DEEP LEARNING ENHANCED COST-AWARE MULTI-FIDELITY UNCERTAINTY QUANTIFICATION OF A COMPUTATIONAL MODEL FOR RADIOTHERAPY

PIERMARIO VITULLO, NICOLA RARES FRANCO AND PAOLO ZUNINO*

MOX, Department of Mathematics, Politecnico di Milano, Italy

ABSTRACT. Forward uncertainty quantification (UQ) for partial differential equations is a many-query task that requires a significant number of model evaluations. The objective of this work is to mitigate the computational cost of UQ for a 3D-1D multiscale computational model of microcirculation. To this purpose, we present a deep learning enhanced multi-fidelity Monte Carlo (DL-MFMC) method that integrates the information of a multiscale full-order model (FOM) with that coming from a deep learning enhanced non-intrusive projection-based reduced order model (ROM). The latter is constructed by leveraging on proper orthogonal decomposition (POD) and mesh-informed neural networks (previously developed by the authors and co-workers), integrating diverse architectures that approximate POD coefficients while introducing fine-scale corrections for the microstructures. The DL-MFMC approach provides a robust estimator of specific quantities of interest and their associated uncertainties, with optimal management of computational resources. In particular, the computational budget is efficiently divided between training and sampling, ensuring a reliable estimation process suitably exploiting the ROM speed-up. Here, we apply the DL-MFMC technique to accelerate the estimation of biophysical quantities regarding oxygen transfer and radiotherapy outcomes. Compared to classical Monte Carlo methods, the proposed approach shows remarkable speed-ups and a substantial reduction of the overall computational cost.

1. **Introduction.** Uncertainty quantification (UQ) is a key aspect of computational modeling [46]. The challenge of UQ becomes even more pronounced in the context of life sciences, where the complexity of physical models is combined with the high uncertainties of the data [11]. A particular case of this broad scenario, and the focus of this work, is the problem of microcirculation and oxygen transfer in biological tissues, here addressed using the multiscale model proposed in [38]. Quantifying how uncertainties propagate through these models, which are governed by complex partial differential equations (PDEs), requires extensive computational resources. In fact, to obtain robust and reliable estimates, numerous simulations are required, which means that the numerical solver has to be queried multiple times (a so-called *many-query* scenario). For this reason, traditional UQ methods that rely solely on high-fidelity simulations are often impractical, especially when

it comes to multiscale models, where different scales need to be resolved accurately, and parametric data are high-dimensional [30].

To address these issues, the state-of-the-art in UQ has been evolving rapidly, with multi-fidelity methods emerging as a promising approach to mitigate the computational burden [36]. Simply put, multi-fidelity methods are based on the concept of combining high-fidelity models with one or more lower-fidelity models. The idea is that, being computationally cheaper but less accurate, lower-fidelity models can be used to inform and accelerate the computation performed by the high-fidelity model [36, 29, 19, 10]. Furthermore, the applicability of multi-fidelity techniques has been further enhanced by recent advances in the reduced-order modeling literature, a research field devoted to the development of suitable surrogate models encompassing accuracy and efficiency [5, 20, 43]. In particular, a major contribution along this direction has been provided by nonintrusive techniques based upon interpolation and regression algorithms [1, 2, 3, 4, 6, 8, 21, 18, 15, 52]. In general, these approaches have gained traction in UQ due to its ability to significantly reduce computational time while maintaining acceptable levels of accuracy and robustness [7, 9].

Over the years, multi-fidelity methods have evolved from simple model hierarchies to more sophisticated techniques that intelligently balance the trade-off between computational cost and accuracy. One such technique is the multi-fidelity Monte Carlo (MFMC), which, complemented by a suitable strategy for optimal management of computational resources, has shown great promise in efficiently estimating the statistics of model outputs under uncertainty [35]. This methodology, provides a comprehensive framework for combining an arbitrary number of surrogate models while also balancing offline and online costs. Scientifically speaking, its derivation is particularly innovative, as it moves away from the traditional reliance on error decay rates. Instead, it introduces an optimization problem that accounts for both errors and costs, acknowledging the fact that ROMs require a training phase in order to be operational. The paper demonstrates that, under certain mild conditions, the optimization problem admits a unique analytic solution, thus providing a practical way to manage the computational budget.

In this work, we contribute to the evolving landscape of MFMC methods by extending the ideas in [35] to the case of deep learning-based ROMs, with a strong emphasis on our motivating application in relation to a multiscale computational model for microcirculation. More precisely, referring to the comprehensive model presented in [40, 38], we address oxygen transfer from microvessels to interstitial tissue. This model provides a description of the oxygen field in the microvascular environment, which directly affects the performance of radiotherapy: see, e.g., [41] and references therein.

Our approach integrates a full-order model (FOM) with a non-intrusive projection-based reduced order model (ROM) enhanced by deep learning, previously developed by the authors in [51]. Simply put, the latter integrates proper orthogonal decomposition (POD) with mesh-informed neural networks [16], effectively relying on a closure modeling technique that resembles a separation of scales approach. This combination not only reduces the computational cost, but also retains the essential features captured by the FOM, spanning both local and global scales. We term our deep learning-enhanced MFMC approach as DL-MFMC.

Building upon [35], we construct the DL-MFMC approach in order to account for two additional sources of computational cost: on the one hand, the training of the neural network architectures; on the other hand, the computational effort

required to generate the input data, which, in microcirculation studies can be non-negligible. By applying our method to the estimation of statistics related to oxygen transfer and radiotherapy in microcirculation, we demonstrate its effectiveness in performing robust and reliable UQ analysis in a computationally efficient manner. In this sense, our work not only contributes to the development of UQ multi-fidelity methods, but it also addresses a specific challenge in the context of microcirculation and radiation therapy.

The paper is organized as follows. First, in Section 2 we set the notation and present the DL-MFMC approach in full generality, discussing the construction of our multi-fidelity estimator and the corresponding optimal management of computational resources. Then, in Section 3 we dive into the details of our motivating application, presenting the FOM, the ROM, and the quantities of interest that we wish to estimate. Finally, we devote Section 4 to the numerical experiments and draw the corresponding conclusions in Section 5.

2. **A deep learning enhanced multi-fidelity Monte Carlo estimator.** In this section, we briefly review two different approaches to estimate the statistics of specific quantities of interest and provide suitable confidence intervals.

2.1. **Problem setup.** We are given a parametrized partial differential equation (PDE) of the form

$$\begin{cases} L_{\boldsymbol{\mu}} u_{\boldsymbol{\mu}} = f_{\boldsymbol{\mu}} & \text{in } \Omega, \\ B_{\boldsymbol{\mu}} u_{\boldsymbol{\mu}} = g_{\boldsymbol{\mu}} & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where $L_{\boldsymbol{\mu}}$ denotes a (semilinear, second order) elliptic operator, whereas $B_{\boldsymbol{\mu}}$ is a boundary operator exemplifying Dirichlet, Neumann or Robin conditions applied to $u_{\boldsymbol{\mu}}$ on $\partial\Omega$, including problems with mixed-type conditions. Clearly, the solution $u = u_{\boldsymbol{\mu}}$ to (1) depends on the parameter vector $\boldsymbol{\mu}$. Here, we assume $\boldsymbol{\mu}$ to be taking values in some parameter space $\mathcal{P} \subset \mathbb{R}^d$, endowed with suitable probability distribution $\mathbb{P}$ modeling uncertainties.

At the continuous level, we assume that the solution to (1) can be sought within a given Hilbert space V. Our main goal is to quantify how uncertainties in the model parameters can propagate through the PDE, ultimately affecting certain outputs of interest. To this end, let us fix the notation and formally introduce a (nonlinear) functional

$$Q : V \to \mathbb{R},$$

representing the quantity of interest (QoI). Then, our objective is to come up with an efficient strategy for computing the average response, that is, $\mathbb{E}_{\boldsymbol{\mu}\sim\mathbb{P}}[Q(u_{\boldsymbol{\mu}})]$.

As a first step, we assume that a suitable high-fidelity discretization of (1) is available (the so-called FOM). For simplicity, we shall assume that the latter consists of a Galerkin projection of (1) onto a Finite Element (FE) space $V_h \subset V$ of dimension $N_h = \dim(V_h)$. As in the continuous case, the FOM defines a map $\mathcal{P} \mapsto V_h$ mapping parameters onto FOM solutions $\boldsymbol{\mu} \mapsto u_{\boldsymbol{\mu}}^{\text{FOM}}$. From the discrete standpoint, the FOM is equivalent to a (large) system of algebraic equations of the form

$$\mathbf{A}_{\boldsymbol{\mu}} \mathbf{u}_{\boldsymbol{\mu}}^{\text{FOM}} = \mathbf{f}_{\boldsymbol{\mu}} \tag{2}$$

where $\mathbf{u}_{\boldsymbol{\mu}}^{\text{FOM}} \in \mathbb{R}^{N_h}$ is the vector of degrees of freedom in the FE approximation. With little abuse of notation, we shall write $Q(\mathbf{u}_{\boldsymbol{\mu}}^{\text{FOM}})$ to intend $Q(u_{\boldsymbol{\mu}}^{\text{FOM}})$, that is: we identify Q with its discrete counterpart, so as to directly operate with vectors

rather than functions.

From a theoretical point of view, the output of interest, namely $\mathbb{E}_{\boldsymbol{\mu} \sim \mathbb{P}}\left[Q(\mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{FOM}})\right]$, can be estimated by classical Monte Carlo sampling. However, this approach would require solving (2) multiple times, resulting in a massive consumption of computational resources. For this reason, we resort to multi-fidelity strategies relying upon ROMs. Mathematically speaking, the ROM is a computational unit that acts as a suitable surrogate of the FOM, that is

$$\mathrm{ROM}: \ \boldsymbol{\mu} \mapsto \mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{ROM}} \qquad \text{with} \qquad \mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{ROM}} \approx \mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{FOM}}.$$

Notice that, since the ROM approximates FOM solutions as a whole, it is not bound to any specific QoI, resulting in a highly flexibile and potentially very useful surrogate. Of note, this approach is further motivated by some recent advancements in the ROM literature, which suggest that approximating the parameter-to-solution map can be more effective compared to directly addressing the parameter-to-QoI map: we refer the interested reader to [22]. Here, we assume that the ROM consists of several neural network architectures, all trained within a supervised learning framework. In other words, to be operational, the ROM requires a preliminary training phase, which is conducted on a selected collection of labeled FOM simulations, $\{(\boldsymbol{\mu}_i, \mathbf{u}_i^{\mathrm{FOM}})\}_{i=1}^n$, where $\mathbf{u}_i^{\mathrm{FOM}} := \mathbf{u}_{\boldsymbol{\mu}_i}^{\mathrm{FOM}}$.

In general, the idea is to train a ROM and construct a multi-fidelity estimator of the QoI, named $\hat{E}_{m_0, m_1}^{\mathrm{MFMC}}$ and defined later on, by combining m_0 high-fidelity simulations, which are accurate but expensive, with m_1 low-fidelity simulations, computed via the ROM, which are cheaper to evaluate but less accurate. Clearly, the whole procedure must be carried out in a suitable way that ensures an actual reduction in the overall computational cost. To better appreciate this, let w_0 and w_1 be the computational times associated with the FOM and ROM simulations, respectively. Assume that the ROM is trained on n high-fidelity samples, whereas the multi-fidelity estimator is constructed using m_0 evaluations of the FOM and m_1 evaluations of the ROM. Then, the overall computational cost is

$$nw_0 + t(n) + m_0 w_0 + m_1 w_1,$$

where $t = t(n)$ is the training time associated with the ROM. In contrast, a classical Monte Carlo estimator $\hat{E}_N^{\mathrm{FOM}}$, constructed using N FOM simulations, entails a computational cost of Nw_0. Assuming the case of unbiased estimators, the uncertainties in the two estimates can be computed as

$$\mathrm{Var}(\hat{E}_{m_0, m_1}^{\mathrm{MFMC}}) \qquad \text{and} \qquad \mathrm{Var}(\hat{E}_N^{\mathrm{FOM}}).$$

It is then clear that, given a computational budget p, a multi-fidelity approach would only advantageous if for $p = nw_0 + t(n) + m_0 w_0 + m_1 w_1 = Nw_0$, one has $\mathrm{Var}(\hat{E}_{m_0, m_1}^{\mathrm{MFMC}}) < \mathrm{Var}(\hat{E}_N^{\mathrm{FOM}})$. We note that the computational cost (generally measured in terms of floating-point operations) and the CPU time (measured in clock ticks or in seconds and highly dependent on the machine architecture) are considered here to be mutually proportional and used interchangeably.

Our purpose is to obtain a reduction of computational cost by proposing a suitable strategy for optimal management of the computational resources, together with an explicit multi-fidelity estimator that leverages on the deep learning nature of the ROM. Before doing so, however, it is worth recalling some of the basic ideas underlying Monte Carlo and multi-fidelity Monte Carlo methods, such as confidence intervals and mean-square-error metrics.

2.2. **Monte Carlo uncertainty quantification.** We start by recalling the standard Monte Carlo estimator for the mean of a QoI. Given a computational budget p, the latter is computed using FOM simulations as

$$\hat{E}_N^{\mathrm{FOM}} = \frac{1}{N} \sum_{i=1}^{N} Q(\mathbf{u}_{\boldsymbol{\mu}_i}^{\mathrm{FOM}}), \tag{3}$$

where N is the number of high-fidelity evaluations, $N = \lfloor p/w_0 \rfloor$, with w_0 being the cost of a single FOM evaluation. Here, we assume that the realizations of the input parameters $\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_N$ are drawn independently and according to $\mathbb{P}$.

By definition, the FOM estimator is unbiased, in the sense that

$$\mathbb{E}\left[\hat{E}_N^{\mathrm{FOM}}\right] = \mathbb{E}\left[Q\left(\mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{FOM}}\right)\right]. \tag{4}$$

Note that here the expected value on the right-hand-side is taken with respect to $\boldsymbol{\mu} \sim \mathbb{P}$. Conversely, the one on the left is taken with respect to all possible outcomes of the sampling procedure, that is, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N \sim \mathbb{P} \otimes \cdots \otimes \mathbb{P}$.

As a direct consequence of (4), the mean-squared-error (MSE) of the estimator

$$\mathrm{MSE}(\hat{E}_N^{\mathrm{FOM}}) := \mathbb{E}\left|\hat{E}_N^{\mathrm{FOM}} - \mathbb{E}\left[Q(\mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{FOM}})\right]\right|^2, \tag{5}$$

is equal to the variance of the estimator. Due to independence, the latter is

$$\mathrm{Var}(\hat{E}_N^{\mathrm{FOM}}) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} Q(\mathbf{u}_{\boldsymbol{\mu}_i}^{\mathrm{FOM}})\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma_0^2 = \frac{\sigma_0^2}{N}, \tag{6}$$

where $\sigma_0^2 := \mathrm{Var}\, Q(\mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{FOM}})$ is the variance of the QoI, taken with respect to $\boldsymbol{\mu} \sim \mathbb{P}$. In practice, σ_0^2 is usually estimated as

$$\widehat{\sigma}_0^2 := \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{E}_N^{\mathrm{FOM}} - Q(\mathbf{u}_{\boldsymbol{\mu}_i}^{\mathrm{FOM}})\right)^2. \tag{7}$$

Using classical results from the theory of statistical estimators, these considerations can be exploited to construct confidence intervals, which are naturally ways to quantify uncertainties of pointwise estimates, such as (3). We report a precise definition below.

Definition 2.1 (**MC-FOM confidence interval**). Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ be N random independent realizations of $\boldsymbol{\mu} \sim \mathbb{P}$. Let Q be a given quantity of interest. Fix a confidence level $\gamma \in (0,1)$. The MC-FOM confidence interval of level γ, $I_{\mathrm{FOM}}^{\gamma} \subset \mathbb{R}$, is

$$I_{\mathrm{FOM}}^{\gamma} := \hat{E}_N^{\mathrm{FOM}} \pm \mathrm{t}_{\frac{1-\gamma}{2}, N-1}\sqrt{\frac{\widehat{\sigma}_0^2}{N}}, \tag{8}$$

where $\hat{E}_N^{\mathrm{FOM}}$ and $\widehat{\sigma}_0^2$ are as in Eq. (3) and (7), respectively. Here, $\mathrm{t}_{\alpha,q}$ denotes the quantile of the level $1 - \alpha$ of a t-student distribution with q degrees of freedom.

The MC-FOM confidence interval provides a better estimate of the QoI, as it enriches the estimate in (3) with a quantification of uncertainties. The confidence level, γ, is related to how *conservative* we want our estimate to be: the higher γ, the larger the interval (since we want to be *more confident* about the fact that the interval captures the actual ground truth $\mathbb{E}[Q(\mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{FOM}})]$). More precisely, the formula in (8) is constructed in such a way that

$$\mathrm{Prob}\left(\mathbb{E}[Q(\mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{FOM}})] \in I_{\mathrm{FOM}}^{\gamma}\right) \approx \gamma,$$

where $\text{Prob} = \mathbb{P} \otimes \cdots \otimes \mathbb{P}$ is the joint probability distribution of the random sample, encoding the stochasticity of the confidence interval.

For a fixed confidence level, (8) clearly shows that the uncertainty in the estimate decreases as a function of the FOM samples N. However, the decay is fairly slow, $\sim N^{-1/2}$. Consequently, a robust estimate may require a large number of FOM simulations. If the computational budget p is limited and the cost of a single simulation w_0 is high, this approach may not be feasible.

2.3. **Multi-fidelity Monte Carlo estimator.** The driving idea behind multi-fidelity Monte Carlo (MFMC) is reduce the uncertainties in the final estimate by integrating Eq. (3) with some additional information, correlated with the QoI, but cheaper to compute. Here, this is achieved by relying on the ROM. For now, let us assume that the ROM is already available and fully operational (no training required). Following our notation in Section 2.1, let w_0 be the computational time of a FOM simulation, and let w_1 be that of a ROM simulation. Let

$$\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{m_0}, \ \boldsymbol{\mu}_{m_0+1}, \ldots, \boldsymbol{\mu}_{m_1},$$

be m_1 independent realization of the input parameter $\boldsymbol{\mu} \sim \mathbb{P}$, where $m_0 < m_1$. The MFMC approach is based on the following (unbiased) estimator,

$$\hat{E}_{m_0,m_1}^{\text{MFMC}} := \hat{E}_{m_0}^{\text{FOM}} + \lambda \left(\hat{E}_{m_1}^{\text{ROM}} - \hat{E}_{m_0}^{\text{ROM}} \right), \tag{9}$$

where $\hat{E}_{m_0}^{\text{FOM}}$ is as in Equation (3), whereas $\hat{E}_{m_k}^{\text{ROM}} := m_k^{-1} \sum_{i=1}^{m_k} Q(\mathbf{u}_{\boldsymbol{\mu}_i}^{\text{ROM}})$ is the ROM counterpart of the FOM estimator. Here, $\lambda > 0$ is a suitable coupling parameter that regulates the impact of the ROM correction over the FOM estimate.

Note that, while we sampled a total of m_1 inputs, only m_0 of those were elaborated by the FOM (and the ROM). Instead, the remaining $m_1 - m_0$ were only processed by the ROM. Thus, the overall computational cost of this procedure is $m_0 w_0 + m_1 w_1$. From a UQ perspective, we also note that

$$\text{MSE}\left(\hat{E}_{m_0,m_1}^{\text{MFMC}}\right) = \frac{\sigma_0^2}{m_0} + \left(\frac{1}{m_0} - \frac{1}{m_1} \right) (\lambda^2 \sigma_1^2 - 2\lambda\rho\sigma_1\sigma_0) \tag{10}$$

where

$$\sigma_1 := \text{Var}\left(Q(\mathbf{u}_{\boldsymbol{\mu}}^{\text{ROM}})\right) \quad \text{and} \quad \rho = \frac{\text{Cov}\left(Q(\mathbf{u}^{\text{FOM}}), Q(\mathbf{u}_{\boldsymbol{\mu}}^{\text{ROM}})\right)}{\sigma_0\sigma_1}.$$

In other words, the uncertainty associated to (9) depends on $m_0, m_1, \lambda, \sigma_0, \sigma_1$ and ρ. In general, given a computational budget p, the idea of the MFMC approach is to choose m_0, m_1 and λ by minimizing the uncertainty in (10), subject to the constraints $0 \leq m_0 \leq m_1$ and $m_0 w_0 + m_1 w_1 = p$.

In practice, this results in an optimization problem, whose optimal solution, m_0^*, m_1^*, λ^*, is known in closed form (at least under suitable mild assumptions: cf. [35]). Notably, the optimal coupling parameter turns out to be $\lambda^* = \rho\sigma_0/\sigma_1$, meaning that $\lambda^* \neq 0$ whenever $\rho \neq 0$. Since the FOM and the ROM are typically correlated, this suggests favoring MFMC over MC-FOM. Indeed, if the ROM is sufficiently cheap to evaluate, it can be shown that, compared to a naive MC-FOM estimator using all computational resources for FOM simulations, i.e. with $N = \lfloor p/w_0 \rfloor$, the optimal MFMC estimator entails a lower MSE (and thus, a lower uncertainty). For the interested reader, we refer to [35, Corollary 3.5]. Here, it is

only worth mentioning that the MSE of the optimal MFMC estimator reads

$$\text{MSE}(\hat{E}_{m_0^*,m_1^*}^{\text{MFMC}}) = \frac{\sigma_0^2}{p}\left(\sqrt{w_0(1-\rho^2)} + \sqrt{w_1\rho^2}\right)^2 \tag{11}$$

This property will come in handy for our construction in Section 2.

While fascinating, this analysis has a major limitation. In fact, by taking the ROM for granted, it completely ignores the *offline* cost regarding the construction and the training of the ROM. In fact, in order to be operational, ROMs typically necessitate of a preliminary training phase, in which they learn to approximate FOM simulations. Thus, part of the computational budget must be devoted to the generation of a suitable training set.

In the literature, this fact was first acknowledged by Farcaș et al. in [14]. There, the authors noted that, if the ROM is trained on n FOM simulations:

i) the remaining budget for a MFMC routine is $p - nw_0$;
ii) the quality of the ROM, especially in terms of QoI correlations, may depend on n.

Consequently, an optimal management of the computational resources requires a careful understanding of how n enters into the equations. In [14], by leveraging on correlation bounds, the authors characterize the optimal training size $n = n_*$ as the solution to a suitable optimization problem involving the MSE of the MFMC estimator.

Here, instead of diving into details of the approach proposed by Farcaș et al., we directly present our adaptation to Deep Learning based ROMs. Compared to [14], our analysis exhibits two major differences. First of all, aside from the generation of the training data, we also include the training time $t = t(n)$ within the offline costs, acknowledging the fact that some of the computational resources must be devoted to the actual training of the neural network architectures. Secondly, we account for the fact that, in complex applications, even the sampling of the input parameters may be computationally demanding. Finally, since Deep Learning based ROMs are extremely efficient, we set $w_1 \equiv 0$: that is, we assume ROM evaluations to have a negligible computational cost.

2.4. **A deep learning enhanced multi-fidelity estimator.** We are now ready to present our deep multi-fidelity estimator (DL-MFMC), specifically tailored for non-intrusive Deep Learning based ROMs. Similarly to the MFMC approach, our objective is to obtain a reduction of the uncertainties while also providing an optimal management of the computational resources. We articulate our presentation into three steps: i) definition of the DL-MFMC estimator, ii) formulation of the optimization problem, and iii) implementation of the optimal policy and construction of confidence intervals.

2.4.1. *Notation.* As for the previous Sections, we assume to have access to a FOM, capable of producing high-quality simulations. In the case of complex multiscale systems, such as the microcirculation, oxygen transfer and radiotherapy model, a FOM simulation consists of two steps: the sampling of the parametric scenario and the consequent evaluation of the FOM solver. In general, the first step may entail a nonnegligible computational cost. For example, in the case of oxygen transfer models, the sampling step may be concerned with the computational synthesis of anatomically realistic vascular networks, encoded as a metric graph. For this reason,

we model the cost of a single FOM simulation as

$$g + w_0,$$

where g and w_0 are the *generation* and the *evaluation* costs, respectively. The first term, g, is intrinsic to the complexity of the problem, while the second one, w_0, is specific of the FOM.

Similarly to Section 2.3, we also assume that a suitable ROM technique has been chosen. For example, it may consist of a collection of multiple neural network architectures interacting in a suitable way, such as in the POD-NN [21], DL-ROM [16] and POD-MINN approaches [51]. As we already mentioned, to be operational, the ROM must be trained on a collection of FOM samples (the so-called *offline* phase). We model the corresponding *offline* cost as

$$n(g + w_0) + t(n),$$

where n is the number of FOM simulations in the training set. The first term, $n(g+w_0)$, is the sampling time, which corresponds to the generation of the training data. The second term, instead, corresponds to the actual training of the neural network architectures, here modeled through some non-decreasing monotone map $t : \mathbb{N}_0 \to \mathbb{R}^+$. Instead, after training, the computational cost of a ROM simulation is $g+w_1$ (input generation plus ROM evaluation). Here, we set $w_1 \equiv 0$ to emphasize the fact that online evaluations of Deep Learning-based ROMs can be carried out at negligible cost.

2.4.2. *Optimal training costs management.* We now reformulate the optimization problem proposed by Farcaș et al. in [14], by adapting it to our framework. Let $p \in \mathbb{R}_+$ be a computational budget. We note that if we were to train a ROM on n FOM simulations, we would be left with a computational budget $p_n = p-(g+w_0)n-t(n)$. At the same time, however, we would also have access to a trained ROM, available for multi-fidelity estimation. Specifically, we could leverage the MFMC paradigm, outlined in Section 2.3, to optimally utilize the remaining resources. In practice, for a fixed n, this procedure would bring us to the following estimator

$$\hat{E}_n^{\text{DL-MFMC}} := \hat{E}_{m_0^*(n)}^{\text{FOM}} + \lambda^*(n) \left(\hat{E}_{m_1^*(n)}^{\text{ROM}} - \hat{E}_{m_0^*(n)}^{\text{ROM}} \right), \tag{12}$$

where $m_0^*(n)$, $m_1^*(n)$, and $\lambda^*(n)$ are the optimal policy associated the computational budget p_n. More precisely,

$$\lambda^*(n) = \rho(n) \frac{\sigma_0}{\sigma_1(n)},$$

where $\rho = \rho(n)$ and $\sigma_1 = \sigma_1(n)$ are the correlation between FOM and ROM QoIs, and the variance of ROM QoIs, respectively. Here, we allow both to depend on n, so as to account for the role played by the training procedure. Similarly,

$$m_0^*(n),\ m_1^*(n) =$$

$$= \operatorname{argmin}_{m_0,m_1} \left\{ \frac{\sigma_0^2}{m_0} - \left(\frac{1}{m_0} - \frac{1}{m_1} \right) \rho^2(n)\sigma_0^2 \quad \text{s.t.} \quad \begin{array}{l} 0 \le m_0 \le m_1, \\ m_0 w_0 + m_1 g = p_n \end{array} \right\},$$

which is nothing but (10) up to substituting λ with $\lambda^*(n)$. Note that, differently from Section 2.3, the multi-fidelity sampling cost is now $m_0 w_0 + m_1 g$, rather than $m_0 w_0 + m_1 w_1$. In fact, the multi-fidelity routine entails:

   i) generating m_1 parameter instances $\mapsto$ cost: $m_1 g$;
   ii) evaluating the FOM on m_0 of such instances $\mapsto$ cost: $m_0 w_0$;
   iii) evaluating the ROM on all parameter instances, $\mapsto$ cost: negligible.

By leveraging on the theory of MFMC estimators, one can then easily prove the following.

Proposition 2.2. *Let $p > 0$ be a given budget and let $n \in \mathbb{N}_+$ be an admissible training size such that $p - (g + w_0)n - t(n) > 0$. Assume that $w_0 \rho^2(n) > g(1 - \rho^2(n))$. Then*

$$\mathrm{MSE}(\hat{E}_n^{\mathrm{DL\text{-}MFMC}}) = \frac{\sigma_0^2}{p - (g + w_0)n - t(n)} \left(\sqrt{w_0(1 - \rho(n)^2)} + \sqrt{g\rho(n)^2} \right)^2. \quad (13)$$

Proof. As we noted previously, for a fixed training size n, and a corresponding trained ROM, constructing the DL-MFMC estimator is equivalent to constructing a MFMC estimator with budget $p_n = p - (g + w_0)n - t(n)$ and cost function $m_0, m_1 \mapsto m_0 w_0 + m_1 g$. Then, (13) can be easily derived from (11) up to substituting p with p_n and w_1 with g. Here, the formula can be applied, since condition $w_0 \rho^2(n) > g(1 - \rho^2(n))$ precisely translates into the efficiency condition required in the MFMC paradigm: see, e.g., [35, Theorem 3.4]. $\square$

Since Eq. (13) only depends on n, this suggests the possibility of finding an optimal sample size, $n = n_*$, by minimizing (13) with respect to n. Indeed, this is what we are going to do. To this end, we start by characterizing the dependency of $t = t(n)$ and $\rho = \rho(n)$ on n. We do this by relying on the following assumptions.

Assumption 2.3. $\exists \zeta, c_1, c_2 > 0$ *such that* $1 - \rho^2(n) \leq c_1 n^{-\zeta} + c_2$ *for all* $n \in \mathbb{N}$, $n \geq 1$.

Assumption 2.4. $\exists c_3, c_4 > 0$ *such that* $t(n) \leq c_3 n + c_4$ *for all* $n \in \mathbb{N}$, $n \geq 1$.

Assumption 2.5. $\forall n \in \mathbb{N}_+$, *one has* $w_0 \rho^2(n) > (1 - \rho^2(n))g$.

The first assumption states that the quality of ROM QoIs increases for larger datasets. The bounding expression is similar to the one proposed in [14], but also comes with an additional term, namely $c_2 > 0$, which accounts for potential limitations inherent in the ROM. In fact, even if provided with an infinite amount of data, certain ROMs might still be incapable of replicating FOM simulations in their entirety.

The second assumption, instead, states that the training time is linear in the sample size n. In practice, deep learning-based ROMs always satisfy this assumption. In fact, the training of a neural network model $\phi : \mathcal{X} \to \mathcal{Y}$ typically involves the minimization of a loss function of the form

$$\mathscr{L}(\phi) = \sum_{i=1}^{n} \ell(\phi(x_i), y_i), \quad (14)$$

where $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ denote an abstract training set, whereas $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$ is a suitable discrepancy measure. Clearly, up to fixing a total number of training epochs, minimizing (14) requires at most $O(n)$ operations: in fact, each term in the sum can be tackled separately, even when computing gradients. Here, the constant c_4 models a fixed cost, in relation, for example, to the initialization of the ROM.

Finally, the third assumption states that most of the cost lies in the evaluation of the FOM solver, rather than in the generation of the input data. In fact, whenever $w_0 > g$, which is the typically scenario, Assumption 2.5 reduces to $\rho^2(n) \geq 1/2$.

Under these assumptions, the MSE in Eq. (13) can be bounded as follows.

Lemma 2.6. *Let Assumptions 2.3 - 2.5 hold. For all $n \in \mathbb{N}$, $n \geq 1$, the* MSE *of $\hat{E}_n^{\text{DL-MFMC}}$ can be bounded as*

$$\text{MSE}(\hat{E}_n^{\text{DL-MFMC}}) \leq \frac{2\sigma_0^2}{p - (g + w_0)n - c_3 n - c_4} \left(c_1 w_0 n^{-\zeta} + c_2 w_0 + g \right). \qquad (15)$$

Proof. By Assumptions 2.3-2.5, we have

$$\text{MSE}(\hat{E}_n^{\text{DL-MFMC}}) = \frac{\sigma_0^2}{p - (g + w_0)n - t(n)} \left(\sqrt{w_0(1 - \rho^2(n))} + \sqrt{g\rho^2(n)} \right)^2$$

$$\leq \frac{\sigma_0^2}{p - (g + w_0)n - c_3 n - c_4} \left(\sqrt{c_1 w_0 n^{-\zeta} + c_2 w_0} + \sqrt{g} \right)^2$$

$$\leq \frac{2\sigma_0^2}{p - (g + w_0)n - c_3 n - c_4} \left(c_1 w_0 n^{-\zeta} + c_2 w_0 + g \right),$$

where we exploited the fact that $\rho(n)^2 \leq 1$, and $(a + b)^2 \leq 2(a^2 + b^2)$ for all $a, b \in \mathbb{R}$. $\qquad\square$

Now, the idea is choose the training size n by minimizing the upper bound in Lemma 2.6 rather than the MSE of the estimator itself. As we shall prove in a moment, this results in a minimization problem admitting a unique minimizer n_*, consistent with the budget constraint. We formalize these considerations in Proposition 2.8, right after the auxiliary Lemma 2.7. As a side note, we point out that such semplification, although commonly adopted within the literature [14, 35, 36], may lead to suboptimal results, especially if the bounds in Assumptions 2.3 and 2.4 are not tight enough.

Lemma 2.7. *Let $A > 0$ and $\zeta > 0$. The functions*

$$\varphi_1(x) := \frac{1}{A - x}, \qquad \varphi_2(x) := \frac{x^{-\zeta}}{A - x}$$

are convex in $(0, A) \subset \mathbb{R}$. Specifically, $\varphi_1''(x) > 0$ and $\varphi_2''(x) \geq 0$ for all $x \in (0, A)$.

Proof. It is straightforward to see that

$$\varphi_1''(x) = 2(A - x)^{-3}$$

for all $x \in (0, A)$. Thus, $\varphi_1'' > 0$ in $(0, A)$. As for φ_2, a direct computation shows that

$$\varphi_2''(x) = \frac{(\zeta + 1)\zeta(A - x)^2 x^{-\zeta-2} - 2x^{-\zeta-1}\zeta(A - x) + 2x^{-\zeta}}{(A - x)^3}.$$

The numerator in the above can be rewritten and bounded as

$$x^{-\zeta-2} \left[(\zeta + 1)\zeta(A - x)^2 - 2x\zeta(A - x) + 2x^2 \right]$$

$$\geq x^{-\zeta-2} \left[\zeta^2(A - x)^2 - 2x\zeta(A - x) + 2x^2 \right]$$

$$= x^{-\zeta-2} \left[\zeta(A - x) - x \right]^2.$$

The conclusion follows. $\qquad\square$

Proposition 2.8 (Existence and uniqueness of the global minimum). *Let $p > 0$ be a given budget. Fix any $w_0, g, c_1, c_2, c_3, c_4, \zeta > 0$ and let $n_{\max} := (p-c_4)/(g+w_0+c_3)$. Then, the following function is strictly convex in $(0, n_{\max})$:*

$$F : n \mapsto \frac{2\sigma_0^2}{p - (g + w_0)n - c_3 n - c_4} \left(c_1 w_0 n^{-\zeta} + c_2 w_0 + g \right),$$

Furthermore, it admits a unique global minimum n_* within $(0, n_{\max})$.

Proof. We note that

$$F(n) = \frac{2\sigma_0^2 c_1 w_0}{g + w_0 + c_3} \cdot \frac{n^{-\zeta}}{n_{\max} - n} + \frac{2\sigma_0^2(c_2 w_0 + g)}{g + w_0 + c_3} \cdot \frac{1}{n_{\max} - n}.$$

Thus, by Lemma 2.7, F is the positive sum of two convex functions, one of which is strictly convex. Consequently, $F'' > 0$ in $(0, n_{\max})$. Since F goes to infinity at the boundaries, $F(n) \to +\infty$ for $n \to 0$ and $n \to n_{\max}$, it follows that F admits a unique global minimum within $(0, n_{\max})$. $\square$

Using Proposition 2.8, we finally propose the following optimal management policy:

1. Fix a computational budget $p > 0$;

2. Generate a preliminary collection of FOM samples n_0, by consuming a small portion of the budget, $n_0(w_0 + g) \ll p$;

3. Initialize and train the ROM for different sample sizes $n_1 < \cdots < n_k \leq n_0$, as to obtain auxiliary records about training times, $t(n_1), \ldots, t(n_k)$, and QoI correlations, $\rho(n_1), \ldots, \rho(n_k)$;

4. Estimate the coefficients $c_1, c_2, c_3, c_4, \zeta$ by leveraging on $\{n_j, t(n_j), \rho(n_j)\}_{j=1}^k$;

5. Find the optimal sample size n_* by solving the minimization problem in Proposition 2.8. Without loss of generality, we assume $n_* > n_0$;

6. Augment the training set by generating $\Delta n = n_* - n_0$ new FOM simulations. The remaining budget is now $p - (g + w_0)n_*$;

7. Train a final ROM surrogate by using all n_* FOM simulations, and compute the corresponding QoI correlation coefficient $\rho = \rho(n_*)$. The remaining budget is now $p - (g + w_0)n_* - t(n_*)$;

8. Compute $m_0^* = m_0^*(n_*)$, $m_1^* = m_1^*(n_*)$ and $\lambda_* = \lambda_*(n_*)$ according to the MFMC paradigm, that is,

$$\begin{cases} m_0^* = \dfrac{p - (g + w_0)n_* - t(n_*)}{w_0 + w_1 r}, \\[2mm] m_1^* = r m_0^*, \\[2mm] \lambda_* = \dfrac{\rho(n_*)\sigma_0}{\sigma_1}, \end{cases} \tag{16}$$

where $r^2 := (w_0 \rho^2(n_*))/(g(1 - \rho^2(n_*)))$;

9. Construct the DL-MFMC estimator

$$\hat{E}^{\text{DL-MFMC}} := \hat{E}_{n_*}^{\text{DL-MFMC}}, \tag{17}$$

by sampling the required FOM-ROM simulations, thus exhausting the computational budget.

Remark 2.9. We stress that, in practice, the values of m_0^*, m_1^* and λ_* are approximated by leveraging on empirical estimates of $\rho(n_*)$, σ_0 and $\sigma_1(n_*)$: however, in order to keep the notation lighter, we choose not to make this distinction explicit. Operationally, the idea goes as follows. First, we exploit the training data to calculate a preliminary estimate of $\rho(n_*)$, which we use to calculate the optimal sample sizes m_0^*, m_1^*. Then, following the MFMC paradigm, we compute m_0^* new independent FOM simulations. On these simulations we derive the final estimates of $\rho(n_*)$, σ_0, $\sigma_1(n_*)$, and consequently, λ_*.

Remark 2.10. Step 3 in the optimal policy pipeline concerns estimating the correlation coefficients $\rho(n_1), \ldots, \rho(n_k)$. One way to achieve this is to directly utilize the training data. For any $j \in 1, \ldots, k$, let ROM_j denote the ROM in its jth training iteration. Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{n_j}$ be the input parameters observed during training. Then

$$\frac{\sum\limits_{i=1}^{n_j}\left[Q(\mathbf{u}_{\boldsymbol{\mu}_i}^{\mathrm{FOM}})Q(\mathbf{u}_{\boldsymbol{\mu}_i}^{\mathrm{ROM}_j}) - \left(\frac{1}{n_j}\sum\limits_{b=1}^{n_j}Q(\mathbf{u}_{\boldsymbol{\mu}_b}^{\mathrm{FOM}})\right)\left(\frac{1}{n_j}\sum\limits_{b=1}^{n_j}Q(\mathbf{u}_{\boldsymbol{\mu}_b}^{\mathrm{ROM}_j})\right)\right]}{\sqrt{\sum\limits_{i=1}^{n_j}\left[Q(\mathbf{u}_{\boldsymbol{\mu}_i}^{\mathrm{FOM}}) - \left(\frac{1}{n_j}\sum\limits_{b=1}^{n_j}Q(\mathbf{u}_{\boldsymbol{\mu}_b}^{\mathrm{ROM}_j})\right)\right]^2 \sum\limits_{i=1}^{n_j}\left[Q(\mathbf{u}_{\boldsymbol{\mu}_i}^{\mathrm{ROM}_j}) - \left(\frac{1}{n_j}\sum\limits_{b=1}^{n_j}Q(\mathbf{u}_{\boldsymbol{\mu}_b}^{\mathrm{ROM}_j})\right)\right]^2}} \tag{18}$$

can be a crude approximation of $\rho(n_j)$. However, (18) would likely result in biased estimate: indeed, $Q(\mathbf{u}_{\boldsymbol{\mu}_1}^{\mathrm{ROM}_j}), \ldots, Q(\mathbf{u}_{\boldsymbol{\mu}_{n_j}}^{\mathrm{ROM}_j})$ are, in general, statistically correlated (notice, in fact, that ROM_j itself depends on the whole training set). To account for this, a better approach is to rely on a different set of data, independent of the training set. Operationally, this can be achieved by picking n_k such that $n_k < n_0$. In this way, the remaining $n_0 - n_k$ observations can be used as a common "test set", shared within the j trainings. In other words, in iteration j, the ROM is trained on $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{n_j}$, but formula (18) is evaluated on $\boldsymbol{\mu}_{n_0-n_k}, \ldots, \boldsymbol{\mu}_{n_0}$.

Concerning these preliminary training stages, we also mention that all trainings should be carried out independently by optimizing the ROM from scratch, as to favor unbiased estimates of the regression coefficients. Furthermore, the computational cost $\delta > 0$ of this procedure should be nearly negligible compared to the overall computational budget p (for instance, for our application in Section 4 we had $\delta \approx 10$ minutes and $p \geq 9$ hours). If that is not the case, then the budget p should be replaced with $\tilde{p} = p - \delta$ in all the formulas from Step 4 on.

2.4.3. *DL-MFMC estimator and uncertainty quantification.* Given a computational budget $p > 0$, the DL-MFMC estimator $\hat{E}^{\mathrm{DL\text{-}MFMC}}$ can be constructed following the optimal management policy in Section 2.4.2. By combining FOM and ROM simulations, the latter provides an efficient and robust estimate of $\mathbb{E}_{\boldsymbol{\mu}\sim\mathbb{P}}[Q(\mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{FOM}})]$, characterized by a significant reduction of the uncertainties. In practice, this fact becomes apparent when considering confidence intervals, rather than crude pointwise estimates.

Definition 2.11 (**DL-MFMC confidence interval**). Let Q be a given quantity of interest. For a given computational budget $p > 0$, let $\hat{E}^{\mathrm{DL\text{-}MFMC}}$ be the DL-MFMC estimator, computed as in Section 2. Fix a confidence level $\gamma \in (0,1)$. The DL-MFMC confidence interval of level γ is

$$I_{\mathrm{DL\text{-}MFMC}}^{\gamma} := \hat{E}^{\mathrm{DL\text{-}MFMC}} \pm z_{\frac{1-\gamma}{2}}\sqrt{\frac{\hat{\sigma}_0^2}{m_0^*(n_*)} - \left(\frac{1}{m_0^*(n_*)} - \frac{1}{m_1^*(n_*)}\right)\hat{\rho}^2(n_*)\hat{\sigma}_0^2}, \tag{19}$$

where z_γ denotes the quantile of level $1 - \gamma$ of the standard Gaussian distribution $\mathcal{N}(0, 1)$.

Similarly to the MC-FOM case, the DL-MFMC confidence interval is constructed such that

$$\text{Prob}\left(\mathbb{E}[Q(\mathbf{u}_{\boldsymbol{\mu}}^{\text{FOM}})] \in I_{\text{DL-MFMC}}^\gamma\right) \approx \gamma.$$

Notice, however, that Definition 2.11 uses normal quantiles rather than t-student ones. In fact, while the latter appear naturally when considering uncorrelated Monte Carlo samples, they do not extend to multi-fidelity estimators (which, in contrast, leverage on correlated observations). For this reason, we rely on more general quantiles, derived from the Gaussian distribution. The heuristics behind this choice lies in the Central Limit Theorem, according to which the independent sum of identically distributed random variables is asymptotically normally distributed. Since

$$\hat{E}^{\text{DL-MFMC}} = \hat{E}_{m_0^*(n_*)}^{\text{FOM}} + \lambda_*(n_*)\hat{E}_{m_1^*(n_*)}^{\text{ROM}} - \lambda_*(n_*)\hat{E}_{m_0^*(n_*)}^{\text{ROM}},$$

the DL-MFMC estimator can be seen as the sum of three random variables,

$$X = \hat{E}_{m_0^*(n_*)}^{\text{FOM}}, \quad Y = \lambda_*(n_*)\hat{E}_{m_1^*(n_*)}^{\text{ROM}}, \quad Z = -\lambda_*(n_*)\hat{E}_{m_0^*(n_*)}^{\text{ROM}},$$

each of which is asymptotically normal. Then, since the sum of normal random variables is also normal, we can approximate the distribution of $\hat{E}^{\text{DL-MFMC}}$ as

$$\mathcal{N}\Big(\mathbb{E}[X + Y + Z],\ \text{Var}(X + Y + Z)\Big) = \mathcal{N}\Big(\mathbb{E}[Q(\mathbf{u}_{\boldsymbol{\mu}}^{\text{FOM}})],\ \text{MSE}(\hat{E}^{\text{DL-MFMC}})\Big),$$

thus motivating the formula in Definition 2.11.

3. **A model for oxygen transport in microcirculation and radiotherapy.** In this Section and in the following, we present an application of our approach to a comprehensive mathematical model of microcirculation, with a specific focus on the intricate interplay between oxygen transport and its implications for radiotherapy. Initially, we introduce the biophysical model of reference along with its associated quantities of interest; then, we provide a succinct overview of the FOM and its corresponding surrogate model. We point out that these three stages modeling of the physical phenomenon, high-fidelity discretization, and model order reduction were previously undertaken by the authors and their collaborators in earlier works: the interested reader can refer to [40], [38], and [51], respectively.

Having established the groundwork, we then move to the actual application of the DL-MFMC approach, offering a comprehensive discussion in Section 4.

3.1. **Oxygen transport in microcirculation: Mechanistic model and quantities of interest.** To model oxygen transport, we rely on the model presented by Possenti et al. in [40, 38], which encompasses blood flow, hematocrit transport coupled with interstitial flow, and oxygen transport in both blood and tissue through vascular-tissue exchange.

The general model describes flow in two distinct domains: the tissue domain ($\Omega \subset \mathbb{R}^3$ with $\dim(\Omega) = 3$), where the unknowns encompass fluid pressure p_t, fluid velocity $\mathbf{u}_t$ and oxygen concentration C_t; and the vascular domain ($\Lambda \subset \mathbb{R}^3$ with $\dim(\Lambda) = 1$), representing a metric graph that describes a network of connected one-dimensional channels. In this domain, unknowns involve blood pressure p_v, blood velocity $\mathbf{u}_v$, and vascular oxygen concentration C_v. The model for oxygen transport uses velocity fields $\mathbf{u}_v$ and $\mathbf{u}_t$ to describe blood flow in the vascular network and

plasma flow in tissue. The governing equations for the oxygen transfer model are as follows:

$$
\begin{cases}
\nabla \cdot (-D_t \nabla C_t + \mathbf{u}_t\ C_t) + V_{max}\ \dfrac{C_t}{C_t + \alpha_t\ p_{m_{50}}} = \phi_{O_2}\, \delta_\Lambda & \text{on } \Omega \\[4mm]
\pi R^2 \dfrac{\partial}{\partial s}\left(-D_v \dfrac{\partial C_v}{\partial s} + v_v\ C_v + v_v\ k_1\ H\ \dfrac{C_v^\gamma}{C_v^\gamma + k_2} \right) = -\phi_{O_2} & \text{on } \Lambda \\[4mm]
\phi_{O_2} = 2\pi R\ P_{O_2}(C_v - C_t) + (1 - \sigma_{O_2})\ \left(\dfrac{C_v + C_t}{2} \right)\ \phi_v & \text{on } \Lambda \\[4mm]
\phi_v = 2\pi R L_p \big((p_v - p_t) - \sigma(\pi_v - \pi_t) \big) & \\[3mm]
C_v = C_{in} & \text{on } \partial\Lambda_{\text{in}} \\[3mm]
-D_v \dfrac{\partial C_v}{\partial s} = 0 & \text{on } \partial\Lambda_{\text{out}} \\[3mm]
-D_t \nabla C_t \cdot \mathbf{n} = \tau_{O_2}\ (C_t - C_{0,t}) & \text{on } \partial\Omega.
\end{cases}
\tag{20}
$$

Specifically, the first equation governs the oxygen distribution within the tissue, the second outlines how oxygen is transported through the bloodstream, and the third defines the transfer of oxygen between the two domains, Ω and Λ. In particular, the flux ϕ_{O_2} is derived under the assumption that the vascular wall acts as a semi-permeable membrane. Complementing this model is a set of boundary conditions detailed in the final three equations: at the vascular inlets $\partial\Lambda_{\text{in}}$, we prescribe the oxygen concentration; at the vascular endpoints $\partial\Lambda_{\text{out}}$, null diffusive flux is enforced; and for the boundary of the tissue domain $\partial\Omega$, we simulate the presence of an adjacent tissue domain with boundary conductivity τ_{O_2} and a concentration in the far field $C_{0,t}$.

The symbols $D_t, D_v, V_{max}, \alpha_t, p_{m_{50}}, k_1, k_2, C_v^\gamma, P_{O_2}, \sigma_{O_2}, L_p, \sigma, \pi_v, \pi_t$ represent constants independent of the model solution. For a complete explanation of the physical significance of these variables, refer, for example, to [38].

Our main interest is to perform a reliable analysis of certain quantities of interest, relevant for radiotherapy applications, when the topology of the vascular network Λ, and the values of the physical parameters V_{max}, C_{in}, P_{O_2} are uncertain. The choice of these parameters is motivated by the sensitivity analysis study recently performed in [50]. As we shall see in a moment, these quantities of interest can be expressed as certain functionals of the tissue oxygenation map C_t, herein measured in mL_{O_2}/mL_B. We refer to Figure 1 for a visual example. In this sense,

$$
u_{\boldsymbol{\mu}} = C_t \quad \text{and} \quad \boldsymbol{\mu} = [\Lambda, V_{max}, C_{in}, P_{O_2}],
$$

according to our notation in Section 2. Here, we shall focus on three quantities of interest: average partial pressure, partial pressure variability, and tumor control probability. We detail them below.

Average oxygen partial pressure $\overline{pO_2}$. To start, we consider the spatial average of the oxygen partial pressure, $\overline{pO_2}$, computed as

$$
\overline{pO_2} = \frac{1}{\|\Omega\|} \int_\Omega \frac{C_t(\mathbf{x})}{\alpha_t},
\tag{21}
$$

where α_t is the oxygen solubility in the tissue. We note that $\overline{pO_2} = Q(C_t)$, where $Q : L^2(\Omega) \to \mathbb{R}$ is linear and continuous with respect to the L^2-norm.
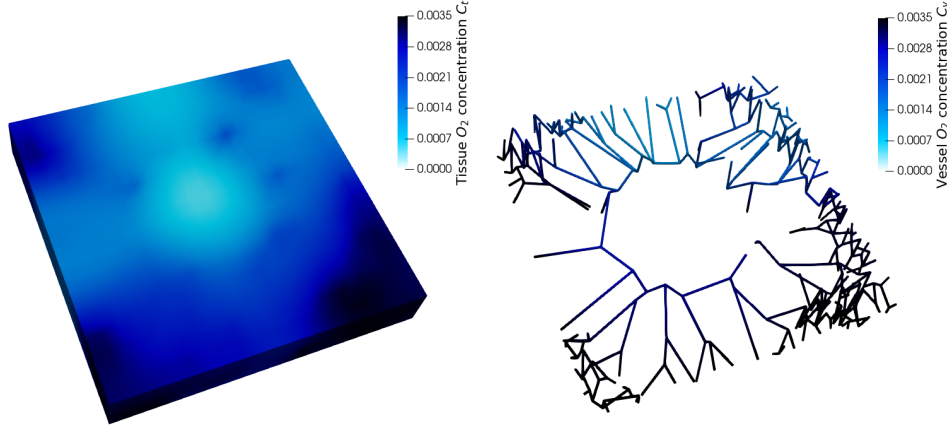
FIGURE 1. The tissue oxygenation map, measured in mL_O2/mL_B, is visually depicted on the left panel through the FOM solution (light blue corresponds to low oxygen). On the right panel, we showc the 1D embedded vascular microstructure, which visibly impacts the oxygen map. Furthermore, oxygen concentration in the blood, indicated as C_v, is also reported.

Oxygen partial pressure range ΔpO_2. Since the average partial pressure provides a global perspective on tissue perfusion, it is also interesting to consider different QoIs which are more sensitive to local fluctuations. To this end, we introduce a further QoI, ΔpO_2, measuring the difference between the maximum and the minimum partial pressure. We call the latter *oxygen partial pressure range.* Operationally, this is computed as

$$\Delta pO_2 = \frac{\max_{\mathbf{x}\in\Omega} C_t(\mathbf{x}) - \min_{\mathbf{x}\in\Omega} C_t(\mathbf{x})}{\alpha_t}. \tag{22}$$

From a clinical point of view, ΔpO_2 quantifies the abundance of local hypoxic effects in the tissue. We note that, unlike the average partial pressure, this QoI is non-smooth and not even defined on $L^2(\Omega)$ as a whole. However, it can be regarded as continuous functional with respect to the L^∞-norm, that is: $\Delta pO_2 = Q(C_t)$ with $Q : L^\infty(\Omega) \subset L^2(\Omega) \to \mathbb{R}$.

Tumor Control Probability (TCP). As a final example, we consider a radiotherapy-related QoI regarding the probability of tumor eradication (Tumor Control Probability, TCP for short). To this end, we model radiotherapy treatment using the linear-quadratic (LQ) model, which is the most widely used radiological model. The latter is based on two different parameters that describe the radiosensitivity of cells or tissue. The first parameter α describes the lethal damage resulting from a single hit, while β is related to multiple hits, namely the interaction of multiple radiation tracks [27]. Combining these two parameters and the dose administered (D), we model the surviving fraction S_f as a spatially dependent map, adopting the model proposed by Tinganelli et al. in [49] that accounts for the effect of oxygen

on radiotherapy:

$$S_f(D, pO_2, \mathbf{x}) = \exp\left(-\alpha\frac{D}{\mathrm{OER}(\mathbf{x}, pO_2,\ LET)} - \beta\left(\frac{D}{\mathrm{OER}(\mathbf{x}, pO_2,\ LET)}\right)^2\right),$$
(23)

where $pO_2 = C_t/\alpha_t$ is the oxygen partial pressure, whereas the Oxygen Enhancement Ratio (OER) is a suitable transformation operating a change of scale:

$$\mathrm{OER}(0, LET)\ =\ \frac{LET^\delta + M\ a}{a + LET^\delta},$$

$$\mathrm{OER}(\mathbf{x}, pO_2, LET)\ =\ \frac{b\ \mathrm{OER}(0,\ LET) + pO_2(\mathbf{x})}{b + pO_2(\mathbf{x})},$$

$$D_{\mathrm{OER}}(\mathbf{x}, pO_2, LET)\ =\ \frac{D}{\mathrm{OER}(\mathbf{x}, pO_2,\ LET)}.$$

Here, M, a, δ, $b > 0$ are the model parameters fitted to the experimental data, LET is the linear energy transfer of ionizing radiation, and $D_{\mathrm{OER}}(\mathbf{x}, pO_2, LET)$ is the dose corrected for the oxygen effect, to be included in the LQ model. The admissible values of S_f range from 0 to 1, representing the fraction of cells that survived treatment with the specified dose D, assumed as a constant.

Based on that, we define the Tumor Control Probability (TCP), which describes the probability of successful treatment. It is based on the number of cells that survive treatment S_f and the distribution of the initial number of clonogenic cells $N = N(\mathbf{x})$. For a fixed radiation dose D, and under the assumption of unicellular independence, following the approach proposed by Strigari et al. [47], the TCP can be expressed as

$$\mathrm{TCP} = \exp\left(-\int_\Omega N(\mathbf{x})S_f(D, C_t/\alpha_t, \mathbf{x})d\mathbf{x}\right). \tag{24}$$

Although the definition of the TCP is far more involved compared to the one of the average partial pressure, the TCP can be realized as a Lipschitz continuous nonlinear functional of the oxygen concentration C_t with respect to the L^2-norm. More precisely, $\mathrm{TCP} = Q(C_t)$ for some smooth operator

$$Q : L^2_+(\Omega) \to \mathbb{R},$$

where $L^2_+(\Omega) := \{u \in L^2(\Omega) \text{ s.t. } u \geq 0\}$. The interested reader can find a formal derivation of this fact in Remark 3.1.

Remark 3.1. We first notice that, for every L-Lipschitz continuous real-valued map $f : [0, +\infty) \to [0, +\infty)$, the composition operator $T_f : g \mapsto f \circ g$ is L-Lipschitz from $L^2_+(\Omega)$ onto itself. In fact,

$$\|f \circ g_1 - f \circ g_2\|_{L^2(\Omega)}$$

$$= \sqrt{\int_\Omega |f(g_1(\mathbf{x})) - f(g_2(\mathbf{x}))|^2 d\mathbf{x}} \leq$$

$$\leq \sqrt{\int_\Omega L^2|g_1(\mathbf{x}) - g_2(\mathbf{x})|^2 d\mathbf{x}} = L\|g_1 - g_2\|_{L^2(\Omega)},$$

for all $g_1, g_2 \in L^2_+(\Omega)$. From here, up to fixing $M, a, \delta, b, D > 0$ and $N = N(\mathbf{x}) \in L^\infty(\Omega)$, the Lipschitz continuity of the TCP is easily proven, as it can be derived by composition. In fact, we notice that the TCP is computed as

$$Q(C_t) = f_3 \circ \mathscr{N} \circ T_{f_2} \circ T_{f_1}(C_t/\alpha_t), \tag{25}$$

where

$$f_1 : z \mapsto D(b + z)/(b\mathrm{OER}(0, LET) + z), \qquad f_2 : z \mapsto \exp\left(-\alpha z - \beta z^2\right),$$

$$\mathscr{N} : g \mapsto \int_\Omega Ng, \qquad f_3 : z \mapsto \exp(-z).$$

Due boundness of their derivatives over $[0, +\infty)$, the maps f_1, f_2, f_3 are clearly Lipschitz continuous: in turn, this implies the smoothness of T_{f_1} and T_{f_2}, as argued previously. Finally, since the operators $\mathscr{N}$ and $C_t \mapsto C_t/\alpha_t$ are both linear and continuous, Eq. (25) immediately implies the Lipschitz continuity of Q.

3.2. **Description of the FOM: A high-fidelity model leveraging on the finite element method.** As a high-fidelity model, we consider a FOM based off a Finite Element discretization of problem (20). Following the approach outlined in [40], we achieve this by discretizing both the interstitial domain Ω, representing a 3D slab with dimensions of $1, mm$ in edge length and $0.15, mm$ in thickness, and the embedded 1D vascular network Λ.

For the interstitial domain, we utilize a structured mesh composed of tetrahedral elements arranged in a $20 \times 20 \times 3$ grid, resulting in $N_h = 1764$ degrees of freedom. Over this mesh, we define a corresponding space of piecewise linear continuous Lagrangian finite elements, denoted as $V_{t,h} = X_h^1(\Omega)$. To discretize the vascular network, instead, we subdivide each vascular branch Λ_i into multiple linear segments. Subsequently, we associate each branch with a corresponding finite element space $V_{v,h}^i = X_h^1(\Lambda_i)$, comprising piecewise linear continuous Lagrangian elements. With this setup, we instruct the FOM solver to produce a numerical solution to (20) as

$$(C_t, C_v) \in V_{t,h} \times V_{v,h},$$

where $V_{v,h} = \left(\bigcup_{i=1}^{N_b} V_{v,h}^i\right) \bigcap \mathcal{C}^0(\Lambda)$.

Notice, however, that in order to assemble (and solve) the oxygen transfer model, one first needs to solve the underlying fluid flow problem, which takes place in the vascular microenvironment [40, 38]. Figure 2 illustrates the sequential calculation of velocity, pressure, and discharge hematocrit in both tissue and vascular networks using the finite element method. We refer to [38] for more details on the derivation and validation of the model.

As we already mentioned, our main interest is to characterize the behavior of certain quantities of interest, related to C_t, under suitable uncertainties in the model parameters. In particular, following an earlier sensitivity analysis performed by Vitullo et al. [50], we focus our attention on the role played by $V_{max}, C_{v,in}, P_{O_2}$ and Λ.

In order to explore different scenarios, we sample the model parameters as follows. For the physical parameters —specifically $V_{max}, C_{v,in}$, and P_{O_2}— we randomly select values within the physiological range of variation in a uniform manner (see Table 1). For the vascular network, instead, we utilize a random generator implementing a biomimetic algorithm that emulates the natural process of new blood
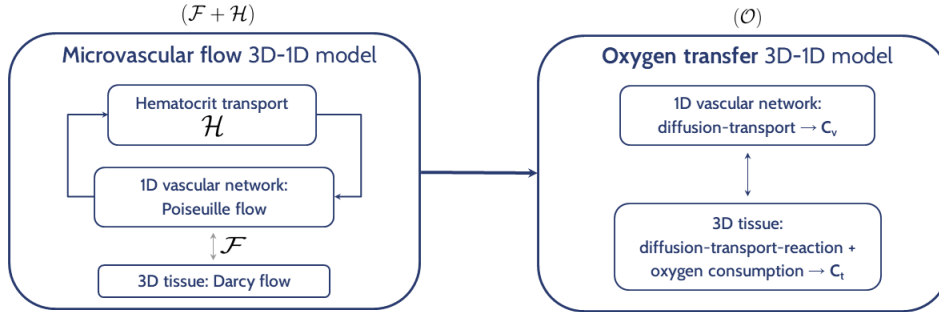
FIGURE 2. General layout of the full order model for the whole vascular microenvironment.

vessel formation. The output of this algorithm is driven by the value of two geometrical parameters: the vascular surface area per unit volume of tissue, denoted as S/V, and the percentage of seeds for angiogenesis. Essentially, the former represents the overall density of the vascular network, affecting the number and spacing of blood vessels within the tissue, while the latter governs the distribution of initial point seeds, serving as starting points for the algorithm to initiate new blood vessel growth. Together, these two parameters allow the generation of complex vascular networks, with different configurations of the blood vessels and extravascular regions of varying dimensions: see, e.g., Figure 3. Similarly to the physical parameters, the values of the geometrical parameters are sampled randomly within the corresponding range of variation.

Symbol	Parameter	Unit	Range of variation
P_{O_2}	O_2 wall permeability	m/s	$0.35 \cdot 10^{-4} - 3.00 \cdot 10^{-4}$
V_{max}	O_2 consumption rate	$\frac{mL_{O_2}}{cm^3 \cdot s}$	$0.40 \cdot 10^{-4} - 2.40 \cdot 10^{-4}$
$C_{v,in}$	O_2 concentration at the inlets	$\frac{mL_{O_2}}{mL_B}$	$2.25 \cdot 10^{-3} - 3.75 \cdot 10^{-3}$
$\% \frac{SEEDS_{(-)}}{SEEDS_{(+)}}$	Seeds for angiogenesis	$\%$	$0 - 75$
S/V	Vascular surface per unit volume	m^{-1}	$5 \cdot 10^3 - 7 \cdot 10^3$

TABLE 1. In the first three rows, the biophysical parameters of the ROM are presented along with their respective ranges of variation. The last two rows outline the hyper-parameters utilized to initialize the algorithm responsible for generating the vascular network.

3.3. **Description of the ROM: A nonintrusive surrogate model based on the POD-MINN+ approach.** To construct the ROM, we exploit the POD-MINN+ approach, a nonintrusive model order reduction technique recently presented in [51], which the authors specifically designed to address for parametrized problems with embedded microstructure. The goal of this strategy is to derive a parameter-to-solution map using physical and geometric inputs and to generate
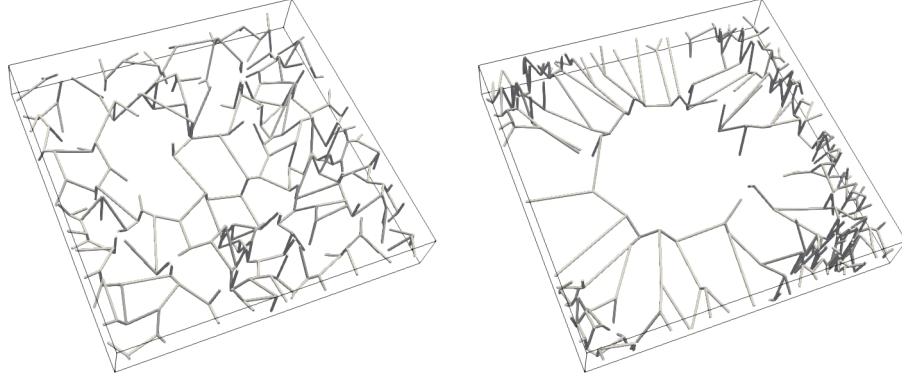
(A) Low vascularization, homogeneous distribution

(B) Low vascularization, high extravascular distance
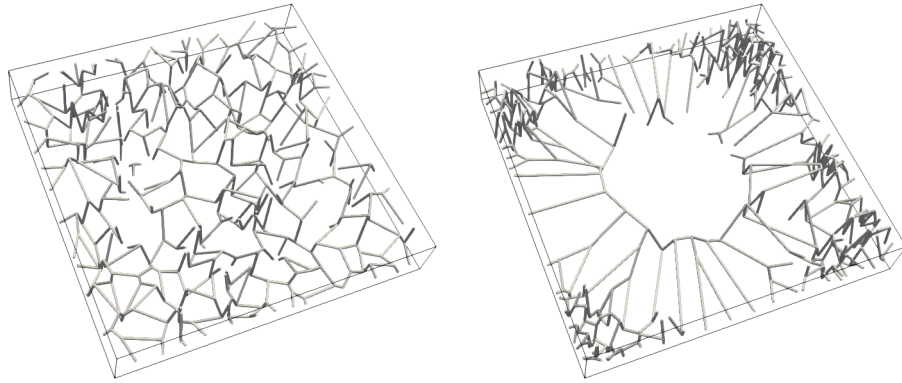
(C) High vascularization, homogeneous distribution

(D) High vascularization, high extravascular distance

FIGURE 3. Examples of architectures with maximum extravascular distance increasing from left to right and higher vascular density from top to bottom.

a reconstruction of the high-fidelity oxygen concentration map in the tissue from which we calcuate the QoIs.

Simply put, the POD-MINN+ technique consists of a projection-based regression model, incorporating neural networks and Proper Orthogonal Decomposition (POD), as in [21], combined with a closure model that captures high frequency components by introducing suitable local corrections. For this purpose, the POD-MINN+ approach leverages on a specific class of neural network architectures,

termed Mesh-Informed Neural Networks (MINNs) [16]. A sketch of the methodology is shown in Figure 4. In formulas, the whole idea can be synthesized as

$$\mathbf{u}_{\boldsymbol{\mu}}^{\mathrm{ROM}} := \mathbb{V}\mathcal{M}_{rb}(\boldsymbol{\mu}) + \mathcal{M}_c(\Lambda), \tag{26}$$

where $\mathbb{V} \in \mathbb{R}^{N_h \times k}$ is the POD basis, while $\mathcal{M}_{rb}$ and $\mathcal{M}_c$ are suitable neural network architectures. The first one maps the model parameters, $\boldsymbol{\mu}$, onto the corresponding POD coefficients, while the second one introduces a local correction that only depends on the vascular network Λ.

In general, both $\mathcal{M}_{rb}$ and $\mathcal{M}_c$ need to process information about the vascular graph. To this end, we parametrize Λ in terms of the extravascular distance $\mathbf{d} \in \mathbb{R}^{N_h}$ and the inlet characteristic function $\boldsymbol{\eta} \in \mathbb{R}^{N_h}$: the former maps each point in the tissue domain to its distance from the nearest point in the vascular network; the latter, instead, assigns a unit value to the nodes located near the inlets within the computational mesh of the tissue domain.

We avoid considering a large number of POD modes for the low-fidelity approximation of the solution manifold in the framework of complex microstructures, where the Kolmogorov n-width can be slow decaying. In this case, we exploit the information contained in the degrees of freedom corresponding to the extravascular distance $\mathbf{d}$ and the inlet characteristic function $\boldsymbol{\eta}$.

Coherently with our presentation in Section 2, the POD-MINN+ approach is a model order reduction technique that leverages on a supervised learning strategy. In particular, it necessitates of some training data. The idea goes as follows. First, we exploit the FOM solver to construct a suitable training set $\{(\boldsymbol{\mu}_i, \mathbf{u}_i^{\mathrm{FOM}})\}_{i=1}^n$. As a second step, we collect all high-fidelity simulations in a matrix of snapshots and perform a Singular Value Decomposition in order to capture the most relevant modes: this results in the construction of the POD basis $\mathbb{V} \in \mathbb{R}^{N_h \times k}$, with $k \ll N_h$. In general, although problems featuring complex microstructures have a slow-decay in the Kolmogorov n-width [51], we avoid considering a large number of POD modes, as the higher frequencies will be incorporated in the closure term.

Once the POD matrix has been constructed, we use it to project the solution manifold onto a linear trial subspace of dimension $k \ll N_h$: thanks to this maneuver, we can move our attention from FOM solutions to POD coefficients. At this point, we assemble a neural network unit responsible for the approximation of the parameter-to-POD-coefficients map. Following [51], we construct the latter using two MINNs, $\mathcal{M}_{rb,\eta}$ and $\mathcal{M}_{rb,d}$, and a deep feed forward neural network, $\mathcal{N}_{rb,ph}$. The first two account account for embedded 1D structure, while the last one models the effects of the physical parameters $\boldsymbol{\mu}_{ph} = [V_{max}, C_{v,in}, P_{O_2}]$. The three components act as

$$\mathcal{M}_{rb}(\boldsymbol{\mu}) = \Big(\mathcal{M}_{rb,\eta}(\boldsymbol{\eta}) \odot \mathcal{M}_{rb,d}(\mathbf{d})\Big)\mathcal{N}_{rb,ph}(\boldsymbol{\mu}_{ph}), \tag{27}$$

where we recall that $\boldsymbol{\mu} = [\Lambda, V_{max}, C_{v,in}, P_{O_2}] = [\mathbf{d}, \boldsymbol{\eta}, \boldsymbol{\mu}_{ph}]$. Here $\odot$ is the Hadamard product, modeling a suitable interaction between the inlets function $\boldsymbol{\eta}$ and the extravascular distance $\mathbf{d}$. Operationally, we design the three neural network architectures, in terms of width, depth, and nonlinearities, following [51]. In particular, we rely on mesh-informed layers and hyperbolic tangent activations: for further details, we refer to [51].
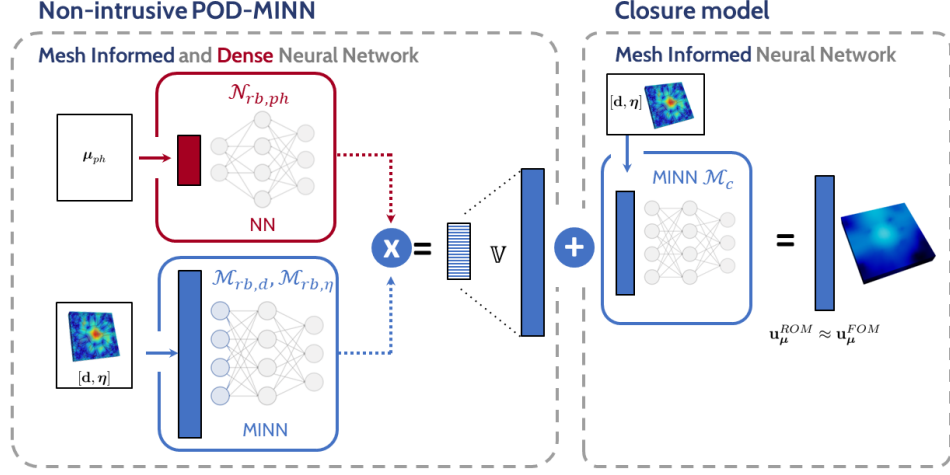
FIGURE 4. A sketch of the POD-MINN+ method. The macroscale parameters and the microscale ones are fed to two separate architectures, whose outputs are later combined to approximate the POD coefficients. The coefficients are then expanded over the POD basis, $\mathbb{V}$, and the ROM solution is further corrected with a closure term computed by a third network that accounts for the local features related to the high frequencies.

We train the parameter-to-POD-coefficient network, $\mathcal{M}_{rb}$, by minimizing the loss function

$$\mathcal{E}(\mathcal{M}_{rb}) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} \|\mathbb{V}^T \mathbf{u}_i^{\text{FOM}} - \mathcal{M}_{rb}(\boldsymbol{\mu}_i)\|.$$

To this end, we rely on the L-BFGS optimizer (default learning rate, no batching), combined with an ensamble learning strategy, i.e. by initializing and optimizing the network multiple times.

The final step of the POD-MINN+ technique concerns the training of the closure model. Similarly to the case of POD coefficients, we process the information concerning the vascular graph using two distinct MINNs, $\mathcal{M}_{c,d}$ and $\mathcal{M}_{c,\eta}$, so that we can effectively isolate the individual impact of $\mathbf{d}$ and $\boldsymbol{\eta}$. Mathematically speaking, we let

$$\mathcal{M}_c(\mathbf{d}, \boldsymbol{\eta}) = \mathcal{M}_{c,\eta}(\boldsymbol{\eta}) \odot \mathcal{M}_{c,d}(\mathbf{d}). \tag{28}$$

Both architectures consist of a suitable combination of mesh-informed and dense layers: we refer to [51] for further details about the design of $\mathcal{M}_{c,d}$ and $\mathcal{M}_{c,\eta}$.

In order to train the closure model, the following loss function is minimized:

$$\mathcal{E}(\mathcal{M}_c) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} \left[\xi \|\mathbf{u}_i^{\text{FOM}} - \mathbb{V}\mathcal{M}_{rb}(\boldsymbol{\mu}_i) - \mathcal{M}_c(\mathbf{d}, \boldsymbol{\eta})\|_{2, N_h} \right.$$

$$\left. + (1 - \xi) \|\mathbf{u}_i^{\text{FOM}} - \mathbb{V}\mathcal{M}_{rb}(\boldsymbol{\mu}_i) - \mathcal{M}_c(\mathbf{d}, \boldsymbol{\eta})\|_{\infty, N_h} \right]$$

Here, only the weights and biases associated with $\mathcal{M}_c$ undergo the optimization process, while those of $\mathcal{M}_{rb}$ are frozen. Notice also the hyperparameter ξ, which

controls the trade-off between 2-norm and ∞-norm regularization. As before, we rely on the L-BFGS optimizer, combined with an ensamble learning strategy, for the minimization of the loss function.

4. **Application of the DL-MFMC method to oxygen transport and radiotherapy.**

4.1. **Computational setup for numerical simulations.** We implemented the FOM solver using a C++ based *in-house* code [40], developed using the GetFem++ library [44]. On our workstation, consisting of an AMD EPYC 7301 16-Core Processor with 2 sockets and 16 cores, each FOM simulation required roughly $w_0 = 25$ minutes, while the generation of a single vascular network took approximately $g = 1$ second. For the implementation of the ROM, instead, we rely on a custom Python library combining Pytorch and FEniCS. In this case, we conducted all trainings and evaluations on a Tesla V100-PCIE-32GB GPU accelerator. Each online evaluation of the Deep Learning-based ROM requires approximately $w_1 = 0.01s$ with the available computational resources. This shows consistency with respect to the assumption to set $w_1 \equiv 0$ in the optimal policy presented in the previous section.

TABLE 2. Prescribed values of input parameters in the comprehensive computational model for the high-fidelity approximation of the solution through the finite element method.

Symbol	Parameter	Unit	Value	Ref. #
L	characteristic length	m	$1 \cdot 10^{-3}$	–
R	average radius	m	$4 \cdot 10^{-6}$	[40]
K	tissue hydraulic conductivity	m^2	$1 \cdot 10^{-18}$	[24, 40]
μ_t	interstitial fluid viscosity	cP	1.2	[48]
μ_v	blood viscosity	cP	3.0	[42]
L_p	wall hydraulic conductivity	$m^2 \, \mathrm{s} \, \mathrm{kg}^{-1}$	$1 \cdot 10^{-12}$	[40]
$\delta\pi$	oncotic pressure gradient	$mmHg$	25	[40]
σ	reflection coefficient	–	0.95	[25]
D_v	vascular diffusion coefficient	m^2/s	$2.18 \cdot 10^{-9}$	[26]
$N \cdot MCHC$	max. hemoglobin-bound O_2	$\frac{mL_{O_2}}{mL_{RBC}}$	0.46	[45]
γ	Hill constant	–	2.64	[53, 26]
$p_{s_{50}}$	O_2 at half-saturation	$mmHg$	27	[53, 28]
α_t	O_2 solubility coefficient	$\frac{mL_{O_2}/mL}{mmHg}$	$3.89 \cdot 10^{-5}$	[45]
D_t	tissue diffusion coefficient	m^2/s	$2.41 \cdot 10^{-9}$	[26]
C	Characteristic O_2 concentration	$\frac{mL_{O_2}}{mL_B}$	$1.50 \cdot 10^{-3}$	–

The POD-MINN+ approach has been implemented projecting the discrete solution manifold over a trial linear subspace consisting of $k = 10$ POD basis functions,

as, given the diffusive nature of the problem, those were sufficient to capture the main global features of the PDE solutions. As discussed in Section 3.3, the training phase consisted of two distinct steps: (i) the approximation of the POD reduced coefficients, and (ii) the correction with the closure model. During first phase, the neural networks were trained for at most 50 epochs, before undergoing a lifting with respect to the original finite element space of dimension $N_h = 1764$. Then, in the second and final step, we trained the closure model for a total of 10 epochs. The rationale behind this decision revolves around balancing the computational resources needed to construct the estimator with the intricate complexity of the neural network architecture defining the closure model. In fact, the considerable number of degrees of freedom in the closure map $\mathcal{M}_c$, stemming from the high-dimensionality of the input and output data, can pose significant challenges. Hence, it becomes essential to impose appropriate constraints on the training time to prevent over-fitting phenomena, and, consequently, a substantial degradation of the ROM as a whole. For what concerns the loss function, we set the regularization parameter to $\xi = 0.75$, as that provides an acceptable fit. Refer to [51] for more details on the empirical tests supporting this choice.

We conclude with a final remark on the definition of the radiotherapy-related QoI that we considered for this study, that is, the Tumor Control Probability (TPC). As discussed in Section 3.1, we computed the TCP by leveraging on the tissue oxygen concentration map C_t and the linear-quadratic model. Here, we initialized the latter using the values in Table 3, where the ratio α/β has been prescribed for a tumor tissue scenario. Finally, we set the radiation dose to $D = 20 Gy$.

Symbol	Parameter	Unit	Value	Ref. #
D	radiation dose	Gy	20	–
α	radiosensitivity parameter for 'single' hit	Gy^{-1}	0.178	[23]
β	radiosensitivity parameter for 'multiple' hits	Gy^{-2}	0.0455	[23]
δ	TIN parameter	-	1.38	[41]
M	TIN parameter	-	2.81	[41]
a	TIN parameter	$keV/\mu m$	522.45	[41]
b	TIN parameter	$mmHg$	1.24	[41]
N_c	Clonogenic cells in the interstitial volume	-	10^8	[12]
LET	Linear Energy Transfer in photons	$keV/\mu m$	2	[13]

TABLE 3. Input parameters values assigned in the linear-quadratic model to compute the TCP QoI.

4.2. **Optimal management of the computational budget.** As we detailed in Section 2.4.2, the implementation of the DL-MFMC estimator entails a preliminary analysis (steps 2 - 5), necessary for the estimation of the optimal sample size n_*, and the optimal sampling policy m_0^*, m_1^*, λ_*. In particular, one of the first steps concerns the estimation of the trend coefficients $\zeta, c_1, c_2, c_3, c_4$, modeling the behavior of training times $t = t(n)$ and QoI correlations for varying sample size $\rho = \rho(n)$, cf.

Assumptions 2.4-2.3. Of note, this analysis can be conducted on a small pool of FOM simulations, regardless of the computational budget p.

Here, we performed this preliminary analysis on a restricted pool of $n_0 = 300$ FOM simulations, recording the behavior of t and ρ for varying sample sizes, $40 \leq n_1 < \ldots n_6 \leq 200 < n_0$, with the n_j's forming a uniform partition of $[40, 200]$. For each n_j we trained the ROM multiple times as to obtain additional records for $t(n_j)$ and $\rho(n_j)$, with correlations being estimated over a test set of size $n_0 - n_j$; then, we estimated the trend coefficients by fitting a suitable regression model. Results are in Figure 5 and 6. Note that, although the regression curves do not bound all the data points in a strict sense, they do bound nearly all the observations referring to successful trainings (i.e., for each n_k, those reporting the lowest $1 - \rho^2(n_k)$). In this sense, we consider our estimates to be satisfactory.
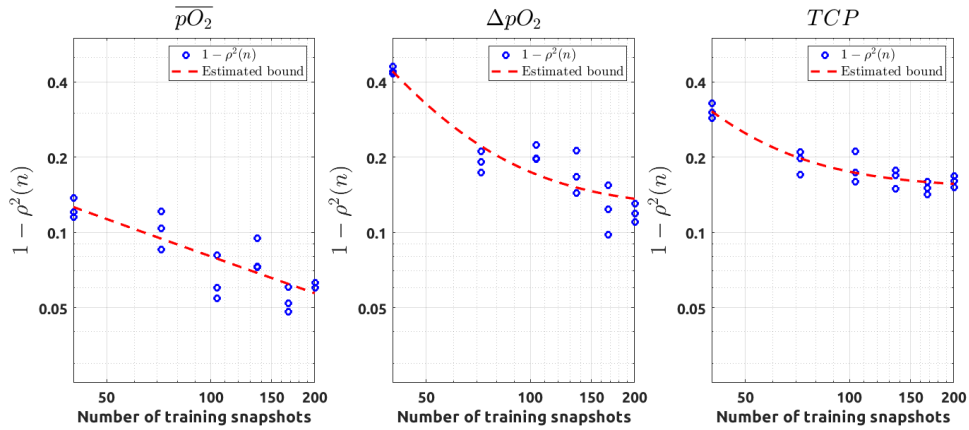
FIGURE 5. Regression models for the law $1 - \rho^2(n_j) \leq c_1 n_j^{-\zeta} + c_2$ for each QoI, $\overline{pO_2}$, ΔpO_2, TCP, varying the sample sizes as $j = 1, ..., k$.

Among the three QoIs, the strongest correlation between FOM and ROM is observed for the average oxygen partial pressure, $\overline{pO_2}$. In general, this is not very surprising, in fact, deep learning-based ROMs are known to yield better approximations when it comes to smooth operators, see, e.g., [15, 17].

On the other hand, the performances of the ROM are considerably more limited when considering less smooth operators, such as ΔpO_2, or operators that result in nearly unimodal distributions, as the TCP. Further evidence of this claim lies in the estimates obtained for c_2, i.e., in the coefficient modeling the intrinsic correlation gap. In fact, we obtained higher values for ΔpO_2 and TCP, respectively 0.126 and 0.151, as opposed to the smaller estimate associated with $\overline{pO_2}$, where $c_2 = 0.003$. In particular, this goes to show that the ROM would perform better on smoother operators even when provided with an infinite amount of data. Another interesting aspect concerns the estimated decay rate of $1 - \rho(n)^2$, namely ζ. In fact, although the functions settle to a higher limit value, the correlation between ROM and FOM for the variability of the partial pressure spatial and the TCP increases faster ($\zeta \sim -2$) than the corresponding map related to the spatial average ($\zeta \sim -1/2$).

Let us now discuss the results for the linear upper bound concerning the training time. As shown in in Figure 6, the initialization of the ROM constitutes the

most relevant source of computational cost. In fact, the actual computational time required for the optimization of the loss function is barely affected by the number of training samples n. The main reason for this lies in our design choice of limiting the number of training epochs below 50. Running the optimization process for a larger number of iterations would make the curve in Figure 6 become steeper, with c_3 gaining dominance over c_4.
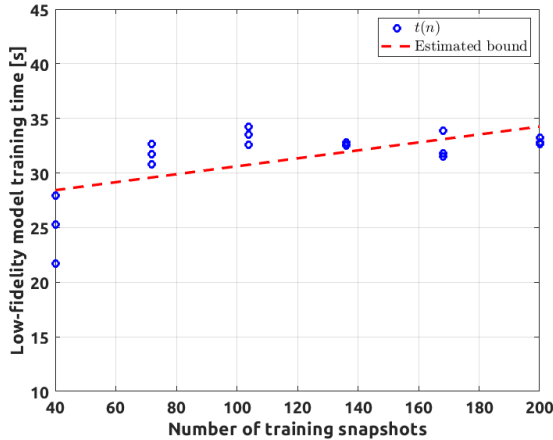

FIGURE 6. Regression model for the law describing the ROM training time $t(n_j) \leq c_3 n_j + c_4$, varying the sample sizes as $j = 1, ..., k$.

After estimating the regression coefficients $\zeta, c_1, c_2, c_3, c_4$, we focused on the retrieval of the optimal sample size n_*, obtained by minimizing the upper bound in Lemma 2.6. The results of this step, however, depend on the overall computational budget $p > 0$, as that appears explicitly in the objective function to be minimized. For this reason, we postpone the discussion right below, together with the UQ analysis.

4.3. **Results of the uncertainty quantification analysis.** For the actual UQ analysis, we consider a variable computational budget of $p \in \{9, 12, 15, 18, 21, 24\}$ hours. As a first step, we leverage on the previous results to estimate the optimal sample size n^*, which we use to train the final ROM surrogate through the aforementioned POD-MINN+ approach. Subsequently, following steps 7-8 in the optimal policy pipeline, we estimate the number of new high-fidelity and low-fidelity evaluations, m_0^* and m_1^*, required for acceleration of QoI statistics and their variance reduction. Table 4 shows a synthetic overview of such analysis, reporting the results for three budgets of reference, namely 12, 18 and 24 hours.

First of all, we note that the computational speed-up achieved through model order reduction and the short training time are effectively utilized: in fact, there is a difference of $1 - 2$ orders of magnitude between the number of reduced order model (ROM) and full order model (FOM) simulations. Additionally, we observe a significant distinction in the outcomes when comparing the three QoIs. Once again, the greater smoothness of the spatial-average operator, $\overline{pO_2}$, results in a increased reliance on the high-fidelity solver. The described trend is observed also when comparing the different allocation of FOM simulations between the training

QoI	Budget p	n^*	m_0^*	m_1^*	$\%\frac{n^*}{n^*+m_0^*}$
$\overline{pO_2}$	$12\,h$	136	232	8311	36.96%
	$18\,h$	199	342	13522	36.78%
	$24\,h$	259	451	19082	36.48%
ΔpO_2	$12\,h$	126	266	6037	32.14%
	$18\,h$	151	426	10053	26.17%
	$24\,h$	171	590	14202	22.47%
TCP	$12\,h$	98	292	6168	25.13 %
	$18\,h$	116	462	9996	20.07%
	$24\,h$	130	635	13914	16.99%

TABLE 4. Number of FOM and ROM simulations employed for a UQ analysis of the oxygen transfer processes with the DL-MFMC estimator for three fixed reference computational budgets. In particular in the last column we report the percentage of FOM simulations required in the training phase.

and the sampling phase, since we obtain higher percentages for $\overline{pO_2}$. As we shall see in a moment, these considerations will also be repeated in the analyses that follow.

Given m_0^* and m_1^*, we finally moved to the sampling phase, which allowed us to estimate the coupling coefficient λ_* and eventually construct the DL-MFMC estimator. As we discussed in Section 2, our main interest is to provide a quantification of the uncertainties, and thus construct suitable confidence intervals based on our estimator. To appreciate the reduction in the uncertainties, we report the confidence intervals obtained by the DL-MFMC approach in comparison with those obtained via standard Monte Carlo (MC-FOM): we refer to Definition 2.11 and 2.1, respectively. Hereon, the confidence level has ben set to $\gamma = 99\%$.

To start, we report below the results obtained for $p = 24h$, where the robustness of the DL-MFMC estimator is particularly evident.

$$\overline{pO_2} : \begin{cases} I_{\text{DL-MFMC}}^{\gamma} & = 29.99\,mmHg \pm 0.33\,mmHg\,, \\ I_{\text{FOM}}^{\gamma} & = 30.09\,mmHg \pm 0.58\,mmHg\,. \end{cases}$$

$$\Delta pO_2 : \begin{cases} I_{\text{DL-MFMC}}^{\gamma} & = 16.92\,mmHg \pm 0.48\,mmHg\,, \\ I_{\text{FOM}}^{\gamma} & = 17.55\,mmHg \pm 0.75\,mmHg\,. \end{cases}$$

$$TCP : \begin{cases} I_{\text{DL-MFMC}}^{\gamma} & = 50.06\% \pm 1.35\%, \\ I_{\text{FOM}}^{\gamma} & = 49.75\% \pm 2.20\%. \end{cases}$$

In general, both approaches report consistent pointwise estimates. However, the DL-MFMC estimator entails a much smaller uncertainty when compared to the classical one based on standard Monte Carlo.
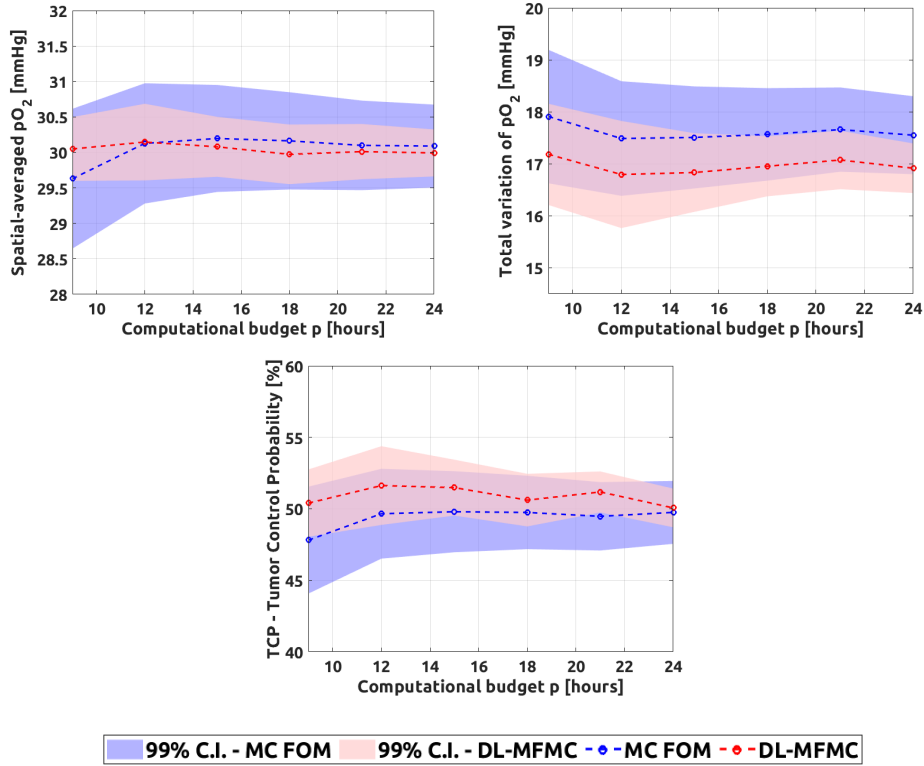
FIGURE 7. Confidence intervals estimates for all the QoIs, comparing the DL-MFMC estimator with the standard Monte-Carlo FOM-based, fixed $\gamma = 99\%$.

Notably, the same behavior was also reported for the other computational budgets, as depicted in Figure 7 and Table 5. In Figure 7 we highlight for each QoI the trend of the point estimates $\hat{E}^{\text{DL-MFMC}}$ and $\hat{E}^{\text{FOM}}$ with respect to the available computational budget p and the contours of the uncertainties of the associated 99% confidence intervals. The plots reveal various noteworthy aspects:

  i) the confidence intervals associated with the DL-MFMC estimator are consistently smaller than the ones obtained via standard Monte Carlo;
 ii) in the DL-MFMC approach, the point estimate of $\mathbb{E}[Q(\mathbf{u}^{\text{FOM}})]$ is more stable for $\overline{pO_2}$ than for ΔpO_2 and TCP;
iii) the stability of the point estimate of $\overline{pO_2}$ is reached with a smaller budget p.

As before, (ii) and (iii) are easily motivated by the additional regularity of the spatial average operator $\overline{pO_2}$, which a linear and continuous functional of C_t, as stated in Section 3.2.

These qualitative considerations are further confirmed by the quantitative analysis in Table 5. There, we focused on three reference budgets, $p = 12, 18, 24$ hours, and we reported the uncertainties attained by the DL-MFMC approach (computed as

QoI	Budget p	FOM simulations		Uncertainty	
$\overline{pO_2}$	$12\,h$	368	(-19.12%)	$0.54\ mmHg$	(-36.24%)
	$18\,h$	541	(-20.79%)	$0.42\ mmHg$	(-38.68%)
	$24\,h$	710	**(-22.06%)**	$0.33\ mmHg$	**(-43.41%)**
ΔpO_2	$12\,h$	392	**(-13.85%)**	$1.03\ mmHg$	**(-6.50%)**
	$18\,h$	577	(-15.52%)	$0.58\ mmHg$	(-34.98%)
	$24\,h$	761	(-16.47%)	$0.48\ mmHg$	(-36.33%)
TCP	$12\,h$	390	(-14.29%)	2.75%	(-12.33%)
	$18\,h$	578	(-15.37%)	1.84%	(-28.30%)
	$24\,h$	765	(-16.03%)	1.35%	(-38.44%)

TABLE 5. Computational cost and UQ for the DL-MFMC estimator. FOM simulations = total number of high-fidelity simulations required by the computational pipeline, namely $n_* + m_0^*$. Uncertainty = amplitude of the DL-MFMC confidence interval, Eq. (2.11). In parentheses, the comparison with the standard Monte Carlo estimator. E.g.: in the first row, we see that, compared to standard Monte Carlo, the DL-MFMC estimator reduced the uncertainty by 36.26% while simultaneously requiring 19.12% less FOM simulations.

the amplitude of the associated confidence interval), together with the total number of FOM simulations required by the computational pipeline (i.e., encompassing both the training and sampling phases). The results are reported in comparison with the MC-FOM approach, emphasizing the reduction in uncertainties and number of FOM evaluations.

In general, the advantage of the DL-MFMC approach becomes more and more pronounced for higher budgets. This is not particularly surprising if we consider that the ROM becomes more reliable when provided with a larger amount of data. In particular, as the correlation between FOM and ROM increases, the multi-fidelity approach starts to rely more and more on the surrogate model, ultimately reducing the number of FOM evaluations. Once again, this phenomenon is particularly evident for the average partial pressure, the QoI associated with the smoothest operator. However, it is interesting to note that, while the computational gain is nearly constant for $\overline{pO_2}$, the trend is much steeper for ΔpO_2 and TCP. For instance, when increasing the computational budget, the reduction in the uncertainties for ΔpO_2 goes from -6.5% to -36.33%, highlight a substantial boost in the performances.

Further confirmation of this fact is provided in Figure 8, where we compared computational resources with model uncertainties. Coherently with our previous observation, we note that for $\overline{pO_2}$ there is a constant computational gap between the Monte Carlo and the DL-MFMC estimators. Instead, for QoIs associated to
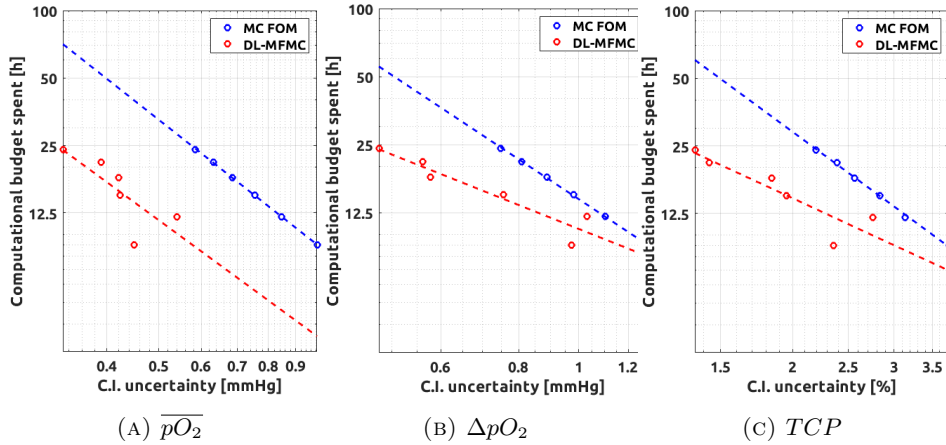
FIGURE 8. Computational budget spent to achieve a fixed level of uncertainty for both the considered estimators, considering three different QoIs.

either less regular or strongly nonlinear operators, such as ΔpO_2 and TCP, the DL-MFMC estimator becomes more advantageous when increasingly accurate estimates are required. Still, the improvement compared to a standard MC estimator remains consistent across a wide range of uncertainties.

5. **Conclusion.** This work addresses the significant computational demands associated with forward UQ for partial differential equations in multi-physics models. In particular, we propose an improvement of multi-fidelity methods [35, 36, 29], integrating state-of-the-art deep learning techniques [15, 16] to increase the efficiency and robustness of predictions.

The primary contribution of our research lies in the development of the Deep Learning-Enhanced Multi-Fidelity Monte Carlo (DL-MFMC) method. This approach, building upon several foundational works in multi-fidelity methods and deep learning, introduces a novel synergy between full-order models (FOM) and reduced-order models (ROM), based on a supervised learning approach enhanced by deep neural networks. The result is a significant reduction in computational costs while retaining the critical features necessary for accurate modeling. By applying our DL-MFMC method to oxygen transfer in the microcirculation, using quantities of interest related to radiotherapy, we have demonstrated its ability to perform robust and reliable UQ analysis in complex, real-world scenarios. We have compared our method against traditional Monte Carlo approaches, demonstrating substantial speed-ups and increased robustness.

Looking ahead, further research should focus on refining the integration of deep learning with multi-fidelity methods, as well as improving the interpretability and scalability of these models remains a crucial area for continued development. In conclusion, our research contributes to the rapidly evolving field of computational modeling and uncertainty quantification in life sciences. The DL-MFMC approach offers a promising path forward in tackling the inherent challenges of computational expense and model accuracy in multi-scale, multi-fidelity scenarios.

REFERENCES

[1] F. Alsayyari, Z. Perkó, D. Lathouwers and J. L. Kloosterman, A nonintrusive reduced order modelling approach using proper orthogonal decomposition and locally adaptive sparse grids, *Journal of Computational Physics*, **399** (2019), 108912.

[2] D. Amsallem, M. J. Zahr and C. Farhat, Nonlinear model order reduction based on local reduced-order bases, *International Journal for Numerical Methods in Engineering*, **92** (2012), 891-916.

[3] C. Audouze, F. De Vuyst and P. B. Nair, Nonintrusive reduced-order modeling of parametrized time-dependent partial differential equations, *Numerical Methods for Partial Differential Equations*, **29** (2013), 1587-1628.

[4] M. Barrault, Y. Maday, N. C. Nguyen and A. T. Patera, An 'empirical interpolation' method: Application to efficient reduced-basis discretization of partial differential equations, *Comptes Rendus Mathematique*, **339** (2004), 667-672.

[5] P. Benner, Peter S. Gugercin and K. Willcox, A survey of projection-based model reduction methods for parametric dynamical systems, *SIAM Review*, **57** (2015), 483-531.

[6] A. Bērziņš, J. Helmig, F. Key and S. Elgeti, Standardized non-intrusive reduced order modeling using different regression models with application to complex flow problems, *Thesis (Ph.D.)–University of Rochester ProQuest LLC, Ann Arbor, MI*, (1970) 55 pp.

[7] P. Chen, A. Quarteroni and G. Rozza, Reduced basis methods for uncertainty quantification, *SIAM/ASA Journal on Uncertainty Quantification*, **5** (2017), 813–869,

[8] W. Chen, J. S. Hesthaven, J. B. Junqiang, Y. Qiu, Z. Yang and Y. Tihao, Greedy nonintrusive reduced order model for fluid dynamics, *AIAA Journal*, **56** (2018), 4927-4943.

[9] L. Cicci, S. Fresca, M. Guo, A. Manzoni and P. Zunino, Uncertainty quantification for nonlinear solid mechanics using reduced order models with Gaussian process regression, *Computers Mathematics with Applications*, **149** (2023), 1-23.

[10] P. Conti, M. Guo, A. Manzoni and J. S. Hesthaven, Multi-fidelity surrogate modeling using long short-term memory networks, *Computer Methods in Applied Mechanics and Engineering*, **404** (2023), 115811.

[11] M. Croci, V. Vinje and M. E. Rognes, Fast uncertainty quantification of tracer distribution in the brain interstitial fluid with multilevel and quasi Monte Carlo, *International Journal for Numerical Methods in Biomedical Engineering*, **37** (2021).

[12] U. Del Monte, Does the cell number 10^9 still really fit one gram of tumor tissue?, *Cell Cycle*, **8** (2009), 505-506.

[13] M. Durante, J. Debus and J. S. Loeffler, Physics and biomedical challenges of cancer therapy with accelerated heavy ions, *Nature Reviews Physics*, **3** (2021), 777-790.

[14] I.-G. Farcaş, B. Peherstorfer, T. Neckel, F. Jenko and H.-J. Bungartz, Context-aware learning of hierarchies of low-fidelity models for multi-fidelity uncertainty quantification, *Computer Methods in Applied Mechanics and Engineering*, **406** (2023), 115908.

[15] N. Franco, A. Manzoni and P. Zunino, A deep learning approach to reduced order modelling of parameter dependent partial differential equations, *Mathematics of Computation*, **92** (2023), 483-524.

[16] N. R. Franco, A. Manzoni and P. Zunino, Mesh-Informed Neural Networks for Operator Learning in Finite Element spaces, *Journal of Scientific Computing*, **97** (2023), 41 pp.

[17] N. R. Franco, S. Fresca, A. Manzoni and P. Zunino, Approximation bounds for convolutional neural networks in operator learning, *Neural Networks*, **161** (2023), 129-141.

[18] Z. Gao, Q. Liu, J. S. Hesthaven, B. S. Wang, W. S. Don and X. Wen, Non-intrusive reduced order modeling of convection dominated flows using artificial neural networks with application to Rayleigh-Taylor instability, *Commun. Comput. Phys.*, **30** (2021), 97-123.

[19] M. Guo, A. Manzoni, M. Amendt, P. Conti and J. S. Hesthaven, Multi-fidelity regression using artificial neural networks: Efficient approximation of parameter-dependent output quantities, *Comput. Methods Appl. Mech. Engrg.*, **389** (2022), 114378.

[20] J. S. Hesthaven, G. Rozza and B. Stamm,, *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*, SpringerBriefs Math. BCAM SpringerBriefs Springer, Cham; BCAM Basque Center for Applied Mathematics, Bilbao, 2016.

[21] J. S. Hesthaven and S. Ubbiali, Non-intrusive reduced order modeling of nonlinear problems using neural networks, *Journal of Computational Physics*, **363** (2018), 55-78.

[22] D. Z. Huang, N. H. Nelsen and M. Trautner, An operator learning perspective on parameter-to-observable maps, preprint, `arXiv:2402.06031`, 2024.

[23] A. M. Kellerer and H. H. Rossi, RBE and the primary mechanism of radiation action, *Radiation Research*, **47** (1971), 15-34, `http://www.ncbi.nlm.nih.gov/pubmed/5559377`.

[24] T. Köppl, E. Vidotto and B. Wohlmuth, A 3D-1D coupled blood flow and oxygen transport model to generate microvascular networks, *International Journal for Numerical Methods in Biomedical Engineering*, **36** (2020), 30 pp.

[25] J. R. Levick, Capillary filtration-absorption balance reconsidered in light of dynamic extravascular factors, *Experimental Physiology*, **76** (1991), 825-857.

[26] A. Lücker, B. Weber and P. Jenny, A dynamic model of oxygen transport from capillaries to tissue with moving red blood cells, *American Journal of Physiology - Heart and Circulatory Physiology*, **308** (2015), 206-216.

[27] S. J. McMahon, The linear quadratic model: Usage, interpretation and challenges, *Physics in Medicine and Biology*, **64** (2019).

[28] T. E. Moschandreou, C. G. Ellis and D. Goldman, Influence of tissue metabolism and capillary oxygen supply on arteriolar oxygen transport: A computational model, *Mathematical Biosciences*, **232** (2011), 1-10.

[29] L. W. T. Ng and K. E. Willcox, Multifidelity approaches for optimization under uncertainty, *International Journal for Numerical Methods in Engineering*, **100** (2014), 746-772.

[30] T. O'Leary-Roseberry, U. Villa, P. Chen and O. Ghattas, Derivative-informed projected neural networks for high-dimensional parametric maps governed by PDEs, *Computer Methods in Applied Mechanics and Engineering*, **388** (2022), 114199.

[31] B. Peherstorfer, Multifidelity monte carlo estimation with adaptive low-fidelity models, *SIAM/ASA J. Uncertain. Quantification*, **7** (2019), 579-603.

[32] B. Peherstorfer, Breaking the kolmogorov barrier with nonlinear model reduction, *Notices of the American Mathematical Society*, **69** (2022), 725-733.

[33] B. Peherstorfer, D. Butnaru, K. Willcox and H. J. Bungartz, Localized discrete empirical interpolation method, *SIAM Journal on Scientific Computing*, **36** (2014), 168-192.

[34] B. Peherstorfer and K. Willcox, Data-driven operator inference for nonintrusive projection-based model reduction, *Computer Methods in Applied Mechanics and Engineering*, **306** (2016), 196-215.

[35] B. Peherstorfer, K. Willcox and M. Gunzburger, Optimal model management for multifidelity Monte Carlo estimation, *SIAM Journal on Scientific Computing*, **38** (2016), A3163-A3194.

[36] B. Peherstorfer, K. Willcox and M. Gunzburger, Survey of multifidelity methods in uncertainty propagation, inference, and optimization, *Siam Review*, **60** (2018), 550-591.

[37] L. Possenti, G. Casagrande, S. Di Gregorio, P. Zunino and M. L. Costantino, Numerical simulations of the microvascular fluid balance with a non-linear model of the lymphatic system, *Microvascular Research*, **122** (2019), 101-110.

[38] L. Possenti, A. Cicchetti, R. Rosati, D. Cerroni, M. L. Costantino, T. Rancati and P. Zunino, A Mesoscale Computational Model for Microvascular Oxygen Transfer, *Annals of Biomedical Engineering*, **29** (2021), 3356-3373.

[39] L. Possenti, S. Di Gregorio, G. Casagrande, M. L. Costantino, T. Rancati and P. Zunino, A global sensitivity analysis approach applied to a multiscale model of microvascular flow, *Computer Methods in Biomechanics and Biomedical Engineering*, (2020), 1215-1224.

[40] L. Possenti, S. di Gregorio, F. M. Gerosa, G. Raimondi, G. Casagrande, M. L. Costantino and P. Zunino, A computational model for microcirculation including Fahraeus-Lindqvist effect, plasma skimming and fluid exchange with the tissue interstitium, *Journal for Numerical Methods in Biomedical Engineering*, **35** (2019), 27 pp.

[41] L. Possenti, P. Vitullo, A. Cicchetti, P. Zunino, and T. Rancati, Modeling hypoxia-induced radiation resistance and the impact of radiation sources, *Computers in Biology and Medicine*, **173** (2024), 1-16.

[42] A. R. Pries, T. W. Secomb, T. Gessner, M. B. Sperandio, J. F. Gross and P. Gaehtgens, Resistance to blood flow in microvessels in vivo, *Circulation Research*, **75** (1994), 904-915.

[43] A. Quarteroni, A. Manzoni and F. Negri, *Reduced Basis Methods for Partial Differential Equations: An Introduction*, Springer, Cham, 2016.

[44] Y. Renard and K. Poulios, etFEM: Automated FE modeling of multiphysics problems based on a generic weak form language, *ACM Trans. Math. Software*, **47** (2021), 31 pp.

[45] T. W. Secomb, R. Hsu, E. Y. H. Park and M. W. Dewhirst, Green's function methods for analysis of oxygen delivery to tissue by microvascular networks, *Annals of Biomedical Engineering*, **32** (2004), 1519-1529.

[46] R. C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2014.

[47] L. Strigari, F. Torriani, L. Manganaro, T. Inaniwa, F. Dalmasso, R. Cirio and A. Attili, Tumour control in ion beam radiotherapy with different ions in the presence of hypoxia: An oxygen enhancement ratio model based on the microdosimetric kinetic model, *Physics in Medicine and Biology*, **63** (2018), article number: 065012.

[48] M. A. Swartz and M. E. Fleury, Interstitial flow and its effects in soft tissues, *Annual Review of Biomedical Engineering*, **9** (2007), 229-256.

[49] W. Tinganelli, M. Durante, R. Hirayama, M. Krämer, A. Maier, W. Kraft-Weyrather, Y. Furusawa, T. Friedrich and E. Scifoni, Kill-painting of hypoxic tumours in charged particle therapy, *Scientific Reports*, **5** (2015).

[50] P. Vitullo, L. Cicci, L. Possenti, A. Coclite, M. L. Costantino and P. Zunino, Sensitivity analysis of a multi-physics model for the vascular microenvironment, *International Journal for Numerical Methods in Biomedical Engineering*, **107** (2023), 351-352.

[51] P. Vitullo, A. Colombo, N. R. Franco, A. Manzoni and P. Zunino, Nonlinear model order reduction for problems with microstructure using mesh informed neural networks, *Finite Elements in Analysis and Design*, **229** (2024), article number: 104068.

[52] Q. Wang, J. S. Hesthaven and D. Ray, Non-intrusive reduced order modeling of unsteady flows using artificial neural networks with application to a combustion problem, *Journal of Computational Physics*, **384** (2019), 289-307.

[53] M. Welter, T. Fredrich, H. Rinneberg and H. Rieger, Computational model for tumor oxygenation applied to clinical data on breast tumor hemoglobin concentrations suggests vascular dilatation and compression, *PLoS ONE*, **11** (2016).