Appendix

A Gaussian process

Gaussian process (GP) is a typical choice for the surrogate model because of its model capacity for complicated black-box functions and uncertainty quantification. Consider, for the time being, a simplified scenario in which we have noise-contaminated observations $\{y_i = g(\mathbf{x}_i) + \epsilon_i\}_{i=1}^N$. In a GP model, a prior distribution is placed over $f(\mathbf{x})$, indexed by $\mathbf{x}$:

$$\eta(\mathbf{x})|\boldsymbol{\theta} \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'|\boldsymbol{\theta})\right), \tag{A.1}$$

with mean and covariance functions:

$$\begin{array}{ll} m_0(\mathbf{x}) & = \mathbb{E}[f(\mathbf{x})], \\ k(\mathbf{x}, \mathbf{x}'|\boldsymbol{\theta}) & = \mathbb{E}[(f(\mathbf{x}) - m_0(\mathbf{x}))(f(\mathbf{x}') - m_0(\mathbf{x}'))], \end{array} \tag{A.2}$$

where $\mathbb{E}[\cdot]$ is the expectation and $\boldsymbol{\theta}$ are the hyperparameters that control the kernel function. By centering the data, the mean function may be assumed to be an equal constant, $m_0(\mathbf{x}) \equiv m_0$. Alternative options are feasible, such as a linear function of $\mathbf{x}$, but they are rarely used until previous knowledge of the shape of the function is provided. The covariance function can take several forms, with the automated relevance determinant (ARD) kernel being the most popular.

$$k(\mathbf{x}, \mathbf{x}'|\boldsymbol{\theta}) = \theta_0 \exp\left(-(\mathbf{x} - \mathbf{x}')^T \text{diag}(\theta_1^{-2}, \ldots, \theta_l^{-2})(\mathbf{x} - \mathbf{x}')\right). \tag{A.3}$$

From this point on, we eliminate the explicit notation of $k(x, x')$'s reliance on $\boldsymbol{\theta}$. In this instance, the hyperparameters $\theta_1, \ldots, \theta_l$ are referred to as length-scales. For constant parameter $\mathbf{x}$, $f(\mathbf{x})$ is its random variable. In contrast, a collection of values, $f(\mathbf{x}_i)$, $i = 1, \ldots, N$, is a partial realization of the GP. GP's realizations are functions of x that are deterministic. The primary characteristic of GPs is that the joint distribution of $eta(xi), i = 1, hdots, N$ is multivariate Gaussian.

Assuming the model deficiency $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is likewise Gaussian, we can derive the model likelihood using the prior (A.1) and available data.

$$\begin{aligned} \mathcal{L} \triangleq p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) &= \int (f(\mathbf{x}) + \varepsilon) df = \mathcal{N}(\mathbf{y}|m_0\mathbf{1}, \mathbf{K} + \sigma^2\mathbf{I}) \\ &= -\frac{1}{2}(\mathbf{y} - m_0\mathbf{1})^T (\mathbf{K} + \sigma^2\mathbf{I})^{-1} (\mathbf{y} - m_0\mathbf{1}) \\ &\quad - \frac{1}{2}\ln|\mathbf{K} + \sigma^2\mathbf{I}| - \frac{N}{2}\log(2\pi), \end{aligned} \tag{A.4}$$

where $\mathbf{K} = [K_{ij}]$ is the covariance matrix, in which $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \ldots, N$. The hyperparameters $\boldsymbol{\theta}$ is often derived from point estimations using the maximum likelihood (MLE) of Eq. (A.4) w.r.t. $\boldsymbol{\theta}$. The joint distribution of $\mathbf{y}$ and $f(\mathbf{x})$ is also a joint Gaussian distribution with mean value $m_0\mathbf{1}$ and covariance matrix.

$$\mathbf{K}' = \left[\begin{array}{c|c} \mathbf{K} + \sigma^2\mathbf{I} & \mathbf{k}(\mathbf{x}) \\ \hline \mathbf{k}^T(\mathbf{x}) & k(\mathbf{x}, \mathbf{x}) + \sigma^2 \end{array}\right], \tag{A.5}$$

where $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}_1, \mathbf{x}), \ldots, k(\mathbf{x}_N, \mathbf{x}))^T$. Conditioning on $\mathbf{y}$, the conditional predictive distribution at $\mathbf{x}$ is obtained.

$$\begin{aligned} \hat{f}(\mathbf{x})|\mathbf{y} &\sim \mathcal{N}\left(\mu(\mathbf{x}), v(\mathbf{x}, \mathbf{x}')\right), \\ \mu(\mathbf{x}) &= m_0\mathbf{1} + \mathbf{k}(\mathbf{x})^T \left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1} (\mathbf{y} - m_0\mathbf{1}), \\ v(\mathbf{x}) &= \sigma^2 + k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^T(\mathbf{x}) \left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1} \mathbf{k}(\mathbf{x}). \end{aligned} \tag{A.6}$$

The expected value $\mathbb{E}[f(\mathbf{x})]$ is given by $\mu(\mathbf{x})$ and the predictive variance by $v(\mathbf{x})$. From Eq. (A.5) to Eq. (A.6) is crucial since the prediction posterior of this wake is based on a comparable block covariance matrix.

B Proof of Theorem

Lemma 1. *[16] If $\mathbf{X}^h \subset \mathbf{X}^l$, the joint likelihood of AR can be decomposed into two independent likelihoods of the low- and high-fidelity.*

This lemma has been proven by [15]. However, the notation and derivation is not easy to follow. To layout the foundations of GAR, we prove it using a clearer way with friendly notations.

Proof. Following Eq. (2), the inversion of the covariance matrix is

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} (\mathbf{K}^l)^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \rho^2(\mathbf{K}^r)^{-1} \end{pmatrix} & -\begin{pmatrix} 0 \\ \rho(\mathbf{K}^r)^{-1} \end{pmatrix} \\ -\begin{pmatrix} 0 & \rho(\mathbf{K}^r)^{-1} \end{pmatrix} & (\mathbf{K}^r)^{-1} \end{pmatrix}.$$

We can write down the log-likelihood for all the low- and high-fidelity observations as,

$$\begin{aligned}
&\log p(\mathbf{Y}^l, \mathbf{Y}^h) \\
=&-\frac{N^h+N^l}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{Y}^l, \rho\mathbf{E}^T\mathbf{Y}^l + \mathbf{Y}^r)^T\boldsymbol{\Sigma}^{-1}\begin{pmatrix} \mathbf{Y}^l \\ \rho\mathbf{E}^T\mathbf{Y}^l + \mathbf{Y}^r \end{pmatrix} \\
=&-\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{N^h+N^l}{2}\log(2\pi) - \frac{1}{2}[(\mathbf{Y}^l)^T(\mathbf{K}^l)^{-1}\mathbf{Y}^l + (\mathbf{Y}^l)^T\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \rho^2(\mathbf{K}^r)^{-1} \end{pmatrix}\mathbf{Y}^l \\
&-\rho(\mathbf{Y}^l)^T\mathbf{E}\big(0,\ \rho(\mathbf{K}^r)^{-1}\big)\mathbf{Y}^l - \Big(0,\ (\mathbf{Y}^r)^T\rho(\mathbf{K}^r)^{-1}\Big)\mathbf{Y}^l - (\mathbf{Y}^l)^T\mathbf{E}\rho{\mathbf{K}_r}^{-1}(\rho\mathbf{E}^T\mathbf{Y}^l + \mathbf{Y}^r) \\
&+\rho\mathbf{Y}^l\mathbf{E}(\mathbf{K}^r)^{-1}(\rho\mathbf{E}^T\mathbf{Y}^l + \mathbf{Y}^r) + \mathbf{Y}^r(\mathbf{K}^r)^{-1}(\rho\mathbf{E}^T\mathbf{Y}^l + \mathbf{Y}^r)] \\
=&-\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{N^h+N^l}{2}\log(2\pi) - \frac{1}{2}[(\mathbf{Y}^l)^T(\mathbf{K}^l)^{-1}\mathbf{Y}^l - \Big(0,\ (\mathbf{Y}^r)^T\rho(\mathbf{K}^r)^{-1}\Big)\mathbf{Y}^l \\
&+\mathbf{Y}^r(\mathbf{K}^r)^{-1}\rho\mathbf{E}^T\mathbf{Y}^l + \mathbf{Y}^r(\mathbf{K}^r)^{-1}\mathbf{Y}^r] \\
=&-\frac{1}{2}\log|\mathbf{K}^l| - \frac{1}{2}\log|\mathbf{K}^r| - \frac{N^l+N^h}{2}log(2\pi) - \frac{1}{2}(\mathbf{Y}^l)^T(\mathbf{K}^l)^{-1}\mathbf{Y}^l - \frac{1}{2}\mathbf{Y}^r(\mathbf{K}^r)^{-1}\mathbf{Y}^r \\
=&\underbrace{-\frac{N^L}{2}log(2\pi) - \frac{1}{2}\log|\mathbf{K}^l| - \frac{1}{2}(\mathbf{Y}^l)^T(\mathbf{K}^l)^{-1}\mathbf{Y}^l}_{\mathcal{L}^l}\underbrace{-\frac{N^h}{2}log(2\pi) - \frac{1}{2}\log|\mathbf{K}^r| - \frac{1}{2}(\mathbf{Y}^r)^T(\mathbf{K}^r)^{-1}\mathbf{Y}^r}_{\mathcal{L}^r}
\end{aligned} \tag{A.7}$$

□

where $\mathbf{Y}^r = \mathbf{Y}^h - \rho\mathbf{E}^T\mathbf{Y}^l$, $\mathcal{L}^l$ is the log-likelihood of the low-fidelity data with the lower fidelity kernel, and $\mathcal{L}^r$ is the log-likelihood of the residual data with the residual kernel; $\mathcal{L}^l$ and $\mathcal{L}^r$ are independent and thus can be trained in parallel.

Based on the joint probability Eq. (2), we can similarly derive the predictive posterior distribution of the high-fidelity using the standard GP posterior derivation. Conditioning on $\mathbf{Y}^h$ and $\mathbf{Y}^l$, the predictive high-fidelity posterior for a new input $\mathbf{x}_*$ is also a Gaussian $\mathcal{N}(\mu_*^h, \sigma_*^h)$:

$$\begin{aligned}
\mu_*^h =& \Big(\rho\mathbf{k}^l(\mathbf{x}_*, \mathbf{X}^l),\ \ \rho^2\mathbf{k}^l(\mathbf{x}_*, \mathbf{X}^h) + \mathbf{k}^r(\mathbf{x}_*, \mathbf{X}^h)\Big)\mathbf{K}^{-1}\begin{pmatrix} \mathbf{Y}^l \\ \rho\mathbf{E}^T\mathbf{Y}^l + \mathbf{Y}^r \end{pmatrix} \\
=&\rho\mathbf{k}^l(\mathbf{x}_*, \mathbf{X}^l)\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l)^{-1}\mathbf{Y}^l \\
&+\rho^3\mathbf{k}^l(\mathbf{x}_*, \mathbf{X}^l)\mathbf{E}(\mathbf{K}^r)^{-1}\mathbf{E}^T\mathbf{Y}^l - \rho^3\mathbf{k}^l(\mathbf{x}_*, \mathbf{X}^h)(\mathbf{K}^r)^{-1}\mathbf{E}^T\mathbf{Y}^l \\
&-\rho\mathbf{k}^r(\mathbf{x}_*, \mathbf{X}^h)(\mathbf{K}^r)^{-1}\mathbf{E}^T\mathbf{Y}^l - \rho^2\mathbf{k}^l(\mathbf{x}_*, \mathbf{X}^l)\mathbf{E}(\mathbf{K}^r)^{-1}\left[\rho\mathbf{E}^T\mathbf{Y}^l + \mathbf{Y}^r\right] \\
&+\rho^2\mathbf{k}^l(\mathbf{x}_*, \mathbf{X}^h)(\mathbf{K}^r)^{-1}\left[\rho\mathbf{E}^T\mathbf{Y}^l + \mathbf{Y}^r\right] + \mathbf{k}^r(\mathbf{x}_*, \mathbf{X}^h)(\mathbf{K}^r)^{-1}\left[\rho\mathbf{E}^T\mathbf{Y}^l + \mathbf{Y}^r\right] \\
=&\left[\rho\mathbf{k}^l(\mathbf{x}_*, \mathbf{X}^l)(\mathbf{K}^l)^{-1}\right]\mathbf{Y}^l + \mathbf{k}^r(\mathbf{x}_*, \mathbf{X}^h)(\mathbf{K}^r)^{-1}\mathbf{Y}^r \\
&-\left[\rho\mathbf{k}^r(\mathbf{x}_*, \mathbf{X}^h)(\mathbf{K}^r)^{-1}\right]\mathbf{E}\mathbf{Y}^l + \left[\rho\mathbf{k}^r(\mathbf{x}_*, \mathbf{X}^h)(\mathbf{K}^r)^{-1}\right]\mathbf{E}\mathbf{Y}^l \\
=&\left[\rho\mathbf{k}^l(\mathbf{x}_*, \mathbf{X}^l)(\mathbf{K}^l)^{-1}\right]\mathbf{Y}^l + \mathbf{k}^r(\mathbf{x}_*, \mathbf{X}^h)(\mathbf{K}^r)^{-1}\mathbf{Y}^r
\end{aligned} \tag{A.8}$$

and

$$\begin{aligned}
\sigma_*^h =& \Big(\rho^2\mathbf{k}^l(\mathbf{x}_*, \mathbf{x}_*) + \mathbf{k}^r(\mathbf{x}_*, \mathbf{x}_*)\Big) - (\rho\mathbf{k}_*^l, \rho^2\mathbf{k}_*^l(\mathbf{X}^h) + \mathbf{k}_*^r)^T\mathbf{K}^{-1}(\rho\mathbf{k}_*^l, \rho^2\mathbf{k}_*^l(\mathbf{X}^h) + \mathbf{k}_*^r) \\
=&\Big(\rho^2\mathbf{k}^l(\mathbf{x}_*, x_*) + \mathbf{k}^r(\mathbf{x}_*, \mathbf{x}_*)\Big) - \Big(\rho(\mathbf{k}_*^l)^T(\mathbf{K}^l)^{-1}\rho\mathbf{k}_*^l\Big) + \Big(\mathbf{0}, \rho(\mathbf{k}_*^r)^T(\mathbf{K}^r)^{-1}\Big)\rho\mathbf{k}_*^l \\
&-(\mathbf{k}_*^r)^T(\mathbf{K}^r)^{-1}\rho^2\mathbf{k}_*^l(\mathbf{X}^h) - (\mathbf{k}_*^r)^T(\mathbf{K}^r)^{-1}\mathbf{k}_*^r \\
=&\rho^2\Big(\mathbf{k}^l(\mathbf{x}_*, \mathbf{x}_*) - (\mathbf{k}_*^l)^T(\mathbf{K}^l)^{-1}\mathbf{k}_*^l\Big) + \Big(\mathbf{k}^r(\mathbf{x}_*, \mathbf{x}_*) - (\mathbf{k}_*^r)^T(\mathbf{K}^r)^{-1}\mathbf{k}_*^r\Big)
\end{aligned} \tag{A.9}$$

where, $\mathbf{k}_*^l(\mathbf{X}^h) = \mathbf{k}^l(\mathbf{x}_*, \mathbf{X}^h)$ is the covariance vector between the new inputs $\mathbf{x}_*$ and $\mathbf{X}^h$. Notice that the predictive posterior is also decomposed into two independent parts that related to the low-fidelity GP and the residual GP, which is convenient for parallel computing and saving computational resources.

Lemma 2. *Given tensor GP priors for* $\mathbf{Y}^l(\mathbf{x}, \mathbf{x}')$ *and* $\mathbf{Y}^r(\mathbf{x}, \mathbf{x}')$ *and the Tucker transformation of Eq. (3), the joint probability for* $\mathbf{y} = [\text{vec}(\mathbf{Y}^l)^T, \text{vec}(\mathbf{Y}^h)^T]^T$ *is* $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$*, where* $\mathbf{\Sigma} =$

$$\begin{pmatrix} \mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l) \otimes \left(\bigotimes_{m=1}^M \mathbf{S}_m^l\right) & \mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^h) \otimes \left(\bigotimes_{m=1}^M \mathbf{S}_m^l \mathbf{W}_m^T\right) \\ \mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^l) \otimes \left(\bigotimes_{m=1}^M \mathbf{W}_m \mathbf{S}_m^l\right) & \mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^h) \otimes \left(\bigotimes_{m=1}^M \mathbf{W}_m \mathbf{S}_m^l \mathbf{W}_m\right) + \mathbf{K}^r(\mathbf{X}^h, \mathbf{X}^h) \otimes \left(\bigotimes_{m=1}^M \mathbf{S}_m^r\right) \end{pmatrix}$$

Proof. Since the $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{y}$, it can be expressed in block form as:

$$\mathbf{\Sigma} = \begin{pmatrix} \text{cov}(\text{vec}(\mathbf{Y}^l), \text{vec}(\mathbf{Y}^l)) & \text{cov}(\text{vec}(\mathbf{Y}^l), \text{vec}(\mathbf{Y}^h)) \\ \text{cov}(\text{vec}(\mathbf{Y}^h), \text{vec}(\mathbf{Y}^l)) & \text{cov}(\text{vec}(\mathbf{Y}^h), \text{vec}(\mathbf{Y}^h)) \end{pmatrix},$$

where $\text{cov}(\text{vec}(\mathbf{Y}^l), \text{vec}(\mathbf{Y}^h)) = \text{cov}(\text{vec}(\mathbf{Y}^l), \text{vec}(\mathbf{Y}^h))^T$ is the cross covariance between $\mathbf{Y}^l$ and $\mathbf{Y}^h$. Assuming $\mathbf{Y}^h \in \mathbb{R}^{N^h \times d_1^h \times \ldots \times d_M^h}$ and $\mathbf{Y}^l \in \mathbb{R}^{N^l \times d_1^l \times \ldots \times d_M^l}$, together with the property of the Tucker operator in Eq. (3), the high-fidelity data and low-fidelity data have the following transformation,

$$\begin{aligned} \mathbf{Y}^h &= \mathbf{Y}^l \times_1 \mathbf{E} \times_2 \mathbf{W}_1 \times_3 \ldots \times_M \mathbf{W}_{M-1} \times_{M+1} \mathbf{W}_M \\ \text{vec}(\mathbf{Y}^h) &= \left[\mathbf{E} \otimes \left(\bigotimes_{m=1}^M \mathbf{W}_m\right)\right] \text{vec}(\mathbf{Y}^l) + \text{vec}(\mathbf{Y}^r), \end{aligned} \tag{A.10}$$

where $\forall i = 1, 2, ..., M, \mathbf{W}_i \in \mathbb{R}^{d_i^h \times d_i^l}$, and $\mathbf{E}^T = (\mathbf{0}, \mathbf{I}_{N^h}) \in \mathbb{R}^{N^h \times N^l}$ is the selection matrix such that $\mathbf{X}^h = \mathbf{E}^T \mathbf{X}^l$. By definition our GP prior, the low-fidelity data has the joint probability:

$$\text{vec}(\mathbf{Y}^l) \sim \mathcal{N}\left(0, \mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l) \otimes \left(\bigotimes_{m=1}^M \mathbf{S}_m^l\right)\right)$$

Thus the covariance matrix of low-fidelity data is $\text{cov}(\text{vec}(\mathbf{Y}^l), \text{vec}(\mathbf{Y}^l)) = \mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l) \otimes \left(\bigotimes_{m=1}^M \mathbf{S}_m^l\right)$. After that, we can derive the other part of the $\mathbf{\Sigma}$. Firstly, assuming the residual information $\text{vec}(\mathbf{Y}^r)$ is independent from $\text{vec}(\mathbf{Y}^l)$, the covariance between $\text{vec}(\mathbf{Y}^h)$ and $\text{vec}(\mathbf{Y}^l)$ is

$$\begin{aligned} \text{cov}(\text{vec}(\mathbf{Y}^l), \text{vec}(\mathbf{Y}^h)) =& \text{cov}\left(\text{vec}(\mathbf{Y}^l), \left[\mathbf{E} \otimes \left(\bigotimes_{m=1}^M \mathbf{W}_m\right)\right] \text{vec}(\mathbf{Y}^l) + \text{vec}(\mathbf{Y}^r)\right) \\ =& \text{cov}\left(\text{vec}(\mathbf{Y}^l), \text{vec}(\mathbf{Y}^r)\right) + \text{cov}\left(\text{vec}(\mathbf{Y}^l), \left[\mathbf{E} \otimes \left(\bigotimes_{m=1}^M \mathbf{W}_m\right)\right] \text{vec}(\mathbf{Y}^l)\right) \\ =& \text{cov}(\text{vec}(\mathbf{Y}^l), \text{vec}(\mathbf{Y}^l)) \left[\mathbf{E} \otimes \left(\bigotimes_{m=1}^M \mathbf{W}_m\right)\right]^T \\ =& \left[\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l) \otimes \left(\bigotimes_{m=1}^M \mathbf{S}_m^l\right)\right] \left[\mathbf{E}^T \otimes \left(\bigotimes_{m=1}^M \mathbf{W}_m^T\right)\right] \\ =& \mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^h) \otimes \left(\bigotimes_{m=1}^M \mathbf{S}_m^l \mathbf{W}_m^T\right). \end{aligned} \tag{A.11}$$

Since $\text{cov}(\text{vec}(\mathbf{Y}^l), \text{vec}(\mathbf{Y}^h))$ is the transpose of $\text{cov}(\text{vec}(\mathbf{Y}^h), \text{vec}(\mathbf{Y}^l))$, so the upper right part of $\mathbf{\Sigma}$ is

$$\text{cov}(\text{vec}(\mathbf{Y}^h), \text{vec}(\mathbf{Y}^l)) = \text{cov}(\text{vec}(\mathbf{Y}^l), \text{vec}(\mathbf{Y}^h))^T = \mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^l) \otimes \left(\bigotimes_{m=1}^M \mathbf{W}_m \mathbf{S}_m^l\right). \tag{A.12}$$

For the lower and right part of $\mathbf{\Sigma}$, the covariance between $\mathrm{cov}(\mathrm{vec}(\mathbf{Y}^h), \mathrm{vec}(\mathbf{Y}^h))$ is

$$
\begin{aligned}
&\mathrm{cov}(\mathrm{vec}(\mathbf{Y}^h), \mathrm{vec}(\mathbf{Y}^h)) \\
=&\mathrm{cov}\left(\left[\mathbf{E}\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\right)\right]\mathrm{vec}(\mathbf{Y}^l)+\mathrm{vec}(\mathbf{Y}^r), \left[\mathbf{E}\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\right)\right]\mathrm{vec}(\mathbf{Y}^l)+\mathrm{vec}(\mathbf{Y}^r)\right) \\
=&\mathrm{cov}\left(\left[\mathbf{E}\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\right)\right]\mathrm{vec}(\mathbf{Y}^l), \left[\mathbf{E}\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\right)\right]\mathrm{vec}(\mathbf{Y}^l)\right)+\mathrm{cov}\left(\mathrm{vec}(\mathbf{Y}^r), \mathrm{vec}(\mathbf{Y}^r)\right) \\
&+\mathrm{cov}\left(\left[\mathbf{E}\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\right)\right]\mathrm{vec}(\mathbf{Y}^l), \mathrm{vec}(\mathbf{Y}^r)\right)+\mathrm{cov}\left(\mathrm{vec}(\mathbf{Y}^r), \left[\mathbf{E}\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\right)\right]\mathrm{vec}(\mathbf{Y}^l)\right) \\
=&\left[\mathbf{E}\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\right)\right]\left(\mathrm{cov}(\mathrm{vec}(\mathbf{Y}^l), \mathrm{vec}(\mathbf{Y}^l))\right)\left[\mathbf{E}\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\right)\right]^T+\mathrm{cov}\left(\mathrm{vec}(\mathbf{Y}^r), \mathrm{vec}(\mathbf{Y}^r)\right) \\
=&\mathbf{K}^l(\mathbf{X}^h,\mathbf{X}^h)\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\mathbf{S}_m^l\mathbf{W}_m^T\right)+\mathbf{K}^r(\mathbf{X}^h,\mathbf{X}^h)\left(\bigotimes_{m=1}^{M}\mathbf{S}_m^r\right).
\end{aligned}
\tag{A.13}
$$

Assembling the several parts together, we have the joint covariance matrix $\mathbf{\Sigma}$:

$$
\begin{pmatrix}
\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)\otimes\left(\bigotimes_{m=1}^{M}\mathbf{S}_m^l\right) & \mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^h)\otimes\left(\bigotimes_{m=1}^{M}\mathbf{S}_m^l\mathbf{W}_m^T\right) \\
\mathbf{K}^l(\mathbf{X}^h,\mathbf{X}^l)\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\mathbf{S}_m^l\right) & \mathbf{K}^l(\mathbf{X}^h,\mathbf{X}^h)\otimes\left(\bigotimes_{m=1}^{M}\mathbf{W}_m\mathbf{S}_m^l\mathbf{W}_m\right)+\mathbf{K}^r(\mathbf{X}^h,\mathbf{X}^h)\otimes\left(\bigotimes_{m=1}^{M}\mathbf{S}_m^r\right)
\end{pmatrix}
\tag{A.14}
$$

□

Before we move on to the next proof, we introduce the matrix inversion property, which will become handy later.

Property 1. *For any invertible block matrixes* $\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{pmatrix}$, *where the sub-matrixes are also invertible, we have* $(\mathbf{B}^T,\mathbf{C})\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{pmatrix}^{-1}=(\mathbf{0},\mathbf{I})$ *and* $\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{B}\\ \mathbf{C}\end{pmatrix}=\begin{pmatrix}\mathbf{0}\\ \mathbf{I}\end{pmatrix}$.

Proof. The inversion of a block matrix (if it is invertible) following the Sherman-Morrison formula is

$$
\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{pmatrix}^{-1}=\begin{pmatrix}\mathbf{P^{-1}} & -\mathbf{P^{-1}BC^{-1}}\\ -\mathbf{C^{-1}B}^T\mathbf{P^{-1}} & \mathbf{C}^{-1}+\mathbf{C^{-1}B}^T\mathbf{P^{-1}BC^{-1}}\end{pmatrix},a
\tag{A.15}
$$

where $\mathbf{P}=\mathbf{A}-\mathbf{BC}^{-1}\mathbf{B}^T$, we can then derive the multiplication in Property 1 by the rule of block matrix multiplication:

$$
\begin{aligned}
&(\mathbf{B}^T,\mathbf{C})\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{pmatrix}^{-1} \\
=&(\mathbf{B}^T,\mathbf{C})\begin{pmatrix}\mathbf{P^{-1}} & -\mathbf{P^{-1}BC^{-1}}\\ -\mathbf{C^{-1}B}^T\mathbf{P^{-1}} & \mathbf{C}^{-1}+\mathbf{C^{-1}B}^T\mathbf{P^{-1}BC^{-1}}\end{pmatrix} \\
=&\left(\mathbf{B}^T\mathbf{P^{-1}}-\mathbf{C}(\mathbf{C^{-1}B^TP^{-1}}),-\mathbf{B^TP^{-1}BC^{-1}}+\mathbf{CC^{-1}}+\mathbf{C}(\mathbf{C^{-1}B^TP^{-1}BC^{-1}})\right) \\
=&\left(\mathbf{B}^T\mathbf{P^{-1}}-\mathbf{B^TP^{-1}},-\mathbf{B^TP^{-1}BC^{-1}}+\mathbf{I}+\mathbf{B^TP^{-1}BC^{-1}}\right) \\
=&(\mathbf{0},\mathbf{I}).
\end{aligned}
\tag{A.16}
$$

Similarly, the other part of the conclusion can also be derived

$$
\begin{aligned}
&\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{B}\\ \mathbf{C}\end{pmatrix} \\
=&\begin{pmatrix}\mathbf{P^{-1}} & -\mathbf{P^{-1}BC^{-1}}\\ -\mathbf{C^{-1}B}^T\mathbf{P^{-1}} & \mathbf{C}^{-1}+\mathbf{C^{-1}B}^T\mathbf{P^{-1}BC^{-1}}\end{pmatrix}\begin{pmatrix}\mathbf{B}\\ \mathbf{C}\end{pmatrix} \\
=&\begin{pmatrix}\mathbf{P^{-1}B}-\mathbf{P^{-1}BC^{-1}C}\\ -\mathbf{C^{-1}B}^T\mathbf{P^{-1}B}+(\mathbf{C^{-1}}+\mathbf{C^{-1}B^TP^{-1}BC^{-1}})\mathbf{C}\end{pmatrix} \\
=&\begin{pmatrix}\mathbf{P^{-1}B}-\mathbf{P^{-1}B}\\ -\mathbf{C^{-1}B}^T\mathbf{P^{-1}B}+\mathbf{I}+\mathbf{C^{-1}B^TP^{-1}B}\end{pmatrix} \\
=&\begin{pmatrix}\mathbf{0}\\ \mathbf{I}\end{pmatrix},
\end{aligned}
\tag{A.17}
$$

□

which seems quite obvious and intuitive if we assume that the matrix is symmetric.

Lemma 3. *Generalization of Lemma 1 in GAR. If* $\mathbf{X}^h \subset \mathbf{X}^l$*, the joint likelihood* $\mathcal{L}$ *for* $\mathbf{y} = [\mathrm{vec}(\mathbf{Y}^l)^T, \mathrm{vec}(\mathbf{Y}^h)^T]^T$ *admits two independent separable likelihoods* $\mathcal{L} = \mathcal{L}^l + \mathcal{L}^r$*, where*

$$\mathcal{L}^l = -\frac{1}{2}\mathrm{vec}\left(\mathbf{Y}^l\right)^T (\mathbf{K}^l \otimes \mathbf{S}^l)^{-1}\mathrm{vec}\left(\mathbf{Y}^l\right) - \frac{1}{2}\log|\mathbf{K}^l \otimes \mathbf{S}^l| - \frac{N^l D^l}{2}\log(2\pi),$$

$$\mathcal{L}^r = -\frac{1}{2}\mathrm{vec}\left(\mathbf{Y}^h - \mathbf{Y}^l \times \hat{\mathbf{W}}\right)^T (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1}\mathrm{vec}\left(\mathbf{Y}^h - \mathbf{Y}^l \times \hat{\mathbf{W}}\right) - \frac{1}{2}\log|\mathbf{K}^r \otimes \mathbf{S}^r| - \frac{N^h D^h}{2}\log(2\pi),$$

where $\hat{\mathbf{W}} = [\mathbf{E}, \mathbf{W}]$ *is the original weight tensor concatenated with an selection matrix* $\mathbf{X}^h = \mathbf{E}^T\mathbf{X}^l$*.*

Proof. Let the kernel matrix be partitioned into four blocks. We again make use of the matrix inversion of Sherman-Morrison formula in Property 1, with a slight modification as follows:

$$\begin{pmatrix} \mathbf{T} & \mathbf{U} \\ \mathbf{V} & \mathbf{M} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{T}^{-1} + \mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1} & -\mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1} \\ -\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1} & \mathbf{Q}^{-1} \end{pmatrix}$$

where

$$\mathbf{Q} = \mathbf{M} - \mathbf{V}\mathbf{T}^{-1}\mathbf{U}.$$

We begin this proof with the matrix $\mathbf{V}\mathbf{T}^{-1}\mathbf{U}$ within Property 1, which gives us:

$$\mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^l)\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l)^{-1}\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^h) = (\mathbf{0}, \mathbf{I})\,\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^h) = \mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^h).$$

Therefore, the last part $\mathbf{V}\mathbf{T}^{-1}\mathbf{U}$ of matrix $\mathbf{Q}$ is

$$\begin{aligned}
&\mathbf{V}\mathbf{T}^{-1}\mathbf{U} \\
=&\left[\mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^l) \otimes \mathbf{W}\mathbf{S}^l\right]\left[\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l) \otimes \mathbf{S}^l\right]^{-1}\left[\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^h) \otimes \mathbf{S}^l\mathbf{W}^T\right] \\
=&\left[\mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^l) \otimes \mathbf{W}\mathbf{S}^l\right]\left[\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l)^{-1} \otimes (\mathbf{S}^l)^{-1}\right]\left[\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^h) \otimes \mathbf{S}^l\mathbf{W}^T\right] \\
=&\left[\mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^l)\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l)^{-1}\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^h)\right] \otimes \left[(\bigotimes_{m=1}^{M}\mathbf{W}_m\mathbf{S}_m^l)(\bigotimes_{m=1}^{M}(\mathbf{S}_m^l)^{-1})(\bigotimes_{m=1}^{M}\mathbf{S}_m^l\mathbf{W}_m^T)\right] \\
=&\left[\mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^l)\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l)^{-1}\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^h)\right] \otimes \left[\bigotimes_{m=1}^{M}(\mathbf{W}_m\mathbf{S}_m^l(\mathbf{S}_m^l)^{-1}\mathbf{S}_m^l\mathbf{W}_m^T)\right] \\
=&\mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^h) \otimes \left[\bigotimes_{m=1}^{M}(\mathbf{W}_m\mathbf{S}_m^l\mathbf{W}_m^T)\right].
\end{aligned} \tag{A.18}$$

Substituting Eq. (A.18) back into the matrix inversion, we can derive matrix $\mathbf{Q}^{-1}$, $-\mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1}$, $\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1}$, and $\mathbf{T}^{-1} + \mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1}$ as

$$\begin{aligned}
&\mathbf{Q}^{-1} \\
=&(\mathbf{M} - \mathbf{V}\mathbf{T}^{-1}\mathbf{U})^{-1} \\
=&\left(\mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^h) \otimes \mathbf{W}\mathbf{S}^l\mathbf{W}^T + \mathbf{K}^r(\mathbf{X}^h, \mathbf{X}^h) \otimes \mathbf{S}^r - \mathbf{K}^l(\mathbf{X}^h, \mathbf{X}^h) \otimes \mathbf{W}\mathbf{S}^l\mathbf{W}^T\right)^{-1} \\
=&\left(\mathbf{K}^r(\mathbf{X}^h, \mathbf{X}^h) \otimes \mathbf{S}^r\right)^{-1},
\end{aligned} \tag{A.19}$$

$$\begin{aligned}
&-\mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1} \\
=&-\left[\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l)^{-1} \otimes (\mathbf{S}^l)^{-1}\right]\left[\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^h) \otimes \mathbf{S}^l\mathbf{W}^T\right]\left[\mathbf{K}^r(\mathbf{X}^h, \mathbf{X}^h)^{-1} \otimes (\mathbf{S}^r)^{-1}\right] \\
=&-\left[\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^l)^{-1}\mathbf{K}^l(\mathbf{X}^l, \mathbf{X}^h)\mathbf{K}^r(\mathbf{X}^h, \mathbf{X}^h)^{-1}\right] \otimes \left[(\mathbf{S}^l)^{-1}\mathbf{S}^l\mathbf{W}^T(\mathbf{S}^r)^{-1}\right] \\
=&-\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{I} \end{pmatrix}\mathbf{K}^r(\mathbf{X}^h, \mathbf{X}^h)^{-1}\right]\left[\mathbf{W}^T(\mathbf{S}^r)^{-1}\right] \\
=&-\begin{pmatrix} \mathbf{0} \\ \mathbf{K}^r(\mathbf{X}^h, \mathbf{X}^h)^{-1} \otimes \mathbf{W}^T(\mathbf{S}^r)^{-1} \end{pmatrix},
\end{aligned} \tag{A.20}$$

$$
\begin{aligned}
&\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1} \\
=&-\left[\mathbf{K}^r(\mathbf{X}^h,\mathbf{X}^h)^{-1}(\mathbf{S}^r)^{-1}\right]\left[\mathbf{K}^l(\mathbf{X}^h,\mathbf{X}^l)\otimes\mathbf{W}\mathbf{S}^l\right]\left[\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)^{-1}\otimes(\mathbf{S}^l)^{-1}\right] \\
=&-\left[\mathbf{K}^r(\mathbf{X}^h,\mathbf{X}^h)^{-1}\mathbf{K}^l(\mathbf{X}^h,\mathbf{X}^l)\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)^{-1}\right]\otimes\left[(\mathbf{S}^r)^{-1}\mathbf{W}\mathbf{S}^l(\mathbf{S}^l)^{-1}\right] \\
=&-\left(\mathbf{0},\quad \mathbf{K}^r(\mathbf{X}^h,\mathbf{X}^h)^{-1}\otimes(\mathbf{S}^r)^{-1}\mathbf{W}\right),
\end{aligned}
\tag{A.21}
$$

and

$$
\begin{aligned}
&\mathbf{T}^{-1}+\mathbf{T}^{-1}\mathbf{U}\mathbf{Q}^{-1}\mathbf{V}\mathbf{T}^{-1} \\
=&\left[\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)^{-1}\otimes(\mathbf{S}^l)^{-1}\right]+\left[\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)^{-1}\otimes(\mathbf{S}^l)^{-1}\right]\left[\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^h)\otimes\mathbf{S}^l\mathbf{W}^T\right] \\
&\times\left[\mathbf{K}^r(\mathbf{X}^h,\mathbf{X}^h)^{-1}\otimes(\mathbf{S}^r)^{-1}\right]\left[\mathbf{K}^l(\mathbf{X}^h,\mathbf{X}^l)\otimes\mathbf{W}\mathbf{S}^l\right]\left[\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)^{-1}\otimes(\mathbf{S}^l)^{-1}\right] \\
=&\left[\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)^{-1}\otimes(\mathbf{S}^l)^{-1}\right]+\left[\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)^{-1}\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^h)\mathbf{K}^r(\mathbf{X}^h,\mathbf{X}^h)^{-1}\mathbf{K}^l(\mathbf{X}^h,\mathbf{X}^l)\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)^{-1}\right] \\
&\otimes\left[(\mathbf{S}^l)^{-1}\mathbf{S}^l(\mathbf{S}^r)^{-1}\mathbf{S}^l(\mathbf{S}^l)^{-1}\right] \\
=&\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)^{-1}\otimes(\mathbf{S}^l)^{-1}+\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{K}^r(\mathbf{X}^h,\mathbf{X}^h)^{-1}\otimes\mathbf{W}^T\mathbf{S}^r\mathbf{W}\end{pmatrix}.
\end{aligned}
\tag{A.22}
$$

Putting all these elements together, we get the inversion of joint kernel matrix $\mathbf{\Sigma}^{-1}=$

$$
\begin{bmatrix}
(\mathbf{K}^l)^{-1}\otimes(\mathbf{S}^l)^{-1}+\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & (\mathbf{K}^r)^{-1}\otimes\mathbf{W}^T(\mathbf{S}^r)^{-1}\mathbf{W}\end{pmatrix} & -\begin{pmatrix}\mathbf{0}\\ (\mathbf{K}^r)^{-1}\otimes\mathbf{W}^T(\mathbf{S}^r)^{-1}\end{pmatrix} \\
-\left(\mathbf{0},(\mathbf{K}^r)^{-1}\otimes(\mathbf{S}^r)^{-1}\mathbf{W}\right) & (\mathbf{K}^r)^{-1}\otimes(\mathbf{S}^r)^{-1}
\end{bmatrix},
\tag{A.23}
$$

where $\mathbf{S}^l=\bigotimes_{m=1}^M\mathbf{S}_m^l$, $\mathbf{S}^r=\bigotimes_{m=1}^M\mathbf{S}_m^r$, $\mathbf{W}=\bigotimes_{m=1}^M\mathbf{W}_m$, $\mathbf{K}^l=\mathbf{K}^l(\mathbf{X}^l,\mathbf{X}^l)$, and $\mathbf{K}^r=\mathbf{K}^r(\mathbf{X}^h,\mathbf{X}^h)$ as defined in the main paper. With the property in Eq. (A.10), and defining $\mathbf{y}=[\mathrm{vec}(\mathbf{Y}^l)^T,\mathrm{vec}(\mathbf{Y}^h)^T]^T$, we can substitute Eq. (A.23) into the joint likelihood to derive the data fitting part of the joint likelihood

$$
\begin{aligned}
&\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} \\
=&\left(\mathrm{vec}(\mathbf{Y}^l)^T,\mathrm{vec}(\mathbf{Y}^l)^T(\mathbf{E}\otimes\mathbf{W}^T)+\mathrm{vec}(\mathbf{Y}^r)^T\right)\mathbf{\Sigma}^{-1}\begin{pmatrix}\mathrm{vec}(\mathbf{Y}^l)\\ (\mathbf{E}^T\otimes\mathbf{W})\mathrm{vec}(\mathbf{Y}^l)+\mathrm{vec}(\mathbf{Y}^r)\end{pmatrix} \\
=&\mathrm{vec}(\mathbf{Y}^l)^T\left(\mathbf{K}^l\otimes\mathbf{S}^l\right)^{-1}\mathrm{vec}(\mathbf{Y}^l)+\mathrm{vec}(\mathbf{Y}^l)^T\mathbf{E}\left((\mathbf{K}^r)^{-1}\otimes\mathbf{W}^T(\mathbf{S}^r)^{-1}\mathbf{W}\right)\mathbf{E}^T\mathrm{vec}(\mathbf{Y}^l) \\
&-\mathrm{vec}(\mathbf{Y}^l)^T\left(\mathbf{E}\otimes\mathbf{W}^T\right)\left((\mathbf{K}^r)^{-1}\mathbf{E}^T\otimes(\mathbf{S}^r)^{-1}\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l) \\
&-\mathrm{vec}(\mathbf{Y}^r)^T\left((\mathbf{K}^r)^{-1}\mathbf{E}^T\otimes(\mathbf{S}^r)^{-1}\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l) \\
&-\mathrm{vec}(\mathbf{Y}^l)^T\left(\mathbf{E}(\mathbf{K}^r)^{-1}\otimes\mathbf{W}^T(\mathbf{S}^r)^{-1}\right)\left[\left(\mathbf{E}^T\otimes\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l)+\mathrm{vec}(\mathbf{Y}^r)\right] \\
&+\mathrm{vec}(\mathbf{Y}^l)^T\left(\mathbf{E}\otimes\mathbf{W}^T\right)(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left[\left(\mathbf{E}^T\otimes\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l)+\mathrm{vec}(\mathbf{Y}^r)\right] \\
&+\mathrm{vec}(\mathbf{Y}^r)^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left[\left(\mathbf{E}^T\otimes\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l)+\mathrm{vec}(\mathbf{Y}^r)\right] \\
=&\mathrm{vec}(\mathbf{Y}^l)^T\left(\mathbf{K}^l\otimes\mathbf{S}^l\right)^{-1}\mathrm{vec}(\mathbf{Y}^l)+\mathrm{vec}(\mathbf{Y}^r)^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\mathrm{vec}(\mathbf{Y}^r) \\
&-\mathrm{vec}(\mathbf{Y}^r)^T\left((\mathbf{K}^r)^{-1}\mathbf{E}^T\otimes(\mathbf{S}^r)^{-1}\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l)+\mathrm{vec}(\mathbf{Y}^r)^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left(\mathbf{E}^T\otimes\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l) \\
=&\mathrm{vec}(\mathbf{Y}^l)^T\left(\mathbf{K}^l\otimes\mathbf{S}^l\right)^{-1}\mathrm{vec}(\mathbf{Y}^l)+\mathrm{vec}(\mathbf{Y}^r)^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\mathrm{vec}(\mathbf{Y}^r).
\end{aligned}
\tag{A.24}
$$

With the block matrix's determinant formula, we can also derive the determinant of joint kernel matrix,

$$
|\mathbf{\Sigma}|=\left|\mathbf{K}^l\otimes\mathbf{S}^l\right|\times|\mathbf{Q}|=\left|\mathbf{K}^l\otimes\mathbf{S}^l\right|\times|\mathbf{K}^r\otimes\mathbf{S}^r|\,.
\tag{A.25}
$$

where we do not decompose them further with the purpose of forming to independent GPs for the low- and high-fidelity. With the conclusion of Eq. (A.24), the full joint log-likelihood is

$$
\begin{aligned}
&\log p(\mathbf{Y}^l, \mathbf{Y}^h) \\
=& -\frac{1}{2}\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{d^l N^l + d^h N^h}{2}\log(2\pi) \\
=& \underbrace{-\frac{1}{2}\mathrm{vec}(\mathbf{Y}^l)^T\left(\mathbf{K}^l \otimes \mathbf{S}^l\right)^{-1}\mathrm{vec}(\mathbf{Y}^l) - \frac{1}{2}\log\left|\mathbf{K}^l \otimes \mathbf{S}^l\right| - \frac{N^l d^l}{2}\log(2\pi)}_{TGP\ for\ low-fidelity\ data} \\
& \underbrace{-\frac{1}{2}\mathrm{vec}(\mathbf{Y}^r)^T\left(\mathbf{K}^r \otimes \mathbf{S}^r\right)^{-1}\mathrm{vec}(\mathbf{Y}^r) - \frac{1}{2}\log|\mathbf{K}^r \otimes \mathbf{S}^r| - \frac{N^h d^h}{2}\log(2\pi)}_{TGP\ for\ residual\ information} \\
=& \log p(\mathbf{Y}^l) + \log p(\mathbf{Y}^h|\mathbf{Y}^l)
\end{aligned}
\tag{A.26}
$$

The meaning of $\mathbf{S}^l$, $\mathbf{S}^r$, $\mathbf{W}$, and $\mathbf{K}^r$ remain the same as defined in Eq. A.23.

B.1 Posterior distribution

For the posterior distribution we compute the mean function and covariance matrix with the assumption that the low- and high-fidelity data have a very strict subset requirement, $\mathbf{X}^h \subseteq \mathbf{X}^l$. With the conclusion in Lemma 2 and rule of block matrix multiplication, the mean function and covariance matrix have the following expression,

$$
\begin{aligned}
&\mathrm{vec}(\mathbf{Z}_*^h) \\
&= \left(\mathbf{k}_*^l \otimes \mathbf{W}\mathbf{S}^l, \mathbf{k}_*^l(\mathbf{X}^h) \otimes \mathbf{W}\mathbf{S}^l\mathbf{W}^T + \mathbf{k}_*^r \otimes \mathbf{S}^r\right)\boldsymbol{\Sigma}^{-1}\begin{pmatrix}\mathrm{vec}(\mathbf{Y}^l)\\ \mathrm{vec}(\mathbf{Y}^h)\end{pmatrix} \\
&= \left(\mathbf{k}_*^l \otimes \mathbf{W}\mathbf{S}^l\right)\left(\mathbf{K}^l \otimes \mathbf{S}^l\right)^{-1}\mathrm{vec}(\mathbf{Y}^l) + \left(\mathbf{k}_*^l \otimes \mathbf{W}\mathbf{S}^l\right)\left(\mathbf{E}(\mathbf{K}^r)^{-1}\mathbf{E}^T \otimes \mathbf{W}^T(\mathbf{S}^r)^{-1}\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l) \\
&\quad - \left(\mathbf{k}_*^l(\mathbf{X}^h) \otimes \mathbf{W}\mathbf{S}^l\mathbf{W}^T\right)\left(\mathbf{K}^r\mathbf{E}^T \otimes (\mathbf{S}^r)^{-1}\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l) \\
&\quad - (\mathbf{k}_*^r \otimes \mathbf{S}^r)\left(\mathbf{K}^r\mathbf{E}^T \otimes (\mathbf{S}^r)^{-1}\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l) \\
&\quad - \left(\mathbf{k}_*^l \otimes \mathbf{W}\mathbf{S}^l\right)\left(\mathbf{E}(\mathbf{K}^r)^{-1} \otimes \mathbf{W}^T(\mathbf{S}^r)^{-1}\right)\mathrm{vec}(\mathbf{Y}^h) \\
&\quad + \left(\mathbf{k}_*^l(\mathbf{X}^h) \otimes \mathbf{W}\mathbf{S}^l\mathbf{W}^T\right)\left(\mathbf{K}^r \otimes (\mathbf{S}^r)^{-1}\right)\mathrm{vec}(\mathbf{Y}^h) + (\mathbf{k}_*^r \otimes \mathbf{S}^r)\left(\mathbf{K}^r \otimes (\mathbf{S}^r)^{-1}\right)\mathrm{vec}(\mathbf{Y}^h) \\
&= \left(\mathbf{k}_*^l \otimes \mathbf{W}\mathbf{S}^l\right)\left(\mathbf{K}^l \otimes \mathbf{S}^l\right)^{-1}\mathrm{vec}(\mathbf{Y}^l) - (\mathbf{k}_*^r \otimes \mathbf{S}^r)\left(\mathbf{K}^r\mathbf{E}^T \otimes (\mathbf{S}^r)^{-1}\mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l) \\
&\quad + (\mathbf{k}_*^r \otimes \mathbf{S}^r)\left(\mathbf{K}^r \otimes (\mathbf{S}^r)^{-1}\right)\left(\mathbf{E}^T \otimes \mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l) + (\mathbf{k}_*^r \otimes \mathbf{S}^r)\left(\mathbf{K}^r \otimes (\mathbf{S}^r)^{-1}\right)\mathrm{vec}(\mathbf{Y}^r) \\
&= \left(\mathbf{k}_*^l(\mathbf{K}^l)^{-1} \otimes \mathbf{W}\right)\mathrm{vec}(\mathbf{Y}^l) + \left(\mathbf{k}_*^r(\mathbf{K}^r)^{-1} \otimes \mathbf{I}_r\right)\mathrm{vec}(\mathbf{Y}^r),
\end{aligned}
\tag{A.27}
$$

$$
\begin{aligned}
\mathbf{S}_*^h =& \left(\mathbf{k}^l(\mathbf{x}_*,\mathbf{x}_*)\otimes\mathbf{W}\mathbf{S}^l\mathbf{W}^T+\mathbf{k}^r(\mathbf{x}_*,\mathbf{x}_*)\otimes\mathbf{S}^r\right)- \\
&\left(\mathbf{k}_*^l\otimes\mathbf{W}\mathbf{S}^l,\mathbf{k}_*^l(\mathbf{X}^h)\otimes\mathbf{W}\mathbf{S}^l\mathbf{W}^T+\mathbf{k}_*^r\otimes\mathbf{S}^r\right)\boldsymbol{\Sigma}^{-1}\begin{pmatrix}(\mathbf{k}_*^l)^T\otimes\mathbf{S}^l\mathbf{W}^T\\ \mathbf{k}_*^l(\mathbf{X}^h)^T\otimes\mathbf{W}^T\mathbf{S}^l\mathbf{W}+(\mathbf{k}_*^r)^T\otimes\mathbf{S}^r\end{pmatrix} \\
=&\left(\mathbf{k}^l(\mathbf{x}_*,\mathbf{x}_*)\otimes\mathbf{W}\mathbf{S}^l\mathbf{W}^T+\mathbf{k}^r(\mathbf{x}_*,\mathbf{x}_*)\otimes\mathbf{S}^r\right) \\
&-\left(\mathbf{k}_*^l\otimes\mathbf{W}\mathbf{S}^l\right)\left(\mathbf{K}^l\otimes\mathbf{S}^l\right)^{-1}\left((\mathbf{k}_*^l)^T\otimes\mathbf{S}^l\mathbf{W}^T\right) \\
&-\left(\mathbf{k}_*^l\otimes\mathbf{W}\mathbf{S}^l\right)\left(\mathbf{E}(\mathbf{K}^l)^{-1}\mathbf{E}^T\otimes(\mathbf{S}^l)^{-1}\right)\left((\mathbf{k}_*^l)^T\otimes\mathbf{S}^l\mathbf{W}^T\right) \\
&+\left(\mathbf{k}_*^l(\mathbf{X}^h)\otimes\mathbf{W}\mathbf{S}^l\mathbf{W}^T\right)\left(\mathbf{K}^r\mathbf{E}^T\otimes(\mathbf{S}^r)^{-1}\mathbf{W}\right)\left((\mathbf{k}_*^l)^T\otimes\mathbf{S}^l\mathbf{W}^T\right) \\
&+(\mathbf{k}_*^r\otimes\mathbf{S}^r)\left(\mathbf{K}^r\mathbf{E}^T\otimes(\mathbf{S}^r)^{-1}\mathbf{W}\right)\left((\mathbf{k}_*^l)^T\otimes\mathbf{S}^l\mathbf{W}^T\right) \\
&+\left(\mathbf{k}_*^l\otimes\mathbf{W}\mathbf{S}^l\right)\left(\mathbf{E}(\mathbf{K}^r)^{-1}\otimes\mathbf{W}^T(\mathbf{S}^r)^{-1}\right)\left(\mathbf{k}_*^l(\mathbf{X}^h)^T\otimes\mathbf{W}^T\mathbf{S}^l\mathbf{W}+(\mathbf{k}_*^r)^T\otimes\mathbf{S}^r\right) \\
&-\left(\mathbf{k}_*^l(\mathbf{X}^h)\otimes\mathbf{W}\mathbf{S}^l\mathbf{W}^T\right)(\mathbf{K}^r\otimes(\mathbf{S}^r)^{-1})\left(\mathbf{k}_*^l(\mathbf{X}^h)^T\otimes\mathbf{W}^T\mathbf{S}^l\mathbf{W}+(\mathbf{k}_*^r)^T\otimes\mathbf{S}^r\right) \\
&-(\mathbf{k}_*^r\otimes\mathbf{S}^r)(\mathbf{K}^r\otimes(\mathbf{S}^r)^{-1})\left(\mathbf{k}_*^l(\mathbf{X}^h)^T\otimes\mathbf{W}^T\mathbf{S}^l\mathbf{W}+(\mathbf{k}_*^r)^T\otimes\mathbf{S}^r\right) \\
=&\left(\mathbf{k}^l(\mathbf{x}_*,\mathbf{x}_*)\otimes\mathbf{W}\mathbf{S}^l\mathbf{W}^T+\mathbf{k}^r(\mathbf{x}_*,\mathbf{x}_*)\otimes\mathbf{S}^r\right) \\
&-\left(\mathbf{k}_*^l\otimes\mathbf{W}\mathbf{S}^l\right)\left(\mathbf{K}^l\otimes\mathbf{S}^l\right)^{-1}\left((\mathbf{k}_*^l)^T\otimes\mathbf{S}^l\mathbf{W}^T\right) \\
&+(\mathbf{k}_*^r\otimes\mathbf{S}^r)\left(\mathbf{K}^r\mathbf{E}^T\otimes(\mathbf{S}^r)^{-1}\mathbf{W}\right)\left((\mathbf{k}_*^l)^T\otimes\mathbf{S}^l\mathbf{W}^T\right) \\
&-(\mathbf{k}_*^r\otimes\mathbf{S}^r)(\mathbf{K}^r\otimes(\mathbf{S}^r)^{-1})\left(\mathbf{k}_*^l(\mathbf{X}^h)^T\otimes\mathbf{W}^T\mathbf{S}^l\mathbf{W}\right) \\
&-(\mathbf{k}_*^r\otimes\mathbf{S}^r)(\mathbf{K}^r\otimes(\mathbf{S}^r)^{-1})\left((\mathbf{k}_*^r)^T\otimes\mathbf{S}^r\right) \\
=&\left(k_{**}^l-(\mathbf{k}_*^l)^T(\mathbf{K}^l)^{-1}\mathbf{k}_*^l\right)\otimes\mathbf{W}\mathbf{S}^l\mathbf{W}^T+\left(k_{**}^r-(\mathbf{k}_*^r)^T(\mathbf{K}^r)^{-1}\mathbf{k}_*^r\right)\otimes\mathbf{S}^r,
\end{aligned}
\tag{A.28}
$$

where the $\mathbf{W}$, $\mathbf{S}^l$ and $\mathbf{S}^r$ have the same meaning with the main paper.

B.2 Joint Likelihood for Non-Subset Multi-Fidelity Data

In the main paper and the subset section, we decompose the joint likelihood $\log p(\mathbf{Y}^h,\mathbf{Y}^l)$ into two parts as

$$
\log p(\mathbf{Y}^l,\mathbf{Y}^h)=\log p(\mathbf{Y}^l)+\log\int p(\mathbf{Y}^h|\hat{\mathbf{Y}}^l,\mathbf{Y}^l)p(\hat{\mathbf{Y}}^l|\mathbf{Y}^l)d\hat{\mathbf{Y}}^l
$$

where $p(\mathbf{Y}^h|\hat{\mathbf{Y}}^l,\mathbf{Y}^l)$ is the derived predictive posterior probability if the high-fidelity data are subset to the low-fidelity data.

$$
\begin{aligned}
&p(\mathbf{Y}^h|\hat{\mathbf{Y}}^l,\mathbf{Y}^l)=2\pi^{-\frac{N^hd^h}{2}}\times|\mathbf{K}^r\otimes\mathbf{S}^r|^{-\frac{1}{2}} \\
&\times\exp\left[-\frac{1}{2}\left[\begin{pmatrix}\mathrm{vec}(\check{\mathbf{Y}}^h)\\ \mathrm{vec}(\hat{\mathbf{Y}}^h)\end{pmatrix}-\hat{\mathbf{W}}\begin{pmatrix}\mathrm{vec}(\mathbf{Y}^l)\\ \mathrm{vec}(\hat{\mathbf{Y}}^l)\end{pmatrix}\right]^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left[\begin{pmatrix}\mathrm{vec}(\check{\mathbf{Y}}^h)\\ \mathrm{vec}(\hat{\mathbf{Y}}^h)\end{pmatrix}-\hat{\mathbf{W}}\begin{pmatrix}\mathrm{vec}(\mathbf{Y}^l)\\ \mathrm{vec}(\hat{\mathbf{Y}}^l)\end{pmatrix}\right]\right]
\end{aligned}
\tag{A.29}
$$

where we define $\hat{\mathbf{W}}=\mathbf{E}\otimes\bigotimes_{m=1}^M\mathbf{W}_m$. Based on the low-fidelity training data, we also have $p(\hat{\mathbf{Y}}^l|\mathbf{Y}^l)\sim\mathcal{N}(\bar{\mathbf{Y}}^l,\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)$ being a Gaussian.

$$
\begin{aligned}
&p(\hat{\mathbf{Y}}^l|\mathbf{Y}^l) \\
=&2\pi^{-\frac{N^md^l}{2}}\times\left|\hat{\mathbf{S}}^l\otimes\mathbf{S}^l\right|^{-\frac{1}{2}}\times\exp\left[-\frac{1}{2}\left(\mathrm{vec}(\hat{\mathbf{Y}}^l)-\mathrm{vec}(\bar{\mathbf{Y}}^l)\right)^T(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\left(\mathrm{vec}(\hat{\mathbf{Y}}^l)-\mathrm{vec}(\bar{\mathbf{Y}}^l)\right)\right],
\end{aligned}
\tag{A.30}
$$

where the $\hat{\mathbf{S}}^l \otimes \mathbf{S}^l$ is the posterior covariance matrix of the $\hat{\mathbf{Y}}^l$. We can combine Eq. (A.29) and Eq. (A.30) to derive the integral part of the joint likelihood

$$
\begin{aligned}
&\log \int p(\mathbf{Y}^h|\hat{\mathbf{Y}}^l, \mathbf{Y}^l) p(\hat{\mathbf{Y}}^l|\mathbf{Y}^l) d\hat{\mathbf{Y}}^l \\
=& -\frac{N^h d^h + N^m d^l}{2} \log(2\pi) - \frac{1}{2}\log|\mathbf{K}^r \otimes \mathbf{S}^r| - \frac{1}{2}\log\left|\hat{\mathbf{S}}^l \otimes \mathbf{S}^l\right| \\
&+ \log \int \exp\left\{ -\frac{1}{2}\left[\begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^h) \\ \mathrm{vec}(\hat{\mathbf{Y}}^h) \end{pmatrix} - \left(\mathbf{E}_n^T \otimes \mathbf{W}\right) \begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^l) \\ \mathrm{vec}(\hat{\mathbf{Y}}^l) \end{pmatrix} \right]^T (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \right. \\
&\quad \left[\begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^h) \\ \mathrm{vec}(\hat{\mathbf{Y}}^h) \end{pmatrix} - \left(\mathbf{E}_n^T \otimes \mathbf{W}\right) \begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^l) \\ \mathrm{vec}(\hat{\mathbf{Y}}^l) \end{pmatrix} \right] \\
&\quad \left. -\frac{1}{2}\left(\mathrm{vec}(\hat{\mathbf{Y}}^l)^T - \mathrm{vec}(\bar{\mathbf{Y}}^l)^T\right) (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} \left(\mathrm{vec}(\hat{\mathbf{Y}}^l) - \mathrm{vec}(\bar{\mathbf{Y}}^l)\right) \right\} d\mathrm{vec}(\hat{\mathbf{Y}}^l),
\end{aligned}
\tag{A.31}
$$

where the $\mathbf{X}^h = \mathbf{E}_n^T[\mathbf{X}^l, \hat{\mathbf{X}}^h]$. Since we know that $\hat{\mathbf{X}}^h = \hat{\mathbf{E}}^T \mathbf{X}^h$ and we assume that $\check{\mathbf{X}}^h = \check{\mathbf{E}}^T \mathbf{X}^h$, we can derive

$$
\begin{aligned}
&\begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^h) \\ \mathrm{vec}(\hat{\mathbf{Y}}^h) \end{pmatrix} - (\mathbf{E}_n^T \otimes \mathbf{W}) \begin{pmatrix} \mathrm{vec}(\mathbf{Y}^l) \\ \mathrm{vec}(\hat{\mathbf{Y}}^l) \end{pmatrix} \\
=& \begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^h) \\ \mathbf{0} \end{pmatrix} - \tilde{\mathbf{W}} \begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^l) \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathrm{vec}(\hat{\mathbf{Y}}^h) \end{pmatrix} - \hat{\mathbf{W}} \begin{pmatrix} \mathbf{0} \\ \mathrm{vec}(\hat{\mathbf{Y}}^l) \end{pmatrix} \\
=& \left(\check{\mathbf{E}} \otimes \mathbf{I}_h\right) \mathrm{vec}(\check{\mathbf{Y}}^h) - \left(\check{\mathbf{E}} \otimes \mathbf{W}\right) \mathrm{vec}(\check{\mathbf{Y}}^l) + \left(\hat{\mathbf{E}} \otimes \mathbf{I}_h\right) \mathrm{vec}(\check{\mathbf{Y}}^h) - \left(\hat{\mathbf{E}} \otimes \mathbf{W}\right) \mathrm{vec}(\check{\mathbf{Y}}^l).
\end{aligned}
\tag{A.32}
$$

in which the $\tilde{\mathbf{W}} = \mathbf{I}_{N^h} \otimes \mathbf{W}$, and $\check{\mathbf{Y}}^l$ denotes the corresponding part of the $\check{\mathbf{Y}}^h$. For convenience, we choose to compute the exponential part in Eq. (A.31) as our first step. We try to decompose it into the subset part, i.e., ($\check{\mathbf{Y}}^h$ and $\check{\mathbf{Y}}^l$) and the non-subset part, i.e., ($\hat{\mathbf{Y}}^h$ and $\hat{\mathbf{Y}}^l$). Eq. (A.32) will become handy for the later derivation. Let's first consider the data fitting part by substituting Eq. (A.32) into Eq. (A.31),

$$
\begin{aligned}
&-\frac{1}{2}\left[\begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^h) \\ \mathrm{vec}(\hat{\mathbf{Y}}^h) \end{pmatrix} - \hat{\mathbf{W}} \begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^l) \\ \mathrm{vec}(\hat{\mathbf{Y}}^l) \end{pmatrix} \right]^T (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \left[\begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^h) \\ \mathrm{vec}(\hat{\mathbf{Y}}^h) \end{pmatrix} - \hat{\mathbf{W}} \begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^l) \\ \mathrm{vec}(\hat{\mathbf{Y}}^l) \end{pmatrix} \right] \\
=& -\frac{1}{2}\left[\mathrm{vec}(\check{\mathbf{Y}}^h)^T \left(\check{\mathbf{E}}^T \otimes \mathbf{I}_h\right) - \mathrm{vec}(\check{\mathbf{Y}}^l)^T \left(\check{\mathbf{E}}^T \otimes \mathbf{W}^T\right)\right] (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \left[\left(\check{\mathbf{E}} \otimes \mathbf{I}_h\right) \mathrm{vec}(\check{\mathbf{Y}}^h) - \left(\check{\mathbf{E}} \otimes \mathbf{W}\right) \mathrm{vec}(\check{\mathbf{Y}}^h)\right] \\
&- \left[\mathrm{vec}(\check{\mathbf{Y}}^h)^T \left(\check{\mathbf{E}}^T \otimes \mathbf{I}_h\right) - \mathrm{vec}(\check{\mathbf{Y}}^l)^T \left(\check{\mathbf{E}}^T \otimes \mathbf{W}^T\right)\right] (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \left[\left(\hat{\mathbf{E}} \otimes \mathbf{I}_h\right) \mathrm{vec}(\hat{\mathbf{Y}}^h) - \left(\hat{\mathbf{E}} \otimes \mathbf{W}\right) \mathrm{vec}(\hat{\mathbf{Y}}^l)\right] \\
&- \frac{1}{2}\left[\mathrm{vec}(\hat{\mathbf{Y}}^h)^T \left(\hat{\mathbf{E}}^T \otimes \mathbf{I}_h\right) - \mathrm{vec}(\hat{\mathbf{Y}}^l)^T \left(\hat{\mathbf{E}}^T \otimes \mathbf{W}\right)\right] (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \left[\left(\hat{\mathbf{E}} \otimes \mathbf{I}_h\right) \mathrm{vec}(\hat{\mathbf{Y}}^h) - \left(\hat{\mathbf{E}} \otimes \mathbf{W}\right) \mathrm{vec}(\hat{\mathbf{Y}}^l)\right],
\end{aligned}
\tag{A.33}
$$

which gives us the decomposition as the subset part, the non-subset part, and the interaction part between them. Now we can substitute Eq. (A.33) into the integral part in Eq. (A.31),

$$
\begin{aligned}
&\log \int \exp[-\frac{1}{2}\left[\mathrm{vec}(\check{\mathbf{Y}}^h)^T\left(\check{\mathbf{E}}^T\otimes\mathbf{I}_h\right)-\mathrm{vec}(\check{\mathbf{Y}}^l)^T\left(\check{\mathbf{E}}^T\otimes\mathbf{W}^T\right)\right](\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left[\left(\check{\mathbf{E}}\otimes\mathbf{I}_h\right)\mathrm{vec}(\check{\mathbf{Y}}^h)-\left(\check{\mathbf{E}}\otimes\mathbf{W}\right)\mathrm{vec}(\check{\mathbf{Y}}^h)\right]\\
&\quad-\left[\mathrm{vec}(\check{\mathbf{Y}}^h)^T\left(\check{\mathbf{E}}^T\otimes\mathbf{I}_h\right)-\mathrm{vec}(\check{\mathbf{Y}}^l)^T\left(\check{\mathbf{E}}^T\otimes\mathbf{W}^T\right)\right](\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left[\left(\hat{\mathbf{E}}\otimes\mathbf{I}_h\right)\mathrm{vec}(\hat{\mathbf{Y}}^h)-\left(\hat{\mathbf{E}}\otimes\mathbf{W}\right)\mathrm{vec}(\hat{\mathbf{Y}}^l)\right]\\
&\quad-\frac{1}{2}\left[\mathrm{vec}(\hat{\mathbf{Y}}^h)^T\left(\hat{\mathbf{E}}^T\otimes\mathbf{I}_h\right)-\mathrm{vec}(\hat{\mathbf{Y}}^l)^T\left(\hat{\mathbf{E}}^T\otimes\mathbf{W}\right)\right](\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left[\left(\hat{\mathbf{E}}\otimes\mathbf{I}_h\right)\mathrm{vec}(\hat{\mathbf{Y}}^h)-\left(\hat{\mathbf{E}}\otimes\mathbf{W}\right)\mathrm{vec}(\hat{\mathbf{Y}}^l)\right]\\
&\quad-\frac{1}{2}\mathrm{vec}(\hat{\mathbf{Y}}^l)^T(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\mathrm{vec}(\hat{\mathbf{Y}}^l)-\frac{1}{2}\mathrm{vec}(\bar{\mathbf{Y}}^l)^T(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\mathrm{vec}(\bar{\mathbf{Y}}^l)+\mathrm{vec}(\hat{\mathbf{Y}}^l)^T(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\mathrm{vec}(\bar{\mathbf{Y}}^l)]d\mathrm{vec}(\hat{\mathbf{Y}}^l)\\
=&-\frac{1}{2}\left[\mathrm{vec}(\check{\mathbf{Y}}^h)^T\left(\check{\mathbf{E}}^T\otimes\mathbf{I}_h\right)-\mathrm{vec}(\check{\mathbf{Y}}^l)^T\left(\check{\mathbf{E}}^T\otimes\mathbf{W}^T\right)\right](\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left[\left(\check{\mathbf{E}}\otimes\mathbf{I}_h\right)\mathrm{vec}(\check{\mathbf{Y}}^h)-\left(\check{\mathbf{E}}\otimes\mathbf{W}\right)\mathrm{vec}(\check{\mathbf{Y}}^h)\right]\\
&\quad-\left[\mathrm{vec}(\check{\mathbf{Y}}^h)^T\left(\check{\mathbf{E}}^T\otimes\mathbf{I}_h\right)-\mathrm{vec}(\check{\mathbf{Y}}^l)^T\left(\check{\mathbf{E}}^T\otimes\mathbf{W}^T\right)\right](\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left(\hat{\mathbf{E}}\otimes\mathbf{I}_h\right)\mathrm{vec}(\hat{\mathbf{Y}}^h)\\
&\quad-\frac{1}{2}\mathrm{vec}(\hat{\mathbf{Y}}^h)^T\left(\hat{\mathbf{E}}^T\otimes\mathbf{I}_h\right)(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left(\hat{\mathbf{E}}\otimes\mathbf{I}_h\right)\mathrm{vec}(\hat{\mathbf{Y}}^h)-\frac{1}{2}\mathrm{vec}(\bar{\mathbf{Y}}^l)^T(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\mathrm{vec}(\bar{\mathbf{Y}}^l)\\
&\quad+\log\int\exp[\left[\mathrm{vec}(\check{\mathbf{Y}}^h)^T\left(\check{\mathbf{E}}^T\otimes\mathbf{I}_h\right)-\mathrm{vec}(\check{\mathbf{Y}}^l)^T\left(\check{\mathbf{E}}^T\otimes\mathbf{W}^T\right)\right](\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left(\hat{\mathbf{E}}\otimes\mathbf{W}\right)\mathrm{vec}(\hat{\mathbf{Y}}^l)\\
&\quad+\mathrm{vec}(\hat{\mathbf{Y}}^h)^T\left(\hat{\mathbf{E}}^T\otimes\mathbf{I}_h\right)(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left(\hat{\mathbf{E}}\otimes\mathbf{W}\right)\mathrm{vec}(\hat{\mathbf{Y}}^l)\\
&\quad-\frac{1}{2}\mathrm{vec}(\hat{\mathbf{Y}}^l)^T\left(\hat{\mathbf{E}}^T\otimes\mathbf{W}^T\right)(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\left(\hat{\mathbf{E}}\otimes\mathbf{W}\right)\mathrm{vec}(\hat{\mathbf{Y}}^l)\\
&\quad-\frac{1}{2}\mathrm{vec}(\hat{\mathbf{Y}}^l)^T(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\mathrm{vec}(\hat{\mathbf{Y}}^l)+\mathrm{vec}(\bar{\mathbf{Y}}^l)^T(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\mathrm{vec}(\hat{\mathbf{Y}}^l)]d\mathrm{vec}(\hat{\mathbf{Y}}^l)\\
=&-\frac{1}{2}\boldsymbol{\phi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\phi}-\frac{1}{2}\mathrm{vec}(\bar{\mathbf{Y}}^l)^T(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\mathrm{vec}(\bar{\mathbf{Y}}^l)\\
&\quad+\frac{1}{2}\left(\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\phi}+(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\mathrm{vec}(\bar{\mathbf{Y}}^l)\right)^T\left(\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\phi}+(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\right)^{-1}\\
&\quad\left(\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\phi}+(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\mathrm{vec}(\bar{\mathbf{Y}}^l)\right)\\
&\quad+\frac{N^md^l}{2}\log 2\pi+\frac{1}{2}\log\left|\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\Psi}+(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\right|\\
=&\frac{N^md^l}{2}\log 2\pi+\frac{1}{2}\log\left|\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\Psi}+(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\right|\\
&\quad-\frac{1}{2}\boldsymbol{\phi}^T\underbrace{\left[(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}-(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\Psi}\left(\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\Psi}+(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\right)^{-1}\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\right]}_{\text{part 1}}\boldsymbol{\phi}\\
&\quad-\frac{1}{2}\mathrm{vec}(\bar{\mathbf{Y}}^l)^T\underbrace{\left[(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}-(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\left(\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\Psi}+(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\right)^{-1}(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\right]}_{\text{part 2}}\mathrm{vec}(\bar{\mathbf{Y}}^l)\\
&\quad+\boldsymbol{\phi}^T\underbrace{(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\Psi}\left(\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\Psi}+(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\right)^{-1}(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}}_{\text{part 3}}\mathrm{vec}(\bar{\mathbf{Y}}^l)
\end{aligned}
\tag{A.34}
$$

where $\boldsymbol{\phi}$ is defined by Eq. (A.35) ,

$$
\begin{aligned}
\boldsymbol{\phi}&=\left((\mathrm{vec}(\check{\mathbf{Y}}^h)^T,\mathrm{vec}(\hat{\mathbf{Y}}^h)^T)-(\mathrm{vec}(\check{\mathbf{Y}}^l)^T,\mathbf{0})\tilde{\mathbf{W}}^T\right)\\
\boldsymbol{\Psi}&=\hat{\mathbf{E}}\otimes\mathbf{W}.
\end{aligned}
\tag{A.35}
$$

With Sherman-Morrison formula, we can further simplify part 1 in Eq. (A.34) as

$$
\begin{aligned}
&(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}-(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\Psi}\left(\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\boldsymbol{\Psi}+(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)^{-1}\right)^{-1}\boldsymbol{\Psi}^T(\mathbf{K}^r\otimes\mathbf{S}^r)^{-1}\\
=&\left(\mathbf{K}^r\otimes\mathbf{S}^r+\boldsymbol{\Psi}(\hat{\mathbf{S}}^l\otimes\mathbf{S}^l)\boldsymbol{\Psi}^T\right)^{-1},
\end{aligned}
\tag{A.36}
$$

part 2 in Eq. (A.34) as

$$\begin{aligned}
&(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} - (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} \left(\mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\right)^{-1} (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} \\
=&(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} - \left((\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l) + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\right)^{-1} \\
=&(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} - (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} + \mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}^T (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T)^{-1} \mathbf{\Psi} \\
=&\mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T)^{-1} \mathbf{\Psi},
\end{aligned} \tag{A.37}$$

and part 3 in Eq. (A.34) as

$$\begin{aligned}
&(\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} \left(\mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\right)^{-1} (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} \\
=&(\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} \left(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l - (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T)^{-1} \mathbf{\Psi} (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\right) (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} \\
=&(\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} - (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T)^{-1} \mathbf{\Psi} \\
=&(\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} - (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \left((\mathbf{K}^r \otimes \mathbf{S}^r) + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T - (\mathbf{K}^r \otimes \mathbf{S}^r)\right) \\
&(\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T)^{-1} \mathbf{\Psi} \\
=&(\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} - (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \left(\mathbf{I} - (\mathbf{K}^r \otimes \mathbf{S}^r)(\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T)^{-1}\right) \mathbf{\Psi} \\
=&(\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} - (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} + \left(\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T\right)^{-1} \mathbf{\Psi} \\
=&\left(\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T\right)^{-1} \mathbf{\Psi}.
\end{aligned} \tag{A.38}$$

With the simplifications for part 1, 2, and 3 we get in Eq. (A.36), Eq. (A.37) and Eq. (A.38), the integral part will be more compact by substituting Eq. (A.36), Eq. (A.37) and Eq. (A.38) back to Eq. (A.34) which is equal to

$$\begin{aligned}
&\frac{N^m d^l}{2} \log 2\pi + \frac{1}{2} \log \left|\mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\right| - \frac{1}{2} \boldsymbol{\phi}^T \left(\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T\right)^{-1} \boldsymbol{\phi} \\
&- \frac{1}{2} \mathrm{vec}(\bar{\mathbf{Y}}^l)^T \mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T)^{-1} \mathbf{\Psi} \mathrm{vec}(\bar{\mathbf{Y}}^l) \\
&+ \boldsymbol{\phi}^T \left(\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T\right)^{-1} \mathbf{\Psi} \mathrm{vec}(\bar{\mathbf{Y}}^l) \\
=&\frac{N^m d^l}{2} \log 2\pi + \frac{1}{2} \log \left|\mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\right| \\
&- \frac{1}{2} (\boldsymbol{\phi} - \mathbf{\Psi} \mathrm{vec}(\bar{\mathbf{Y}}^l))^T \left(\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T\right)^{-1} (\boldsymbol{\phi} - \mathbf{\Psi} \mathrm{vec}(\bar{\mathbf{Y}}^l)).
\end{aligned} \tag{A.39}$$

The determinant part of Eq. (A.31) of the matrix can also be decomposed in the following way,

$$\begin{aligned}
&-\frac{1}{2} \log |\mathbf{K}^r \otimes \mathbf{S}^r| - \frac{1}{2} \log \left|\hat{\mathbf{S}}^l \otimes \mathbf{S}^l\right| + \frac{1}{2} \log \left|\mathbf{\Psi}^T (\mathbf{K}^r \otimes \mathbf{S}^r)^{-1} \mathbf{\Psi} + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\right| \\
=&-\frac{1}{2} \log |\mathbf{K}^r \otimes \mathbf{S}^r| - \frac{1}{2} \log \left|\hat{\mathbf{S}}^l \otimes \mathbf{S}^l\right| \\
&+\frac{1}{2} \log |\mathbf{K}^r \otimes \mathbf{S}^r| + \frac{1}{2} \log \left|\hat{\mathbf{S}}^l \otimes \mathbf{S}^l\right| - \frac{1}{2} \log \left|\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T\right| \\
=&-\frac{1}{2} \log \left|\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T\right|.
\end{aligned} \tag{A.40}$$

Putting everything we have derived up to this point, the joint likelihood for the non-subset data is:

$$\begin{aligned}
&\log p(\mathbf{Y}^l, \mathbf{Y}^h) \\
=&\log p(\mathbf{Y}^l) - \frac{N^h d^h}{2} \log(2\pi) - \frac{1}{2} \log \left|\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T\right| \\
&- \frac{1}{2} (\boldsymbol{\phi} - \mathbf{\Psi} \mathrm{vec}(\bar{\mathbf{Y}}^l))^T \left(\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T\right)^{-1} (\boldsymbol{\phi} - \mathbf{\Psi} \mathrm{vec}(\bar{\mathbf{Y}}^l))
\end{aligned} \tag{A.41}$$

where $\boldsymbol{\phi} - \mathbf{\Psi} \mathrm{vec}(\bar{\mathbf{Y}}^l) = \begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^h) \\ \mathrm{vec}(\hat{\mathbf{Y}}^h) \end{pmatrix} - \tilde{\mathbf{W}} \begin{pmatrix} \mathrm{vec}(\check{\mathbf{Y}}^l) \\ \mathrm{vec}(\bar{\mathbf{Y}}^l) \end{pmatrix}$, and $\mathbf{K}^r \otimes \mathbf{S}^r + \mathbf{\Psi}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Psi}^T = \mathbf{K}^r \otimes \mathbf{S}^r + \hat{\mathbf{E}}\hat{\mathbf{S}}^l\hat{\mathbf{E}}^T \otimes \mathbf{W}^T \mathbf{S}^l \mathbf{W}$.

B.3 Posterior Distribution for Non-Subset Multi-Fidelity Data

We then explore the posterior distribution of this non-subset data structure. For the first, we use the integration to express the posterior distribution,

$$\begin{aligned} p\left(\mathbf{Y}^*|\mathbf{Y}^l,\mathbf{Y}^h\right) &= \int p(\mathbf{Y}^*,\hat{\mathbf{Y}}^l|\mathbf{Y}^h,\mathbf{Y}^l)d\hat{\mathbf{Y}}^l \\ &= \int p(\mathbf{Y}^*|\mathbf{Y}^h,\mathbf{Y}^l,\hat{\mathbf{Y}}^l)p(\hat{\mathbf{Y}}^l)d\hat{\mathbf{Y}}^l \end{aligned} \tag{A.42}$$

We try to express the integral by different parts. Once $\hat{\mathbf{Y}}^l$ is decided, the predictive posterior $p(\mathbf{Y}^*|\mathbf{Y}^h,\mathbf{Y}^l,\hat{\mathbf{Y}}^l)$ can be described using the standard subset way, which is $p(\mathbf{Y}^*|\mathbf{Y}^h,\mathbf{Y}^l,\hat{\mathbf{Y}}^l) \sim \mathcal{N}\left(\text{vec}(\bar{\mathbf{Z}}_*^h),\mathbf{S}_*^h\right)$, where the mean function and covariance matrix are

$$\begin{aligned} \text{vec}(\bar{\mathbf{Z}}_*^h) &= \left(\mathbf{k}_*^l(\mathbf{K}^l)^{-1}\otimes\mathbf{W}\right)\begin{pmatrix}\text{vec}(\mathbf{Y}^l)\\ \text{vec}(\hat{\mathbf{Y}}^l)\end{pmatrix} + \left(\mathbf{k}_*^r(\mathbf{K}^r)^{-1}\otimes\mathbf{I}^h\right)\text{vec}(\mathbf{Y}^r), \\ \mathbf{S}_*^h &= \left(k_{**}^l-(\mathbf{k}_*^l)^T(\mathbf{K}^l)^{-1}\mathbf{k}_*^l\right)\otimes\mathbf{W}\mathbf{S}^l\mathbf{W}^T + \left(k_{**}^r-(\mathbf{k}_*^r)^T(\mathbf{K}^r)^{-1}\mathbf{k}_*^r\right)\otimes\mathbf{S}^r. \end{aligned} \tag{A.43}$$

We further simplify the situation and introduce definitions:

$$\begin{aligned} \mathbf{K}_*^l &= k_{**}^l-(\mathbf{k}_*^l)^T(\mathbf{K}^l)^{-1}\mathbf{k}_*^l, \\ \mathbf{K}_*^r &= k_{**}^r-(\mathbf{k}_*^r)^T(\mathbf{K}^r)^{-1}\mathbf{k}_*^r \end{aligned}$$

Which simplify Eq. (A.43) as

$$\mathbf{S}_*^h = \mathbf{K}_*^l\otimes\mathbf{W}\mathbf{S}^l\mathbf{W}^T + \mathbf{K}_*^r\otimes\mathbf{S}^r.$$

At the same time, since $\hat{\mathbf{Y}}^l$ is the sample from $\mathbf{Y}^l$, so it also follows the posterior distribution in subset way, which means $\hat{\mathbf{Y}}^l \sim \mathcal{N}\left(\text{vec}(\bar{\mathbf{Y}}^l),\hat{\mathbf{S}}^l\otimes\mathbf{S}^l\right)$ where the $\text{vec}(\bar{\mathbf{Y}}^l)$ and $\mathbf{S}_*^l\otimes\mathbf{S}^l$ are

$$\begin{aligned} \text{vec}(\bar{\mathbf{Y}}^l) &= \left(\mathbf{k}_*^l(\mathbf{K}^l)^{-1}\otimes\mathbf{I}^l\right)\text{vec}(\mathbf{Y}^l), \\ \hat{\mathbf{S}}^l\otimes\mathbf{S}^l &= \left(k_{**}^l-(\mathbf{k}_*^l)^T(\mathbf{K}^l)^{-1}\mathbf{k}_*^l\right)\otimes\mathbf{S}^l. \end{aligned}$$

Therefore the posterior distribution of non-subset data structure is

$$
\begin{aligned}
& p\left(\mathbf{Y}^{*}|\mathbf{Y}^{l},\mathbf{Y}^{h}\right)\\
=&\int p(\mathbf{Y}^{*},\hat{\mathbf{Y}}^{l}|\mathbf{Y}^{h},\mathbf{Y}^{l})d\hat{\mathbf{Y}}^{l}\\
=&\int p(\mathbf{Y}^{*}|\mathbf{Y}^{h},\mathbf{Y}^{l},\hat{\mathbf{Y}}^{l})p(\hat{\mathbf{Y}}^{l})d\hat{\mathbf{Y}}^{l}\\
=&\int 2\pi^{-\frac{N^{p}d^{h}}{2}}\times\left|\mathbf{S}_{*}^{h}\right|^{-\frac{1}{2}}\\
&\times\exp\left[-\frac{1}{2}\left(\mathrm{vec}(\mathbf{Y}^{*})-\left(\mathbf{k}_{*}^{l}(\hat{\mathbf{K}}^{l})^{-1}\otimes\mathbf{W}\right)\begin{pmatrix}\mathrm{vec}(\mathbf{Y}^{l})\\ \mathrm{vec}(\hat{\mathbf{Y}}^{l})\end{pmatrix}-\left(\mathbf{k}_{*}^{r}(\mathbf{K}^{r})^{-1}\otimes\mathbf{I}^{h}\right)\mathrm{vec}(\mathbf{Y}^{r})\right)^{T}\right.\\
&\left.(\mathbf{S}_{*}^{h})^{-1}\left(\mathrm{vec}(\mathbf{Y}^{*})-\left(\mathbf{k}_{*}^{l}(\hat{\mathbf{K}}^{l})^{-1}\otimes\mathbf{W}\right)\begin{pmatrix}\mathrm{vec}(\mathbf{Y}^{l})\\ \mathrm{vec}(\hat{\mathbf{Y}}^{l})\end{pmatrix}-\left(\mathbf{k}_{*}^{r}(\mathbf{K}^{r})^{-1}\otimes\mathbf{I}^{h}\right)\mathrm{vec}(\mathbf{Y}^{r})\right)\right]\\
&\times 2\pi^{-\frac{N^{m}d^{l}}{2}}\times\left|\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l}\right|^{-\frac{1}{2}}\times\exp[-\frac{1}{2}(\mathrm{vec}(\hat{\mathbf{Y}}^{l})-\mathrm{vec}(\bar{\mathbf{Y}}^{l}))^{T}(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}(\mathrm{vec}(\hat{\mathbf{Y}}^{l})-\mathrm{vec}(\bar{\mathbf{Y}}^{l}))]d\hat{\mathbf{Y}}^{l}\\
=&2\pi^{-\frac{N^{p}d^{h}+N^{m}d^{l}}{2}}\times\left|\mathbf{S}_{*}^{h}\right|^{-\frac{1}{2}}\times\left|\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l}\right|^{-\frac{1}{2}}\times\exp\left[-\frac{1}{2}\tilde{\mathbf{Y}}^{T}(\mathbf{S}_{*}^{h})^{-1}\tilde{\mathbf{Y}}-\frac{1}{2}\mathrm{vec}(\bar{\mathbf{Y}}^{l})^{T}(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\mathrm{vec}(\bar{\mathbf{Y}}^{l})\right]\\
&\times\int\exp[+\tilde{\mathbf{Y}}^{T}(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}\mathrm{vec}(\hat{\mathbf{Y}}^{l})+\mathrm{vec}(\bar{\mathbf{Y}}^{l})^{T}(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\mathrm{vec}(\hat{\mathbf{Y}}^{l})\\
&-\frac{1}{2}\mathrm{vec}(\hat{\mathbf{Y}}^{l})^{T}\mathbf{\Gamma}^{T}(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}\mathrm{vec}(\hat{\mathbf{Y}}^{l})-\frac{1}{2}\mathrm{vec}(\hat{\mathbf{Y}}^{l})^{T}(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\mathrm{vec}(\hat{\mathbf{Y}}^{l})]d\hat{\mathbf{Y}}^{l}\\
=&2\pi^{-\frac{N^{p}d^{h}+N^{m}d^{l}}{2}}\times\left|\mathbf{S}_{*}^{h}\right|^{-\frac{1}{2}}\times\left|\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l}\right|^{-\frac{1}{2}}\times\exp\left[-\frac{1}{2}\tilde{\mathbf{Y}}^{T}(\mathbf{S}_{*}^{h})^{-1}\tilde{\mathbf{Y}}-\frac{1}{2}\mathrm{vec}(\bar{\mathbf{Y}}^{l})^{T}(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\mathrm{vec}(\bar{\mathbf{Y}}^{l})\right]\\
&\times 2\pi^{\frac{N^{m}d^{l}}{2}}\times\left|\mathbf{\Gamma}^{T}(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}+(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\right|^{-\frac{1}{2}}\times\exp\left[\frac{1}{2}\left(\tilde{\mathbf{Y}}^{T}(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}+\mathrm{vec}(\bar{\mathbf{Y}}^{l})^{T}(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\right)\right.\\
&\left.\left(\mathbf{\Gamma}^{T}(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}+(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\right)^{-1}\left(\tilde{\mathbf{Y}}^{T}(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}+\mathrm{vec}(\bar{\mathbf{Y}}^{l})^{T}(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\right)^{T}\right]\\
=&2\pi^{-\frac{N^{p}d^{h}}{2}}\times\underbrace{\left|\mathbf{S}_{*}^{h}\right|^{-\frac{1}{2}}\times\left|\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l}\right|^{-\frac{1}{2}}\times\left|\mathbf{\Gamma}^{T}(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}+(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\right|^{-\frac{1}{2}}}_{\text{part d}}\\
&\times\exp\left[-\frac{1}{2}\tilde{\mathbf{Y}}^{T}\underbrace{\left((\mathbf{S}_{*}^{h})^{-1}-(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}+(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\right)^{-1}\mathbf{\Gamma}^{T}(\mathbf{S}_{*}^{h})^{-1}\right)}_{\text{part a}}\tilde{\mathbf{Y}}\right.\\
&-\frac{1}{2}\mathrm{vec}(\bar{\mathbf{Y}}^{l})^{T}\underbrace{\left((\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}-(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\left(\mathbf{\Gamma}^{T}(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}+(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\right)^{-1}(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\right)}_{\text{part b}}\mathrm{vec}(\bar{\mathbf{Y}}^{l})\\
&\left.+\tilde{\mathbf{Y}}^{T}\underbrace{(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}\left(\mathbf{\Gamma}^{T}(\mathbf{S}_{*}^{h})^{-1}\mathbf{\Gamma}+(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}\right)^{-1}(\hat{\mathbf{S}}^{l}\otimes\mathbf{S}^{l})^{-1}}_{\text{part c}}\mathrm{vec}(\bar{\mathbf{Y}}^{l})\right]
\end{aligned}
\tag{A.44}
$$

where $\tilde{\mathbf{Y}}$ and $\mathbf{\Gamma}$ is defined by the following equation,

$$\begin{aligned}\tilde{\mathbf{Y}} &= \left(\text{vec}(\mathbf{Y}^*) - \left(\mathbf{k}_*^l(\hat{\mathbf{K}}^l)^{-1} \otimes \mathbf{W}\right)\begin{pmatrix}\text{vec}(\mathbf{Y}^l)\\ \mathbf{0}\end{pmatrix} - \left(\mathbf{k}_*^r(\mathbf{K}^r)^{-1} \otimes \mathbf{I}^h\right)\left(\text{vec}(\mathbf{Y}^h) - \begin{pmatrix}\text{vec}(\check{\mathbf{Y}}^l)\\ \mathbf{0}\end{pmatrix}\right)\right),\\ \mathbf{\Gamma} &= \left([\mathbf{k}_*^r(\mathbf{K}^r)^{-1}\mathbf{E}_n^T - \mathbf{k}_*^l(\hat{\mathbf{K}}^l)^{-1}] \otimes \mathbf{W}\right)\mathbf{E}_m \otimes \mathbf{I}^l.\end{aligned} \tag{A.45}$$

We then utilize the Sherman-Morrison formula to simplify part a, b, and c in Eq. (A.44) as follows. For part a in Eq. (A.44),

$$\begin{aligned}&(\mathbf{S}_*^h)^{-1} - (\mathbf{S}_*^h)^{-1}\mathbf{\Gamma}\left(\mathbf{\Gamma}^T(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma} + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\right)^{-1}\mathbf{\Gamma}^T(\mathbf{S}_*^h)^{-1}\\ =&\left(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T\right)^{-1},\end{aligned} \tag{A.46}$$

for part b in Eq. (A.44),

$$\begin{aligned}&(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} - (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\left(\mathbf{\Gamma}^T(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma} + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\right)^{-1}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\\ =&(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} - \left((\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l) + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\right)^{-1}\\ =&(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} - \left((\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1} - \mathbf{\Gamma}^T(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T)^{-1}\mathbf{\Gamma}\right)\\ =&\mathbf{\Gamma}^T(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T)^{-1}\mathbf{\Gamma},\end{aligned} \tag{A.47}$$

and for part c in Eq. (A.44),

$$\begin{aligned}&(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma}\left(\mathbf{\Gamma}^T(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma} + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\right)^{-1}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\\ =&(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma}\left((\hat{\mathbf{S}}^l \otimes \mathbf{S}^l) - (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T)^{-1}\mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\right)(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\\ =&(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma} - (\mathbf{S}_*^h)^{-1}\mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T)^{-1}\mathbf{\Gamma}\\ =&(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma} - (\mathbf{S}_*^h)^{-1}\left(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T - \mathbf{S}_*^h\right)(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T)^{-1}\mathbf{\Gamma}\\ =&(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma} - (\mathbf{S}_*^h)^{-1}\left(\mathbf{I} - \mathbf{S}_*^h(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T)^{-1}\right)\mathbf{\Gamma}\\ =&(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma} - (\mathbf{S}_*^h)^{-1}\mathbf{\Gamma} - (\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T)^{-1}\mathbf{\Gamma}\\ =&-(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T)^{-1}\mathbf{\Gamma}.\end{aligned} \tag{A.48}$$

And the determinant (part d in Eq. (A.44)) can also use the Sherman-Morrison formula to derive a more compact version,

$$\begin{aligned}&\left|\mathbf{S}_*^h\right|^{-\frac{1}{2}} \times \left|\hat{\mathbf{S}}^l \otimes \mathbf{S}^l\right|^{-\frac{1}{2}} \times \left|\mathbf{\Gamma}^T(\mathbf{S}_*^h)^{-1}\mathbf{\Gamma} + (\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\right|^{-\frac{1}{2}}\\ =&\left|\mathbf{S}_*^h\right|^{-\frac{1}{2}} \times \left|\hat{\mathbf{S}}^l \otimes \mathbf{S}^l\right|^{-\frac{1}{2}} \times \left|(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\right|^{-\frac{1}{2}} \times \left|(\mathbf{S}_*^h)^{-1}\right|^{-\frac{1}{2}} \times \left|(\mathbf{S}_*^h)^{-1} + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\mathbf{\Gamma}^T\right|^{-\frac{1}{2}}\\ =&\left|(\mathbf{S}_*^h)^{-1} + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\mathbf{\Gamma}^T\right|^{-\frac{1}{2}}\end{aligned} \tag{A.49}$$

Taking part a, b, c, and d back into Eq. (A.44), we have the compact form

$$\begin{aligned}&p\left(\mathbf{Y}^*|\mathbf{Y}^l, \mathbf{Y}^h\right)\\ =&2\pi^{-\frac{N^p d^h}{2}} \times \left|(\mathbf{S}_*^h)^{-1} + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\mathbf{\Gamma}^T\right|^{-\frac{1}{2}} \times \exp[-\frac{1}{2}\tilde{\mathbf{Y}}^T\left(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T\right)^{-1}\tilde{\mathbf{Y}}\\ &-\frac{1}{2}\text{vec}(\bar{\mathbf{Y}}^l)^T\mathbf{\Gamma}^T\left(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T\right)^{-1}\mathbf{\Gamma}\text{vec}(\bar{\mathbf{Y}}^l) - \tilde{\mathbf{Y}}^T\left(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T\right)^{-1}\mathbf{\Gamma}\text{vec}(\bar{\mathbf{Y}}^l)]\\ =&2\pi^{-\frac{N^p d^h}{2}} \times \left|(\mathbf{S}_*^h)^{-1} + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)^{-1}\mathbf{\Gamma}^T\right|^{-\frac{1}{2}}\\ &\times \exp[-\frac{1}{2}\left(\tilde{\mathbf{Y}} - \mathbf{\Gamma}\text{vec}(\bar{\mathbf{Y}}^l)\right)^T\left(\mathbf{S}_*^h + \mathbf{\Gamma}(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l)\mathbf{\Gamma}^T\right)^{-1}\left(\tilde{\mathbf{Y}} - \mathbf{\Gamma}\text{vec}(\bar{\mathbf{Y}}^l)\right)]\\ =&2\pi^{-\frac{d^h}{2}} \times \left|\mathbf{S}_*^h + \mathbf{\Gamma}\left(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l\right)\mathbf{\Gamma}^T\right|^{-\frac{1}{2}}\\ &\times \exp\left[-\frac{1}{2}\left(\text{vec}(\mathbf{Z}_*^h) - \text{vec}(\bar{\mathbf{Z}})\right)^T\left(\mathbf{S}_*^h + \mathbf{\Gamma}\left(\hat{\mathbf{S}}^l \otimes \mathbf{S}^l\right)\mathbf{\Gamma}^T\right)^{-1}\left(\text{vec}(\mathbf{Z}_*^h) - \text{vec}(\bar{\mathbf{Z}})\right)\right].\end{aligned} \tag{A.50}$$

We can see the joint likelihood ends up with a elegant formulation about the low-fidelity TGP and residual TGP.

B.4 CIGAR

As we mentioned in the Section 3.4, we assume the output covariance matrixes $\mathbf{S}_m^h$ and $\mathbf{S}_m^l$ are identical matrixes and orthogonal weight matrixes, i.e., $\mathbf{W}_m^T\mathbf{W}_m = \mathbf{I}$. Substituting these assumptions into (A.41), we get the simplified covariance matrix,

$$\begin{aligned}&\mathbf{K}^r \otimes \mathbf{I}^r + \hat{\mathbf{E}}\hat{\mathbf{S}}^l\hat{\mathbf{E}}^T \otimes \mathbf{W}^T\mathbf{I}^l\mathbf{W}\\ =&\mathbf{K}^r \otimes \mathbf{I}^r + \hat{\mathbf{E}}\hat{\mathbf{S}}^l\hat{\mathbf{E}}^T \otimes \mathbf{W}^T\mathbf{W}\\ =&\mathbf{K}^r \otimes \mathbf{I}^r + \hat{\mathbf{E}}\hat{\mathbf{S}}^l\hat{\mathbf{E}}^T \otimes \mathbf{I}^h\\ =&(\mathbf{K}^r + \hat{\mathbf{E}}\hat{\mathbf{S}}^l\hat{\mathbf{E}}^T) \otimes \mathbf{I}^h\end{aligned} \tag{A.51}$$

where $\mathbf{I}^r$ is a identical matrix of size $d^r \times d^r$; the same rules apply to $\mathbf{I}^r$; and $\mathbf{I}^r = \mathbf{I}^h$. The joint likelihood of non-subset data becomes

$$\begin{aligned}\log p(\mathbf{Y}^l, \mathbf{Y}^h) = \log p(\mathbf{Y}^l) - \frac{N^h d^h}{2}\log(2\pi) - \frac{1}{2}\log\left|(\mathbf{K}^r + \hat{\mathbf{E}}\hat{\mathbf{S}}^l\hat{\mathbf{E}}^T) \otimes \mathbf{I}^h\right|\\ - \frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\Psi}\mathrm{vec}(\bar{\mathbf{Y}}^l))^T\left((\mathbf{K}^r + \hat{\mathbf{E}}\hat{\mathbf{S}}^l\hat{\mathbf{E}}^T) \otimes \mathbf{I}^h\right)^{-1}(\boldsymbol{\phi} - \boldsymbol{\Psi}\mathrm{vec}(\bar{\mathbf{Y}}^l))\end{aligned} \tag{A.52}$$

where $\left(\boldsymbol{\phi} - \boldsymbol{\Psi}\mathrm{vec}(\bar{\mathbf{Y}}^l)\right) = \begin{pmatrix}\mathrm{vec}(\check{\mathbf{Y}}^h)\\ \mathrm{vec}(\hat{\mathbf{Y}}^h)\end{pmatrix} - \tilde{\mathbf{W}}\begin{pmatrix}\mathrm{vec}(\bar{\mathbf{Y}}^l)\\ \mathrm{vec}(\bar{\mathbf{Y}}^l)\end{pmatrix}$.

We can see that the complexity of kernel matrix inversion is reduced to $\mathcal{O}((N^h)^3)$.

B.5 τ-Fidelity Autoregression Model

As we mentioned in Section 2.1, we can apply the AR to more levels of fidelity, so the GAR does. In this section, we try to expand the GAR into more levels of fidelity. Assuming the $\mathbf{F}^\tau(\mathbf{x}) = \mathbf{F}^{\tau-1}(\mathbf{x}) \times_1 \mathbf{W}_1^{\tau-1} \times_2 \cdots \times_M \mathbf{W}_M^{\tau-1} + \mathbf{F}_\tau^r(\mathbf{x})$, we can derive the joint covariance matrix,

$$\boldsymbol{\Sigma}^\tau = \begin{pmatrix}\mathbf{K}^{\tau-1}(\mathbf{X}^{\tau-1}, \mathbf{X}^{\tau-1}) \otimes \mathbf{S}^{\tau-1} & \mathbf{K}^{\tau-1}(\mathbf{X}^{\tau-1}, \mathbf{X}^\tau) \otimes \mathbf{S}^{\tau-1}(\mathbf{W}^{\tau-1})^T\\ \mathbf{K}^{\tau-1}(\mathbf{X}^\tau, \mathbf{X}^{\tau-1}) \otimes \mathbf{W}^{\tau-1}\mathbf{S}^{\tau-1} & \mathbf{K}^{\tau-1}(\mathbf{X}^\tau, \mathbf{X}^\tau) \otimes \mathbf{W}^{\tau-1}\mathbf{S}^{\tau-1}(\mathbf{W}^{\tau-1})^T + \mathbf{K}_\tau^r(\mathbf{X}^\tau, \mathbf{X}^\tau) \otimes \mathbf{S}_\tau^r\end{pmatrix},$$

where $\mathbf{S}^{\tau-1} = \bigotimes_{m=1}^M \mathbf{S}_m^{\tau-1}$ and $\mathbf{W}^{\tau-1} = \bigotimes_{m=1}^M \mathbf{W}_m^{\tau-1}$.
As same as the proof of GAR, we can derive the inversion of the joint covariance matrix, $(\boldsymbol{\Sigma}^\tau)^{-1} = \begin{bmatrix}(\mathbf{K}^{\tau-1})^{-1} \otimes (\mathbf{S}^{\tau-1})^{-1} + \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & (\mathbf{K}_\tau^r)^{-1} \otimes \mathbf{W}^{\tau-1T}(\mathbf{S}_\tau^r)^{-1}\mathbf{W}^{\tau-1}\end{pmatrix} & -\begin{pmatrix}\mathbf{0}\\ (\mathbf{K}_\tau^r)^{-1} \otimes \mathbf{W}^{\tau-1T}(\mathbf{S}_\tau^r)^{-1}\end{pmatrix}\\ -\left(\mathbf{0}, (\mathbf{K}_\tau^r)^{-1} \otimes (\mathbf{S}_\tau^r)^{-1}\mathbf{W}^{\tau-1}\right) & (\mathbf{K}_\tau^r)^{-1} \otimes (\mathbf{S}_\tau^r)^{-1}\end{bmatrix}$
Therefore, we have shown here that building an s-level TGP is equivalent to building s independent TGPs. We present the mean function and covariance matrix of the posterior distribution,

$$\begin{aligned}\mathrm{vec}(\mathbf{Z}_*^\tau) &= \left(\mathbf{k}_*^{\tau-1}(\mathbf{K}^{\tau-1})^{-1} \otimes \mathbf{W}^{\tau-1}\right)\mathrm{vec}(\mathbf{Y}^{\tau-1}) + \left((\mathbf{k}_\tau^r)_*(\mathbf{K}_\tau^r)^{-1} \otimes \mathbf{I}_r\right)\mathrm{vec}(\mathbf{Y}_\tau^r)\\ \mathbf{S}_*^\tau &= \left(k_{**}^{\tau-1} - (\mathbf{k}_*^{\tau-1})^T(\mathbf{K}^{\tau-1})^{-1}\mathbf{k}_*^{\tau-1}\right) \otimes \mathbf{W}^{\tau-1}\mathbf{S}^{\tau-1}(\mathbf{W}^{\tau-1})^T + \left((k_\tau^r)_{**} - (\mathbf{k}_\tau^r)_*^T(\mathbf{K}_\tau^r)^{-1}(\mathbf{k}_\tau^r)_*\right) \otimes \mathbf{S}_\tau^r.\end{aligned} \tag{A.53}$$

C Summary of the SOTA methods

We compare and conclude the capability and complexity of the SOTA methods, GAR, and CIGAR in Table 1.

D Implementation and Complexity

We now present the training and prediction algorithm for GAR and CIGAR using tensor algebra so that the full covariance matrix is never assembled or explicitly computed to improve computational efficiency. We use a normal TGP as example, given the dataset $(\mathbf{X}, \mathbf{Y})$, $\mathrm{vec}(\mathbf{Y}) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}(\mathbf{X}, \mathbf{X}) \otimes \left(\bigotimes_{m=1}^M \mathbf{S}_m\right)\right)$. The inference needs to estimate all the covariance matrix $\bigotimes_{m=1}^M \mathbf{S}_m$ and $\mathbf{K}(\mathbf{X}, \mathbf{X})$. For compactness, we use $\mathbf{S}$ and $\mathbf{K}$ to denote $\bigotimes_{m=1}^M \mathbf{S}_m$ and $\mathbf{K}(\mathbf{X}, \mathbf{X})$, and $\boldsymbol{\Sigma} = \mathbf{K} \otimes \mathbf{S} + \epsilon^{-1}\mathbf{I}$. We estimate parameters by minimizing the negative log likelihood of the model,

$$\mathcal{L} = \frac{Nd}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\mathrm{vec}(\mathbf{Y})^T\boldsymbol{\Sigma}^{-1}\mathrm{vec}(\mathbf{Y}).$$

Table 1: Comparison of SOTA multi-fidelity fusion for high-diemsnosional problems

Model	Arbitrary outputs?	Non-subset data?	Complexity
NAR [16]	Yes	No	$\mathcal{O}(\sum_i (N^i)^3)$
ResGP[9]	No	No	$\mathcal{O}(\sum_i (N^i)^3)$
MF-BNN [13]	Yes	Yes	$\mathcal{O}(\sum_i (N^i)(A_i^2+\omega))$*
DC [12]	Yes	No	$\mathcal{O}(\sum_i (N^i)^3)$
AR [3]	No	No	$\mathcal{O}(\sum_i (N^i d^i)^3)$
GAR	Yes	Yes	$\mathcal{O}(\sum_i \sum_{m=1}^M (d_m^i)^3 + (N^i)^3)$
CIGAR	Yes	Yes	$\mathcal{O}(\sum_i (N^i)^3)$

*A_i is the total weight size of NN for i-th fidelity and ω is the number of all parameters

However, since the $\mathbf{S}$ is a matrix of size $Nd \times Nd$, when the size of outputs is large, it will be unable to compute the inversion of $\mathbf{K} \otimes \mathbf{S}$. So for the TGP, we exploit the Kronecker product in $\mathbf{K} \otimes \mathbf{S}$ to calculate the negative log-likelihood efficiently. Firstly, we use eigendecomposition to denote the joint kernel matrix, $\mathbf{K} = \mathbf{U}^T \text{diag}(\lambda)\mathbf{U}$ and $\mathbf{S}_m = \mathbf{U}_m^T \text{diag}(\lambda_m)\mathbf{U}_m$. Then we use $\mathbf{\Sigma}$ to denote the joint kernel matrix, $\mathbf{\Sigma} = \mathbf{K} \otimes \mathbf{S} + \epsilon^{-1}\mathbf{I} = \left(\mathbf{U}^T \text{diag}(\lambda)\mathbf{U}\right) \otimes \left(\mathbf{U}_1^T \text{diag}(\lambda_1)\mathbf{U}_1\right) \otimes \cdots \otimes \left(\mathbf{U}_M^T \text{diag}(\lambda_M)\mathbf{U}_M\right) + \epsilon^{-1}\mathbf{I}$. With the Kronecker product property, we can have that

$$\mathbf{\Sigma} = \mathbf{P}^T \Lambda \mathbf{P} + \epsilon^{-1}\mathbf{I} \tag{A.54}$$

where $\mathbf{P} = \mathbf{U} \otimes \mathbf{U}_1 \otimes \cdots \otimes \mathbf{U}_M$ and $\Lambda = \text{diag}(\lambda \otimes \lambda_1 \otimes \cdots \otimes \lambda_M)$ since $\mathbf{U}$ and $\mathbf{U}_m$ is eigenvectors and orthogonal, so $\mathbf{P}^T\mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{I}$. Therefore, we can have that

$$\log|\mathbf{\Sigma}| = \log\left|\mathbf{P}^T\Lambda\mathbf{P} + \epsilon^{-1}\mathbf{I}\right| = \log\left|\mathbf{P}^T(\Lambda + \epsilon^{-1}\mathbf{I})\mathbf{P}\right| = \log\left|\Lambda + \epsilon^{-1}\mathbf{I}\right|. \tag{A.55}$$

Therefore, we only need to compute Nd diagonal elements to calculate part of the negative log-likelihood.

After that, we compute the $\text{vec}(\mathbf{Y})^T\mathbf{\Sigma}^{-1}\text{vec}(\mathbf{Y})$ part in the negative log likelihood. First, we have $\mathbf{A} = \lambda \circ \lambda_1 \circ \cdots \circ \lambda_M + \epsilon^{-1}\mathbb{1}$, where $\mathbb{1}$ is a tensor of full ones and $\circ$ is the Kruskal operator. Then we have

$$\begin{aligned}
\text{vec}(\mathbf{Y})^T\mathbf{\Sigma}^{-1}\text{vec}(\mathbf{Y}) &= \text{vec}(\mathbf{Y})^T\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}^{-\frac{1}{2}}\text{vec}(\mathbf{Y}) \\
&= \text{vec}(\mathbf{Y})^T\mathbf{P}^T(\Lambda + \epsilon^{-1}\mathbf{I})^{-\frac{1}{2}}\mathbf{P}\mathbf{P}(\Lambda + \epsilon^{-1}\mathbf{I})^{-\frac{1}{2}}\mathbf{P}^T\text{vec}(\mathbf{Y}) \\
&= \eta^T\eta,
\end{aligned} \tag{A.56}$$

where $\eta = \mathbf{P}(\Lambda + \epsilon^{-1}\mathbf{I})^{-\frac{1}{2}}\mathbf{P}^T\text{vec}(\mathbf{Y})$. Since $\mathbf{P}$ is a Kronecker product matrix, we can apply the property of Tucker operator [19] to compute $\mathbf{b}$.

$$\begin{aligned}
&\mathbf{T}_1 = \mathbf{Y} \times_1 \mathbf{U}^T \times_2 \mathbf{U}_1^T \times_3 \cdots \times_{M+1} \mathbf{U}_M^T \\
&\mathbf{T}_2 = \mathbf{T}_1 \odot \mathbf{A}^{\cdot -\frac{1}{2}} \\
&\mathbf{T}_3 = \mathbf{T}_2 \times_1 \mathbf{U} \times_2 \mathbf{U}_1 \times_3 \cdots \times_{M+1} \mathbf{U}_M \\
&\eta = \text{vec}(\mathbf{T}_3)
\end{aligned} \tag{A.57}$$

where $\odot$ means element-wise product, and $(\cdot)^{\cdot -\frac{1}{2}}$ means take power of $-\frac{1}{2}$ element wisely. Therefore the complexity of negative log likelihood is $\mathcal{O}(\sum_{m=1}^M (d_m)^3 + (N)^3)$.
Based on the above conclusions, we can also calculate the GAR more efficiently. According to Lemma 3, the joint likelihood admits two separable likelihoods $\mathcal{L}^l$ and $\mathcal{L}^r$. For each of these two, we can use the tricks to reduce the complexity to $\mathcal{O}(\sum_{m=1}^M (d_m^l)^3 + (N^l)^3) + \mathcal{O}(\sum_{m=1}^M (d_m^r)^3 + (N^r)^3)$. Since,

$$\begin{aligned}
&\log\left|\mathbf{K}^l \otimes \mathbf{S}^l\right| = \log\left|\Lambda^l + \epsilon^{-1}\mathbf{I}^l\right|, \qquad \log|\mathbf{K}^r \otimes \mathbf{S}^r| = \log\left|\Lambda^r + \epsilon^{-1}\mathbf{I}^r\right|, \\
&\text{vec}(\mathbf{Y}^l)^T(\mathbf{K}^l \otimes \mathbf{S}^l)^{-1}\text{vec}(\mathbf{Y}^l) = (\eta^l)^T\eta^l; \quad \text{vec}(\mathbf{Y}^r)^T(\mathbf{K}^r \otimes \mathbf{S}^r)^{-1}\text{vec}(\mathbf{Y}^r) = (\eta^r)^T\eta^r,
\end{aligned} \tag{A.58}$$

in which η^h, η^l and Λ^h, Λ^l are low-fidelity data and residuals corresponding vectors and eigenvalues. Therefore, the joint log-likelihood will be,

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}^l + \mathcal{L}^r \\
&= \text{const} - \frac{1}{2}\log\left|\Lambda^l + \epsilon^{-1}\mathbf{I}^l\right| - \frac{1}{2}(\eta^l)^T\eta^l - \frac{1}{2}\log\left|\Lambda^r + \epsilon^{-1}\mathbf{I}^r\right| - \frac{1}{2}(\eta^r)^T\eta^r
\end{aligned} \tag{A.59}$$

Given a new input $\mathbf{x}_*$, the prediction of the output tensorized as $\text{vec}(\mathbf{Z}_*^h)$ is a conditional Gaussian distribution $\text{vec}(\mathbf{Z}_*) \sim \mathcal{N}(\text{vec}(\bar{\mathbf{Z}}_*), \mathbf{S}_*)$, where

$$\begin{aligned}
\text{vec}(\bar{\mathbf{Z}}_*) &= \left(\mathbf{k}_* (\mathbf{K})^{-1} \otimes \mathbf{I}\right)\text{vec}(\mathbf{Y}) \\
\mathbf{S}_* &= \left(k_{**} - (\mathbf{k}_*)^T (\mathbf{K})^{-1} \mathbf{k}_*\right) \otimes \mathbf{S}.
\end{aligned} \tag{A.60}$$

We can use the Tucker operator to compute the predictive mean $\mathrm{vec}(\bar{\mathbf{Z}}_*)$ and $\mathbf{S}_*$ in a more efficient way. Using the eigendecomposition of kernel matrix, we can derive that $\mathbf{S}_* = k_{**} \otimes \mathbf{S} - \mathbf{L}\mathbf{L}^T$, where $\mathbf{L} = ((\mathbf{k}_*)^T (\mathbf{K})^{-1} \mathbf{U} \otimes \mathbf{U}_1 \otimes \cdots \otimes \mathbf{U}_M)(\Lambda(\Lambda + \epsilon^{-1}\mathbf{I})^{-\frac{1}{2}})$. Therefore, the $\mathrm{diag}(\mathbf{S}_*) = k_{**} \otimes \mathrm{diag}(\mathbf{S}) - \mathrm{diag}(\mathbf{L}\mathbf{L}^T)$. We can also use tensor algebra to calculate the predictive covariance matrix

$$\mathrm{diag}(\mathbf{S}_*) = \mathrm{vec}(\mathbf{M}),$$

where $\mathbf{M} = k_{**}(\mathrm{diag}(\mathbf{S}_1) \circ \cdots \circ \mathrm{diag}(\mathbf{S}_M)) + \left((\lambda \circ \lambda_1 \circ \cdots \circ \lambda_M) \odot \mathbf{A}^{\cdot -\frac{1}{2}}\right)^{\cdot 2} \times_1 (\mathbf{k}_* \mathbf{K}^{-1} \mathbf{U})^{\cdot 2} \times_2 (\mathbf{U}_1)^{\cdot 2} \times_3 \cdots \times_{M+1} (\mathbf{U}_M)^{\cdot 2}$. Therefore, we can also compute the predictive covariance matrix $\mathbf{S}_*^h$ in GAR efficiently.

$$\mathrm{diag}(\mathbf{S}_*^h) = \mathrm{vec}(\mathbf{M}^l) + \mathrm{vec}(\mathbf{M}^r) \tag{A.61}$$

where the $\mathrm{vec}(\mathbf{M}^l)$ and $\mathrm{vec}(\mathbf{M}^r)$ are vectors for low-fidelity and residual data. When we calculate the $\mathrm{vec}(\mathbf{M}^l)$, we need to be careful that the output kernel matrix should be $\mathbf{W}\mathbf{S}^l\mathbf{W}^T$.

E Experiment in Detail

E.1 Canonical PDEs

We consider three canonical PDEs: Poisson's equation, the heat equation, and Burger's equation, These PDEs have crucial roles in scientific and technological applications [62, 56, 66]. They offer common simulation scenarios, such as high-dimensional spatial-temporal field outputs, nonlinearities, and discontinuities, and are frequently used as benchmark issues for surrogate models [12, 51–53]. x and y denote the spatial coordinates, and t specifies the time coordinate, which contradicts the notation in the main paper. This notation in the appendix serves merely to make the information clear; it has no bearing on or connections to the main article.

Burgers' equation is regarded as a standard nonlinear hyperbolic PDE; it is commonly used to represent a variety of physical phenomena, including fluid dynamics [56], nonlinear acoustics [57], and traffic flows [58]. It serves as a benchmark test case for several numerical solvers and surrogate models [59–61] since it can generate discontinuities (shock waves) based on a normal conservation equation. The viscous version of this equation is given by

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = v\frac{\partial^2 u}{\partial x^2},$$

where u indicates volume, x represents a spatial location, t indicates the time, and v denotes the viscosity. We set $x \in [0, 1]$, $t \in [0, 3]$, and $u(x, 0) = \sin(x\pi/2)$ with homogeneous Dirichlet boundary conditions. We uniformly sampled viscosities $v \in [0.001, 0.1]$ as the input parameter to generate the solution field.

In the space and time domains, the problem is solved using finite elements with hat functions and backward Euler, respectively. For the first (lowest-fidelity) solution, the spatial-temporal domain is discretized into 16×16 regular rectangular mesh. Higher-fidelity solvers double the number of nodes in each dimension of the mesh, e.g., 32×32 for the second fidelity and 64×64 for the third fidelity. The result fields (i.e., outputs) are calculated using a $128by128$ regular spatial-temporal mesh.

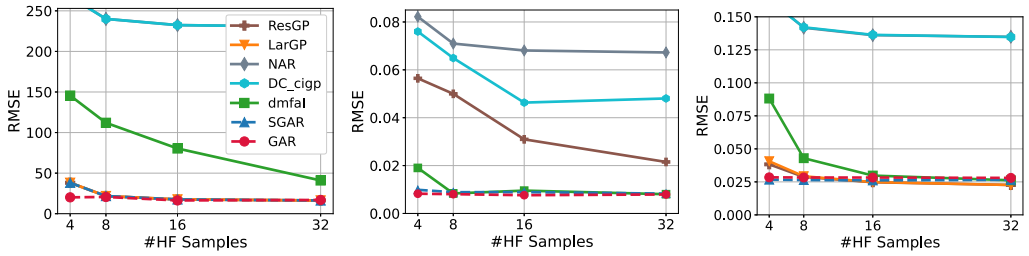
Poisson's equation is a typical elliptic PDE in mechanical engineering and physics for modeling potential fields, such as gravitational and electrostatic fields [62]. Written as

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

It is a generalization of Laplace's equation [63]. Despite its simplicity, Poisson's equation is commonly encountered in physics and is regularly used as a fundamental test case for surrogate models [51, 64]. In our experiments, we impose Dirichlet boundary conditions on a 2D spatial domain with $\mathbf{x} \in [0, 1] \times [0, 1]$. The input parameters consist of the constant values of the four borders and the center of the rectangular domain, which vary from 0.1 to 0.9 each. We sample the input parameters equally in order to create the matching potential fields as outputs. Using the finite difference approach with a first-order center differencing scheme and regular rectangular meshes, the PDE is solved. For the coarsest level solution, we utilized an 8×8 mesh. The improved solver employs a finer mesh with twice as many nodes in each dimension. The resultant potential fields are estimated using a spatial-temporal regular grid of 32×32 cells.

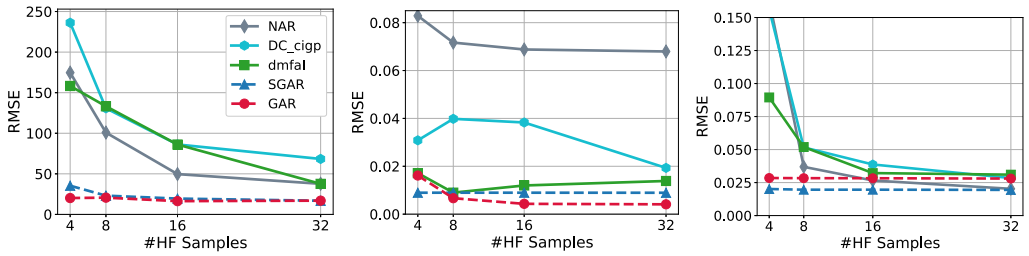
Heat equation is a fundamental PDE that defines the time-dependent evolution of heat fluxes. Despite having been established in 1822 to describe just heat fluxes, the heat equation is prevalent in many scientific domains, including probability theory [65, 66] and financial mathematics [67]. Consequently, it is commonly utilized as a stand-in model. This is the heat equation:

$$\frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left(k\frac{\partial T}{\partial y}\right) + \frac{\partial}{\partial z}\left(k\frac{\partial T}{\partial z}\right) + q_V = \rho c_p \frac{\partial T}{\partial t}$$



(a) Poisson's (b) Burger's (c) Heat's

Figure 6: RMSE against increasing number of high-fidelity training samples with training samples increased using Sobol sequence and aligned (interpolated) outputs.



(a) Poisson's (b) Burger's (c) Heat's

Figure 7: RMSE against increasing number of high-fidelity training samples with training samples increased using Sobol sequence and unaligned outputs.

where k is the materials conductivity q_V is the rate at which energy is generated per unit volume of the medium ρ is the density and c_p is the specific heat capacity. The input parameters are the flux rate of the left boundary at $x = 0$ (ranging from 0 to 1), the flux rate of the right boundary at $x = 1$ (ranging from -1 to 0), and the thermal conductivity (ranging from 0.01 to 0.1).

We establish a 2D spatial-temporal domain $x \in [0, 1], t \in [0, 5]$ with the Neumann boundary condition at$x = 0$ and $x = 1$, and $u(x, 0) = H(x - 0.25) - H(x - 0.75)$, where $H(\cdot)$ is the Heaviside step function.

The equation is solved using the finite difference in space and backward Euler in time domains. The spatial-temporal domain is discretized into a 16×16 regular rectangular mesh for the first (lowest) fidelity solver. A refined solver uses a 32×32 mesh for the second fidelity. The result fields are computed on a 100×100 spatial-temporal grid.

The equation is solved using a finite difference in the spatial domain and reverse Euler in the temporal domain. The spatial-temporal domain is discretized into an 8×8 regular rectangular mesh for the first (least accurate) solution. The second fidelity of an improved solver's mesh is a 32×32 grid. On a 100×100 spatial-temporal grid, the result fields are calculated.

E.2 Multi-Fidelity Fusion for Canonical PDEs

We use the same experimental setup as in Section 5.1 for these experiments with the only difference being that the training data is generated using a Sobol sequence. We generated 256 data samples for testing and 32 samples for training. We increased the number of high-fidelity training data gradually from 4 to 32 with the high-fidelity training data fixed to 32. Fig. 6 and Fig. 7 show the RMSE statistical results for aligned outputs using interpolated and original unaligned outputs. GAR and CIGAR outperform the competitors with a large margin with scarce high-fidelity training data as in the main paper. Similarly, the advantage of GAR and CIGAR are more obvious when dealing with non-aligned outputs, where GAR and CIGAR demonstrate a 5x reduction in RMSE with 4 and 8 high-fidelity training samples, surpassing the competitors by a wide margin.

E.3 Multi-Fidelity Fusion for Topology Optimization

We use GAR in a topology structure optimization problem, where the output is the best topology structure (in terms of maximum mechanical metrics like stiffness) of a layout of materials, such as alloy and concrete, given some design parameters like external force and angle. Topology structure optimization is a significant approach

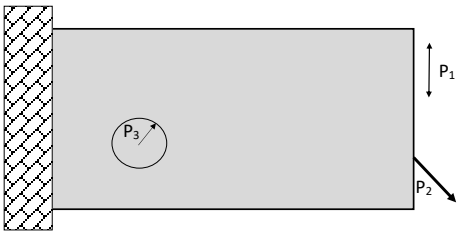
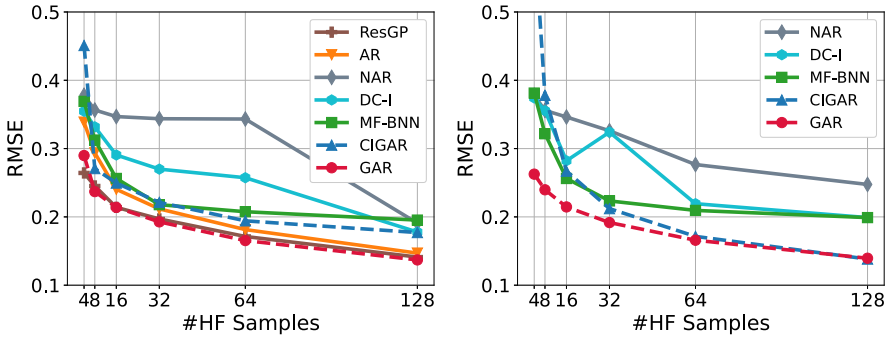


Figure 8: Geometry, boundary conditions, and simulation parameters for cantilever beam

in mechanical designs, such as airfoils and slab bridges, especially with recent 3D printing processes in which material is deposited in minute quantities. However, it is well known that topology optimization is computationally intensive due to the gradient-based optimization and simulations of the mechanical characteristics involved. A high-fidelity solution, which necessitates a huge discretization mesh and imposes a significant computing overhead in space and time, makes matters worse.

Utilizing data-driven ways to aid in the process by offering the appropriate structures [68, 13] is subsequently gaining popularity. Here, we investigate the topology optimization of a cantilever beam (shown in the appendix). We employ the rapid implementation [69] to carry out density-based topology optimization by reducing compliance C subject to volume limitations $V \leq \bar{V}$.

The SIMP scheme [70] is used to convert continuous density measurements to discrete, optimal topologies. We set the position of point load $P1$, the angle of point load $P2$, and the filter radius $P3$ [55] as system input. We solve this challenge for low-fidelity with a 40×80 regular mesh and high-fidelity with a 40×80 regular mesh. This experiment only includes techniques that can process arbitrary outputs.



(a) Aligned outputs

(b) Raw outputs

Figure 9: RMSE against increasing number of high-fidelity training samples for topology optimization using Sobol sequence.

As with the early experiments, we generate 128 testing samples and 64 training samples using a Sobol sequence to approximately assess the ance in active learning. The results are shown in Figure 9. We can see that all available methods show similar performance for both raw outputs that are not aligned by interpolation and the aligned outputs. Nevertheless, GAR consistently outperforms the competitors with a clear margin. CIGAR, in contrast, performs better for the raw outputs.

E.4 Multi-Fidelity Fusion for Solid Oxide Fuel Cell

In this test problem, a steady-state 3-D solid oxide fuel cell model is considered. Fig 10 illustrates the geometry. The model incorporates electronic and ionic charge balances (Ohm's law), flow distribution in gas channels (Navier-Stokes equations), flow in porous electrodes (Brinkman equation), and gas-phase mass balances in both gas channels and porous electrodes (Maxwell-Stefan diffusion and convection). Butler-Volmer charge transfer kinetics is assumed for reactions in the anode ($H_2 + O^{2-} \rightarrow H_2O + 2e^-$) and cathode ($O_2 + 4e^- \rightarrow 2O^{2-}$). The cell functions in a potentiostat manner (constant cell voltage). COMSOL Multiphysics[7] (Application ID: 514), which uses the finite-element approach, was used to solve the model.

[7] https://www.comsol.com/model/current-density-distribution-in-a-solid-oxide-fuel-cell-514

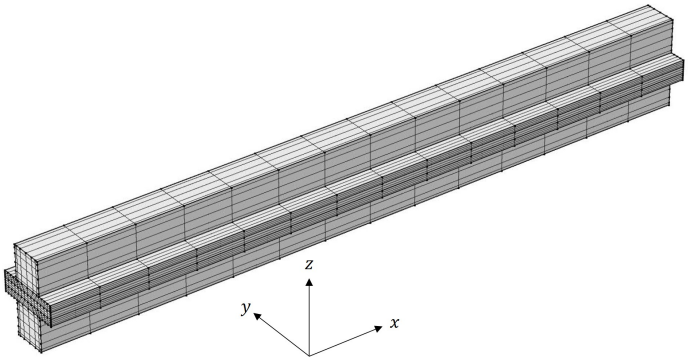


Figure 10: The cathode is at the top of the computational domain for the SOFC example, which consists of gas channels, electrodes, and electrolyte. The layers are, from top to bottom, a channel, an electrode, an electrolyte, an electrode, and a channel. The dimensions of the channel are (x * y * z) 1 cm * 0.5 mm * 0.5 mm, the dimensions of the electrode are 1 cm * 1 mm * 0.1 mm, and the dimensions of the electrolyte are 1 cm * 1 mm * 0.1 mm. The cathode intake is placed at $x = 1$ cm while the anode inlet is located at $x = 0$ cm.

The assumed inputs are the electrode porosities $\epsilon \in [0.4, 0.85]$, the cell voltage $E_c \in [0.2, 0.85]$ V, the temperature $T \in [973, 1273]$ K, and the channel pressure $P \in [0.5, 2.5]$ atm. A Sobol sequence is used to choose 60 inputs within the ranges specified for the low-fidelity and high-fidelity simulations. 40 high-fidelity test points are chosen at random (from the ranges above) to complete the test. The low-fidelity F1 model used 3164 mapped elements and relative tolerance of 0.1, while the high-fidelity model employed 37064 elements and relative tolerance of 0.001. Additionally, the COMSOL model employs a V cycle geometric multigrid. The quantities of interest are profiles of electrolyte current density (A m-2) and ionic potential (V) in the $x - z$ plane centered on the channels (Fig. 10). In both instances, $d = 100 \times 50 = 5000$ points are captured, and both profiles are vectorized to provide the training and test outputs.

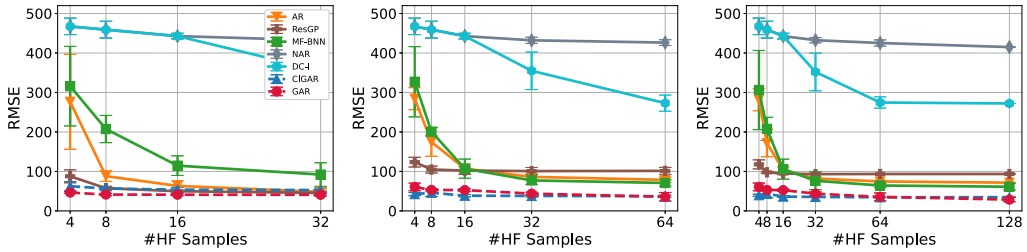


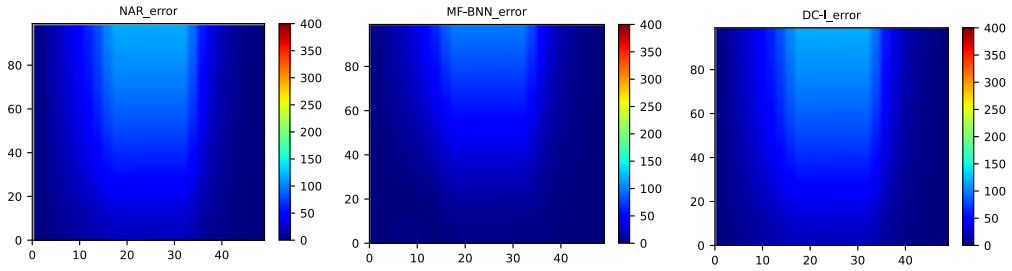
Figure 11: RMSE against increasing number of high-fidelity training samples for SOFC with low-fidelity training sample number fixed to {32,64,128}.

We add the classic experiment where the number of low-fidelity training samples was fixed to {32,64,128} and the high-fidelity training samples are gradually increased from 4 to {32,64,128}. The outputs are aligned using interpolation, and the experimental results are shown in Fig. 11. We can see that the GAR and CIGAR methods always perform better than the other methods, especially when only a few high-fidelity training data are used. This is consistent with the previous experiment. We can also see that AR also performs well indicating that these data are not highly nonlinear and complex, making it relatively easy to solve. However, both AR and MF-BNN converge to a higher error whereas GAR and CIGAR converge to a lower error bound.

To investigate the prediction error in detail, we define the average RMSE field $\mathbf{Z}^{(\mathrm{AEF})}$ by

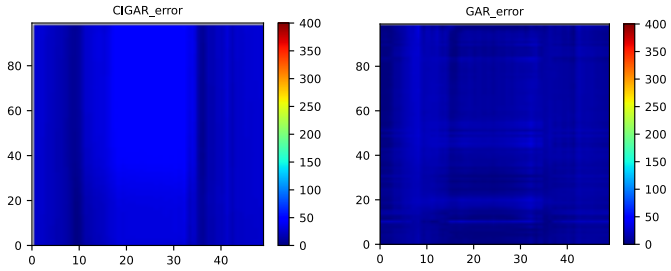
$$\mathbf{Z}^{(\mathrm{AEF})} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{Z}_i - \tilde{\mathbf{Z}}_i)^2},$$

where $\tilde{\mathbf{Z}}_i$ is the prediction, $\mathbf{Z}_i$ is the ground true value, and the square root is element-wise operation. Fig. 12 shows the average RMSE field of NAR, MF-BNN, and DC methods on the ECD in SOFC data with 32 low-fidelity training samples, 16 high-fidelity training samples, and 128 test samples. To highlight the advantage of GAR and CIGAR, Fig. 13 shows the average RMSE field of the same setup with **only 4 high-fidelity training samples**. It can be seen clearly that GAR and CIGAR have a smaller error field even with only 4 high-fidelity



(a) NAR (b) MF-BNN (c) DC

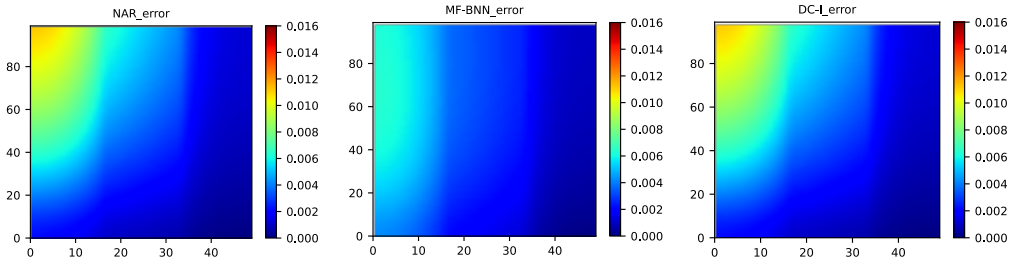
Figure 12: RMSE fields of ECD for 128 testing samples, using 32 low-fidelity and 16 high-fidelity training samples.



(a) CIGAR (b) GAR

Figure 13: RMSE fields of ECD for 128 testing samples, using 32 low-fidelity and **4** high-fidelity training samples.

training samples compared to NAR, MF-BNN, and DC with 16 high-fidelity training samples. Also note that GAR seems to have some checkerboard artifacts, which might be caused by the over-parameterization using a full transfer matrix. We leave this issue to our further work to resolve. CIGAR have fewer checkerboard artifacts with the price of a slight increase in the RMSE.



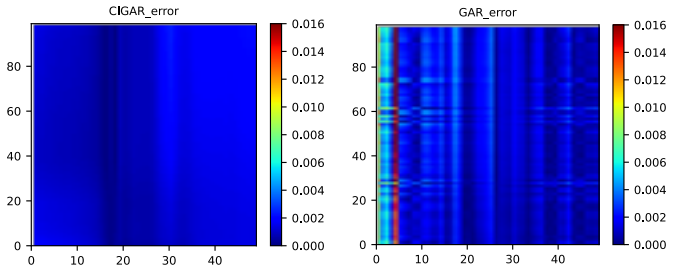
(a) NAR (b) MF-BNN (c) DC

Figure 14: RMSE fields of IP for 128 testing samples, using 32 low-fidelity and 16 high-fidelity training samples.

In Fig. 14 and Fig. 15, similar to the previous experimental setup, we draw the average RMSE with 128 testing samples on the IP fields from the SOFC dataset. The NAR, MF-BNN and DC are trained with 16 high-fidelity samples, while GAR and CIGAR are trained with only 4 high-fidelity samples. We can see that our methods outperform other methods by a clear margin. However, the checkerboard artifact is even worse for GAR in this case, whereas CIGAR successfully reduces such an artifact with also low error.

E.5 Plasmonic Nanoparticle Arrays Simulations

In the final example, we calculate the extinction and scattering efficiencies Q_{ext} and Q_{sc} for plasmonic systems with varying numbers of scatterers using the Coupled Dipole Approximation (CDA) approach. CDA is a method for mimicking the optical response of an array of similar, non-magnetic metallic nanoparticles with dimensions



(a) CIGAR (b) GAR

Figure 15: RMSE fields of IP for 128 testing samples, using 32 low-fidelity and **4** high-fidelity training samples.

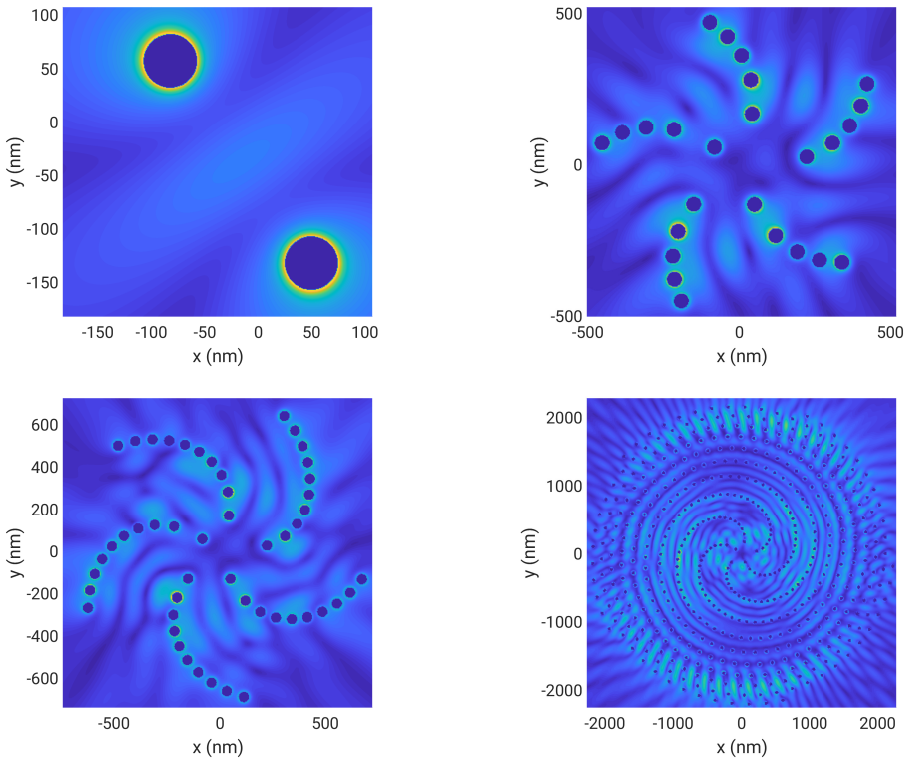


Figure 16: Sample configurations of Vogel spirals with $\{2, 25, 50, 500\}$ particles.

far smaller than the wavelength of light (here 25 nm). Q_{ext} and Q_{sc} are defined as the QoIs in this document. We construct surrogate models for efficiency with up to three fidelities using our proposed method. We examine particle arrays resulting from Vogel spirals. Since the number of interactions of incident waves from particles influences the magnetic field, the number of nanoparticles in a plasmonic array has a substantial effect on the local extinction field caused by plasmonic arrays. The configurations of Vogel spirals with particle numbers in the set $\{2, 25, 50\}$ that define fidelities F1 through F3 are depicted in Fig. 16. $\lambda \in [200, 800]$ nm, $\alpha_{vs} \in [0, 2\pi]$ rad, and $a_{vs} \in (1, 1500)$ are determined to be the parameter space. These are, respectively, the incidence wavelength, the divergence angle, and the scaling factor. A Sobol sequence is utilized to choose inputs. The computing time requires to execute CDA increases exponentially as the number of nanoparticles increases. Consequently, the proposed sampling approach results in significant reductions in computational costs.

The response of a plasmonic array to electromagnetic radiation is calculable using the solution of the local electric fields, $\mathbf{E}_{loc}(\mathbf{r}_j)$, for each nano-sphere. Considering N metallic particles defined by the same volumetric polarizability $\alpha(\omega)$ and situated at vector coordinates $\mathbf{r}_i$, it is possible to calculate the local field $\mathbf{E}_{loc}(\mathbf{r}_j)$ by solving [71] the corresponding linear equation.

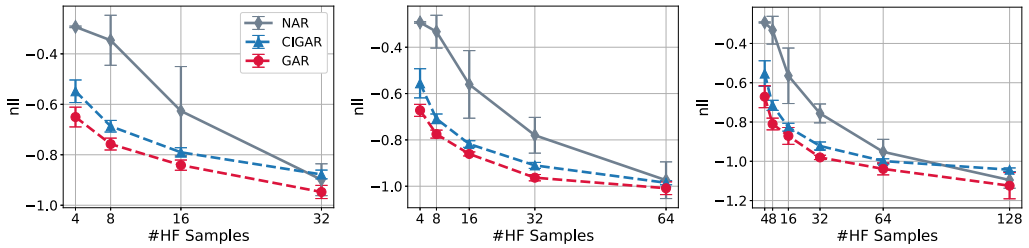
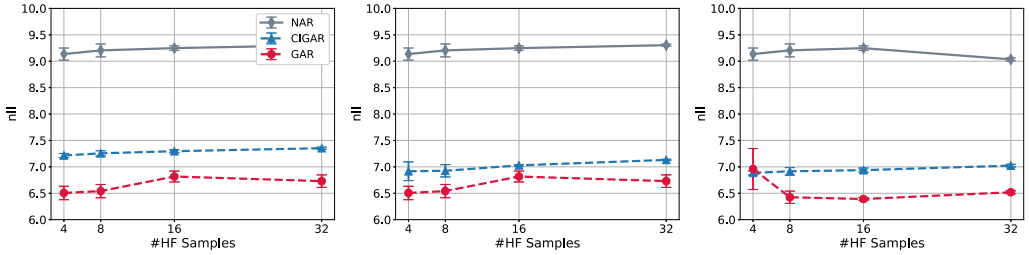


Figure 17: NLL with low-fidelity training sample number fixed to {32,64,128} for topology structure predictions.



(a) ECD+IP (b) ECD (c) IP

Figure 18: NLL for SOFC with low-fidelity training sample number fixed to 32.

$$\mathbf{E}_{loc}(\mathbf{r}_i) = \mathbf{E}_0(\mathbf{r}_i) - \frac{\alpha k^2}{\epsilon_0} \sum_{j=1, j\neq i}^{N} \tilde{\mathbf{G}}_{ij} \mathbf{E}_{loc}(\mathbf{r}_j) \tag{A.62}$$

in which $\mathbf{E}_0(\mathbf{r}_i)$ is the incident field, k is the wave number in the background medium, ϵ_0 denotes the dielectric permittivity of vacuum ($\epsilon_0 = 1$ in the CGS unit system), and $\tilde{\mathbf{G}}_{ij}$ is constructed from 3×3 blocks of the overall $3N \times 3N$ Green's matrices for the ith and jth particles. $\tilde{\mathbf{G}}_{ij}$ is a zero matrix when $j = i$, and otherwise calculated as

$$\tilde{\mathbf{G}}_{ij} = \frac{\exp(ikr_{ij})}{r_{ij}} \left\{ \mathbf{I} - \hat{\mathbf{r}}_{ij}\hat{\mathbf{r}}_{ij}^T - \left[\frac{1}{ikr_{ij}} + \frac{1}{(kr_{ij})^2} (\mathbf{I} - 3\hat{\mathbf{r}}_{ij}\hat{\mathbf{r}}_{ij}^T) \right] \right\} \tag{A.63}$$

where $\hat{\mathbf{r}}_{ij}$ denotes the unit position vector from particles j to i and $r_{ij} = |\mathbf{r}_{ij}|$. By solving Eqs. A.62 and A.63, the total local fields $\mathbf{E}_{loc}(\mathbf{r}_i)$, and as a result the scattering and extinction cross-sections, are computed. Details of the numerical solution can be found in [72].

Q_{ext} and Q_{sc} are derived by normalizing the scattering and extinction cross-sections relative to the array's entire projected area. We considered the Vogel spiral class of particle arrays, which is described by [73]

$$\rho_n = \sqrt{n} a_{vs} \quad \text{and} \quad \theta_n = n\alpha_{vs}, \tag{A.64}$$

where ρ_n and θ_n represent the radial distance and polar angle of the n-th particle in a Vogel spiral array, respectively. Therefore, the Vogel spiral configuration may be uniquely defined by the incidence wavelength λ, the divergence angle α_{vs}, the scaling factor a_{vs}, and the number of particles n.

E.6 Metircs for the Predictive Uncertainty

Despite that RMSE has been used as a standard metric for evaluating the performance of a multi-fidelity fusion algorithm [9, 12, 13, 16], a metric that considers the predictive uncertainty is also important [47], particularly when the downstream applications rely heavily on the quality of the predictive confidence, e.g., in MFBO [23]. To assess the proposed method more comprehensively, we evaluate the quality of the predictive posterior using the most commonly used metric, negative-log-likelihood (nll).

We repreduce Figs. 2 and 5 using exactly the same experimental setups but with the nll metric, and the results are shown in Figs. 17 and 18. Note that the nll of DC and MF-BNN is every poor, probably due to our implementations, and cannot be fitted into the figures. Thus they are not shown in the figures. Also note that some figures show negative nll. This is because our computation of the nll omits the constant term. Nevertheless, this modification does not affect the comparison. We can see that for the topology structure posterior in Fig. 17, the results are consistent with the conclusion drawn on the RMSE results. Since the CIGAR ignores the

inter-output correlations, it will overestimate the covariance determinant, leading to higher nll than GAR. The NAR starts with poor performance with a small number of training data. It consistently improves with increasing number of training data and end up with similar perform as GAR and CIGAR. Similarly, the SOFC results are consistent with the conclusion for the RMSE results. However, all methods demonstrated do not improve their performance significantly with more training data. This is caused by the calculations of the nll and the data itself. More specifically, in the ECD and IP fields, there are a few spatial locations where the recorded values are almost constant (caused by the Dirichlet boundary conditions). In this case, the nll will be dominated by the logarithm of variance and becomes less informative for the quality of the predictive variance. We thus see that the nll in Fig. 18 fluctuates around the same values no matter how many training points are used. We leave investigating the uncertainty metric using more advance metric (e.g., [74]) more in depth in the future considering the scope of this work.