Supplementary Material

Table 1: Summary of participant demographics in iSTAGING consortium. Age is described in format: mean ± std [min, max]. F and M in gender represent female and male separately. Field indicates the magnetic strength of the MRI scanners.

Study	Subject	Age	Gender (F/M)	Field
BLSA-1.5T	157	69.1 ± 8.5 [48.0, 85.0]	66 / 91	1.5T
BLSA-3T	960	65.0 ± 14.7 [22.0, 93.0]	525 / 435	3T
UKBB	2202	62.8 ± 7.3 [45.0, 79.0]	1189 / 1013	3T
SHIP	2739	52.6 ± 13.7 [21.2, 90.4]	1491 / 1248	1.5T

Table 2: Summary of participant demographics in ADNI dataset. Age is described in format: mean ± std [min, max]. F and M in gender represent female and male separately. Field indicates the magnetic strength of the MRI scanners.

Study	Subject	CN	AD	Age	Gender (F/M)	Field
ADNI-1	422	229	193	75.5 ± 6.2 [55.0, 90.9]	201 / 221	1.5T
ADNI-2/GO	441	294	147	73.4 ± 6.8 [55.4, 90.3]	221 / 220	3T

Table 3: Multi-layer perceptron (MLP) network implementation details. The network is used for age regression and AD classification tasks. The output size k of the final layer is depends on the task.

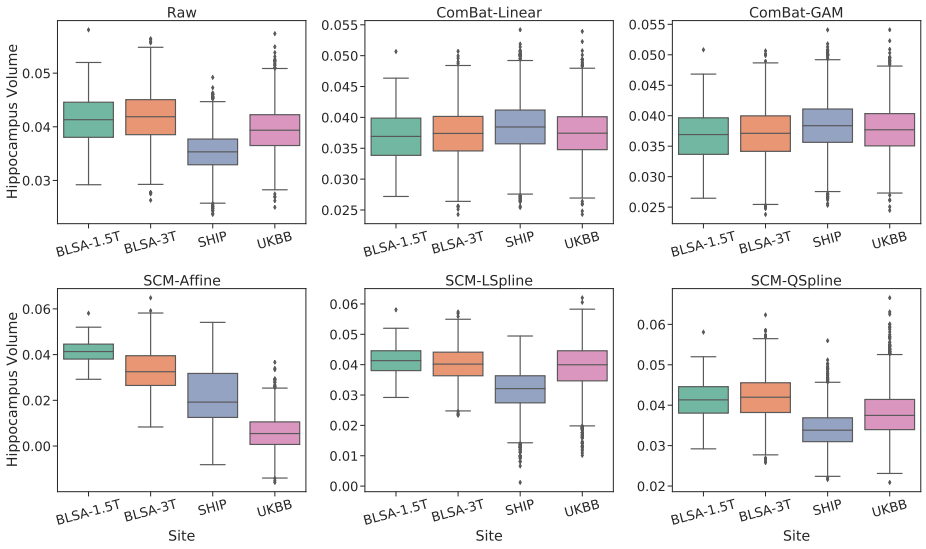
Layer	Input Size	LeakyReLU α	Output Size
Linear + LeakyReLU	145	0.1	72
Linear + LeakyReLU	72	0.1	36
Linear	36	-	k

Table 4: Flow-based SCM implementation details. We directly learn the binary probability of sex s and categorical probability of site t. p_θ^S and p_θ^T are the learnable mass functions of the categorical distribution for variables sex s and site t, and K is the number of site t. The modules indicated with θ are parameterized using neural networks. We constrain age a variable with lower bound (exponential transform) and rescale it with fixed affine transform for normalization. Spline_θ transformation refers to the linear neural spline flows [**?**]. The $\text{ConditionalTransform}_\theta(\cdot)$ can be conditional affine or conditional spline transform, which reparameterizes the noise distribution into another Gaussian distribution. We use linear [**?**] and quadratic [**?**] autoregressive neural spline flows for the conditional spline transform, which are more expressive compared to the affine flows. The transformation parameters of the $\text{ConditionalTransform}_\theta(\cdot)$ are predicted by a context neural network taking $\cdot$ as input. The context networks are implemented as fully-connected networks for affine and spline flows.

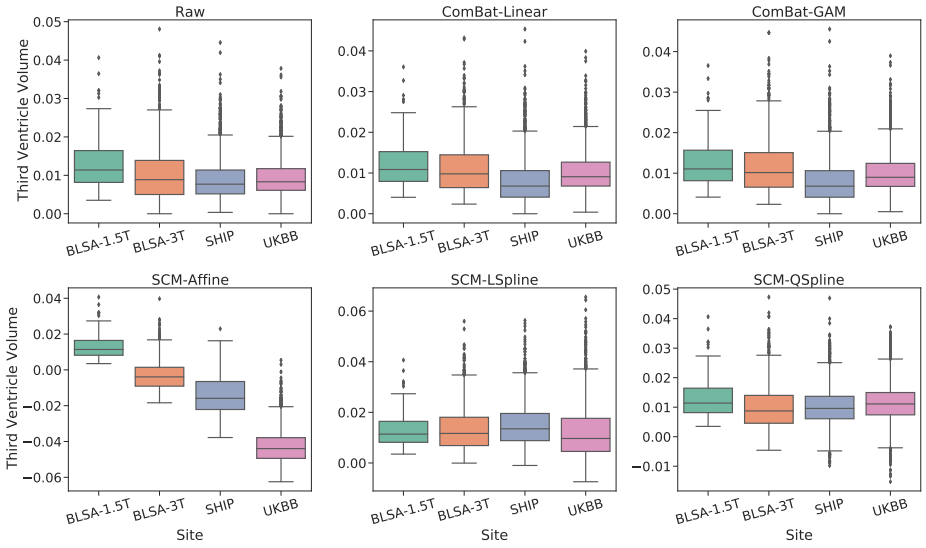
Observations	Exogenous noise
$s := \epsilon_S$	$\epsilon_S \sim \text{Ber}(p_\theta^S)$
$a := f_A(\epsilon_A) = (\text{Spline}_\theta \circ \text{Affine} \circ \text{Exp})(\epsilon_A)$	$\epsilon_A \sim \mathcal{N}(0, 1)$
$t := \epsilon_T$	$\epsilon_T \sim \text{Cat}(K, p_\theta^T)$
$x := f_X(\epsilon_X; s, a, t) = (\text{ConditionalTransform}_\theta([s, a, t]))(\epsilon_X)$	$\epsilon_X \sim \mathcal{N}(0, 1)$

Table 5: Comparison of associative abilities of different type of flows on iSTAGING consortium and ADNI dataset. We observe that spline flows achieved higher log-likelihood compared to that of affine flow for both datasets. This indicates that a flow with higher expressive power helps for density estimation.

Study	Model	Log-likelihood
iSTAGING	Affine	1.8817
	Linear Spline	17.2204
	Quadratic Spline	17.2397
ADNI	Affine	1.8963
	Linear Spline	15.2715
	Quadratic Spline	15.2055

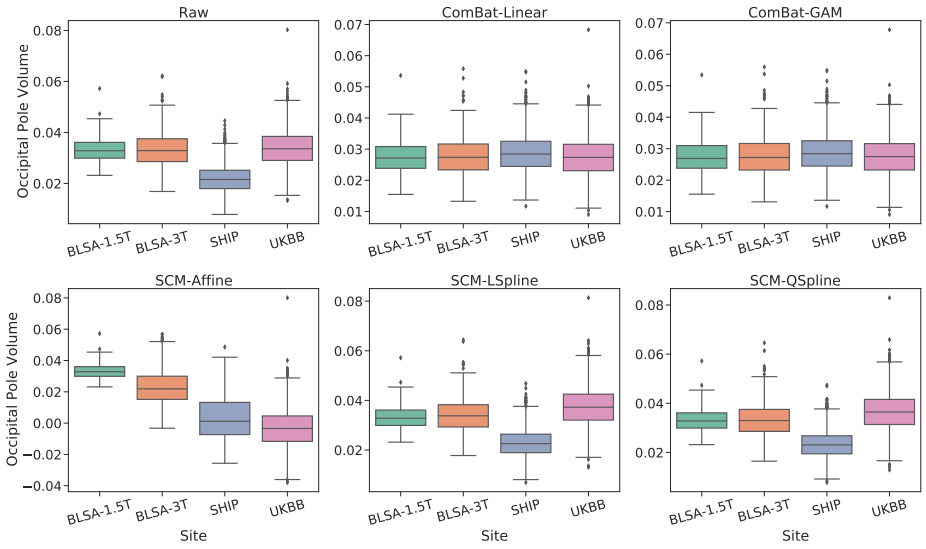


(a) Hippocampus (Right)

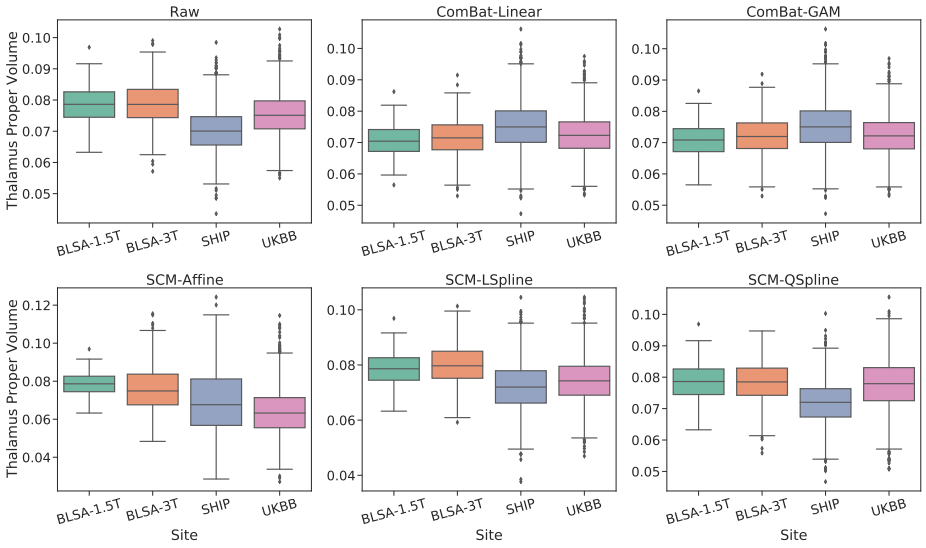


(b) Third Ventricle (Right)

Fig. 1: Comparison of normalized feature distributions cross-site in iSTAGING consortium before and after apply the ComBat methods (ComBat-Linear and ComBat-GAM) and the proposed methods (SCM-Affine, SCM-LSpline, and SCM-QSpline). The distributions of the features harmonized by ComBat methods are aligned cross-site, whereas those harmonized by our proposed method (Q-Spline) are unchanged compared to the raw features. We preserve the unknown cofounders (subject-specific information due to biological variability, such as race, gene, and pathology AD/CN) instead of removing them as site-effects, which is beneficial for downstream analysis, such as AD diagnosis.



(a) Occipital Pole



(b) Thalamus Proper (Right)

Fig. 2: Continued comparison of normalized feature distributions cross-site in iSTAGING consortium before and after apply the ComBat methods (ComBat-Linear and ComBat-GAM) and the proposed methods (SCM-Affine, SCM-LSpline, and SCM-QSpline). The distributions of the features harmonized by ComBat methods are aligned cross-site, whereas those harmonized by our proposed method (Q-Spline) are unchanged compared to the raw features. We preserve the unknown cofounders (subject-specific information due to biological variability, such as race, gene, and pathology AD/CN) instead of removing them as site-effects, which is beneficial for downstream analysis, such as AD diagnosis.