Conformal Prediction for Tabular Prior-Data Fitted Networks with Missing Data

Florian D. van Leeuwen
Department of Methodology and Statistics
Utrecht University
f.d.vanleeuwen@uu.nl

Abstract

Tabular Prior-Data Fitted Networks (PFNs) achieve state-of-the-art performance for prediction tasks on small tabular datasets. Beyond point predictions, PFNs provide full posterior predictive distributions for uncertainty quantification. Among these models, TabPFN has emerged as particularly powerful due to its ability to handle missing data through internal deterministic imputation. However, we demonstrate that its uncertainty estimates are not valid conditional on the missing data pattern.

In this paper, we adapt the conformal prediction method CP-MDA-exact to TabPFN, providing a practical framework for obtaining mask-conditional valid uncertainty estimates. Our experiments on simulated data demonstrate that this approach successfully corrects the coverage across different missing patterns, even with small calibration sets.

1 Introduction

The landscape of tabular data analysis has been transformed by Prior-Data Fitted Networks (PFNs). PFNs leverage synthetic data generated from prior information to train models for supervised learning tasks, unlike traditional approaches that encode priors as parameter distributions. PFNs use in-context learning to rapidly adapt to new small datasets without requiring retraining [1]. Among these models, TabPFN has emerged as a particularly powerful tool, achieving state-of-the-art performance on tabular prediction tasks [2, 3, 4, 5] while providing not just point predictions but full posterior predictive distributions (PPDs) for uncertainty quantification.

However, real-world tabular data frequently contain missing values, presenting a fundamental challenge for uncertainty quantification. In traditional prediction pipelines, practitioners typically employ single imputation to simplify workflows, as research suggests this approach might maintain predictive accuracy, particularly when using expressive models or missing-value indicators [6, 7, 8]. Yet accuracy alone is insufficient: valid uncertainty estimates must properly reflect the information loss from missing data, ideally producing wider prediction intervals for observations with more missingness [9].

While TabPFN accepts missing data as input and performs internal imputation, a critical question remains: are its uncertainty estimates valid conditional on the missing data pattern? This is not merely a theoretical concern but a practical necessity for reliable decision-making. Recent advances in conformal prediction (CP) offer a promising solution, with methods specifically designed to provide valid coverage guarantees even in the presence of missing data [10].

In this paper, we make the following contributions: (1) We demonstrate that TabPFN's uncertainty estimates are not valid conditional on the missing pattern. (2) We adapt a CP with Missing Data Augmentation algorithm (CP-MDA-exact) to TabPFN, providing a practical framework for obtaining

AI for Tabular Data workshop at EurIPS 2025.

valid mask-conditional uncertainty estimates. (3) We investigate the performance of TabPFN + CP-MDA-exact under realistic constraints, particularly when calibration set sizes are limited.

2 Preliminaries

In this paper, the prediction setting is as follows. The data consists of i.i.d. pairs of features, a missing mask and a response: (X_i, M_i, Y_i), i = 1,..., n, with X being a matrix (n x p) with n the number of samples and p the number of predictors. Only the scenario with missing values in the features is considered, and thus the mask or missing pattern (M) is a matrix with the same dimensions as X (n x p). The mask value is 1 if a cell is missing and 0 if not. For an observation with $X_i = (5, \text{NA})$ the mask will be $M_i = (0, 1)$. Y is a vector with a continuous outcome (n x 1), resulting in a regression problem. The goal is to obtain a prediction model that can map the features to the outcome well for a new data point (X_{new}). Specifically, the focus of this paper is the predictive uncertainty, which is quantified with the prediction interval; in CP terms, this is considered the prediction set (C) [11]. In CP, the goal is to find a marginally valid prediction set:

$$P(Y_{new} \in C_\alpha(X_{new}, M_{new})) \geq 1 - \alpha. \tag{1}$$

With α being the preferred confidence level. In words, if α is set to 5%, the goal is to find a prediction interval that contains the true value at least (and preferably also at most) 95% of the time.

2.1 Missing data types

Missing data can have many causes; some are random (e.g., a weight scale runs out of battery), and some are not (e.g., an unconscious patient does not conduct a verbal test). To separate missing data, three categories can be used [12]: Missing completely at random (MCAR), Missing at random (MAR), Missing not at random (MNAR). In the case of MCAR, the cause of the missing data is unrelated to the data. MAR is more general and relates to missing values that have a cause that is observed (e.g., the conscious state of a patient is noted down along with the results of the verbal test). In the case that neither MCAR nor MAR hold, then we end up in the situation of MNAR. The probability of being missing then varies based on reasons unknown to us [13]. Approaches for prediction with missing data, and the effect of missing data on the uncertainty in the predicted values are further elaborated in Appendix A.1.

2.2 Prior fitted networks

PFNs have surged in popularity due to the good performance across different domains and the ability to handle diverse tabular data types [5, 2, 4, 3]. The general idea is that a large (transformer) model is trained to approximate the posterior predictive distribution (PPD) for a new data point given the training data [14]. The PPD is estimated directly, without having a posterior over the latent variables, which is used in classical Bayesian models [15, p. 357]. The training data for these PFN models consists of many synthetic datasets generated from Structural Causal Models [16]. After the model is trained, the model's parameters are fixed. The model now can to take a dataset as input, and in a single forward pass, learn a mapping from the features to the outcome (in-context learning). For more details on in-context learning for PFNs, see [14, 17]. Unlike a PPD obtained through draws from the latent parameters, a PFN needs to output parameters of a distribution describing the PPD. TabPFN utilises a Riemann distribution, where the data is discretely placed in buckets with the tail buckets being half-normal distributions for unbounded support, to approximate the PPD (details in [1]).

The first version of TabPFN [18] could not internally handle missing values, so an impute-then-regress regime was employed. The second version of TabPFN [2] overcame this limitation in two ways. Practically, the model allowed for missing data as input. The supplied data first undergoes a z-normalisation, after which the missing values are set to zero. This results in mean-imputation [1]. Furthermore, a missing data mask is appended to the data, indicating if a cell was missing or observed

[1]This might seem like a naive approach, as it does not consider the conditional distribution of the missing value. However, it has been empirically shown that the quality of the imputation matters less for point prediction accuracy when using very flexible prediction models [8]. The transformer architecture of TabPFN might thus be flexible enough to deal with "bad" imputations.

[2]. Theoretically, the datasets in the prior contained (MCAR) missing values, thereby creating some support.

2.2.1 Conformal prediction (with missing data)

Conformal prediction is a framework that can be utilized to create non-parametric predictive intervals (or sets) for any predictive algorithm, under the assumption of exchangeability between the training and test data [19]. A popular, computationally friendly approach is split-CP, where the training data (X_i, Y_i) is split into a training and a calibration set. The training set is used to create a prediction model that outputs a prediction interval, while the calibration set is used to create correction terms for the uncertainty bounds. After the correction terms are created, they can be applied to new data. A calibration set of size n = 1000 is sufficient for most purposes [11], which would exclude a serious number of tabular datasets [5].

Missing data complicated the story, as the more missing values an observation has, the higher the predictive uncertainty [9] (see Model 1 in Appendix A.1 for an example). So even if a prediction model can obtain the right marginal coverage (Equation 1), it should also be able to obtain the right coverage conditional on the missing mask. This would result in mask-conditional-validity (MCV) [10]:

$$P(Y \in C_\alpha(X, M)|M) \geq 1 - \alpha. \tag{2}$$

A natural way to use CP would be to create calibration sets based on the missingness masks. Unfortunately, the number of masks grows exponentially with the number of predictors, so chances are that the size of the calibration sets becomes rather small. One solution is to use Missing Data Augmentation (MDA), in which observations of the calibration set are masked for each missingness pattern to obtain more samples with the same mask [10]. There are two main variants, exact masking and nested masking. The general framework, named CP-MDA-Nested*, is proposed and explained in detail in [9]. In this paper, the focus will be on the exact variant, as the CP-MDA-Nested approach can result in prediction intervals that are too wide. The exact approach does come at the cost that fewer observations can be used in the calibration set. The nested versions might be preferred in a scenario with many different masks. The algorithm works under the assumption that the missing values are MCAR, although empirical results indicate that it might also work under the MAR assumption [9]. Moreover, the algorithm works under the assumption that Y is not a direct function of M.

3 Conformal predictions for TabPFN with missing data

Obtaining the correction terms using CP-MDA-exact with TabPFN can be described in 8 steps. The full Algorithm 1 is in Appendix A.2. Contrary to the CP-MDA method proposed in [10], the data does not need to be imputed by a user, as TabPFN does this internally. 1) The data is split into a training and calibration set. 2) TabPFN uses the training set to generate a mapping from X to Y. Note that TabPFN automatically generates a distribution for the outcomes, so there is no need to specify the quantities during training. 3) A set of unique masks is obtained, and steps 4-7 are performed for all elements in this set. 4) All observations in the calibration set with the same or a nested mask are selected as a subset, as explained in Figure 2 (Appendix A.2). 5) The fitted TabPFN model is used to obtain the $\alpha/2$ and $1 - \alpha/2$ bounds for the prediction interval. 6) The maximum error between the prediction interval bounds and the observed outcomes is calculated and stored in a set S. 7) The empirical quantile[2] of the error set S is taken, using the user-specified α level, to obtain the correction terms for each mask. 8) Finally, the unique masks accompanied by the correction terms are returned.

When a new value (x_{new}) is observed, the trained TabPFN model is used to obtain the prediction and uncertainty bounds. The correction term is then selected based on the mask (m_{new}) for the new value and added to both sides of the prediction interval:

$$\hat{C}_\alpha(x_{new}, m_{new}) = \left[\hat{\text{TabPFN}}_{\alpha/2}(x_{new}) - \mathcal{Q}_\alpha[m_{new}], \hat{\text{TabPFN}}_{1-\alpha/2}(x_{new}) + \mathcal{Q}_\alpha[m_{new}]\right] \tag{3}$$

[2]Using the empirical quantile does require that $|S| \geq (1 - \alpha)/\alpha$, for $1 - \tilde{\alpha} \leq 1$ to hold. This limits how small the calibration sets can be.

Note that this does not work if $m_{new} \notin \mathcal{M}_{\text{unique}}$. Furthermore, the correction terms are symmetric for the lower and upper bounds and could thus still be improved.

4 Experiments

The experiments, all repeated 1000 times, entail a linear model between X and Y, with 50% of the rows in X containing MCAR missing values. The missing pattern can be seen in the missing matrix 15 (Appendix A.3); each pattern with missing values had an equal probability of being assigned to a row. All predictors had the same effect on the outcome. Training, calibration, and test sets contained 500, 500 (unless otherwise specified), and 1000 observations, respectively. Calibration sets were used only for models with CP-MDA-exact. Details and code are in Appendix A.3 and on Github.

4.1 Results

TabPFN produces PPDs that are too narrow (Figure 3, Appendix A.4), leading to mild undercoverage; a phenomenon observed in variational inference [20]. Classical split conformal prediction can correct this to achieve nominal coverage levels. However, a significant issue arises when examining coverage conditional on missing data patterns. Figure 1**A** reveals systematic bias: complete observations exhibit overcoverage, while undercoverage increases as more values become missing. TabPFN + CP-MDA-exact resolves this conditional miscoverage by adjusting prediction interval widths (Figure 1**B**). TabPFN's prediction intervals do widen as missingness increases (Figure 1**C**), demonstrating that the model appropriately captures some uncertainty. Moreover, the overall proportion of missing data does not affect the marginal coverage of TabPFN (Figure 4, Appendix A.4).

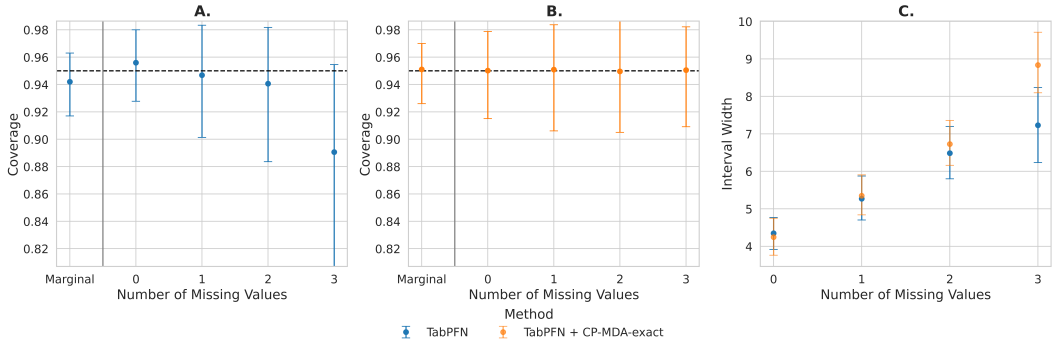

Figure 1: A prediction model with four predictors, based on a missing pattern that has increasingly more missing values (see Appendix A.3 for details). The mean marginal and conditional coverages, with the 95% empirical interval, of TabPFN before and after the CP-MDA-exact correction are shown in plot **A/B**. The dashed line indicates the correct coverage. Plot **C** visualizes the width of the prediction interval of TabPFN before and after CP.

Figure 5 (Appendix A.4) shows how coverage varies with calibration set size. Even with a calibration set of 100 (50 for the smallest mask), the coverages are close to 95%. There is only modest overcoverage, as is expected when the calibration set is small [11].

5 Conclusion

This study investigates the relation between missing values and uncertainty quantification of TabPFN. The results show that TabPFN, without any adjustment, will give prediction intervals that are too large for cases with no missing values or too small for cases with many missing values. Using the CP-MDA-exact algorithm can elevate this problem, even with small calibration sets, under certain assumptions of the missing type.

Future work should examine whether PFN priors can be adjusted to directly induce MCV, and how CP-MDA-exact could be extended to MAR mechanisms possibly using CP with conformal guarantees [21].

References

[1] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers Can Do Bayesian Inference, August 2024. arXiv:2112.10510 [cs].

[2] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. Publisher: Nature Publishing Group UK London.

[3] Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A Closer Look at TabPFN v2: Understanding Its Strengths and Extending Its Capabilities, June 2025. arXiv:2502.17361 [cs].

[4] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. June 2025.

[5] Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A Living Benchmark for Machine Learning on Tabular Data, June 2025. arXiv:2506.16791 [cs].

[6] Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What'sa good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.

[7] Alexandre Perez-Lebel, Gaël Varoquaux, Marine Le Morvan, Julie Josse, and Jean-Baptiste Poline. Benchmarking missing-values approaches for predictive models on health databases. *GigaScience*, 11:giac013, 2022. Publisher: Oxford University Press.

[8] Marine Le Morvan and Gaël Varoquaux. Imputation for prediction: beware of diminishing returns, February 2025. arXiv:2407.19804 [cs].

[9] Margaux Zaffran, Julie Josse, Yaniv Romano, and Aymeric Dieuleveut. Predictive Uncertainty Quantification with Missing Covariates, May 2024. arXiv:2405.15641 [stat].

[10] Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. Conformal prediction with missing values. In *International Conference on Machine Learning*, pages 40578–40604. PMLR, 2023.

[11] Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, December 2022. arXiv:2107.07511 [cs].

[12] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. Publisher: Oxford University Press.

[13] Stef van Buuren. *Flexible Imputation of Missing Data, Second Edition*. Chapman and Hall/CRC, New York, 2 edition, July 2018.

[14] Thomas Nagler. Statistical Foundations of Prior-Data Fitted Networks. In *Proceedings of the 40th International Conference on Machine Learning*, pages 25660–25676. PMLR, July 2023. ISSN: 2640-3498.

[15] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. Bayesian Data Analysis. *Bayesian Data Analysis*, 2013. ISBN: 9781439840955.

[16] Judea Pearl. *Causality*. Cambridge university press, 2009.

[17] Samuel Müller, Noah Hollmann, and Frank Hutter. Bayes' Power for Explaining In-Context Learning Generalizations, October 2024. arXiv:2410.01565 [cs].

[18] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second, September 2023. arXiv:2207.01848 [cs].

[19] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Nature, 2005. Google-Books-ID: PfChEAAAQBAJ.

[20] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017.

[21] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, September 2025.

[22] Adam Kapelner and Justin Bleich. Prediction with missing data via Bayesian Additive Regression Trees. *Canadian Journal of Statistics*, 43(2):224–239, June 2015.

[23] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45:1–67, 2011.

A Appendix

A.1 Predicting with missing data

In the prediction context, there are multiple ways to deal with missing values. An often used method is the impute-then-regress procedure [6], where the missing values are first imputed and then passed along to the prediction model. The quality of the imputed value will depend on the type of missing data and the imputation method. Imputation is necessary for some models, e.g., linear regression, as they cannot naturally handle the missing values. For tree-based methods, imputing is not necessary; the missing values can just be passed along one side of the split in a leaf, or more advanced splitting schemes can be used [22]. For the models that do not have this luxury, a researcher has to choose which imputation method to use. There are two main methods: deterministic and stochastic imputation methods. The first attempts to find the best possible value for the missing cell by minimizing some loss, often using the MSE. In stochastic imputation, the goal is to obtain draws for the conditional distribution and often multiple draws are used. There are then multiple data sets, each one a realization of the conditional distributions, on which the prediction model can then be fitted separately. To combine the results of normally distributed outcomes, Rubin's pooling rules can be used, see [23] for more details.

Using a toy example, we can show that missing data can lead to more uncertainty in the predicted values:

Model 1. (Gaussian Linear Model) Consider the following model:

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}_+^{p \times p}$
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$, where $\sigma \in \mathbb{R}_+$
- $Y = \boldsymbol{\beta}^T X + \epsilon$

Let $p = 2$, so $X = (X_1, X_2)^T$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$. Assume $|\Sigma_{12}| < \sigma_1 \sigma_2$ (correlation less than 1 in absolute value). The coefficients $\boldsymbol{\beta}$ and variance σ^2 are assumed known and fixed and nonzero.

Proposition 1. When predicting the response for a new observation with missing data, the prediction variance is strictly larger compared to the same fully observed observation, provided that perfect imputation is impossible (i.e., $\text{Var}(X_2|X_1) > 0$).

Proof.

Case 1: Fully observed data

For a new observation with known values $X_1 = x_1$ and $X_2 = x_2$, the conditional distribution of Y given the predictors is:

$$(Y|X_1) = x_1, X_2 = x_2 \sim \mathcal{N}(\beta_1 x_1 + \beta_2 x_2, \sigma^2) \tag{4}$$

Therefore, the prediction variance is:

$$\text{Var}(Y|X_1 = x_1, X_2 = x_2) = \sigma^2 \tag{5}$$

Case 2: Missing data for X_2

For a new observation where $X_1 = x_1$ is observed but X_2 is missing, we must predict Y using only the information in X_1. The conditional distribution of Y given only X_1 is obtained by integrating over the conditional distribution of X_2:

$$(Y|X_1) = x_1 \sim \mathcal{N}\left(\beta_1 x_1 + \beta_2 \mathbb{E}[X_2|X_1 = x_1], \text{Var}(Y|X_1 = x_1)\right) \tag{6}$$

To find the prediction variance, we use the law of total variance:

$$\text{Var}(Y|X_1 = x_1) = \mathbb{E}[\text{Var}(Y|X_1, X_2)|X_1 = x_1] + \text{Var}(\mathbb{E}[Y|X_1, X_2]|X_1 = x_1) \tag{7}$$

The first term is:

$$\mathbb{E}[\sigma^2|X_1 = x_1] = \sigma^2 \tag{8}$$

The second term is:

$$\text{Var}(\beta_1 x_1 + \beta_2 X_2 | X_1 = x_1) = \beta_2^2 \text{Var}(X_2 | X_1 = x_1) \tag{9}$$

Therefore:

$$\text{Var}(Y | X_1 = x_1) = \sigma^2 + \beta_2^2 \text{Var}(X_2 | X_1 = x_1) \tag{10}$$

Combining this result with Equation 5 and given that a perfect imputation is not possible ($\text{Var}(X_2 | X_1 = x_1) > 0$), we get:

$$\text{Var}(Y | X_1 = x_1) = \sigma^2 + \beta_2^2 \text{Var}(X_2 | X_1 = x_1) > \sigma^2 = \text{Var}(Y | X_1 = x_1, X_2 = x_2) \tag{11}$$

The additional term $\beta_2^2 \text{Var}(X_2 | X_1 = x_1)$ represents the extra uncertainty introduced by not knowing the true value of X_2. We thus see that missing values can lead to more uncertainty in the predicted value.

A.2 TabPFN + CP-MDA-Exact algorithm

In Figure 2, the masking procedure is shown. For each mask, the subset of data points for which a nested mask is selected.

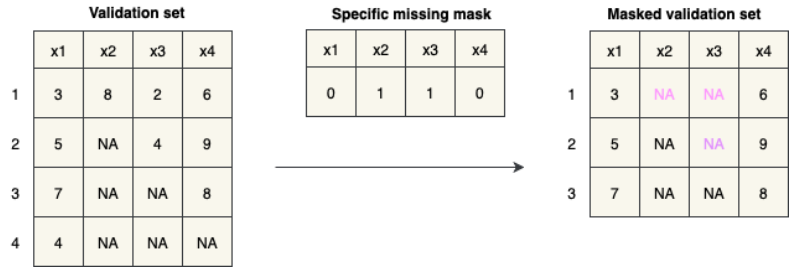

Figure 2: A subset of the calibration set is used during the CP based on the missing mask. In the example, the missing mask has two missing values for x2/x3. The rows with missing values in x1/x4 are left out (row 4), and for the other rows, the values for x2/x3 are set to NA if they were observed.

The TabPFN + CP-MDA-Exact algorithm is explained in 1. The output is the correction terms paired with the unique masks in the calibration set. These can then be applied to new data points if the new mask has been seen in the calibration set. The algorithm is implemented at Github.

Algorithm 1 TabPFN + CP-MDA-Exact

Input: TabPFN algorithm, significance level α, training set $\{(x_i, m_i, y_i)\}_{i=1}^{n}$
Output: Correction terms $\{\hat{Q}^{m}_{1-\tilde{\alpha}} : m \in \mathcal{M}_{\text{unique}}\}$
Randomly split $\{1, \ldots, n\}$ into two disjoint sets: Tr and Cal
Fit TabPFN:
    $\hat{\text{TabPFN}}(\cdot) \leftarrow$ TabPFN$(\{(x_i, y_i)\}_{i \in \mathsf{Tr}}, \alpha/2)$
Obtain the set of unique masks: $\mathcal{M}_{\text{unique}} = \{m_i \mid i \in \mathsf{Cal}\}$
Initialize correction terms dictionary: $\mathcal{Q} = \{\}$
for each unique mask $m \in \mathcal{M}_{\text{unique}}$ **do**
    Generate augmented calibration set: $\mathsf{Cal}(m) = \{i \in \mathsf{Cal} \text{ such that } m_i \subset m\}$
    for $i \in \mathsf{Cal}(m)$ **do**
        $m_i \leftarrow m$
    end for
    for $i \in \mathsf{Cal}(m)$ **do**
        Compute score: $s_i = \max\left(\hat{\text{TabPFN}}_{\alpha/2}(x_i) - y_i,\ y_i - \hat{\text{TabPFN}}_{1-\alpha/2}(x_i)\right)$
    end for
    Set $S = \{s_i \mid i \in \mathsf{Cal}(m)\}$
    Compute empirical quantile:
      $1 - \tilde{\alpha} = (1 - \alpha) \cdot (1 + 1/|S|)$
      $\hat{Q}^{m}_{1-\tilde{\alpha}}(S)$ is the $(1 - \tilde{\alpha})$-th empirical quantile of S
    Store correction term: $\mathcal{Q}_{\alpha}[m] = \hat{Q}^{m}_{1-\tilde{\alpha}}(S)$
end for
Return $\mathcal{Q}$

A.3 Experiments setup

The TabPFN model was trained using 500 samples. The size of the calibration set changed depending on the experiment, but it was 500 if not otherwise specified. In all scenarios, the model was evaluated using a test set of 1000 observations. Each simulation scenario was run 1000 times.

A.3.1 Data generating mechanism

In our experiments, we used a multivariate Gaussian distribution to generate the features:

$$X \sim \mathcal{N}(\mu, \Sigma). \tag{12}$$

With:

$$\mu = \mathbf{0}_p, \quad \Sigma = \rho + (1 - \rho) * I_p \tag{13}$$

Where ρ was set to 0.5. A linear function is used to model X on the outcome Y, where the effect sizes for the parameters were equal.

$$Y = X * B + N(0, \sigma^2) \tag{14}$$

Where B is a vector of coefficients that are always set to 1, σ^2 is also set to 1.

A.3.2 Missing data generating mechanism

The missing values were generated under the MCAR mechanism using a unique missingness pattern matrix $\mathcal{M}$. The missing pattern matrix indicates the unique missing masks in the data. For all experiments in Figure 1, the following unique missing matrix $\mathcal{M}$ is used:

$$\mathcal{M}_{\text{unique}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \tag{15}$$

For the experiments in Figures 5, a simpler example is used with only two predictors. Since the calibration set size depends on the number of nested masks, this makes it easier to investigate the effect of calibration sample size on the coverage. The unique missing matrix is given by:

$$\mathcal{M}_{\text{unique}} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \tag{16}$$

In all cases with missing data, the proportion of missing values is set to 50%, indicating that the first row of the matrix $\mathcal{M}_{\text{unique}}$ is used for half the data points and the other rows are equally split under the rest of the observations.

A.3.3 Outcomes

The outcome of the study was the coverage rate, where the significance level was set to $\alpha = 0.05$, resulting in a 95% prediction interval.

The coverage is empirically calculated based on the test set:

$$\text{coverage} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\text{TabPFN}}_{\alpha/2}(x_i) \leq y_i \leq \hat{\text{TabPFN}}_{1-\alpha/2}(x_i)) \tag{17}$$

The simulations were run on an A100 GPU through Google Colab. All analyses performed during this project resulted in a cost of 60 computing units.

A.4 Additional Results

The mean coverage at different thresholds of alpha is shown in Figure 3. The dashed line indicates perfect coverage, and this is closely aligned with the Tab-PFN combined with classical CP.

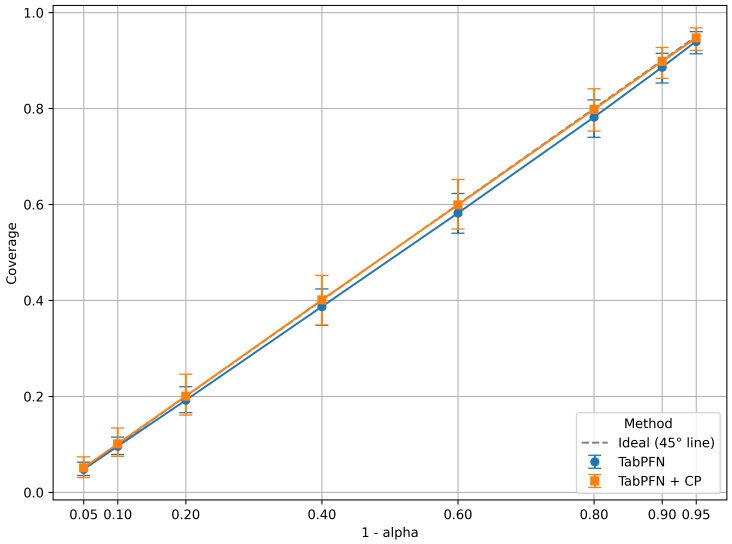

Figure 3: The mean coverages are shown with the 95% empirical interval over the simulations. The striped line indicates good coverage over all interval widths. There seems to be a small undercoverage of TabPFN, which can be corrected with CP.

The mean coverage of TabPFN is shown for different proportions of missing values in Figure 4. Even with 50% of the rows containing missing values, the marginal coverage is not very bad.

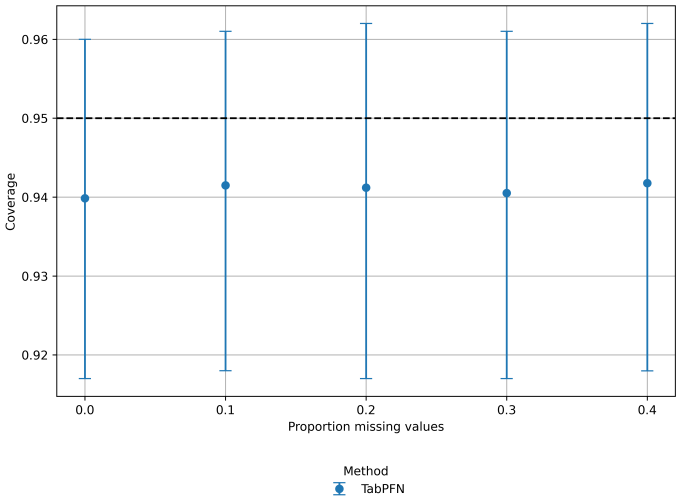

Figure 4: The mean coverages are shown with the 95% empirical interval over the simulations. The proportion missing value versus the marginal coverage for TabPFN.

The relation between the calibration set and the coverage is shown in Figure 5. Since the sample size is different based on the number of nested masks, a simple example is used with only two predictions and two masks. This means that if a row has no missing, about 50% of the datapoints are in the calibration set, and if there are missing, all datapoints can be used. It seems that even for a small calibration set of 100, the coverage is almost correct. The variance around the coverage is still relatively large with small sample sizes.

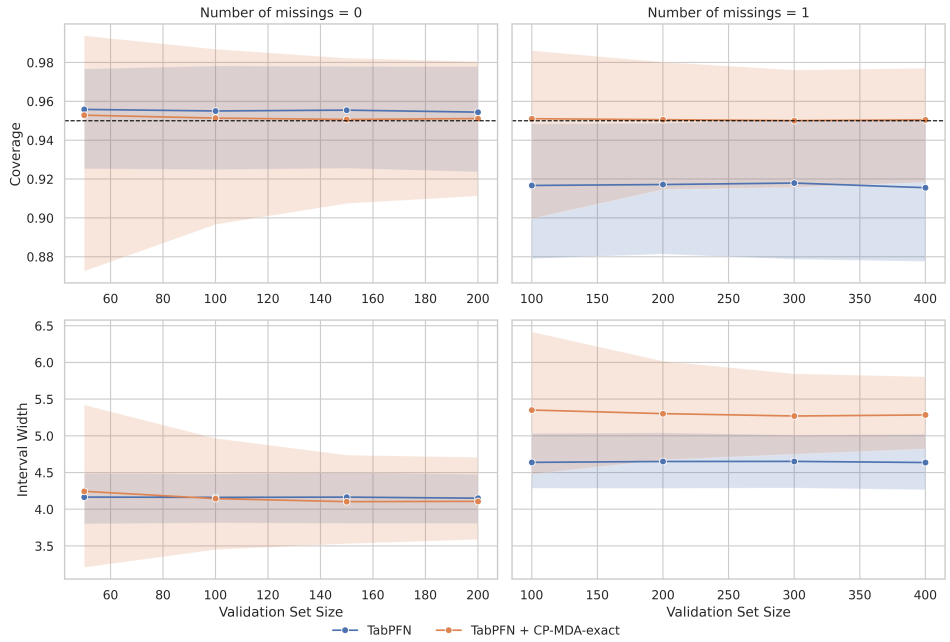

Figure 5: The mean converge and prediction interval width are shown with the 95% empirical interval over the simulations. The model only has two features, where one is always observed and the other is missing half of the time (see unique missing matrix 16). The coverage is almost conditionally correct, even with a calibration set of just 100 samples.