffman-Yee Award. F.N. was supported by the Stanford Graduate Fellowship, Stanford NeuroTech Graduate Training Program Fellowship, Stanford EDGE Fellowship, and Stanford Pathways to Neuroscience Fellowship.

References

[1] Mohammad Hassan Abbasi and Ehsan Adeli. sMRI processing pipeline: A lightweight, end-to-end workflow for structural brain MRI preprocessing and quality control, 2025. 5

[2] Ahmed Abdulaal et al. An x-ray is worth 15 features: Sparse autoencoders for interpretable radiology report generation. *arXiv:2410.03334*, 2024. 1, 3

[3] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644*, 2016. 1, 2

[4] Carlo Alberto Barbano et al. Anatomical foundation models for brain mris. *Pattern Recognition Letters*, 2025. 1, 3

[5] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995. 4

[6] Trenton Bricken et al. Towards monosemanticity: Decomposing language models with dictionary learning. Technical report, Anthropic AI Research, 2023. 1, 3, 4, 6

[7] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 5

[8] Davide Coluzzi, Valentina Bordin, Massimo W Rivolta, Igor Fortel, Liang Zhan, Alex Leow, and Giuseppe Baselli. Biomarker investigation using multiple brain measures from mri through explainable artificial intelligence in alzheimer's disease classification. *Bioengineering*, 12(1):82, 2025. 1, 3

[9] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv:2309.08600*, 2023. 1, 3

[10] Muhammed Furkan Dasdelen, Hyesu Lim, Michele Buck, Katharina S Götze, Carsten Marr, and Steffen Schneider. Cytosae: Interpretable cell embeddings for hematology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 77–86. Springer, 2025. 1, 3, 4

[11] Patricia Diaz-Galvan, Giulia Lorenzon, Rosaleena Mohanty, Gustav Mårtensson, Enrica Cavedo, Simone Lista, Andrea Vergallo, Kejal Kantarci, Harald Hampel, Bruno Dubois, et al. Differential response to donepezil in mri subtypes of mild cognitive impairment. *Alzheimer's Research & Therapy*, 15 (1):117, 2023. 7

[12] Kathryn A Ellis et al. The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. *International psychogeriatrics*, 21(4):672–687, 2009. 2, 5

[13] Louisa Fay, Erick Cobos, Bin Yang, Sergios Gatidis, and Thomas Küstner. Avoiding shortcut-learning by mutual information minimization in deep learning-based image processing. *IEEE Access*, 11:64070–64086, 2023. 1, 3

[14] Leo Gao et al. Scaling and evaluating sparse autoencoders. *arXiv:2406.04093*, 2024. 1, 3

[15] Yibo Gao, Hangqi Zhou, Zheyao Gao, Bomin Wang, Shangqi Gao, Sihan Wang, and Xiahai Zhuang. Learning concept-driven logical rules for interpretable and generalizable medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 291–300. Springer, 2025. 1, 3

[16] Edith Natalia Villegas Garcia and Alessio Ansuini. Interpreting and steering protein language models through sparse autoencoders. *arXiv:2502.09135*, 2025. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6

[18] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016. 1, 3, 6

[19] Sai Sumedh R Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. *arXiv:2503.01822*, 2025. 3

[20] Clifford R Jack Jr et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008. 2, 4

[21] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012. 6

[22] Ian T Joliffe and BJT Morgan. Principal component analysis and exploratory factor analysis. *Statistical methods in medical research*, 1(1):69–95, 1992. 6

[23] Wooseok Jung, Joonhyuk Park, and Won Hwa Kim. Disclose the neurodegeneration dynamics: Individualized ode discovery for alzheimer's disease precision medicine. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 177–186. Springer, 2025. 1, 3

[24] Emily Kaczmarek, Justin Szeto, Brennan Nichyporuk, and Tal Arbel. Building a general simclr self-supervised foundation model across neurological diseases to advance 3d brain mri diagnoses. In *IEEE ICCV*, pages 1310–1319, 2025. 1, 3

[25] Been Kim et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, pages 2668–2677. PMLR, 2018. 1, 2

[26] Nhat Minh Le et al. Learning biologically relevant features in a pathology foundation model using sparse autoencoders. *arXiv:2407.10785*, 2024. 1, 3, 4

[27] Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. *arXiv preprint arXiv:2412.05276*, 2024. 3

[28] Soroosh Safari Loaliyan and Greg Ver Steeg. Comparative analysis of generalization and harmonization methods for 3d brain fmri images: A case study on openbhb dataset. In *2024 IEEE CVPR workshop*, pages 4915–4923. IEEE, 2024. 3

[29] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 2

[30] Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv:1312.5663*, 2013. 3, 4, 6

[31] Matthew Lyle Olson, Musashi Hinck, Neale Ratzlaff, Changbai Li, Phillip Howard, Vasudev Lal, and Shao-Yen Tseng. Probing the representational power of sparse autoencoders in vision models. In *Proceedings of the IEEE ICCV*, pages 6167–6177, 2025. 1, 3

[32] Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv:2504.02821*, 2025. 1, 3

[33] Vincent Planche, José V Manjon, Boris Mansencal, Enrique Lanuza, Thomas Tourdias, Gwenaëlle Catheline, and Pierrick Coupé. Structural progression of alzheimer's disease over decades: the mri staging scheme. *Brain communications*, 4 (3):fcac109, 2022. 7

[34] Sebastian Pölsterl, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, and Japanese Alzheimer's Disease Neuroimaging Initiative. Identification of causal effects of neuroanatomy on cognitive decline requires modeling unobserved confounders. *Alzheimer's & Dementia*, 19(5):1994–2005, 2023. 1, 3

[35] Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv:2407.14435*, 2024. 3

[36] Shaohao Rui, Lingzhi Chen, Zhenyu Tang, Lilong Wang, Mianxin Liu, Shaoting Zhang, and Xiaosong Wang. Multimodal vision pre-training for medical image analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5164–5174, 2025. 3

[37] 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 *IEEE ICCV*, pages 618–626, 2017. 1, 2

[38] Elana Simon and James Zou. Interplm: discovering interpretable features in protein language models via sparse autoencoders. *Nature methods*, 22(10):2107–2117, 2025. 1, 3, 4

[39] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Interpretable and testable vision features via sparse autoencoders. *arXiv preprint arXiv:2502.06755*, 2025. 3

[40] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328. PMLR, 2017. 1, 2

[41] Divyanshu Tak et al. A generalizable foundation model for analysis of human brain mri. *Nature Neuroscience*, pages 1–12, 2026. 1, 2, 3, 5, 6

[42] Adly Templeton. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic, 2024. 1, 3

[43] Joseph Therriault, Tharick A Pascoal, Firoza Z Lussier, Cécile Tissot, Mira Chamoun, Gleb Bezgin, Stijn Servaes, Andrea L Benedet, Nicholas J Ashton, Thomas K Karikari, et al. Biomarker modeling of alzheimer's disease using pet-based braak staging. *Nature aging*, 2(6):526–535, 2022. 2, 7

[44] Tassilo Wald, Constantin Ulrich, Stanislav Lukyanenko, Andrei Goncharov, Alberto Paderno, Maximilian Miller, Leander Maerkisch, Paul Jaeger, and Klaus Maier-Hein. Revisiting mae pre-training for 3d medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5186–5196, 2025. 3

[45] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9677–9696, 2024. 3