CGCP: CLINICALLY-GUIDED CONTRASTIVE PRETRAINING WITH MIXED-TYPE CLINICAL SIGNALS FOR 3D BRAIN MRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Neuroimaging cohorts often pair brain MRI with structured clinical assessments, yet representation pretraining either ignores these signals or incorporates them without establishing whether they provide information beyond the downstream diagnostic target. We introduce **Clinically-Guided Contrastive Pretraining (CGCP)**, a representation-learning framework for 3D brain MRI that integrates heterogeneous clinical supervision through continuous cognitive-severity learning, discrete stage-aware contrastive learning with a label-tracking momentum memory bank, and variance regularisation. A Bayesian optimisation procedure balances these objectives while retaining a minimum contribution from clinically meaningful staging supervision. Beyond proposing a pretraining strategy, we use CGCP as an analytical probe to determine when mixed-type clinical supervision provides genuinely complementary information. We evaluate longitudinal-to-cross-sectional transfer against label-free self-supervised, supervised, and label-aware alternatives under subject-level evaluation, complemented by targeted loss ablations, frozen and nonlinear probing, layer-wise analysis, calibration, data-efficiency experiments, and sensitivity studies. The resulting evidence shows that clinically guided representations can support strong downstream classification, but their benefit is not uniform: correlated clinical instruments can act as label proxies, measurement ceilings can restrict continuous supervision, representation rankings can change substantially after fine-tuning, and memory-constrained optimisation can complicate multi-objective balancing. We consolidate these observations into a validity framework for evaluating clinically guided representation learning before improvements are attributed to additional clinical supervision.

1 INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia, affecting an estimated 57 million people worldwide in 2019, with numbers projected to reach 153 million by 2050 (Nichols et al., 2022). Structural MRI is a key non-invasive biomarker for AD, capturing patterns of cortical atrophy that correlate with disease progression (Jack Jr et al., 2018; Frisoni et al., 2010; Scahill et al., 2002). Large clinically characterised neuroimaging cohorts remain comparatively scarce, while acquisition, longitudinal follow-up, and expert clinical phenotyping are costly, limiting the amount of task-specific supervision available for representation learning.

Self-supervised learning (SSL) addresses this bottleneck through pretext tasks: contrastive methods such as SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and BYOL (Grill et al., 2020) enforce augmentation invariance, while masked approaches such as MAE (He et al., 2022) reconstruct missing regions. These methods transfer well in natural images (Oquab et al., 2024) and medical imaging (Azizi et al., 2021; Taleb et al., 2020; Zhou et al., 2021; VanBerlo et al., 2024), and augmenting contrastive SSL with a metadata signal such as age can further improve transfer (Dufumier et al., 2021).

A notable asymmetry nonetheless persists in neuroimaging: label-free SSL treats every scan as unlabelled, discarding the clinical metadata that longitudinal studies routinely collect. In OASIS-2 (Marcus et al., 2010), scans are accompanied by structured clinical assessments including MMSE,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reflecting cognitive severity on a continuous 0–30 scale (Folstein et al., 1975), and CDR, encoding clinical dementia status (Morris, 1993). These annotations encode exactly the pathological structure downstream tasks aim to recover, yet standard pretraining objectives are not designed to exploit them.

We propose **Clinically-Guided Contrastive Pretraining (CGCP)**, which retains the momentum encoder and memory-bank machinery of contrastive SSL but augments it with clinical supervision, occupying a middle ground between label-free SSL and fully supervised pretraining. CGCP jointly optimises MMSE regression (continuous severity), CDR memory-bank contrastive learning (discrete staging), and VICReg variance regularisation (Bardes et al., 2021) (anti-collapse), with loss weights chosen by constrained Bayesian optimisation (Akiba et al., 2019). Because such a design is exposed to label-proxy redundancy, heterogeneous measurement quality, optimisation imbalance, and small-sample evaluation effects, we use CGCP not only as a pretraining method but also as an analytical probe of when clinical guidance provides genuinely complementary supervision. We make four contributions:

- **C1. Mixed-type clinically guided representation learning.** We introduce CGCP, a 3D brain MRI pretraining framework that combines continuous cognitive-severity supervision, discrete stage-aware contrastive learning through a label-tracking momentum memory bank, and variance regularisation within a shared representation space (§3).
- **C2. Controlled comparison of clinical supervision strategies.** We compare CGCP against label-free self-supervised, supervised, and label-aware alternatives under a common 3D backbone and downstream protocol, and complement fine-tuning with frozen, nonlinear, and layer-wise probing (§4).
- **C3. Validity analysis of mixed clinical signals.** Through targeted ablations and psychometric analyses, we show that additional clinical supervision need not provide additional information: correlated clinical instruments can act as label proxies, while ceiling effects can restrict the usable continuous severity signal (§4.4, §5).
- **C4. Evaluation and optimisation diagnostics.** We identify substantial sensitivity to evaluation stage, readout complexity, and memory-constrained multi-objective optimisation, motivating four validity checks—label-proxy redundancy, instrument ceiling effects, evaluation-protocol sensitivity, and small-batch architectural constraints—for clinically guided representation learning (§4.2, §5).

We organize our findings around four categories of validity risk—*label-proxy redundancy*, *instrument ceiling effects*, *evaluation-protocol sensitivity*, and *small-batch architectural constraints*—established before results are presented (§4.2) and populated with direct evidence in §5.

2 Related Work

Label-free and medical SSL. SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) contrast augmented views via InfoNCE (Oord et al., 2018); BYOL (Grill et al., 2020) removes negatives through asymmetric prediction; Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2021) counter dimensional collapse; MAE (He et al., 2022) and SparK (Tian et al., 2023) reconstruct masked patches; and DINOv2 (Oquab et al., 2024) combines self-distillation with discriminative objectives at scale. In medical imaging, Azizi et al. (2021) use multi-instance contrastive learning, Taleb et al. (2020) extend five proxy tasks to 3D, and Tang et al. (2022) pretrain Swin UNETR on 5,050 CT scans. VanBerlo et al. (2024) note that integrating clinical knowledge with SSL remains underexplored, and FOMO60K (Cerri et al., 2025) highlights the need for large-scale neuroimaging SSL benchmarks. In all these works, clinical metadata is unused during pretraining.

Metadata- and label-guided contrastive learning. ConVIRT (Zhang et al., 2022) and CLIP (Radford et al., 2021) align images with text, but require paired reports unavailable in most neuroimaging cohorts. SupCon (Khosla et al., 2020) uses discrete labels for positive selection but does not accommodate continuous annotations. Most closely related, y-Aware InfoNCE (Dufumier et al., 2021) incorporates a single continuous proxy (age) into contrastive pretraining for brain MRI, and Rank-N-Contrast (Zha et al., 2023) learns order-preserving representations for continuous targets. CGCP extends this family along two axes: joint use of mixed-type (continuous MMSE + discrete CDR) annotations, and integration of variance regularisation with principled loss balancing.

Table 1: **Positioning of CGCP.** 3D brain MRI evaluation, continuous clinical signal, discrete clinical staging, momentum memory bank, and constrained BO for loss balancing.

Method	3D brain MRI	Continuous	Discrete	Memory bank	Constrained BO
SimCLR (Chen et al., 2020)	✗	✗	✗	✗	✗
MoCo v2 (He et al., 2020)	✗	✗	✗	✓	✗
BYOL / MAE / VICReg	✗	✗	✗	✗	✗
SupCon (Khosla et al., 2020)	✗	✗	✓	✗	✗
ConVIRT (Zhang et al., 2022)	✗	✗	✗	✗	✗
y-Aware (Dufumier et al., 2021)	✓	✓	✗	✗	✗
CGCP (ours)	✓	✓	✓	✓	✓

Multi-task loss balancing. GradNorm (Chen et al., 2018) equalises gradient magnitudes across tasks, uncertainty weighting (Kendall et al., 2018) uses homoscedastic uncertainty, gradient-conflict methods such as PCGrad (Yu et al., 2020) and CAGrad (Liu et al., 2021) project conflicting gradients, and multi-task learning has been cast as multi-objective optimisation (Sener & Koltun, 2018). Bayesian optimisation is standard for hyperparameter search (Snoek et al., 2012), but constrained BO for *clinical* loss-weight selection has, to our knowledge, not been studied. Table 1 positions CGCP.

3 METHOD

3.1 WHY JOINT DISCRETE–CONTINUOUS SUPERVISION?

Why the discrete objective should organize representations. Contrastive losses asymptotically optimize *alignment* of positive pairs and *uniformity* of the embedding distribution on the hypersphere, and both properties transfer to downstream performance (Wang & Isola, 2020). CDR-based contrastive learning targets alignment along a clinically meaningful axis (subjects sharing a diagnostic stage are pulled together), following Khosla et al. (2020), who show that label-defined positives improve alignment specifically for downstream classification.

Why the discrete objective alone should be insufficient. Zha et al. (2023) show that classification-style objectives on inherently continuous targets yield fragmented representations—clustered by category but failing to preserve continuous ordering within and across categories. A binary CN/demented contrastive pressure has no mechanism to preserve *where* within a category a subject falls, motivating MMSE regression as a complementary, order-preserving signal.

3.2 OVERVIEW

CGCP operates in two phases (Figure 1): (1) **pretraining** on OASIS-2 with three jointly optimised clinical losses and BO-selected weights, and (2) **fine-tuning** on OASIS-1 for AD vs. CN classification and MMSE regression. CGCP retains contrastive machinery associated with label-free SSL—momentum encoder, FIFO memory bank, InfoNCE-style loss—but is *not* self-supervised: both clinical objectives are built from annotations closely tied to the downstream task. We therefore compare against label-aware baselines using the same annotations (§4.1).

3.3 ENCODER AND MEMORY BANK

The backbone $f_\theta : \mathbb{R}^{160\times192\times160} \to \mathbb{R}^{512}$ is a MONAI (Cardoso et al., 2022) 3D ResNet-18 (layers $[2, 2, 2, 2]$, in-planes $[64, 128, 256, 512]$, global average pooling). A projector g_ϕ (Linear(512,256)–ReLU–Linear(256,512)) and an MMSE head h_ψ (Linear(512,128)–ReLU–Dropout(0.1)–Linear(128,1)) share this representation. A momentum encoder $(f_{\theta'}, g_{\phi'})$ is an EMA of the online encoder ($m = 0.999$), and a FIFO memory bank of size $K = 256$ stores ℓ_2-normalised momentum embeddings *together with their CDR labels*. Following MoCo (He et al., 2020) but with clinical label tracking, this enables clinically-grouped contrastive selection against a large pool despite the batch size $B = 2$ forced by 3D volumes. The bank is warmed up with a full pass before training (Algorithm 1, Appendix A); a sweep over K is in Appendix G.1.

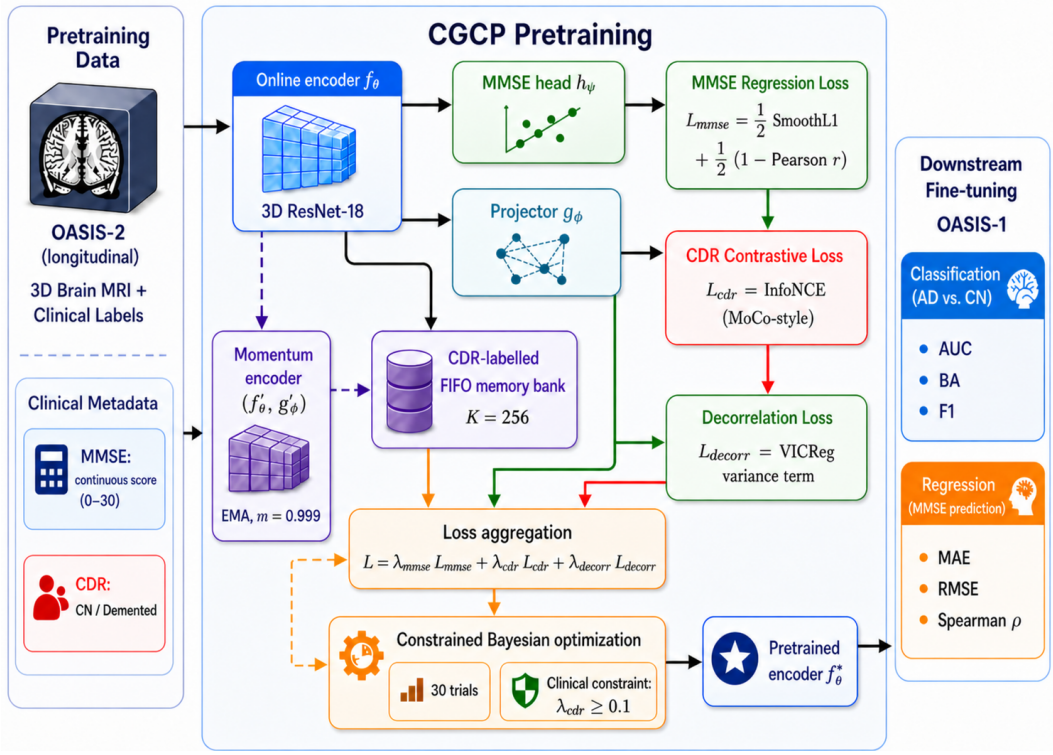


Figure 1: **CGCP pipeline.** *Left:* pretraining on OASIS-2 with a 3D ResNet-18 encoder and three clinical losses—MMSE regression ($\mathcal{L}_{\text{mmse}}$), CDR memory-bank contrastive learning ($\mathcal{L}_{\text{cdr}}$), and VICReg variance regularisation ($\mathcal{L}_{\text{decorr}}$); a momentum encoder maintains the memory bank. *Centre:* constrained BO ($\lambda_{\text{cdr}} \geq 0.1$, 30 trials) selects loss weights. *Right:* fine-tuning on OASIS-1 for classification and regression.

3.4 Clinical Loss Functions

CGCP optimises a convex combination of three losses:

$$\mathcal{L} = \lambda_{\text{mmse}}\mathcal{L}_{\text{mmse}} + \lambda_{\text{cdr}}\mathcal{L}_{\text{cdr}} + \lambda_{\text{decorr}}\mathcal{L}_{\text{decorr}}, \qquad \lambda_{\text{mmse}} + \lambda_{\text{cdr}} + \lambda_{\text{decorr}} = 1. \tag{1}$$

(i) MMSE regression ($\mathcal{L}_{\text{mmse}}$). MMSE in OASIS-2 is heavily right-skewed (most subjects score 26–30), and plain Smooth-L1 collapses to mean prediction. We add a Pearson term that penalises this collapse:

$$\mathcal{L}_{\text{mmse}} = \tfrac{1}{2}\,\text{SmoothL1}(\hat{y}, y) + \tfrac{1}{2}\,(1 - r_P(\hat{y}, y)), \tag{2}$$

computed over non-missing MMSE values (seven OASIS-2 scans lack MMSE; they still contribute to $\mathcal{L}_{\text{cdr}}$). With $B = 2$, $r_P \in \{-1, +1\}$ when predictions differ and is set to 0 when they coincide, so the term acts as a discrete anti-collapse signal rather than a correlation estimate.

(ii) CDR memory-bank contrastive loss ($\mathcal{L}_{\text{cdr}}$). CDR is binarised into CN (CDR$= 0$) and demented (CDR≥ 0.5), following standard practice for binary AD screening. For each query $\mathbf{z}_i^q$ from the online projector,

$$\mathcal{L}_{\text{cdr}} = -\frac{1}{|Q|}\sum_{i \in Q} \log \frac{\sum_{j \in P(i)} \exp(\mathbf{z}_i^q \cdot \mathbf{k}_j / \tau)}{\sum_{j \in V} \exp(\mathbf{z}_i^q \cdot \mathbf{k}_j / \tau)}, \tag{3}$$

where V is the set of filled bank entries, $P(i) \subset V$ those sharing query i's CDR group, $\mathbf{k}_j$ are normalised momentum embeddings, and $\tau = 0.07$. A stratified sampler guarantees at least one CN and one demented scan per batch while preserving the class ratio ($\sim$55/45).

(iii) Variance regularisation ($\mathcal{L}_{\text{decorr}}$). We adopt VICReg's variance term, $\mathcal{L}_{\text{decorr}} = \frac{1}{D}\sum_{d=1}^{D}\max(0, 1 - \text{std}_d(\mathbf{z}))$ with $D = 512$, rather than Barlow Twins' cross-correlation (Zbontar et al., 2021), whose $D \times D$ estimate is degenerate at $B = 2$; the per-dimension standard deviation remains well-defined and acts as a coarse anti-collapse pressure.

3.5 Bayesian Optimisation with a Clinical Constraint

Loss weights are selected with Optuna (Akiba et al., 2019) (TPE sampler, 30 trials). Each trial trains a fresh model for 5 epochs and scores frozen embeddings with a Ridge probe for MMSE (Spearman ρ_S) and a balanced logistic probe for CDR (AUC):

$$\text{proxy} = 0.5\,(-\rho_S) + 0.5\,(1 - \text{AUC}) \quad \text{(minimised; optimum } -0.5). \tag{4}$$

Raw weights $\text{raw}_{\text{mmse}}, \text{raw}_{\text{cdr}} \in [0.1, 1]$, $\text{raw}_{\text{decorr}} \in [0.01, 0.1]$ are sampled and normalised to sum to one.

Clinical constraint. We enforce $\lambda_{\text{cdr}} \geq 0.1$ after normalisation; violating trials receive penalty 1.0 without training. The constraint addresses *objective marginalisation under label-proxy redundancy*: because MMSE and CDR capture substantially overlapping aspects of cognitive status, optimisation of the continuous objective alone can also yield CDR-discriminative representations. The minimum-weight safeguard therefore ensures that the discrete staging objective remains explicitly represented during the search. The threshold is permissive by design; a sweep over $c \in \{0.05, 0.10, 0.15, 0.25\}$ is in Appendix G.2.

Outcome. Median pruning (from epoch 2) removed 20 of 30 trials ($\sim$67% compute savings); no trial violated the constraint. The best configuration has proxy -0.349 ($\rho_S = 0.765$, AUC$= 0.932$) with $\lambda_{\text{mmse}} = 0.298$, $\lambda_{\text{cdr}} = 0.651$, $\lambda_{\text{decorr}} = 0.052$—a CDR-dominant configuration more than $6\times$ the constraint floor. The best configuration has proxy -0.349 ($\rho_S = 0.765$, AUC$= 0.932$) with $\lambda_{\text{mmse}} = 0.298$, $\lambda_{\text{cdr}} = 0.651$, and $\lambda_{\text{decorr}} = 0.052$. Because no sampled trial violated the constraint and the selected λ_{cdr} lies substantially above its floor, the constraint functions as a permissive safeguard rather than an active determinant of the selected solution (landscape and full log in Appendix B).

3.6 Investigative Variants

We construct two variants, not as competing methods but as targeted interrogations of design choices; all other components are unchanged.

CGCP-GradNorm replaces the outer-loop BO search with GradNorm (Chen et al., 2018), which adapts weights during training by equalising gradient-normalised training rates. We choose GradNorm over uncertainty weighting (Kendall et al., 2018) and gradient-conflict methods (Yu et al., 2020; Liu et al., 2021) because it most directly solves the same problem as BO (choosing $\lambda_{\text{mmse}}, \lambda_{\text{cdr}}, \lambda_{\text{decorr}}$) and requires no additional per-task backward passes, which would be costly under the memory constraint.

CGCP-RNC tests the motivation in §3.1 by replacing $\mathcal{L}_{\text{mmse}}$ with a memory-bank adaptation of Rank-N-Contrast (Zha et al., 2023), which enforces monotonicity between embedding distance and label distance. RNC's comparison-based construction reuses the memory bank already maintained for $\mathcal{L}_{\text{cdr}}$, and in its original formulation, provably yields order-preserving embeddings.

4 Experiments

4.1 Datasets and Protocol

OASIS-2 (Marcus et al., 2010) (pretraining): 150 subjects aged 60–96, 373 sessions (2–5 visits each) with paired MMSE and CDR ($\sim$55% CN). **OASIS-1** (Marcus et al., 2007) (evaluation): of 416 subjects (ages 18–96), CDR was administered predominantly to older adults; after excluding subjects without complete CDR/MMSE or usable scans, 212 subjects remain, each with a single session of 3–6 T1-weighted runs (822 runs). Because each subject has one session, subject-level averaging never mixes disease-stage timepoints. OASIS-2 and OASIS-1 are distributed as separate cohorts with

Table 2: **Validity-risk framework.** Categories fixed before results are presented, with the evidence gathered for each.

Category	Guiding question	Evidence
1. Label-proxy redundancy	Is the pretraining signal independent of the downstream label, or does it overlap with how the label is defined?	MMSE–CDR correlation; MMSE-only predictive power (§5); MMSE ablation (§4.4); SupCon comparison (§4.3)
2. Instrument ceiling effects	Does the instrument's own precision bound achievable performance, independent of representation quality?	MMSE variance decomposition by CDR group (App. E.2); cross-method Spearman convergence; UMAP silhouettes (§4.6); residuals (App. G.3)
3. Evaluation-protocol sensitivity	Does the validation scheme, training stage, or readout complexity change conclusions independent of the model?	5-seed vs. 5-fold (App. F.5); frozen vs. fine-tuned (§4.5, App. F.1); nonlinear probes (App. F.2); layer-wise probing (App. F.3); data efficiency (App. F.6); variant significance (App. F.4)
4. Small-batch constraints	Do memory-driven batch limits (B=2) distort the multi-objective loss landscape?	Bank-size sweep (App. G.1); constraint sweep (App. G.2); CDR ablation trade-off (§4.4); GradNorm misallocation and variance diagnostic (App. G.4)

distinct subject identifiers. Volumes are skull-stripped, registered to MNI152 (1 mm), reoriented to RAS, cropped/padded to $160 \times 192 \times 160$, and z-scored (FSL, TorchIO (Pérez-García et al., 2021); Appendix C.1).

Evaluation. We use subject-level 5-fold cross-validation on OASIS-1: every subject is tested exactly once, checkpoint selection uses validation data, and the test fold is touched once. All metrics are computed at subject level (predictions averaged across runs). A single seed (42) is used; fold-to-fold variation replaces the seed variation of single-split protocols (a comparison with our earlier 5-seed single-split protocol is in Appendix F.5).

Compared configurations. All methods share the same 3D ResNet-18 and *identical* fine-tuning protocol, so differences are attributable to the pretraining objective alone. (1) *Supervised-only* (random init.). (2) *Label-free SSL*: MoCo v2 (He et al., 2020), BYOL (Grill et al., 2020), SimCLR (Chen et al., 2020), MAE (He et al., 2022), all pretrained on OASIS-2 without labels. (3) *Label-aware baselines with the same annotations as CGCP*: y-Aware InfoNCE (Dufumier et al., 2021) (continuous MMSE kernel), Multitask supervised (direct MMSE and CDR heads, no contrastive machinery), and SupCon (Khosla et al., 2020) (CDR only). (4) *CGCP* (BO) and its two variants. (5) *Ablations* removing each loss, each with its own 30-trial BO run.

Training. Pretraining: AdamW (Loshchilov & Hutter, 2017) (lr 10^{-4}, wd 10^{-3}), batch 2 with 8-step accumulation (effective 16), 20 epochs, gradient clipping at 1.0. Fine-tuning: AdamW (lr 10^{-5}, wd 10^{-5}), batch 2, 20 epochs. Heads, compute, and other details are in Appendices C.2–C.3.

4.2 Analytical Framework

Before presenting results, we fix four categories of validity risk that any clinically-guided pretraining study on a small imaging cohort should anticipate (Table 2); §5 synthesises the evidence per category.

4.3 Main Results

Table 3 reports downstream results for all eleven pretraining configurations. No method wins more than two of six metrics. CGCP (BO) attains the best BA (0.732) and F1 (0.695); CGCP-RNC attains the numerically highest AUC (0.784), not statistically distinguishable from CGCP-BO (0.780; paired t-test $p = 0.797$, Appendix F.4); CGCP-GradNorm attains the lowest MAE (2.193); Multitask supervised the lowest RMSE (3.348); and SupCon—which never observes MMSE—the highest Spearman r (0.446).

A Friedman test across methods does not reach significance at $\alpha = 0.05$ on any metric, which reflects limited power at $n = 5$ folds rather than demonstrated equivalence, and CGCP-BO is not statistically distinguishable from its strongest competitors. CGCP is thus a strong but not dominant classification method under rigorous evaluation, and no method shows a comparable advantage on regression.

Table 3: **Downstream results on OASIS-1** (mean ± std, subject-level 5-fold CV). **Best** per column. Same backbone and fine-tuning for all methods; only pretraining differs.

	AD vs. CN Classification			MMSE Regression		
Method	**AUC ↑**	**BA ↑**	**F1 ↑**	**MAE ↓**	**RMSE ↓**	**Spearman r ↑**
Supervised-only	0.737±.099	0.552±.043	0.302±.274	2.249±.269	3.394±.448	0.407±.127
Label-free SSL baselines						
MoCo v2 (He et al., 2020)	0.755±.059	0.649±.067	0.599±.080	2.248±.302	3.501±.480	0.418±.091
BYOL (Grill et al., 2020)	0.760±.047	0.671±.074	0.603±.083	2.298±.222	3.472±.413	0.366±.210
SimCLR (Chen et al., 2020)	0.771±.052	0.659±.063	0.579±.067	2.204±.254	3.389±.437	0.418±.106
MAE (He et al., 2022)	0.777±.062	0.674±.111	0.550±.210	2.260±.163	3.495±.358	0.433±.099
Label-aware baselines (same clinical annotations as CGCP)						
y-Aware InfoNCE (Dufumier et al., 2021)	0.761±.074	0.652±.079	0.546±.141	2.247±.234	3.412±.291	0.433±.099
Multitask supervised	0.758±.092	0.698±.091	0.638±.099	2.250±.299	**3.348±.484**	0.433±.110
SupCon (CDR) (Khosla et al., 2020)	0.750±.106	0.622±.146	0.538±.283	2.230±.231	3.414±.343	**0.446±.102**
CGCP variants						
CGCP-GradNorm	0.742±.072	0.689±.046	0.640±.046	**2.193±.242**	3.390±.396	0.425±.115
CGCP-RNC	**0.784±.076**[†]	0.698±.036	0.653±.060	2.214±.262	3.370±.394	0.431±.090
CGCP (ours, BO)	0.780±.077	**0.732±.071**	**0.695±.075**	2.242±.284	3.386±.453	0.444±.110

[†]Not significantly different from CGCP (BO) on AUC (paired t-test, $p = 0.797$, $n = 5$ folds).

Table 4: **Ablations on OASIS-1** (mean ± std, subject-level 5-fold CV). **Best** per column.

Config.	**Active losses**	**AUC ↑**	**BA ↑**	**F1 ↑**	**MAE ↓**	**RMSE ↓**	**Spearman r ↑**
Supervised	—	0.737±.099	0.552±.043	0.302±.274	2.249±.269	3.394±.448	0.407±.127
CGCP (full)	$\mathcal{L}_{\text{mmse}} + \mathcal{L}_{\text{cdr}} + \mathcal{L}_{\text{decorr}}$	**0.780±.077**	**0.732±.071**	**0.695±.075**	2.242±.284	3.386±.453	**0.444±.110**
w/o $\mathcal{L}_{\text{mmse}}$	$\mathcal{L}_{\text{cdr}} + \mathcal{L}_{\text{decorr}}$	0.745±.082	0.696±.103	0.627±.148	2.216±.242	3.428±.388	0.431±.103
w/o $\mathcal{L}_{\text{cdr}}$	$\mathcal{L}_{\text{mmse}} + \mathcal{L}_{\text{decorr}}$	0.759±.093	0.656±.133	0.536±.286	**2.189±.273**	**3.264±.379**	0.414±.115
w/o $\mathcal{L}_{\text{decorr}}$	$\mathcal{L}_{\text{mmse}} + \mathcal{L}_{\text{cdr}}$	0.777±.060	0.705±.057	0.647±.084	2.231±.237	3.383±.328	0.437±.101

4.4 Ablation Study

Table 4 removes one loss at a time (remaining weights re-optimised by a separate 30-trial BO run), and Table 5 reports fold-matched bootstrap CIs; paired t-test and Wilcoxon p-values are in Appendix D.

Across eighteen comparisons, only one reaches the Wilcoxon floor achievable at $n = 5$ ($p = 0.0625$): removing $\mathcal{L}_{\text{cdr}}$ improves RMSE in 5 of 5 folds with an entirely positive bootstrap CI, even though the t-test alone ($p = 0.102$) would not flag it. MAE moves in the same direction (2.242→2.189) but without support (3 of 5 folds, CI crossing zero). On Spearman r the same ablation shows the reverse pattern (full CGCP better, CI $[+.003, +.056]$), giving CDR removal a genuine *trade-off* signature. Two further CIs narrowly exclude zero without reaching significance on either test—AUC for $\mathcal{L}_{\text{mmse}}$ removal and F1 for $\mathcal{L}_{\text{decorr}}$ removal; the remaining comparisons show no reliable effect. Notably, removing $\mathcal{L}_{\text{mmse}}$ has no statistically distinguishable effect on any metric.

4.5 Frozen-Embedding Evaluation

Table 6 probes frozen 512-d embeddings with Ridge (MMSE) and logistic regression (CDR) under the same 5-fold protocol, isolating pretraining-stage quality from fine-tuning capacity. SimCLR attains the best frozen AUC, BA, and F1, while CGCP-GradNorm attains the best frozen MAE, RMSE, and Spearman r. This ordering does *not* persist after fine-tuning: SimCLR's fine-tuned performance is worse than its own frozen performance on four of six metrics (AUC, BA, F1, RMSE)—the most pronounced frozen-to-fine-tuned reversal of any method—while CGCP-GradNorm barely moves (frozen AUC 0.740 vs. fine-tuned 0.742). CGCP (BO) shows the opposite profile: a mid-pack frozen starting point with the largest fine-tuning gains on BA and F1 (Appendix F.1). Rankings also depend on readout complexity: under SVM, random-forest, and gradient-boosting probes, Multitask supervised—undistinguished elsewhere—is the consistent best classifier (Appendix F.2).

Table 5: **Bootstrap 95% CI of the mean fold-wise difference (full − ablation)**, $n = 5$ folds. Intervals excluding zero in **bold**. For MAE/RMSE, a positive difference means the ablation has *lower* error.

Ablation	AUC	BA	F1	MAE	RMSE	Spearman r
w/o $\mathcal{L}_{\text{mmse}}$	**[+.002,+.069]**	[−.030,+.104]	[−.028,+.190]	[−.045,+.102]	[−.161,+.077]	[−.009,+.034]
w/o $\mathcal{L}_{\text{cdr}}$	[−.018,+.060]	[−.031,+.193]	[−.034,+.463]	[−.036,+.148]	**[+.037,+.236]**	**[+.003,+.056]**
w/o $\mathcal{L}_{\text{decorr}}$	[−.022,+.020]	[−.013,+.051]	**[+.002,+.091]**	[−.079,+.097]	[−.147,+.132]	[−.027,+.041]

Table 6: **Frozen-embedding linear probes on OASIS-1** (mean ± std, subject-level 5-fold CV; never fine-tuned). **Best** per column.

	AD vs. CN Classification			MMSE Regression		
Method	**AUC ↑**	**BA ↑**	**F1 ↑**	**MAE ↓**	**RMSE ↓**	**Spearman r ↑**
Supervised-only	0.724±.041	0.614±.080	0.544±.112	3.537±.248	4.755±.609	0.236±.092
MoCo v2	0.747±.068	0.684±.073	0.633±.091	3.109±.535	5.338±1.960	0.329±.070
BYOL	0.717±.052	0.669±.065	0.623±.073	3.168±.805	5.756±.259	0.365±.110
SimCLR	**0.789±.070**	**0.699±.057**	**0.652±.063**	2.462±.310	3.291±.432	0.386±.122
MAE	0.714±.096	0.635±.098	0.568±.123	3.397±.269	4.389±.330	0.203±.079
y-Aware InfoNCE	0.714±.047	0.666±.032	0.616±.050	2.991±.407	4.037±.828	0.273±.091
Multitask supervised	0.714±.039	0.661±.036	0.611±.042	3.464±.266	4.695±.437	0.130±.211
SupCon (CDR)	0.724±.068	0.661±.066	0.617±.065	2.678±.353	3.686±.585	0.372±.118
w/o $\mathcal{L}_{\text{mmse}}$	0.721±.048	0.660±.054	0.615±.060	2.690±.390	3.802±.593	0.291±.144
w/o $\mathcal{L}_{\text{cdr}}$	0.720±.072	0.659±.083	0.603±.095	3.037±.325	4.204±.543	0.279±.196
w/o $\mathcal{L}_{\text{decorr}}$	0.691±.086	0.636±.072	0.580±.086	2.905±.372	3.919±.505	0.270±.175
CDR-only	0.658±.027	0.621±.030	0.562±.038	4.779±.662	10.120±3.698	0.161±.166
CGCP-GradNorm	0.740±.094	0.694±.084	0.652±.090	**2.398±.320**	**3.172±.455**	**0.408±.161**
CGCP-RNC	0.748±.077	0.694±.066	0.648±.077	2.706±.188	3.719±.338	0.391±.120
CGCP (ours)	0.736±.082	0.642±.092	0.586±.109	2.546±.490	3.478±.634	0.365±.144

4.6 Embedding Geometry

Figure 2 shows UMAP (McInnes et al., 2018) projections of frozen embeddings with silhouette scores (Rousseeuw, 1987) on CN/AD labels. CGCP (full) attains the highest silhouette of all fourteen objectives (0.142; Appendix E.1), but all values are far below what would indicate clean separation (> 0.5), and CN/AD overlap persists everywhere. Visual clustering and diagnostic separation are not the same property: w/o $\mathcal{L}_{\text{mmse}}$ shows the most visually distinct clusters yet the lowest silhouette (0.012), because each cluster is a near-even CN/AD mixture. Similarly, CGCP-RNC has competitive frozen AUC (0.748) but low silhouette (0.036), consistent with RNC organising embeddings by severity rank rather than diagnosis (MMSE–CDR $r = -0.70$, not -1). Continuous-MMSE colouring shows no visible low-MMSE gradient for any method, a consequence of the ceiling effect (§5) rather than of any objective.

5 Discussion

CGCP's strong classification point estimates are not the full picture once all experiments are considered together. We organise the evidence around the four categories of §4.2 (extended version in Appendix I).

Label-proxy redundancy. A 5-fold cross-validated logistic regression using MMSE alone predicts CDR-based diagnosis at AUC$= 0.875$—higher than any imaging pipeline evaluated, including fine-tuned CGCP (0.780). MMSE correlates with CDR at $r = -0.70$ (Spearman) and with the binary label at $r = -0.63$ (point-biserial). Consistently, removing $\mathcal{L}_{\text{mmse}}$ has no statistically distinguishable effect on any metric (§4.4), and SupCon, which never sees MMSE, attains the best Spearman r. The

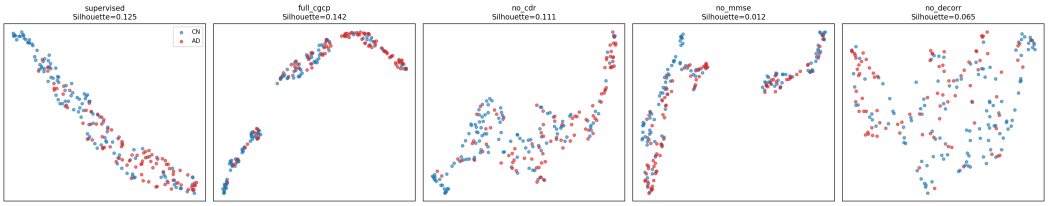


Figure 2: **UMAP of frozen pretrained embeddings** on OASIS-1 ($n = 212$, one point per subject). Blue: CN; red: AD. Left to right: supervised, CGCP (full), w/o $\mathcal{L}_{\text{cdr}}$, w/o $\mathcal{L}_{\text{mmse}}$, w/o $\mathcal{L}_{\text{decorr}}$; silhouette scores above each panel. All fourteen objectives and MMSE colouring in Appendix E.1.

two clinical signals substantially measure the same construct, and MMSE regression is closer to indirect regression toward the label than to an independent auxiliary signal.

Instrument ceiling effects. CN subjects cluster at the ceiling (MMSE 29.09 ± 1.11) while demented subjects spread widely (24.56 ± 4.08); CDR group explains only 39.8% of MMSE variance (Appendix E.2). Fine-tuned Spearman r converges to a narrow band (0.37–0.45) across every method in Table 3, and residuals are systematically negative for CN and positive for demented subjects *identically* with or without $\mathcal{L}_{\text{cdr}}$ or $\mathcal{L}_{\text{mmse}}$ (Appendix G.3)—a property of the target distribution, not of pretraining.

Evaluation-protocol sensitivity. Frozen and fine-tuned rankings disagree in both directions (§4.5); probe complexity changes the best frozen classifier (Appendix F.2); the three CGCP weighting/formulation variants are indistinguishable on seventeen of eighteen comparisons (Appendix F.4); the task-optimal layers coincide for only 7 of 14 methods (Appendix F.3); and moving from a 5-seed single split to 5-fold CV shrinks the $\mathcal{L}_{\text{decorr}}$ ablation effect on AUC by 96% and the $\mathcal{L}_{\text{cdr}}$ effect by 71%, while the $\mathcal{L}_{\text{mmse}}$ effect changes by only 5% (Appendix F.5).

Small-batch architectural constraints. The only statistically supported ablation effect—removing $\mathcal{L}_{\text{cdr}}$ improves RMSE but reduces Spearman r—reappears in CGCP-RNC's frozen embeddings via a mechanistically distinct comparison-based objective. A variance diagnostic rules out batch-of-2 sampling noise as the cause of GradNorm's over-weighting of $\mathcal{L}_{\text{decorr}}$ (coefficient of variation < 0.0004 at $n = 2$); instead, the loss sits within 0.001 of its structural maximum on ℓ_2-normalised 512-d embeddings, so a rate-based heuristic plausibly misreads a near-saturated term as a stalled one (Appendix G.4).

Synthesis and limitations. The first two categories are properties of this dataset pair and must be re-established, not assumed, elsewhere; the latter two are general lessons for memory-constrained multi-signal contrastive methods. CGCP's classification estimates are real, but they show what this configuration achieved on this dataset under this protocol, not that joint discrete–continuous supervision reliably outperforms simpler alternatives. Key limitations include limited inferential power at $n = 5$ folds, use of a single optimisation seed for the primary cross-validation experiments, possible participant overlap between the anonymised OASIS releases, a BO proxy computed on frozen embeddings despite frozen/fine-tuned divergence, untuned variant hyperparameters, a single backbone, binarised CDR, and two small cohorts; these issues are detailed in Appendix I.6.

6 Conclusion

We presented CGCP as both a framework for mixed-type clinically guided 3D brain MRI pretraining and a controlled probe of when such supervision is genuinely informative. Across label-free, supervised, label-aware, and clinically guided objectives, no single formulation dominates all downstream criteria, while targeted analyses reveal that the usefulness of additional clinical signals is governed by their redundancy, measurement range, evaluation protocol, and optimisation behaviour. In particular, clinical metadata should not be assumed to provide complementary supervision merely because multiple instruments are available. Our results therefore motivate a broader evaluation principle for clinically guided representation learning: the incremental information carried by each clinical signal, and the robustness of any resulting improvement to the evaluation protocol, should be established before attributing gains to multi-signal supervision.

AI USE STATEMENT

Large language models were used to assist with LaTeX formatting and editing for clarity. All experiments, analyses, and scientific claims are the authors' own; the authors verified all text and take full responsibility for the content.

ETHICS STATEMENT

This work uses the publicly available, de-identified OASIS-1 and OASIS-2 datasets under their data-use terms. The models are research artefacts and are not validated for clinical decision-making; as discussed in §5, their performance is bounded by properties of the cohort, and calibration is imperfect (Appendix H).

REPRODUCIBILITY STATEMENT

Architectures, losses, and the BO procedure are specified in §3 and Algorithm 1; the full BO log is in Appendix B; preprocessing, augmentation, fine-tuning heads, and compute are in Appendices C.1–C.3. All results use subject-level 5-fold CV with a fixed seed (42). Code and pretrained weights will be released upon acceptance.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3478–3488, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

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.

Stefano Cerri, Asbjørn Munk, Jakob Ambsdorf, Julia Machnio, Sebastian Nørgaard Llambias, Vardan Nersesjan, Christian Hedeager Krag, Peirong Liu, Pablo Rocamora García, Mostafa Mehdipour Ghazi, et al. A large-scale heterogeneous 3d magnetic resonance brain imaging dataset for self-supervised learning. *arXiv preprint arXiv:2506.14432*, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm Mcdonald, Camille Marie Piguet, et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 58–68. Springer, 2021.

Marshal F Folstein, Susan E Folstein, and Paul R McHugh. Mini-mental state. *Journal of psychiatric research*, 12(3):189–198, 1975.

Giovanni B Frisoni, Nick C Fox, Clifford R Jack Jr, Philip Scheltens, and Paul M Thompson. The clinical use of structural mri in alzheimer disease. *Nature reviews neurology*, 6(2):67–77, 2010.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: toward a biological definition of alzheimer's disease. *Alzheimer's & dementia*, 14(4):535–562, 2018.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12):2677–2684, 2010.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

John C Morris. The clinical dementia rating (cdr) current version and scoring rules. *Neurology*, 43 (11):2412–2412, 1993.

Emma Nichols, Jaimie D Steinmetz, Stein Emil Vollset, Kai Fukutaki, Julian Chalek, Foad Abd-Allah, Amir Abdoli, Ahmed Abualhasan, Eman Abu-Gharbieh, Tayyaba Tayyaba Akram, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. *The Lancet Public Health*, 7(2):e105–e125, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer methods and programs in biomedicine*, 208:106236, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Rachael I Scahill, Jonathan M Schott, John M Stevens, Martin N Rossor, and Nick C Fox. Mapping the evolution of regional atrophy in alzheimer's disease: unbiased analysis of fluid-registered serial mri. *Proceedings of the National Academy of Sciences*, 99(7):4703–4707, 2002.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in neural information processing systems*, 33:18158–18172, 2020.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20730–20740, 2022.

Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. *arXiv preprint arXiv:2301.03580*, 2023.

Blake VanBerlo, Jesse Hoey, and Alexander Wong. A survey of the impact of self-supervised pretraining for diagnostic tasks in medical x-ray, ct, mri, and ultrasound. *BMC medical imaging*, 24(1):79, 2024.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36:17882–17903, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

A CGCP PSEUDOCODE

Algorithm 1 gives the BO phase and the full pretraining phase, including memory-bank warm-up, stratified sampling, momentum updates, and gradient accumulation.

Algorithm 1 CGCP: Clinically-Guided Contrastive Pretraining

Require: OASIS-2 dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i^{\text{mmse}}, y_i^{\text{cdr}})\}$, BO budget T=30, epochs E=20, EMA momentum m=0.999, temperature τ=0.07, bank size K=256
Ensure: Pretrained encoder f_θ^*
— Phase 1: Bayesian Optimisation of Loss Weights —
for trial $t = 1, \ldots, T$ **do**
  Sample $\text{raw}_{\text{mmse}} \sim [0.1, 1.0]$, $\text{raw}_{\text{cdr}} \sim [0.1, 1.0]$, $\text{raw}_{\text{decorr}} \sim [0.01, 0.1]$ via TPE
  Normalise: $\lambda_k \leftarrow \text{raw}_k / \sum_j \text{raw}_j$
  if $\lambda_{\text{cdr}} < 0.1$ **then**
    Record $\text{proxy}_t \leftarrow 1.0$ {constraint violation; skip training}
    continue
  end if
  Initialise fresh encoder; train 5 epochs with $(\lambda_{\text{mmse}}, \lambda_{\text{cdr}}, \lambda_{\text{decorr}})$
  Extract frozen validation embeddings; fit Ridge $\rightarrow \rho_S$, logistic $\rightarrow$ AUC
  $\text{proxy}_t \leftarrow 0.5 \cdot (-\rho_S) + 0.5 \cdot (1 - \text{AUC})$
  MedianPruner: prune if proxy worse than median of completed trials (from epoch 2)
end for
$(\lambda_{\text{mmse}}^*, \lambda_{\text{cdr}}^*, \lambda_{\text{decorr}}^*) \leftarrow \arg\min_t \text{proxy}_t$
— Phase 2: Full Pretraining —
Initialise f_θ, $f_{\theta'}$, g_ϕ, $g_{\phi'}$, h_ψ; FIFO bank $\mathcal{M}$ (capacity K, labels -1)
Warm up $\mathcal{M}$: pass $\mathcal{D}$ through $f_{\theta'}, g_{\phi'}$ in shuffled order until full
for epoch $= 1, \ldots, E$ **do**
  for each batch $\mathcal{B} \sim$ `StratifiedCDRSampler`$(\mathcal{D})$ **do**
    $\mathbf{e} \leftarrow f_\theta(\mathbf{x})$; $\mathbf{z} \leftarrow \ell_2\text{-norm}(g_\phi(\mathbf{e}))$; $\hat{y} \leftarrow h_\psi(\mathbf{e})$
    $\mathbf{k} \leftarrow \ell_2\text{-norm}(g_{\phi'}(f_{\theta'}(\mathbf{x})))$ {no gradient}
    Compute $\mathcal{L}_{\text{mmse}}$ (Eq. 2, non-missing MMSE), $\mathcal{L}_{\text{cdr}}$ (Eq. 3), $\mathcal{L}_{\text{decorr}}$
    $\mathcal{L} \leftarrow \lambda_{\text{mmse}}^* \mathcal{L}_{\text{mmse}} + \lambda_{\text{cdr}}^* \mathcal{L}_{\text{cdr}} + \lambda_{\text{decorr}}^* \mathcal{L}_{\text{decorr}}$
    Accumulate gradients; AdamW step every 8 accumulation steps
    $\theta' \leftarrow m\theta' + (1-m)\theta$; $\phi' \leftarrow m\phi' + (1-m)\phi$
    Enqueue $\mathbf{k}$ with CDR labels into $\mathcal{M}$
  end for
  Evaluate on validation set; save checkpoint if best Spearman r
end for
return f_θ^*

B BAYESIAN OPTIMISATION: LANDSCAPE AND FULL TRIAL LOG

Figure 3 visualises the BO landscape and Table 7 the full 30-trial log. The best trial selects $\lambda_{\text{cdr}} = 0.651$, far above the 0.1 floor, and no trial violated the constraint, so the constraint acts as a safety bound rather than an active limit. Configurations with low λ_{cdr} (e.g., Trial 3, $\lambda_{\text{cdr}} = 0.124$) reach competitive AUC but weaker ρ_S. This preference for a strong CDR contribution is consistent with $\mathcal{L}_{\text{cdr}}$ being the only ablation with a statistically supported effect (§4.4).

Pruned-trial proxy values. For pruned trials, the proxy is the value at pruning time (typically epoch 2–3), not after 5 epochs; values such as Trial 13's -0.270 are snapshot estimates not directly comparable to completed trials. The MedianPruner compares against the median of completed trials at the same epoch (standard Optuna behaviour). Trial numbering is arbitrary under TPE.

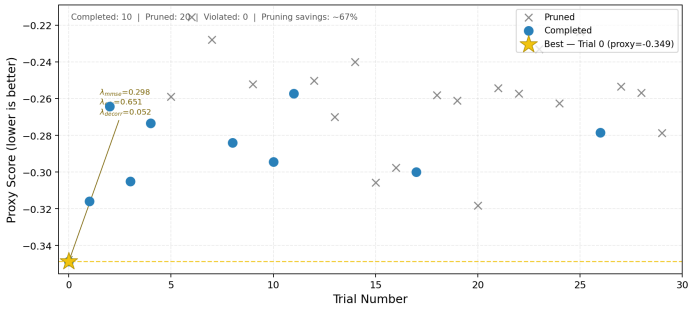


Figure 3: **BO landscape** (30 trials, TPE, MedianPruner). Blue: completed; grey ×: pruned; gold star: best trial (proxy -0.349, $\lambda_{\text{mmse}} = 0.298$, $\lambda_{\text{cdr}} = 0.651$, $\lambda_{\text{decorr}} = 0.052$).

Table 7: Complete BO trial log. **Best** in green; pruned in grey. Pruned-trial proxies are reported at pruning time.

T	State	λ_{mmse}	λ_{cdr}	λ_{decorr}	Proxy	ρ_S	AUC
0	**Complete**	**.298**	**.651**	**.052**	**−.349**	**.765**	**.932**
1	Complete	.707	.266	.027	−.316	.689	.943
2	Complete	.139	.803	.058	−.264	.638	.891
3	Complete	.774	.124	.102	−.305	.656	.954
4	Complete	.728	.250	.023	−.273	.640	.907
5	Pruned	.381	.537	.082	−.259	—	—
6	Pruned	.534	.395	.071	−.216	—	—
7	Pruned	.357	.575	.068	−.228	—	—
8	Complete	.380	.600	.021	−.284	.649	.919
9	Pruned	.465	.523	.012	−.252	—	—
10	Complete	.321	.605	.074	−.295	.633	.956
11	Complete	.388	.566	.045	−.257	.597	.917
12	Pruned	.640	.333	.027	−.250	—	—
13	Pruned	.779	.132	.089	−.270	—	—
14	Pruned	.409	.577	.014	−.240	—	—
15	Pruned	.575	.357	.068	−.306	—	—
16	Pruned	.444	.500	.057	−.298	—	—
17	Complete	.440	.540	.019	−.300	.652	.948
18	Pruned	.530	.441	.030	−.258	—	—
19	Pruned	.577	.295	.128	−.261	—	—
20	Pruned	.555	.425	.020	−.318	—	—
21	Pruned	.753	.147	.101	−.254	—	—
22	Pruned	.764	.156	.081	−.257	—	—
23	Pruned	.649	.264	.088	−.233	—	—
24	Pruned	.618	.327	.055	−.263	—	—
25	Pruned	.471	.464	.065	−.224	—	—
26	Complete	.553	.293	.154	−.278	.645	.912
27	Pruned	.751	.165	.084	−.253	—	—
28	Pruned	.659	.268	.073	−.257	—	—
29	Pruned	.330	.619	.051	−.279	—	—

C Implementation Details

C.1 Preprocessing and Augmentation

1. **Skull stripping** via FSL BET.
2. **Reorientation** to RAS via `fslreorient2std`.
3. **Affine registration** to MNI152 1 mm via FSL FLIRT.
4. **Spatial normalisation:** centre-crop or mean-pad to $160 \times 192 \times 160$ via TorchIO (Pérez-García et al., 2021).
5. **Intensity normalisation:** z-score at load time via TorchIO.

Training augmentations (pretraining and fine-tuning): left–right flip ($p = 0.5$); random affine with scale $\pm 5\%$, rotation $\pm 5°$, translation ± 3 mm ($p = 0.5$); Gaussian noise with std $\in [0, 0.03]$ ($p = 0.3$); random bias field with coefficient 0.2 ($p = 0.3$). No augmentation is applied at validation

or test time. The strategy is deliberately conservative: aggressive spatial augmentation can destroy the subtle atrophy patterns that distinguish AD from CN.

C.2 Downstream Fine-Tuning Heads

Classification. $f_\theta \to \text{Dropout}(0.1) \to \text{Linear}(512, 2)$, cross-entropy loss, all parameters trainable (AdamW, lr 10^{-5}, wd 10^{-5}, 20 epochs). AUC uses the softmax AD-class probability, averaged per subject.

Regression. $f_\theta \to \text{Dropout}(0.1) \to \text{Linear}(512, 128) \to \text{ReLU} \to \text{Linear}(128, 1)$, Smooth-L1 loss, same optimiser settings; predictions averaged per subject before MAE, RMSE, and Spearman r.

C.3 Compute

Experiments ran on dual NVIDIA RTX PRO 5000 GPUs (48 GB each). One epoch at batch 2 with 8-step accumulation takes ≈4 minutes (4.10 min/epoch, averaged over two measurements). One BO search (30 trials × 5 epochs, with pruning) takes ≈10 GPU-hours; full pretraining (20 epochs) ≈1.3 GPU-hours; 5-fold fine-tuning for both tasks ≈13.3 GPU-hours per objective. BO was run seven times (primary, three ablations, three constraint thresholds); the bank-size sweep reuses the primary weights. Twenty configurations required pretraining and twenty-one were fine-tuned under 5-fold CV. Total compute for BO, pretraining, and fine-tuning is ≈377 GPU-hours, excluding the cheaper frozen probing, UMAP, and data-efficiency analyses.

D Full Paired Significance Tests for Ablations

Table 8: **Paired tests, ablations vs. CGCP (full)** ($n = 5$ folds, fold-matched). **(a)** paired t-test; **(b)** Wilcoxon signed-rank (0.0625 is the minimum attainable at $n = 5$). Bootstrap CIs are in Table 5.

(a) Paired t-test p-values

Ablation	AUC	BA	F1	MAE	RMSE	Spearman r
w/o $\mathcal{L}_{\text{mmse}}$	0.147	0.422	0.345	0.590	0.576	0.364
w/o $\mathcal{L}_{\text{cdr}}$	0.393	0.296	0.353	0.346	0.102	0.124
w/o $\mathcal{L}_{\text{decorr}}$	0.795	0.242	0.122	0.842	0.976	0.736

(b) Wilcoxon signed-rank p-values

Ablation	AUC	BA	F1	MAE	RMSE	Spearman r
w/o $\mathcal{L}_{\text{mmse}}$	0.313	0.625	0.438	1.000	0.625	0.625
w/o $\mathcal{L}_{\text{cdr}}$	0.438	0.625	0.625	0.438	**0.0625**	0.188
w/o $\mathcal{L}_{\text{decorr}}$	0.625	0.375	0.125	1.000	1.000	0.813

E Instrument Ceiling Effects

E.1 Full UMAP Visualisations

Figures 4 and 5 extend Figure 2 to all fourteen objectives; Table 9 gives the complete silhouette ranking.

E.2 MMSE Variance Decomposition

CDR group explains a minority (39.8%) of MMSE variance, so the redundancy in §5 is not reducible to simple between-group separation; the remaining variance is concentrated in the demented group, while CN subjects are compressed against the ceiling.

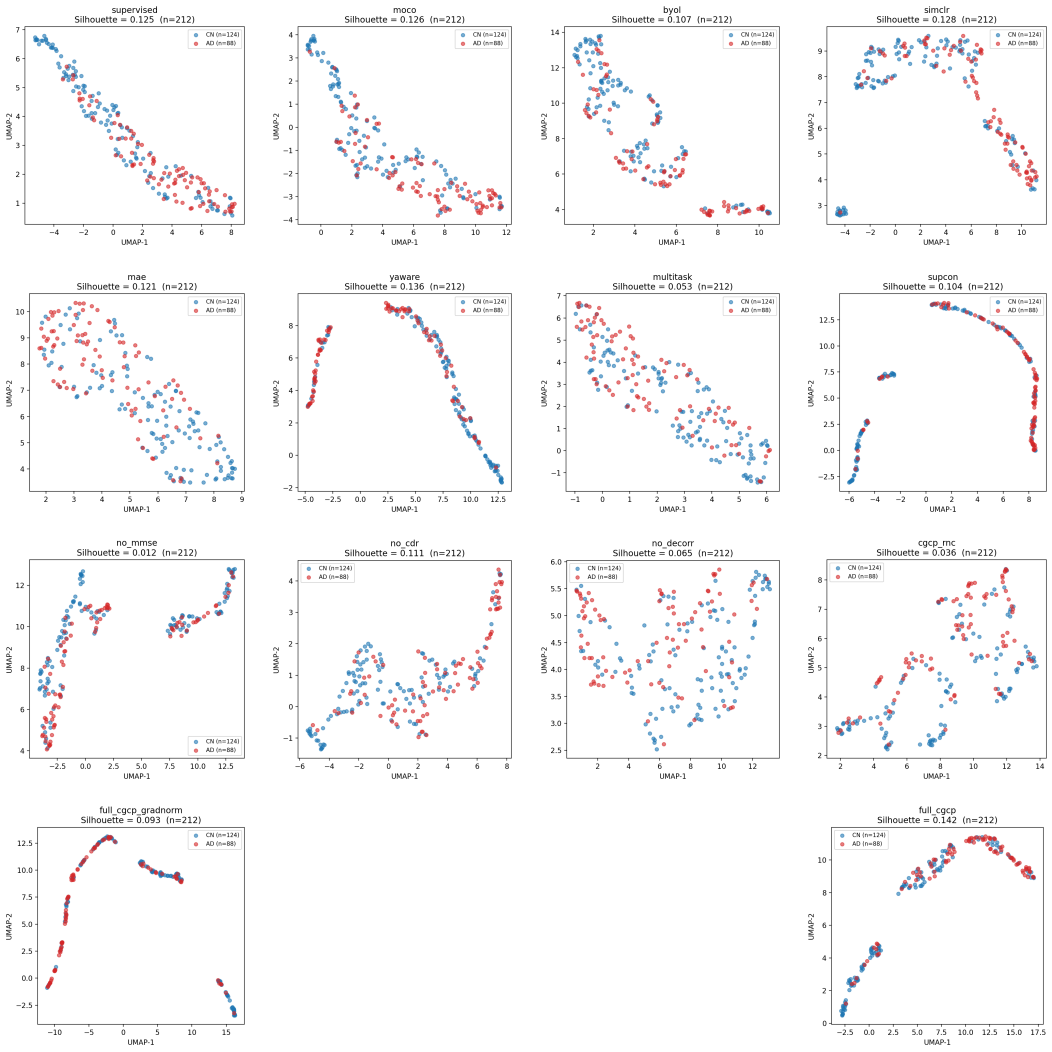


Figure 4: **UMAP, binary CN/AD colouring, all fourteen objectives** (frozen embeddings, $n = 212$). Rows: supervised and label-free SSL; label-aware baselines; ablations and CGCP-RNC; CGCP-GradNorm and CGCP (full). Silhouettes in panel titles.

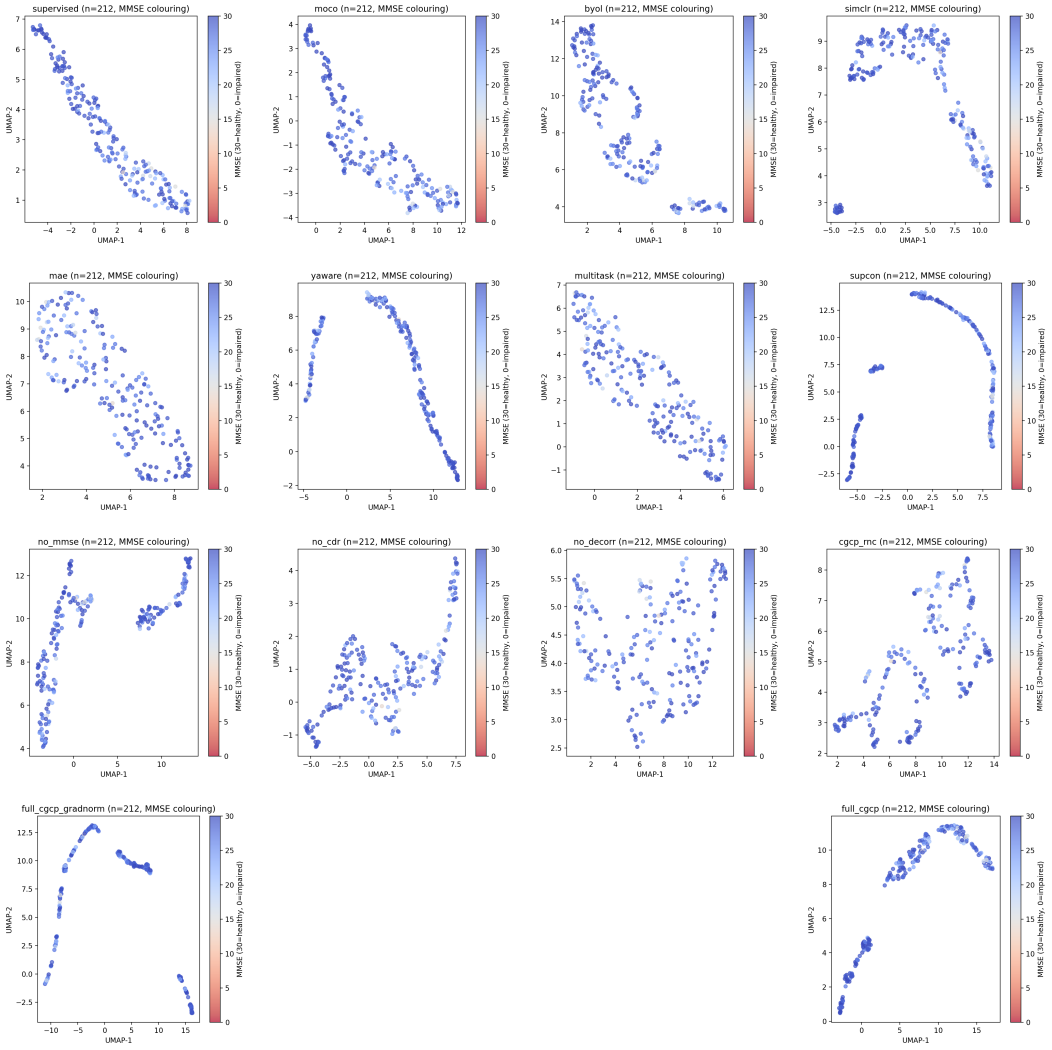


Figure 5: **UMAP, continuous MMSE colouring**, same layout as Figure 4. Blue = 30, red = 0. No embedding shows a gradient toward red, a consequence of the MMSE ceiling effect (Appendix E.2): even demented subjects mostly score above the scale midpoint (mean 24.56).

Table 9: **Silhouette scores, all objectives** (CN/AD labels, $n = 212$, frozen embeddings).

Method	Silh.	Method	Silh.
CGCP (full)	0.142	BYOL	0.107
y-Aware InfoNCE	0.136	SupCon	0.104
SimCLR	0.128	CGCP-GradNorm	0.093
MoCo v2	0.126	w/o $\mathcal{L}_{\text{decorr}}$	0.065
Supervised-only	0.125	Multitask supervised	0.053
MAE	0.121	CGCP-RNC	0.036
w/o $\mathcal{L}_{\text{cdr}}$	0.111	w/o $\mathcal{L}_{\text{mmse}}$	0.012

Table 10: **MMSE by CDR group** (OASIS-1, $n = 212$). Eta-squared (MMSE variance explained by CDR group) $= 0.398$.

CDR group	n	Mean	Std	Range
CN (CDR$= 0$)	124	29.09	1.11	[25, 30]
Demented (CDR≥ 0.5)	88	24.56	4.08	[14, 30]

F Evaluation-Protocol Sensitivity

F.1 Fine-Tuning Gain

Table 11: **Fine-tuning gain** (fine-tuned − frozen; for MAE/RMSE, frozen − fine-tuned, so positive = improvement). Computed from Tables 3 and 6.

Method	ΔAUC	ΔBA	ΔF1	ΔMAE	ΔRMSE	ΔSpearman
MoCo v2	$+0.008$	-0.035	-0.034	$+0.861$	$+1.837$	$+0.089$
BYOL	$+0.043$	$+0.002$	-0.020	$+0.870$	$+2.284$	$+0.001$
SimCLR	$\mathbf{-0.018}$	$\mathbf{-0.040}$	$\mathbf{-0.073}$	$+0.258$	$\mathbf{-0.098}$	$+0.032$
MAE	$+0.063$	$+0.039$	-0.018	$+1.137$	$+0.894$	$+0.230$
y-Aware	$+0.047$	-0.014	-0.070	$+0.744$	$+0.625$	$+0.160$
Multitask	$+0.044$	$+0.037$	$+0.027$	$+1.214$	$+1.347$	$+0.303$
CDR-only	$+0.107$	$+0.062$	$+0.050$	$+2.536$	$+6.663$	$+0.251$
SupCon	$+0.026$	-0.039	-0.079	$+0.448$	$+0.272$	$+0.074$
CGCP (full)	$+0.044$	$\mathbf{+0.090}$	$\mathbf{+0.109}$	$+0.304$	$+0.092$	$+0.079$

SimCLR is the only method with negative gain on four of six metrics. CGCP shows the largest gains on BA and F1 from a mid-pack frozen starting point. CDR-only shows the largest regression gains, consistent with its very poor frozen regression leaving the most room to recover.

F.2 Classical-ML Frozen Probe

The best method changes with probe type: SimCLR is best under linear probes on both tasks, whereas Multitask supervised is the best classifier under all three nonlinear probes. No method is best under every probe for either task.

F.3 Layer-Wise Probing

The classification- and regression-optimal layers coincide for exactly 7 of 14 methods, indistinguishable from chance. Layer 1 is the most common optimum for Spearman r (10 of 14), while AUC optima spread across layers 1–3 (Layer 4 wins AUC only for MoCo v2 and never wins Spearman r). A clean early/late-layer dissociation between tasks therefore does not hold as a general property of this architecture; where it appears for individual methods (e.g., CGCP peaks at Layer 3 for both tasks), it does not generalise.

Table 12: **Classical-ML frozen probes** (AUC and Spearman r; subject-level 5-fold CV) on nine objectives. **Best** per column.

	AUC				Spearman r			
Method	**LogReg**	**SVM**	**RF**	**GB**	**Ridge**	**SVR**	**RF**	**GB**
MoCo v2	0.756	0.750	0.740	0.699	0.349	**0.446**	**0.399**	**0.391**
BYOL	0.721	0.747	0.727	0.682	0.369	0.386	0.397	0.329
SimCLR	**0.778**	0.723	0.745	0.739	**0.397**	0.404	0.332	0.280
MAE	0.710	0.728	0.740	0.757	0.200	**0.446**	0.389	0.369
y-Aware	0.715	0.722	0.710	0.736	0.270	0.396	0.364	0.319
Multitask	0.721	**0.751**	**0.752**	**0.760**	0.138	0.359	0.288	0.244
CDR-only	0.666	0.733	0.713	0.708	0.172	**0.446**	0.384	0.329
SupCon	0.726	0.741	0.738	0.724	0.370	0.444	**0.399**	0.385
CGCP (full)	0.728	0.720	0.741	0.739	0.387	0.374	0.393	0.383

Table 13: **Layer-wise linear probing** of frozen intermediate ResNet-18 features (subject-level 5-fold CV). Best layer per method and task highlighted.

	AUC				Spearman r			
Method	**L1**	**L2**	**L3**	**L4**	**L1**	**L2**	**L3**	**L4**
Supervised	**0.741**	0.703	0.639	0.666	**0.356**	0.273	0.167	0.256
MoCo v2	0.716	0.726	0.730	**0.748**	**0.375**	0.353	0.277	0.306
BYOL	0.741	0.733	**0.743**	0.742	**0.397**	0.393	0.381	0.350
SimCLR	**0.728**	0.726	0.727	0.721	**0.412**	0.388	0.322	0.300
MAE	0.752	0.716	**0.753**	0.733	**0.403**	0.312	0.338	0.304
y-Aware	**0.742**	0.740	0.693	0.689	**0.405**	0.304	0.273	0.255
Multitask	0.727	**0.737**	0.732	0.681	**0.364**	0.254	0.323	0.243
CDR-only	**0.739**	0.725	0.685	0.641	**0.399**	0.327	0.258	0.298
SupCon	0.732	0.742	**0.750**	0.701	0.381	**0.425**	0.412	0.365
w/o $\mathcal{L}_{\text{mmse}}$	**0.723**	0.716	0.723	0.695	**0.365**	0.337	0.363	0.364
w/o $\mathcal{L}_{\text{cdr}}$	0.708	0.724	**0.750**	0.669	**0.351**	0.349	0.324	0.309
w/o $\mathcal{L}_{\text{decorr}}$	**0.744**	0.710	0.723	0.698	0.360	**0.364**	0.301	0.303
CGCP-RNC	0.730	**0.757**	0.732	0.682	0.354	**0.425**	0.385	0.310
CGCP (full)	0.732	0.717	**0.746**	0.743	0.356	0.327	**0.424**	0.419

F.4 Significance Among CGCP Variants

Of eighteen comparisons, only BO vs. GradNorm on AUC reaches $p < 0.05$. GradNorm vs. RNC shows AUC in the same direction (RNC higher) without reaching significance ($p = 0.157$).

F.5 5-Fold CV vs. Single-Split Evaluation

The three ablations respond very differently to the protocol change: the MMSE effect is nearly unchanged, the CDR effect shrinks substantially but remains the largest, and the decorrelation effect—second-largest under the original protocol—becomes statistically unconfirmed under 5-fold CV (bootstrap CI $[-0.022, +0.020]$).

F.6 Data-Efficiency Sweep

G Small-Batch Architectural Constraints

G.1 Memory-Bank Size Sweep

Classification is non-monotonic in K: AUC and BA dip sharply at $K = 128$ before recovering at $K = 256$, with $K = 64$ closer to the default than $K = 128$. $K = 256$ is best on four metrics; $K = 128$ attains the best MAE and RMSE. We have no mechanistic explanation for the dip and report it as an open pattern.

Table 14: **Paired t-test p-values among CGCP variants** ($n = 5$ folds). BO: Bayesian-optimised; GN: GradNorm; RNC: Rank-N-Contrast MMSE loss.

Comparison	AUC	BA	F1	MAE	RMSE	Spearman r
BO vs. GN	**0.030**	0.115	0.101	0.523	0.912	0.457
BO vs. RNC	0.797	0.360	0.407	0.572	0.744	0.677
GN vs. RNC	0.157	0.729	0.746	0.588	0.637	0.839

Table 15: **AUC ablation effects under the original 5-seed single-split protocol vs. subject-level 5-fold CV.** ΔAUC = full − ablation within each protocol.

	5-seed		5-fold CV		
Config.	**AUC**	**ΔAUC**	**AUC**	**ΔAUC**	**Shrinkage**
CGCP (full)	0.744	—	0.780	—	—
w/o $\mathcal{L}_{\text{mmse}}$	0.707	0.037	0.745	0.035	5%
w/o $\mathcal{L}_{\text{cdr}}$	0.671	0.073	0.759	0.021	71%
w/o $\mathcal{L}_{\text{decorr}}$	0.677	0.067	0.777	0.003	96%

G.2 CDR-Constraint Threshold Sweep

The default $c = 0.10$ is best on five of six metrics; only Spearman r is marginally higher at $c = 0.25$ (0.456 vs. 0.444).

G.3 Regression Residual Inspection

To test whether the CDR ablation trade-off reflects a specific bias in absolute-value predictions, we inspect fine-tuned residuals (predicted − true MMSE) pooled across folds ($n = 212$).

Residual patterns are nearly identical across configurations: small negative residuals for CN (mean ≈ -0.40) and large positive residuals for demented subjects (mean $\approx +3.1$). This does not support a CDR-specific mechanism for the trade-off; it is instead consistent with the target distribution's imbalance (Appendix E.2). We report it as a negative result for the hypothesis it was designed to test and as independent confirmation of the ceiling effect.

G.4 Decorrelation Variance Diagnostic

To test whether $B = 2$ makes $\mathcal{L}_{\text{decorr}}$'s signal unstable—a candidate explanation for GradNorm's over-weighting of this term—we compute the loss at sample sizes $n \in \{2, 8, 16, 32, 64, 128\}$ over 2000 resamples from cached frozen embeddings. Because the pool consists of final embeddings rather than mid-training snapshots, this measures estimation noise, not training dynamics.

The coefficient of variation at $n = 2$ is negligible, ruling out small-batch noise: averaging the per-dimension penalty over 512 dimensions cancels sampling noise even at the smallest batch. Both encoders' loss values sit within 0.001 of the loss's maximum of 1.0 (CGCP: 0.98941; w/o $\mathcal{L}_{\text{decorr}}$: 0.98998), a near-unavoidable consequence of per-dimension variance on ℓ_2-normalised 512-d embeddings. CGCP, which trains against this loss, reaches a slightly lower value, confirming the loss responds to training—but from a level so close to its maximum that a rate-based heuristic like GradNorm plausibly reads the near-flat trajectory as a struggling task.

H Calibration

We evaluate Expected Calibration Error (Guo et al., 2017) (ECE, 10 bins) on subject-level predictions pooled across folds ($n = 212$), with 2000-resample bootstrap CIs.

SimCLR and CDR-only are the best calibrated, with overlapping CIs. CGCP is substantially better calibrated than random initialisation (non-overlapping CIs) but not the best among those tested. In Figure 6, supervised-only oscillates between over- and under-confidence (e.g., predicted confidence ≈ 0.34 against an observed rate of 0.89, followed by ≈ 0.76 against 0.25). SimCLR and CDR-only

Table 16: **Data efficiency: classification** as a function of labelled fraction (subject-level 5-fold CV). 100% reproduces Table 3. **Best** per fraction and metric.

	10%			25%			50%			100%		
Method	**AUC**	**BA**	**F1**	**AUC**	**BA**	**F1**	**AUC**	**BA**	**F1**	**AUC**	**BA**	**F1**
Supervised	**0.737±.050**	0.512±.025	0.238±.292	**0.755±.051**	0.603±.070	0.491±.247	0.752±.051	0.548±.051	0.387±.276	0.737±.099	0.552±.043	0.302±.274
SimCLR	0.692±.058	0.618±.033	0.503±.087	0.730±.082	**0.660±.124**	**0.604±.153**	**0.755±.046**	**0.676±.092**	0.534±.268	0.771±.052	0.659±.063	0.579±.067
MAE	0.694±.050	0.584±.059	0.433±.229	0.743±.052	0.588±.083	0.311±.213	0.745±.066	0.654±.088	0.518±.268	0.777±.062	0.674±.111	0.550±.210
y-Aware	0.732±.038	0.590±.096	0.336±.289	0.731±.045	0.640±.057	0.498±.124	0.735±.054	0.607±.084	0.424±.247	0.761±.074	0.652±.079	0.546±.141
SupCon	0.705±.081	**0.619±.080**	**0.507±.151**	0.728±.092	0.639±.066	0.552±.047	0.730±.053	0.596±.045	0.463±.177	0.750±.106	0.622±.146	0.538±.283
CGCP (full)	0.704±.074	0.552±.052	0.369±.157	0.749±.053	0.610±.118	0.422±.264	0.732±.100	0.639±.090	**0.572±.146**	**0.780±.077**	**0.732±.071**	**0.695±.075**

Table 17: **Data efficiency: MMSE regression.** The 10% fraction (≈14 training subjects) produces degenerate near-constant predictions for all methods (MAE, RMSE > 12) and is omitted as a training failure. **Best** per fraction and metric.

	25%			50%			100%		
Method	**MAE**	**RMSE**	**Spearman r**	**MAE**	**RMSE**	**Spearman r**	**MAE**	**RMSE**	**Spearman r**
Supervised	2.378±.253	3.326±.422	**0.409±.095**	2.265±.298	3.465±.417	0.441±.064	2.249±.269	3.394±.448	0.407±.127
SimCLR	2.479±.123	3.553±.253	0.402±.066	2.274±.279	3.490±.368	0.407±.084	**2.204±.254**	3.389±.437	0.418±.106
MAE	2.424±.187	3.349±.413	0.378±.080	2.364±.242	3.545±.392	0.339±.093	2.260±.163	3.495±.358	0.433±.099
y-Aware	**2.347±.148**	**3.310±.403**	0.406±.090	**2.197±.226**	**3.344±.343**	0.443±.080	2.247±.234	3.412±.291	0.433±.099
SupCon	2.505±.104	3.497±.275	0.358±.079	2.249±.262	3.441±.378	**0.446±.075**	2.230±.231	3.414±.343	**0.446±.102**
CGCP (full)	2.350±.183	3.362±.449	0.386±.106	2.302±.211	3.547±.279	0.376±.058	2.242±.284	**3.386±.453**	0.444±.110

track the diagonal closely. CGCP tracks well at low-to-moderate confidence but is overconfident in the highest bin (predicted ≈ 0.93 vs. observed 0.57), consistent with its higher ECE despite stronger discriminative metrics. Bin counts as low as 3–4 subjects illustrate why we report bootstrap CIs rather than point ECE alone.

I Extended Discussion

I.1 Label-Proxy Redundancy

MMSE alone predicts CDR-based diagnosis at AUC$= 0.875$ (5-fold CV logistic regression), exceeding every imaging pipeline evaluated, including fine-tuned CGCP (0.780). With Spearman $r = -0.70$ to CDR and point-biserial $r = -0.63$ to the binary label, MMSE regression during pretraining is closer to indirect regression toward the downstream label than to an independent auxiliary signal. Removing $\mathcal{L}_{\text{mmse}}$ has no statistically distinguishable effect on any metric (Tables 5, 8), and SupCon, with zero exposure to MMSE, attains the highest fine-tuned Spearman r (0.446). Combining MMSE and CDR therefore provides less independent information than "two clinical signals" suggests.

I.2 Instrument Ceiling Effects

CN subjects cluster at the ceiling while demented subjects spread widely; CDR group explains only 39.8% of MMSE variance (Appendix E.2). Continuous-MMSE UMAP colouring shows no low-MMSE gradient for any method (Appendix E.1) because too few subjects score low enough to populate that end of the scale. Residuals (Appendix G.3) are identical with or without $\mathcal{L}_{\text{cdr}}$ or $\mathcal{L}_{\text{mmse}}$, a signature of the target distribution rather than of any pretraining objective.

I.3 Evaluation-Protocol Sensitivity

Frozen rankings do not predict fine-tuned rankings: SimCLR is the strongest frozen classifier but degrades on four of six metrics after fine-tuning, while CGCP-GradNorm barely moves (Appendix F.1). Frozen rankings themselves depend on readout complexity (Appendix F.2). The three CGCP variants are indistinguishable on seventeen of eighteen comparisons (Appendix F.4), and layer-wise optima coincide for only half the methods (Appendix F.3). The protocol itself changes conclusions about individual loss components (Appendix F.5): three components, three different sensitivities to the same change.

Table 18: **Memory-bank size sweep** (subject-level 5-fold CV), weights fixed at the primary BO configuration. **Best** per column.

K	AUC	BA	F1	MAE	RMSE	Spearman r
64	0.765±.109	0.674±.109	0.606±.157	2.235±.233	3.386±.334	0.442±.100
128	0.666±.124	0.644±.112	0.521±.182	**2.196±.263**	**3.370±.486**	0.434±.112
256 (default)	**0.780±.077**	**0.732±.071**	**0.695±.075**	2.242±.284	3.386±.453	**0.444±.110**
512	0.766±.092	0.717±.095	0.677±.118	2.209±.277	3.450±.501	0.422±.114

Table 19: **Constraint threshold sweep** $\lambda_{\text{cdr}} \geq c$; each row uses the best BO configuration under that threshold. **Best** per column.

c	AUC	BA	F1	MAE	RMSE	Spearman r
0.05	0.745±.079	0.611±.076	0.421±.169	2.255±.263	3.449±.468	0.396±.128
0.10 (default)	**0.780±.077**	**0.732±.071**	**0.695±.075**	**2.242±.284**	**3.386±.453**	0.444±.110
0.15	0.710±.077	0.660±.055	0.576±.077	2.247±.230	3.454±.464	0.424±.126
0.25	0.767±.101	0.601±.062	0.438±.126	2.257±.217	3.413±.276	**0.456±.098**

I.4 Small-Batch Architectural Constraints

Removing $\mathcal{L}_{\text{cdr}}$ improves RMSE in 5 of 5 folds while reducing Spearman r; the same signature appears in CGCP-RNC's frozen embeddings via a distinct comparison-based objective, suggesting a property of comparison-based supervision under this batch constraint rather than a CDR-specific artefact. The bank-size sweep is non-monotonic without a mechanistic account (Appendix G.1), whereas the constraint sweep replicates cleanly (Appendix G.2). The variance diagnostic (Appendix G.4) indicates that GradNorm mistook a stable, near-saturated term for one that had stopped learning.

I.5 Generality

Label-proxy redundancy and ceiling effects describe quantifiable properties of OASIS-1/2 that must be re-established on any other cohort. Evaluation-protocol sensitivity and small-batch constraints are more general: studies combining multiple clinical signals in memory-constrained contrastive frameworks should expect frozen/fine-tuned disagreement, probe-dependent rankings, and loss-balancing heuristics confused by regularisers near saturation.

I.6 Limitations

Statistical power at $n = 5$ folds. The Wilcoxon test's minimum two-sided p is 0.0625, so no comparison could reach $p < 0.05$ under it. Borderline effects (AUC for $\mathcal{L}_{\text{mmse}}$ removal, F1 for $\mathcal{L}_{\text{decorr}}$ removal) are suggestive rather than established; repeated k-fold CV would resolve them.

Cross-release participant independence. OASIS-1 and OASIS-2 are separate public releases with anonymised identifiers, but release-specific identifiers do not provide a reliable mechanism for establishing participant identity across datasets. We therefore cannot guarantee complete participant independence between the pretraining and downstream releases. The OASIS-1 experiments should consequently be interpreted as cross-release transfer rather than definitive external validation. Evaluation on an independently identifiable cohort such as ADNI or AIBL is an important next step.

The BO proxy may not track fine-tuned performance. Weights are selected from frozen, 5-epoch embeddings, yet frozen and fine-tuned rankings diverge substantially (Appendix F.1); the same divergence may affect the search itself.

The CDR trade-off mechanism is unexplained. The trade-off is established and replicates in CGCP-RNC, but the residual test designed to explain it did not support its hypothesis.

Investigative variants were not independently tuned. GradNorm's asymmetry and RNC's temperature were not searched with a budget comparable to CGCP's 30-trial BO, so relative performance may partly reflect tuning effort.

Table 20: **Fine-tuned regression residuals by CDR group**, pooled across 5 folds.

	CN		Demented	
Configuration	**Mean**	**Std**	**Mean**	**Std**
CGCP (full)	-0.391	1.453	$+3.054$	3.887
w/o $\mathcal{L}_{\text{cdr}}$	-0.388	1.435	$+3.144$	3.818
w/o $\mathcal{L}_{\text{mmse}}$	-0.408	1.402	$+3.076$	3.768

Table 21: **Coefficient of variation of $\mathcal{L}_{\text{decorr}}$** across 2000 resamples per sample size.

Encoder	$n = 2$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$
CGCP (full)	0.00039	0.00026	0.00019	0.00013	0.00008	0.00004
w/o $\mathcal{L}_{\text{decorr}}$	0.0000115	0.0000076	0.0000056	0.0000037	0.0000024	0.0000013

Single backbone. All results use a 3D ResNet-18; we do not establish whether the findings hold under other architectures.

Partial diagnostic scope. The classical-ML probe covers nine of fourteen objectives, and the variance diagnostic uses final rather than mid-training embeddings.

CDR binarisation discards ordinal structure (0.5/1/2/3) that a multi-class formulation could exploit.

Cohort size. OASIS-1 (212 subjects after filtering) and OASIS-2 (150 subjects) are small; dataset-specific findings require independent re-establishment on cohorts such as ADNI or UK Biobank.

I.7 Future Work

Directions include ViT-based 3D encoders (Dosovitskiy et al., 2020) with masked clinical conditioning; multimodal pretraining with age, sex, APOE $\varepsilon 4$, and amyloid PET as continuous or discrete signals; MCI-to-AD conversion prediction, where ordinal structure should matter most; larger cohorts to test whether BO-selected weights and the constraint floor transfer; and multi-class CDR staging.

Table 22: **Expected Calibration Error** on OASIS-1 (pooled, $n = 212$). Overlapping CIs indicate no reliable difference.

Method	**ECE ↓**	**95% CI**
Supervised-only	0.215	[0.171, 0.285]
CDR-only (ablation)	0.059	[0.054, 0.144]
SimCLR	**0.056**	[0.049, 0.138]
CGCP (ours, BO)	0.105	[0.082, 0.178]

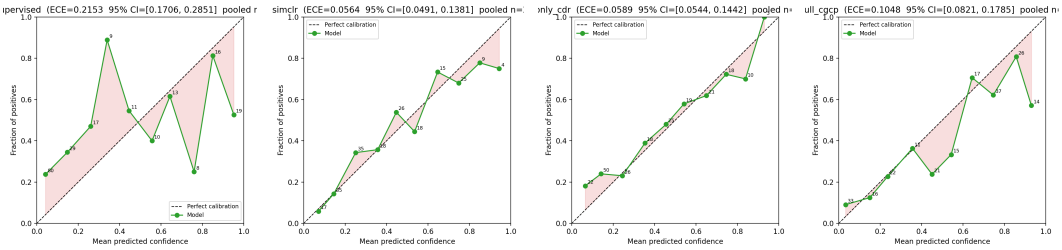


Figure 6: **Reliability diagrams** for the methods in Table 22, pooled across folds. The diagonal indicates perfect calibration; ECE and CI in panel titles.