BRAINGENFLOW: MODELING LONGITUDINAL PROGRESSION VIA RESIDUAL FLOW MATCHING AND GAUSSIAN IMPUTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting individualized brain trajectories from a single baseline scan is essential for prognosis and early intervention in neurodegenerative diseases, yet existing methods either operate on composite biomarkers or require sequences of follow-up visits at inference. We introduce BrainGenFlow, a two-stage framework that produces month-resolved, multi-region brain trajectories from a single scan and clinical covariates. The central design addresses cohort sparsity through Bayesian distillation: a deep kernel Gaussian process performs exact temporal imputation per region of interest, converting sparse irregular visits into dense monthly supervisory trajectories. We utilize these trajectories to train an autoregressive flow matching model through a new parameterization technique called Residual Flow Matching, to model the changes in the biomarkers directly. At inference, BrainGenFlow forecasts decade-long trajectories across 145 brain regions using only the baseline scan, without requiring access to longitudinal data. On long-horizon AD biomarker forecasting, our method achieves state-of-the-art performance with improvements that hold under leave-one-study-out evaluation. Furthermore, predicted regional rate-of-change align with established Alzheimer's disease progression pattern.

1 INTRODUCTION

Accurately forecasting the temporal evolution of brain biomarkers from a single baseline scan is essential for monitoring neurodegenerative disease progression such as Alzheimer's Disease (AD), identifying optimal intervention windows, and supporting prognostic enrichment in clinical trials (Maheux et al., 2023). A clinically useful forecasting model must take a patient's brain measurements at the current visit together with a few clinical covariates, and predict the state of the brain at every month over the following decade. Producing such individualized, month-resolved trajectories from limited per-subject data is the central challenge we address.

Most approaches in AD forecasting operate on a small set of summary measures, such as cognitive scores, total hippocampal and ventricular volumes, and CSF or PET biomarkers. Recurrent neural networks such as RNN-AD (Nguyen et al., 2020) autoregressively predict these endpoints month-by-month using imputation strategies to handle irregular visit intervals. Non-linear mixed-effects models such as AD Course Map (Maheux et al., 2023; Koval et al., 2021b) forecast subject-specific logistic trajectories for the same class of endpoints and have been validated as the strongest performer in the TADPOLE challenge (Marinescu et al., 2018). Other approaches such as the temporal deep kernel regression of (Tassopoulou et al., 2025; 2022) handle irregularly sampled training observations through a learned temporal encoding, but their forecasts are restricted to univariate trajectories. These methods establish that long-horizon forecasting in AD is feasible, but they collapse the brain to a small number of scalar measurements, discarding the spatial heterogeneity that characterizes neurodegeneration across cortical and subcortical regions.

Recent methodologies do not address the particular setting. Latent ODE-style models (Rubanova et al., 2019) handle irregular sampling, but they are designed to encode a sequence of observations into a latent state. They degrade when only a single baseline scan is available, and they target reconstruction rather than month-resolved long-horizon forecasting. More broadly, recent generative time-series methods are evaluated under different sparsity regimes, either endpoint-pair supervision

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.

over observed intervals (Chen et al., 2025) or structured low-dimensional clinical measurements (Zhang et al., 2024), that do not match the extreme month-level sparsity of longitudinal neuroimaging cohorts, where typical observation horizons span several years but each subject contributes only a handful of actual visits. The clinically relevant regime therefore demands a model that simultaneously (i) produces month-resolved predictions over multi-year horizons, (ii) operates over all brain regions , in our case 145, rather than a few summary scores, and (iii) forecasts from a single baseline scan plus clinical covariates. To our knowledge, no prior work addresses all three requirements together.

We introduce BRAINGENFLOW, a two-stage forecasting framework designed for the sparse regime of longitudinal cohorts and propose the following contributions: **1)** we introduce a temporal imputation strategy based on temporal deep kernel Gaussian Processes that converts sparse, irregular longitudinal neuroimaging visits into dense monthly brain trajectories, enabling month-resolved learning without requiring regularly sampled observations, offering a **6.8%** improvement over an otherwise identical flow trained on observed visit pairs from the regime proposed by (Chen et al., 2026), with lower error on all five folds , **2)** we showcase that standard next-state flow matching methods underperform in the volumetric regime and we implement a residual flow-matching formulation for volumetric brain biomarker forecasting that models local month-to-month changes rather than absolute next states, improving long-horizon autoregressive stability, at inference time **0.055s** per subject for a 120-month trajectory and a **59.5%** improvement compared to next-state prediction under an identical architecture **3)** we demonstrate that BRAINGENFLOW forecasts individualized 145-region brain trajectories from a single baseline scan and clinical covariates, achieving a **26.8%** reduction in mean absolute error over an adapted version of RNN-AD (Nguyen et al., 2020) in standardized brain volumes and against state-of-the-art flow matching time series prediction algorithms, namely Trajectory Flow Matching (Zhang et al., 2024) by **69.9%**; performance holds under leave-one-study-out evaluation, and predicted rate-of-change patterns align with established Alzheimer's disease progression.

2 RELATED WORK

Disease progression models for low-dimensional biomarkers. A substantial body of work models the longitudinal progression of Alzheimer's disease biomarkers (Marinescu et al., 2018). Lorenzi et al. (2019) introduced a Gaussian process-based disease progression model for predicting cognitive scores and volumetric measurements, but the formulation imposes assumptions on the number of observations per subject and is restricted to a small number of biomarkers. Koval et al. (2021a) proposed a Bayesian mixed-effects model (AD Course Map) that produces subject-specific biomarker trajectories and was validated as the strongest performer in the TADPOLE challenge (Marinescu et al., 2018), though it operates on low-dimensional inputs. Abi Nader et al. (2020) estimated spatiotemporal biomarker progression at the population level without adapting to individual follow-up data. Manifold learning approaches (Hong et al., 2019; Gruffaz et al., 2021) infer biomarker trajectories from longitudinal data but require at least two acquisitions per subject at inference time. Tassopoulou et al. (2022) proposed a temporal deep kernel regression method that handles irregular observation timing and personalizes its predictions as additional follow-up visits become available for a subject, but it does so for univariate biomarker trajectories. None of these methods simultaneously addresses multivariate biomarker forecasting, month-level temporal resolution, and inference from a single baseline scan.

Recurrent and continuous-time forecasting models. Recurrent neural networks have been the standard tool for forecasting in clinical time series (Lipton et al., 2016; Che et al., 2018; Cao et al., 2018; Rajkomar et al., 2018). In the AD setting, RNN-AD (Nguyen et al., 2018; 2020) is one of the most established baselines. The model autoregressively forecasts cognitive scores and a small number of volumetric and biofluid biomarkers month-by-month, using imputation strategies to handle irregular visit intervals; we adopt it as our primary point of comparison by adapting for multivariate biomarkers. Continuous-time alternatives such as Neural ODEs (Chen et al., 2018; De Brouwer et al., 2019; Dupont et al., 2019) and Latent ODEs (Rubanova et al., 2019) model irregular sampling natively, and Neural SDEs further introduce stochastic drift to capture model uncertainty (Oh et al., 2024). While these methods handle irregular sampling, their formulation is designed for sequence-to-sequence settings rather than single-scan long-horizon forecasting.

Diffusion and flow-based models for longitudinal data Instead of recurrent networks, generative models have recently been applied for downstream medical applications. These models generally follow two families: image-based and biomarker-based. The first forecasts progression in image space, voxel-level synthesis (Yoon et al., 2023; Litrico et al., 2024; Puglisi et al., 2025; Pombo et al., 2021; Liu et al., 2025) or deformation fields between scans (Kim et al., 2022; Balakrishnan et al., 2019), evaluated with image-level metrics; these target a different output space and are not directly comparable to regional trajectory forecasting. Δ-LFM (Chen et al., 2026) regresses a finite-difference velocity between observed longitudinal latents, but relies on an ArcRank-regularized latent space that a fixed-dimensional ROI vector does not provide. What transfers to our setting is its supervision scheme, endpoint-pair targets over observed intervals, which we compare against directly. BrLP (Puglisi et al., 2025) performs forecasting at the image latent level, using a ControlNet module to condition predictions on univariate biomarker trajectories alongside MRI inputs. We choose to produce these trajcetories as they are an effective means of monitoring patient trajectory. ConDOR (Cho et al., 2025) employs a dual regional-temporal diffusion architecture to generate synthetic biomarker trajectories conditioned on ordinal attributes, with the goal of data augmentation between known patient visits, but does not offer support for longitudinal prediction. TFM (Zhang et al., 2024) reparameterizes flow matching with an SDE to control predictive uncertainty for low-dimensional ICU trajectories. TSFlow (Kollovieh et al., 2024) combines Gaussian processes with flow matching, but uses the GP as an informed prior within the generative process rather than as a training-time temporal imputation mechanism, and does not address the extreme observation sparsity that characterizes longitudinal neuroimaging cohorts.

3 PRELIMINARIES

Conditional Flow Matching Flow matching (Lipman et al., 2022) is a simulation-free framework for training continuous normalizing flows (Chen et al., 2018), where the goal is to learn a smooth, time-dependent vector field $v_\theta(t, x)$ that transports samples from a source distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$ to a target distribution $\nu \in \mathcal{P}(\mathbb{R}^d)$. The framework assumes the existence of a ground-truth vector field $v_t(x)$ such that the flow defined by the ODE $dx_t = v_t(x_t)dt$ with $x_0 \sim \mu$ produces $x_1 \sim \nu$. The training procedure regresses v_θ onto v_t via the flow matching objective:

$$\mathcal{L}_{\text{FM}}(\theta) := \mathbb{E}_{t\sim\mathcal{U}(0,1),\, X_t\sim p_t} \left\| v_\theta(t, X_t) - v_t(X_t) \right\|^2, \tag{1}$$

where p_t is the marginal distribution of x_t, i.e., a probability path interpolating between μ and ν. Under mild regularity assumptions, v_t generates p_t if and only if the pair satisfies the continuity equation

$$\frac{\partial p_t}{\partial t} + \nabla \cdot (p_t v_t) = 0, \quad p_0 = \mu, \quad p_1 = \nu. \tag{2}$$

At inference, given a learned approximation v_θ, one draws $\tilde{x}_0 \sim \mu$ and integrates $d\tilde{x}_t = v_\theta(t, \tilde{x}_t)dt$ from $t = 0$ to $t = 1$, yielding $\tilde{x}_1$ approximately distributed according to ν.

Direct optimization of $\mathcal{L}_{\text{FM}}$ is generally intractable because the marginal vector field v_t has no closed form. Conditional Flow Matching (CFM) (Tong et al., 2023) circumvents this by introducing a latent variable z drawn from a prior q, together with a conditional vector field $v_t(x \mid z)$ and a conditional probability path $p_t(x \mid z)$ that satisfy the continuity equation between the conditional marginals $\mu(dx \mid z)$ and $\nu(dx \mid z)$. The CFM objective is as follows:

$$\mathcal{L}_{\text{CFM}}(\theta; q) := \mathbb{E}_{t\sim\mathcal{U}(0,1),\, Z\sim q,\, X_t\sim p_t(\cdot|Z)} \left\| v_\theta(t, X_t) - v_t(X_t \mid Z) \right\|^2. \tag{3}$$

4 METHODOLOGY

We now describe the core components of BRAINGENFLOW. Section 4.1 details the temporal imputation stage that produces dense monthly trajectories from sparse longitudinal observations. Section 4.2 presents the residual flow matching objective that drives the generative process. Section 4.3 describes the history conditioning mechanism for the trajectory generation.

4.1 TEMPORAL IMPUTATION AS BAYESIAN DISTILLATION OF BRAIN DYNAMICS

The challenge of long-horizon forecasting from a single baseline scan is that the training signal itself is sparse: each subject contributes only a few irregularly spaced visits. We address this by

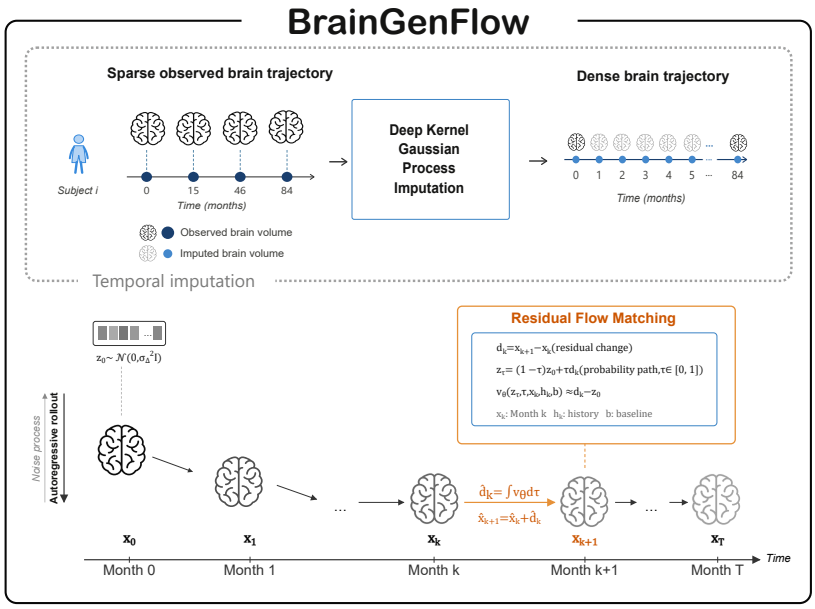

Figure 1: **Conceptual Overview of BRAINGENFLOW.** *Top (Temporal Imputation):* For each subject i, sparse and irregularly sampled brain volumes are converted into a dense monthly trajectory via an ensemble of Deep Kernel Gaussian Process regressors, one per ROI, providing a temporally regular representation over the observed support. *Bottom (Residual Flow Matching):* A velocity field v_θ is trained to predict mean-centered residual changes $d_k = x_{k+1} - x_k$ along a straight probability path $z_\tau = (1-\tau)z_0 + \tau d_k$, conditioned on the current state x_k, a recurrent history embedding h_k, and a static baseline embedding b. At inference, $\hat{d}_k$ is recovered by integrating v_θ in τ and applied autoregressively, $\hat{x}_{k+1} = \hat{x}_k + \hat{d}_k$, to generate individualized long-horizon trajectories from an initial scan x_0.

treating the imputation stage as a form of Bayesian distillation. A deep kernel Gaussian process (DKGP), fit per ROI (Wilson et al., 2015; Tassopoulou et al., 2025), performs exact posterior inference over each subject's latent trajectory using all available longitudinal observations from the training cohort. The resulting dense monthly trajectories serve as supervisory targets for the downstream flow-matching model (Section 4.2), which learns to reproduce the same dynamics in an amortized form. At inference time, the generative model can produce month-resolved forecasts from a single scan, without requiring access to longitudinal data or to the GP itself.

Given a subject i observed at sparse, irregular time points $\mathcal{T}_i = \{t_1^{(i)}, \ldots, t_{m_i}^{(i)}\} \subset \mathbb{R}_{\geq 0}$ in months, with measurements $\{y_d^{(i,j)}\}_{j=1}^{m_i}$ for each ROI $d \in \{1, \ldots, D\}$, we construct a dense monthly trajectory $X^{(i)} = (x_0^{(i)}, x_1^{(i)}, \ldots, x_n^{(i)})$ with $n = \lfloor t_{m_i}^{(i)} \rfloor$, where each $x_k^{(i)} \in \mathbb{R}^D$, via an ensemble of deep kernel Gaussian process regressors, one per Region of Interest (ROI). The Temporal imputation module is depicted in the top row of Figure 1.

Each observation is represented by a subject-specific context vector

$$\tilde{x}^{(i)}(t) = \left[x_0^{(i)};\ c^{(i)};\ t \right], \tag{4}$$

where $x_0^{(i)} \in \mathbb{R}^D$ is the baseline imaging feature vector and $c^{(i)}$ collects static subject covariates. For each ROI d, a dedicated learned transformation $\Phi_d : \mathbb{R}^{\dim(\tilde{x})} \to \mathbb{R}^r$ maps $\tilde{x}^{(i)}(t)$ into a latent space, in which the covariance is computed via an ARD-RBF kernel:

$$k_d\Big(\tilde{x}^{(i)}(t),\, \tilde{x}^{(i)}(t')\Big) = k_{\mathrm{RBF}}\Big(\Phi_d(\tilde{x}^{(i)}(t)),\ \Phi_d(\tilde{x}^{(i)}(t'))\Big). \tag{5}$$

The latent biomarker function for ROI d is then modeled as

$$f_d^{(i)}(t) \sim \mathcal{GP}\Big(\mu_d(\tilde{x}^{(i)}(t)),\ \ k_d\Big(\tilde{x}^{(i)}(t),\, \tilde{x}^{(i)}(t')\Big)\Big), \quad y_d^{(i,j)} = f_d^{(i)}\Big(t_j^{(i)}\Big) + \varepsilon_d,\ \ \varepsilon_d \sim \mathcal{N}(0, \sigma_d^2), \tag{6}$$

with per-ROI mean function μ_d and noise variance σ_d^2. The transformation Φ_d, the mean function μ_d, and the kernel hyperparameters are learned per ROI by maximizing the GP log marginal likelihood. The dense state at month k is then assembled from the per-ROI posterior means:

$$x_k^{(i)} = \left(\mathbb{E}\Big[f_1^{(i)}(k) \mid \mathbf{y}_1^{(i)} \Big], \ldots, \mathbb{E}\Big[f_D^{(i)}(k) \mid \mathbf{y}_D^{(i)} \Big] \right)^\top, \quad k \in \{0, 1, \ldots, n\}. \tag{7}$$

Imputation is restricted to the observed support $[0, n]$, avoiding extrapolation. The resulting trajectories $X^{(i)}$ provide the dense monthly support on which the generative process of Section 4.2 is trained.

4.2 Residual Flow Matching

Given the DKGP-imputed trajectories $X^{(i)} = (x_0, x_1, \ldots, x_n)$ from Section 4.1, we model each patient's progression as a sequence of one-month transitions $x_k \to x_{k+1}$, with $x_k \in \mathbb{R}^D$ the ROI biomarker vector at month k. Flow matching traditionally models a vector field that transports a base distribution, typically a Gaussian distribution, directly to the absolute data distribution (Lipman et al., 2022; Albergo & Vanden-Eijnden, 2022; Albergo et al., 2025; Liu et al., 2022). Because monthly neuroanatomical changes are small relative to absolute ROI volumes, direct next-state generation is a too highly complex task to model, along with retaining a history condition. To address this, we reparameterize the generation target so the network predicts the localized residual change between time steps rather than the absolute next state x_{k+1}, a subtype of the paradigm we call *Residual Flow Matching*, as visible in the bottom part of Figure 1.

The transition residual change is defined as $d_k = x_{k+1} - x_k$. Rather than generating the absolute next state x_{k+1}, the model learns the local residual transition d_k. This reparameterization focuses the generative model on month-to-month progression dynamics, which are small relative to the absolute magnitude of ROI biomarkers. Consequently, the velocity network operates in residual space, $v_\theta(z_\tau, \tau \mid h_k, x_k, E_{\text{baseline}})$, where $\tau \in [0, 1]$ is the continuous flow integration time and z_τ is the intermediate flow state. For the conditional probability path, we adopt the optimal-transport displacement interpolant used in flow matching (Lipman et al., 2022), adapted to the residual target. We sample the initial noise $z_0 \sim \mathcal{N}(0, I)$ independently of the residual target d_k. The probability path forms a straight line in residual space,

$$z_\tau = (1 - \tau) z_0 + \tau d_k, \tag{8}$$

with corresponding target velocity $u_\tau = d_k - z_0$. The model is optimized by mean squared error to match this residual velocity:

$$\mathcal{L}_{\text{RFM}}(\theta) = \mathbb{E}_{\tau,\, q(d_k),\, z_0} \left[\left\| v_\theta\big(z_\tau, \tau \mid h_k, x_k, E_{\text{baseline}}\big) - (d_k - z_0) \right\|_2^2 \right]. \tag{9}$$

By reparameterizing the target, the network focuses on localized disease-progression dynamics rather than reconstruction of the full biomarker state. During inference, sampling is performed by integrating the learned vector field v_θ from $\tau = 0$ to $\tau = 1$ using an off-the-shelf ODE solver, yielding the predicted residual $\hat{d}_k$. The final absolute state is then recovered autoregressively as $\hat{x}_{k+1} = \hat{x}_k + \hat{d}_k$.

Although the network is trained to regress the velocity field, this objective is equivalent up to a time-dependent reweighting to directly predicting the residual target d_k. Given the interpolant in Eq. 8, a target-predicting network $\hat{d}_\theta(\tau, z_\tau)$ induces the velocity $v_\theta(\tau, z_\tau) := (\hat{d}_\theta(\tau, z_\tau) - z_\tau)/(1 - \tau)$. The conditional vector field satisfies $u_\tau(z \mid d_k) = (d_k - z_\tau)/(1 - \tau)$, which reduces to $d_k - z_0$ along the straight path $z_\tau = (1 - \tau) z_0 + \tau d_k$. Letting the target-prediction loss be

$$\mathcal{L}_{\text{target}}(\theta) = \mathbb{E}_{\tau,\, q(d_k),\, p_\tau(z \mid d_k)} \left[\left\| \hat{d}_\theta(\tau, z_\tau) - d_k \right\|_2^2 \right], \tag{10}$$

we have

$$\mathcal{L}_{\text{target}}(\theta) = \mathbb{E}_{\tau,\, q(d_k),\, p_\tau(z \mid d_k)} \left[(1 - \tau)^2 \left\| v_\theta(\tau, z_\tau) - u_\tau(z \mid d_k) \right\|_2^2 \right]. \tag{11}$$

Thus, target prediction corresponds to a time-weighted flow-matching objective with weight $w(\tau) = (1 - \tau)^2$. The two objectives share the same minimizer and differ only in how supervision is allocated across flow time. In implementation, residuals are normalized using training-fold statistics before flow training and transformed back to the original scale during autoregressive rollout; details are provided in Appendix A.

4.3 History Conditioning

The velocity network $v_\theta(z_\tau, \tau \mid h_k, x_k, E_{\text{baseline}})$ introduced in Section 4.2 is conditioned on three signals: the current state x_k, a recurrent history embedding h_k, and a static baseline embedding E_{baseline}. We now describe how the latter two are constructed.

Prior autoregressive generative forecasting models condition next-step generation on temporal context, either through recurrent hidden states in diffusion-based time-series models such as TimeGrad (Rasul et al., 2021), or through fixed-time memory modules in trajectory flow matching (Zhang et al., 2024). BrainGenFlow follows the broader principle of recurrent history conditioning, but differs in that the recurrent state conditions a residual flow-matching model over month-to-month brain-volume changes rather than a diffusion model over absolute next observations. Specifically, we summarize the patient's generated history up to month k with a learned recurrent embedding $h_k \in \mathbb{R}^H$, produced by a multi-layer GRU encoder (Chung et al., 2014). The GRU state is updated at every autoregressive step so that conditioning information propagates continuously through the generated trajectory. During inference, this history is updated using the model-generated states.

To anchor predictions and mitigate error acumulation over long horizons, we additionally extract a static baseline representation E_{baseline} by passing the initial scan x_0 together with clinical covariates C (age, sex, diagnosis, and APOE-ε4 allele count) through a baseline encoder. The history embedding h_k, the baseline embedding E_{baseline}, and a sinusoidal embedding of the flow integration time τ jointly drive Feature-wise Linear Modulation (FiLM) (Perez et al., 2018) of the velocity network's residual blocks, while x_k enters as a concatenated feature alongside the flow state.

5 Experiments

5.1 Prediction of Regional Volume Trajectories

We evaluate BrainGenFlow on a diverse, multi-cohort dataset comprising 3,969 participants and 15,049 neuroimaging sessions from nine longitudinal studies in the iSTAGING Consortium (Habes et al., 2021). Full cohort and preprocessing details are reported in Appendix D. Subjects are split into five folds for subject-level cross-validation, yielding held-out predictions for all participants. Training data are used for temporal imputation of longitudinal brain measurements and for fitting BrainGenFlow. The input X consists of 145 brain volumes plus clinical variables (age, sex, diagnosis, and APOE ε4 allele count). At inference, a single baseline scan is used to produce individualized volumetric trajectories spanning the training time horizon. We use a maximum rollout window of 120 months to assess the model's ability to capture long-term neurodegeneration. Evaluation metrics are computed only at visits with observed measurements. Details on temporal imputation, model architecture, and competing baselines are reported in Appendices E, G.1, G.2, F, and G, respectively and more information on the per-fold performance is reported on Appendix B

Table 1 reports the mean results across all five folds. BrainGenFlow achieves a mean MAE of 0.2167 (95% CI [0.2079, 0.2255]), consistently forecasting brain volumes within a quarter of a standard deviation of ground truth over a 10-year horizon. Performance is stable across folds (0.2105-0.2258), confirming robust generalization to diverse subject cohorts. BrainGenFlow

Table 1: MAE on standardized regional volumes at follow-up visits, averaged over 145 ROIs (5-fold subject-level CV; 95% CIs across folds). Lower-block methods differ only in training objective. †Adapted to 145 ROI outputs. ‡Δ-LFM supervision on consecutive visit pairs.

Method	Supervision	Mean MAE (95% CI)
No-change (persistence)	—	0.2684 [0.2622, 0.2747]
Linear mixed-effects (per ROI)	Observed visits	0.2472 [0.2357, 0.2587]
GAM (per ROI)	Observed visits	0.2384 [0.2331, 0.2436]
Deep Regression	Observed visits	0.3003 [0.2961, 0.3045]
RNN-AD†	Observed visits	0.2961 [0.2918, 0.3003]
Endpoint-pair flow (Δ-LFM) ‡	Observed pairs	<u>0.2326</u> [0.2292, 0.2360]
TFM-ODE	Dense monthly	0.7206 [0.6992, 0.7421]
Det-MLP (residual regression)	Dense monthly	0.3220 [0.1407, 0.5033]
Det-Transformer (residual regression)	Dense monthly	0.2240 [0.2190, 0.2290]
BrainGenFlow (ours)	Dense monthly	**0.2167** [0.2079, 0.2255]

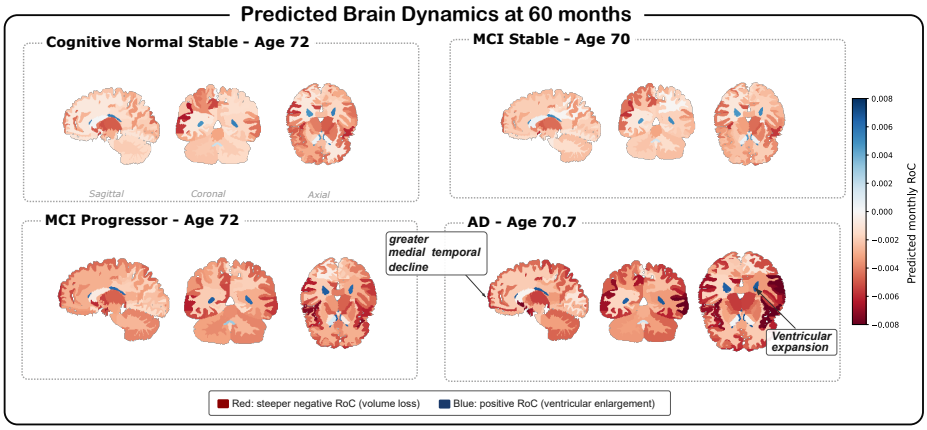

Figure 2: **Representative subject-level predicted volumetric RoC maps.** Predicted monthly RoC for four representative subjects (70-72 years) across progression groups. Red indicates volume loss; blue indicates ventricular enlargement. Atrophy intensity and spatial extent increase monotonically from CN Stable to AD. All maps share the same colorscale ($\pm 8 \times 10^{-3}$ month^{-1}).

outperforms both the Deep Regression and RNN-AD baselines, as well the next-prediction state baselines.

The statistical baselines bound what is achievable without learning individual dynamics: no-change carries baseline anatomy forward and reaches 0.2684, while GAM (0.2384) and a per-ROI mixed-effects model (0.2472) receive the same 145 baseline volumes, covariates and target time as the neural methods, and are univariate only in their output parameterization. To isolate what the generative formulation contributes, we train two deterministic residual predictors on the same dense trajectories, with identical conditioning and optimization, replacing the flow-matching objective with direct regression: at matched capacity the MLP head reaches 0.3220 and is unstable across folds (95% CI [0.141, 0.503]), reflecting drift over the 120-step rollout, while a Transformer head matches our accuracy (0.2240) at roughly three times the parameters of the velocity network. The endpoint-pair flow (0.2326) replaces dense monthly supervision with Δt-conditioned targets over observed visit pairs, the scheme used by prior delta-based flow models.

We report TFM with its original parameterization; see Section 5.4 for an analysis of the residual reparameterization's contribution. TFM-ODE underperforms in this setting under its original next-state parameterization. We hypothesize that, in multivariate ROI space, monthly changes are small relative to absolute biomarker magnitudes, so direct next-state generation forces the model to reconstruct largely unchanged structure at each step. Residual reparameterization instead concentrates model capacity on the clinically relevant local change signal.

A detailed analysis of error by forecast horizon, baseline diagnosis, sex, age group, and progression status is provided in Appendix H.2; in brief, MAE increases monotonically from 0.166 at 0-12 months to 0.331 at 60-70 months, and is primarily driven by clinical trajectory complexity rather than demographic factors. A per-region breakdown (Appendix H.3) confirms that prediction error tracks known biological variability, with ventricles hardest and primary cortices easiest, and remains bounded across the five-year horizon. Qualitative forecasts at arbitrary months across all diagnostic groups are provided in Appendix H.4, illustrating BRAINGENFLOW's ability to be queried at any future month from a single baseline scan. We also present in Appendix C ablations on imputation strategies, to show the sensitivity of the downstream model to the imputation phase of the framework.

5.2 PREDICTED RATE OF CHANGE STRATIFIES COGNITIVE PROGRESSION GROUPS

We assess whether predicted trajectories encode clinically meaningful signals by examining monthly RoC across 145 ROIs in six progression groups (statistical details in Appendix H.1). Figure 2 shows that at the individual level, the spatial signature of disease severity is apparent: CN Stable subjects exhibit mild, diffuse atrophy; MCI Progressors show markedly elevated medial temporal and parietal involvement; and AD presents the most saturated and spatially extensive pattern, with deep red concentrated in the entorhinal and parahippocampal cortices, consistent with the known

early vulnerability of these regions in the AD neurodegeneration cascade (Braak & Braak, 1991). Crucially, these differences emerge from a *single* baseline scan, without longitudinal follow-up at inference time.

A Kruskal-Wallis test across all six groups confirms this statistically (H=66.6, p=5.3×10^{-13} on mean RoC). The discriminative signal resides in the *rate* and spatial heterogeneity of atrophy. Median whole-brain RoC ranges from -4.10×10^{-3} month^{-1} (MCI Stable) to -4.86×10^{-3} month^{-1} (AD) across the full spectrum. Critically, MCI Progressors already show a $\sim$9% faster atrophy rate than MCI Stable subjects (-4.48 vs. -4.10×10^{-3} month^{-1}), demonstrating that the model captures clinically relevant within-diagnosis heterogeneity. Pairwise tests (Table 8 in Appendix H.1) confirm significant separation at every level of the disease hierarchy, including within-MCI discrimination between stable and progressor subgroups (Cliff's δ=0.13, q=6.5×10^{-3}) - a distinction difficult to detect from a single timepoint. The one null result, CN Stable vs. CN→MCI (δ=0.02, q=0.75), is interpretable: prodromal conversion likely precedes detectable structural atrophy. Full pairwise statistics and ROI-level results (128/145 regions significant, q<0.05) are provided in Appendix H.1. We also report model performance with and without diagnosis as a covariate on Appendix I.

5.3 Generalization to External Cohorts

We evaluate BrainGenFlow under a leave-one-study-out (LOSO) protocol, in which each evaluation cohort is held out entirely and the model is trained on the remaining studies. This assesses transfer to an external cohort whose subjects and visits are absent from training. Here, we present per-study results for seven cohorts: ADNI, AIBL, BLSA, HABS, OASIS, PENN, and PREVENTAD. Evaluation is restricted to follow-up observations (month ≥ 1; dataset details in Appendix D).

Figure 3(a) shows cohort MAEs ranging from 0.123 to 0.298 and R^2 from 0.793 to 0.962. PREVENTAD and HABS achieve the lowest errors, with MAEs of 0.123 and 0.157 and corresponding R^2 of 0.962 and 0.926. ADNI, PENN, AIBL, and OASIS form an intermediate group, with MAEs of 0.231–0.282 and R^2 of 0.825–0.891. BLSA has the highest error (MAE = 0.298, R^2 = 0.793). The mean MAE across the seven cohorts is 0.227, compared with the in-fold reference of 0.217. Figure 3(b) compares cohort-specific forecast-horizon profiles with the in-fold 5-fold CV reference (Section 5.1). Individual study errors increase over their available horizons, with several cohorts remaining close to the in-fold reference at 12–60 months. At 60–120 months, BLSA and OASIS show larger errors, while ADNI, PENN, and AIBL remain closer to the reference. These patterns suggest transfer at intermediate forecast horizons, with reduced accuracy in some cohorts at longer horizons.

5.4 Ablation Study of Model Components

We isolate the contribution of each component of BrainGenFlow by modifying a TFM baseline along four axes: the prediction target (residual delta vs. direct next-state), the history representation

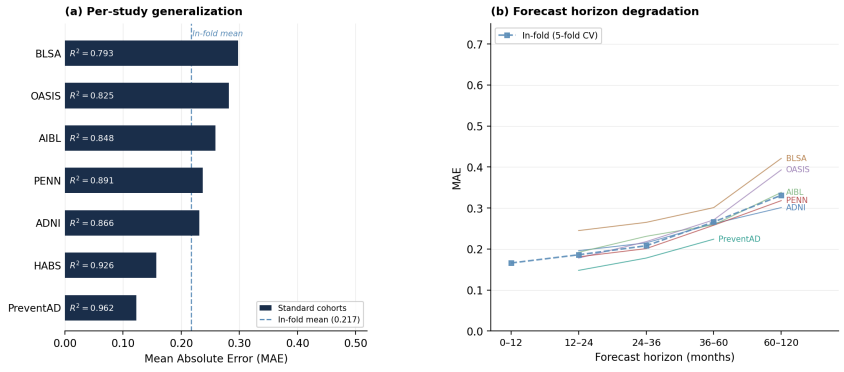

Figure 3: **Generalization to external cohorts under LOSO evaluation. (a)** Per-cohort MAE for seven cohorts, with R^2 annotated inside the bars. The dashed line marks the in-fold mean MAE of 0.217. **(b)** MAE by forecast horizon for cohorts with available observations. Colored lines denote individual cohorts; the dashed blue line shows the in-fold five-fold CV reference.

Table 2: Ablation across all five folds. *Target*: direct next state x_{k+1} or residual $d_k = x_{k+1} - x_k$. *History*: fixed-length memory window or GRU embedding h_k. *Anchor*: static baseline embedding E_{baseline}. *Dynamics*: linear flow matching with ODE sampling or a VP-SDE formulation.

Variant	Target	History	Anchor	Dynamics	Mean MAE (95% CI)
TFM-ODE	Next state	Window	–	ODE	0.7206 [0.6992, 0.7421]
Memory Window (ODE)	Next state	Window	✓	ODE	0.8010 [0.7340, 0.8680]
TFM-SDE	Next state	Window	–	SDE	0.9100 [0.8620, 0.9580]
SDE Direct	Next state	GRU	✓	SDE	1.0350 [0.6360, 1.4340]
Memory Window (SDE)	Next state	Window	✓	SDE	1.7560 [1.5620, 1.9500]
TFM-ODE + Res	Residual	Window	–	ODE	<u>0.2910</u> [0.2630, 0.3190]
BRAINGENFLOW (ours)	Residual	GRU	✓	ODE	**0.2167** [0.2079, 0.2255]

(GRU vs. fixed-length memory window), static baseline achoring and the flow formulation (linear FM/ODE vs. VP-SDE). Results on all folds are reported in Table 2.

Replacing TFM's direct x_{k+1} prediction with residual reparameterization ($d_k = x_{k+1} - x_k$) reduces MAE by 59.5% (0.7206 $\rightarrow$ 0.2910) under an otherwise identical TFM architecture. We hypothesize that, due to the small magnitude of inter-month ROI changes in neurodegeneration, predicting x_{k+1} directly forces the model to reconstruct most of the input from noise at every step, while residual prediction allows the model to focus on the trajectory dynamics. Under the FM/ODE formulation, replacing the GRU with a fixed-length window (with baseline anchor and direct prediction) gives MAE 0.801, worse than TFM-ODE. Combined with residual reparameterization and baseline conditioning, our GRU-based history encoder reduces MAE further to 0.2167 - a 25.5% improvement over TFM-ODE-Res (0.291). The remaining gain is attributable to recurrent history encoding and baseline anchoring. Across all configurations, replacing the FM objective with an SDE formulation substantially degrades performance. The TFM-SDE baseline reaches MAE 0.91, while combining SDE with the memory window diverges to MAE 1.7560. We attribute this to the long ($T = 120$) autoregressive horizon: stochastic noise compounds across rollout steps, while linear FM provides a more stable training and sampling trajectory.

6 CONCLUSION

In this work, we introduce BRAINGENFLOW, a two-stage framework for longitudinal brain trajectory modeling. We adopt a Gaussian Process imputation strategy to densify temporal priors in patients with limited numbers of scans. We then utilize this prior to train a generative process, which scales the forecasting process into multiple dimensions and disentangles the generation strategy from access to the patient data on inference. We project from a single scan and introduce a reparameterization trick to take the small biomarker changes into account and create a realistic training target. The framework could support population-level analyses in dementia and Alzheimer's research, including normative trajectory modeling and cohort-level progression studies. Individual-level prognostic applications would require dedicated validation in a clinical setting.

Limitations The current model takes a single baseline scan as input and is not designed to update its forecasts when additional scans become available. This was a deliberate scoping choice to test whether multivariate long-horizon forecasting is feasible from minimal input, but it leaves room for richer conditioning on partial trajectories. Second, the residual flow-matching stage treats the GP-imputed monthly values as supervisory ground truth, which means any imputation error or trajectory abnormality propagates into the generative model; jointly modeling imputation uncertainty and generative learning is a natural next step.

Future Work We aim to extend BRAINGENFLOW in two directions. Firstly, we will generalize the conditioning mechanism to accommodate repeated measurements and partially observed trajectories, allowing the model to refine its forecasts as new scans arrive. Secondly, we will equip the generative process with calibrated uncertainty estimates over predicted brain trajectories. Beyond these methodological extensions, the framework provides individualized brain trajectory forecasts from a single cross-sectional scan, supporting subject-specific characterization of neurodegenerative progression. Translation to clinical decision-making applications, however, remains a longer-term objective that will require extensive validation in clinically representative cohorts and prospective clinical settings.

REFERENCES

Clément Abi Nader, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. Monotonic Gaussian Process for spatio-temporal disease progression modeling in brain imaging data. *NeuroImage*, 205:116266, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.116266. URL https://www.sciencedirect.com/science/article/pii/S1053811919308572.

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, August 2019. ISSN 1558-254X. doi: 10.1109/tmi.2019.2897538. URL http://dx.doi.org/10.1109/TMI.2019.2897538.

V. M. Bashyam, G. Erus, Y. Cui, D. Wu, G. Hwang, A. Getka, A. Singh, G. Aidinis, K. Baik, R. Melhem, and E. Mamourian. Dlmuse: Robust brain segmentation in seconds using deep learning. *Radiology: Artificial Intelligence*, 7(6):e240299, November 2025.

H. Braak and E. Braak. Neuropathological stageing of alzheimer-related changes. *Acta Neuropathologica*, 82:239–259, 1991. doi: 10.1007/BF00308809.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Hao Chen, Rui Yin, Yifan Chen, Qi Chen, and Chao Li. Learning patient-specific disease dynamics with latent flow matching for longitudinal imaging generation. *arXiv preprint arXiv:2512.09185*, 2025.

Hao Chen, Rui Yin, Yifan Chen, Qi Chen, and Chao Li. Learning patient-specific disease dynamics with latent flow matching for longitudinal imaging generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 82707–82723, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/859c3ab5fe31e8d1c1a70a34fe03dde9-Paper-Conference.pdf.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Hyuna Cho, Ziquan Wei, Seungjoo Lee, Tingting Dan, Guorong Wu, and Won Hwa Kim. Conditional diffusion with ordinal regression: Longitudinal data generation for neurodegenerative disease studies. In *The Thirteenth International Conference on Learning Representations*, 2025.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Alexandra Dagley, Meghan LaPoint, Willem Huijbers, Trey Hedden, Donald G. McLaren, Jasmeer P. Chatwal, Kathryn V. Papp, Rebecca E. Amariglio, Deborah Blacker, Dorene M. Rentz, Keith A. Johnson, Reisa A. Sperling, and Aaron P. Schultz. Harvard aging brain study: Dataset and accessibility. *NeuroImage*, 144(Pt B):255–258, January 2017. doi: 10.1016/j.neuroimage.2015.03.069.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. *Advances in neural information processing systems*, 32, 2019.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. *Advances in neural information processing systems*, 32, 2019.

Kathryn A. Ellis, Ashley I. Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T. Lautenschlager, Nat Lenzo, Ralph N. Martins, Paul Maruff, Colin Masters, Andrew Milner, Kerryn Pike, Christopher Rowe, Greg Savage, Cassandra Szoeke, Kevin Taddei, Victor Villemagne, Michael Woodward, and David Ames. 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. doi: 10.1017/S1041610209009405. URL https://pubmed.ncbi.nlm.nih.gov/19470201/.

Luigi Ferrucci. The Baltimore Longitudinal Study of Aging (BLSA): A 50-year-long journey and plans for the future. *The Journals of Gerontology: Series A: Biological Sciences and Medical Sciences*, 63(12):1416–1419, 2008. doi: 10.1093/gerona/63.12.1416.

Gary D. Friedman, Gary R. Cutter, Richard P. Donahue, Glenn H. Hughes, Stephen B. Hulley, David R. Jacobs, Kiang Liu, and Peter J. Savage. Cardia: study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology*, 41(11):1105–1116, 1988. ISSN 0895-4356. doi: https://doi.org/10.1016/0895-4356(88)90080-7. URL https://www.sciencedirect.com/science/article/pii/0895435688900807.

Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, M Jorge Cardoso, Marc Modat, Sébastien Ourselin, Lauge Sørensen, Alzheimer's Disease Neuroimaging Initiative, et al. Training recurrent neural networks robust to incomplete data: Application to alzheimer's disease progression modeling. *Medical image analysis*, 53:39–46, 2019.

Samuel Gruffaz, Pierre-Emmanuel Poulet, Etienne Maheux, Bruno Jedynak, and Stanley DURRLEMAN. Learning Riemannian metric for disease progression modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23780–23792. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c7b90b0fc23725f299b47c5224e6ec0d-Paper.pdf.

Mohamad Habes, Raymond Pomponio, Haochang Shou, Jimit Doshi, Elizabeth Mamourian, Guray Erus, Ilya Nasrallah, Lenore J Launer, Tanweer Rashid, Murat Bilgel, et al. The Brain Chart of Aging: machine-learning analytics reveals links between brain aging, white matter disease, amyloid burden, and cognition in the iSTAGING consortium of 10,216 harmonized MR scans. *Alzheimer's & Dementia*, 17(1):89–102, 2021.

Sungmin Hong, James Fishbaugh, Jason J. Wolff, Martin A. Styner, and Guido Gerig. Hierarchical multi-geodesic model for longitudinal analysis of temporal trajectories of anatomical shape and covariates. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*, volume 11767, pp. 57–65. Springer, Oct 2019. doi: 10.1007/978-3-030-32251-9_7. URL https://doi.org/10.1007/978-3-030-32251-9_7.

Sterling C Johnson, Rebecca L Koscik, Erin M Jonaitis, Lindsay R Clark, Kimberly D Mueller, Sara E Berman, Barbara B Bendlin, Corinne D Engelman, Ozioma C Okonkwo, Kirk J Hogan, et al. The Wisconsin Registry for Alzheimer's Prevention: a review of findings and current directions. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:130–142, 2018.

Boah Kim, Inhwa Han, and Jong Chul Ye. Diffusemorph: Unsupervised deformable image registration using diffusion model. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 347–364, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19821-2.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with gaussian process priors for probabilistic time series forecasting. *arXiv preprint arXiv:2410.03024*, 2024.

Igor Koval, Alexandre Bône, Maxime Louis, Thomas Lartigue, Simona Bottani, Arnaud Marcoux, Jorge Samper-González, Ninon Burgos, Benjamin Charlier, Anne Bertrand, et al. AD Course Map charts Alzheimer's disease progression. *Scientific Reports*, 2021a. Available at Inria - Institut national de recherche en sciences et technologies du numérique (HAL).

Igor Koval, Alexandre Bône, Maxime Louis, Thomas Lartigue, Simona Bottani, Arnaud Marcoux, Jorge Samper-González, Ninon Burgos, Benjamin Charlier, Anne Bertrand, Stéphane Epelbaum, Olivier Colliot, Stéphanie Allassonnière, and Stanley Durrleman. AD Course Map charts Alzheimer's disease progression. *Scientific Reports*, 11(1):8020, 2021b. ISSN 2045-2322. doi: 10.1038/s41598-021-87434-1. URL https://doi.org/10.1038/s41598-021-87434-1.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Zachary C Lipton, David C Kale, Randall Wetzel, et al. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare*, 56(56):253–270, 2016.

Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Ravì, and Sebastiano Battiato. Tadm: Temporally-aware diffusion model for neurodegenerative progression on brain mri. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 444–453, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72069-7.

Chen Liu, Ke Xu, Liangbo L. Shen, Guillaume Huguet, Zilong Wang, Alexander Tong, Danilo Bzdok, Jay Stewart, Jay C. Wang, Lucian V. Del Priore, and Smita Krishnaswamy. Imageflownet: Forecasting multiscale image-level trajectories of disease progression with irregularly-sampled longitudinal medical images. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890535.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Marco Lorenzi, Maurizio Filippone, Giovanni B. Frisoni, Daniel C. Alexander, and Sebastien Ourselin. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease. *NeuroImage*, 190:56–68, April 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2017.08.059. URL http://dx.doi.org/10.1016/j.neuroimage.2017.08.059.

Etienne Maheux, Igor Koval, Juliette Ortholand, Colin Birkenbihl, Damiano Archetti, Vincent Bouteloup, Stéphane Epelbaum, Carole Dufouil, Martin Hofmann-Apitius, and Stanley Durrleman. Forecasting individual progression trajectories in Alzheimer's disease. *Nature Communications*, 14(1):1–15, 2023. doi: 10.1038/s41467-022-35712-5.

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, 12 2010. ISSN 0898-929X. doi: 10.1162/jocn.2009.21407. URL https://doi.org/10.1162/jocn.2009.21407.

Razvan V. Marinescu, Neil P. Oxtoby, Alexandra L. Young, Esther E. Bron, Arthur W. Toga, Michael W. Weiner, Frederik Barkhof, Nick C. Fox, Stefan Klein, Daniel C. Alexander, and the EuroPOND Consortium. Tadpole challenge: Prediction of longitudinal evolution in alzheimer's disease, 2018. URL https://arxiv.org/abs/1805.03909.

Minh Nguyen, Nanbo Sun, Daniel C Alexander, Jiashi Feng, and BT Thomas Yeo. Modeling alzheimer's disease progression using deep recurrent neural networks. In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pp. 1–4. IEEE, 2018.

Minh Nguyen, Tong He, Lijun An, Daniel C. Alexander, Jiashi Feng, and B.T. Thomas Yeo. Predicting alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, 222:117203, 2020. doi: 10.1016/j.neuroimage.2020.117203.

YongKyung Oh, Dong-Young Lim, and Sungil Kim. Stable neural stochastic differential equations in analyzing irregular time series data. *arXiv preprint arXiv:2402.14989*, 2024.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Guilherme Pombo, Robert J. Gray, Manuel Jorge Cardoso, Sébastien Ourselin, Geraint Rees, John Ashburner, and Parashkev Nachev. Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models. *CoRR*, abs/2111.14923, 2021. URL https://arxiv.org/abs/2111.14923.

Lemuel Puglisi, Daniel C Alexander, Daniele Ravì, Alzheimer's Disease Neuroimaging Initiative, et al. Brain latent progression: Individual-based spatiotemporal disease progression on 3d brain mris via latent diffusion. *Medical image analysis*, pp. 103734, 2025.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew Dai, Nissan Hajaj, Peter Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Gavin Duggan, Gerardo Flores, Michaela Hardt, Jamie Irvine, Quoc Le, Kurt Litsch, Jake Marcus, Alexander Mossin, and Jeff Dean. Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*, 1, 01 2018. doi: 10.1038/s41746-018-0029-1.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International conference on machine learning*, pp. 8857–8868. PMLR, 2021.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Vasiliki Tassopoulou, Fanyang Yu, and Christos Davatzikos. Deep Kernel Learning with Temporal Gaussian Processes for Clinical Variable Prediction in Alzheimer's Disease. In Antonio Parziale, Monica Agrawal, Shalmali Joshi, Irene Y. Chen, Shengpu Tang, Luis Oala, and Adarsh Subbaswamy (eds.), *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pp. 539–551. PMLR, 28 Nov 2022. URL https://proceedings.mlr.press/v193/tassopoulou22a.html.

Vasiliki Tassopoulou, Haochang Shou, and Christos Davatzikos. Adaptive shrinkage estimation for personalized deep kernel regression in modeling brain trajectories. *arXiv preprint arXiv:2504.08840*, 2025.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Jennifer Tremblay-Mercier, Cécile Madjar, Samir Das, Alexa Pichet Binette, Stephanie O. M. Dyke, Pierre Étienne, Marie-Elyse Lafaille-Magnan, Jordana Remz, Pierre Bellec, D. Louis Collins, M. Natasha Rajah, Veronique Bohbot, Jeannie-Marie Leoutsakos, Yasser Iturria-Medina, Justin Kat, Richard D. Hoge, Serge Gauthier, Christine L. Tardif, M. Mallar Chakravarty, Jean-Baptiste Poline, Pedro Rosa-Neto, Alan C. Evans, Sylvia Villeneuve, Judes Poirier, John C. S. Breitner, and PREVENT-AD Research Group. Open science datasets from PREVENT-AD, a longitudinal cohort of pre-symptomatic Alzheimer's disease. *NeuroImage: Clinical*, 31:102733, 2021. doi: 10.1016/j.nicl.2021.102733.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep kernel learning, 2015. URL https://arxiv.org/abs/1511.02222.

Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In Alejandro Frangi, Marleen de Bruijne, Demian Wassermann, and Nassir Navab (eds.), *Information Processing in Medical Imaging*, pp. 388–400, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-34048-2.

Xi Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis L Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. *Advances in Neural Information Processing Systems*, 37:107198–107224, 2024.

TECHNICAL APPENDICES

A RESIDUAL NORMALIZATION

In implementation, residual statistics are computed using only the training trajectories within each fold. Let

$$\mu_\Delta = \mathbb{E}_{\text{train}}[d_k], \qquad \sigma_\Delta^2 = \text{Var}_{\text{train}}[d_k],$$

where the mean and variance are computed per ROI over consecutive imputed monthly states. During training, the residual target is centered as

$$\bar{d}_k = d_k - \mu_\Delta,$$

and the Gaussian source distribution is matched to the empirical residual scale,

$$z_0 \sim \mathcal{N}(0, \text{diag}(\sigma_\Delta^2)).$$

The flow is trained on the centered residual target $\bar{d}_k$ using the same objective as in Section 4.2, with d_k replaced by $\bar{d}_k$. At inference, the predicted centered residual $\hat{\bar{d}}_k$ is transformed back before the autoregressive update:

$$\hat{x}_{k+1} = \hat{x}_k + \hat{\bar{d}}_k + \mu_\Delta.$$

All normalization statistics are estimated from the training fold only and are fixed during validation and testing.

B PER-FOLD RESULTS OF MAIN BASELINES

We present the per-fold results in Table 3 of the main baselines from Table 1 for further clarity on individual performance. Performance is similar to the mean values, with no individual fold being overly favored and the proposed methodology having the best results.

Table 3: Per-fold MAE for the autoregressive methods. BRAINGENFLOW achieves the lowest error in every fold.

Method	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Mean
Deep Regression	0.3047	0.2976	0.3002	0.3023	0.2965	0.3003
RNN-AD	0.3017	0.2947	0.2960	0.2951	0.2927	0.2961
TFM-ODE	0.7265	0.7321	0.7355	0.6925	0.7167	0.7206
Endpoint-pair flow	0.2327	0.2299	0.2306	0.2393	0.2305	0.2326
BRAINGENFLOW (ours)	**0.2258**	**0.2117**	**0.2105**	**0.2230**	**0.2127**	**0.2167**

C IMPUTATION ABLATIONS

We present on Table 4 ablations on different imputation strategies. The standard RBF kernel underperforms compared to the alternative of the deep kernel, due to the deep kernel having realized the features related towards the covariates sufficiently bettter. Curiously, linear interpolation performs better than the standard GP, due to poor conditioning from the GP side.

Table 4: Sensitivity of downstream forecasting to the imputation method. The residual flow and all hyperparameters are held fixed; only the method used to construct the dense monthly training targets changes. MAE at acquired follow-up visits, mean and 95% CI across the five folds.

Dense targets from	Mean MAE (95% CI)
DKGP posterior mean (ours)	**0.216** [0.208, 0.224]
Linear interpolation	0.590 [0.196, 0.984]
Standard RBF-GP	1.030 [0.774, 1.286]

We also ablate the generated trajectories, while holding out certain visits. The residual at acquired visits measures how closely the posterior mean tracks observations it was fitted on, and does not establish that the imputed values between visits reflect the data rather than the prior. We therefore hold out acquired visits. For every subject with at least three acquired visits, we remove one internal visit across all 145 ROIs, refit the ROI-specific DKGPs without those measurements, and predict the removed visit at its exact acquisition time. Results across the five folds are reported in Table 5. Held-out correlation remains high and the mean signed residual is near zero, so the posterior mean does not shrink systematically toward the population trajectory. Predictive coverage is close to nominal, indicating that the posterior uncertainty is calibrated rather than overconfident at the visit spacings present in the data.

Table 5: Held-out intermediate-visit validation of the DKGP supervisory targets. For each subject with ≥ 3 acquired visits, one internal visit is removed across all 145 ROIs, the ROI-specific DKGPs are refit without it, and the removed visit is predicted at its acquisition time. Values are aggregated across the five folds.

Metric	Mean	Fold SD	95% CI
MAE (z-units)	0.1634	0.00075	[0.1624, 0.1643]
RMSE (z-units)	0.2530	0.00471	[0.2471, 0.2588]
Pearson correlation	0.9668	0.00110	[0.9655, 0.9682]
Mean signed residual	-0.0047	0.00100	$[-0.0060, -0.0035]$
95% predictive coverage	96.33%	0.28 pp	[95.99%, 96.68%]

D Datasets and Preprocessing

We curated a large, multi-cohort longitudinal dataset from the iSTAGING Consortium Habes et al. (2021), comprising 3,969 participants and 15,049 neuroimaging sessions across nine studies. The dataset includes neuroimaging, demographic, genetic, and diagnostic variables collected from heterogeneous aging and Alzheimer's disease cohorts. Neuroimaging features consist of 145 anatomical brain ROI volumes extracted using DL-MUSE, a deep learning-based segmentation framework Bashyam et al. (2025). These ROIs include 119 gray matter regions, 20 white matter regions, and 6 ventricular regions. Clinical and demographic covariates include age at baseline, sex, baseline diagnosis, and APOE-$\varepsilon 4$ allele status, the latter being a well-established genetic risk factor for Alzheimer's disease.

We use strict subject-level separation throughout the complete pipeline. For each fold: (1) per-ROI means and standard deviations are estimated from training subjects only; (2) deep feature mappings, GP kernel parameters, and likelihood parameters are learned only from training observations; (3) dense supervisory trajectories are generated using the GP models fitted within that training fold; (4) the fitted standardization parameters and GP models are applied unchanged to held-out subjects; and (5) the residual-flow model is trained only on the fold-specific training trajectories. No validation or test observation contributes to standardization, kernel learning, hyperparameter estimation, imputation-model fitting, or construction of training trajectories. The entire preprocessing, imputation, and forecasting pipeline is repeated independently for all five subject-level folds.

Diagnostic labels were harmonized across cohorts into three primary categories: cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease dementia (AD). Participants diagnosed with alternative dementias, including Lewy body dementia and frontotemporal dementia, were excluded. These exclusions represented a small fraction of the dataset and did not materially affect the overall sample size. Missing diagnostic labels were retained and encoded as unknown (UKN).

The nine cohorts span a broad range of demographic, clinical, and longitudinal sampling characteristics. The Alzheimer's Disease Neuroimaging Initiative (ADNI; $n = 1{,}576$; http://www.adni-info.org/) constitutes the largest subset, with an average of 4.9 visits per participant. Additional cohorts include the Baltimore Longitudinal Study of Aging (BLSA; $n = 675$) (Ferrucci, 2008), the Coronary Artery Risk Development in Young Adults study (CARDIA; $n = 479$) (Friedman et al., 1988), the Open Access Series of Imaging Studies (OASIS; $n = 549$) (Marcus et al., 2010), PENN ($n = 102$), PreventAD ($n = 271$) (Tremblay-Mercier et al., 2021), the Wisconsin Registry for Alzheimer's Prevention

Table 6: **Per-cohort descriptive statistics.** All demographic statistics are computed at baseline, defined as the earliest available visit per subject. Follow-up is reported as the median and maximum time between the first and last available visit, in months. Diagnosis: CN = cognitively normal, MCI = mild cognitive impairment, Dem. = dementia/AD. APOE4+: proportion of subjects carrying at least one $\varepsilon 4$ allele. WRAP (n=14) estimates should be interpreted with caution due to the small sample size.

Study	***N***	**Age (yrs)**				**Diagnosis at baseline (%)**			**Follow-up (mo)**	
		Mean	SD			CN	MCI	Dem.	Median	Max
ADNI	1576	73.8	7.2	45.7	47.1	32.7	49.2	18.1	28	178
AIBL	80	72.5	7.5	45.0	8.8	33.8	35.0	31.2	39	127
BLSA	675	68.5	13.0	55.1	24.3	98.7	0.6	0.7	68	160
CARDIA	479	50.4	3.5	52.2	27.1	100.0	0.0	0.0	58	73
HABS	223	73.4	6.2	59.2	28.3	99.6	0.4	0.0	37	46
OASIS	549	68.0	9.8	56.3	35.9	84.9	2.5	12.6	54	156
PENN	102	70.6	9.1	53.9	30.4	28.4	37.3	34.3	12	110
PREVENTAD	271	63.8	5.1	72.0	37.6	100.0	0.0	0.0	37	73
WRAP[†]	14	61.0	5.9	71.4	35.7	92.9	7.1	0.0	68	164
All	**3,969**	—		—	—	—	—	—	—	—

[†] WRAP comprises only n=14 subjects; estimates should be interpreted with caution.

(WRAP; $n = 14$) (Johnson et al., 2018), the Harvard Aging Brain Study (HABS; $n = 223$) (Dagley et al., 2017), and the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL; $n = 80$) (Ellis et al., 2009). Across all cohorts, participants contributed an average of 3.7 longitudinal visits, enabling modeling of individualized brain trajectories under irregular follow-up.

For downstream analyses, we further define a *CN Stable* subgroup as participants who remained cognitively normal across all available follow-up visits. This group is distinguished from participants who were CN at baseline but subsequently converted to MCI or AD during the observation period. Cohort-level demographic and clinical summaries are provided in Table 6.

E Temporal Interpolation

E.1 DKGP Implementation Details

This section provides the implementation details for the per-ROI deep kernel Gaussian process regressors described in Section 4.1. All models are implemented in `GPyTorch`.

Each ROI biomarker is standardized independently to zero mean and unit variance using statistics computed on the training set of the corresponding fold. Concretely, for ROI d we compute the empirical mean m_d and standard deviation s_d over all observations $\{y_d^{(i,j)}\}$ in the training fold, and replace each observation with $(y_d^{(i,j)} - m_d)/s_d$. Per-ROI standardization places all biomarkers on a common scale, which (i) makes the homoscedastic noise variance σ_d^2 comparable across ROIs, (ii) prevents ROIs with larger absolute volumes from dominating the ARD lengthscale optimization, and (iii) stabilizes training of the shared optimizer hyperparameters across the 145 independent models.

For each ROI d, the transformation $\Phi_d : \mathbb{R}^{\dim(\tilde{x})} \to \mathbb{R}^r$ is a single fully connected layer of width $r = \lfloor \dim(\tilde{x})/2 \rfloor$ with ReLU activation and dropout (rate 0.1). The output is rescaled to $[-1, 1]$ component-wise before entering the kernel, which stabilizes the RBF lengthscales during optimization. The covariance uses an ARD-RBF kernel with one lengthscale per latent dimension, wrapped in a scale kernel to learn the output variance. The mean function μ_d is linear in the latent embedding $\Phi_d(\tilde{x}^{(i)}(t))$. Observation noise is modeled by a homoscedastic Gaussian likelihood with learned variance σ_d^2.

Each per-ROI model is trained as an exact GP by maximizing the log marginal likelihood

$$\log p\Big(\mathbf{y}_d \mid \tilde{X}\Big) = -\frac{1}{2}\,(\mathbf{y}_d - \boldsymbol{\mu}_d)^\top \big(K_d + \sigma_d^2 I\big)^{-1}(\mathbf{y}_d - \boldsymbol{\mu}_d) - \frac{1}{2}\log\big|K_d + \sigma_d^2 I\big| - \frac{N}{2}\log 2\pi, \quad (12)$$

where $\mathbf{y}_d$ stacks the standardized training observations for ROI d across all subjects, $\tilde{X}$ stacks the corresponding context vectors, K_d is the kernel matrix with entries $[K_d]_{ab} = k_d(\tilde{x}_a, \tilde{x}_b)$, and $\boldsymbol{\mu}_d$ stacks the linear mean evaluations. We optimize Φ_d, the mean weights, the ARD lengthscales, the output scale, and σ_d^2 jointly with Adam (learning rate 0.02, weight decay 0.01) for 300 iterations in full-batch mode.

At inference, for each subject i, the input vector $\tilde{x}^{(i)}(t)$ is constructed at every integer month $t \in \{0, 1, \ldots, n\}$ with $n = \lfloor t_{m_i}^{(i)} \rfloor$, and the posterior mean $\mathbb{E}[f_d^{(i)}(t) \mid \mathbf{y}_d]$ is recorded. The predictions are then inverse-standardized via $s_d \cdot \mathbb{E}[f_d^{(i)}(t) \mid \mathbf{y}_d] + m_d$ and stacked across ROIs to produce the dense state $x_k^{(i)} \in \mathbb{R}^D$.

We perform the imputation process using within-training data alone on each fold, ensuring no leakage between training and test subjects in the imputation stage. Subjects are partitioned into five disjoint folds by PTID, and for each fold we train one DKGP per ROI on the corresponding training subjects only (totaling 145×5 models). The resulting imputed trajectories are then used as input to all downstream training and evaluation of the residual flow matching model.

E.2 Examples of Imputed Trajectories

The DKGP imputation stage of Section 4.1 converts each subject's sparse, irregularly sampled visits into a dense monthly trajectory that supervises the downstream flow-matching model. To verify that this stage produces trajectories faithful to the original observations, we visualize the DKGP posterior mean alongside the acquired measurements for twelve representative subjects from the left hippocampus (MUSE ROI 48), split across two panels: Figure 4 shows six densely-acquired subjects (9-14 visits each), and Figure 5 shows six sparsely-acquired subjects (2-3 visits each), reflecting the realistic distribution of follow-up in the cohort, where the median subject has only three acquired scans. Across all $3{,}175$ subjects with sufficient longitudinal coverage to fit a GP, the mean absolute residual between imputed and observed values at the original visit times is 0.091 z-units (RMSE 0.141, mean residual -0.003).

F Details on BrainGenFlow

Implementation. The GRU history encoder uses 4 layers with hidden dimension 128 and dropout 0.2; the flow network uses 3 FiLM-conditioned residual blocks with latent dimension 256 and dropout 0.1; the sinusoidal time embedding has dimension 64. Per-ROI residual statistics $\mu_\Delta, \sigma_\Delta$ are computed from consecutive imputed visits in the training fold and used to center the target and scale the source distribution $z_0 \sim \mathcal{N}(0, \mathrm{diag}(\sigma_\Delta^2))$. To improve robustness to autoregressive drift, we inject Gaussian noise $\tilde{x}_k = x_k + 0.5\,\sigma_\Delta \odot \epsilon$ into the conditioning state during training.

Training. We train for up to 100 epochs with AdamW (lr 10^{-3}, weight decay 10^{-4}), batch size 64, and gradient clipping at norm 1.0. The learning rate is halved after 8 epochs without validation improvement, with early stopping at patience 25. At inference, we integrate the velocity field with a 20-step RK4 solver over a 120-month horizon. Table 7 reports measured cost for each stage on a single NVIDIA RTX A6000. The imputation figure is the aggregate over all five folds, not a per-fold value, and the residual-flow timing was collected separately, so the two are reported over their own scopes rather than summed. Because the 145 regional GPs are independent, the imputation stage

Table 7: Measured computational cost on a single NVIDIA RTX A6000.

Stage	Scope	Wall-clock	Peak memory
DKGP fitting and dense-trajectory generation (145 ROIs)	All five folds	4.53 h (16,293 s)	RAM 13.87 GiB; GPU 15.84 GiB allocated, 17.70 GiB reserved
Residual-flow training	One run, 100 epochs	1.6 s/epoch (≈2.7 min)	GPU ≈8.5 GiB
Inference, 120-month rollout (5 samples per subject)	One held-out fold	43.5 s total (0.055 s per subject)	—

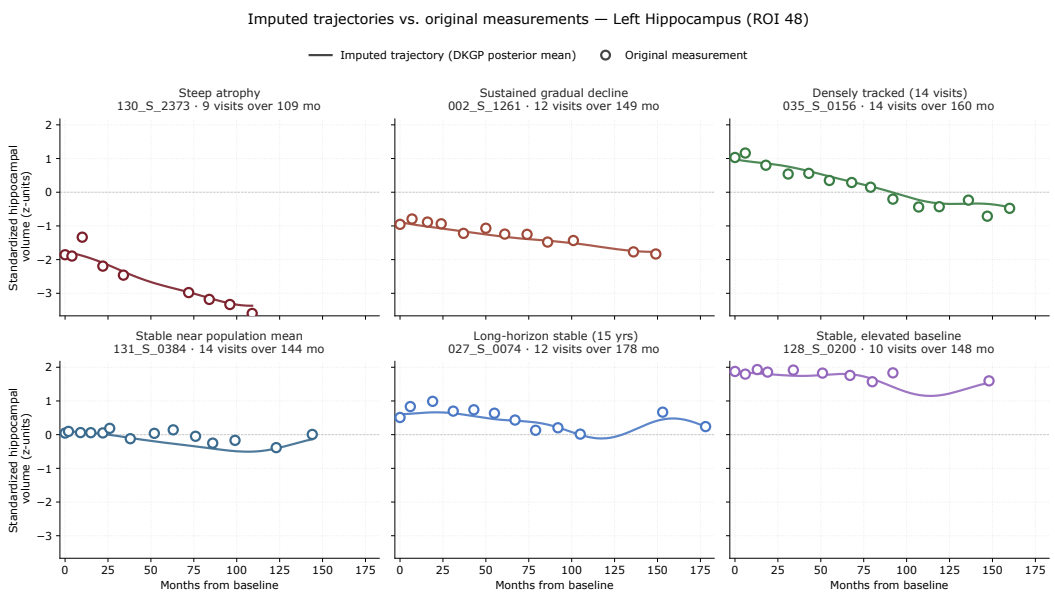

Figure 4: **DKGP imputation under dense longitudinal acquisition.** Six representative subjects from the left hippocampus (MUSE ROI 48), each with 9-14 visits over 109-178 months. Solid lines show the DKGP posterior mean (one value per month); open circles mark the original sparse acquisitions. Each panel reports the subject identifier, visit count, and observation horizon. All panels share the same axes; the dashed horizontal line marks the population mean ($z = 0$). The posterior mean traces the observed acquisitions without overshoot or oscillation across the full range of trajectory archetypes (steep atrophy, sustained gradual decline, near stability) and baseline volumes (-3 to $+2$ z-units).

parallelizes across ROIs and GPUs; we report the directly measured sequential runtime rather than an extrapolated parallel estimate.

G Comparison with Competing Baselines

G.1 Deep Regression

Deep regression is a baseline model that formulates longitudinal brain volume forecasting as a direct input-output mapping. Rather than modeling temporal dynamics sequentially, it takes a subject's baseline scan, clinical covariates, and a target time as a fixed-length input vector and predicts the full volumetric state at that time in a single forward pass. The model makes no use of intermediate visit observations at either training or inference time, serving as a strong non-sequential baseline that quantifies how much of the predictive signal can be recovered from the baseline scan alone, without any recurrent or generative mechanism.

For subject i and target time t^*, the input vector is

$$v^{(i)}(t^*) = \left[x_0^{(i)} \; ; \; a^{(i)} \; ; \; s^{(i)} \; ; \; \mathrm{dx}^{(i)} \; ; \; \mathrm{apoe}^{(i)} \; ; \; \tau^*\right] \in \mathbb{R}^{151}, \tag{13}$$

where $x_0^{(i)} \in \mathbb{R}^{145}$ is the baseline volumetric state, $a^{(i)}$ is age at baseline, $s^{(i)}$ is sex, $\mathrm{dx}^{(i)}$ is baseline diagnosis, $\mathrm{apoe}^{(i)}$ is APOE-ε4 allele count, and $\tau^* = t^*/T_{\max}$ is the target time normalized by the maximum observed visit time in the training fold. The model outputs $\hat{x}_{t^*}^{(i)} \in \mathbb{R}^{145}$, the predicted volumetric state at time t^*.

The MLP is a fully connected feedforward network,

$$z_\ell = \sigma\big(W_\ell \, z_{\ell-1} + b_\ell\big), \quad \ell = 1, \ldots, L, \qquad \hat{x}_{t^*}^{(i)} = W_{L+1} \, z_L + b_{L+1}, \tag{14}$$

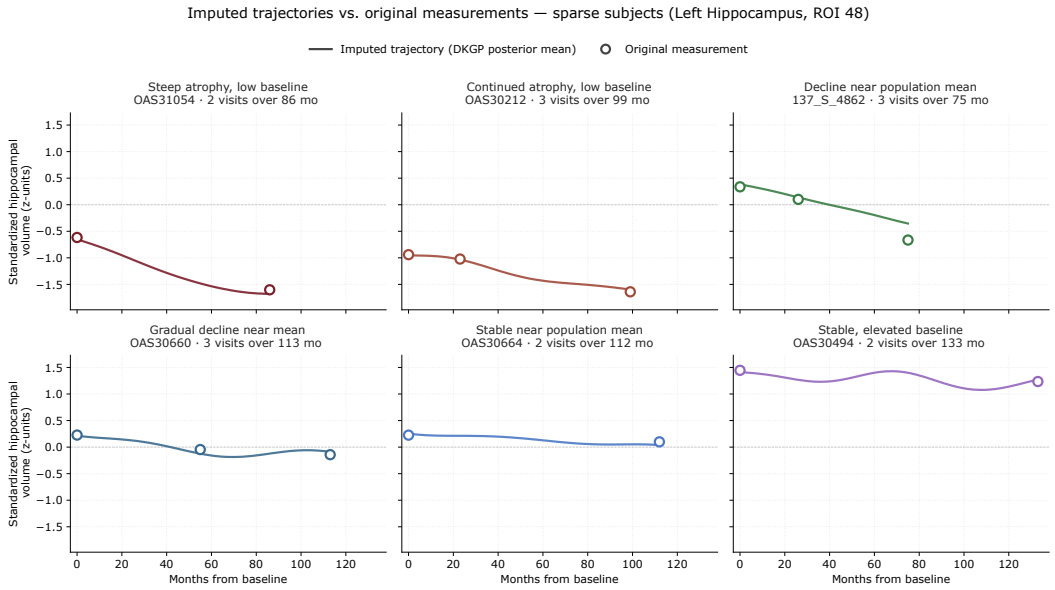

Figure 5: **DKGP imputation under sparsity.** Six subjects with only two or three acquired visits, shown in the same per-panel format as Figure 4. Despite the sparse supervision, the DKGP posterior mean connects the observed acquisitions without sharp transitions or aggressive extrapolation; declining subjects yield monotonic decline and stable subjects yield near-flat trajectories. Mild oscillations visible in the most sparsely-sampled panels reflect the increased contribution of the GP prior when data is limited.

where σ denotes ReLU activation, $z_0 = v^{(i)}(t^*)$, and the final layer is a linear readout to $\mathbb{R}^{145}$. Dropout of rate p is applied after each hidden activation. We use L=3 hidden layers of width 512, p=0.3, matching the parameter scale of the RNN-AD baseline.

Each subject-visit pair (i, j) with $j > 0$ defines one training example: the input is $v^{(i)}(t_j^{(i)})$, constructed using the baseline scan and the visit time τ_j, and the target is the observed volumetric state $x_{t_j}^{(i)}$ at that visit. The model is therefore trained on all observed follow-up visits across subjects, regardless of their temporal spacing. The training loss is the mean squared error over all training pairs,

$$\mathcal{L}_{\text{MLP}} = \frac{1}{N_{\text{pairs}}} \sum_{i=1}^{N} \sum_{j=1}^{m_i - 1} \left\| \hat{x}_{t_j}^{(i)} - x_{t_j}^{(i)} \right\|_2^2. \tag{15}$$

Optimization uses Adam (Kingma & Ba, 2014) with learning rate 10^{-3}, weight decay 10^{-4}, and batch size 64. Training runs for up to 200 epochs with early stopping on validation MSE (patience 20 epochs); the best checkpoint is restored before evaluation.

At inference, the target time τ^* is varied across the 120-month horizon while the baseline input $x_0^{(i)}$ and clinical covariates remain fixed. Each desired forecast timepoint is predicted independently via a single forward pass through the network, yielding an individualized trajectory with no autoregressive rollout and no accumulation of prediction error across time steps.

The Deep Regression follows the same 5-fold subject-level cross-validation scheme as all other methods (Section D), with a 15% validation holdout per fold for early stopping. This ensures that no subject's baseline or follow-up visits appear in both training and evaluation sets for any fold.

G.2 RNN-AD

RNN-AD is a sequence-to-sequence recurrent baseline for longitudinal brain volume forecasting, inspired by recurrent architectures proposed for Alzheimer's disease progression modelling Nguyen et al. (2020); Ghazi et al. (2019). RNN-AD operates directly on the sparse, irregularly sampled longitudinal observations. Irregular visit timing is handled implicitly by encoding the elapsed time of each visit as part of the input vector, allowing the recurrent network to internalize the temporal gap structure without an external imputation step. At inference, a single baseline scan is sufficient to produce a full individualized trajectory over the 120-month horizon.

Subject i contributes m_i visits at observed times $\mathcal{T}_i = \{t_1^{(i)}, \ldots, t_{m_i}^{(i)}\}$. At each visit j, the input vector is

$$u_j^{(i)} = \left[x_{t_j}^{(i)} \;;\; a^{(i)} \;;\; s^{(i)} \;;\; \mathrm{dx}^{(i)} \;;\; \mathrm{apoe}^{(i)} \;;\; \tau_j\right] \in \mathbb{R}^{151}, \tag{16}$$

where $x_{t_j}^{(i)} \in \mathbb{R}^{145}$ is the observed volumetric state at visit j, $a^{(i)}$ is age at baseline, $s^{(i)}$ is sex, $\mathrm{dx}^{(i)}$ is baseline diagnosis, $\mathrm{apoe}^{(i)}$ is APOE-ε4 allele count, and $\tau_j = t_j^{(i)}/T_{\max}$ is the visit time normalized by the maximum observed time in the training fold. The normalized time τ_j serves as the model's internal mechanism for handling irregular and arbitrarily large inter-visit gaps: the recurrent state transitions between consecutive observations are informed by how far apart in time those observations actually are, without any interpolation of missing months. Volumes are standardized per ROI using training-fold statistics.

The model is an 2-layer stacked LSTM with hidden dimension H:

$$h_j = \mathrm{LSTM}_\theta\big(u_j^{(i)},\, h_{j-1}\big), \qquad \hat{x}_{t_{j+1}}^{(i)} = W_o\, h_j + b_o, \tag{17}$$

where $W_o \in \mathbb{R}^{145\times H}$ and $b_o \in \mathbb{R}^{145}$ form a linear readout that maps the recurrent hidden state to the predicted volumetric state at the next visit. Dropout of rate p is applied between recurrent layers. We use the default hyperparameters H=256, L=2, p=0.3.

The model is trained with teacher forcing on the raw visit sequences. At step j, the input $u_j^{(i)}$ uses the *observed* volumes $x_{t_j}^{(i)}$, and the target is the next observed state $x_{t_{j+1}}^{(i)}$. The training loss is the mean squared error over all subjects and visit transitions,

$$\mathcal{L}_{\text{RNN-AD}} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(m_i-1)D} \sum_{j=1}^{m_i-1} \left\| W_o h_j^{(i)} + b_o - x_{t_{j+1}}^{(i)} \right\|_2^2, \tag{18}$$

where N is the number of training subjects and D=145. Optimization uses Adam (Kingma & Ba, 2014) with learning rate 10^{-3}, weight decay 10^{-4}, and batch size 64. Training runs for up to 200 epochs with early stopping on validation MSE (patience 20 epochs); the best checkpoint by validation loss is restored before evaluation.

At inference, only the baseline scan $x_0^{(i)}$, clinical covariates, and the target time τ are provided. The model is unrolled autoregressively for 120 steps: at each step j, the predicted state $\hat{x}_{t_{j+1}}^{(i)}$ is concatenated with the fixed clinical covariates and the next normalized time index τ_{j+1} to form $u_{j+1}^{(i)}$, which is fed back into the LSTM. This autoregressive rollout generates a full individualized trajectory from a single baseline observation, mirroring the inference protocol of BRAINGENFLOW.

RNN-AD follows the same 5-fold subject-level cross-validation scheme as the main experiment (Section D). For each fold, a 15% holdout of training subjects is used for early stopping.

H ADDITIONAL EXPERIMENTAL RESULTS

H.1 PROGRESSION GROUP ANALYSIS: FULL STATISTICAL DETAILS

Cohort. We evaluate six progression groups defined by clinical diagnosis trajectories over the 60-month prediction horizon: cognitively normal stable (CN stable, n=2,488), CN converters to MCI (CN→MCI, n=121), CN converters to AD (CN→AD, n=74), MCI stable (n=531), MCI progressors (n=281), and stable AD (n=401). All subjects have an exact 60-month target timepoint and a real observed baseline scan. RoC is computed as monthly fractional volume change from the real baseline to the model-predicted brain at 60 months.

Global group differences. A Kruskal-Wallis test across all six groups reveals highly significant differences in mean predicted RoC (H=66.6, p=5.3×10^{-13}), median predicted RoC (H=65.2, p=1.0×10^{-12}), mean absolute RoC (H=59.2, p=1.8×10^{-11}), and the spatial standard deviation of RoC across ROIs (H=43.5, p=2.9×10^{-8}). In contrast, the fraction of declining ROIs is not significant (H=6.4, p=0.27), indicating that the groups differ in the *speed* of atrophy rather than its spatial extent: nearly all subjects exhibit predominantly declining volumes (≈94-96% of ROIs).

Group-level RoC ordering. Median whole-brain predicted RoC values are as follows: MCI Stable (-4.10×10^{-3} month^{-1}), CN Stable (-4.14×10^{-3} month^{-1}), CN→MCI (-4.17×10^{-3} month^{-1}), CN→AD (-4.18×10^{-3} month^{-1}), MCI Progressor (-4.48×10^{-3} month^{-1}), and AD (-4.86×10^{-3} month^{-1}). This ordering is clinically coherent and monotonic with disease severity. The spatial standard deviation of RoC also increases monotonically, confirming that advanced disease stages show greater regional vulnerability heterogeneity in addition to faster global atrophy.

Pairwise post-hoc comparisons. Pairwise Mann-Whitney U tests with Benjamini-Hochberg correction are reported in Table 8. All comparisons use the *median* predicted RoC per subject (median across 145 ROIs), consistent with the group-level values reported above. The strongest separation is between CN Stable and AD Stable (Cliff's δ=0.23, q=1.8×10^{-12}). Within-MCI discrimination is significant (δ=0.13, q=6.5×10^{-3}), demonstrating sensitivity to heterogeneity within a single diagnostic category. The null result for CN Stable vs. CN→MCI (δ=0.02, q=0.75) indicates that prodromal MCI conversion is not reflected in volumetric RoC at the 60-month horizon, consistent with evidence that amyloid accumulation precedes detectable structural atrophy.

Table 8: Pairwise Mann-Whitney U tests on *median* predicted RoC (median across 145 ROIs per subject, consistent with the group-level values reported in the text). q-values are Benjamini-Hochberg corrected. Cliff's $\delta > 0$ means group A atrophies more slowly than group B. Medians in units of 10^{-3} month^{-1}.

Group A	Group B	Median$_A$	Median$_B$	Cliff's δ	q-value
CN stable	AD stable	−4.14	−4.86	+0.23	1.8×10^{-12}
MCI stable	AD stable	−4.10	−4.86	+0.23	1.2×10^{-8}
MCI stable	MCI progressor	−4.10	−4.48	+0.13	6.5×10^{-3}
CN stable	CN→MCI	−4.14	−4.17	+0.02	0.75 (n.s.)

ROI-level analysis. At the individual ROI level, 128 of 145 MUSE regions (88%) exhibit significant group differences after FDR correction (q<0.05). The five most discriminative regions are listed in Table 9. The top ROIs correspond to the entorhinal cortex, temporal association cortex, and hippocampus - regions with established early involvement in the AD neurodegeneration cascade Braak & Braak (1991), and consistent with the focal deep-red signal visible in the AD panel of Figure 2.

Table 9: Top 5 most discriminative ROIs by Kruskal-Wallis statistic across six progression groups. q-values are Benjamini-Hochberg corrected. Median RoC values in units of 10^{-3} month^{-1}.

MUSE ROI	Region	H	q-value	AD median	CN Stable median
186	Entorhinal cortex	148.8	3.5×10^{-28}	−3.72	−2.71
198	Temporal association ctx.	124.8	2.2×10^{-23}	−5.56	−4.37
76	Hippocampus	120.3	1.3×10^{-22}	−9.95	−8.11
196	Parahippocampal gyrus	119.1	1.8×10^{-22}	−6.16	−5.22
59	Lateral temporal cortex	113.3	2.4×10^{-21}	−4.72	−5.25

Group-averaged RoC volumes. Group-averaged predicted RoC maps are provided as NIfTI volumes for qualitative spatial inspection (see supplementary data release).

H.2 Error Analysis by Forecast Horizon and Subject-Level Covariates

We provide a detailed characterisation of in-fold prediction error as a function of forecast horizon and four subject-level covariates: baseline diagnosis, sex, age group, and clinical progression status

(first→last diagnosis). All results are derived from the 5-fold cross-validation experiment described in Section 5.1, using observations at month ≥ 1. Per-fold MAE is computed by averaging absolute errors across all 145 MUSE ROIs and all subjects within each fold; cross-fold means and 95% confidence intervals.

Error vs forecast horizon. Figure 6(a) shows the monthly MAE smoothed across folds alongside horizon-bin summaries. MAE increases monotonically from 0.166 (95% CI [0.159, 0.174]) at 0-12 months to 0.331 ([0.312, 0.350]) at 60-70 months. The rate of degradation is gradual up to 36 months and steepens thereafter: the 36-60 month bin (0.266) shows a 28% increase over the 24-36 bin (0.208). The cross-fold standard deviation remains small throughout (ranging from 0.006 to 0.016), confirming that this degradation pattern is stable and consistent across all data splits.

Error by baseline diagnosis. Table 10 reports horizon-stratified MAE by baseline diagnostic group. CN subjects consistently show the lowest error across all horizons, while AD subjects show the highest, particularly at longer horizons. At 36-60 months, the gap between CN (0.250, [0.236, 0.263]) and AD (0.380, [0.316, 0.444]) reaches 0.130 - a 52% relative difference. MCI subjects are intermediate at all horizons. The widening gap with horizon reflects the greater trajectory variability in clinical groups: AD subjects undergo faster and more heterogeneous atrophy, which is harder to predict from a single baseline scan as the horizon extends. The AD 60-120 bin (0.360) is estimated from only 3 folds due to limited long-term observations in this group and should be interpreted with caution.

Table 10: Horizon-stratified MAE by baseline diagnosis. Values are mean (95% CI) across folds. † AD 60-120 month estimate based on n=3 folds only.

Horizon	CN	MCI	AD
0-12	0.154 [0.141, 0.167]	0.170 [0.161, 0.180]	0.182 [0.171, 0.194]
12-24	0.158 [0.150, 0.167]	0.205 [0.195, 0.215]	0.216 [0.203, 0.229]
24-36	0.188 [0.177, 0.198]	0.243 [0.231, 0.256]	0.289 [0.255, 0.324]
36-60	0.250 [0.236, 0.263]	0.323 [0.298, 0.348]	0.380 [0.316, 0.444]
60-120	0.313 [0.297, 0.329]	0.398 [0.364, 0.431]	0.360† [0.252, 0.469]

Error by sex. Sex differences are small and consistent across all horizons (Table 11). Female subjects show slightly higher MAE than male subjects at short horizons (0-12: 0.180 vs 0.151; 12-24: 0.197 vs 0.173), with the gap narrowing at longer horizons (36-60: 0.269 vs 0.264; 60-120: 0.319 vs 0.341). The overall difference across all horizons is 0.236 (Female) vs 0.226 (Male), a relative difference of 4%. These results indicate that BRAINGENFLOWdoes not exhibit systematic sex bias in prediction error.

Table 11: Horizon-stratified MAE by sex. Values are mean (95% CI) across folds.

Horizon	Female	Male
0-12	0.180 [0.166, 0.195]	0.151 [0.142, 0.160]
12-24	0.197 [0.183, 0.211]	0.173 [0.167, 0.179]
24-36	0.216 [0.203, 0.229]	0.201 [0.193, 0.208]
36-60	0.269 [0.247, 0.290]	0.264 [0.250, 0.278]
60-120	0.319 [0.292, 0.346]	0.341 [0.321, 0.362]

Error by baseline age group. Age-stratified results (Table 12) reveal a non-monotonic pattern. The 60-70 age group achieves the lowest overall MAE (0.217), consistent with this group being the most heavily represented in the iSTAGING training pool and therefore best covered by the learned trajectory distribution. Subjects under 60 (0.224) and over 80 (0.237) show moderately higher error, reflecting sparser training coverage at the tails of the age distribution. The 70-80 group (0.243) shows the highest overall error likely because this decade coincides with peak trajectory heterogeneity in the prodromal AD population. These differences are small in absolute terms (range 0.217-0.243) and all within overlapping confidence intervals, suggesting the model generalises reasonably across the observed age spectrum.

Table 12: Horizon-stratified MAE by baseline age group. Values are mean (95% CI) across folds.

Horizon	<60	60-70	70-80	≥80
0-12	0.163 [0.130, 0.196]	0.155 [0.142, 0.167]	0.171 [0.150, 0.191]	0.173 [0.159, 0.188]
12-24	0.169 [0.145, 0.194]	0.174 [0.156, 0.192]	0.196 [0.184, 0.207]	0.186 [0.175, 0.197]
24-36	0.205 [0.160, 0.250]	0.189 [0.173, 0.205]	0.218 [0.204, 0.231]	0.219 [0.206, 0.233]
36-60	0.275 [0.255, 0.296]	0.239 [0.225, 0.254]	0.283 [0.254, 0.311]	0.267 [0.258, 0.277]
60-120	0.308 [0.287, 0.329]	0.330 [0.288, 0.371]	0.350 [0.332, 0.367]	0.338 [0.304, 0.373]

Error by progression status. Progression status provides the most clinically informative stratification (Table 13 and Figure 6(b)). CN Stable subjects achieve the lowest overall MAE (0.210), confirming that stable, normal aging trajectories are the easiest to predict. MCI Progressors (0.302) and AD Stable (0.299) subjects show the highest error, driven primarily by long-horizon performance: MCI Progressors reach MAE = 0.477 [0.422, 0.531] at 60-120 months - the highest value across all subgroups and horizons. This is clinically expected: a subject presenting as MCI at baseline but progressing to AD by the target timepoint undergoes a trajectory that deviates substantially from what would be inferred from their baseline scan alone, making long-range prediction fundamentally harder. CN→AD subjects show a similar pattern (0.353 at 60-120 months) for the same reason. Conversely, CN→MCI subjects show near-CN error across all horizons (0.218 overall), consistent with the group-level null result reported in Section 5.2: structural atrophy in prodromal MCI is not yet detectable from baseline volumetrics alone.

Table 13: Horizon-stratified MAE by progression status (first→last diagnosis). Values are mean (95% CI) across folds. Groups with fewer than 5 folds contributing are marked †.

Horizon	CN Stable	CN→MCI	CN→AD	MCI Stable	MCI Prog.	AD Stable
0-12	0.154	0.143	0.207†	0.170	0.177	0.183
12-24	0.156	0.170	0.198	0.203	0.212	0.217
24-36	0.185	0.195	0.242	0.233	0.264	0.297
36-60	0.245	0.267	0.350	0.285	0.379	0.390
60-120	0.310	0.313	0.353	0.345	0.477	0.406†
Overall	0.210	0.218	0.270	0.247	0.302	0.299

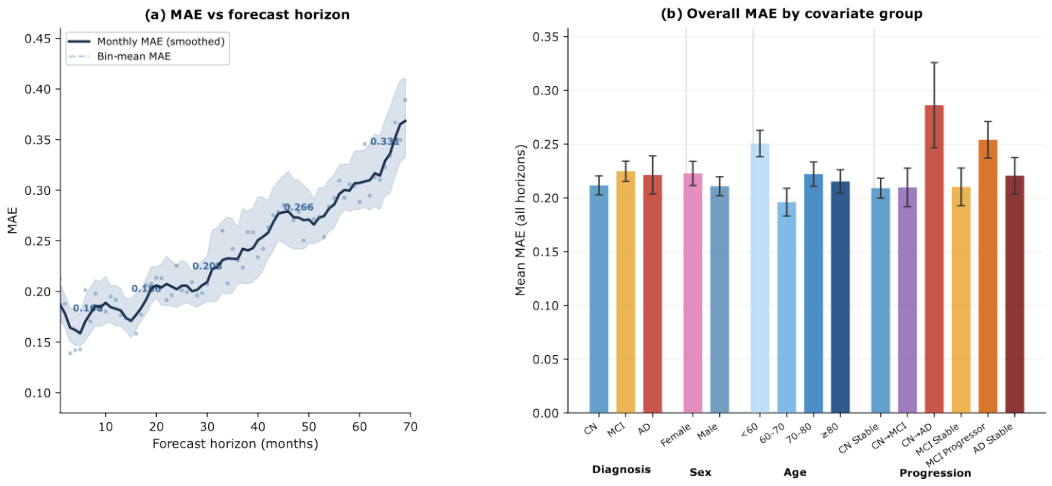

Figure 6: **In-fold prediction error by forecast horizon and subject-level covariates. (a)** Monthly MAE (dots) with smoothed mean ± 1 std across folds (shaded band) and per-bin mean (dashed lines). MAE increases monotonically from 0.166 at 0-12 months to 0.331 at 60-70 months. **(b)** Overall MAE (mean across all horizons) stratified by baseline diagnosis, sex, age group, and progression status, with 95% confidence intervals. Error bars reflect fold-level variability. MCI Progressors and AD Stable subjects show the highest error ($>$0.29), while CN Stable subjects show the lowest (0.210). Sex and age differences are small ($<$0.02) across all groups.

H.3 Per-Region Error Analysis

This subsection provides a per-region breakdown of BrainGenFlow's prediction error across the 145 ROIs, examines how error accumulates over the autoregressive rollout described in Section 4.2, and reports performance on regions that are clinically critical for AD progression. The goal is to characterize where the model is reliable, where it is not, and whether failure modes align with known biological variability rather than reflecting systematic deficiencies of the residual flow-matching parameterization.

Setup. We evaluate on the predictions for $3{,}969$ iSTAGING subjects across 70 months obtained through the 5 folds. For the analysis below we report errors in the standardized scale. By construction, the autoregressive rollout $\hat{x}_{k+1} = \hat{x}_k + \hat{d}_k$ is initialized at the observed baseline x_0, so $\mathrm{MAE}(t = 0) = 0$ for all regions and all subjects; the analysis below characterizes how error accumulates over subsequent monthly transitions. We summarize each ROI's performance with the mean absolute error $\mathrm{MAE}_r = \mathbb{E}[|\hat{y}_r - y_r|]$ and the root mean squared error RMSE_r, both reported in the normalized units defined above.

Aggregate performance. Averaged across the 145 ROIs, BrainGenFlow achieves a mean MAE of 0.150 z-units. The distribution of per-ROI MAEs is unimodal and right-skewed, ranging from 0.080 (left transverse temporal gyrus) to 0.282 (right inferior lateral ventricle); even the largest per-ROI MAE remains well below one within-region standard deviation, indicating that no single region is poorly modeled in absolute terms. The variation we discuss below should therefore be read as relative difficulty rather than failure.

Spatial distribution of error. Figure 7 ranks the easiest and hardest 20 regions. The hardest regions are dominated by the ventricular system-both lateral ventricles, both inferior lateral ventricles, and the third ventricle account for five of the ten worst MAE values (range 0.21-0.28). Among gray-matter regions, the temporal pole, superior temporal gyrus, superior frontal gyrus, supplementary motor cortex, parietal operculum, and basal forebrain are the most error-prone. The easiest regions are concentrated in primary sensorimotor and visual cortex (transverse temporal gyrus, precentral gyrus, occipital pole), in inferior frontal regions (orbital and triangular parts of the inferior frontal gyrus), and in the cerebellum. This pattern is consistent with the biological intuition that ventricular

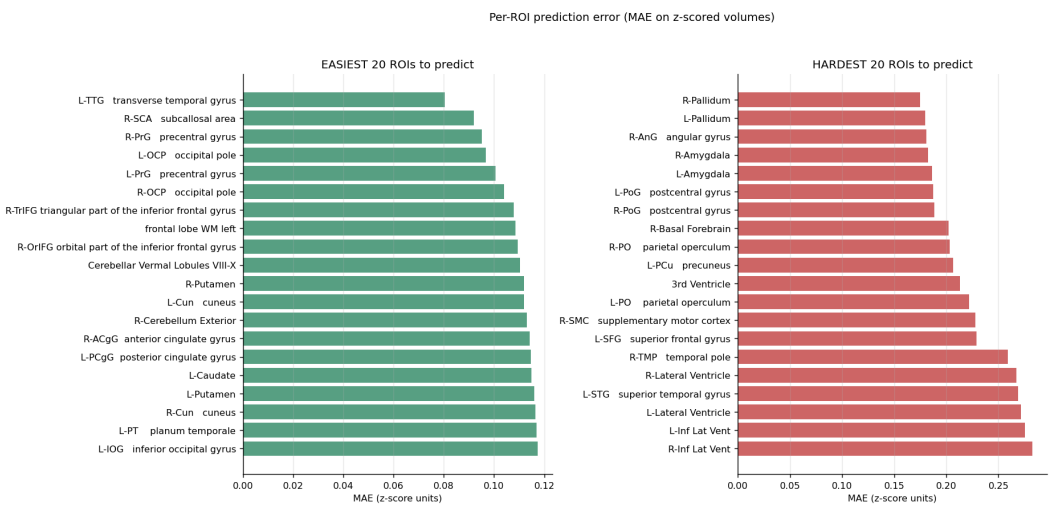

Figure 7: The 20 easiest (left) and 20 hardest (right) ROIs to predict, ranked by MAE in z-score units. Ventricular structures dominate the high-error end; primary sensorimotor and visual cortex dominate the low-error end.

volume exhibits high inter-subject variability and accelerated, non-linear expansion in aging and AD, whereas primary cortices and cerebellar regions remain comparatively stable across the cohort. The residual parameterization of Section 4.2 interacts with this pattern in an intuitive way: regions whose true monthly residuals are small and weakly structured (primary cortices) admit a near-zero residual prediction at low cost, whereas regions with large, systematically signed residuals (ventricular expansion, hippocampal atrophy) require the velocity field to capture both the magnitude and direction of change, and any per-step residual error compounds over the autoregressive rollout.

Figure 8 projects per-ROI MAE back onto the brain atlas. Two complementary views are shown: aggregate MAE across all months (top row) and MAE at the longest horizon $t = 69$ (bottom row). The aggregate map is spatially uniform and dominated by low-error regions, whereas the month-69 map highlights a clear periventricular and medial-temporal pattern, consistent with regions known to drive AD-related neurodegeneration.

Performance on clinically critical regions. Table 14 reports MAE for AD-relevant regions. Bilateral entorhinal cortex (MAE 0.119 R / 0.126 L) and fusiform gyrus (0.134 R / 0.141 L) are predicted most accurately; the hippocampus, while harder than these neighboring structures, retains comparable accuracy (0.161 R / 0.173 L), with the left hippocampus marginally harder than the right. The amygdala is moderately harder still (0.183 R / 0.187 L), and ventricular structures yield the largest errors (0.213-0.282). This ordering-entorhinal $<$ fusiform $<$ parahippocampal $<$ hippocampus $<$ amygdala $<$ ventricles-is consistent with the well-established hierarchy of inter-subject variability in these structures.

Temporal scaling of prediction error. Figure 9 shows that the cohort-averaged MAE grows approximately monotonically with prediction horizon, from 0 at $t = 0$ (by construction; the autoregressive rollout is initialized at x_0) to 0.39 z-units at $t = 69$. This growth is the empirical signature of the error accumulation that the static baseline embedding E_{baseline} of Section 4.3 is designed to mitigate; the fact that error remains bounded across the full five-year horizon-staying well below one within-region standard deviation for every ROI-suggests that the anchoring mechanism is partially effective. The growth is substantially heterogeneous across tissue types: gray and white matter errors grow gradually, while ventricular errors accelerate sharply after roughly two years. Figure 10 plots the trajectories for the AD-critical regions identified above. The inferior lateral and lateral ventricles exhibit the steepest growth (linear slopes of 0.0094-0.0102 z-units/month, computed over $t > 0$), roughly tripling their error from year one to year five. The hippocampus and amygdala grow more gradually (slopes of 0.003-0.005 z-units/month). At the opposite extreme, several white-matter and visual-cortex regions (lingual gyrus, occipital pole, calcarine cortex, frontal-lobe white matter) have negligible slopes

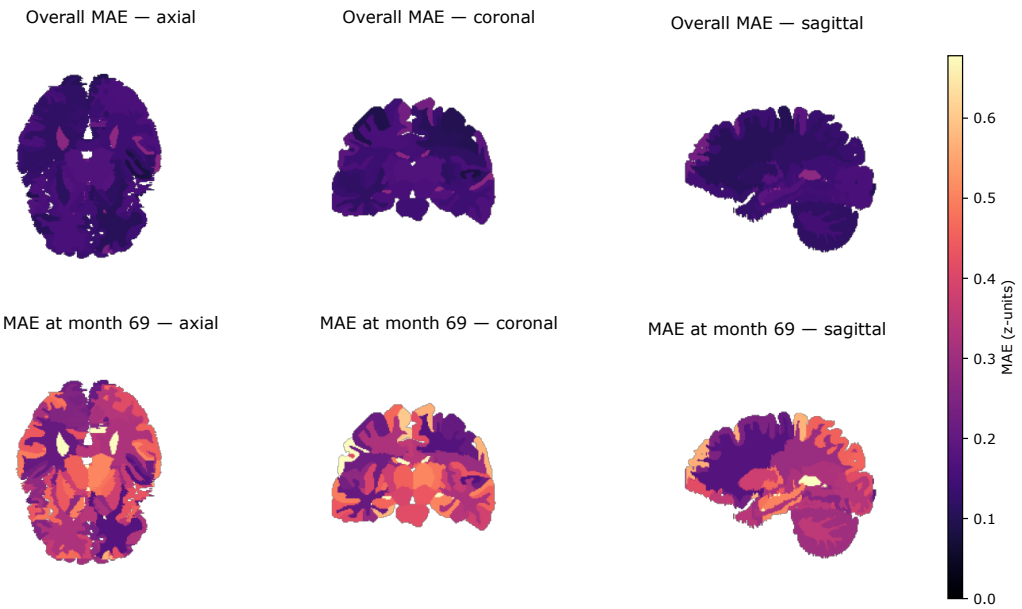

Figure 8: Per-ROI MAE projected onto the MUSE atlas, shown in axial, coronal, and sagittal views. Top row: MAE aggregated across all 70 months. Bottom row: MAE at the longest horizon ($t = 69$ months). The shared color scale highlights the substantial increase in error at long horizons, concentrated in the ventricular system and medial temporal lobe.

Table 14: Per-region MAE for clinically critical AD regions, sorted from easiest to hardest. MAE is reported in normalized units (residual-space z-units, see Section 4.2).

Region	Hemisphere	MAE
Entorhinal cortex	R	0.119
Entorhinal cortex	L	0.126
Fusiform gyrus	R	0.134
Fusiform gyrus	L	0.141
Fourth ventricle	B	0.150
Parahippocampal gyrus	R	0.157
Parahippocampal gyrus	L	0.158
Hippocampus	R	0.161
Hippocampus	L	0.173
Amygdala	R	0.183
Amygdala	L	0.187
Third ventricle	B	0.213
Lateral ventricle	R	0.267
Lateral ventricle	L	0.271
Inferior lateral ventricle	L	0.275
Inferior lateral ventricle	R	0.282

(< 0.001 z-units/month), suggesting that for these structures BRAINGENFLOW's longest-horizon predictions are essentially as reliable as its short-horizon predictions. We interpret this dichotomy as reflecting the interaction between the residual target distribution and the autoregressive rollout: regions with near-zero true residuals accumulate little compounding error, whereas regions with large, structured residuals magnify even small per-step velocity-field errors over 69 integration steps.

Discussion. Three findings stand out. First, the structures that are hardest to predict are precisely those that exhibit the greatest biological variability in this cohort, ventricles in particular, rather than structures of any specific anatomical category, suggesting that the residual flow-matching objective

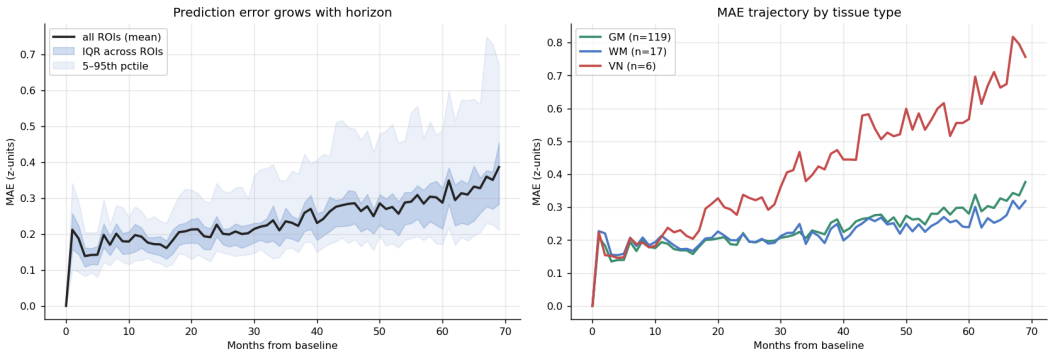

Figure 9: Cohort-averaged MAE versus prediction horizon. **Left:** overall MAE; the solid line is the cross-ROI mean, the darker shaded band is the interquartile range and the lighter band the 5-95% range across ROIs. **Right:** MAE trajectories stratified by tissue type. Ventricular regions (VN) accelerate sharply after $\sim$24 months, whereas gray and white matter grow gradually.

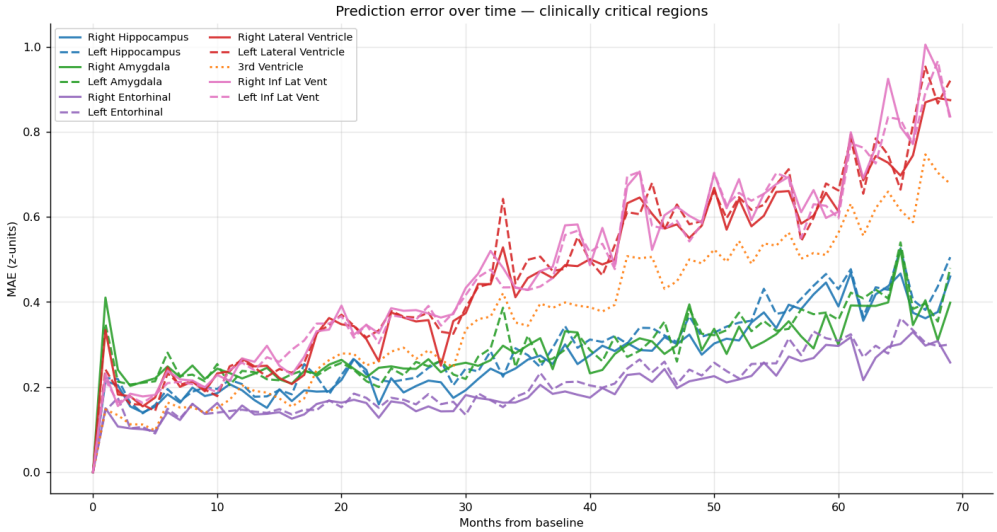

Figure 10: MAE trajectories for clinically critical AD regions. Solid and dashed lines denote right and left hemispheres respectively. Ventricular regions exhibit the steepest growth (slopes of 0.0094-0.0102 z-units/month over $t > 0$); entorhinal cortex is the most stable of the critical regions across the full horizon.

distributes its modeling capacity in proportion to the structure of the underlying dynamics. Second, BRAINGENFLOW degrades slowly with horizon: even at $t = 69$ months, every ROI's MAE remains well below one within-region standard deviation, and error growth is concentrated in regions where the underlying biology is itself accelerating. The combination of the recurrent history embedding h_k and the static baseline embedding E_{baseline} (Section 4.3) appears to bound compounding error effectively across the five-year horizon, although the per-region growth rates we report remain a useful diagnostic for identifying structures where the anchoring is least effective. Third, the AD-critical medial-temporal regions are predicted with MAEs in the 0.12-0.19 z-unit range, comparable to the cohort-wide mean of 0.150, suggesting that BRAINGENFLOW's accuracy on regions of clinical interest is not systematically worse than its accuracy elsewhere.

Limitations. Three caveats apply to the numbers reported above. (i) Ground-truth observations are sparse and non-uniformly distributed in time: while $t = 0$ has near-complete coverage ($n = 3{,}963$), several follow-up months have $n < 50$ subjects with acquired scans, which inflates the variance of the per-month MAE estimates and may produce spurious local fluctuations in Figures 9 and 10.

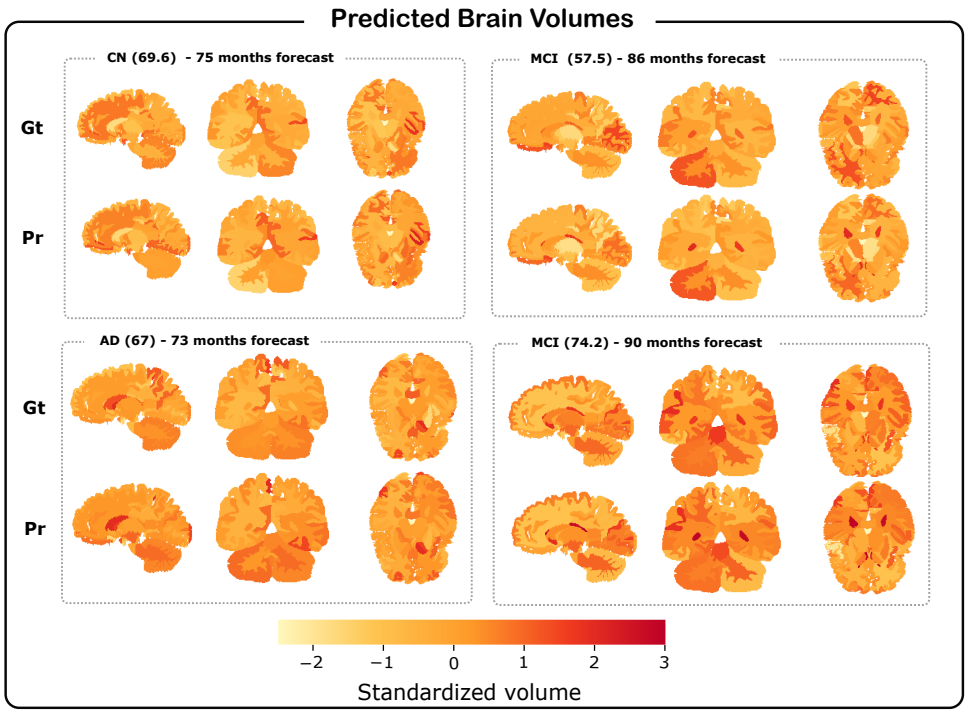

Figure 11: **Qualitative brain volume forecasts at arbitrary horizons.** Predicted (Pr) and ground-truth (Gt) standardized ROI volumes for four held-out subjects spanning the diagnostic spectrum (CN, MCI, AD) and both cohorts (ADNI, AIBL), shown in three orthogonal MUSE-atlas views (sagittal, coronal, axial). Each panel reports the subject's diagnostic group, baseline age (in parentheses), and the forecast horizon in months. BRAINGENFLOW generates month-resolved trajectories from a single baseline scan, so the horizons shown (73-90 months) are queried directly from the autoregressive rollout rather than chosen from a fixed grid. Color encodes standardized ROI volume (z-scored across the population); cooler tones (light yellow) indicate volumes below the population mean, warmer tones (dark red) indicate volumes above. Across all four subjects, predictions recover the spatial pattern of the ground truth, including the periventricular contrast that develops with age.

(ii) Subjects with longer follow-up may differ systematically from those who drop out earlier; the temporal slope of MAE therefore conflates three distinct effects, genuine model degradation under autoregressive rollout, the underlying acceleration of disease-related volume change at later horizons, and cohort drift and should be interpreted as an upper bound on horizon-induced error growth attributable to the model alone. (iii) The five cross-validation folds are aggregated in this analysis; fold-stratified diagnostics (omitted for brevity) showed no substantial fold-to-fold inconsistency.

H.4 QUALITATIVE TRAJECTORY VISUALIZATIONS

To illustrate BRAINGENFLOW's flexibility in forecasting at arbitrary horizons across different diagnostic groups and cohorts, Figure 11 visualizes predicted brain volumes for four representative subjects sampled from the test sets. For each subject, baseline imaging features (145 ROIs) and clinical covariates are passed to BRAINGENFLOW, and the autoregressive rollout of Section 4.2 is unrolled to the target horizon; the predicted ROI volumes (Pr) are shown alongside the corresponding ground-truth scan (Gt) at that horizon, projected onto the MUSE atlas in three orthogonal views.

The four subjects span the diagnostic spectrum and both cohorts used in this work: a cognitively normal (CN) ADNI subject (baseline age 69.6) forecast 75 months ahead; an MCI ADNI subject (baseline age 57.6) forecast 86 months ahead; an AD subject from the AIBL cohort (baseline age 67.0) forecast 73 months ahead; and a second MCI ADNI subject (baseline age 74.2) forecast 90 months ahead. The horizons were not selected from a fixed grid: BRAINGENFLOW generates month-resolved trajectories from a single baseline scan, so any month within the rollout window can

be queried directly. Across all four cases, the predicted standardized volumes recover the spatial pattern of the ground truth, including the periventricular and medial-temporal contrasts characteristic of later-life atrophy.

I Results with and without Diagnosis

We present esults obtained by training the model with and without the diagnosis covariate on Table 15. We added this experiment in order to show that even without diagnosis, the predictive ability is still substantial. However, the covariate adds certain information relative to the rate of change pertaining to differently affected cohorts.

Table 15: Effect of withholding baseline diagnosis. The covariate is held constant during both training and testing, so the two models are architecturally identical. Five-fold protocol of Section 5.1.

	With diagnosis	Diagnosis withheld
Mean MAE (95% CI)	0.2167 [0.2079, 0.2255]	0.2228 [0.2048, 0.2408]
Ordering of mean predicted atrophy	CN $<$ MCI $<$ AD	CN $<$ MCI $<$ AD

The fold-level confidence intervals overlap substantially, and the progression ordering is preserved, indicating that forecasting is not driven by direct access to the diagnosis label.

J Disclaimer of LLM Usage

Large language models were utilized as writing assistants for grammar correction and style refinement. All technical content, methodology, and experiments were developed by the authors.