Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spatio-temporal Neural Surrogates for Agent-Based Burn Wound Models

Roland V. Bumbuc

Department of Plastic, Reconstructive and Hand Surgery, Amsterdam Movement Sciences (AMS) Institute, Amsterdam UMC, Location VUmc, Amsterdam, The Netherlands; Department of Molecular Cell Biology and Immunology, Amsterdam Infection and Immunity (AII) Institute, Amsterdam UMC, Location VUmc, Amsterdam, The Netherlands; Computational Science Lab, Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands https://orcid.org/0009-0004-9173-9798

Julia Kowalczuk

Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands https://orcid.org/0009-0001-1987-9604

Rezo Getsadze

Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands https://orcid.org/0009-0008-1153-2455

Ioannis Papapanagiotou

1st Department of Critical Care Medicine, Medical School, National & Kapodistrian University of Athens, Evangelismos Hospital, Athens, Greece https://orcid.org/0009-0002-7651-588X

H. Ibrahim Korkmaz

h.korkmaz@amsterdamumc.nl

Department of Plastic, Reconstructive and Hand Surgery, Amsterdam Movement Sciences (AMS) Institute, Amsterdam UMC, Location VUmc, Amsterdam, The Netherlands; Department of Molecular Cell Biology and Immunology, Amsterdam Infection and Immunity (AII) Institute, Amsterdam UMC, Location VUmc, Amsterdam, The Netherlands; Alliance of Dutch Burn Care, Burn Center and Department of Plastic and Reconstructive Surgery, Red Cross Hospital, Beverwijk, The Netherlands https://orcid.org/0000-0001-7356-0701

Vivek M. Sheraton

v.s.muniraj@uva.nl

Computational Science Lab, Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands https://orcid.org/0000-0002-6577-6016

Research Article

Keywords: Agent-Based Model, Burn wound healing, Deep Operator Network, Physics-Informed Neural Network, Surrogate modelling

Posted Date: June 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9937754/v1

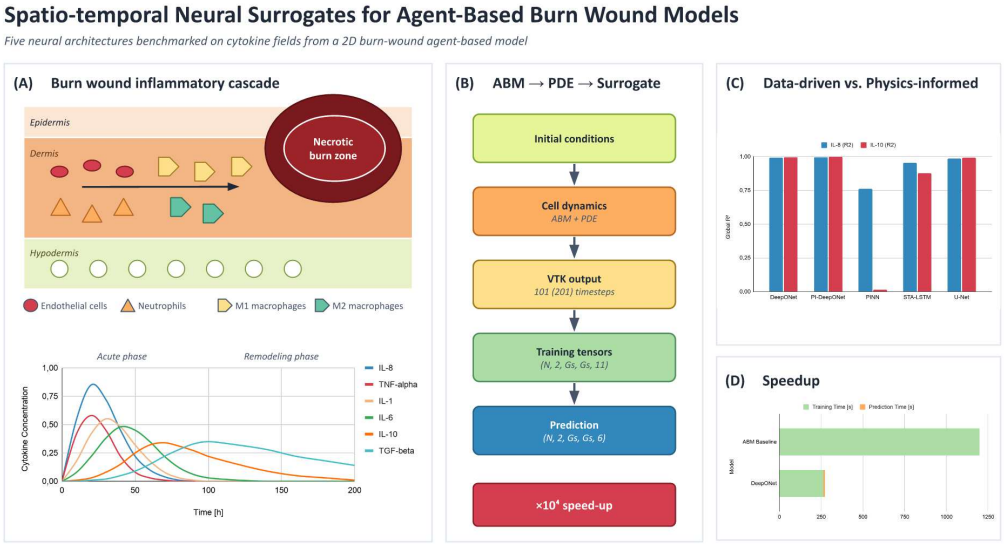
License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Graphical Abstract

Spatio-temporal Neural Surrogates for Agent-Based Burn Wound Models

Roland V. Bumbuc, Julia Kowalczuk, Rezo Getsadze, Ioannis Papapanagiotou, H. Ibrahim Korkmaz, Vivek M. Sheraton



Highlights

Spatio-temporal Neural Surrogates for Agent-Based Burn Wound Models

Roland V. Bumbuc, Julia Kowalczuk, Rezo Getsadze, Ioannis Papapanagiotou, H. Ibrahim Korkmaz, Vivek M. Sheraton

- DeepONet is the most accurate surrogate at 250^2 ($R^2 > 0.99$ for both cytokines).
- U-Net trains 17-41$\times$ faster than DeepONet at 250^2 with near-equal accuracy.
- Surrogates give 10^2-$10^4\times$ per-evaluation speed-ups, amortising within $\sim$1 trajectory.
- Physics-informed losses do not help on this discrete, stochastic ABM.
- Sparse fields (IL-10) expose structural limits that more data cannot fix.

Spatio-temporal Neural Surrogates for Agent-Based Burn Wound Models

Roland V. Bumbuc[a,b,c,1], Julia Kowalczuk[d,1], Rezo Getsadze[d], Ioannis Papapanagiotou[e], H. Ibrahim Korkmaz[a,b,f], Vivek M. Sheraton[c,*]

[a] *Department of Plastic, Reconstructive and Hand Surgery, Amsterdam Movement Sciences (AMS) Institute, Amsterdam UMC, Location VUmc, Amsterdam, The Netherlands*

[b] *Department of Molecular Cell Biology and Immunology, Amsterdam Infection and Immunity (AII) Institute, Amsterdam UMC, Location VUmc, Amsterdam, The Netherlands*

[c] *Computational Science Lab, Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands*

[d] *Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands*

[e] *1st Department of Critical Care Medicine, Medical School, National & Kapodistrian University of Athens, Evangelismos Hospital, Athens, Greece*

[f] *Alliance of Dutch Burn Care, Burn Center and Department of Plastic and Reconstructive Surgery, Red Cross Hospital, Beverwijk, The Netherlands*

Abstract

Agent-based models (ABMs) of the post-burn immune response capture cell-level heterogeneity and spatially resolved cytokine dynamics, but their cost (hundreds to thousands of seconds per trajectory) limits workflows requiring repeated evaluation, such as parameter sweeps, sensitivity analysis, or Bayesian calibration. Neural surrogates promise to remove this bottleneck, but the relative merits of competing architectures remain unclear. This work benchmarks five surrogate architectures – DeepONet, Physics-Informed DeepONet (PI-DeepONet), U-Net, Spatio-Temporal Attention LSTM (STA-LSTM), and Physics-Informed Neural Network (PINN) – for predicting cytokine fields from a 2D burn wound ABM, and quantifies the accuracy-speed trade-off. All architectures are trained on a single NVIDIA A100 GPU at

*Corresponding author

Email addresses: h.korkmaz@amsterdamumc.nl (H. Ibrahim Korkmaz), v.s.muniraj@uva.nl (Vivek M. Sheraton)

[1] These authors contributed equally for this work,.

two grid resolutions (100^2, 250^2) and two cytokines spanning the sparsity spectrum (IL-8: <1%; IL-10: up to ∼13%), with performance assessed under temporal interpolation and extrapolation using five metrics (Dice, SSIM, spatial correlation, R^2, and Masked RMSE). DeepONet achieves the highest accuracy ($R^2 > 0.99$ at 250^2), while U-Net retains comparable accuracy and trains 17-41× faster, giving the best accuracy-speed balance. Physics-informed variants yield no consistent improvement, reflecting a mismatch between the continuum PDE assumption and the ABM's discrete, stochastic dynamics, and all architectures struggle on sparse cytokine fields, with the stand-alone PINN failing most severely. Surrogates yield 10^2-10^4× per-evaluation speed-ups and amortise their training cost within at most two ABM trajectories, bringing Bayesian calibration and digital-twin pipelines into reach; the binding constraint on further gains is structural, not data-driven.

Keywords: Agent-Based Model, Burn wound healing, Deep Operator Network, Physics-Informed Neural Network, Surrogate modelling

1. Introduction

Severe burns are inflammatory injuries as much as physical ones. Beyond the visible wound, they trigger a cascade of local and systemic immune responses whose intensity and duration largely determine whether the injury heals, progresses, or escalates into multi-organ complications [1]. The severity of these responses depends on factors such as anatomical location, burn depth, and total body surface area (TBSA) burned, and they remain difficult to anticipate because the tools used to assess burn patients in clinical practice are essentially static: scores computed at admission summarise injury severity at one instant in time but do not track how the wound's microenvironment evolves over the following hours and days [2, 3]. With burns estimated to cause approximately 180,000 deaths each year, disproportionately in low- and middle-income countries [4], the practical question is not simply "how bad is this injury today?" but rather "where will it be tomorrow?". Answering the latter requires models that simulate how the wound evolves in time - not just describe its state once.

Computational modelling is increasingly used to study immune dynamics during wound healing [5]. Hybrid agent-based models (ABMs) combine discrete cell behaviour with continuous chemical fields and are well suited

to capturing spatial heterogeneity and local interactions [6]. Understanding these heterogeneous fields, both in the pathophysiology of the burn and in a model that can simulate the physical phenomena these fields "control", could be key for treating burn injuries. The computational cost of these models is concrete and substantial: a single 100-hour ABM trajectory under the configuration of [5] takes approximately 541 s at 100^2 resolution and 2503 s ($\sim$42 min) at 250^2 on the CompuCell3D reference implementation. For workflows that involve hundreds to thousands of evaluations – parameter sweeps, global sensitivity analysis, Bayesian calibration, or patient-specific scenario exploration – these costs translate into days to weeks of compute per study, which is impractical in many settings. This has motivated the use of neural network (NN) surrogates to approximate ABM outputs more efficiently [7–9]. In practice, well-trained surrogates can reduce the per-evaluation cost by several orders of magnitude, making the same workflows tractable on a single machine within minutes rather than weeks [10–13]. Several machine learning approaches have been explored for this problem. Earlier work has demonstrated that methods such as random forests and feed-forward neural networks can learn ABM input-output relationships with reasonable accuracy [11], although these models remain largely data-driven. More recent studies have shifted towards incorporating additional structure into the learning process. For example, surrogate models that directly infer governing equations from simulation data have been proposed [14]. Within this broader shift, Physics-Informed Neural Networks (PINNs) represent a natural extension. They incorporate PDE-based constraints in the loss function to encourage physically consistent predictions [15, 16], although their training is known to suffer from gradient pathologies and spectral bias that can render the optimisation problem ill-conditioned in practice [17, 18]. For spatially resolved ABMs, the task is often reformulated as an image-to-image prediction problem, making convolutional architectures a natural choice [19, 20]. These developments have also been considered in the context of digital twins and optimal control, where fast surrogate models are required to evaluate numerous possible intervention strategies in real time [12, 13].

In burn modelling specifically, a similar progression can be observed. Early work focused on mechanistic ABMs of the proliferative phase in simplified 2D settings [21]. More recent frameworks link local tissue dynamics with systemic processes to capture early inflammatory responses [5]. Surrogate models have been introduced to reduce the cost of repeated simulations. Papapanagiotou et al. [7] demonstrated that attention-based LSTM models can

approximate cytokine diffusion and cell dynamics while achieving substantial computational speed-ups. They also reported several limitations, including sensitivity to long-horizon drift and difficulty in capturing sharp spatial fronts or low-signal cytokines. Their experiments with PINNs suggested that, while biologically appealing, these models can be computationally expensive and may not always achieve competitive accuracy. It is therefore not yet clear whether richer architectural families, such as operator networks, can address the long-horizon drift and sparse-signal failures reported so far, nor under which resolution and temporal regimes such surrogates remain reliable.

Deep Operator Networks (DeepONets) are mesh-aware through sampled function representations, rather than magically resolution-free without adequate sensor design [22]. Their branch-trunk architecture enables the evaluation of predictions at arbitrary spatial locations without the need for retraining. This is particularly useful when working with varying resolutions or irregular domains; however, the practical benefit depends on the representation of the input functions. Recent studies indicate that DeepONets can capture long-range dependencies in spatio-temporal systems, including applications to post-burn contraction and related processes [8, 9, 23].

1.1. Biological/medical relevance

From a clinical perspective, the ability to model immune dynamics in burn patients is of interest as a potential complement to existing triage and treatment decision-making tools. Wound healing proceeds through overlapping haemostatic, inflammatory, proliferative, and remodelling phases, whose duration and intensity vary between individuals and injury severity [24]. Dysregulation of these phases contributes to complications in severe burns and limits the predictive value of static assessment tools. Although the present work is purely computational and does not involve patient data, computationally efficient models that approximate underlying biological dynamics may be relevant for future translational applications.

Current clinical burn assessment primarily relies on static prognostic scores such as the Abbreviated Burn Severity Index (ABSI) and the revised Baux score, which combine factors including age, total body surface area (TBSA) burned, and inhalation injury to estimate mortality risk [2, 3]. These scores are well validated and widely used in clinical practice, but they are time-invariant and do not capture spatial or temporal dynamics of the wound micro-environment. Mechanistic models of post-burn immune responses can represent this spatial and temporal heterogeneity, but their computational

cost currently limits their use in time-sensitive or large-scale simulation settings. Hybrid agent-based models (ABMs) capture local cell-cell and cytokine interactions in principle, but are generally too computationally expensive for routine use or repeated simulation. In this context, surrogate models may provide a computationally efficient approximation of ABM outputs. In principle, such models could enable faster exploration of patient- or wound-specific scenarios by approximating spatio-temporal cytokine dynamics under defined input conditions. However, this remains speculative and requires validation before any clinical interpretation. This is particularly relevant in research and computational settings where large parameter sweeps or repeated simulations are required, rather than in direct bedside applications.

This study does not evaluate clinical utility and does not involve patient-level validation. Instead, it focuses on a computational prerequisite: assessing whether neural surrogate models can approximate ABM-generated spatio-temporal cytokine fields with sufficient accuracy and efficiency to justify further translational development.

1.2. Computational relevance

The computational cost of ABMs not only increases runtime; it also limits the types of analyses that can be conducted. Without surrogate models, tasks such as uncertainty quantification, global sensitivity analysis, or Bayesian calibration against patient data quickly become impractical. Additionally, the discretisation-invariant nature of DeepONets addresses a practical issue. ABM outputs are typically defined on fixed grids, whereas clinical data, such as wound images, are available at varying resolutions and geometries. Methods requiring retraining for each new discretisation are challenging to implement in practice. DeepONets, however, operate on functional inputs (e.g., continuous representations of cytokine fields) and can be evaluated at arbitrary spatial locations. This flexibility makes them suitable candidates for applications where input data are heterogeneous. This flexibility is not unconditional, however, the quality of a DeepONet's predictions depends on how the input function is sampled at the sensor locations and encoded by the branch network, so the apparent resolution-independence ultimately rests on design choices that must be matched to the problem at hand.

1.3. Gap

Despite recent progress, the field still lacks a systematic benchmark of neural surrogates for post-burn cytokine-field prediction that accounts for the specific structure of this problem. The relevant features include sharp cytokine fronts that evolve over time, low-abundance species that are biologically important yet numerically sparse, and multi-scale interactions between local cellular behaviour and systemic outcomes. Existing studies of post-burn surrogates rely largely on error metrics averaged over space and time, which can obscure model failures in clinically relevant regions, such as the wound edge. This work constructs a multi-resolution benchmark and evaluates models not only in terms of overall accuracy but also regarding their ability to generalise across spatial resolutions and time horizons beyond the training range.

1.4. Study extension and comparison framework

Concretely, this study extends the work of Papapanagiotou et al. [7], who introduced an attention-based LSTM surrogate for the same ABM, in two directions. First, three additional architectural families are evaluated on the same problem: two operator-learning models (DeepONet, PI-DeepONet) and a U-Net baseline that is widely used for image-to-image surrogates of biological PDEs but had not been benchmarked on this ABM. Second, the standalone PINN and STA-LSTM are re-evaluated under the same metric framework so that all five architectures can be compared head-to-head rather than across heterogeneous protocols. The result is a more complete characterisation of which architectural choices generalise across the regimes of interest for this specific ABM, rather than a claim of broader applicability beyond it.

1.5. Research objectives

In this work, neural network surrogates are developed and evaluated for an agent-based model (ABM) of the post-burn immune response, emphasising predictive performance and computational efficiency.

First, this study assesses whether DeepONets can provide accurate surrogate models for the burn-wound ABM and identifies configurations that capture relevant spatio-temporal dynamics under realistic computational constraints.

Secondly, DeepONets and Physics-Informed DeepONets (PI-DeepONets) are compared with other commonly used architectures (PINNs, STA-LSTMs,

and U-Net), focusing on the influence of their inductive biases on the prediction of cytokine field evolution and cell distributions.

Third, surrogate fidelity is assessed using multiple complementary metrics, including Masked Root Mean Squared Error (Masked RMSE), Dice Similarity Coefficient (DSC), Structural Similarity Index (SSIM), spatial correlation, and the coefficient of determination (R^2), across both interpolation and extrapolation time windows.

Fourth, the accuracy-speed trade-off across architectures is quantified by jointly reporting predictive performance, training cost, and inference latency relative to direct ABM simulation. This objective is treated as a primary outcome rather than an ancillary observation, since computational efficiency is what ultimately determines whether a surrogate is usable in downstream tasks such as parameter sweeps, sensitivity analysis, or near real-time decision support.

2. Methods

2.1. Baseline agent-based model

The work builds directly on the ABM introduced in [5], which is used without modification as a baseline. This model is implemented in CompuCell3D and adheres to the Glazier-Graner-Hogeweg (GGH) framework. In practice, this means that cell-level stochastic dynamics are coupled with a set of diffusion-reaction PDEs describing six cytokines: IL-8, IL-1β, IL-6, IL-10, TNF-α, and TGF-β. For each cytokine u, the field evolves according to

$$\frac{\partial u}{\partial t} = D\,\nabla^2 u \;-\; k\,u \;+\; \underbrace{\sum_{c\in\mathcal{C}_{\text{prod}}} \alpha_c\,\mathbf{1}_c(x,y,t)}_{\text{cell-driven secretion}} \;-\; \underbrace{\sum_{c\in\mathcal{C}_{\text{upt}}} \beta_c\,\mathbf{1}_c(x,y,t)\,u}_{\text{cell-driven uptake}}, \qquad (1)$$

where D is the cytokine-specific diffusion coefficient, k the linear decay rate, and $\mathbf{1}_c(x,y,t) \in \{0,1\}$ is the indicator function for the presence of cell type c at (x,y) at time t. The sets $\mathcal{C}_{\text{prod}}$ and $\mathcal{C}_{\text{upt}}$ contain the cell types that produce or take up the cytokine, with rate constants α_c and β_c inherited from the original ABM [5]. For IL-10, for example, $\mathcal{C}_{\text{prod}}$ contains M2 macrophages and $\mathcal{C}_{\text{upt}}$ contains the receptor-bearing cell populations that internalise the chemokine. Equation 1 is the form to which the physics-informed losses defined in Section 2.5.4 and following are matched. The numerical values of

D, k, α_c, and β_c, together with the full cellular rule set, are taken from [5] without modification.

These cytokine fields evolve over a two-dimensional domain representing wounded tissue and adjacent vasculature. Their dynamics are not solely diffusive. Depending on the local environment, they may be produced, degraded, internalised, or neutralised. The exact rules governing these processes are taken directly from the original implementation and are not revisited here.

On the cellular side, the model comprises a heterogeneous population of endothelial cells, neutrophils in multiple states (resting, activated, and necrotic), monocytes, macrophages (M1 and M2 phenotypes), fibroblasts, and myofibroblasts. Cell behaviour emerges from the local conditions within the simulated environment rather than being prescribed directly. Migration, for instance, is governed by chemotaxis, while phenotype transitions depend on the surrounding cytokine concentrations.

All parameters and governing equations are retained exactly as defined in [5], ensuring that any differences observed in subsequent experiments can be attributed to the surrogate models rather than modifications to the underlying biological system.

2.2. Datasets

All data used in this study are generated from the simulation framework originally introduced by Korkmaz et al. [5], with simulations re-run for the present study using the published code base [25]; the dataset is not inherited from prior surrogate work [7] but generated under matched parameter settings to allow direct comparison. The dataset spans 100 hours and is sampled at hourly intervals (10,000 Monte Carlo steps per hour), resulting in a total of 101 snapshots. Simulations were performed on grids of size 100^2 and 250^2 to examine the effect of spatial resolution. The wall-clock cost of producing a single ABM trajectory under the reference simulator configuration is 541 s at 100^2 and 2503 s at 250^2; these costs serve as the reference baselines for the speed-up factors and amortisation analysis reported in Section 3.4.

From the raw CompuCell3D outputs, binary masks were extracted for five cell types: endothelial cells, resting and activated neutrophils, and M1 and M2 macrophages. These masks enter the surrogate pipeline in two distinct roles. For the data-driven architectures (U-Net, STA-LSTM, DeepONet, PI-DeepONet data term), the masks are concatenated with the cytokine fields as additional input channels at each look-back time step ($t-2$ and $t-1$), so the network sees the contemporaneous cell distribution corresponding to

each input frame. For the physics-informed models (PINN, PI-DeepONet PDE term), the masks are first averaged over the 70 training time-steps to produce a static, time-averaged spatial occupancy map for each cell type. This aggregated map is then multiplied by the cytokine-specific secretion and uptake rate constants from the ABM to construct the spatially varying source terms $s_1(x, y)$, $s_2(x, y)$, and uptake $e(x, y)$ that enter the PDE residual (Eq. 5). The static approximation reduces memory cost and smooths agent-level stochasticity in cell positions; it remains valid because the cell populations evolve slowly relative to the cytokine diffusion timescale. The biological assignment of cells to cytokines (e.g. M2 macrophages as the source term for IL-10) follows the rules embedded in the original ABM [5] and is detailed alongside Eq. 5 in Section 2.5.4.

A simple statistical analysis of the dataset (see Table 1) reveals a pronounced imbalance between cytokines. IL-8 is almost ubiquitous across the domain, with sparsity consistently below 1%. In contrast, IL-10 is significantly more intermittent, reaching sparsity levels of approximately 13% at the higher resolution. Rather than attempting to model all cytokines simultaneously, this study focuses on these two as representative extremes. This approach renders the problem manageable while still probing performance under different signal regimes. The remaining four cytokines are deferred for tractability, as their intermediate sparsity is expected to fall between the IL-8 and IL-10 regimes already covered.

Table 1: Dataset statistics at 100 h simulation time, for the two grid resolutions used in the experiments. Sparsity is the fraction of grid points at which the cytokine concentration is zero.

Grid	**Cytokine**	**Global Min**	**Global Max**	**Sparsity (%)**
100^2	IL-1β	0	4.94×10^{-7}	2.78
100^2	IL-6	0	1.49×10^{-7}	5.75
100^2	IL-8	0	1.83×10^{-6}	0.16
100^2	IL-10	0	3.01×10^{-8}	8.50
100^2	TGF-β	0	1.97×10^{-6}	5.38
100^2	TNF-α	0	6.20×10^{-7}	4.14
250^2	IL-1β	0	2.35×10^{-8}	5.15
250^2	IL-6	0	2.36×10^{-8}	9.10
250^2	IL-8	0	7.61×10^{-8}	0.48
250^2	IL-10	0	5.35×10^{-9}	12.79
250^2	TGF-β	0	3.43×10^{-8}	7.59
250^2	TNF-α	0	4.61×10^{-8}	7.33

Figure 1 illustrates the distributions of cytokines and the structural contrast between dense and highly sparse regimes at $t = 80\,\mathrm{h}$.

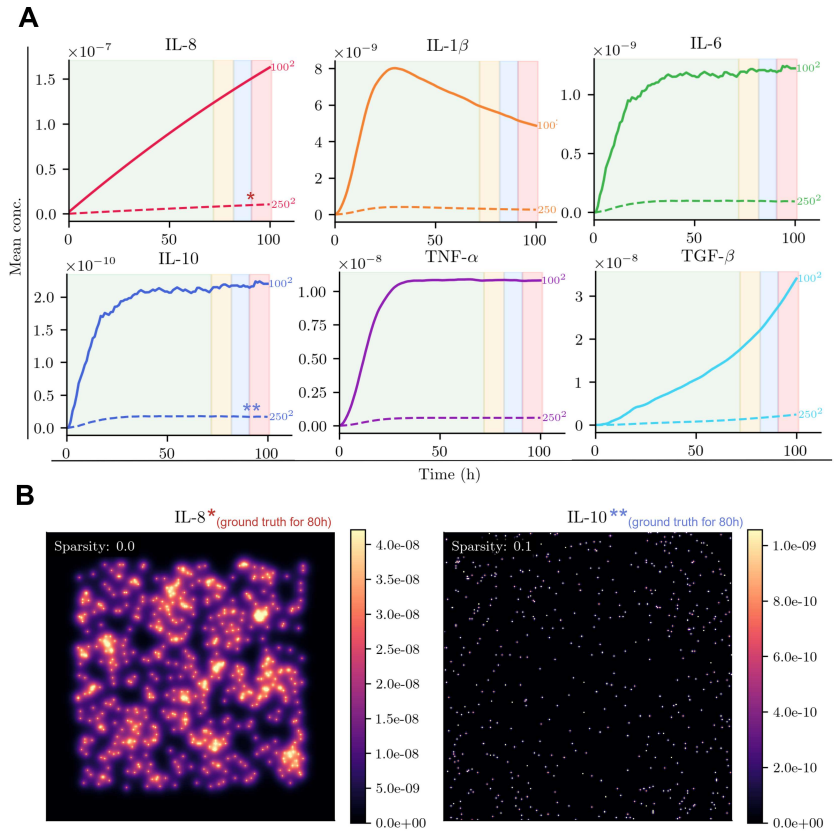


Figure 1: Mean cytokine concentrations over the 100 h simulation, together with example spatial distributions.**(A)** shows the spatially averaged concentration for each cytokine, reflecting the variation in their temporal behaviour. **(B)** shows snapshots of the ground truth at $t = 80\,\mathrm{h}$ for IL-8 and IL-10, emphasising the contrast between a relatively dense field and a highly sparse one. The colour scale is capped at the 99.5th percentile to prevent a small number of large values from compressing the dynamic range and masking spatial structure.

2.3. Data preparation

To incorporate temporal information, a sliding window with a look-back of $w = 2$ is utilised. Each prediction at time t is based on the two preceding steps ($t-2$ and $t-1$), with the corresponding cytokine fields and cell masks stacked to form the input. This results in tensors of shape $(N, 2, G_s, G_s, 11)$

and targets of shape $(N, G_s, G_s, 6)$, where G_s denotes the grid size and the six output channels correspond to the six cytokines (IL-8, IL-1β, IL-6, IL-10, TNF-α, TGF-β). The number of samples N is 99 for the 100-hour dataset.

Although the targets contain all six cytokines, this work focuses on IL-8 and IL-10 as representative endpoints of the sparsity spectrum. For every architecture, a separate model is trained per cytokine: the dataset's six output channels are sliced and only the channel corresponding to the target cytokine is used as ground truth. Joint multi-cytokine training is not used, because the cytokines differ by several orders of magnitude in dynamic range and sparsity (Table 1), and a single shared loss term would otherwise be dominated by the most numerically prominent species. Per-cytokine training also avoids loss-balancing hyperparameters between cytokines and keeps the comparison between architectures clean.

Normalisation is applied separately to each cytokine. Because the distributions are heavy-tailed for sparse species, values are first clipped using a kurtosis-adaptive percentile threshold before scaling to $[-1, 1]$. The clipping percentile q is selected from the excess kurtosis κ of the per-cytokine training trajectory according to a piecewise schedule:

$$q(\kappa) = \begin{cases} 98.0, & \kappa \geq 600, \\ 98.5, & 300 \leq \kappa < 600, \\ 99.0, & 100 \leq \kappa < 300, \\ 99.5, & 20 \leq \kappa < 100, \\ \text{no clipping}, & \kappa < 20. \end{cases} \tag{2}$$

The corresponding upper-bound value $c_{\max} = \text{percentile}(u_{\text{train}}, q(\kappa))$ together with $c_{\min} = \min(u_{\text{train}})$ then defines the linear rescaling

$$u_{\text{scaled}} = 2\,\frac{\text{clip}(u, c_{\min}, c_{\max}) - c_{\min}}{c_{\max} - c_{\min}} - 1, \tag{3}$$

which maps the clipped data to $[-1, 1]$. In practice $c_{\min} = 0$ for all cytokines (concentrations cannot be negative), so the denormalisation simplifies to $u_{\text{phys}} = (u_{\text{scaled}} + 1)/2 \cdot c_{\max}$. The same $c_{\max}$ is reused as the threshold base in the Dice metric (Eq. 14), as the data range L in the SSIM constants, and as the denominator when denormalising predictions back to physical units. Spatial coordinates (x, y) and time t are also rescaled to $[-1, 1]$. The kurtosis-adaptive choice was empirically essential for training stability on

sparser cytokines: a fixed 99.9th percentile retained too many outliers for IL-10 ($\kappa > 600$ at 250^2), whereas a fixed 95th percentile suppressed genuine secretion peaks.

To confirm that the choice of clipping strategy does not bias the architectural comparison, a sensitivity analysis was performed in which the entire pipeline – preprocessing, Optuna hyperparameter search, and final training – was re-run under three preprocessing variants: no clipping (true per-channel maximum), a fixed 95th-percentile clip applied uniformly to all six cytokines, and the kurtosis-adaptive schedule used throughout the main results. The top two architectures (DeepONet and PI-DeepONet) retain their leading position in every configuration of cytokine and horizon; mid-rank ordering for IL-8 between U-Net, STA-LSTM and PINN shifts under the more aggressive clipping but stays within an R^2 band narrower than 0.1, indicating that the clipping choice does not drive the conclusions of Sections 3.1 and 3.2. Full per-metric tables for IL-8 (Table A.9) and IL-10 (Table A.10), together with a discussion of the two cases in which mid-rank ordering changes, are reported in Appendix Appendix A.5.

From this shared representation, inputs are adapted according to the model architecture. The STA-LSTM operates directly on the full sequential tensor $(N, 2, G_s, G_s, 11)$, with the look-back dimension treated as a recurrent axis. The U-Net flattens the look-back dimension into the channel axis to match its 2D convolutional structure, yielding inputs of shape $(N, G_s, G_s, 22)$ where each spatial location now carries 22 features (six cytokines plus five cell-type masks at each of the two look-back frames). The DeepONet variants separate inputs into global and pointwise components: the Trunk network receives the full 24-dimensional per-point feature vector consisting of two normalised spatial coordinates $(x, y) \in [-1, 1]^2$ and the 22 stacked cytokine and cell-mask values (six cytokines plus five cell-type masks at each of the two look-back frames). The Branch network receives a compact 7-dimensional summary of the wound state at the first look-back frame, consisting of: (i) the maximum normalised concentration of the target cytokine, (ii) its spatial mean, (iii) its spatial standard deviation, (iv-v) the x- and y-coordinates of the cell-mask centroid (averaged over all five cell types), (vi) the spatial extent of cell occupancy (fraction of grid points containing any cell), and (vii) the normalised time. Items (i)-(iii) capture the per-cytokine signal envelope, (iv)-(vi) capture the wound geometry, and (vii) tells the network where in the simulation timeline the prediction is being made. The PINN formulation entirely disregards field inputs and learns each cytokine

as a function of (x, y, t) alone, with separate networks trained per cytokine.

2.4. Data splits

The dataset is divided chronologically using a fixed 70/10/19 split for training, validation, and testing. This chronological design tests temporal generalisation: the model is trained exclusively on early-time dynamics and evaluated on later, fully held-out time windows that it has never observed during training or hyperparameter selection. This is a deliberately stricter evaluation than random hold-out interpolation, in which test samples would be drawn from the same temporal range as the training data and the surrogate would only need to interpolate between nearby observed states. Under the present protocol, by contrast, the surrogate must extrapolate forward in time, which is the regime relevant for the downstream applications motivating this work (forecasting wound evolution, parameter sweeps over future trajectories, and near real-time decision support).

Performance is evaluated across two regions of the test set. Evaluation is conducted over two time intervals: a near-horizon window ($t = 82$-$91\,\mathrm{h}$) and a far-horizon window ($t = 92$-$100\,\mathrm{h}$). The near/far split allows the rate of error growth with forecast distance to be quantified separately from absolute accuracy.

2.5. Architectures

Implementation details and training configurations of the architectures introduced in Figure 2 can be found in Table A.4 (Appendix Appendix A). The following discussion emphasises the aspects important for interpreting the results.

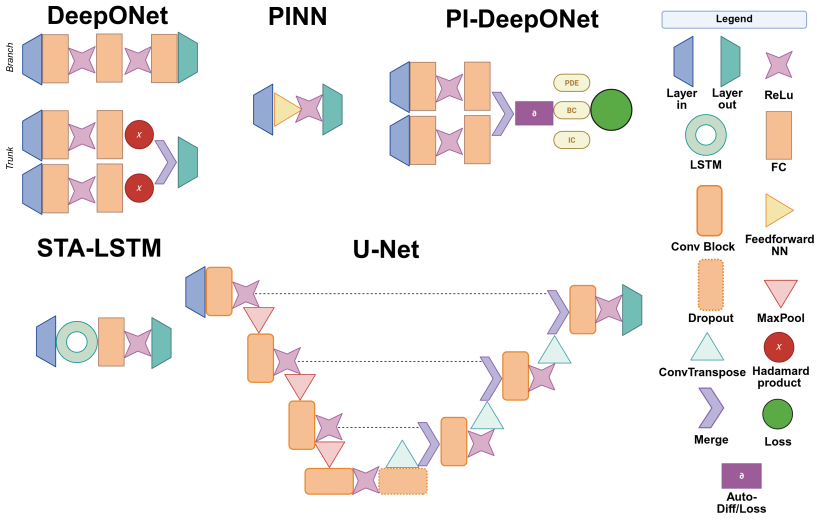


Figure 2: Simplified architectural diagrams of neural networks: DeepONet, PI-DeepONet, PINN, STA-LSTM, and U-Net. Adapted from [7].

2.5.1. U-Net

A U-Net architecture is used as the grid-based baseline. The original U-Net of Ronneberger, Fischer, and Brox [26] was developed for biomedical image segmentation and remains the canonical reference for encoder-decoder networks with skip connections; subsequent applications to ABM and PDE surrogates [20] have inherited its architecture largely unchanged. The implementation here follows the same encoder-decoder design, with skip connections that preserve spatial detail lost during downsampling. This is particularly relevant for the present problem, as some cytokine patterns are localised at the wound margin.

The input consists of two consecutive time steps, providing a total of 22 channels. This is a relatively minimal method for incorporating temporal information, but it is sufficient for stable training. A systematic sensitivity sweep over $w \in \{1, 2, 3, 4, 5, 10\}$ (Appendix Appendix A.2, Table A.5) confirms this choice: $w = 2$ is the strict optimum for IL-10 on every reported metric, while for IL-8 the best alternative ($w = 4$) yields only marginal accuracy gains at a 64-fold increase in model size and roughly twice the training time. The simpler setup is therefore retained.

2.5.2. Spatio-Temporal Attention LSTM

The STA-LSTM extends a convolutional LSTM [27] with a spatial attention layer applying softmax-normalised weights over feature maps, conceptually following the additive attention of Bahdanau et al. [28] adapted to two-dimensional spatial fields. This enables the model to focus on specific regions rather than treating all spatial locations equally at each time step. Although the look-back window is short ($w = 2$), the recurrent structure remains useful because the LSTM gating mechanism propagates information about cell-secretion events whose effects on cytokine fields persist beyond the immediate hourly step.

This becomes important in practice because activity is not evenly distributed across the domain. Large regions contribute little, while dynamics are concentrated near the wound boundary and in areas of active inflammation. The attention mechanism provides a way to account for this without explicitly defining regions of interest, which would necessitate additional assumptions.

2.5.3. Deep Operator Network

DeepONet [29] is formulated in terms of operators on function spaces. The mapping is represented through a branch-trunk decomposition:

$$G(u)(\xi) \approx \sum_{k=1}^{p} b_k(u) \cdot t_k(\xi) + c \tag{4}$$

where the bias $c \in \mathbb{R}$ is a single scalar offset shared across all query locations ξ. A separate DeepONet is trained per target cytokine, so the model produces a single-channel output and each cytokine has its own scalar bias; multi-cytokine prediction is therefore handled by training independent networks rather than by extending Eq. 4 with vector-valued outputs. This per-cytokine specialisation reflects the substantial differences in dynamic range and sparsity between cytokines (Table 1) and avoids the loss-balancing problems that arise when a single network must simultaneously fit signals differing by several orders of magnitude.

In this work, the trunk network adopts a gated structure [30], introducing element-wise interactions via Hadamard products. This modification reduces the model's tendency to smooth higher-frequency components, an effect that becomes more apparent when examining spatial predictions, particularly in regions with sharper gradients. More broadly, the operator-learning formulation decouples the spatial query points (handled by the trunk) from the

function representation (handled by the branch), which in principle allows the surrogate to be evaluated on grids different from those seen during training. This flexibility is contingent on adequate branch sampling: if the sensors used to summarise the input function fail to capture its essential structure, no amount of trunk querying will recover that information at inference time.

2.5.4. Physics-Informed Neural Network

A separate PINN is trained per cytokine, taking normalised coordinates $(x, y, t) \in [-1, 1]^3$ as input and producing a single scalar concentration as output. The network is a feed-forward architecture with tanh activations and Glorot uniform initialisation, with the number of hidden layers and units treated as tunable hyperparameters (Table A.4).

The continuum approximation underlying the PINN combines diffusion, linear decay, and spatially varying source/uptake terms [15, 16], and is the residual form of Eq. 1 with the cell-driven sums collapsed into static spatial maps. For a single cytokine, the residual takes the form:

$$F(\hat{u}; x, y, t) = \frac{\partial \hat{u}}{\partial t} - D\left(\frac{\partial^2 \hat{u}}{\partial x^2} + \frac{\partial^2 \hat{u}}{\partial y^2}\right) + k\,\hat{u} - s_1(x, y) - s_2(x, y) + e(x, y)\,\hat{u}, \quad (5)$$

where $\hat{u}$ is the network output, D and k are the cytokine-specific diffusion coefficient and decay rate inherited from the ABM, and the spatially varying terms s_1, s_2, e encode cell-driven secretion and endocytosis.

The residual in Eq. 5 is evaluated on the normalised output $\hat{u} \in [-1, 1]$ rather than on the physical concentration. The strict residual that follows from substituting $u = c_{\max}(\hat{u} + 1)/2$ into Eq. 1 would carry an additive offset of the form $(\hat{u} + 1)$ in the decay and uptake terms, reflecting the $[0, c_{\max}] \to [-1, 1]$ rescaling. The simplified form above drops that offset, and the PINN trained in this study therefore tests a *softened* physics constraint rather than the strict residual of the original PDE. The motivation is both empirical and structural: in our ABM, cytokine sources are inherently discrete and point-like at the grid scale, representing individual secreting agents. Enforcing a strict residual against such discontinuous, high-frequency signals tends to produce severe gradient instabilities during backpropagation, preventing the optimiser from successfully fitting the supervised data. Concretely, with $\hat{u} \approx -1$ over the predominantly empty domain (the zero-background of sparse cytokines such as IL-10), the constant terms produced by the $(\hat{u}+1)$ offset generate a large systematic PDE residual that dominates

the physics loss and biases the network toward the trivial zero-field solution. The softened residual acts as a milder physics regulariser that penalises non-physical configurations without attempting to force a smooth continuum solution onto a discrete generative process. We treat this as appropriate because the PDE is itself an approximation of an agent-based generative process rather than an exact governing equation. We emphasise this choice explicitly: a faithful PINN with the strictly rescaled residual was not evaluated in this study, and the conclusions drawn here about physics-informed methods on this ABM (Sections 3.2 and 4.2) should be read within that scope. A formal comparison of the two formulations is identified in Section 4.5 as a direction for follow-up work.

Diffusion coefficients are taken directly from the ABM parameter set (e.g. $D = 2.09 \times 10^{-6}$ for IL-8 in physical units) and converted to grid units via $D_{\text{grid}} = D \cdot S_{\text{MCS}}/a_{\text{conv}}$, with $S_{\text{MCS}} = 60$ Monte Carlo steps per hour and $a_{\text{conv}} = (5/G)^2$ for a $G \times G$ grid. Decay rates k are converted analogously using $h_{\text{MCS}} = 1/S_{\text{MCS}}$. All derivatives in Eq. 5 are computed by automatic differentiation through the network.

The spatially varying coefficients s_1, s_2, e are constructed from the binary cell-presence masks for endothelial cells, resting/activated neutrophils, and M1/M2 macrophages. To keep the source tensor static (rather than time-dependent), each mask is averaged over the 70 training time-steps before being multiplied by the cytokine-specific secretion/uptake constants from the ABM [5]. The static approximation reduces memory cost and smooths agent-level stochasticity, which empirically improves convergence on the 100 h dataset; it remains valid because the cell populations evolve slowly relative to the cytokine diffusion timescale. For IL-10 specifically, the M2 macrophage mask is used as the source term, consistent with the established biology of M2-driven anti-inflammatory secretion [31]. The source maps s_1 and s_2 are scaled by $2/c_{\max}$ to match the $[-1, 1]$ normalisation of the data term; the uptake map e is left unscaled because it multiplies $\hat{u}$ in Eq. 5 and therefore inherits the rescaling implicitly.

The spatial domain is the unit square $[-1, 1]^2$ and the temporal domain spans the training window $t \in [t_{\min}, t_{\max}]$ in normalised time. Homogeneous Neumann conditions $\partial\hat{u}/\partial n = 0$ are enforced on the entire boundary. The ABM uses mixed Dirichlet/Neumann conditions in principle (Dirichlet $u = 0$ on a corner sub-region of width $G/10$), but for the circular wound geometry studied here the active region never reaches the corners; the Dirichlet patches are therefore inactive and the simpler uniform Neumann formulation is used

to free a loss-weight slot for the data and PDE terms. The initial condition is enforced as a soft constraint $\hat{u}(x, y, t_{\min}) = u_0(x, y)$ at all G^2 spatial locations, where u_0 is the ground-truth field at $t = t_{\min}$.

At each training iteration, $N_d = 5\,000$ collocation points are drawn uniformly at random from the space-time domain to evaluate the PDE residual, and $N_b = 1\,000$ points are sampled on the spatial boundary for the Neumann constraint. Data observations are taken from a uniform random subsample of 2 500 spatial locations per training time-step (or the full grid when $G^2 \leq 2\,500$), giving roughly 1.75×10^5 supervised points across the 70 training time-steps. During Optuna hyperparameter search, the collocation budget is reduced to $N_d = 1\,000$ and $N_b = 200$ to make 20-trial searches tractable; the larger budget is used for the final training run.

The total loss combines the four terms with separate weights:

$$\mathcal{L}_{\text{PINN}} = \mathcal{L}_{\text{data}} + \lambda_{\text{pde}}\,\mathcal{L}_{\text{pde}} + \lambda_{\text{ic}}\,\mathcal{L}_{\text{ic}} + \lambda_{\text{bc}}\,\mathcal{L}_{\text{bc}}, \tag{6}$$

each computed as a mean squared residual over its respective sample set. The weights λ_{pde}, λ_{ic}, and λ_{bc} are tuned by Optuna over $[10^{-3}, 1]$, $[0.1, 10]$, and $[10^{-3}, 1]$ respectively (Table A.4). Optimisation proceeds in three stages: 5 000 Adam iterations on $\mathcal{L}_{\text{data}}$ alone (warm-up), 10 000 Adam iterations on the full composite loss, and 5 000 L-BFGS-B iterations for refinement. The staged schedule is necessary because jointly minimising data and physics residuals from random initialisation tends to produce ill-conditioned gradients and to let one term dominate the other, a well-documented PINN pathology [17, 18]. Model selection uses the best validation MSE checkpoint across the full 20 000-iteration schedule.

2.5.5. Physics-Informed DeepONet

The PI-DeepONet shares the branch-trunk architecture and gated trunk of the data-driven DeepONet (Section above) and extends it with the same physics constraints used by the standalone PINN. A separate network is trained per cytokine; the branch encodes the same 7-scalar wound-state features, the trunk receives the same 24-dimensional per-point input as the data-driven DeepONet, and the diffusion coefficient D, decay rate k, and source/uptake terms s_1, s_2, e (including the same simplified residual form discussed in Section 2.5.4) are inherited unchanged from the PINN setup.

The composite loss is calculated as:

$$\mathcal{L}_{\text{PI-DeepONet}} = \mathcal{L}_{\text{data}} + \lambda_{\text{pde}}\,\mathcal{L}_{\text{pde}} + \lambda_{\text{bc}}\,\mathcal{L}_{\text{bc}} + \lambda_{\text{ic}}\,\mathcal{L}_{\text{ic}}, \tag{7}$$

with each term computed as a mean squared residual:

$$\mathcal{L}_{\text{data}} = \frac{1}{N_d} \sum_{i=1}^{N_d} \big(\hat{u}(\xi_i; u_i) - y_i\big)^2, \tag{8}$$

$$\mathcal{L}_{\text{pde}} = \frac{1}{N_p} \sum_{j=1}^{N_p} F\big(\hat{u}(\,\cdot\,; u_j);\, x_j, y_j, t_j\big)^2, \tag{9}$$

$$\mathcal{L}_{\text{bc}} = \frac{1}{N_b} \sum_{l=1}^{N_b} \left(\frac{\partial \hat{u}(\,\cdot\,; u_l)}{\partial n}\bigg|_{(x_l, y_l, t_l)} \right)^2, \tag{10}$$

$$\mathcal{L}_{\text{ic}} = \frac{1}{N_i} \sum_{m=1}^{N_i} \big(\hat{u}(x_m, y_m, t_{\min}; u_m) - u_0(x_m, y_m)\big)^2, \tag{11}$$

where $\hat{u}(\xi; u)$ denotes the PI-DeepONet output at trunk query point $\xi = (x, y, t)$ conditioned on the branch input u (the wound-state encoding for that sample), and F is the same diffusion-decay-source-uptake residual defined in Eq. 5, evaluated on the trunk via two nested automatic-differentiation passes (one for $\partial_t \hat{u}$ and the spatial first derivatives, one for the second spatial derivatives in $\Delta\hat{u}$). The boundary loss enforces homogeneous Neumann conditions $\partial\hat{u}/\partial n = 0$ on the entire spatial boundary, and the initial loss matches the PI-DeepONet output at $t_{\min}$ to the ground-truth field u_0.

At each physics step, $N_p = 2000$ collocation points are sampled uniformly in $[-1, 1]^2 \times [t_{\min}, t_{\max}]$, $N_b = 200$ points are sampled on the spatial boundary (evenly distributed across the four sides), and $N_i = 500$ initial-condition points are sampled without replacement from the G^2 spatial locations at $t_{\min}$. Because the PDE residual depends on the branch input through $\hat{u}$, the physics terms must be evaluated for individual wound-state encodings rather than for the batch as a whole; we therefore draw a small random subset of $N_{\text{phys}} = 3$ training samples per physics step and average the resulting losses. Data and physics steps are interleaved within each epoch, so the supervised data loss and the physics losses are optimised jointly throughout training rather than in separate phases as in the stand-alone PINN.

The loss weights $\lambda_{\text{pde}}, \lambda_{\text{bc}}, \lambda_{\text{ic}}$ are tuned by Optuna over $[10^{-4}, 1]$, $[10^{-4}, 0.1]$, and $[0.1, 10]$ respectively (log-uniform), with the best configuration selected per cytokine and grid resolution. The selected values are reported in Table A.4 (Appendix Appendix A). The relative weighting strongly influences behaviour; balance is particularly sensitive for cytokines with sparse spatial

support, where the PDE residual evaluated at predominantly empty grid points can otherwise dominate the data term and bias the PI-DeepONet toward the trivial zero-field solution.

2.6. Performance metrics

Assessing model performance in this setting is not straightforward, as different metrics capture various aspects of prediction. In particular, agreement in overall magnitude does not necessarily imply that the spatial structure is well reproduced. The five metrics defined below are computed on predictions denormalised from the training domain $[-1, 1]$ back to physical concentration units, with non-physical negative values clipped to zero before evaluation.

The masked RMSE evaluates the squared error only at spatial locations where any cell agent is present. The active region at time t is defined as:

$$\Omega_{\text{active}}(t) = \left\{ j \;\middle|\; \max_{c} m_{j,t}^{(c)} = 1 \right\}, \tag{12}$$

where $m_{j,t}^{(c)} \in \{0, 1\}$ is the binary mask indicating the presence of cell type c at location j and time t, and the maximum is taken across the five tracked cell types (endothelial cells, resting/activated neutrophils, M1/M2 macrophages). This anatomical definition is independent of the predicted or ground-truth cytokine concentrations, and reflects the biological reality that cytokine activity is meaningful only where cells are present to produce or respond to it. The masked RMSE is then:

$$\text{Masked RMSE} = \sqrt{\frac{\sum_{t,j} (y_{j,t} - \hat{y}_{j,t})^2 \cdot \mathbf{1}[j \in \Omega_{\text{active}}(t)]}{\sum_{t,j} \mathbf{1}[j \in \Omega_{\text{active}}(t)] + \varepsilon}}, \tag{13}$$

with $\varepsilon = 10^{-12}$ for numerical stability. The sum runs jointly over space and time within the evaluation horizon. Because $\Omega_{\text{active}}(t)$ is fixed by the cell-mask input, a model cannot reduce its denominator by predicting empty fields; an underactive prediction enlarges the residual inside $\Omega_{\text{active}}(t)$ without changing the support over which the error is averaged. Throughout the paper, masked RMSE values are reported in absolute (physical) concentration units. Within a single column of the result tables (fixed cytokine, grid resolution, and evaluation horizon) all architectures are evaluated against the same ground truth and the same active region $\Omega_{\text{active}}(t)$, so per-column ranking between models is valid. Values are *not* comparable across columns, however: different cytokines differ by several orders of magnitude in dynamic range (Table 1),

different grid resolutions imply different per-cell concentration scales, and the normalisation constant $c_{\max}$ used to denormalise predictions also changes between configurations. Cross-column comparisons should therefore rely on the dimensionless metrics (R^2, SSIM, Dice, spatial correlation) rather than on raw RMSE values.

The Dice coefficient measures the overlap of predicted and true active regions after binarisation with a fixed physical threshold. Both fields are thresholded at 5% of the per-cytokine clipping value $c_{\max}$ used during normalisation:

$$\tilde{y}_{j,t} = \mathbf{1}[y_{j,t} > 0.05\, c_{\max}], \qquad \hat{\tilde{y}}_{j,t} = \mathbf{1}[\hat{y}_{j,t} > 0.05\, c_{\max}], \tag{14}$$

and the per-frame Dice is:

$$\mathrm{Dice}_t = \frac{2 \sum_j \tilde{y}_{j,t} \hat{\tilde{y}}_{j,t}}{\sum_j \tilde{y}_{j,t} + \sum_j \hat{\tilde{y}}_{j,t} + \varepsilon}. \tag{15}$$

The threshold is therefore absolute in physical units and global (the same value across all time-steps and locations within a given configuration), which avoids drift in the binarisation criterion as the field evolves and ensures consistency across cytokines whose concentrations differ by several orders of magnitude (Table 1). The choice of 5% of $c_{\max}$ provides a margin above numerical noise while remaining low enough to retain genuine low-amplitude features. Frames in which both $\tilde{y}_{\cdot,t}$ and $\hat{\tilde{y}}_{\cdot,t}$ are entirely empty are excluded from the time-step average, and the reported value is the mean of Dice_t over the retained time-steps in the evaluation horizon.

Structural similarity is assessed using the Structural Similarity Index:

$$\mathrm{SSIM}(y_t, \hat{y}_t) = \frac{(2\mu_y \mu_{\hat{y}} + C_1)(2\sigma_{y\hat{y}} + C_2)}{(\mu_y^2 + \mu_{\hat{y}}^2 + C_1)(\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2)}. \tag{16}$$

The implementation uses scikit-image [32] with its default uniform sliding window of size 7×7. While the original SSIM formulation [33] specifies a Gaussian window of standard deviation $\sigma = 1.5$, the uniform-window variant has been adopted as the standard in scientific computing pipelines and produces values that are consistently within a few thousandths of the Gaussian formulation for the field magnitudes encountered here [34]. The choice has no effect on the relative ordering of architectures, which is the central concern of this benchmark. The stabilisation constants are $C_1 = (k_1 L)^2$ and

$C_2 = (k_2 L)^2$ with $k_1 = 0.01$, $k_2 = 0.03$ (scikit-image defaults, matching [33]), and the data range L is fixed to the per-cytokine clipping value $c_{\max}$ rather than recomputed from each frame. Using a fixed L keeps SSIM scores comparable across time-steps and models; per-frame L values would otherwise rescale the constants and produce time-step-to-time-step instability. A frame is excluded from the SSIM average when its ground-truth dynamic range $\max_j y_{j,t} - \min_j y_{j,t}$ falls below 10^{-12}, which indicates a near-constant field for which SSIM is mathematically degenerate.

The spatial correlation coefficient is computed per time-step, with summations running over the spatial indices of a single frame:

$$\rho_t = \frac{\sum_j (y_{j,t} - \bar{y}_t)(\hat{y}_{j,t} - \bar{\hat{y}}_t)}{\sqrt{\sum_j (y_{j,t} - \bar{y}_t)^2 \cdot \sum_j (\hat{y}_{j,t} - \bar{\hat{y}}_t)^2}}, \tag{17}$$

where $\bar{y}_t$ denotes the spatial mean at time t and j runs over all $G \times G$ spatial locations of that frame. To produce a single horizon-level summary, per-frame correlations are aggregated using the Fisher z-transform: each ρ_t is mapped to $z_t = \frac{1}{2} \ln\big[(1+\rho_t)/(1-\rho_t)\big]$ (with ρ_t clipped to $[-0.9999, 0.9999]$ for numerical stability), the mean $\bar{z}$ is computed across time-steps, and the result is mapped back via $\bar{\rho} = \tanh(\bar{z})$. The Fisher transform stabilises the variance of correlation estimates close to ± 1, which is relevant here because well-performing surrogates routinely produce per-frame $\rho_t > 0.99$, where direct arithmetic averaging is biased [35]. Frames with degenerate variance ($\sigma_y < 10^{-12}$ or $\sigma_{\hat{y}} < 10^{-12}$) are excluded from the average.

The coefficient of determination (R^2) is reported as a global summary statistic computed jointly over space and time:

$$R^2 = 1 - \frac{\sum_{j,t} (y_{j,t} - \hat{y}_{j,t})^2}{\sum_{j,t} (y_{j,t} - \bar{y})^2}, \tag{18}$$

where the sums run over all spatial locations j and all time-steps t in the evaluation horizon, and $\bar{y}$ is the mean over the same space-time domain. This space-time pooled definition treats the prediction task as a single regression problem and matches the convention used in prior surrogate-modelling work on this ABM [7]. Per-time-step R^2 trajectories, computed by restricting the sums in Eq. 18 to a single frame, are reported separately in the temporal degradation analysis (Figure 4); frames in which $\sigma_{y_t} < 10^{-12}$ are reported as NaN to avoid undefined values. Although widely used, R^2 should be

interpreted alongside the spatial and structural metrics above, as a high R^2 obtained largely from agreement in the zero-background can coexist with poor reconstruction of the active region.

Each metric is computed within an evaluation horizon (near or far) for a fixed seed, then averaged across the three random seeds (1, 42, 100). Mean values appear in the main result tables; per-seed standard deviations (population, ddof = 0) are reported in Tables A.7 and A.8. Considering the metrics collectively provides a more complete picture of model performance: models that appear competitive in terms of RMSE often perform worse when evaluated using spatial or structural measures, motivating the multi-metric framework adopted throughout this work. This caveat applies in both directions for sparse fields: when most of the domain is zero, overlap metrics (Dice) become highly sensitive to the binarisation threshold and to small mislocalisations of the active region, while variance-based metrics (R^2, spatial correlation) can be inflated by agreement in the zero-background or, conversely, collapse when the ground-truth variance itself is small. Reporting the full set of metrics together is intended to expose, rather than average over, these regime-dependent behaviours.

2.7. Implementation

Experiments were conducted on the Dutch National Supercomputer Snellius [36]. Surrogate models were implemented in Python 3.11 using TensorFlow 2.15.1 with CUDA 12.1 on NVIDIA A100 GPUs. The PINN utilised the DeepXDE library with float64 precision. Hyperparameter optimisation was performed using Optuna 4.6.0 [37] with 20 trials per configuration (seed 42), and results were validated across two additional random seeds (1, 100) to assess reproducibility.

Training mini-batch sizes are as follows. For the U-Net and STA-LSTM, the batch size was tuned by Optuna over $\{4, 8, 16\}$ and applied to whole spatial frames; final-stage prediction was always performed with a fixed batch size of 2 to match the shape of the held-out test windows. For the DeepONet and PI-DeepONet, the operator-learning formulation requires two batching levels: the outer batch over wound-state inputs is tuned over $\{4, 8\}$, and within each input the trunk is evaluated in spatial chunks of $\{2048, 4096\}$ points to bound peak GPU memory. The PINN does not use mini-batches in the conventional sense: at each iteration, the full set of $N_d = 5\,000$ collocation points and $N_b = 1\,000$ boundary points is processed jointly, alongside the full

supervised observation set; this matches the standard DeepXDE training loop [15].

All architectures use convergence-based stopping rather than fixed-iteration training, so that reported training times reflect each model's actual time-to-convergence rather than a uniform compute budget. The U-Net and STA-LSTM use Keras `EarlyStopping` on validation loss with patience 20 and `restore_best_weights=True`, combined with `ReduceLROnPlateau` (patience 10, factor 0.5); the maximum-epoch ceilings of 200 are upper bounds that are rarely reached. The DeepONet and PI-DeepONet use a custom training loop with the same logic (patience 40 for early stopping, patience 15 for learning-rate reduction); the best validation weights are explicitly restored before evaluation, and the 400-epoch ceiling acts as an upper bound. For the PINN, the three-phase schedule (5 000 Adam warm-up, 10 000 Adam joint-loss iterations, 5 000 L-BFGS-B refinement) uses DeepXDE's `ModelCheckpoint` with `save_better_only=True`, and final evaluation is performed on the weights restored from the best validation step rather than the last step. Because each architecture stops when its own validation loss plateaus rather than after a fixed number of iterations, the relative training times in Tables 2 and 3 reflect convergence behaviour and are not confounded by an artificial cap on optimisation effort.

Reported training and prediction times follow a consistent wall-clock protocol across architectures. The training timer starts immediately before the first forward pass of final training (after data loading, preprocessing, and Optuna search are complete) and stops immediately after the final optimiser step, including all early-stopping, learning-rate-reduction, and checkpoint-restoration logic; for the PINN, this interval covers the full three-phase schedule (warm-up, joint Adam, L-BFGS-B). Prediction times are measured separately, starting before the first inference forward pass on the held-out test set and stopping after the last; they are reported as the mean over five repeated inference passes to dampen short-term variance, and they exclude metric computation, denormalisation, and any disk I/O. Both training and prediction timings exclude data loading, hyperparameter search, and post-processing such as JSON serialisation. All surrogate wall-clock measurements are taken on a single NVIDIA A100 GPU with no concurrent jobs and reflect the seed-42 run unless otherwise stated. The reference ABM runtimes used in Table A.6 and in the amortisation analysis of Section 3.4 (541 s at 100^2, 2503 s at 250^2, reported in the Datasets section) are wall-clock costs measured on the reference simulator configuration. Speed-up factors and

break-even evaluation counts reported in this paper are therefore wall-clock ratios under the deployments actually used; they should be interpreted as the per-evaluation cost reduction a downstream user would observe in practice, not as a like-for-like algorithmic comparison on identical hardware.

3. Results

Tables 2 and 3 summarise all five evaluation metrics for each architecture on the 100 h dataset (means across three random seeds), while per-seed variability is reported in Appendix Appendix A (Tables A.7 and A.8). Because R^2 saturates near unity for several models on the dense IL-8 field, structural metrics (SSIM, Dice) are given primary weight when claims of practical superiority are made; R^2 and Masked RMSE are used to confirm those claims and to flag cases where high pooled accuracy coexists with poor reconstruction of the active region. As noted in the Methods, masked RMSE values are tabulated in absolute physical units: within a single column they support direct ranking between architectures, but they should not be compared quantitatively across different cytokines, resolutions, or clipping configurations because the normalisation scale changes between columns.

3.1. Which architecture achieves the highest prediction accuracy?

Performance on the 100 h dataset (Tables 2 and 3) varies across models; however, a consistent ordering emerges when both cytokines are considered together. DeepONet remains at the top, with PI-DeepONet and U-Net forming a second tier, while STA-LSTM and PINN fall notably behind.

For IL-8, where the signal is effectively dense, R^2 values bunch in the 0.97-0.998 range across the leading models, so the structural metrics carry the discrimination. DeepONet attains the highest SSIM (0.972 at 100^2 near horizon) and Dice (0.984), with PI-DeepONet within the seed-level noise on both. U-Net trails by roughly 3-4 SSIM points and 4 Dice points at 100^2, recovering some of the gap at 250^2 but never matching DeepONet on structure. STA-LSTM's SSIM drops to 0.879 at 100^2 near horizon and degrades further at the far horizon, indicating that its local pattern reconstruction is markedly weaker than $R^2 = 0.974$ would suggest in isolation.

The separation is more evident for the remaining models. STA-LSTM at 100^2 near horizon reaches $R^2 = 0.974 \pm 0.003$, while PINN does not exceed $R^2 \approx 0.86$, despite requiring substantially longer optimisation times.

A different pattern appears for IL-10. Increasing sparsity (8.5-12.8%) results in a significantly wider performance spread, and the metric set must be read as a whole because sparsity affects each metric differently. Overlap metrics such as Dice are sensitive to the chosen binarisation threshold and to small spatial misregistrations of the active region; variance-based metrics (R^2, spatial correlation) can either be inflated by correct prediction of the dominant zero background or collapse when the ground-truth variance over a near-empty frame is small. Across both grids, DeepONet maintains $R^2 > 0.98$, and its Dice scores indicate that active regions are reliably identified. U-Net achieves comparable R^2 at lower resolution, but the markedly reduced Dice (0.778 at 100^2, Table 3) indicates that boundary definition is less precise.

The STA-LSTM demonstrates a less stable pattern, with a sharp drop in performance at 100^2, followed by a partial recovery at 250^2. In contrast, PINN fails to replicate the field's structure, with both R^2 and Dice remaining low. In this regime, the differences between models are no longer incremental but qualitative.

Moreover, the temporal extrapolation introduces an additional source of error. All models degrade at the far horizon, although the rate of degradation varies. DeepONet remains comparatively stable, while STA-LSTM exhibits a more pronounced decline. U-Net retains accuracy while sacrificing some spatial detail, situated between these extremes.

Across seeds 1, 42, and 100, with hyperparameters tuned on seed 42 and then fixed, the data-driven baselines (U-Net, STA-LSTM) and the DeepONet show low between-seed variability on the dimensionless metrics (Tables A.7 and A.8). The PI-DeepONet is a clear exception: its SSIM standard deviation reaches 0.136 on IL-10 at 250^2 (Table A.8), roughly two orders of magnitude above the DeepONet value at the same configuration (0.0012), and individual runs at the higher resolution show visibly poorer structural reconstruction even when R^2 remains close to unity. The PINN exhibits the largest between-seed variability overall, consistent with the optimisation-difficulty diagnosis in Section 4.2. Taken together, the main accuracy ordering is not driven by initialisation alone, but the physics-informed variants – and PI-DeepONet on sparse high-resolution targets in particular – carry a reliability cost that should be weighed against their marginal accuracy gains.

Table 2: IL-8, 100 h. Values are means across three random seeds (1, 42, 100). Best per column in bold. Tr = training time, Pr = prediction time. Masked RMSE is in absolute physical units; within-column values can be ranked between models, but across-column comparisons (different cytokines, resolutions, or clipping configurations) are not meaningful because the normalisation scale changes.

	100^2 Near					**100^2 Far**					**Time**	
	R^2	SSIM	Dice	Corr	RMSE	R^2	SSIM	Dice	Corr	RMSE	Tr	Pr
DeepONet	**0.9981**	**0.9718**	**0.9840**	**0.9996**	1.7e-8	**0.9924**	**0.9479**	**0.9643**	**0.9988**	2.8e-8	549 s	<1 s
PI-DeepONet	0.9969	0.9651	0.9760	0.9993	1.6**e-8**	0.9918	0.9418	0.9620	0.9978	2.3**e-8**	523 s	2 s
U-Net	0.9798	0.9378	0.9458	0.9901	7.1e-8	0.9762	0.9254	0.9463	0.9885	8.9e-8	39 s	<1 s
STA-LSTM	0.9737	0.8786	0.9505	0.9899	1.2e-7	0.9491	0.7847	0.9297	0.9861	1.5e-7	29 s	<1 s
PINN	0.8443	0.7919	0.9188	0.9189	4.5e-7	0.8488	0.7862	0.9204	0.9215	4.5e-7	2314 s	<1 s
	250^2 Near					**250^2 Far**					**Time**	
	R^2	SSIM	Dice	Corr	RMSE	R^2	SSIM	Dice	Corr	RMSE	Tr	Pr
DeepONet	**0.9984**	**0.9750**	**0.9902**	**0.9997**	4.8e-10	**0.9945**	**0.9305**	**0.9831**	**0.9994**	9.9**e-10**	2781 s	5 s
PI-DeepONet	0.9966	0.9482	0.9868	0.9996	4.6**e-10**	0.9848	0.8922	0.9684	0.9986	1.2e-9	3241 s	13 s
U-Net	0.9851	0.9274	0.9810	0.9928	4.3e-9	0.9832	0.9207	0.9820	0.9923	4.6e-9	68 s	<1 s
STA-LSTM	0.9590	0.8733	0.9851	0.9898	1.3e-8	0.9402	0.8367	0.9807	0.9890	1.4e-8	46 s	<1 s
PINN	0.7579	0.5768	0.9461	0.8709	2.4e-8	0.7720	0.5742	0.9487	0.8789	2.4e-8	2655 s	<1 s

RMSE = Masked RMSE. Corr = Spatial Correlation.

Table 3: IL-10, 100 h. Values are means across three random seeds (1, 42, 100). Best per column in bold. Tr = training time, Pr = prediction time. Masked RMSE is in absolute physical units; within-column values can be ranked between models, but across-column comparisons (different cytokines, resolutions, or clipping configurations) are not meaningful because the normalisation scale changes.

	100^2 Near					**100^2 Far**					**Time**	
	R^2	SSIM	Dice	Corr	RMSE	R^2	SSIM	Dice	Corr	RMSE	Tr	Pr
DeepONet	**0.9884**	**0.9901**	**0.9975**	**0.9956**	8.0**e-11**	**0.9810**	**0.9868**	**0.9965**	**0.9931**	1.0**e-10**	222 s	<1 s
PI-DeepONet	0.9860	0.9780	**0.9975**	0.9942	1.3e-10	0.9756	0.9700	0.9950	0.9902	1.5e-10	657 s	2 s
U-Net	0.9632	0.8965	0.7777	0.9832	1.3e-10	0.9429	0.8767	0.7432	0.9746	1.4e-10	40 s	<1 s
STA-LSTM	0.6227	0.6219	0.6804	0.8059	3.2e-10	0.4867	0.5618	0.6089	0.7415	4.0e-10	24 s	<1 s
PINN	0.1240	0.4030	0.2346	0.3714	7.1e-10	0.0678	0.3729	0.2177	0.3181	6.2e-10	2310 s	<1 s
	250^2 Near					**250^2 Far**					**Time**	
	R^2	SSIM	Dice	Corr	RMSE	R^2	SSIM	Dice	Corr	RMSE	Tr	Pr
DeepONet	**0.9988**	**0.9972**	0.9797	**0.9996**	4.5**e-13**	**0.9960**	**0.9931**	0.9343	**0.9987**	9.8**e-13**	2762 s	5 s
PI-DeepONet	0.9976	0.9024	**0.9910**	**0.9997**	8.4e-13	0.9910	0.8458	**0.9638**	**0.9989**	1.7e-12	3083 s	14 s
U-Net	0.9941	0.9538	0.9434	0.9973	1.2e-12	0.9900	0.9514	0.9281	0.9957	1.5e-12	159 s	3 s
STA-LSTM	0.8909	0.7925	0.8043	0.9447	5.3e-12	0.8573	0.7515	0.7953	0.9264	5.5e-12	43 s	<1 s
PINN	0.0226	0.1221	0.2512	0.1622	1.4e-11	0.0177	0.1148	0.2531	0.1548	1.5e-11	3011 s	<1 s

RMSE = Masked RMSE. Corr = Spatial Correlation.

The spatial implications of these performance metrics are illustrated in Figure 3.

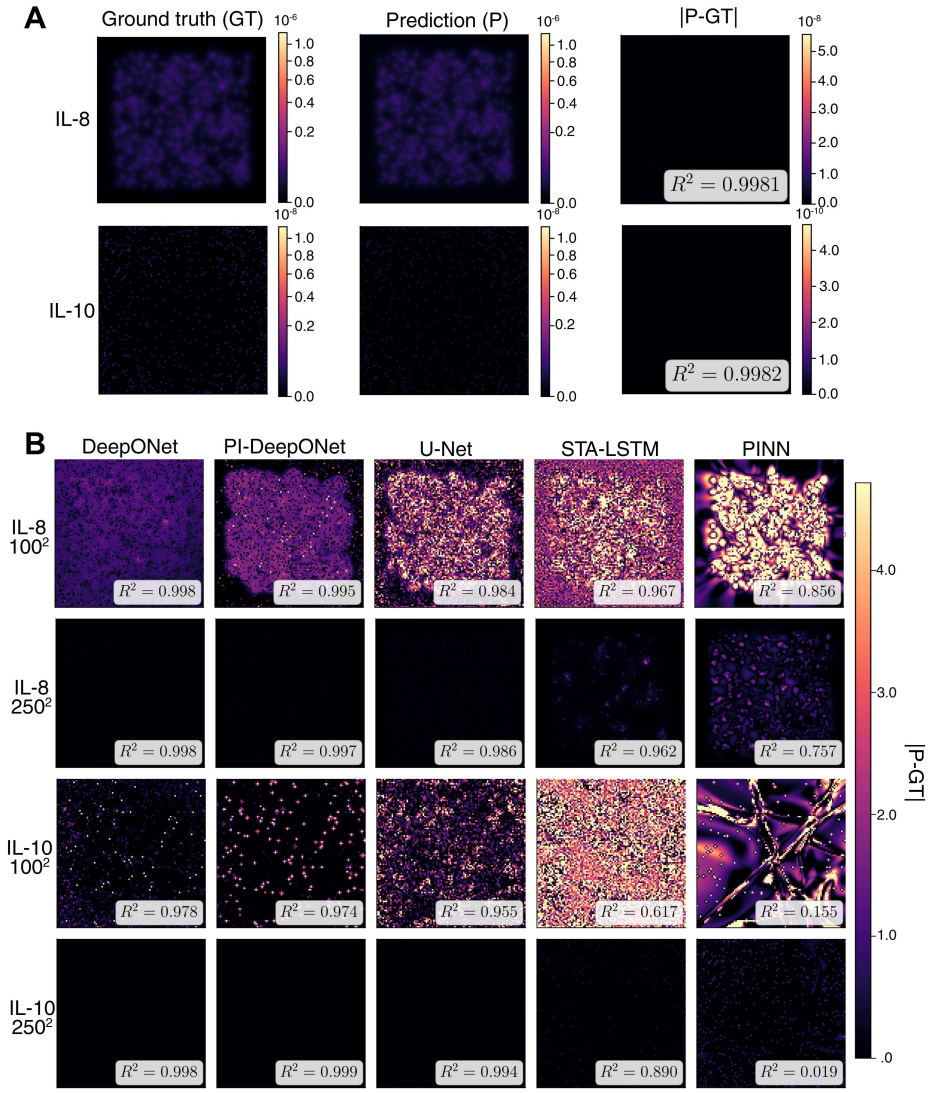


Figure 3: Spatial distribution of cytokine concentrations and prediction errors at $t = 88\,\mathrm{h}$ (on 100 hours of data). **(A)** Comparison of ground truth, DeepONet prediction, and absolute error for IL-8 and IL-10 at 250^2 resolution. **(B)** Absolute error maps ($|P - GT|$) across all models, both cytokines (IL-8 and IL-10) and grid resolutions. R^2 values (for the run with seed 42) for each specific configuration are indicated in the bottom corner of the respective panels.

3.2. Does physics-informed regularisation improve surrogate accuracy?

The introduction of physics-informed regularisation does not result in a consistent improvement in performance. For IL-8, DeepONet and PI-DeepONet yield essentially indistinguishable results, with differences falling within the observed variance, as shown in Figure 4.

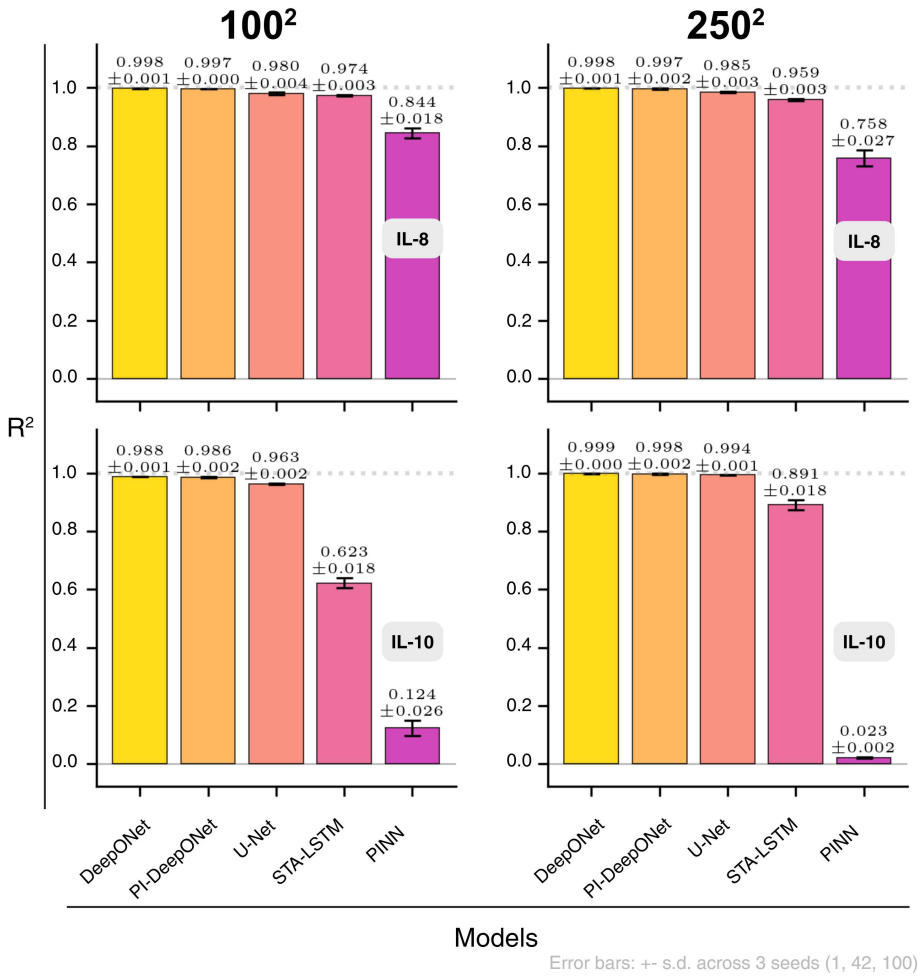


Figure 4: Global accuracy (R^2) for surrogate models across grid resolutions (100^2 and 250^2) for IL-8 (top) and IL-10 (bottom). Bars show three-seed means on the near-horizon evaluation window; error bars indicate ±1 standard deviation across seeds 1, 42, and 100 (see Tables A.7 and A.8).

Results for IL-10 are less consistent. Although PI-DeepONet slightly

outperforms DeepONet in one configuration (seed 42 at 250^2), this does not reflect the overall trend. Run-to-run variability is also noticeably worse: the PI-DeepONet SSIM standard deviation on IL-10 at 250^2 is 0.136 (Table A.8), roughly two orders of magnitude above the corresponding DeepONet value (0.0012), and individual seed runs occasionally collapse to structurally poor solutions even when pooled R^2 remains close to unity. Examination of these cases suggests that the additional loss terms can, in some seeds, guide the PI-DeepONet towards suboptimal solutions; the marginal accuracy gains of the physics-informed variant therefore come with a real reliability cost on sparse, high-resolution targets.

The standalone PINN behaves differently. Across all configurations, it remains substantially less accurate than the data-driven models. The discrepancy becomes most apparent for IL-10 at 250^2, where predictions degrade to near-zero R^2.

This can be attributed to a mismatch in representation. The PINN enforces a smooth reaction-diffusion process with static source terms, whereas the ABM generates cytokine dynamics through discrete, stochastic events. For IL-8, where the signal is spatially diffuse, the approximation remains acceptable. For IL-10, it fails, as isolated secretion events produce highly localised peaks that, on the scale of the grid, are effectively approximated as δ-like contributions and are not captured by a smooth PDE formulation.

3.3. Temporal degradation

Per-time-step R^2 trajectories on the 100 h test windows illustrate how prediction quality evolves with forecast distance (Figure 5). Across all architectures, accuracy at the near-horizon window remains close to its training-time level, with the most pronounced degradation occurring at the far horizon. DeepONet maintains the most stable trajectory across both cytokines, while STA-LSTM exhibits the steepest decline, consistent with the long-horizon drift previously reported for attention-augmented LSTMs on this ABM.

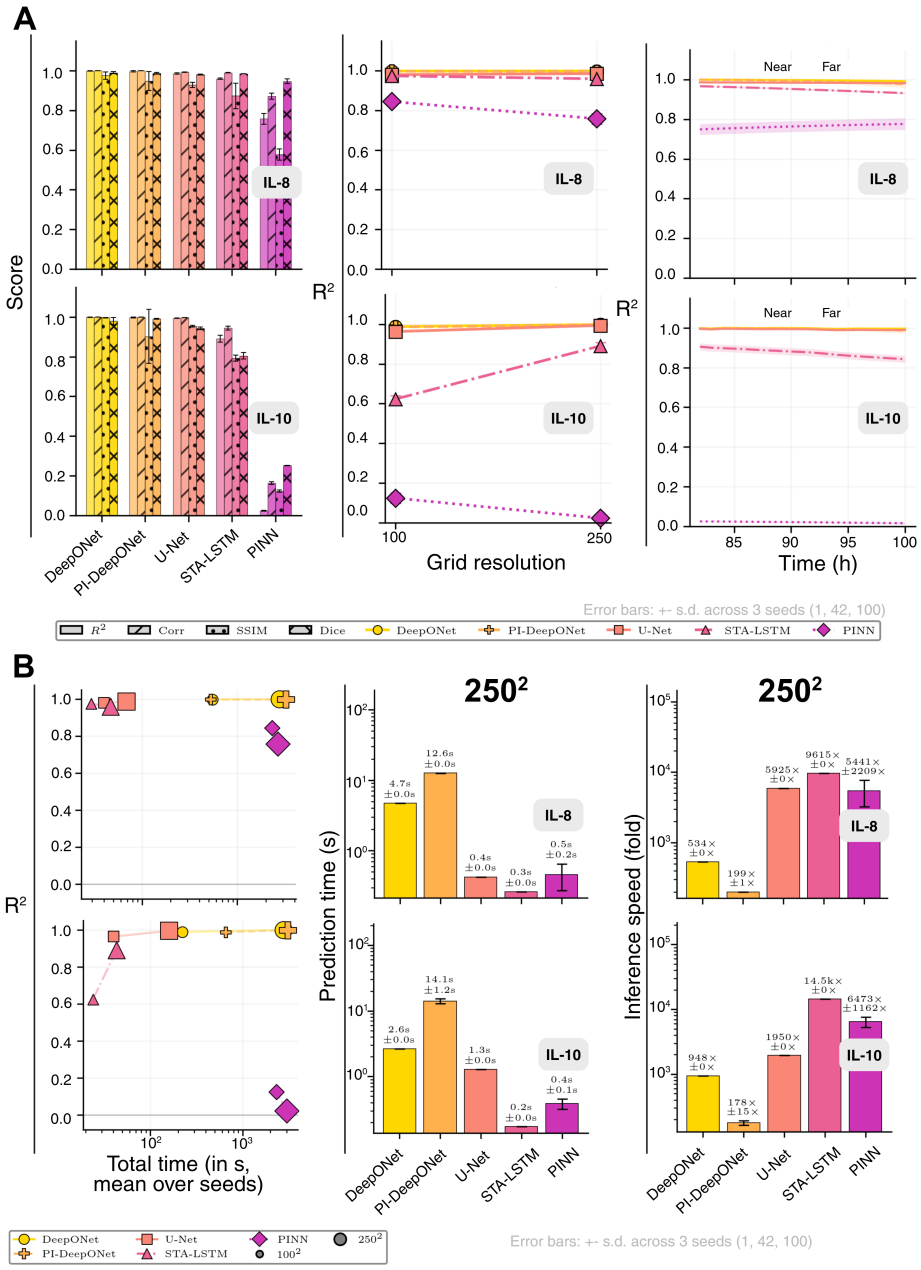


Figure 5: **(A)** Surrogate model evaluation across grid resolutions on the 100 h dataset. Panels show metric breakdowns and stability over time. Bar plots show three-seed means with error bars indicating ±1 standard deviation across seeds 1, 42, and 100; line plots show per-time-step R^2 trajectories with shaded bands representing ±1 standard deviation across the same three seeds. Detailed per-seed values are reported in Tables A.7 and A.8. **(B)** Accuracy-speed trade-off across architectures. R^2 values are three-seed means (seeds 1, 42, 100) on the near-horizon evaluation window; error bars represent ±1 standard deviation across seeds. The ABM reference runtime ($\sim$ 541 s at 100^2, $\sim$ 2503 s at 250^2) is indicated as a vertical reference line and reflects the wall-clock cost of the reference simulator configuration. Detailed numerical values, including break-even evaluation counts $N^\star$, are reported in Appendix Appendix A, Table A.6.

3.4. Accuracy-speed trade-off

The computational cost varies significantly between models. At 250^2, U-Net trains in 68 s, whereas DeepONet and PINN require 2781 s and 2655 s, respectively. PINN training scales particularly poorly with resolution.

Inference times show a different ordering. PINN is the fastest, but its low accuracy limits its practical usefulness. The operator-learning models are the slowest at inference – PI-DeepONet most of all (13.3 s at 250^2), followed by DeepONet (4.6 s) – reflecting the cost of evaluating the trunk network at each query point. U-Net and STA-LSTM achieve sub-second inference at moderate resolutions. When both accuracy and computational cost are considered, U-Net emerges as the most balanced option, offering strong predictive performance at relatively low cost (Figure 5).

Per-evaluation speed-ups omit training cost and therefore overstate the practical benefit when only a handful of evaluations are needed. A more useful break-even metric is the number of independent ABM evaluations $N^\star$ at which the total surrogate cost (training plus $N^\star$ inferences) equals the cost of running $N^\star$ ABM trajectories, computed as $N^\star = T_{\text{train}}/(T_{\text{ABM}} - T_{\text{infer}})$. The full per-architecture break-even counts are listed in Table A.6. At 250^2, the convolutional and recurrent baselines amortise almost immediately ($N^\star = 0.02$ for STA-LSTM, 0.05 for U-Net), DeepONet and PI-DeepONet break even after approximately one to one-and-a-quarter ABM trajectories ($N^\star = 1.11$ and 1.27), and the PINN reaches break-even after $\sim$1.1 trajectories but at a regime in which its accuracy is too low to recommend ($R^2 < 0.4$). The standalone PINN at 100^2 is the only configuration where break-even exceeds a few trajectories ($N^\star = 4.28$). In other words, the surrogate framing pays off in absolute wall-clock terms as soon as the downstream workflow plans more than a handful of independent evaluations, which is the regime of every use case motivating this work (sensitivity analysis, Bayesian calibration, patient-specific scenario exploration).

Detailed speed-up factors across grid resolutions are provided in Appendix Appendix A, Table A.6.

4. Discussion

The binding constraint on the next generation of surrogates for this ABM is not training data or compute, but the choice of inductive bias against the discrete, multi-scale structure of the underlying biology. The benchmark presented above produces two main outcomes that, taken together, define

what neural surrogates can and cannot do for the ABM studied here. First, operator-learning and convolutional models are accurate enough at 100^2-250^2 to be inserted into downstream inference and optimisation pipelines that the ABM alone is too slow to support, with DeepONet leading on accuracy and U-Net offering the more favourable accuracy-cost balance. Second, two failure modes – the underperformance of physics-informed variants and the systematic loss of fidelity on sparse cytokine fields - recur across architectures and resolutions, and both point to representational rather than data-related limits. The remainder of this section interprets these outcomes, explains why the failure modes arise, and discusses what they imply for surrogate design on hybrid agent-based models more broadly. Limitations of the benchmark itself are then addressed in Section 4.5.

4.1. Operator networks suit the ABM, but the U-Net baseline is the more balanced choice

The accuracy lead of DeepONet is consistent across metrics and resolutions, but the practical reading depends on how one weights accuracy against compute. The branch-trunk decomposition appears well matched to the ABM's structure: the branch encodes a low-dimensional summary of the wound state (cytokine envelope, cell-mask geometry, time), while the trunk evaluates the field at arbitrary coordinates. This decoupling lets the DeepONet represent both the global state of the wound and its local spatial structure without forcing one to dominate the loss. The gated trunk [30] adds high-frequency representational capacity at the cost of additional trunk-side parameters; removing it or removing the cytokine fields from the per-point input visibly smooths sparse features, suggesting both choices contribute meaningfully.

U-Net gives up only a small accuracy margin while training one to two orders of magnitude faster. For workflows in which the surrogate is retrained frequently – per-patient calibration, parameter sweeps with rolling re-fits – this is the more useful trade-off. The locality of U-Net's pixel-wise feature hierarchies, with skip connections preserving boundary detail, appears to scale naturally with grid resolution in the range tested, consistent with reports from cellular Potts surrogates that have adopted essentially the same encoder-decoder design unchanged [20]. STA-LSTM occupies a third position: it recapitulates the long-horizon drift and low-signal failure modes previously reported for attention-augmented LSTMs on this ABM [7], and

broadening the architectural comparison does not correct them. Two architectural families that were not benchmarked here are worth flagging because they target the failure modes identified below: Fourier neural operators [38], which achieve discretisation-convergent representations on regular grids but tend to suppress the high-frequency components that distinguish active from quiescent regions in sparse fields, and graph-network operators [39], which replace fixed grids with cell-adjacency graphs and are a closer structural match to an ABM driven by discrete cell-cell interactions; the latter are the family most likely to break the IL-10 ceiling that the grid-based methods evaluated here cannot.

4.2. Why physics-informed regularisation does not help

The expectation that encoding the ABM's underlying PDE would improve surrogate accuracy is not borne out. Two distinct causes compound here, and disentangling them clarifies what physics-informed methods can and cannot offer on problems of this kind.

The first is optimisation. PINN training is known to suffer from gradient pathologies, spectral bias, and ill-conditioned competition between data and physics losses [17, 18]. A critical methodological finding of this study, beyond these established issues, is the necessity of employing a softened residual in the physics loss (Section 2.5.4). This was not a procedural oversight but a deliberate architectural choice required to manage the fundamental discrepancy between continuum physics and agent-level stochasticity: in our ABM, cytokine sources are inherently discrete and point-like at the grid scale, and enforcing a strict residual against such discontinuous, high-frequency signals tends to produce gradient instabilities that prevent the optimiser from successfully fitting the supervised data. The softened residual acts as a milder regulariser that penalises non-physical configurations without attempting to force a smooth continuum solution onto a discrete generative process. The staged warm-up schedule used here is, similarly, a response to optimisation difficulties rather than a remedy for them; the additional optimisation cost buys only modest physical consistency in regions where the data loss already constrains the solution.

The PI-DeepONet results sharpen this point. Even with the operator-learning architecture absorbing most of the representational burden, the physics-informed variant shows roughly two orders of magnitude higher between-seed SSIM variability than the data-driven DeepONet on IL-10 at 250^2 (Table A.8), with individual seeds occasionally collapsing into structurally poor

solutions even when pooled R^2 remains high. This sensitivity to initialisation, despite identical data and identical hyperparameter budgets, indicates that the physics loss is interacting badly with the optimisation landscape rather than gently regularising it.

The second cause is model-form mismatch, and it is the more fundamental of the two. The continuum reaction-diffusion residual assumes smooth source distributions, deterministic kinetics, and length scales large compared to individual cells. The ABM produces fields through discrete, pulsed agent-level secretion events whose effective regularity is set by the cell radius. For dense cytokines such as IL-8, many overlapping secretion events combine into a near-continuous field, spatial averaging conceals the underlying discreteness, and the PDE remains a useful approximation; optimisation difficulties bound performance from above but the residual itself is not the limiting factor. For sparse cytokines such as IL-10, the situation reverses: isolated M2-macrophage secretion events behave on the grid as δ-like contributions, and the smooth residual cannot reconcile them with the data without violating one assumption or the other. The physics loss then actively biases the PINN toward smooth solutions that fit neither the discrete pulses nor the zero background well, and accuracy collapses.

This is a specific conclusion about the smooth reaction-diffusion template applied to a discrete generative process, not a general verdict on physics-informed methods. It is further qualified by the methodological choice flagged in Section 2.5.4: the PINN and PI-DeepONet evaluated here use a softened residual (the offset from the $[0, c_{\max}] \to [-1, 1]$ normalisation is dropped), not the strict rescaled residual of the original PDE. The empirical effect of the softened residual is to weaken the physics constraint relative to a faithful PINN, which could in principle change the magnitude of the gap observed against the data-driven baseline. We do not expect it to change the direction of the conclusion – the strict residual produces a larger systematic mismatch precisely in the IL-10 regime where the failure is concentrated – but a head-to-head comparison of softened and strict formulations on this ABM has not been performed and is identified in Section 4.5 as a target for follow-up work. Physics constraints remain useful when the encoded physics genuinely reflects the data-generating process [15, 16]; on hybrid ABMs of this kind, more structurally honest routes include learning effective PDEs from the ABM rather than imposing them [14], or multi-fidelity formulations that learn a relative weighting between cheap continuum approximations and expensive ABM observations [40].

4.3. Sparsity is a representational property, not a sample-size one

The IL-10 results expose a representational bottleneck that affects every grid-based architecture evaluated. A naive reading would attribute the gap to insufficient exposure to the sparse class; this is inconsistent with the observed pattern, however, because U-Net and DeepONet are trained on identical data yet produce qualitatively different reconstructions of the same active regions, and increasing resolution from 100^2 to 250^2 does not close the gap.

A more consistent reading is structural. IL-10's sparsity is not a sample-size artefact but a consequence of how the secretion is encoded in the grid representation. Most of the domain is zero, most of the loss budget is spent matching the zero background, and active regions are encoded as small clusters of pixels that any grid-based model must reconstruct by fitting the same kernels that also have to produce zeros elsewhere. The model is asked to do two incompatible things at once with the same parameters, and increasing the resolution of the grid only enlarges the zero background without adding representational capacity for the active regions. The methodological implication is that the IL-10 difficulty cannot be expected to resolve through more training data within the same architectural class. Architectures whose inductive bias matches the discrete, cell-driven origin of the signal – the graph-operator family flagged above is the natural candidate – are needed to break this ceiling, and this is identified here rather than implemented.

4.4. Speed-up enables a different class of analyses

The accuracy and speed axes have to be weighted together because they govern different downstream uses, and the most balanced architecture depends on which axis is binding. Two framings of the speed-up are worth keeping separate. The first is the per-evaluation ratio, which is in the 10^2-10^4 range and is the quantity most often quoted for surrogates of expensive simulators; it is the relevant cost once training is paid for. The second is the amortised cost, which folds training time into the comparison and is therefore the relevant cost for the user deciding whether to train a surrogate at all. Both convolutional and operator-learning surrogates amortise within at most a small number of ABM trajectories at the resolutions tested (Table A.6), and what makes the surrogate framing useful is less the headline ratio than the class of analyses it newly makes tractable. With per-evaluation costs reduced by two to four orders of magnitude relative to the simulator, global sensitivity analyses and Bayesian calibration against patient-level data move from impractical to routine. Approximate Bayesian computation [41] has

been applied to biological ABMs such as tumour-growth and neuronal-growth models [42, 43] but has been gated for spatially resolved cytokine ABMs by per-simulation cost; the surrogates here lift that gate. The same surrogates fit naturally into the Kennedy-O'Hagan calibration framework [44], which jointly accounts for parameter and model-form uncertainty, and integration into a digital-twin pipeline [12, 13, 45] would make patient-specific parameter inference a routine by-product of clinical observation rather than a separate, week-long computational task.

4.5. Limitations

Several aspects of the present benchmark qualify how broadly the findings should be applied.

Scope of what is surrogated. The surrogates evaluated here predict cytokine fields conditional on observed cell-mask inputs and do not model the agent-level dynamics themselves. Specifically, the following ABM components are not surrogated and remain inputs taken from ground-truth simulation: (i) cell motility governed by the Glazier-Graner-Hogeweg Hamiltonian (adhesion energies, volume and surface constraints, chemotactic terms), (ii) cell phenotype transitions (M1 to M2 macrophage polarisation, neutrophil resting to activated to necrotic), (iii) cell proliferation and apoptosis, and (iv) the closed-loop feedback from local cytokine concentrations onto cell behaviour. For the data-driven models, cell masks at the two look-back time-steps are concatenated as input channels; for the physics-informed models, the same masks are time-averaged and enter the PDE residual through the static source/uptake terms s_1, s_2, e. The current framework therefore approximates the reaction-diffusion phenomena "controlled" by the agents' positions at any given time, rather than the cell dynamics themselves. This hierarchical decomposition is a common and necessary step in building multi-scale surrogates for biological ABMs; a complete surrogate-based digital twin would require an additional architectural component to predict cell-level displacements and phenotype transitions. The present benchmark targets the forecasting scenario in which cell positions are known (e.g. from imaging or from a separate cell-population surrogate) and only cytokine concentrations need to be inferred.

Softened physics-informed residual. As detailed in Section 2.5.4, both the standalone PINN and the PI-DeepONet are trained against a softened residual that drops the $(\hat{u} + 1)$ offset arising from the $[0, c_{\max}] \to [-1, 1]$ rescaling. This is a deliberate design choice – the strict residual produces a

dominant systematic component in the predominantly empty parts of sparse cytokine fields that the data loss cannot recover from – but it does mean that the negative result about physics-informed methods (Sections 3.2 and 4.2) is established against a softened, rather than a strict, formulation. A faithful PINN with the strictly rescaled residual was not evaluated here. The expected direction of the strict-form result is a larger gap relative to the data-driven baselines, particularly on IL-10, because the additional systematic component compounds the model-form mismatch with the ABM's discrete generative process; an empirical confirmation – ideally including loss-curve diagnostics and gradient-pathology checks – is the cleanest follow-up to the present work.

Predictive uncertainty. None of the architectures evaluated here provides predictive uncertainty bands, which is a real gap for the clinical applications motivating the work – a confidently wrong prediction is more dangerous than a flagged ambiguous one. Three families of methods would address this with limited engineering effort: deep ensembles [46], MC dropout [47], and conformal prediction [48]. Each fits the present pipeline; evaluating calibration quality and reliability at the wound edge where errors concentrate (Figure 5) is a self-contained extension and is left to dedicated follow-up work.

Interpretability. All five architectures are black-box predictors in the sense that none provides a mechanistic decomposition of how cytokine values are computed from inputs. Post-hoc explainability tools (gradient attribution, integrated gradients, SHAP) are not applied here because, on spatially resolved cytokine fields, they typically produce saliency maps that highlight regions of high cell density without distinguishing genuinely mechanistic dependencies from collinearity in the inputs; for a benchmark whose primary goal is architectural comparison, attribution outputs are not informative enough to justify the additional methodological apparatus. Interpretability and uncertainty quantification together represent the two main translational gaps that downstream work will need to close before such surrogates can be used in decision-support contexts.

Geometry. The ABM is two-dimensional, whereas real burn wounds are volumetric, with depth-dependent cellular composition, vascular architecture, and oxygen gradients that a planar slice cannot represent without information loss [24]. Three-dimensional ABMs of skin tissue and epidermal wound healing exist [49], so whether the architectural ranking reported here transfers to volumetric domains is an open question rather than a closed one.

Resolution and cytokine coverage. The resolution range examined is

restricted to 100^2 and 250^2. Whether the DeepONet advantage observed here persists at substantially higher resolutions remains an open question that would require dedicated multi-seed experiments to settle; the branch-network's seven-scalar input summary may stop carrying enough information about the underlying function as spatial variability per grid point grows. Only IL-8 and IL-10 were examined in detail; the remaining four cytokines have intermediate sparsity and may behave differently. The ABM itself encodes a specific set of biological assumptions [5] that the surrogate inherits but does not validate.

Biological interpretation. Surrogate fidelity is strongly species-dependent in a way that tracks the spatial structure of secretion rather than biological importance. IL-8 is produced broadly by activated neutrophils and endothelial cells [24] and forms dense fields that all data-driven architectures capture well; IL-10 is secreted in localised pockets primarily by anti-inflammatory M2-polarised macrophages [1, 31], consistent with the cell-cluster pattern encoded in the ABM, although a quantitative correspondence to *in vivo* immunohistochemistry would require independent experimental data. The wound-edge regions where surrogate errors concentrate (Figure 5) coincide with areas that clinicians associate with conversion, the progression from partial- to full-thickness injury [2], so the regions that are hardest to predict are also the most clinically consequential. This connection is identified as future translational relevance rather than current evidence.

5. Conclusions and Future Work

The binding constraint on neural surrogates for this ABM is not training data or compute, but the choice of inductive bias against the discrete, multi-scale structure of the underlying biology. This work benchmarked five neural surrogate architectures – operator-learning, convolutional, recurrent, and physics-informed – for predicting spatio-temporal cytokine dynamics in a 2D agent-based model of the post-burn immune response, and three findings substantiate that claim. Operator networks produce the most accurate predictions in the resolution range tested, and amortise their training cost within at most a small number of ABM trajectories, so the surrogate framing already pays off in any workflow that plans more than a handful of independent evaluations. A convolutional baseline retains most of that accuracy at a fraction of the training cost, and is the more balanced option when both axes are weighted. Physics-informed regularisation, in the form used here, does

not deliver a consistent benefit on this problem – a smooth reaction-diffusion template cannot reconcile itself with a discrete, stochastic generative process, particularly for sparse cytokine fields whose secretion is encoded in localised, pulsed events. This conclusion is itself qualified: the physics-informed loss used here drops a normalisation-aware offset in the residual (Section 2.5.4) and so tests a softened physics constraint rather than the strict residual of the original PDE, and conclusions about physics-informed methods on this ABM should be understood within that scope.

This is, to our knowledge, the first head-to-head benchmark of these architectural families on a hybrid agent-based model of post-burn immune dynamics, and the protocol – chronological splits, multi-metric evaluation including structural fidelity, and joint reporting of per-evaluation and amortised cost – is intended to be reusable for future ABM surrogate studies in this and adjacent domains.

Closing the remaining gaps will require directions identified here rather than implemented. The most informative next step on the modelling side is a faithful PINN evaluation with the strict normalised residual, to settle the scope of the negative result about physics-informed methods on this ABM. Architectures whose inductive bias matches the discrete, cell-driven origin of the signal – in particular graph-network operators that encode locality through the cell-adjacency graph and pay no representational tax for empty regions [39] – are the most promising route to the IL-10 sparsity ceiling, with adaptive-resolution and multi-fidelity formulations [40] as complementary alternatives. Uncertainty quantification through deep ensembles, MC dropout, or conformal calibration is a prerequisite for the digital-twin and triage applications motivating this work. Extending the framework to three-dimensional geometries [49] and integrating the surrogate into parameter-inference pipelines such as SMoRe ParS [50, 51] that explicitly map ABM inputs to clinical observables would close the remaining gap toward translational deployment.

Looking further out, surrogates of the kind benchmarked here move the bottleneck for mechanism-aware burn modelling away from compute and toward biology. Once each evaluation costs milliseconds rather than minutes, the practical question stops being "can we afford to simulate this patient?" and becomes "which model components do we still need to refine to make these predictions clinically actionable?". That is the question this benchmark is intended to enable, and the one we believe future work in this area should be organised around.

Appendix A. Supplementary results

Appendix A.1. Hyperparameter configuration

Table A.4: Architecture, hyperparameters, and training configuration (100 h dataset). † = tuned with Optuna (20 trials, seed 42). B = Branch network, T = Trunk network. G denotes the grid size.

	U-Net	**STA-LSTM**	**DeepONet**	**PI-DeepONet**	**PINN**
Architecture					
Input	$(G, G, 22)$	$(2, G, G, 11)$	B:(7); T:$(G^2, 24)$	B:(7); T:$(G^2, 24)$	(x, y, t)
Output	$(G, G, 1)$	$(G, G, 1)$	$(G^2, 1)$	$(G^2, 1)$	(1)
Activation	ReLU	ReLU + sigmoid	B:ReLU; T:tanh	B:ReLU; T:tanh	tanh
Gated layers	-	-	2 Hadamard	2 Hadamard	-
Tuned hyperparameters†					
Filters/hidden	{16,32,64}	{32,64}	{128,256}	{128,256}	{50,64,128}
Depth/layers	{2,3,4}	-	2+2	2+2	{3,4,5}
Latent p	-	-	{64,128,256}	{64,128,256}	-
Learning rate	$[10^{-5}, 10^{-3}]$	$[10^{-5}, 10^{-3}]$	$[10^{-5}, 10^{-3}]$	$[10^{-5}, 10^{-3}]$	$[10^{-4}, 10^{-3}]$
Training					
Loss	MSE	MSE	Masked MSE	Data+PDE+BC+IC	Data+PDE+BC+IC
Max epochs	200	200	400	400	20 000*
Optimiser	Adam	Adam	Adam	Adam	Adam + L-BFGS
Precision	float32	float32	float32	float32	float64

*PINN schedule: 5k warm-up (data only) + 10k physics (Adam) + 5k L-BFGS.

Appendix A.2. U-Net look-back window sensitivity

Table A.5: U-Net look-back window sensitivity at 250^2 (seed 42). Bold rows mark the best far-horizon R^2 per cytokine. Parameter counts and training times correspond to the Optuna-selected configuration for each window.

Cyt.	w	Params	Time (s)	Near R^2	Far R^2	Near SSIM	Far SSIM	Near Dice	Far Dice
IL-8	1	486 001	62.87	0.9814	0.9803	0.9137	0.9032	0.9757	0.9760
IL-8	2	487 585	70.26	0.9860	0.9862	0.9127	0.9098	0.9794	0.9815
IL-8	3	489 169	73.93	0.9869	0.9879	0.9098	0.9097	0.9809	0.9820
IL-8	**4**	**31 078 913**	**146.43**	**0.9905**	**0.9897**	**0.9264**	**0.9099**	**0.9847**	**0.9853**
IL-8	5	1 946 209	107.07	0.9702	0.9698	0.8429	0.8325	0.9612	0.9620
IL-8	10	31 116 929	176.44	0.9870	0.9821	0.8858	0.8409	0.9769	0.9722
IL-10	1	31 059 905	131.07	0.9894	0.9821	0.8827	0.8636	0.8777	0.8368
IL-10	**2**	**31 066 241**	**146.13**	**0.9940**	**0.9907**	**0.9388**	**0.9364**	**0.9338**	**0.9230**
IL-10	3	489 169	66.93	0.9878	0.9810	0.8778	0.8652	0.9082	0.9002
IL-10	4	31 078 913	144.77	0.9918	0.9866	0.9180	0.9082	0.9081	0.8719
IL-10	5	31 085 249	188.89	0.9826	0.9702	0.8257	0.8135	0.8205	0.7715
IL-10	10	31 116 929	177.47	0.9926	0.9847	0.9111	0.8840	0.9260	0.8576

Appendix A.3. Speed-up and amortisation

Table A.6: Inference performance and amortisation across spatial scales (100^2 and 250^2). R^2 values (mean ± standard deviation) are aggregated across the three random seeds, both cytokines, and both evaluation horizons. Training times are means across seeds 1, 42, 100 and over the two cytokines per architecture. Prediction times are wall-clock measurements as defined in Section 2 (Implementation), averaged over the two cytokines and over the available timed seeds (three for PI-DeepONet and PINN; seed 42 for DeepONet, U-Net, and STA-LSTM, for which only the seed-42 inference pass was timed). Speed-up is the per-evaluation wall-clock ratio between the reference ABM (541 s at 100^2, 2503 s at 250^2) and the trained surrogate; it does not include training cost. Break-even is the number of independent ABM evaluations at which the total surrogate cost (training + N inferences) equals the cost of running N ABM trajectories, computed as $N^\star = T_{\text{train}}/(T_{\text{ABM}} - T_{\text{infer}})$. Below $N^\star$ the ABM is cheaper end-to-end; above it the surrogate is. Reported values are wall-clock ratios under the deployments actually used in this study and should be read as the practical cost reduction a user would observe, not as a like-for-like algorithmic comparison.

Grid	**Model**	R^2 **(Mean ± SD)**	**Train. Time (s)**	**Pred. Time (s)**	**Speed-up**	**Break-even** $N^\star$
100^2	DeepONet	0.990 ± 0.007	385	0.822	658×	0.71
	PI-DeepONet	0.988 ± 0.009	590	2.397	226×	1.10
	U-Net	0.966 ± 0.016	40	0.179	3 022×	0.07
	STA-LSTM	0.758 ± 0.219	26	0.100	5 421×	0.05
	PINN	0.471 ± 0.394	2 312	0.168	3 214×	4.28
250^2	DeepONet	0.997 ± 0.003	2 771	4.640	539×	1.11
	PI-DeepONet	0.992 ± 0.010	3 162	13.330	188×	1.27
	U-Net	0.988 ± 0.005	113	1.522	1 644×	0.05
	STA-LSTM	0.912 ± 0.044	45	0.249	10.1k×	0.02
	PINN	0.393 ± 0.390	2 833	0.423	5 913×	1.13

Appendix A.4. Seed stability analysis

Table A.7: Seed variability for IL-8: standard deviation across seeds 1, 42, 100 (population, ddof = 0; near horizon, 100 h dataset). Lower values indicate higher reproducibility.

	100^2				250^2			
	R^2	SSIM	Dice	Corr	R^2	SSIM	Dice	Corr
DeepONet	0.0007	0.0168	0.0048	0.0002	0.0009	0.0183	0.0064	0.0002
PI-DeepONet	0.0005	0.0194	0.0040	0.0002	0.0025	0.0476	0.0044	0.0001
U-Net	0.0042	0.0142	0.0038	0.0021	0.0031	0.0130	0.0019	0.0013
STA-LSTM	0.0029	0.0215	0.0053	0.0011	0.0034	0.0641	0.0004	0.0015
PINN	0.0177	0.0292	0.0026	0.0096	0.0272	0.0285	0.0123	0.0155

Masked RMSE omitted for brevity. Corr = Spatial Correlation.

Table A.8: Seed variability for IL-10: standard deviation across seeds 1, 42, 100 (population, ddof = 0; near horizon, 100 h dataset). Lower values indicate higher reproducibility.

	100^2				250^2			
	R^2	SSIM	Dice	Corr	R^2	SSIM	Dice	Corr
DeepONet	0.0006	0.0033	0.0000	0.0005	0.0005	0.0012	0.0185	0.0002
PI-DeepONet	0.0020	0.0186	0.0000	0.0008	0.0021	0.1356	0.0013	0.0001
U-Net	0.0021	0.0290	0.0929	0.0013	0.0007	0.0052	0.0070	0.0003
STA-LSTM	0.0177	0.1204	0.2252	0.0276	0.0179	0.0159	0.0168	0.0097
PINN	0.0262	0.0398	0.0384	0.0310	0.0022	0.0055	0.0020	0.0067

Masked RMSE omitted for brevity. Corr = Spatial Correlation.

Appendix A.5. Clipping sensitivity analysis

The kurtosis-adaptive clipping schedule of Eq. 2 was introduced as a numerical-stability device for heavy-tailed cytokine distributions, but a downstream concern is that the schedule could implicitly favour the operator-learning architectures whose loss formulation it was originally tuned against. To address this, a sensitivity sweep was performed in which the entire pipeline – preprocessing, Optuna hyperparameter search, and final training – was rerun at 50^2 resolution (the most challenging regime for all surrogates, with the fewest active grid points per cytokine signal) for both cytokines (IL-8, IL-10) and a single seed (42), under three preprocessing variants:

- **no-clip**: $c_{\max}$ set to the true per-channel maximum on the training trajectory; no values are clipped. This is the null hypothesis that any clipping is unnecessary.
- **p95**: $c_{\max}$ set to the fixed 95th percentile of the per-channel training trajectory, applied uniformly across all six cytokines. This is deliberately more aggressive than the adaptive schedule for dense cytokines (where it discards genuine dynamics) and tests whether the architectural ranking is robust to over-suppression of high-amplitude features.
- **adaptive**: the kurtosis-driven schedule used throughout the main paper (Eq. 2), reported here as reference. Note that the absolute R^2 values at 50^2 are not directly comparable to the $100^2/250^2$ numbers in Tables 2 and 3 because the grid resolution itself was reduced for this experiment.

Three clipping strategies × five architectures × two cytokines × one seed = 30 independent training runs, each with its own Optuna hyperparameter

search (20 trials per run). For consistency with the main analysis the dimensionless metrics (R^2, SSIM, Dice, spatial correlation) are reported across both evaluation horizons (near $t = 82$-91 h, far $t = 92$-100 h). Masked RMSE is omitted from this experiment because each clipping strategy changes the normalisation constant $c_{\max}$ and therefore the absolute concentration scale to which predictions are denormalised, making cross-row RMSE comparisons within the same table mathematically inconsistent.

Tables A.9 and A.10 report full per-metric results. Three observations summarise the outcome:

DeepONet and PI-DeepONet occupy the top two positions for every (cytokine, horizon, clipping) combination tested. The relative ordering of the two within the top tier interchanges occasionally between configurations, but always within seed-level noise ($\Delta R^2 < 0.005$). This is the central finding: the architectural recommendation of the main paper (operator networks for this ABM) does not depend on the choice of clipping strategy.

The IL-8 ranking of U-Net, STA-LSTM and PINN does change between clippings. Specifically: under *adaptive* clipping the ordering is U-Net $\geq$ STA-LSTM $\geq$ PINN for IL-8 at 50^2, but under *no-clip* the ordering becomes U-Net $>$ PINN $>$ STA-LSTM, and under *p95* it becomes PINN $>$ STA-LSTM $>$ U-Net. All three configurations cluster within an R^2 band narrower than 0.1, so the reshuffling reflects movement of close competitors rather than a qualitative change in performance. Note that this band is itself smaller than the gap separating the top two architectures from the remainder, which preserves the main paper's qualitative conclusion.

For IL-10 the architectural ordering DeepONet $\geq$ PI-DeepONet $\gg$ U-Net $>$ STA-LSTM $\gg$ PINN is reproduced exactly under all three clipping strategies, on both horizons. The standalone PINN remains the clear failure mode ($R^2 < 0.4$ across all configurations) and the operator networks remain the clear winners ($R^2 > 0.97$ across all configurations), confirming that the sparse-signal limitation discussed in Section 4.3 is a structural property of the architectures rather than an artefact of the preprocessing choice.

The *p95* variant produces uniformly higher R^2 for all architectures than the other two clippings. This is a known consequence of aggressive clipping: $c_{\max}$ is reduced, the target variance shrinks, and the proportion of error explained relative to the residual rises mechanically. The effect operates symmetrically across architectures (all five gain a comparable amount) and therefore does not constitute evidence that *p95* is the *better* strategy; what matters for the sensitivity check is that the relative ranking among

architectures is conserved, which it is.

Taken together, these results address the concern that the kurtosis-adaptive schedule could be implicitly favouring the operator-learning architectures: the top-two ranking holds under all three preprocessing strategies (including the deliberately aggressive *p95* variant), and the only sensitivity observed concerns mid-rank architectures that are not central to the paper's main recommendation.

Table A.9: Clipping sensitivity at 50^2 (seed 42) for IL-8. Four dimensionless metrics under three clipping strategies. *no-clip* retains the true per-channel maximum; *p95* applies a fixed 95th-percentile clip to every cytokine; *adaptive* is the kurtosis-driven schedule used throughout the main paper (Eq. 2). Masked RMSE is omitted because each clipping strategy changes $c_{\max}$ and therefore the absolute concentration scale, making within-table comparisons inconsistent.

		Near horizon				**Far horizon**			
Model	**Clip**	R^2	SSIM	Dice	Corr	R^2	SSIM	Dice	Corr
DeepONet	no-clip	0.999	0.976	0.989	1.000	0.997	0.967	0.971	0.999
	p95	0.999	0.963	0.981	1.000	0.999	0.956	0.969	1.000
	adaptive	0.999	0.991	0.989	1.000	0.998	0.985	0.972	0.999
PI-DeepONet	no-clip	0.999	0.995	0.983	1.000	0.994	0.977	0.941	0.999
	p95	1.000	0.997	0.990	1.000	0.999	0.994	0.983	1.000
	adaptive	0.998	0.985	0.979	0.999	0.993	0.980	0.947	0.998
U-Net	no-clip	0.988	0.972	0.930	0.994	0.979	0.958	0.907	0.990
	p95	0.949	0.853	0.838	0.978	0.914	0.817	0.815	0.965
	adaptive	0.891	0.776	0.744	0.947	0.846	0.736	0.728	0.924
STA-LSTM	no-clip	0.714	0.674	0.814	0.954	0.679	0.648	0.786	0.951
	p95	0.961	0.911	0.970	0.988	0.930	0.880	0.963	0.979
	adaptive	0.959	0.967	0.917	0.986	0.943	0.952	0.892	0.984
PINN	no-clip	0.925	0.908	0.878	0.963	0.911	0.892	0.874	0.956
	p95	0.993	0.992	0.986	0.997	0.986	0.983	0.974	0.993
	adaptive	0.976	0.971	0.918	0.988	0.966	0.961	0.909	0.984

Table A.10: Clipping sensitivity at 50^2 (seed 42) for IL-10. Four dimensionless metrics under three clipping strategies. *no-clip* retains the true per-channel maximum; *p95* applies a fixed 95th-percentile clip to every cytokine; *adaptive* is the kurtosis-driven schedule used throughout the main paper (Eq. 2). Masked RMSE is omitted because each clipping strategy changes $c_{\max}$ and therefore the absolute concentration scale, making within-table comparisons inconsistent.

		Near horizon				**Far horizon**			
Model	**Clip**	R^2	SSIM	Dice	Corr	R^2	SSIM	Dice	Corr
DeepONet	no-clip	0.983	0.963	0.997	0.998	0.982	0.993	0.997	0.996
	p95	0.998	0.994	0.995	1.000	0.998	0.997	0.992	1.000
	adaptive	0.989	0.987	0.997	0.999	0.985	0.992	0.997	0.999
PI-DeepONet	no-clip	0.982	0.974	0.997	0.997	0.975	0.985	0.997	0.996
	p95	0.997	0.979	0.996	0.999	0.996	0.978	0.976	0.999
	adaptive	0.988	0.987	0.997	0.998	0.982	0.989	0.997	0.996
U-Net	no-clip	0.909	0.812	0.824	0.959	0.824	0.744	0.686	0.918
	p95	0.943	0.852	0.607	0.976	0.908	0.842	0.576	0.962
	adaptive	0.804	0.653	0.286	0.907	0.714	0.611	0.267	0.858
STA-LSTM	no-clip	0.549	0.714	0.736	0.742	0.476	0.678	0.716	0.696
	p95	0.886	0.808	0.688	0.943	0.837	0.765	0.673	0.918
	adaptive	0.557	0.633	0.570	0.748	0.479	0.588	0.551	0.700
PINN	no-clip	0.021	0.305	0.173	0.171	−0.014	0.289	0.142	0.131
	p95	0.388	0.562	0.374	0.744	−0.536	0.460	0.342	0.536
	adaptive	0.082	0.448	0.367	0.478	−0.114	0.401	0.341	0.404

CRediT authorship contribution statement

Roland V. Bumbuc: Conceptualization, Methodology, Software, Investigation, Data curation, Validation, Project administration, Resources, Visualization, Writing - review and editing, Supervision

Julia Kowalczuk: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing - original draft, Writing - review and editing

Rezo Getsadze: Visualization, Writing - review and editing

Ioannis Papapanagiotou: Conceptualization, Software, Writing - review and editing

H. Ibrahim Korkmaz: Conceptualization, Validation, Supervision, Funding acquisition, Writing - review and editing

Vivek M. Sheraton: Conceptualization, Methodology, Supervision, Project administration, Data curation, Writing - review and editing

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Claude Opus 4.8 (Anthropic) to verify compliance with journal guidelines, refine LaTeX formatting, and assist with consistency checks. After using this tool, the authors reviewed and edited the content as necessary and take full responsibility for the content of the published article.

Data availability statement

The code, data and necessary scripts to reproduce the results in this research article are available in Zenodo [52].

Ethics Statement

This research does not require an ethics statement as it involves only computational simulation and the use of an agent-based model. No human subjects, animal experiments, or clinical trials were involved in this study.

Acknowledgments

The authors declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Dutch Burns Foundation (DBF) - Health-Holland, Beverwijk, project PPS 22.01. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-18360.

References

[1] H. I. Korkmaz, G. Flokstra, M. Waasdorp, A. Pijpe, S. G. Papendorp, E. de Jong, T. Rustemeyer, S. Gibbs, P. P. van Zuijlen, The complexity of the post-burn immune response: an overview of the associated local and systemic complications, Cells 12 (2023) 345. doi:10.3390/cells12030345.

[2] A. Wardhana, et al., Predicting mortality in burn patients: literature review of risk factors for burn mortality and application in scoring systems, Annals of Burns and Fire Disasters 36 (2023) 3–10. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11044732/.

[3] G. Roberts, M. Lloyd, M. Parker, R. Martin, B. Philp, O. Shelley, P. Dziewulski, The baux score is dead. Long live the Baux score: a 27-year retrospective cohort study of mortality at a regional burns service, Journal of Trauma and Acute Care Surgery 72 (2012) 251–256. doi:10.1097/TA.0b013e31824052bb.

[4] World Health Organization, Burns: Fact sheet, https://www.who.int/news-room/fact-sheets/detail/burns, 2024. Accessed: 2025-02-14.

[5] H. I. Korkmaz, V. M. Sheraton, R. V. Bumbuc, M. Li, A. Pijpe, P. P. Mulder, B. K. Boekema, E. de Jong, S. G. Papendorp, R. Brands, et al., An in silico modeling approach to understanding the dynamics of the post-burn immune response, Frontiers in Immunology 15 (2024) 1303776. doi:10.3389/fimmu.2024.1303776.

[6] F. Chiacchio, M. Pennisi, G. Russo, S. Motta, F. Pappalardo, Agent-based modeling of the immune system: Netlogo, a promising framework, BioMed Research International 2014 (2014) 907171. doi:10.1155/2014/907171.

[7] I. Papapanagiotou, R. V. Bumbuc, H. I. Korkmaz, V. Krzhizhanovskaya, V. M. Sheraton, From simulations to surrogates: Neural networks enhancing burn wound healing predictions, Journal of Computational Science 89 (2025) 102593. doi:10.1016/j.jocs.2025.102593.

[8] S. Husanovic, G. Egberts, A. Heinlein, F. Vermolen, Deep operator network models for predicting post-burn contraction, Clinical Biome-

chanics (2025). doi:10.1016/j.clinbiomech.2025.106558, special issue: Wound care technologies.

[9] G. Egberts, F. Vermolen, P. van Zuijlen, High-speed predictions of post-burn contraction using a neural network trained on 2d-finite element simulations, Frontiers in Applied Mathematics and Statistics 9 (2023) 1098242. doi:10.3389/fams.2023.1098242.

[10] K.-A. Norton, D. Bergman, H. V. Jain, T. Jackson, Advances in surrogate modeling for biological agent-based simulations: Trends, challenges, and future prospects, arXiv preprint arXiv:2504.11617 (2025). doi:10.48550/arXiv.2504.11617.

[11] C. Angione, E. Silverman, E. Yaneske, Using machine learning as a surrogate model for agent-based simulations, PLOS ONE 17 (2022) e0263150. doi:10.1371/journal.pone.0263150.

[12] L. L. Fonseca, L. Böttcher, B. Mehrad, R. C. Laubenbacher, Surrogate modeling and control of medical digital twins, arXiv preprint arXiv:2402.05750 (2024). doi:10.48550/arXiv.2402.05750, pubMed: PMID 38827450; PMC11142319.

[13] L. L. Fonseca, L. Böttcher, B. Mehrad, R. C. Laubenbacher, Optimal control of agent-based models via surrogate modeling, PLOS Computational Biology 21 (2025) e1012138. doi:10.1371/journal.pcbi.1012138.

[14] K. Burrage, P. M. Burrage, J. N. Kreikemeyer, A. M. Uhrmacher, H. N. Weerasinghe, Learning surrogate equations for the analysis of an agent-based cancer model, Frontiers in Applied Mathematics and Statistics 11 (2025) 1578604. doi:10.3389/fams.2025.1578604.

[15] M. Raissi, P. Perdikaris, G. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics 378 (2019) 686–707. doi:10.1016/j.jcp.2018.10.045.

[16] C. Rao, P. Ren, Q. Wang, O. Buyukozturk, H. Sun, Y. Liu, Encoding physics to learn reaction-diffusion processes, arXiv (2021). doi:10.48550/arXiv.2106.04781.

[17] S. Wang, Y. Teng, P. Perdikaris, Understanding and mitigating gradient flow pathologies in physics-informed neural networks, SIAM Journal on Scientific Computing 43 (2021) A3055–A3081. doi:10.1137/20M1318043.

[18] A. S. Krishnapriyan, A. Gholami, S. Zhe, R. M. Kirby, M. W. Mahoney, Characterizing possible failure modes in physics-informed neural networks, in: Advances in Neural Information Processing Systems, volume 34, 2021, pp. 26548–26560. URL: https://papers.nips.cc/paper/2021/hash/df438e5206f31600e6ae4af72f2725f1-Abstract.html.

[19] P. Chavoshnejad, L. Chen, X. Yu, J. Hou, N. Filla, D. Zhu, T. Liu, G. Li, M. J. Razavi, X. Wang, An integrated finite element method and machine learning algorithm for brain morphology prediction, Cerebral Cortex 33 (2023) 9354–9366. doi:10.1093/cercor/bhad208.

[20] T. Comlekoglu, J. Q. Toledo-Marín, T. Comlekoglu, D. W. DeSimone, S. M. Peirce, G. Fox, J. A. Glazier, Surrogate modeling of cellular-potts agent-based models as a segmentation task using the u-net neural network architecture, PLOS Computational Biology 21 (2025) e1013626. doi:10.1371/journal.pcbi.1013626.

[21] R. V. Bumbuc, H. I. Korkmaz, P. van Zuijlen, L. Vermeulen, V. M. Sheraton, Understanding the dynamics of the proliferative phase in local burn wound healing: A computational model, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2023, pp. 3676–3683. doi:10.1109/BIBM58861.2023.10385875.

[22] K. Shukla, V. Oommen, A. Peyvan, M. Penwarden, N. Plewacki, L. Bravo, A. Ghoshal, R. M. Kirby, G. E. Karniadakis, Deep neural operators as accurate surrogates for shape optimization, Engineering Applications of Artificial Intelligence 129 (2024) 107615. doi:10.1016/j.engappai.2023.107615.

[23] J. Tang, P. T. Brewick, R. G. McClarren, C. Sweet, Using deep operators to create spatio-temporal surrogates for dynamical systems under uncertainty, arXiv preprint arXiv:2506.11761 (2025). URL: https://arxiv.org/abs/2506.11761. doi:10.48550/arXiv.2506.11761.

[24] J. M. Reinke, H. Sorg, Wound repair and regeneration, European Surgical Research 49 (2012) 35–43. doi:10.1159/000339613.

[25] I. Papapanagiotou, R. Bumbuc, H. I. Korkmaz, V. Krzhizhanovskaya, V. M. Sheraton, From simulations to surrogates: Neural networks enhancing burn wound healing predictions, 2024. doi:10.5281/zenodo.13867716, zenodo repository, all versions.

[26] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 9351 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

[27] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-C. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in: Advances in Neural Information Processing Systems (NeurIPS), volume 28, 2015, pp. 802–810. URL: https://arxiv.org/abs/1506.04214.

[28] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: International Conference on Learning Representations (ICLR), 2015. URL: https://arxiv.org/abs/1409.0473.

[29] L. Lu, P. Jin, G. E. Karniadakis, Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators, CoRR abs/1910.03193 (2019). URL: https://arxiv.org/pdf/1910.03193.

[30] S. Wang, H. Wang, P. Perdikaris, Improved architectures and training strategies for deep operator networks, Journal of Scientific Computing 92 (2022) 35. doi:10.1007/s10915-022-01881-0.

[31] P. J. Murray, Macrophage polarization, Annual Review of Physiology 79 (2017) 541–566. doi:10.1146/annurev-physiol-022516-034339.

[32] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, scikit-image contributors, scikit-image: image processing in Python, PeerJ 2 (2014) e453. doi:10.7717/peerj.453.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612. doi:10.1109/TIP.2003.819861.

[34] G. P. Renieblas, A. T. Nogués, A. M. n. González, N. Gómez-León, E. G. del Castillo, Structural similarity index family for image quality assessment in radiological images, Journal of Medical Imaging 4 (2017) 035501. doi:10.1117/1.JMI.4.3.035501.

[35] N. C. Silver, W. P. Dunlap, Averaging correlation coefficients: should Fisher's z transformation be used?, Journal of Applied Psychology 72 (1987) 146–148. doi:10.1037/0021-9010.72.1.146.

[36] H. Bal, D. Epema, C. de Laat, R. Van Nieuwpoort, J. Romein, F. Seinstra, C. Snoek, H. Wijshoff, A medium-scale distributed system for computer science research: Infrastructure for the long term, Computer 49 (2016) 54–63. doi:10.1109/MC.2016.127.

[37] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, pp. 2623–2631. doi:10.1145/3292500.3330701.

[38] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, A. Anandkumar, Fourier neural operator for parametric partial differential equations, in: International Conference on Learning Representations (ICLR), 2021. URL: https://openreview.net/forum?id=c8P9NQVtmnO.

[39] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, et al., Relational inductive biases, deep learning, and graph networks, arXiv preprint arXiv:1806.01261 (2018). URL: https://arxiv.org/abs/1806.01261.

[40] X. Meng, G. E. Karniadakis, A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems, Journal of Computational Physics 401 (2020) 109020. doi:10.1016/j.jcp.2019.109020.

[41] M. A. Beaumont, Approximate Bayesian computation in evolution and ecology, Annual Review of Ecology, Evolution, and Systematics 41 (2010) 379–406. doi:10.1146/annurev-ecolsys-102209-144621.

[42] T. Duswald, L. Breitwieser, T. Thorne, B. Wohlmuth, R. Bauer, Calibration of stochastic, agent-based neuron growth models with approximate Bayesian computation, Journal of Mathematical Biology 89 (2024) 50. doi:10.1007/s00285-024-02144-2.

[43] X. Wang, A. L. Jenner, R. Salomone, D. J. Warne, C. Drovandi, Calibration of agent-based models for monophasic and biphasic tumour growth using approximate Bayesian computation, Journal of Mathematical Biology 88 (2024) 28. doi:10.1007/s00285-024-02045-4.

[44] M. C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2001) 425–464. doi:10.1111/1467-9868.00294.

[45] B. Björnsson, et al., Digital twins to personalize medicine, Genome Medicine 12 (2020) 4. doi:10.1186/s13073-019-0701-3.

[46] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems (NeurIPS), volume 30, 2017, pp. 6402–6413. URL: https://arxiv.org/abs/1612.01474.

[47] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: Proceedings of the 33rd International Conference on Machine Learning (ICML), volume 48 of *PMLR*, 2016, pp. 1050–1059. URL: https://proceedings.mlr.press/v48/gal16.html.

[48] A. N. Angelopoulos, S. Bates, Conformal prediction: A gentle introduction, Foundations and Trends in Machine Learning 16 (2023) 494–591. doi:10.1561/2200000101.

[49] T. Sun, S. Adra, R. Smallwood, M. Holcombe, S. MacNeil, Exploring hypotheses of the actions of TGF-β1 in epidermal wound healing using a 3D computational multiscale model of the human epidermis, PLoS ONE 4 (2009) e8515. doi:10.1371/journal.pone.0008515.

[50] H. V. Jain, K.-A. Norton, B. B. Prado, T. L. Jackson, SMoRe ParS: A novel methodology for bridging modeling modalities and experimental data applied to 3D vascular tumor growth, Frontiers in Molecular Biosciences 9 (2022) 1056461. doi:10.3389/fmolb.2022.1056461.

[51] D. R. Bergman, K.-A. Norton, H. V. Jain, T. Jackson, Connecting agent-based models with high-dimensional parameter spaces to multi-dimensional data using SMoRe ParS: A surrogate modeling approach, Bulletin of Mathematical Biology 86 (2024) 11. doi:10.1007/s11538-023-01240-6.

[52] R. Bumbuc, J. Kowalczuk, I. Papapanagiotou, R. Getsadze, youwillfindinfinity/spatio-temporal-neural- surrogates-for-agent-based-burn-wound-models: v1.0.0 — initial release, 2026. URL: https://doi.org/10.5281/zenodo.20457663. doi:10.5281/zenodo.20457663.