APPENDIX / SUPPLEMENTAL MATERIAL

A DATASETS AND PREPROCESSING

We use the iSTAGING consortium (Habes et al., 2021) that consolidated and harmonized imaging and clinical data from multiple cohorts. Our real data consists of neuroimaging and demographic measures taken from subjects in the iSTAGING consortium. Specifically, the neuroimaging measures are the 145 anatomical brain ROI volumes (119 ROIs in gray matter, 20 ROIs in white matter and 6 ROIs in ventricles) extracted using a multi-atlas label fusion method (Doshi et al., 2016). Phase-level harmonization was applied on these 145 ROI volumes to remove site effects (Pomponio et al., 2020). Specifically, we use the Alzheimer's Disease Neuroimaging Initiative (ADNI,http://www.adni-info.org/), which is a public-private collaborative longitudinal cohort study and has recruited participants categorized as Cognitively Normal (CN), Mild Cognitive Impairment (MCI) and diagnosed with Alzheimer's Disease (AD) through 4 phases (ADNI1, ADNIGO and ADNI2) (Weiner et al., 2017). We also use Baltimore Longitudinal Study of Aging (BLSA) (Ferrucci, 2008), which has been following participants who are cognitively normal at enrollment with imaging and cognitive exams since 1993.

We also extracted additional studies from the iSTAGING cohort, including the OASIS dataset Marcus et al. (2010), the Australian Imaging, Biomarker, and Lifestyle (AIBL) study (Ellis et al., 2009), and the PreventAD study (Tremblay-Mercier et al., 2021). These studies were exclusively reserved as held-out datasets for evaluating our method on external neuroimaging data.

Our analysis incorporates subjects across all identified progression trajectories: Cognitively Normal (CN) stables, individuals with Mild Cognitive Impairment (MCI), and those progressing to Alzheimer's Disease (AD) from either CN or MCI stages. For the clinical variables, we utilize Age at Baseline, Sex, Years of Education, and APOE4 Allele status, the latter being a known risk factor for Alzheimer's Disease (AD). Diagnostic categories were designated as Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). Subjects diagnosed with alternative forms of dementia, such as Lewy Body Dementia and Frontotemporal Dementia, were excluded from the study. These exclusions were minimal and did not significantly impact the overall sample size. Missing diagnostic information was classified as unknown (UKN). Furthermore, Years of Education was dichotomized: subjects with more than 16 years of education were coded as '1', while those with 16 years or fewer were coded as '0'. Detailed demographic and clinical characteristics of the diverse cohort are presented in Table 2.

Table 2: Summary of longitudinal studies with demographic and clinical Information. OASIS, AIBL, and PreventAD studies are used as held-out neuroimaging studies. For age, the mean and the standard deviation are reported. For sex, the number of males and the percentage is presented.

Study	**Subjects**	**Obs./Subject**	**#Obs.**	**Age**	**Male (%)**	**Diagnosis (%)**		
						CN	**MCI**	**AD**
ADNI	1616	5.0±2.0	7867	73.6±7.0	55.5	44.7	34.6	20.6
BLSA	584	3.0±1.0	1843	74.9±11.1	45.7	95.8	2.8	1.4
OASIS	548	3.0±1.0	1562	67.8±9.0	42.4	88.9	1.9	12.2
AIBL	82	3.0±1.0	247	75±7.7	56.14	33.74	28.81	37.45
PreventAD	271	4.2±1.4	1141	65.3±5.5	28.5	98.6	1.4	0.0

B ARCHITECTURAL DESIGN AND TRAINING

B.1 ROI VOLUME MODELS

For each ROI Volume biomarker, we build a separate deep kernel regression model with adaptive shrinkage. The deep kernel models (p-DKGP and ss-DKGP) take as input 145 volumetric ROIs along with the following covariates: Age at Baseline, Sex, Diagnosis at Baseline, APOE4 Alleles, Education Years, and Time. The transformation function Φ is implemented as a multilayer perceptron (MLP) composed of a sequence of linear layers. Φ reduces the input dimensionality from 151

(145 imaging features + 5 covariates and Time) to 64. Based on empirical validation, further reduction degrades predictive performance. The Gaussian Process (GP) is initialized with a zero mean function and an RBF kernel.

The p-DKGP is trained for 500 epochs with a learning rate of 0.01 using the Adam optimizer (Kingma & Ba, 2014) (with a weight decay of 0.01) and a dropout rate of 0.2 for regularization. Upon completion, we save the weights $(\mathbf{W}_p, \mathbf{b}_p)$ and the GP hyperparameters (variance and lengthscale) for inference on new test subjects and for transfer learning in the subject-specific model (ss-DKGP). For the subject-specific model, we initialize the ss-DKGP with the saved weights $(\mathbf{W}_p, \mathbf{b}_p)$ and the hyperparameters of the population GP. Then, we train the ss-DKGP for 100 epochs with a learning rate of 0.01, during which the deep kernel is frozen; only the subject-specific GP hyperparameters are updated. The Adam optimizer with a weight decay of 0.05 is used in this stage.

B.2 SPARE MODELS

For each SPARE biomarker, SPARE-AD and SPARE-BA, we build a separate deep kernel regression model with adaptive shrinkage estimation. The input features include the same 145 volumetric ROIs, along with the following covariates: Age at Baseline, Sex, Diagnosis at Baseline, APOE4 Alleles, Education, SPARE-BA, and SPARE-AD at baseline, in addition to Time.

As in the ROI Volume models, the transformation function Φ is a multilayer perceptron that projects the 153-dimensional input to a 64-dimensional feature space. We employ a GP with a zero mean function and an RBF kernel. The p-DKGP is trained for 500 epochs with a learning rate of 0.01, using the Adam optimizer with a weight decay of 0.01 and a dropout rate of 0.2. The learned weights $(\mathbf{W}_p, \mathbf{b}_p)$ and GP hyperparameters (variance and lengthscale) are then saved for subsequent inference and for initializing transfer learning in the subject-specific model. Transfer learning is performed by initializing the ss-DKGP with the saved weights $(\mathbf{W}_p, \mathbf{b}_p)$ and the population GP hyperparameters. The ss-DKGP is then trained for 100 epochs with a learning rate of 0.01, during which the deep kernel is detached from the optimization process and only the subject-specific GP hyperparameters are updated using the Adam optimizer with a weight decay of 0.05.

B.3 DETAILS ON THE COMPETING BASELINES

We compare our method against various baselines, including Linear Mixed Effects (LMM) models, Generalized Additive Models (GAMs), Deep Regression, and the Deep Mixed Effects (DME) (Chung et al., 2019). Each model was trained and tested using the same population dataset of ADNI and BLSA cohort, as our method. For LMM, we use the 145 ROI Volumes at first visit and clinical covariates (Age, Sex, Diagnosis, Education Years, APOE4 Alleles, and Time). The Subject ID served as a random intercept and the interaction term Time:Subject ID as a slope. For GAMs, personalization involved fitting a GAM to population data, supplemented with each test subject's partially observed trajectory. The second non-linear baseline is the Deep Regression. At first, we train the Deep Regression on the population dataset. Then on the personalization, we freeze the first layers of the deep network and we fine tune only the last layer with the subject data. The architecture of the Deep Network is an MLP that consists of an input layer, three hidden layers, and an output layer. The first hidden layer contains 100 neurons, the second hidden layer has 50 neurons, and the third hidden layer again contains 100 neurons. Each hidden layer uses the Rectified Linear Unit (ReLU) activation function, which introduces non-linearity into the model and helps it learn complex data patterns. The MLP is trained using the Stochastic Gradient Descent (SGD) optimization algorithm to minimize the Mean Squared Error (MSE) loss function. For the Deep Mixed Effects (Chung et al., 2019), we used the publicly available code in order to apply the DME method to our data. As a warping mean function, we use a MLP. Additionally, we experimented with a Transformer model (Vaswani et al., 2017) utilizing positional encoding along the temporal dimension and implemented LSTM models Hochreiter & Schmidhuber (1997). However, both models faced convergence issues during training and did not yield satisfactory results on our sparse temporal dataset. Theoretically, Transformer models rely on self-attention mechanisms to capture dependencies across sequences, which assume the availability of comprehensive and densely sampled sequential data. In the context of sparse temporal data, the self-attention mechanism cannot function optimally due to insufficient temporal information, leading to suboptimal performance. Similarly, LSTM models require temporally aligned and regularly sampled data to maintain the sequential relationships inherent in time

series. Without prior preprocessing, such as data imputation to handle irregularities and missing values, LSTMs struggle to learn effectively from sparse temporal data. As a result, we omitted these models from the quantitative comparisons in the current work.

C Analysis on Posterior Correction

Our goal is to determine the oracle shrinkage parameter α in Equation equation 11, which combines the predictions from the population model (p-DKGP) and the subject-specific model (ss-DKGP). To achieve this, we propose minimizing the Mean Squared Error (MSE) between the combined prediction y_c and the ground truth y_t over all time points. The objective function is defined as:

$$J(\alpha) = \sum_{t=0}^{t_n} \left(y_t - (\alpha y_{p_t} + (1-\alpha) y_{s_t})\right)^2. \tag{11}$$

In this section, we provide a theoretical justification for this formulation, explaining why the independence assumption between the models' errors does not affect the estimation of α using this objective function.

Both the p-DKGP and ss-DKGP models provide predictive means y_{p_t} and y_{s_t} for the ROI value at each time point t. We aim to find the oracle α that minimizes the MSE between the combined prediction y_c and the ground truth y_t. The combined prediction is given by:

$$y_c = \alpha y_{p_t} + (1-\alpha) y_{s_t}. \tag{12}$$

To find the optimal α, we take the derivative of $J(\alpha)$ with respect to α and set it to zero:

$$\frac{dJ}{d\alpha} = -2 \sum_{t=0}^{t_n} \left(y_t - (\alpha y_{p_t} + (1-\alpha) y_{s_t})\right) \left(y_{p_t} - y_{s_t}\right) = 0. \tag{13}$$

Simplifying, we get:

$$\sum_{t=0}^{t_n} \left(y_t - (\alpha y_{p_t} + (1-\alpha) y_{s_t})\right) \left(y_{p_t} - y_{s_t}\right) = 0. \tag{14}$$

Solving for α, we find:

$$\alpha^* = \frac{\sum_{t=0}^{t_n} (y_t - y_{s_t})(y_{p_t} - y_{s_t})}{\sum_{t=0}^{t_n} (y_{p_t} - y_{s_t})^2}. \tag{15}$$

This expression shows that the optimal α depends on the covariance between $y_t - y_{s_t}$ and $y_{p_t} - y_{s_t}$, and the variance of $y_{p_t} - y_{s_t}$.

To gain further insight into the dependence of the optimal α^* on statistical properties of the data, we relate Equation 15 to the concepts of covariance and variance. Let us define:

$$X_t = y_{p_t} - y_{s_t}, \quad Y_t = y_t - y_{s_t}. \tag{16}$$

With these definitions, Equation equation 15 becomes:

$$\alpha^* = \frac{\sum_{t=0}^{t_n} Y_t X_t}{\sum_{t=0}^{t_n} X_t^2}. \tag{17}$$

The numerator and denominator in Equation equation 17 are related to the sample covariance and variance, respectively. Specifically, the numerator is proportional to the covariance between Y_t and X_t, and the denominator is proportional to the variance of X_t:

$$\text{Cov}(Y, X) = \frac{1}{n} \sum_{t=0}^{t_n} (Y_t - \bar{Y})(X_t - \bar{X}), \tag{18}$$

$$\text{Var}(X) = \frac{1}{n} \sum_{t=0}^{t_n} (X_t - \bar{X})^2, \tag{19}$$

Table 3: Correlation between the errors of p-DKGP and ss-DKGP models for different ROIs and Observations

ROI	3 Observations	4 Observations	5 Observations	6 Observations
Hippocampus R	0.237	0.337	0.374	0.318
Thalamus Proper R	0.136	0.348	0.302	0.344
Lateral Ventricle R	0.201	0.247	0.319	0.354
Hippocampus L	0.341	0.300	0.348	0.208
Amygdala R	0.262	0.325	0.355	0.372
Amygdala L	0.292	0.356	0.331	0.394

where $\bar{Y}$ and $\bar{X}$ are the sample means of Y_t and X_t, respectively, and $n = t_n + 1$ is the number of time points.

Assuming that Y_t and X_t are centered (i.e., $\bar{Y} = 0$ and $\bar{X} = 0$), which is valid if we consider deviations from their means, Equation equation 17 simplifies to:

$$\alpha^* = \frac{n \cdot \text{Cov}(Y, X)}{n \cdot \text{Var}(X)} = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}. \tag{20}$$

This expression shows that the optimal α^* is the coefficient that minimizes the residual sum of squares in a simple linear regression of Y_t on X_t without an intercept. In other words, α^* is the scaling factor that best relates the difference between the population and subject-specific predictions (X_t) to the residuals of the subject-specific model (Y_t).

- If $\text{Cov}(Y, X)$ is large and positive, it indicates that when the subject-specific model underpredicts or overpredicts (Y_t deviates from zero), the difference between the population and subject-specific predictions (X_t) tends to be in the same direction. In this case, a larger α (giving more weight to the population model) helps reduce the overall error.

- If $\text{Cov}(Y, X)$ is small or negative, it suggests that the population model does not provide useful information to correct the subject-specific model's errors, and a smaller α (giving more weight to the subject-specific model) is preferable.

This analysis confirms that the optimal α^* depends on the covariance between $y_t - y_{s_t}$ and $y_{p_t} - y_{s_t}$, and the variance of $y_{p_t} - y_{s_t}$. Understanding this dependence provides valuable insight into how the differences between the models' predictions relate to the residuals and how to optimally combine them to minimize the prediction error.

C.1 Independence Assumption and Its Impact

The combined predictive mean y_c is a deterministic function of y_{p_t}, y_{s_t}, and α, as given in Equation 12. It does not involve the errors or variances associated with the predictions. As a result, the independence or correlation between the models' errors does not influence the calculation of y_c. While the independence assumption does not affect the estimation of α or the calculation of y_c, it does impact the calculation of the combined predictive variance v_c. The variance of the combined prediction is given by:

$$v_c = \alpha^2 v_{p_t} + (1 - \alpha)^2 v_{s_t} + 2\alpha(1 - \alpha)\text{Cov}(y_{p_t}, y_{s_t}). \tag{21}$$

If the errors of the two models are assumed to be independent, the covariance term $\text{Cov}(y_{p_t}, y_{s_t})$ is zero, simplifying v_c to:

$$v_c = \alpha^2 v_{p_t} + (1 - \alpha)^2 v_{s_t}. \tag{22}$$

Empirical analysis indicates that the errors of the two models are midly correlated, with correlation to range between 0.136 to 0.394. Therefore, the inclusion the covariance term in the calculation of v_c to accurately quantify the uncertainty of the combined prediction.

Overall, the theoretical justification demonstrates that the MSE objective function is appropriate for estimating the shrinkage parameter α in our context. It avoids the need for the independence assumption during α estimation and simplifies the optimization process. However, when calculating the predictive variance v_c, it is essential to account for the covariance between the models' predictions to accurately quantify uncertainty.

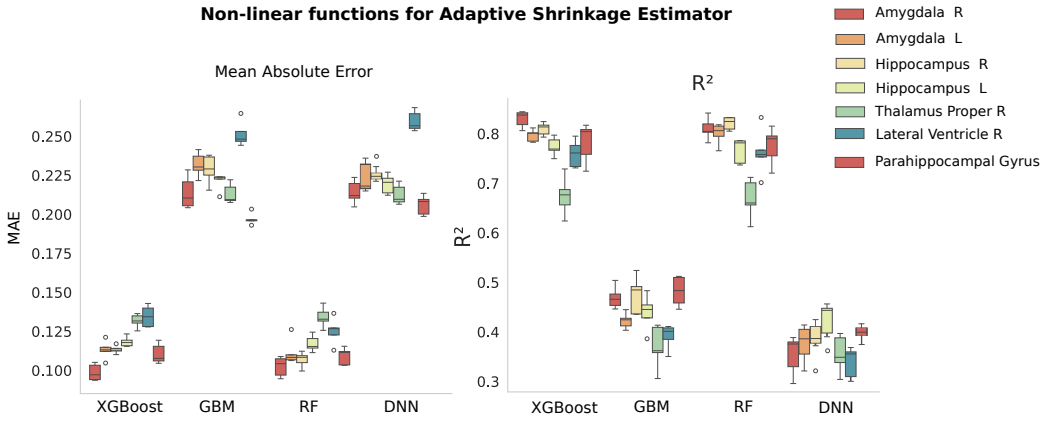

Figure 5: We present MAE and R^2 from 5-fold cross-validation using the 200 held-out subjects from ADNI and BLSA subjects for the Adaprive Shrinkage estimator using XGBoost, GBM, RF and DNN as non-linear functions

To address this issue, we do:

- **Estimating Covariance:** Empirically estimate $\text{Cov}(y_{p_t}, y_{s_t})$ using validation data.
- **Adjusting Variance Calculations:** Include the covariance term in the calculation of v_c as per Equation 21.
- **Reassessing Prediction Intervals:** Recompute prediction intervals using the adjusted v_c to ensure improved coverage.

C.2 Alternatives of Non-Linear functions for Adaptive Shrinkage Estimator

We experiment with several non-linear functions to determine which one learns best the adaptive shrinkage mapping, namely the mapping between a and $y_p, y_s, V_p, V_s, T_{obs}$. We conduct 5-fold cross-validation using XGBoost Regression (XGBoost), Random Forest (RF), Gradient Boosting Machine (GBM), and a Deep Neural Network (DNN). The DNN architecture includes a linear layer (5x16), ReLU activation, a linear layer (16x8), ReLU activation, and a final linear layer (8x1). It is trained with MSE loss and optimized using Adam with a learning rate of 0.01. Results, presented in Figure 5 indicate that XGBoost Regression and Random Forest achieve the best performance in terms of mean absolute error and r^2 score on the test set, with both models achieving an average r^2 score greater than 0.75 across the majority of the ROI Volumes.

Table 4: XGBoost performance on predicting the adaptive shrinkage α for 7 ROI Volume biomarkers

ROI Volume	MAE	R²
Amygdala R	0.099	0.830
Amygdala L	0.113	0.796
Hippocampus R	0.113	0.810
Hippocampus L	0.118	0.774
Lateral Ventricle R	0.132	0.675
Thalamus Proper R	0.135	0.759
PHG R	0.111	0.783

D Experiments

D.1 Stratified Performance Analysis by Covariates

To thoroughly evaluate our method, we perform stratification of prediction errors across key demographic and clinical factors: sex, APOE4 Allele status, and education level. This stratification allows

us to examine the model's ability on varying subpopulations. We report the Mean Absolute Error and corresponding 95% confidence intervals (CIs) for pers-DKGP, alongside with the competing baselines.

Stratification by Sex. Our results indicate that pers-DKGP consistently achieves the lowest Mean AE for both male and female groups. In males, pers-DKGP attains a Mean AE of 0.135 (95% CI: $[0.120, 0.150]$), significantly outperforming LMM, which yields a Mean AE of 0.187 (CI: $[0.160, 0.214]$). Similarly, for females, pers-DKGP reports a Mean AE of 0.145 (CI: $[0.130, 0.160]$), compared to GAM's Mean AE of 0.198 (CI: $[0.165, 0.231]$). Although prediction errors are slightly higher in females—likely due to increased biomarker variability—the consistently narrower CIs of pers-DKGP underscore its enhanced reliability across sexes.

Stratification by APOE4 Alleles Status. Considering the crucial role of the APOE4 Allele in Alzheimer's Disease progression, we examine model performance for Non-Carriers, Heterozygous and Homozygous separately. For APOE4 homozygotes, pers-DKGP achieves a Mean AE of 0.142 (CI: $[0.128, 0.156]$), markedly lower than DME's Mean AE of 0.210 (CI: $[0.176, 0.244]$). For non-carriers, pers-DKGP obtains a Mean AE of 0.130 (CI: $[0.118, 0.142]$), outperforming DeepRegr, which records a Mean AE of 0.192 (CI: $[0.162, 0.222]$).

Stratification by Education Education level, serving as a proxy for cognitive reserve, introduces additional variability in disease progression predictions. In the subgroup with education levels below 16 years, pers-DKGP achieves a mean AE of 0.155 (CI: $[0.140, 0.170]$), outperforming LMM, which exhibits a mean AE of 0.225 (CI: $[0.195, 0.255]$). Among subjects with 16 or more years of education, pers-DKGP maintains its advantage, recording a mean AE of 0.120 (CI: $[0.110, 0.130]$), whereas GAM shows a mean AE of 0.175 (CI: $[0.145, 0.205]$).

Overall, the stratification of AE demonstrates that pers-DKGP outperforms baseline methods in all subpopulations. Its lower mean AE and narrower confidence intervals indicate not only higher predictive accuracy but also greater reliability, even in challenging subgroups such as APOE4 carriers, and individuals with lower education levels.

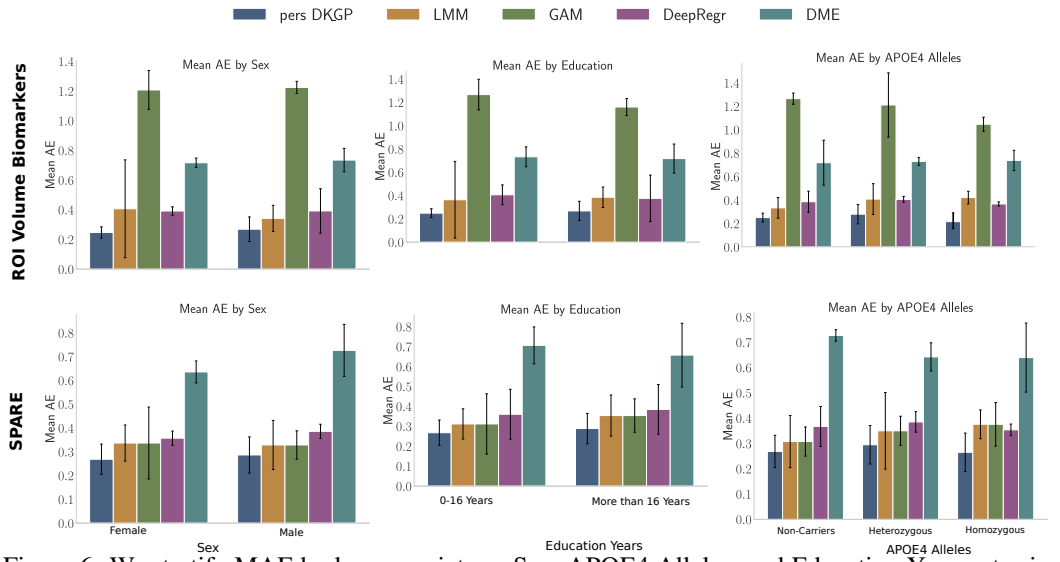

Figure 6: We stratify MAE by key covariates—Sex, APOE4 Alleles, and Education Years—to rigorously assess model performance across different subpopulations. Error bars denote the 95% confidence intervals of the MAE. The top row aggregates metrics for seven ROI Volume biomarkers, while the bottom row summarizes the MAE for both SPARE-AD and SPARE-BA.

D.2 Performance with Number of Observations

Error with Number of Observations for SPARE-AD Score. Table 5 presents the mean absolute error and 95% confidence interval for the SPARE-AD biomarker across different numbers of observations (history). A history of 1 corresponds to using the population model prediction, which

we employ when only a single acquisition of the subject is available; in this case, we have $\alpha = 1$. As we increase the number of observations, we apply posterior correction with adaptive shrinkage α inferred by the adaptive shrinkage estimator, allowing us to adjust the model based on the subject's individual history. Notably, the mean AE decreases as more observations are included. This demonstrates the benefit of applying Adaptive Shrinkage with increased subject history to improve the accuracy of the SPARE-AD biomarker prediction.

Table 5: Mean Absolute Error and 95% Confidence Interval for the SPARE-AD biomarker with increasing number of observations

Observations	**Mean AE**	**95% CI**
1 ($\alpha = 1$)	0.227	0.003
2	0.233	0.008
3	0.219	0.008
4	0.153	0.010
5	0.148	0.010

Error Analysis. Figure 7**a** illustrates the distribution of absolute errors across history levels (1 to 6) using boxplots. The median error is indicated by the central line within each box, with the interquartile range (IQR) defining the edges and whiskers extending to 1.5 times the IQR. Outliers are depicted as individual points beyond the whiskers. A red line represents the mean absolute error, providing an overview of the central tendency.

The results demonstrate a marked reduction in mean absolute error with increasing history, particularly during the earlier transitions: a 21.96% decrease from history 1 to 2 and a further 15.92% decrease from history 2 to 3. This underscores the significance of incorporating additional longitudinal observations. However, the improvements plateau at higher history levels, reflecting diminishing returns. It is important to note that the error will never practically reach zero, owing to the inherent noise and variability of neuroimaging biomarkers. Nevertheless, the results highlight the necessity of subject-specific personalization, as individual trajectories often deviate from population-level SPARE-AD estimates. With additional follow-up observations, these deviations are better captured, resulting in more accurate and individualized SPARE-AD trajectories. This emphasizes the critical role of model adaptation in clinical practice, as refined SPARE-AD estimates can provide valuable insights for predicting disease progression, including transitions to dementia or, more specifically, progression from MCI to Alzheimer's Disease.

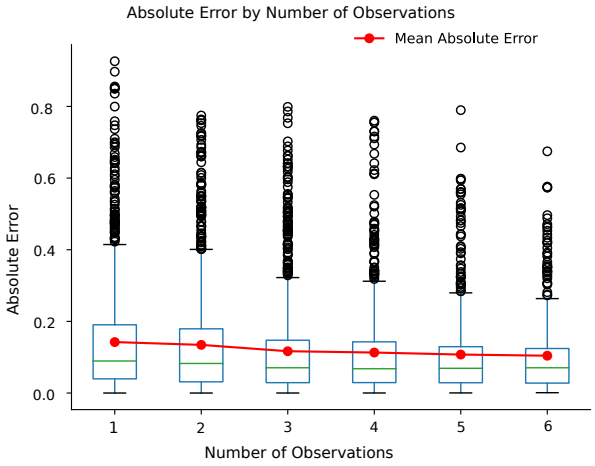

Figure 7: Boxplots show the distribution of absolute errors across history levels (1 to 6), with the central line indicating the median, the box edges representing the interquartile range (IQR), and whiskers extending to 1.5 times the IQR. Outliers are shown as points beyond the whiskers. The red line connects the mean absolute error for each level.

D.3 QUALITATIVE EXAMPLES OF ROI VOLUME AND SPARE BIOMARKERS

In this section we provide additional qualitative results on test subjects. We present results for the ROI volume biomarkers as well as the SPARE AD biomarker. The ROI progression models use as input the imaging scan (145 Volumetric ROIs), demographics and clinical variables. The SPARE-AD progression model uses the 145 Volumetric ROIs, demographics and clinical variables as well as the SPARE-AD score at baseline.

Empirical Evidence of Predicted SPARE-AD Trajectories for MCI Progressor. In figure 8 we present an example of a subject that starts as Cognitive Normal at the Age of 74 years old. We use our model (pers-DKGP) in order to predict the longitudinal SPARE-AD changes from the 145 Volumetric ROIs as well as the demographics (Age, Sex, Education Years) and clinical variables such as the Clinical Diagnosis and the APOE4 Alleles. At the first visit of the subject, we extrapolate a SPARE-AD trajectory that indicates no changes related to progression. Within the 2 and a half years of observations the MCI the predicted trajectory of the SPARE-AD biomarker indicates no significant longitudinal change in the SPARE-AD trajectory. In the 42 months of observations, the predicted SPARE-AD trajectory indicates an increasing trend in the SPARE-AD values that indicates increased AD releated patterns in the brain. Increased AD-like patterns indicate higher risk of conversion to MCI or Dementia (AD). In almost 5 years of observation, the predicted trajectory indicates a steeper increase in the future SPARE-AD values indicating againg high risk of MCI or AD. The subject finaly is clinically diagnosed with MCI after 80 months of observation. Our method is able to predicted changes of biomarker values that are indicative of Progression and this highlights also the clinical usage of our method as a stong predictive tool for progression prediction either for use in the clinical practice or the design of clinical trials. For example, this subject with an increasing trend of SPARE-AD trajectory would be an ideal subject for recruitment in a clinical trial as it converts to demonstrates inclining biomarker trajectory making it a subject that is highly likely to be part of a clinical trial.

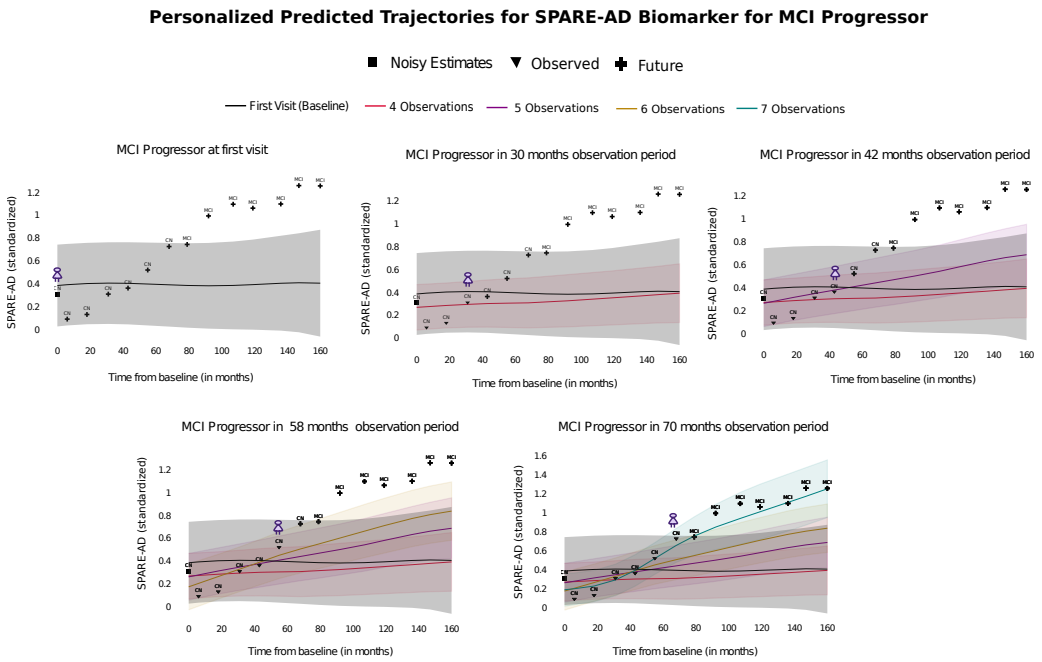

Figure 8: We present predicted SPARE-AD trajectories for a Cognitive Normal subject at the baseline age of 74 years old. After 7 years the subject is diagnosed with Mild Cognitive Impairment. The predicted SPARE-AD trajectories predict the increasing attrophy-like patterns 3 years prior the clinical diagnosis of conversion to MCI. This highlights the potential clinical application of our tool for progression prediction and clinical trial design.

Empirical Evidence from a Healthy Control and and MCI Progressor. In Figure 9, we present a qualitative comparison of predicted trajectories for two subjects who begin the study at similar ages—74 (left) and 71 (right), respectively—and are cognitively normal at baseline. We analyze the volumetric loss in three brain regions: the amygdala, hippocampus, and lateral ventricle. The volumetric loss is modeled as a function of MRI scans alongside clinical and demographic covariates, including age, sex, diagnosis, APOE4 allele status, and years of education.

At the initial visit, both subjects exhibit minimal hippocampal atrophy. However, over successive follow-up observations, the subject on the left (9**b**) demonstrates a markedly steeper decline in hippocampal volume compared to the subject on the right, who maintains a more stable hippocampal trajectory. The predicted accelerated decline in hippocampal volume for the subject on the left suggests an elevated risk of progressing to mild cognitive impairment (MCI) or dementia, potentially due to underlying pathology such as Alzheimer's disease (AD) or accelerated brain aging. In contrast, the subject on the right (9**a**), who remains a healthy control throughout the observation period, exhibits only minimal hippocampal volume loss.

This example illustrates the practical application of our method in predicting disease progression, which has significant implications for clinical practice, clinical trial design, and treatment effect estimation. Specifically, in the context of clinical trial design, identifying subjects with steep hippocampal atrophy trajectories can inform the recruitment of individuals who are more likely to exhibit disease progression, thereby enhancing the efficiency and efficacy of the study.

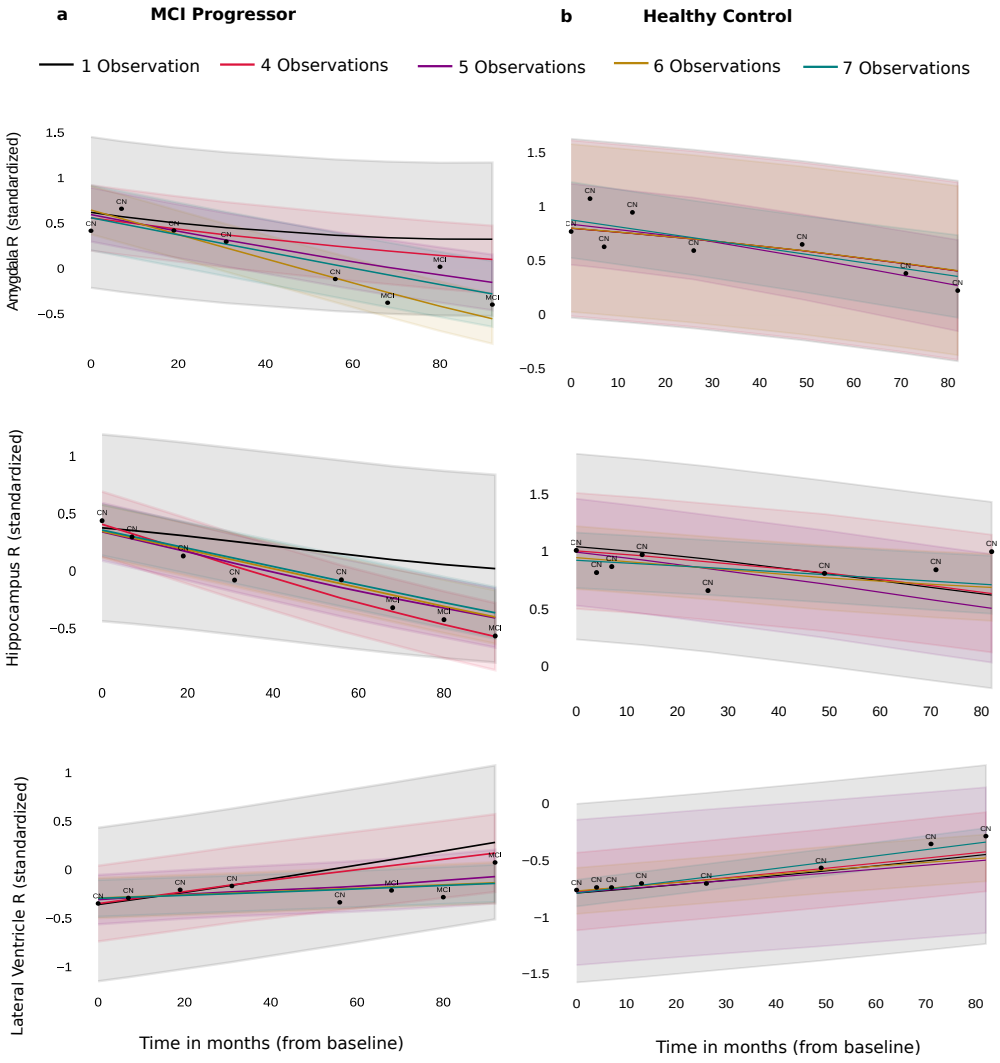

Figure 9: We present predicted Amygdala and Hippocampal Volume trajectories and Ventricular Enlargement for a Healthy Control and and MCI Progressor. MCI Progressor exhibits steeper volume loss in Amygdala and Hippocampus in comparison with the Healthy Control. MCI Progressor exhibits either accelarated brain aging or is in the onset of AD which justifies its faster volume loss.

Empirical Evidence of the Personalization in Test Subjects. To further validate the efficacy of our method, we provide empirical evidence through qualitative analysis in scenarios where individual trajectories either diverge from or align with the true underlying trend. In Figure 10, we present a cohort of test subjects (panels **(a)**–**(h)**) exhibiting variability in progression status, alongside the corresponding adaptive shrinkage parameter α—depicted in the second row—utilized at each personalization step. Consistently across all examples, we observe that the shrinkage parameter α progressively decreases as the number of observations increases. In several cases, the adjustments remain more conservative, with α staying closer to 1, which aligns with the foundational intuition of our method. This pattern suggests that an adequate accumulation of evidence regarding a subject's trajectory is necessary to shift the adaptive shrinkage parameter toward zero, thereby placing greater trust in the ss-DKGP predictions. This rationale is well-founded, as substantial evidence is crucial for the ss-DKGP to generate meaningful trajectories and mitigate the noise variations inherent in neuroimaging data acquisitions. Additional examples are visualized in Figure 11.

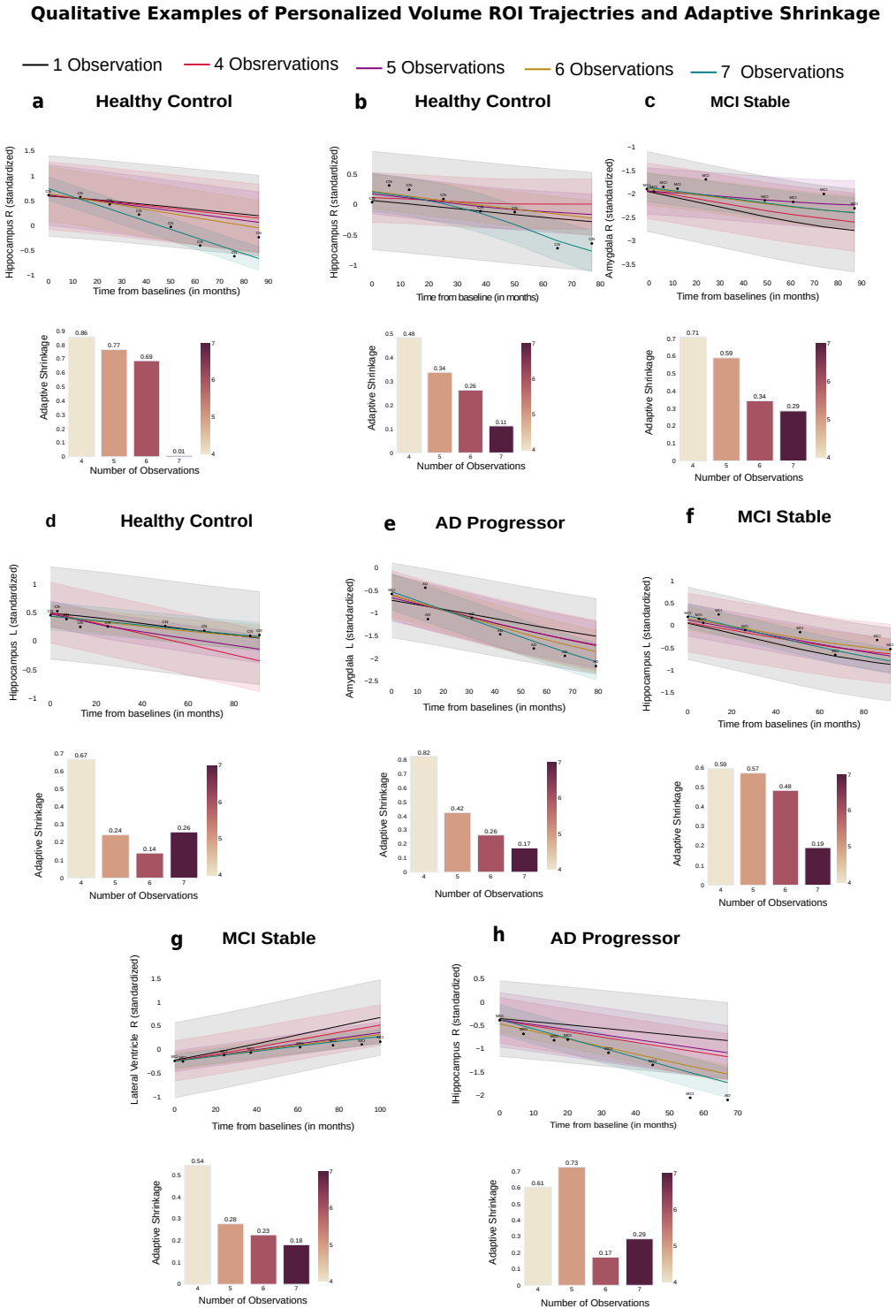

Figure 10: We present qualitative examples where population trajectories deviate from the subject's observed trajectory throughout the observation period (in years). Evidence is provided from eight distinct test subjects. In the first row of each panel (**a**)-(**h**), we present the adapted trajectories. The second row of each subfigure visualizes the corresponding adaptive shrinkage for posterior correction for each observation, ranging from 4 to 7 observations.

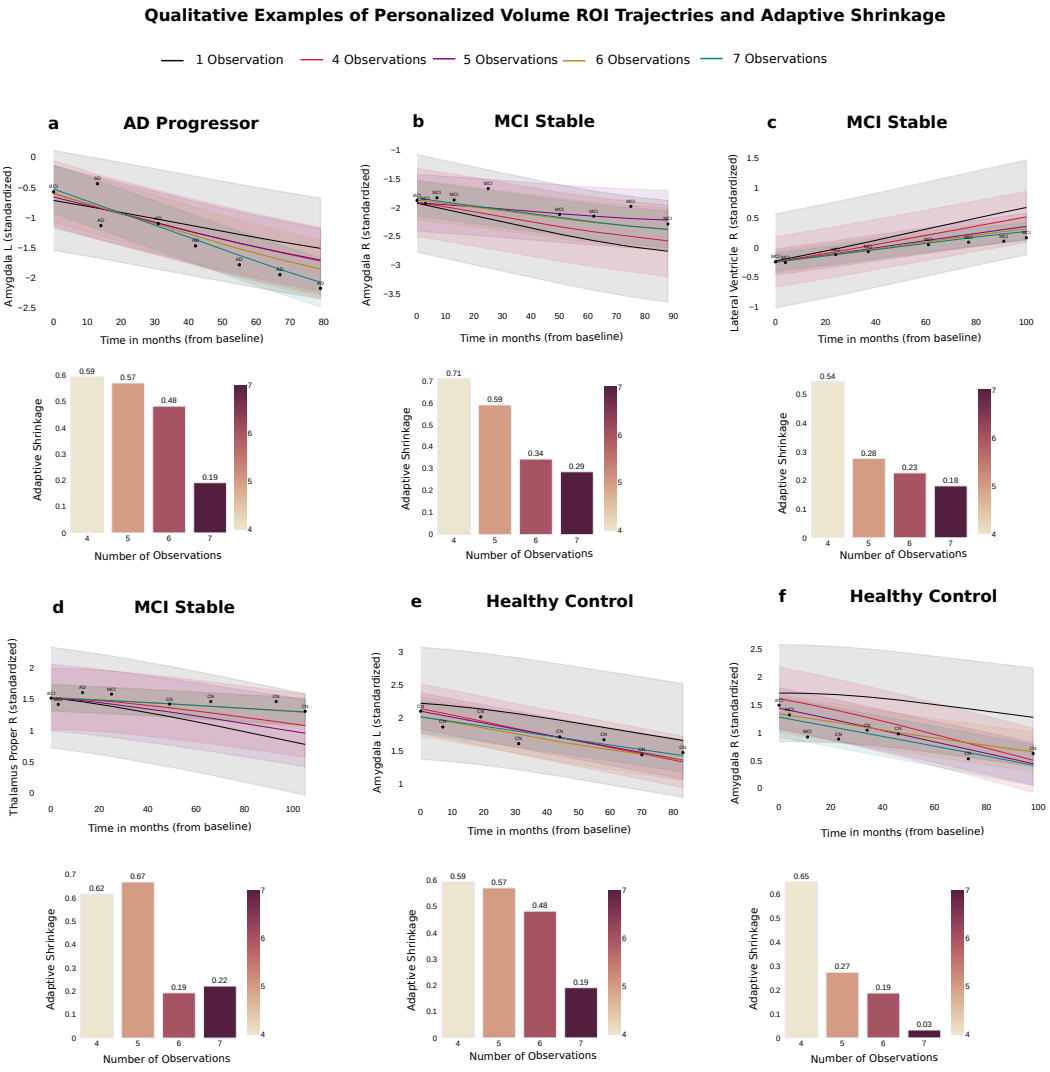

Figure 11: We present qualitative examples where population trajectories deviate from the subject's observed trajectory throughout the observation period (in years). Evidence is provided from six distinct test subjects. In the first row of each panel (**a**)-(**f**), we present the adapted trajectories. The second row of each subfigure visualizes the corresponding adaptive shrinkage for posterior correction for each observation, ranging from 4 to 7 observations.

D.4 Analysis of Adaptive Shrinkage Estimator

D.4.1 Ablation on shrinkage parameter α

Determining the shrinkage for each ROI Volume is non-trivial, particularly for predicting long-term atrophy trajectories. This is a particularly difficult task because either a subject's trajectory would deviate from population trends, or a subject would have limited acquisitions, making it difficult for the subject-specific model to extrapolate its ROI Volume trajectory. Volume loss, or differently atrophy development, in the brain is a slow process, especially for a subject who is young or has not yet developed any pathology. Thus, in the case of limited acquisitions for a subject, which are also close in time to the baseline, the additional observations are rather noisy copies of the baselines and do not contain any "signal" of the trajectory of developing atrophy. In that case, the ss-DKGP model would not have enough evidence to extrapolate future values. As a result, we should find the optimal shrinkage of the two predictors to leverage both the population's ability to make reliable long-term predictions and the subject-specific model's ability to learn short-term predictions. We show that Adaptive Shrinkage provides the best results compared to any other weighting scheme, as we also present in table 6.

Table 6: Ablation study on the shrinkage parameter α. We report the Mean AE along with its 95% percentile CI, Mean Coverage, and Mean Interval Width

ROI	Mean AE (CI)	Mean Coverage	Mean Interval
Best Constant			
Hippocampus R	0.257 (0.209)	0.808	0.843
Lateral Ventricle R	0.143 (0.182)	0.853	0.507
Thalamus Proper R	0.241 (0.214)	0.934	1.127
Amygdala R	0.349 (0.317)	0.742	0.918
Hippocampus L	0.274 (0.245)	0.805	0.850
PHG R	0.423 (0.360)	0.582	0.844
Deterministic			
Hippocampus R	0.308 (0.275)	0.480	0.459
Lateral Ventricle R	0.156 (0.192)	0.620	0.310
Thalamus Proper R	0.308 (0.287)	0.512	0.492
Amygdala R	0.418 (0.400)	0.503	0.650
Hippocampus L	0.314 (0.290)	0.487	0.478
PHG R	0.487 (0.457)	0.459	0.681
Adaptive Shrinkage			
Hippocampus R	**0.243** (0.191)	0.795	0.902
Lateral Ventricle R	**0.131** (0.186)	0.855	0.626
Thalamus Proper R	**0.219** (0.216)	0.849	0.911
Amygdala R	**0.312** (0.283)	0.762	0.964
Hippocampus L	**0.258** (0.241)	0.790	0.901
PHG R	**0.389** (0.344)	0.745	0.908

D.4.2 Interpretation of Adaptive Shrinkage Estimator

As we increase the number of observations, we see that, no matter the biomarker, the alpha tends to zero. This aligns with the domain expectation that the longer the time from the baseline of the last observation Tobs, the more likely we are to have observed a trajectory trend from the subject's data.

In Figure 12, we visualize the distribution of Adaptive Shrinkage in the test set as well as in the three external clinical studies. This demonstrates that Adaptive Shrinkage has learned to assign greater trust to the subject-specific model as the number of follow-ups increases for a subject. This aligns perfectly with domain expectations and the explainability analysis we implemented for the Adaptive Shrinkage function. This property makes the Adaptive Shrinkage function a transparent method for performing posterior correction in the two predictive distributions, p-DKGP and ss-DKGP.

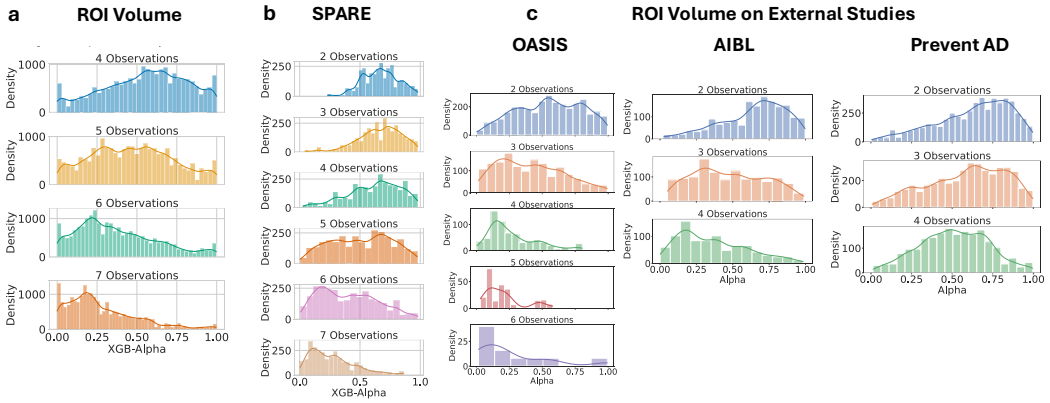

Figure 12: We visualize the distribution of adaptive shrinkage α for **a)** the 7 ROI Volumes, the **b)** SPARE-BA and SPARE-AD biomarkes and **c)** the 7 ROI Volumes in the external neuroimaging studies: OASIS, AIBL and PreventAD

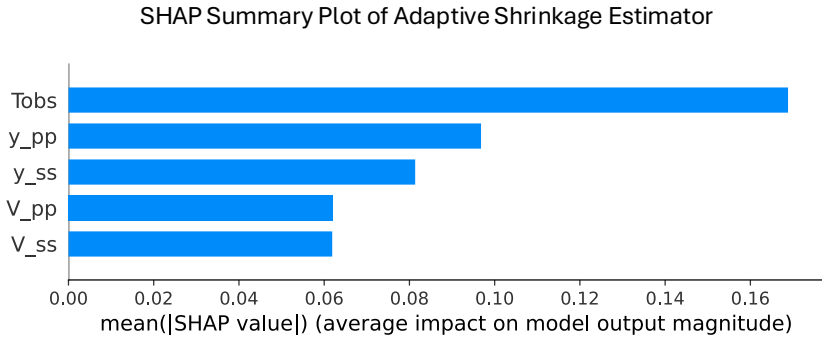

Figure 13: We calculate SHAP values for the Adaptive Shrinkage estimator for the SPARE-AD biomarker. As expected, the time of observation Tobs emerges as the most influential feature of the Adaptive Shrinkage estimator.

Table 7: Correlation Analysis between Deviation (δ_y) and Predicted α, and between T_{obs} and Predicted α for Large Deviation

Biomarker	**Correlation between T_{obs} and Predicted α for Large δ_y**
SPARE-BA	-0.640
SPARE-AD	-0.529
Lateral Ventricle	-0.484
Hippocampus L	-0.401
Hippocampus R	-0.381
Thalamus Proper R	-0.555
PHG R	-0.479
Amygdala R	-0.439

D.5 COMPARISON ON ALTERNATIVE GP PERSONALIZATION APPROACHES

In this section, we conduct a comparative analysis with other personalized GPs that align with our formulation. Specifically, within the ss-DKGP framework, we do an ablation study to see how the Φ transformation, learned from the population model, affects the subject-specific process. To achieve this, we train the ss-DKGP for each subject on the test set without initializing the deep kernel (ss-DKGP no init). We also train a standard subject-specific Gaussian Process (ss-GP) with a zero mean and RBF kernel. This comparison demonstrates the effectiveness of transferring the Φ from the p-DKGP when training the ss-DKGP. Additionally, we explore an alternative personalization approach where the population dataset D_p is augmented with the subject's observed trajectory D_s. In this setting, we again employ the Φ transformation learned from the p-DKGP. This approach is referred to as the ft-DKGP (fine-tuned DKGP). We perform transfer learning by initializing the weights of the deep kernel with $(\mathbf{W}_p, \mathbf{b}_p)$. The ft-DKGP is trained for 500 epochs using the same learning rate as the p-DKGP. During this process, the deep kernel is detached from the optimization procedure, and only the hyperparameters of the subject-specific GP are updated. The Adam optimizer with a weight decay of 0.05 is utilized.

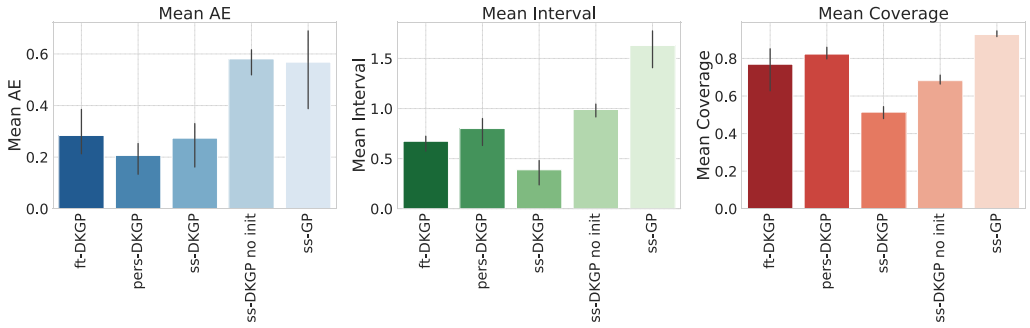

Figure 14: Comparison of predictive performance and uncertainty quantification across various GP models, averaged over three regions of interest (ROIs): Hippocampus, Lateral Ventricle, and Thalamus Proper, indicating that the personalized DKGP models achieve the best prediction accuracy and highest coverage. The bar plot displays the Mean performance metrics (AE, interval length and coverage) across these ROIs, while the line represents the standard deviation.

Among the ss-DKGP, ss-DKGP no init, and ss-GP models, we observe that ss-DKGP achieves the lowest Absolute Error (AE) with a significant margin compared to the other two settings. This indicates that leveraging the population transformation Φ is crucial for the effective training of ss-DKGP. This finding supports our hypothesis that the transformation Φ successfully captures the most predictive features for ROI progression, which are beneficial for ss-DKGP training. We observe that ft-DKGP model achieves performance that is close to the ss-DKGP model. However, ft-DKGP fails to personalize on unseen times, since the predicted trajectory falls back to the population trend. This is not the optimal way to personalize since the trajectory does not adapt to the subject specific trend. Additionally, it is not computationally efficient to retrain the model with the entire population data every time we need to personalize a subject.

Furthermore, the pers-DKGP model achieves the lower AE, which is an additional indication in favor of our approach. It highlights the strength of including the p-DKGP model in the final personalized prediction. Knowing solely the observed trajectory of a subject is not enough in case of limited and noisy observations. In that case we should trust the p-DKGP model more, which translates to an α parameter close to 1. Interestingly, this intuition aligns with the predicted α that we got during the personalization from the XGBoost regression. To verify that, we gathered the predicted α from the personalization process from the 7 ROIs. We plotted the distribution of α with the number of observations ranging from 4 till 7. The plot is attached on the Appendix 12. It clearly depicts that, as the number of observations increases, the distributions tend to show more noticeable skewness to the right, with higher densities in the lower α ranges and decreasing densities towards higher α values. This trend suggests that as more observations are taken into account in personalization, the shrinkage parameter α tends to be smaller. That translates to more trust to the ss-DKGP prediction. This is highly intuitive because as observation time T_{obs} increases, more acquisitions we obtain for

a subject and thus the more information the ss-DKGP captures about the progression of a ROI over time.

Determining the optimal α for new subjects is a complex task, which motivates us to explore various possibilities for α values. To demonstrate the effectiveness of our shrinkage formulation, we present an ablation study comparing different α approaches in the following subsection. This analysis aims to showcase the superiority of our approach in optimizing model performance for unseen data.