Bridging Deterministic and Probabilistic Deep Learning Imputation via Post-Hoc Uncertainty Quantification

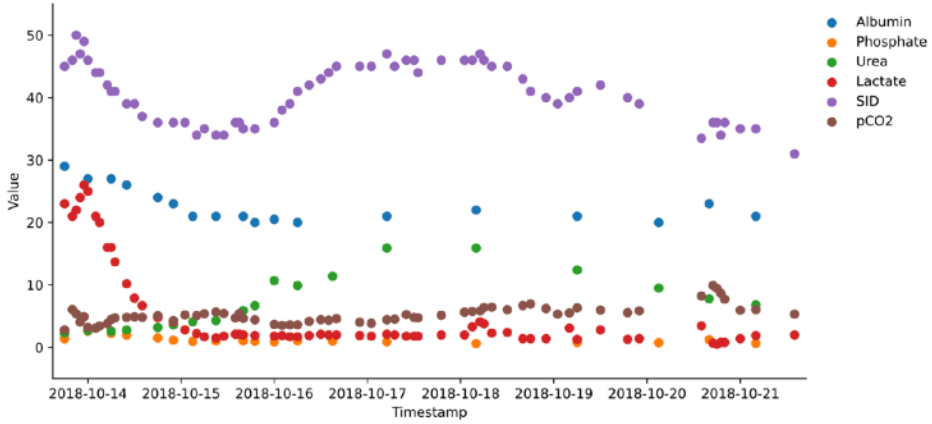
Ali Septiandri

3 September 2026

Background

- In ICU settings, data come from multiple sources and are inherently related
- Measurements collected at irregular intervals (informative sampling)—aligning them will result in missing values
- Cannot always get more samples! Some measurements are invasive (Siegal et al., 2023)

Challenges



- We want to impute missing values…
- but traditional imputation often ignores temporal structure (e.g. MICE) & uncertainty (e.g. deep learning)
- Need for robust, uncertainty-aware imputation in critical care datasets

On uncertainty quantification (UQ)

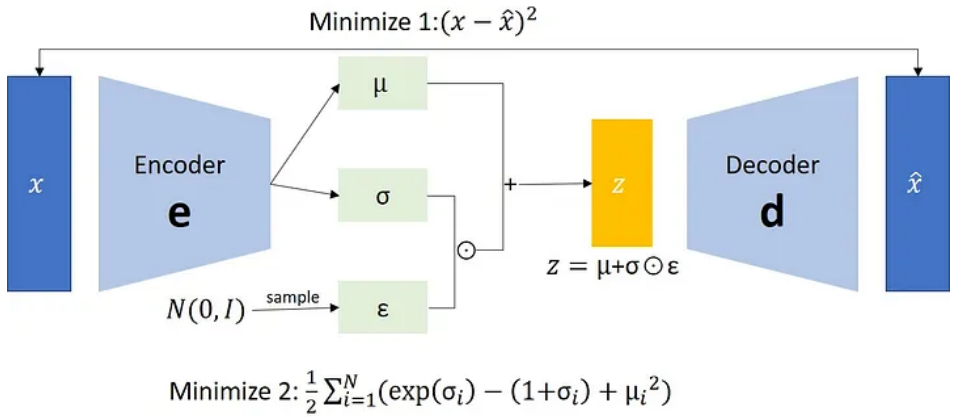
- Medical observations are inherently uncertain
  - Coming from measurement errors
  - Using surrogate markers
  - Leading to unreliable model predictions (Cabitza et al., 2017)
- *Alarm fatigue* (Embi & Leonard, 2012; Umscheid et al., 2015)
  - Alerts triggered by prediction tools are often not accompanied by a clinically actionable change

VAEs for imputation

- For UQ in critical care data imputation, previous studies mainly used variational autoencoders (VAEs)
- Diffusion models have been explored more recently, but they are expensive to train

Model	Year	Architecture	Component Bias	Uncertainty Quantification
Modified Single Architectures				
MRNN [43]	2017	RNN	Sequential	✘
GRUD [10]	2018	RNN	Sequential	✘
BRITS [42]	2018	RNN	Sequential	✘
GLIMA [47]	2020	Attention	Globality	✘
MTSIT [48]	2022	Attention	Globality	✘
SAITS [49]	2023	Attention	Globality	✘
Tiled CNN [50]	2015	CNN	Locality	✘
TimesNet [51]	2023	CNN	Locality	✘
TSI-GNN [54]	2021	GNN	Relational	✘
VAEs				
MIWAE [57]	2019	CNN	Locality, Stochasticity	✔
GP-VAE [58]	2020	CNN	Locality, Stochasticity	✔
HI-VAE [59]	2020	MLP	Stochasticity	✔
V-RIN [60]	2021	RNN	Sequential, Stochasticity	✔
Shi-VAE [61]	2022	RNN	Sequential, Stochasticity	✔
supnot-MIWAE [62]	2023	CNN, Attention	Locality, Globality, Stochasticity	✔
MDNs				
CDNet [63]	2022	RNN	Sequential, Mixture	✔
GAN				
VIGAN [64]	2017	CNN	Locality, Adversariality	✘
GRUI-GAN [65]	2018	RNN	Sequential, Adversariality	✘
E^2GAN [66]	2019	RNN	Sequential, Adversariality	✘
NAOMI [67]	2019	MLP	Adversariality	✘
US-GAN [68]	2021	RNN	Sequential, Adversariality	✘
Sim-GAN [69]	2022	CNN	Locality, Adversariality	✘
Diffusion				
CSDI [70]	2021	Attention	Globality, Gradualism	✔
SSSD [71]	2023	CNN	Locality, Gradualism	✔
CSBI [72]	2023	CNN, Attention	Locality, Globality, Gradualism	✔
DA-TASWDM [73]	2023	Attention	Globality, Gradualism	✘

Issues in VAEs



- Usually assuming Gaussian distribution in the latent space
- Instability in optimisation—balancing reconstruction and latent space regularisation (Dehaene & Brossard, 2021; Yacoby et al., 2020)
- Some deterministic models could achieve lower errors

Better performance with deterministic models

Model	MAE	MSE
BRITS	0.297 (0.001)	0.338 (0.001)
MRNN	0.708 (0.029)	0.921 (0.044)
GRUD	0.450 (0.004)	0.492 (0.006)
TimesNet	0.353 (0.003)	0.355 (0.004)
SAITS	0.257 (0.019)	0.276 (0.018)
CSDI	0.252 (0.004)	0.313 (0.039)
GP-VAE	0.445 (0.006)	0.499 (0.011)
US-GAN	0.310 (0.003)	0.318 (0.005)

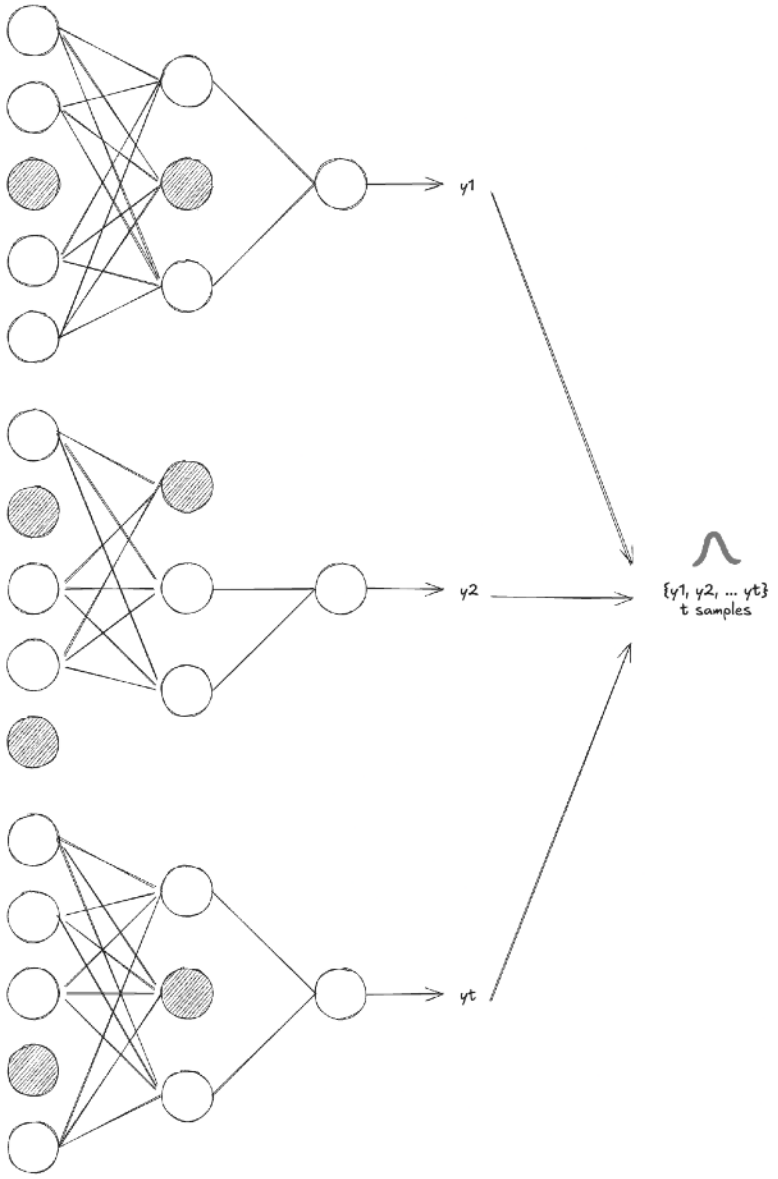
- BRITS (Cao et al., 2018) & SAITS (Du et al., 2023) achieve low errors, but do not come with UQ by default
- While CSDI (a diffusion-based model) could achieve a slightly lower MAE, the training time is nearly 500x longer (Qian et al., 2025)

Applying post-hoc UQ to best performing models

- We propose to apply two UQ methods to SAITS and BRITS
  - Monte Carlo (MC) Dropout
  - Last-layer Laplace Approximation (LLLA)
- The error performance should be relatively undisturbed, but we can still get UQ

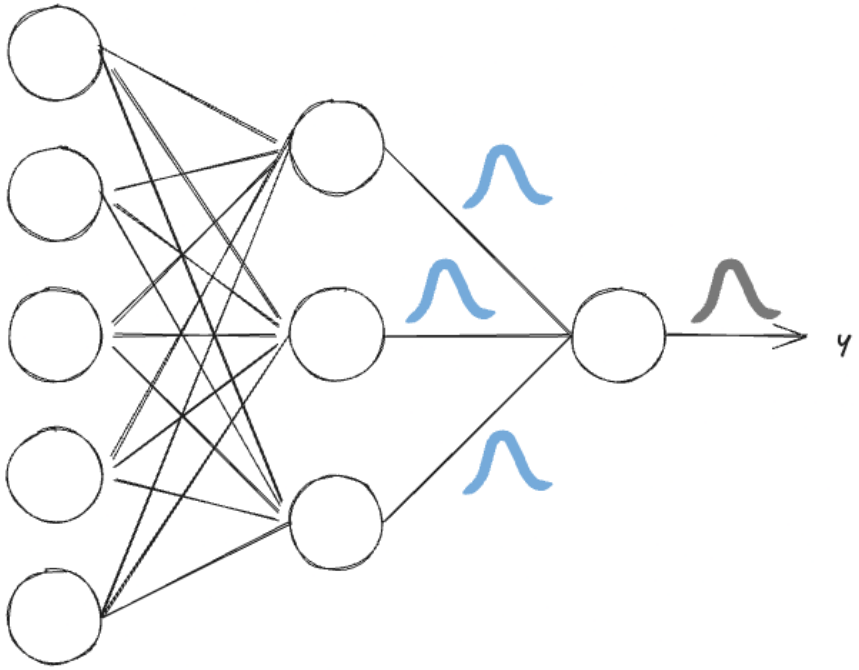
Monte Carlo Dropout (Gal & Ghahramani, 2016)

- Using an already trained model with dropout layers
- Doing inference while randomly turning off hidden units with probability p
- From t samples, we can compute
  - Predictive mean μ
  - Predictive variance σ^2



Is the Last Layer Sufficient for Uncertainty Quantification?

- In short: yes (Kristiadi et al., 2020; Wilson et al., 2026)
- Linearising DNNs to form Bayesian Generalised Linear Models (GLMs)
- LL-GLM matches or *beats* the full-network version on 6 of 9 datasets by NLL, and is never meaningfully worse



Experiments

- **Datasets used**
  - PhysioNet 2012
  - HiRID
- **Preprocessing:** Hourly discretisation, z-score normalisation, block masking to simulate missingness
- **Benchmarks:**
  - Last observation carried forward (LOCF)
  - Median
  - GP-VAE
  - CSDI

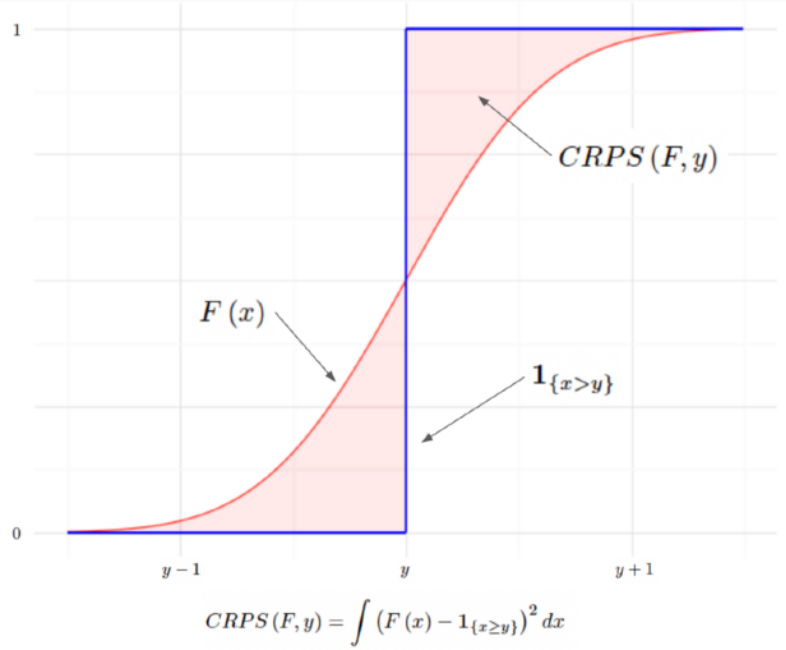
Model evaluation

- Three levels of missingness: 10%, 20%, 30%
- Two evaluation metrics
  - Mean absolute error – imputation accuracy

$$\text{MAE} = \frac{1}{N \times D} \sum_{i=1}^{N} \sum_{d \in D} |Y_{id} - \hat{Y}_{id}|$$

  - Continuous Ranked Probability Score – uncertainty quantification

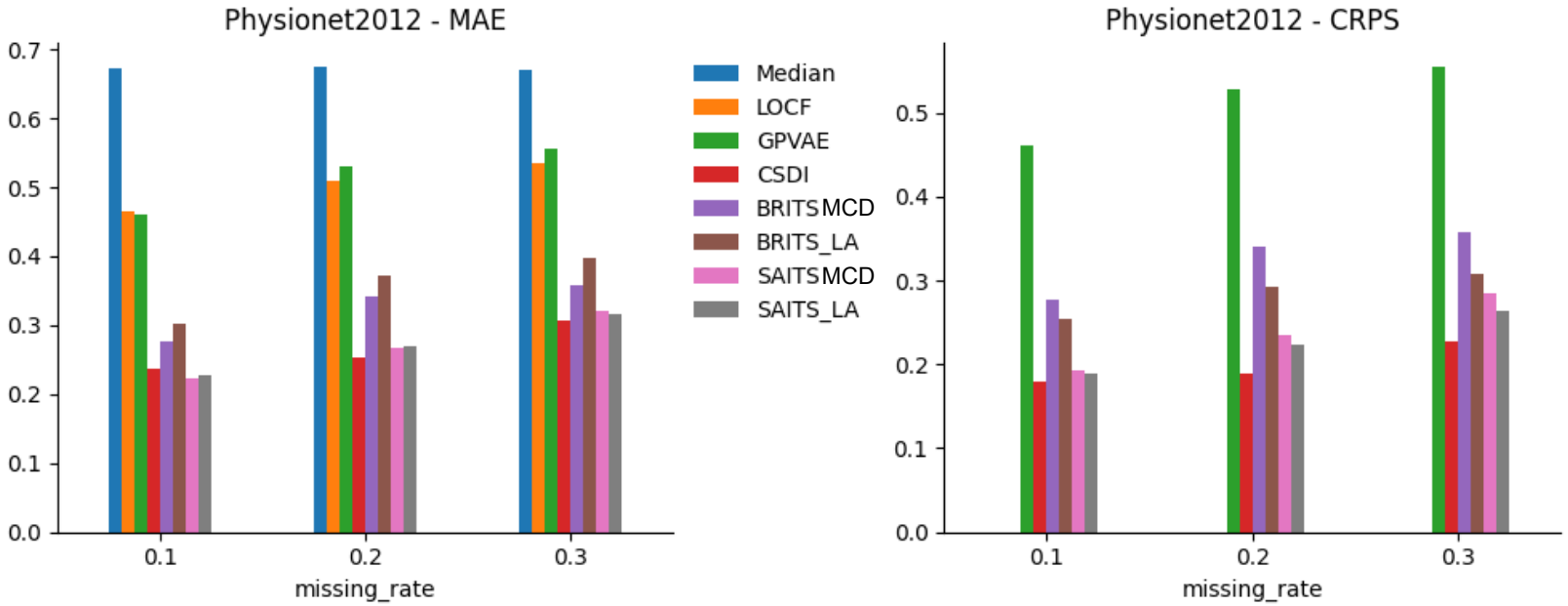
$$\text{CRPS}(F, y) = \int (F(x) - 1_{\{x \geq y\}})^2 dx$$



Visualisation of the CRPS. The predicted distribution is marked in red, and the ground truth's degenerate distribution is marked in blue. The CRPS is the (squared) area trapped between the two CDFs.

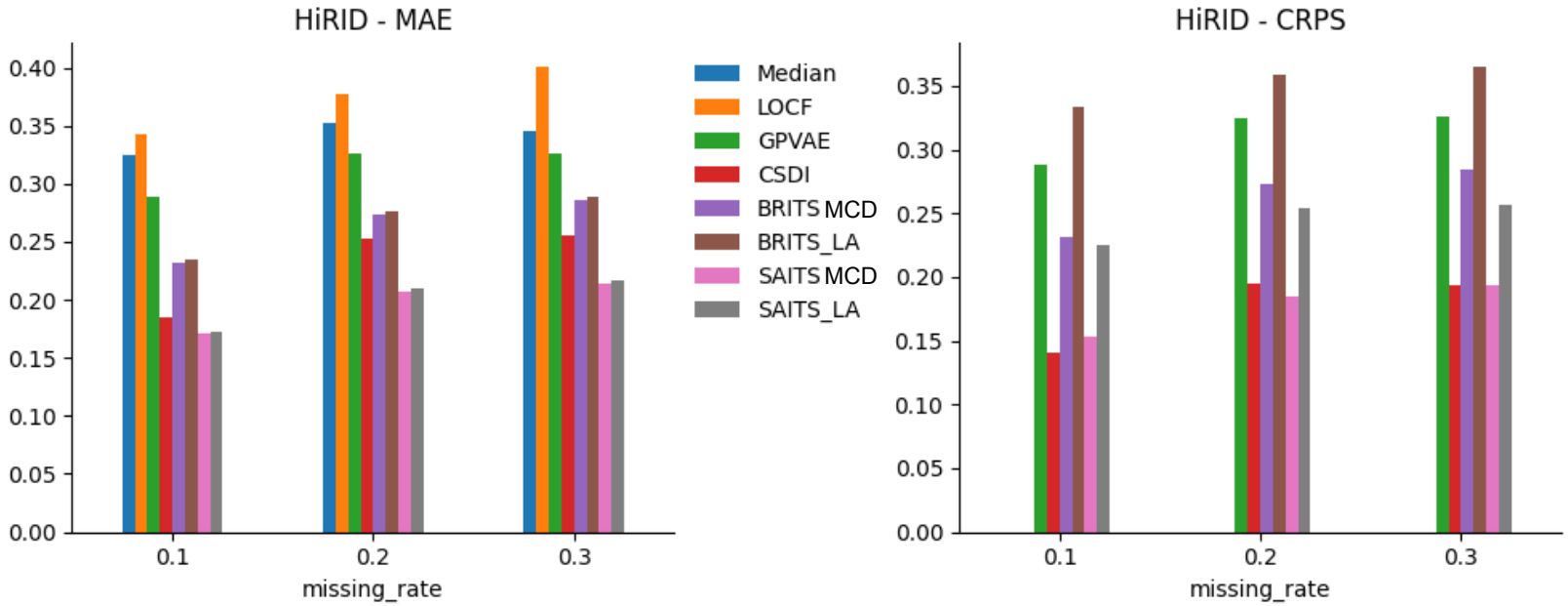
PhysioNet 2012

- Errors are increasing with higher missingness for each model
- CSDI outperforms other models
- SAITS can get close to CSDI
- LLLA produces better UQ compared to MC Dropout



HiRID

- CSDI only achieves the best CRPS at missing rate = 0.1
- SAITS outperforms CSDI
- SAITS + MC Dropout is better than + LLLA



Contribution

Because of the mixed results in the two post-hoc UQ methods, can we combine them instead?

Combining UQ from both post-hoc methods

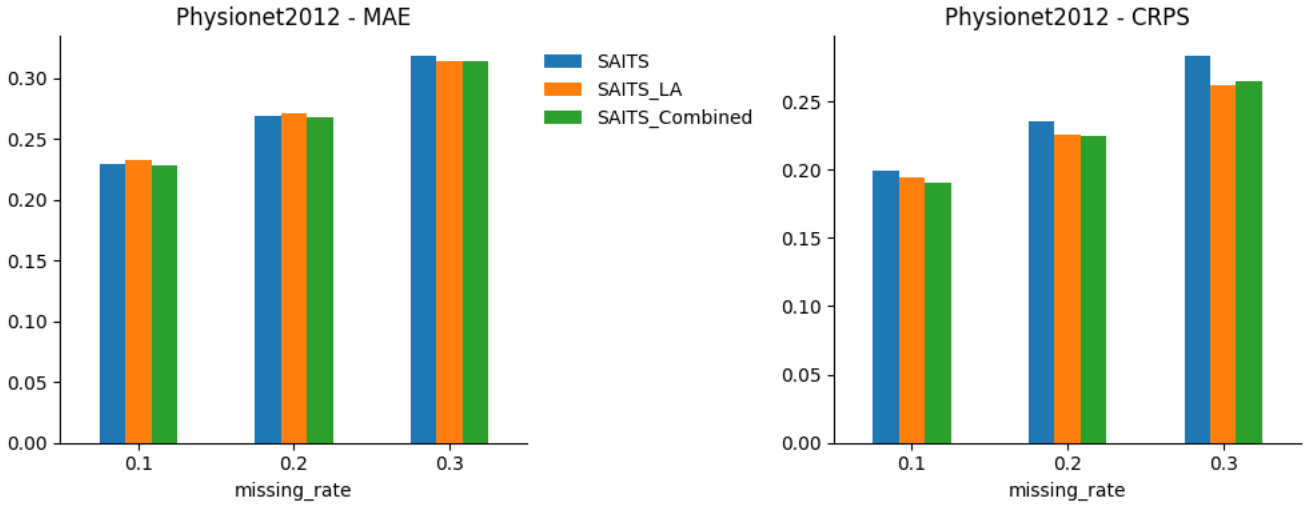
1. Train the model once
2. For inference
    a. Run the inference with MC Dropout (100 samples)
    b. Run the inference without MC Dropout, but with LLLA
    c. Combine the results from the two methods
    d. Repeat 10 times with different masking

Ensemble as uniformly-weighted mixture model

- Inspired by uncertainty estimation using deep ensembles (Lakshminarayanan et al., 2017)
- Assuming uniform weights
  - $\mu_* = M^{-1} \sum_m \mu_m$
  - $\sigma_*^2 = M^{-1} \sum_m (\theta_m^2 + \mu_m^2) - \mu_*^2$

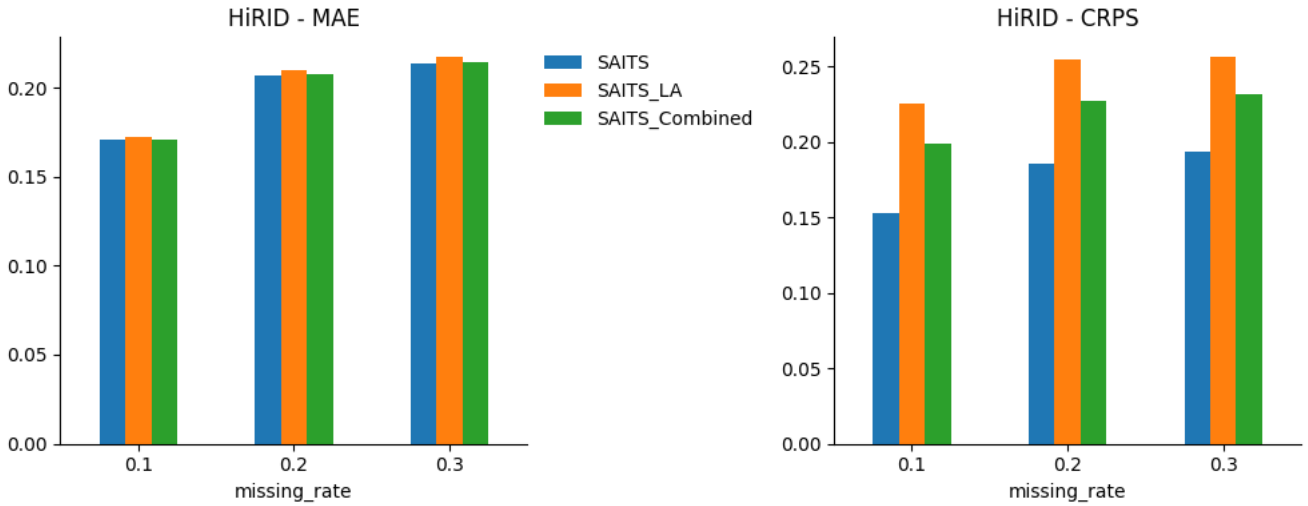
PhysioNet 2012

- Lowest MAE when combining MC Dropout & LLLA
- Lowest CRPS for combined at 10% and 20%, slightly higher than LLLA in 30%



HiRID

- Improved CRPS by combining compared to LLLA
- A better trade off than relying solely on LLLA

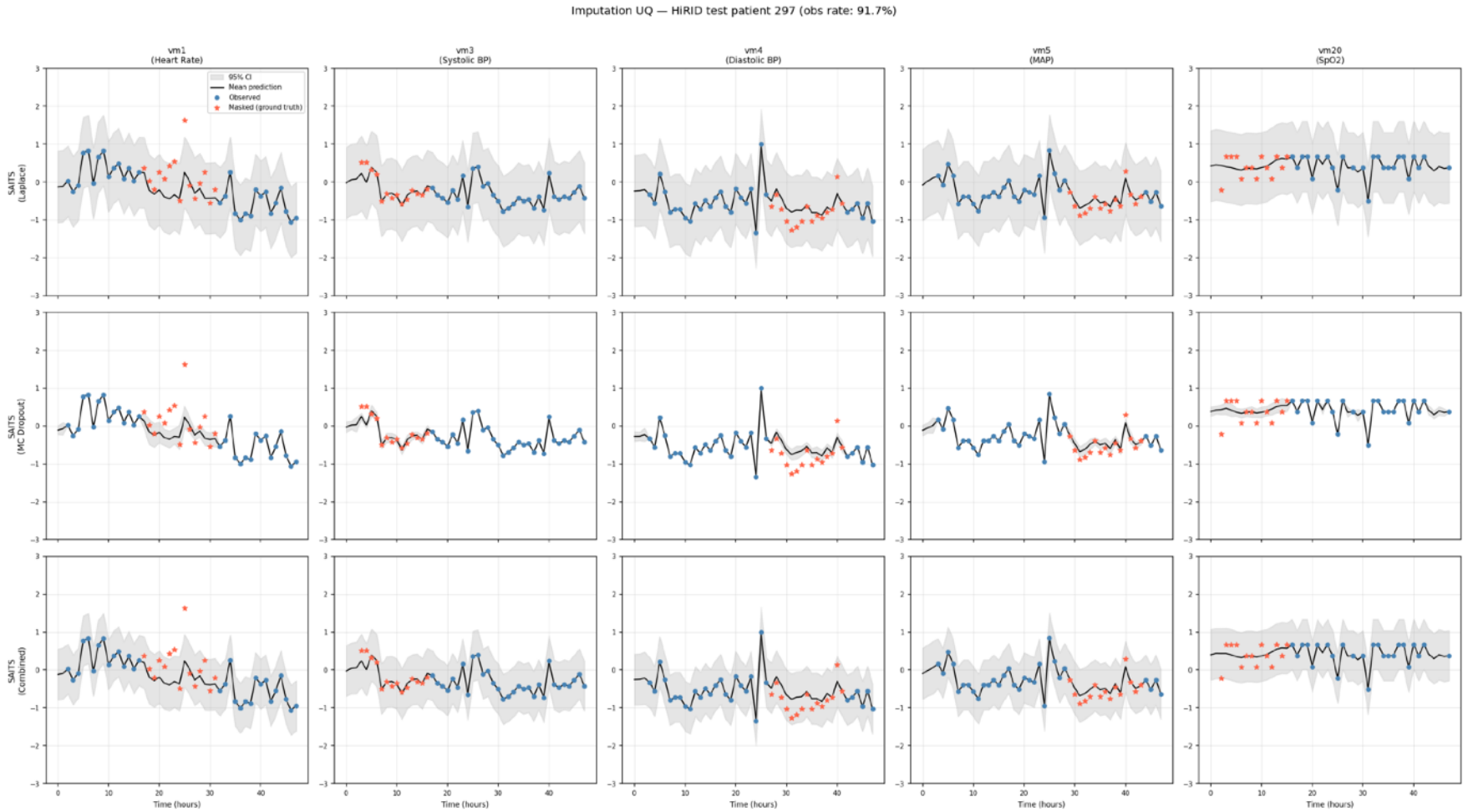


Visual analysis of imputation using SAITS with UQ on HiRID

LLLA

MCD

Combined



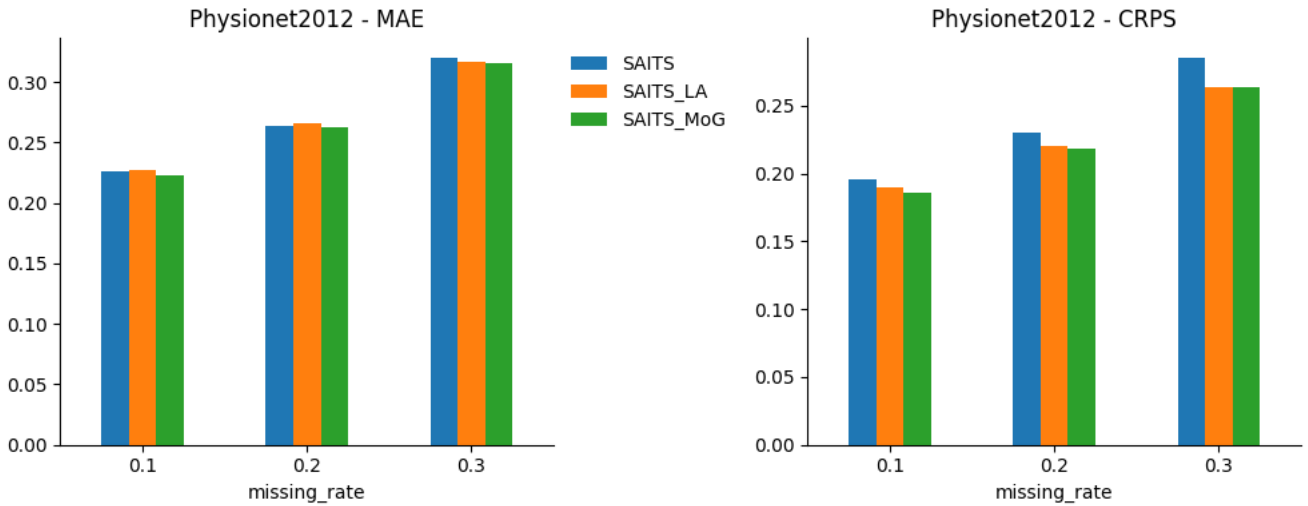
While MCD is clearly overconfident, LLLA is underconfident, and the combination addresses these issues

Finding the best way to combine LLLA and MC Dropout

1. Train the model once
2. Run the inference with MC Dropout (100 samples)
3. Run the inference without MC Dropout, but with LLLA
4. Use a calibration set to grid search the best α, using CRPS as the metric to optimise
   - $\mu_* = \alpha \cdot \mu_{MCD} + (1 - \alpha) \cdot \mu_{LA}$
   - $\sigma_*^2 = \alpha \cdot \sigma_{MCD}^2 + (1 - \alpha) \cdot \sigma_{LA}^2 + \alpha \cdot (1 - \alpha) \cdot (\mu_{MCD} - \mu_{LA})^2$

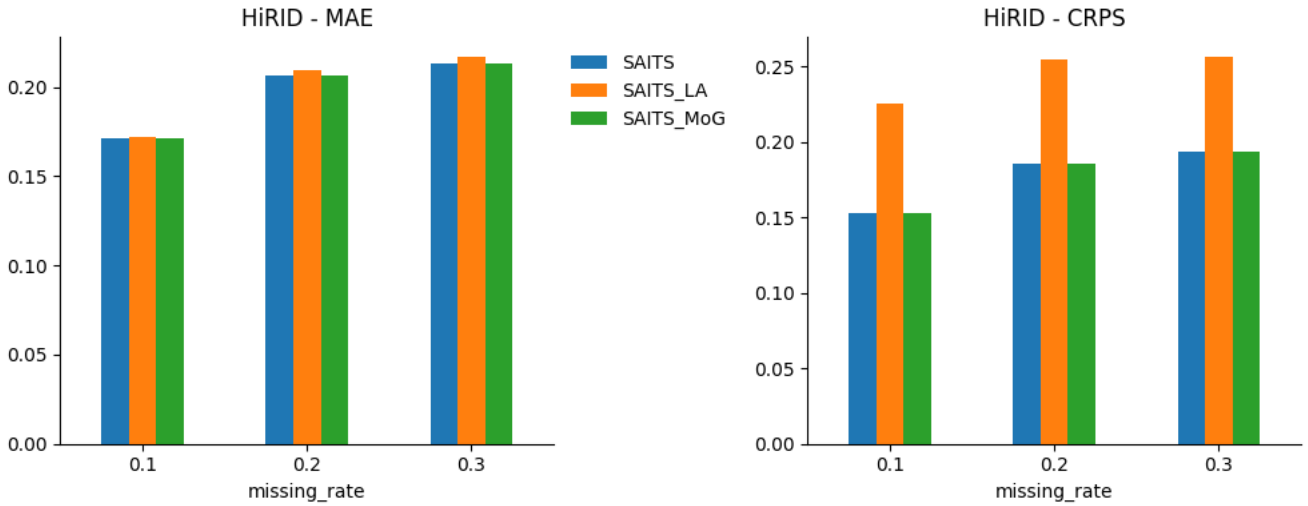
PhysioNet 2012

- Lowest MAE when combining MC Dropout & LLLA
- Lowest CRPS for all missingness rate



HiRID

- Since MC Dropout has betters CRPS, the calibration step will give the full weight to MC Dropout



Conclusions

- A post-hoc UQ method on a low-error deterministic model could match dedicated probabilistic models and with much cheaper inference
- LLLA and MC Dropout each perform better depending on dataset context
- Combining LLLA and MC Dropout yields the best overall results, capturing complementary uncertainty

Future work

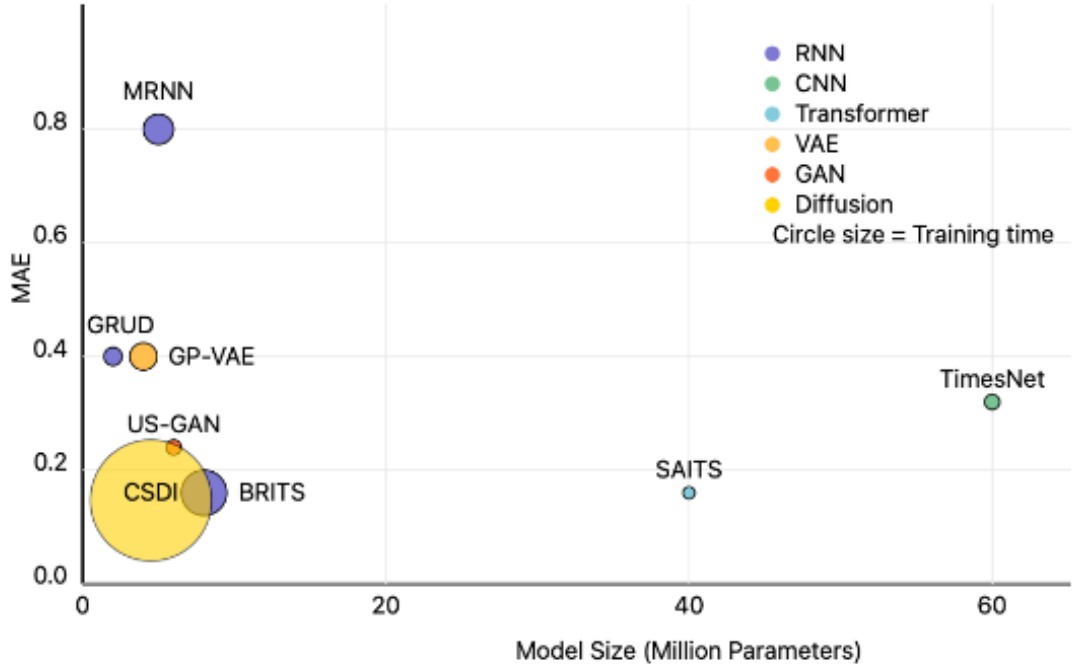
- Explore learned mixture weighting, e.g. via Bayesian optimisation
- Evaluate on datasets outside of the critical care domain
- Broaden evaluation to additional base architectures beyond SAITS and BRITS

Thank you

Reach out to discuss more!

- aliakbars.id/rsc26
- ali.septiandri.21@ucl.ac.uk

Error, model size, and training time



While CSDI can be very competitive, the training time is much slower (Qian et al., 2025)

Experiment with a synthetic dataset

Data

- $y = \sin(x) + \varepsilon, \varepsilon \sim \mathcal{N}(0, 0.1^2)$
- Generate 100 data points
- Mask 30% of data points

Model

- Two hidden layers, 50 hidden units each + ReLU activation
- Dropout units in each layer, $p = 0.1$
- N epochs = 100
- Learning rate = 0.01

Train the model once

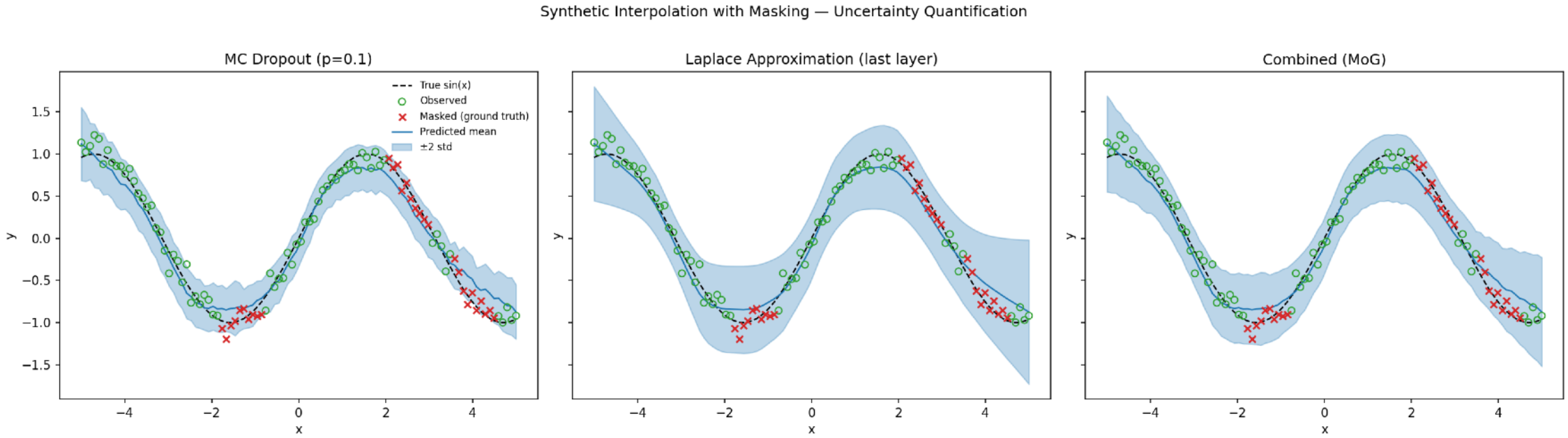
For inference

- Run the inference with MC Dropout (100 samples)
- Run the inference without MC Dropout, but with LLLA
- Combine the results from the two methods
- Repeat 10 times with different masking

Synthetic dataset result

The combination produces a slightly higher MAE

compared to only MC Dropout, but lower CRPS and NLPD overall

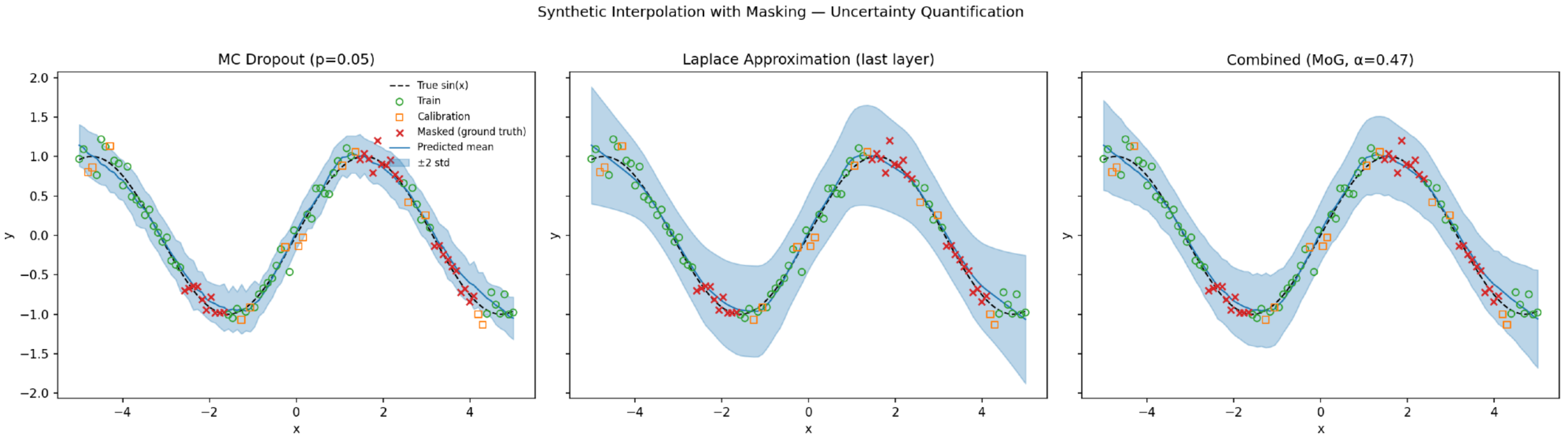


Method	MAE	CRPS	NLPD
MC Dropout	0.1676±0.0454	0.1202±0.0359	-0.0169±0.4694
Laplace	0.1698±0.0536	0.1216±0.0335	-0.0891±0.2078
Combined (MoG)	0.1683±0.0492	0.1179±0.0334	-0.1506±0.2496

Metrics from 10 different runs

Synthetic dataset result

Using grid search to find the best α



Method	MAE	CRPS	NLPD
MC Dropout	0.1424±0.0305	0.1027±0.0229	-0.0818±0.5010
Laplace	0.1427±0.0315	0.1095±0.0145	-0.1146±0.1051
Combined (MoG)	0.1418±0.0316	0.1032±0.0209	-0.0938±0.4440

Lowest overall MAE, second lowest CRPS and NLPD