A Appendix

A.1 Proof of Continuous and Unbounded Case

Note that p and P denotes distribution and probability respectively. As $P(X_i{=}\bar{x}_i) \approx 0$ implies $P(X_i{=}\bar{x}_i|\cdot) \approx 0$, as long as the conditioning is on a set of observations that is not extremely improbable. Then:

$$P(X_i'{=}\bar{x}_i|\cdot) = P(X_i'{=}\bar{x}_i, M_i = 1|\cdot) + P(X_i'{=}\bar{x}_i, M_i = 0|\cdot) \tag{12}$$
$$= P(M_i = 1)P(X_i'{=}\bar{x}_i|M_i = 1, \cdot) + P(M_i = 0)P(X_i'{=}\bar{x}_i|M_i = 0, \cdot) \tag{13}$$
$$= P(M_i = 1) + P(M_i = 0)P(X_i{=}\bar{x}_i|\cdot) \approx P(M_i = 1) \tag{14}$$
$$\Rightarrow p(Y|X_i'{=}\bar{x}_i, \cdot) = \frac{p(Y|\cdot)P(X_i'{=}\bar{x}_i|Y, \cdot)}{P(X_i'{=}\bar{x}_i|\cdot)} \approx \frac{p(Y|\cdot)P(M_i = 1)}{P(M_i = 1)} \approx p(Y|\cdot) \tag{15}$$

A similar analysis can be performed to derive the approximate versions of equations of Eq. (5) to (6), which are special cases.

A.2 Suboptimal Choice of Placeholder Values

If $\bar{x}_i \in \text{support}(X_i)$, then $X_i' = \bar{x}_i$ is either because of two mutually exclusive cases:

$$\begin{cases} M_i = 1 \text{ or} \\ M_i = 0 \text{ and } X_i = \bar{x}_i \end{cases}$$

Let $P(M_i = 0) = r$ and $\boldsymbol{x}_j \neq \bar{x}_j, \forall j \neq i$. Assuming that $P(X_i'{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) > 0$, then:

$$p(Y|X_i'{=}\bar{x}_i, \boldsymbol{X}'_{-i}{=}\boldsymbol{x}_{-i}) = p(Y|X_i'{=}\bar{x}_i, \boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) = \frac{p(Y, X_i'{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})}{P(X_i'{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})} \tag{16}$$
$$P(X_i'{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) = P(M_i = 1) + P(M_i = 0, X_i{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) \tag{17}$$
$$= 1 - r + P(X_i{=}\bar{x}_i|M_i = 0, \boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})P(M_i = 0|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) \tag{18}$$
$$= 1 - r + r \times P(X_i{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) \tag{19}$$
$$p(Y, X_i'{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) = p(Y|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})P(X_i'{=}\bar{x}_i|Y, \boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) \tag{20}$$
$$= p(Y|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})(1 - r + r \times P(X_i{=}\bar{x}_i|Y, \boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})) \tag{21}$$

From Eq. (19) and (21),

$$p(Y|X_i'{=}\bar{x}_i, \boldsymbol{X}'_{-i}{=}\boldsymbol{x}_{-i}) = p(Y|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})\frac{1 - r + r \times P(X_i{=}\bar{x}_i|Y, \boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})}{1 - r + r \times P(X_i{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})} \tag{22}$$

Thus, $p(Y|X_i'{=}\bar{x}_i, \boldsymbol{X}'_{-i}{=}\boldsymbol{x}_{-i}) = p(Y|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})$ is equivalent to

$$\Leftrightarrow 1 - r + r \times P(X_i{=}\bar{x}_i|Y, \boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) = 1 - r + r \times P(X_i{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) \tag{23}$$
$$\Leftrightarrow P(X_i{=}\bar{x}_i|Y, \boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) = P(X_i{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i}) \tag{24}$$
$$\Leftrightarrow Y \perp\!\!\!\perp X_i{=}\bar{x}_i|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i} \tag{25}$$

It is difficult to find $\bar{x}_i$ to satisfy Eq. 25, most likely $p(Y|X_i'{=}\bar{x}_i, \boldsymbol{X}'_{-i}{=}\boldsymbol{x}_{-i}) \neq p(Y|\boldsymbol{X}_{-i}{=}\boldsymbol{x}_{-i})$.

For example, consider the case where $E_X = \{x_1, x_2\}, E_Y = \{0, 1\}, x_0 := x_1$, and $\alpha = 0.5$.

$$\mathrm{P}(X = x_1) = 0.3, \quad \mathrm{P}(X = x_2) = 0.7 \tag{26}$$

$$\mathrm{P}(Y = 0|X = x_1) = \mathrm{P}(Y = 1|X = x_2) = 1 \tag{27}$$

$$\Rightarrow \mathrm{P}(Y = 0) = 0.3, \quad \mathrm{P}(Y = 1) = 0.7 \tag{28}$$

$$\mathrm{P}(Y = 0|X' = x_1) = \frac{\mathrm{P}(Y = 0, X' = x_1)}{\mathrm{P}(X' = x_1)} \tag{29}$$

$$= \frac{\mathrm{P}(Y = 0, M_X = 1) + \mathrm{P}(Y = 0, M_X = 0, X = x_1)}{\mathrm{P}(M_X = 1) + \mathrm{P}(M_X = 0, X' = x_1)} \tag{30}$$

$$= \frac{0.3 * 0.5 + 0.5 * \mathrm{P}(X = x_1)\mathrm{P}(Y = 0|X = x_1)}{0.5 + 0.5 * 0.3} \tag{31}$$

$$= \frac{0.3 * 0.5 + 0.5 * 0.3 * 1}{0.5 + 0.5 * 0.3} = \frac{2 * 0.3 * 0.5}{0.5 + 0.5 * 0.3} = \frac{0.6}{1.3} \tag{32}$$

$$\Rightarrow \mathrm{P}(Y = 0|X' = x_1) \neq \mathrm{P}(Y = 0) = 0.3 \tag{33}$$

$$\Rightarrow \mathrm{P}(Y = 0|X' = x_1) \neq \mathrm{P}(Y = 0|X = x_1) = 1 \tag{34}$$

A.3 Analysis of Placeholders for Structured Knockout

In this subsection, we analyze the effect of different placeholder values on a structured Knockout task, specifically the multi-modal tumor segmentation task from Section 4.4. We scale image intensity values to $[0, 1]$ per image. We train three Knockout models with the following placeholders: a constant image of -1s, a constant image of 0s, and the mean of all images per modality. At inference time, we evaluate on all modality missingness patterns. In the event that all images are missing, we randomly select one so that the model sees at least one image. For Knockout-trained models, the corresponding placeholder is imputed for missing images.

Fig. 6 shows the results. Interestingly, we observe the mean placeholder (Knockout*) performs better than constant-image placeholders, and the constant image of 0s generally outperforms the constant image of -1s. We hypothesize that in the context of structured inputs like images in conjunction with limited data and model capacity, placeholders which balance feasibility with practical considerations like causing unstable gradients due to out-of-range inputs is an important consideration.

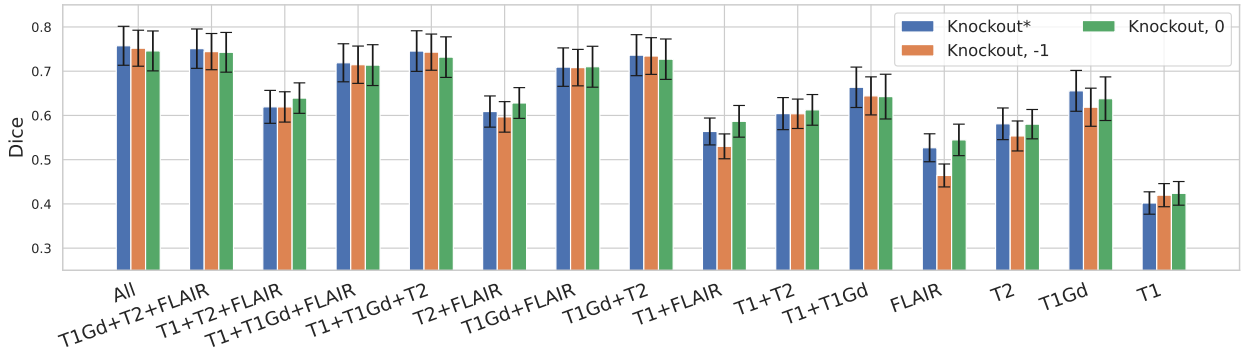


Figure 6: Dice performance of multi-modal tumor segmentation across varying missingness patterns of modality images. Knockout-trained models only. We observe mean placeholders perform better than constant-image placeholders. Error bars depict the 95% confidence interval over test subjects.

B Experimental Details

All experiments were performed with access to a machine equipped with an AMD EPYC 7513 32-Core processor and an Nvidia A100 GPU. All code is written in PyTorch.

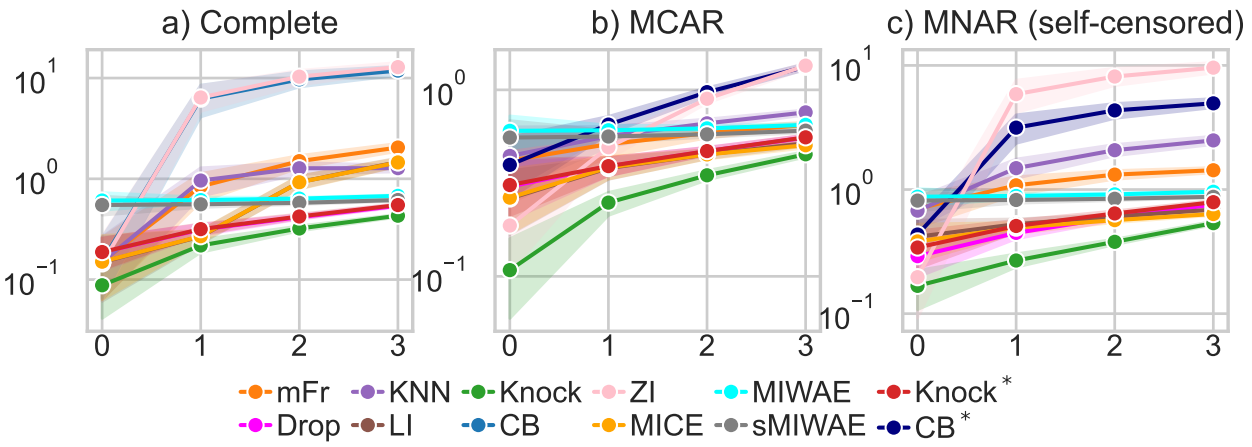


Figure 7: Test MSE evaluated against observations (Y) from 10 repetitions of the regression simulation. Lower is better. X axis indicates the number of missing variables at inference time. a) Complete training data. b) Missing completely at random (MCAR) training data. c) Missing not at random (MNAR) training data

B.1 Simulations

B.1.1 Regression

In each repetition, the data are sampled from a 10-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The mean vector $\boldsymbol{\mu}$ is sampled uniformly from the interval $[0, 1]$, i.e. $\boldsymbol{\mu} \sim \mathsf{Uniform}(0, 1) \in \mathbb{R}^{10}$. The covariance matrix is sampled as $\boldsymbol{\Sigma} := \boldsymbol{W}^T\boldsymbol{W}$, whereby $\boldsymbol{W} \sim \mathsf{Uniform}(0, 1) \in \mathbb{R}^{10\times 10}$. The full training and test datasets are then generated using the covariance matrix (fixed). This ensures consistency and avoids introducing bias that could affect predictive imputation methods. The first 9th variables of the multivariate Gaussian are assigned as $\boldsymbol{X}$ ($\boldsymbol{X} \in \mathbb{R}^9$) and the 10th variable is assigned as Y ($Y \in \mathbb{R}$). The Knockout rate is set at 0.0741 so that half of the mini-batches have no induced missing variable.

In addition to the MMSE-minimizing Bayes optimal predictions: $\mathbb{E}[Y|\boldsymbol{X}]$, we also evaluate the models' predictions against the observed values of Y (Fig. 7). To assess model robustness across varying data availability, we experiment with training set sizes of 300, 1000, and 15000 samples (compared to the original 3000-sample setting). This ablation focuses exclusively on fully observed training data. As illustrated in Fig. 8, Knockout consistently outperforms all baseline methods across all levels of missingness. With only 300 samples, Knockout achieves the lowest mean squared error (MSE) at every level of missingness. This performance trend continues at 1000 samples, with Knockout again leading. Although all methods benefit from the increased training data at 15000 samples, Knockout maintains a clear and consistent performance advantage. These results underscore Knockout's robustness and strong generalization, especially under high missingness and limited data conditions.

B.1.2 Binary Classification

We evaluate the prediction error rate with full features ($\boldsymbol{X}$) and missing feature (only X_1 or X_2 as input). We also evaluate how close the models' predicted probability distributions with missing feature are against the marginal distributions (Fig. 9a and Fig. 9a top and left panels) using Jensen-Shannon divergence. The marginals distributions are estimated empirically using all data.

Continuous Inputs. All input variables are continuous ($\boldsymbol{X} \in \mathbb{R}^2$). Knockout achieves similar error rate compared to standard training but much better performance when a variable is missing (Table 8). Knockout* performs worse than Knockout due to the sub-optimal choice of placeholders.

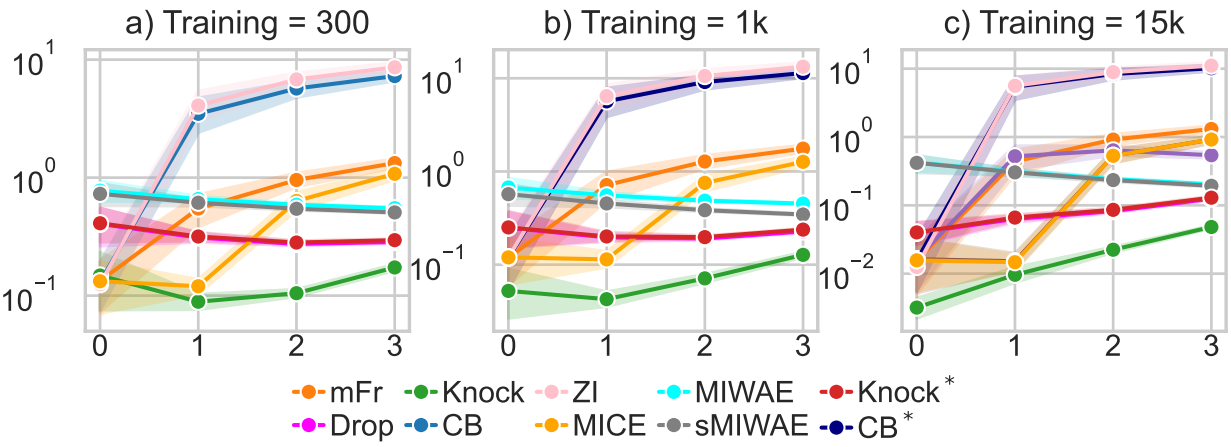
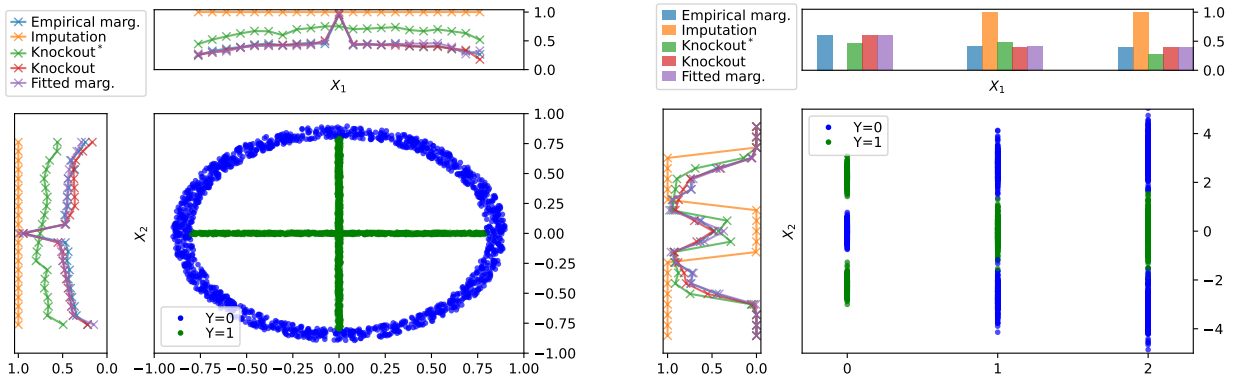


Figure 8: Knockout at low-data and high-data settings. Test MSE evaluated against Bayes optimal prediction ($\mathbb{E}[Y|\boldsymbol{X}]$) averaged over 10 repetitions (lower is better). X axis indicates the number of missing variables at inference time.



(a) Continuous inputs

(b) Mixed inputs

Figure 9: Visualization of the two classification simulations. Knockout's estimates of the marginal distributions (i.e. $P(Y|X_1)$ and $P(Y|X_2)$, denoted by red lines) are closer to the empirical estimates (blue lines) than baselines'. Top: $P(Y{=}1|X_1)$ estimated empirically and estimated by various approaches. Left: Various estimates of $P(Y{=}1|X_2)$. Bottom Right: Data visualization.

Mixed Inputs. $\boldsymbol{X}$ consists of a binary variable and a continuous variable (i.e. $X_1 \in \{0, 1\}$, $X_2 \in \mathbb{R}$). We also include a common baseline (CB) approach that imputes using $\bar{x}$. CB ($\bar{x}$) is similar to CB but impute missing variables using new categories and out-of-support values during inference. Since CB ($\bar{x}$) lacks a learned representation for these values, its performance is less reliable and may not generalize well. In contrast, Knockout offers greater robustness by explicitly modeling missingness during training. Knockout achieves better results than baselines in all scenarios (Table 8).

B.2 Alzheimer's Forecasting

All participants used in this work are from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and

Table 8: Classification simulations. Best results are in bold. Err.: Proportion of test error. JSD: Jensen–Shannon divergence of the estimated and empirical marginal.

	Missing Rate = 0	Missing Rate = 1/2			
Method	Err. ($\boldsymbol{X}$)	Err. (X_1)	JSD (X_1)	Err. (X_2)	JSD (X_2)
Continuous inputs					
Common baseline	**0.0003**	0.3970	∞	0.4001	∞
Knockout* (Ours)	0.0210	0.3549	0.0179	0.3727	0.0214
Knockout (Ours)	0.0007	**0.2559**	**0.0003**	**0.2563**	**0.0007**
Fitted Marginals	N/A	0.2531	0.0006	0.2600	0.0006
Mixed inputs					
Common baseline	0.0032	0.5972	∞	0.5410	∞
Common baseline ($\bar{x}$)	0.0032	0.4843	∞	0.4137	∞
Knockout* (Ours)	0.0187	0.4843	0.0038	0.3410	0.0073
Knockout (Ours)	**0.0031**	**0.4028**	**0.0001**	**0.2844**	**0.0009**
Fitted Marginals	N/A	0.4028	0.0000	0.2809	0.0008

Table 9: Summary statistics of the participants at baseline in the Alzheimer's Disease data. Mean ± standard deviations are listed. APOE4 row represents the number of alleles.

Characteristic	(n=789)
Female/Male	324/465
Age (yr)	73.46 ± 7.39
Education (yr)	15.93 ± 2.81
APOE4 (0/1/2)	371/313/98
CDR	1.55 ± 0.89
MMSE	27.52 ± 1.82

neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

We select the participants who have mild cognitive impairment (MCI) at the baseline (screening) visit and had at least one follow-up diagnosis within the next five years. We excluded participants who were diagnosed as CN in a later follow-up year (n=284) since these subjects might have been diagnosed incorrectly at some point. After this exclusion, we are left with 789 participants. Table 9 lists summary statistics for the participants; including sex, age, number of years of education completed, count of Apolipoprotein E4 (APOE4) allele, Clinical Dementia Rating(CDR), and Mini Mental State Examination (MMSE) scores at baseline. As is common in many real-world longitudinal studies, ADNI experiences missing follow-up visits, irregular timings, and high dropout rates before the study's planned end. Table 10 shows the number of subjects available in each diagnostic category for annual follow-ups. In Table 10 and all analyses, any subject who progressed from MCI to AD before withdrawing was considered to remain in the AD state until the fifth year, reflecting AD's irreversible nature. We employed the reweighted cross-entropy loss scheme introduced in (Karaman et al., 2022) during training to account for the imbalance in diagnoses.

We use features from several domains. Demographics include age and years of education (PTEDUCAT), and genotype is represented by APOE4 allele count. Cognitive assessments include CDR, FAQ, ADAS (11, 13, Q4), MMSE, RAVLT, LDELTOTAL, TRABSCOR, mPACCdigit and mPACCtrailsB. Biomarkers include CSF measures (ABETA, TAU, PTAU), MRI-derived brain volumes (ventricles, hippocampus, whole brain, entorhinal, fusiform, mid-temporal, and ICV) from FreeSurfer (Desikan et al., 2006; Fischl et al., 2004),

Table 10: The number of available subjects in each diagnostic group for annual follow-up visits in the Alzheimer's Disease data. The follow-up diagnoses are not actually exactly 12 months apart. They have been rounded to the nearest time horizon in years.

Follow-up year	1	2	3	4	5
MCI	674	431	317	202	127
AD	110	218	261	286	292

Table 11: The degree of missingness (%) in different data modalities in the Alzheimer's Disease data.

Data Type	Missingness Rate (%)
Demographics	0.06
Genotype	0.89
Cognitive assessments	0.20
CSF	37.14
MRI	21.22
FDG	24.08

and PET SUVR for the FDG tracer. All features are numerical and z-score normalized using training-set statistics. Missing rates per modality are shown in Table 11.

We note that we train our models using the hyperparameters stated in (Karaman et al., 2022). Fig. 10 shows the AUROC scores obtained using the complete portion of the dataset ($n = 256$ subjects). In this experiment, the training data has no observed missing variables. These results are similar to the results included in the main text, where Knockout outperforms the baseline and the choice of the appropriate placeholder has an impact on the performance.

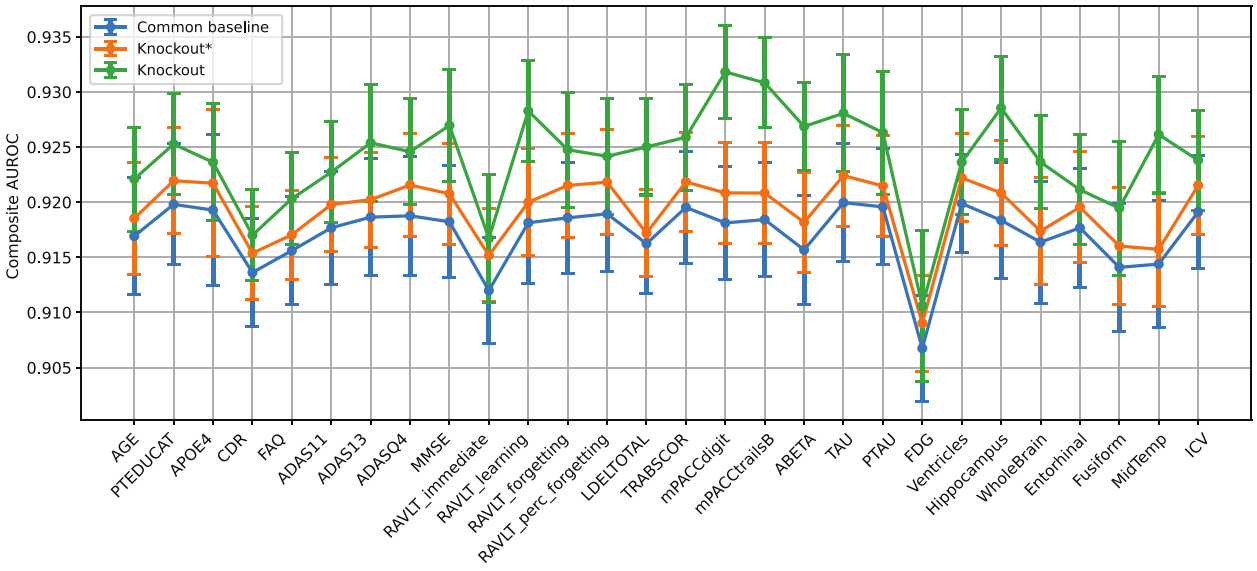


Figure 10: AUROC scores obtained for the three model variants when each input feature is missing during inference (x-axis) for the complete data case in the Alzheimer's Disease forecasting experiment. Displayed are averages of 10 train-test splits. Error bars indicate the standard error across these splits.

B.3 Multi-modal Tumor Segmentation

The RSNA-ASNR-MICCAI BraTS (Baid et al., 2021) challenge releases a dataset of 1251 subjects with multi-institutional routine clinically-acquired multi-parametric MRI scans of glioma. Each subject has 4 modalities: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). All the imaging datasets have been annotated manually, by one to four raters, following the same annotation protocol, and their annotations were approved by experienced neuro-radiologists. Annotations comprise the GD-enhancing tumor (ET - label 3), the peritumoral edematous/invaded tissue (ED - label 2), and the necrotic tumor core (NCR - label 1).

The following pre-processing is applied: co-registration to the same anatomical template, interpolation to the same resolution (1 mm3), skull-stripped, and min-max normalized to the range [0, 1]. The ground truth data were created after their pre-processing. For training, we use 80%/5%/15% data split of the subjects for training/validation/testing.

For the segmentation model, we use a 3D UNet with 4 downsampling layers and 2 convolutional blocks per resolution (Ronneberger et al., 2015). We minimize a sum of cross-entropy loss and Dice loss with equal weighting and use an Adam optimizer with a learning rate of 1e-3.

B.4 Prostate Cancer Detection

A common clinical workflow for the diagnosis of prostate cancer is to detect and localize abnormalities from 3 MR modalities: T2-weighted (T2w), diffusion-weighted (DWI) and apparent diffusion coefficient (ADC) images (Turkbey et al., 2019). T2w images provide anatomical details, while DWI and ADC highlight restricted diffusion, which can be a sign of malignancy.

We divided 1500 biparametric MR image sets provided from Prostate Imaging: Cancer AI (PICAI) challenge (Saha et al., 2022) "training" dataset into training, validation, test sets in a 0.6/0.2/0.2 ratio. Among the 1500 cases, 425 were confirmed as cancer by biopsy. DWI and ADC images are registered to T2w images and all images are cropped around prostate and resized to $100 \times 100 \times 40$. For the modality-wise classification tasks, we used 3D CNN with 4 blocks, each with a convolution layer, BatchNorm, leakly ReLU activation and average pooling layer, followed by fully connected layer. We trained the models to predict PCa using binary cross entropy loss and an Adam optimizer with a learning rate of $1e-3$.

Table 12: AUC performance for prostate cancer detection from the ensemble baseline, common baseline, and Knockout, across varying missingness patterns at inference time. Each column represents non-missing modalities. Best results in **bold**, second-best underlined.

	T2	ADC	DWI	ADC +DWI	T2 +DWI	T2 +ADC	All
Ensemble	$0.683_{\pm 0.013}$	$\mathbf{0.786}_{\pm 0.010}$	$0.718_{\pm 0.007}$	$0.780_{\pm 0.005}$	$0.722_{\pm 0.005}$	$\underline{0.766}_{\pm 0.006}$	$\underline{0.766}_{\pm 0.004}$
Common	$\underline{0.687}_{\pm 0.011}$	$\underline{0.771}_{\pm 0.011}$	$\underline{0.720}_{\pm 0.007}$	$\underline{0.784}_{\pm 0.003}$	$\underline{0.727}_{\pm 0.006}$	$\mathbf{0.771}_{\pm 0.008}$	$\mathbf{0.774}_{\pm 0.004}$
Knockout	$\mathbf{0.694}_{\pm 0.009}$	$0.730_{\pm 0.019}$	$\mathbf{0.736}_{\pm 0.009}$	$\mathbf{0.789}_{\pm 0.005}$	$\mathbf{0.744}_{\pm 0.004}$	$0.753_{\pm 0.011}$	$\mathbf{0.774}_{\pm 0.007}$

B.5 Privileged information for noisy label learning

We briefly introduce two datasets we used for this experiment: CIFAR-10H (Peterson et al., 2019) and CIFAR-10/100N (Wei et al., 2021). CIFAR-10H relabels the original CIFAR-10 10K test set with multiple annotators and provides high-quality sample-wise annotation information such as annotator ID, reaction time and annotator confidence as PI. Following a previous setup (Wang et al., 2023), we test on the high-noise version of CIFAR-10H, by selecting incorrect labels when available, denoted as "CIFAR-10H Worst". The estimated noise rate is 64.6%. While we train on the high-noise version, testing is conducted on the original CIFAR-10 50K training set. CIFAR-10/100N provides multiple annotations for CIFAR-10/100 training set. The raw data also includes information about annotation process. But this information is

Datasets	PI quality	SOP	Common baseline	Knockout*	Knockout
CIFAR-10H (Worst)	High	$51.3_{\pm 1.9}$	$55.2_{\pm 0.8}$	$\underline{56.9_{\pm 0.59}}$	$\mathbf{57.4_{\pm 0.6}}$
CIFAR-10N (Worst)	Low	$\mathbf{85.0_{\pm 0.8}}$	$82.3_{\pm 0.3}$	$83.6_{\pm 0.7}$	$\underline{84.7_{\pm 0.7}}$
CIFAR-100N (Fine)	Low	$61.9_{\pm 0.6}$	$60.7_{\pm 0.6}$	$\underline{61.6_{\pm 0.6}}$	$\mathbf{62.1_{\pm 0.3}}$

Table 13: Test accuracy of different methods on noisy label dataset with PI. Best results in **bold**, second-best underlined.

provided as averages over batches of examples rather than sample-wise. The estimated noise rate is 40.2% for CIFAR-10/100N.

For all CIFAR experiments and baselines, we use the Wide-ResNet-10-28 (Zagoruyko & Komodakis, 2016) architecture. We use SGD optimizer with 0.9 Nesterov momentum, a batch size of 256, 0.1 learning rate and 1e-3 weight decay and minimized the cross-entropy loss with respect to the provided labels. The total training epoch is 90, and the learning rate decayed by a factor of 0.2 after 36, 72 epochs. For the PI features, we use annotator ID and annotation reaction time. In PI features are normalized to [0, 1] for preprocessing. For Knockout, during training, we randomly knock out all PI features at 50% rate and use -1 as placeholder value. All experiments are performed on one A6000. In Table 13, we further show results for common baseline, Knockout* and Knockout.