ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Lu F, Zlobina K, Rondoni NA, Teymoori S, Gomez M. 2024 Enhancing wound healing through deep reinforcement learning for optimal therapeutics. *R. Soc. Open Sci.* **11**: 240228.
https://doi.org/10.1098/rsos.240228

Received: 16 February 2024
Accepted: 10 June 2024

Subject Category:
Computer science and artificial intelligence

Subject Areas:
bioengineering, biomathematics, biotechnology

Keywords:
deep learning, reinforcement learning, optimal adaptive control, wound healing, optimal treatment regime

Author for correspondence:
Fan Lu
e-mail: flu16@ucsc.edu

THE ROYAL SOCIETY
PUBLISHING

Enhancing wound healing through deep reinforcement learning for optimal therapeutics

Fan Lu, Ksenia Zlobina, Nicholas A. Rondoni, Sam Teymoori and Marcella Gomez

Applied Mathematics, Baskin School of Engineering, University of California, Santa Cruz, CA, USA

FL, 0000-0001-6490-6416; MG, 0000-0001-9709-5015

Finding the optimal treatment strategy to accelerate wound healing is of utmost importance, but it presents a formidable challenge owing to the intrinsic nonlinear nature of the process. We propose an adaptive closed-loop control framework that incorporates deep learning, optimal control and reinforcement learning to accelerate wound healing. By adaptively learning a linear representation of nonlinear wound healing dynamics using deep learning and interactively training a deep reinforcement learning agent for tracking the optimal signal derived from this representation without the need for intricate mathematical modelling, our approach has not only successfully reduced the wound healing time by 45.56% compared to the one without any treatment, but also demonstrates the advantages of offering a safer and more economical treatment strategy. The proposed methodology showcases a significant potential for expediting wound healing by effectively integrating perception, predictive modelling and optimal adaptive control, eliminating the need for intricate mathematical models.

1. Introduction

Personalized precision treatments have become an emerging research topic in modern medicine owing to the recent advances in artificial intelligence [1–3]. The necessity of precision treatment arises from the fact that different patients exhibit different responses to a given medication. These variations stem from the molecular disparities among different patients and within the same patient at different times [4]. Personalized treatment aims to determine personalized dosages of drugs, drug types and the optimal timing of drug delivery for each patient according to current and predicted patient responses based on experimental data and statistical analysis [5]. In this article, we focus on developing an online adaptive controller using deep learning and reinforcement learning (RL) based on the patient's real-time response to the administered treatments. This controller has been designed to expedite wound healing, but its applicability extends to other nonlinear dynamics.

Wound healing is a dynamic and continuous process that can unfold through a series of overlapping stages: haemostasis, inflammation, proliferation and maturation [6]. The process involves nonlinear transformations of different cells (platelets, neutrophils, macrophages, myofibroblasts, fibroblasts, keratinocytes and others) and biomolecules (blood coagulation factors, pro- and anti-inflammatory cytokines, polymers and enzymes of extracellular matrix) [7].

Determining the optimal timing and precise dosage for administering each drug presents a challenge, particularly when considering the different nonlinear dynamics of drug digestion and biological transformations targeted by the drug. The mechanism involved in drug distribution can be elucidated using mathematical models [8,9]. However, the complexity of biological systems and disparities within organisms reduces the reliability of model-based controllers.

Even if the model is accurate, the inherent nonlinearity further complicates the task of establishing the optimality and safety of the prescribed control policy for drug administration. This emphasizes the need to formulate controller design strategies that furnish optimal and adaptable control solutions while considering the individualized requirements of patients with specific health conditions, all under the guidance of analytically optimal and safe solutions.

Several closed-loop control strategies, such as model predictive control, optimal control and adaptive disturbance rejection control, have been suggested to control drug administration [8,10–12]. The control strategies currently in use for regulating patient drug dosing have focused on optimal drug infusion with respect to given performance measures or adaptive drug infusion that addresses patient parameter uncertainty. The main advantage of adaptive controllers is that they can derive patient-specific infusion profiles even without an accurate patient model. However, such controllers may not account for certain desired performance constraints. On the other hand, optimal controllers are predicated on nominal patient models, leading to suboptimal performance or even instability of the closed-loop system in the face of drug titration for actual patients.

The challenge here is to design an optimal treatment that accounts for gender, age, weight, pharmacokinetic and pharmacodynamic intrapatient and interpatient variability, and health conditions of the patient under treatment. In contrast to standard controller design methods, RL-based approaches allow the development of control algorithms that can be used in real time to affect optimal and adaptive drug dosing in the presence of pharmacokinetic and pharmacodynamic patient variability. The method presented in this article can be used to derive patient-specific treatment profiles, such as generating a desired patient drug response without requiring an accurate patient model. Specifically, we use a learning-based controller design strategy that can facilitate patient-specific and optimal drug titration.

Learning-based control strategies have found applications in various medical settings, enhancing the precision of drug dosing and optimizing treatment regimens. These applications include devising dynamic treatment plans for lung cancer patients [13], optimizing erythropoietin dosing during haemodialysis [14], facilitating cytotoxin delivery during chemotherapy [15], aiding insulin regulation for diabetic individuals [16] and administering anaesthetic drugs to maintain desired sedation levels [17]. Recent studies, such as those discussed in Moore *et al.* and Padmanabhan *et al.* [17,18], have delved into clinical and computational trials employing RL to enhance the precision of anaesthetic drug infusion.

However, it is worth noting that these approaches do not account for safety exploration in RL. Owing to the inherent non-convexity of objective functions and the complexity of deep neural networks, achieving a globally optimal control policy is not always assured. In the study by Padmanabhan *et al.* [19], the utilization of RL to inform drug dosing is suggested. Nevertheless, this approach still relies on prior knowledge of reference signals.

Compared to the studies by Zhao *et al.*, Martín-Guerrero *et al.*, Padmanabhan *et al.*, Daskalaki *et al.* and Moore *et al.* [13–17,19], the proposed approach in this article presents a distinct advantage by formulating a leader–follower paradigm solved by deep reinforcement learning (DRL) which has demonstrated success in Zhou *et al.* [20]. We begin by learning a mapping from nonlinear wound dynamics to its linear representation. From this linear model, the optimal control law is derived, allowing the calculation of the subsequent optimal linear state. This linear state is then used by the decoder in the DeepMapper, as illustrated in figure 1, to predict the next optimal nonlinear state. This prediction serves as a reference signal for the RL agent, guiding the formulation of a treatment strategy, such as real-time drug dosages, aimed at closely matching the actual next nonlinear state to this reference state. Thanks to DRL, the regime eliminates the need for modelling the nonlinear wound dynamics or the treatment effects within it.

Learning linear representations of nonlinear systems is crucial owing to the fact that nonlinear systems are prevalent in nature, with most systems of practical interest exhibiting nonlinear behaviour, and the control of such systems is challenging with no general and scalable solution [21–23]. On the other hand, the study of linear systems is well developed with scalable design, analysis, control and optimization of linear systems thoroughly detailed within the literature [24,25].

Finding the mapping between nonlinear systems and linear models is challenging. The Koopman operator theory, as explored in seminal works in Koopman, Mezić & Banaszuk and Mezić [26–28], offers a promising avenue by enabling the representation of a nonlinear system as an infinite-dimensional linear system. However, the optimization of this approach primarily operates within the realm of functional space, rendering it often intractable in practical applications. Moreover, it does not account for the effect of control inputs in nonlinear systems.

In recent years, many advances have been made to generalize the Koopman operator theory for the control of nonlinear systems [21,29–31]. These extensions find linear representations of nonlinear systems with finite-dimensional function approximations, which can subsequently be used for the tractable control of the system. This motivated us to design a deep neural network-based algorithm called DeepMapper to learn a mapping from an unknown nonlinear system to its linear representation, similar to the goal of the Koopman Operator. The work is primarily motivated by Kaiser *et al.* and Ahmed *et al.* [30,31]. The major difference is that neither addresses the issue of overfitting during learning, which is mitigated through DRL in this article.

The main contributions of this article include the following. (i) We propose an adaptive closed-loop control framework using deep learning, optimal control and RL to enhance wound healing, as schematized in figure 1. This framework eliminates the need for mathematical modelling of nonlinear dynamics or the treatment effects within it. (ii) We propose an autoencoder-like mechanism called DeepMapper to learn a linear representation of the nonlinear wound healing dynamics, which provides an optimal reference signal for the DRL agent to track. It is shown that this regime not only improves the precision of the linear representation in modelling the wound dynamics under optimal treatments but also ensures the efficiency of the DRL agent. (iii) The experimental results show that our approach has successfully reduced the wound healing time by 45.56% compared with the one without any treatment, as well as outperforming the one with DRL directly optimized over a nonlinear system without DeepMapper. The proposed framework showcases the significant potential for expediting wound healing by effectively integrating optimal control and data-driven methods. By leveraging advanced algorithms that adapt in real time to changing conditions, this system offers a more accurate and reliable means of promoting faster recovery without relying on the limitations of conventional models.

The remainder of this article is organized as follows: §2 presents an overview of the closed-loop control framework, followed by the detailed design of the deep learning-based algorithm for finding the linear representation for nonlinear wound healing dynamics, as well as a DRL-based algorithm for accelerating wound healing. Implementation details, simulation results and a detailed discussion of these results are given in §§3 and 4. Finally, in §5, we present conclusions and future research directions.

2. Approach

Nonlinear dynamics characterize the vast majority of systems of practical interest. One example of this is the wound healing dynamics depicted in this article. The control and optimization of such systems, particularly in scenarios like devising the most effective wound treatment strategy, is challenging owing to their inherently nonlinear behaviour.

Conventional methods, such as linearizing around a fixed point, frequently fall short when applied to complex systems with nonlinear behaviours. These methods assume that the system's behaviour near the fixed point can be approximated by a linear model, which is not always effective for systems exhibiting significant nonlinearity, multiple equilibria or chaotic dynamics. Consequently, they require alternative approaches that can accurately model and predict the behaviour of such systems across a

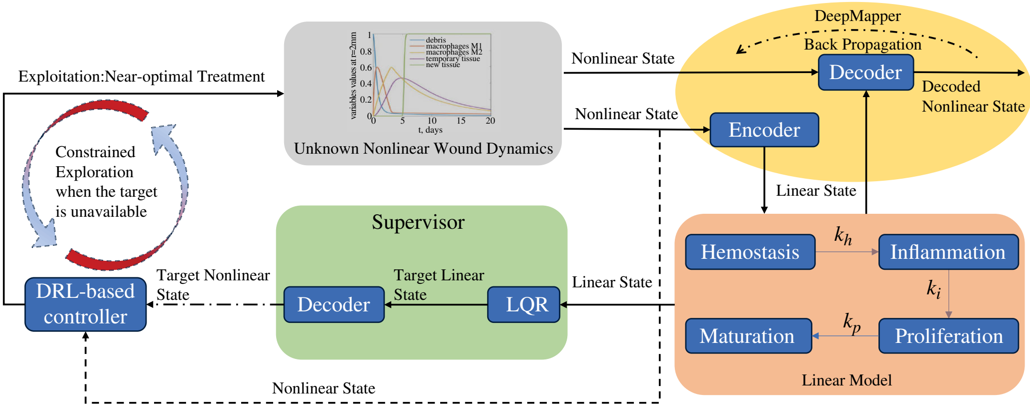

Figure 1. DRL-based closed-loop control to accelerate wound healing pipeline. The pipeline consists of five major blocks for performing the real-world wound state estimation in nonlinear dynamics, finding a linear representation of wound dynamics, calculating optimal reference signal in the learned linear model, supervision of the real-world wound target state and constraint exploration and exploitation of DRL agents.

broader range of conditions, bypassing the limitations inherent to linearization techniques. Alternative methods, such as machine learning and deep learning, can capture the intricate nature of these systems more efficiently through data.

For this reason, we propose a deep learning framework called DeepMapper to learn a mapping from a nonlinear wound healing dynamics to its linear representation. We show that the learned linear model can be used to provide a near-optimal reference signal to a RL controller, which will, in turn, refine the DeepMapper. This controller learns the best treatment strategy for wound care without heavily relying on mathematical interpretation of the treatment into any dynamic model. The learning framework is schematized in figure 1.

2.1. DeepMapper: linearization of nonlinear wound healing dynamics

Consider a nonlinear wound healing dynamical system with treatment inputs defined by:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x) + \mathbf{Bu}, \tag{2.1}$$

where $x \in \mathbb{R}^{d_x}$, $\mathbf{B} \in \mathbb{R}^{d_x \times d_u}$, $\mathbf{u} \in \mathbb{R}^{d_u}$ and $f : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$.

As discussed in Zlobina *et al.* [9], the function f can be an unmanageable nonlinear function that defines different cell transitions during wound healing. The nonlinear state x associated with wound healing surveyed in the literature may include variables such as pH, temperature [32–37] or visual representations captured through images of the wound [38]. We assume that x can be measured by some sensor, but f is unknown to the control algorithm.

Solving for the optimal control input $\mathbf{u}^{\star}$ is often difficult, particularly when the dynamics evolve nonlinearly. As extensive research and literature have been dedicated to the study of linear systems, encompassing scalable design, analysis, control and optimization [24,39], we propose the utilization of a deep learning approach to model a linear system that best approximates the behaviour of the underlying nonlinear system.

Note that equation (2.1) defines a control-affine system. As discussed in Kaiser *et al.* [30], the decoupling of the states and inputs allows us to find a transformation of the states alone:

$$z = h(x), \tag{2.2}$$

where $x \in \mathbb{R}^{d_x}$ is the state that evolves subject to nonlinear dynamics, $z \in \mathbb{R}^{d_z}$ evolves linearly and $h : \mathbb{R}^{d_x} \to \mathbb{R}^{d_z}$ is a function that maps the nonlinear state x to linear state z.

From the chain rule, the relationship between the new state z and the original state x can be defined:

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\mathrm{d}z}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t} = \mathbf{J}_h(x)\frac{\mathrm{d}x}{\mathrm{d}t}, \tag{2.3}$$

where $\mathbf{J}_h(x) \in \mathbb{R}^{d_z \times d_x}$ is the Jacobian matrix of h.

We first seek to find a linear representation of equation (2.1) without control $\mathbf{u}$ satisfying the condition that the dynamics of the new state z are linear in z:

$$\frac{\mathrm{d}z}{\mathrm{d}t} = Az, \tag{2.4}$$

where $A \in \mathbb{R}^{d_z \times d_z}$

By expanding equation (2.3) through substitution of equation (2.1) with $\mathbf{u} = 0$ for the time derivative term and accounting for equation (2.4) that we want to satisfy, we have

$$\begin{aligned}\frac{\mathrm{d}z}{\mathrm{d}t} &= \mathbf{J}_h(x)f(x)\\ &= Az\,.\end{aligned}$$

Then by plugging the control input $\mathbf{u}$ into equation (2.1) and following the same procedure, we get a linear representation with control:

$$\begin{aligned}\frac{\mathrm{d}z}{\mathrm{d}t} &= \mathbf{J}_h(x)(f(x) + \mathbf{Bu})\\ &= Az + \mathbf{J}_h(x)\mathbf{Bu},\end{aligned} \tag{2.5}$$

which defines an underdetermined system of d_z equations with d_z unknown transformations and d_z^2 unknown coefficients in matrix A. However, for some special cases, a closed-form solution can be found directly, such as when $d_x = d_z = 1$. In general, however, equation (2.5) cannot be solved directly.

Alternatively, we propose to solve it by reformulating equation (2.5) into an optimization problem that can be solved using data measured from the system. Specifically, an objective function can be defined as the squared Euclidean norm of the difference between equations (2.3) and (2.5):

$$L(x, \mathbf{u}; h; A) = \| \mathbf{J}_h(x)\frac{\mathrm{d}x}{\mathrm{d}t} - [Az + \mathbf{J}_h(x)\mathbf{Bu}] \|_2^2,$$

where $L \in \mathbb{R}$ and $\| \cdot \|_2 : \mathbb{R}^{d_z} \to \mathbb{R}$.

The unconstrained optimization problem can be defined as finding $h^\star$ and $A^\star$ such that

$$h^\star, A^\star \in \underset{h \in F, A \in \mathbb{R}^{d_z \times d_z}}{\arg\ \min}\ L(x, \mathbf{u}; h; A)\,. \tag{2.6}$$

However, optimization over function spaces as in equation (2.6) is often difficult and intractable. Therefore, an alternative strategy involves parametrizing the space of functions or establishing a set of basis functions from which the broader function space can be derived, as discussed in Sasane [40]. In this article, we adopt this alternative approach. Specifically, we employ a deep neural network to parametrize the function h in equation (2.2):

$$z = h^\theta(x), \tag{2.7}$$

where $\theta \in \mathbb{R}^{d_n}$ is the weight of neural networks. With A also parameterized by neural networks denoted as A^ω through $\omega \in \mathbb{R}^{d_z \times d_z}$, the optimization problem defined in equation (2.6) can be reformulated into a manageable form, solely involving the optimization of parameters:

$$\theta^\star, \omega^\star \in \underset{\theta \in \mathbb{R}^{d_n}, \omega \in \mathbb{R}^{d_z \times d_z}}{\arg\min}\ L(x, \mathbf{u}; \theta, \omega) \tag{2.8}$$

with

$$L(x, \mathbf{u}; \theta; \omega) :=\ \| \mathbf{J}_\theta(x)\frac{\mathrm{d}x}{\mathrm{d}t} - \left[A^\omega h^\theta(x) - \mathbf{J}_\theta(x)\mathbf{Bu}\right] \|_2^2, \tag{2.9}$$

where J_θ is the Jacobian matrix of the deep neural network with regard to the parameter θ.

Note that equation (2.8) is trivially minimized by the solution $\theta = 0$ and $\omega = 0$. To avoid such issues, a regularization term was added to the objective function. In essence, the additional term defines a 'decoder' network [41] to perform the inverse transformation from the new state z, back to a reconstruction of x, which we denote as $\hat{x}$ with

$$\hat{x} = \hat{h}^{\hat{\theta}}(z), \tag{2.10}$$

where $\hat{h}$ is the neural network with weight $\hat{\theta} \in \mathbb{R}^{d_z \times d_{\hat{h}}}$. Combining equations (2.7) and (2.10), the decoder's objective is then defined to minimize

$$\hat{L}\left(x;\hat{\theta}\right) = \| x - \hat{x} \|_2^2 = \| x - \hat{h}^{\hat{\theta}}\left(h^{\theta}(x)\right) \|_2^2 . \tag{2.11}$$

The overall objective function consists of a weighted sum of equations (2.9) and (2.11), which gives rise to the optimization problem:

$$\min_{\theta,\hat{\theta},\omega} L(x,\mathbf{u};\theta;\omega) + \alpha\hat{L}\left(x;\hat{\theta}\right), \tag{2.12}$$

where α defines a scalar weighting factor applied to the decoder term.

The optimization problem can now be solved, provided that data measured from the system are available. In the case of equation (2.12), additional data in the form of the input measurements, $\mathbf{u}$, through time are necessary. This gives the final optimization solved numerically over the data:

$$\begin{gathered} \theta^{\star}, \hat{\theta}^{\star}, \omega^{\star} \in \arg \min_{\theta,\hat{\theta},\omega} \frac{1}{T}\sum_{t=1}^{T} L^{s}\left(\mathbf{s}_t;\theta,\omega,\hat{\theta}\right) \\ L^{s}\left(\mathbf{s}_t;\theta;\omega,\hat{\theta}\right): = L(\mathbf{s}_t;\theta,\omega) + \alpha\hat{L}\left(x_t;\hat{\theta}\right), \end{gathered} \tag{2.13}$$

where $\mathbf{s}_t$: = $(x_t, \mathbf{u}_t)$ and T is the number of samples in the data and superscript defines the tth sample.

The learned linear dynamics of equation (2.5) can be represented as:

$$\frac{\mathrm{d}z}{\mathrm{d}t} = A^{\omega^{\star}} z + \mathbf{J}_{\theta^{\star}}(x)\mathbf{B}\mathbf{u}. \tag{2.14}$$

The optimal control problem can be solved to control nonlinear dynamic systems of the form of equation (2.1), using the learned linear representation of the form of equation (2.14), by solving the Riccati equation for the optimal gain matrix, $\mathbf{K} \in \mathbb{R}^{d_u \times d_z}$, giving the optimal control law:

$$\mathbf{u}^{\star} = -\mathbf{K}z, \tag{2.15}$$

referred to as the linear quadratic regulator [39].

Note that equation (2.14) is linear in z, but not necessarily jointly linear in the inputs and states owing to the Jacobian term, $J_{\theta^*}(x)$, which may be dependent on the nonlinear state x. As discussed in Kaiser *et al.* [30], though the nonlinear state-dependent term does not pose any major issues with regard to control of the nonlinear system or the linear system, in practice, the Jacobian matrix can be ill-conditioned during the initial phase of learning, making equation (2.15) unavailable. In this article, we propose a DRL agent to track the reference signal incurred by equation (2.15) whenever it is available and penalize the DRL agent whenever it is not. We show that the control law learned by this DRL agent is better than the one directly optimize it over a nonlinear system without a mapping.

2.2. Reinforcement learning algorithm design

In this section, we introduce the use of a DRL algorithm to explore possible policies that will cover as many scenarios of the nonlinear dynamics with inputs as possible in the case when the optimal control input from the learned linear representation is not available. Meanwhile, such exploration should adhere to constraints that account for the physical and biological limitations inherent to the wound healing system, while ensuring ethical considerations are not compromised.

When the optimal control input is accessible, the DRL algorithm should be able to exploit its acquired knowledge to generate a policy that closely approximates the resulting nonlinear state to the one achieved through control based on the optimal control. The exploration and exploitation of the DRL algorithm do not require knowledge of either nonlinear or linear dynamics, and thus it not only alleviates the burden of mathematical interpretation in real-world treatment scenarios but also significantly expedites the healing process. To realize this, we first formulate the wound healing dynamics as the Markov decision process (MDP) problem and subsequently solve it using the famous Deep Q-learning [42].

We consider a MDP defined by (X, U, P, r, γ), where X represents the state space, U represents the input/action space, P represents the transition probability matrix, r represents the reward function and γ represents the discount factor. In MDP, an autonomous agent makes sequential discrete-time decisions as time passes. Generally speaking, the MDP problem conforms to the decision-making process of physicians in wound care. Based on the state $x_t \in X$, the agent selects action $u_t \in U$ at time t, then it observes the next state x_{t+1} and receives the reward $r(x_t, u_t) \in \mathbb{R}$. To collect more state information in wound management, the agent can perform state observation more frequently, such as a state observation every hour and action selection every 20 min [9,38]. The state s_t transits to the next state x_{t+1} following the transition probability matrix $P(x_{t+1} | x_t, u_t)$, which represents the dynamics of the operating environment. The transition probability matrix satisfies the Markovian (or memoryless) property since a transition to the next state x_{t+1} depends only on the current state x_t and action u_t rather than a historical series of states and actions. The agent learns the optimal policy $\phi^\star : X \to U$, which maps $x \in X$ to optimal actions $u \in U$ over trial and error interaction with the environment. Nevertheless, the transition probability matrix and the probability distribution of the reward function are generally unknown in reality.

Deep Q-learning The goal of an RL agent is to interact with the environment by selecting actions to maximize cumulative future rewards. We make the standard assumption that future rewards are discounted by a factor of γ per time step and define the optimal action-value Q-function as the maximum total discounted expected reward over all possible action sequences $\mathscr{U} : = \{u_t : t \geq 1\}$:

$$\begin{aligned} Q^\star(x, u) &= \max_U \sum_{t=0}^{\infty} \gamma^t E[r(x_t, u_t) \,|\, x_0 = s, u_0 = u] \\ &= \max_U \sum_{t=0}^{\infty} \sum_{x' \in X} P(s' \,|\, s_t, u_t)\left(r(x_t, u_t) + \gamma \max_{u'} Q^\star(x', u')\right), \end{aligned}$$

with $x \in X$ and $u \in U$.

Let P_u denote the state transition matrix when action $u \in U$ is taken. It is known that the Q-function is the unique solution to the Bellman equation [43]:

$$Q^\star(x, u) = r(x, u) + \gamma \sum_{x' \in X} P_u(x, x') \underline{Q}^\star(x'),$$

where $\underline{Q}(x) : = \max_{u \in U} Q(x, u)$ for any function $Q : X \times U \to \mathbb{R}$.

Consider a parametrized family of approximations $\left\{Q^\vartheta : \vartheta \in \mathbb{R}^d\right\}$, wherein $Q^\vartheta : X \times U \to \mathbb{R}$ and ϑ may represent the weights from deep neural networks. The associated family of policies is defined as

$$\phi^\vartheta(x) \in \arg\max_{u \in U} Q^\vartheta(x, u), \quad x \in X. \tag{2.16}$$

The goal of the Deep Q-network (DQN) algorithm is to find $\vartheta^\star$ such that the mean square Bellman error is minimized:

$$\vartheta^\star \in \arg\min_{\vartheta \in \mathbb{R}^d} E\left[\| D_{t+1}(\vartheta) \|_2^2\right], \tag{2.17}$$

where $\mathscr{D}_{t+1}(\vartheta) : = r(x_t, u_t) + \gamma \underline{Q}^\vartheta(x_{t+1}) - Q^\vartheta(x_t, u_t)$, and the expectation is in a steady state.

To balance the trade-off between exploration and exploitation, we adopt the ε-greedy policy approach, where ε follows the following updating rule:

$$\varepsilon_{t+1} = \max\,(\varepsilon_{\min}, \nu \varepsilon_t), \tag{2.18}$$

with $\varepsilon_{\min}$ the minimum value that ε can achieve and $0 < \nu < 1$ the decay rate.

Initially, ε_0 is set to a value close to 1, such as 0.99. This initial value encourages a higher probability of random selection of action $u_t \in U(x_t)$ with $U(x) \subseteq U$ constrained by the current state. This choice aligns with the early stages of training when both the transformation h^θ and the DRL agent are still in the learning process and are not yet well-versed. As they are continuously updated through trajectories collected from real-world experiments or simulated wound dynamics, they gain more confidence in the learned linear representation and nonlinear dynamics. Thus, we should gradually decrease ε and guide the policy towards more deterministic actions.

The optimization problems of equations (2.12) and (2.17) are then solved iteratively using data collected by interacting with some wound dynamics as per Algorithm 1.

The steps for obtaining optimal solutions are summarized as follows:

Step 1: Hyperparameter setting The weighting faction α, applied to the decoder term in equation (2.12), is initialized with value α°. The exploration rate ε is set to value ε°.

Step 2: Deep neural network initialization The initial weights of the deep Q networks ϑ°, transformer θ°, decoder $\widehat{\theta}^\circ$ and A^ω matrix ω° are randomly selected by the Kaiming uniform method [44,45].

Step 3: Learning through data A while loop is initiated until the termination criteria are satisfied. That is, either the optimal parameters between iterations are similar, where similarity is measured with a Euclidean distance metric, or the maximum number of iterations is exceeded. Within this while loop, we have another while loop that keeps interacting with the nonlinear wound dynamics using control inputs either obtained from equation (2.15) or the randomly selected one based on some constrained input space $U(x)$. This interaction will stop until the wound has healed. All the data during the interaction will be stored for optimizing the parameters. The optimization problems considered in this article were solved using the Adam optimizer [46].

Step 4: Return The optimal of parameters, $\vartheta^\star$, $\theta^\star$, $\widehat{\theta^\star}$, and $\omega^\star$ will be obtained by using the Polyak–Ruppert averaging method defined in equation (3.4).

Algorithm 1: Closed-loop control of wound healing

```
$\varepsilon_0 \leftarrow \varepsilon^\circ$;
$\alpha \leftarrow \alpha^\circ$;
$n \leftarrow 0$;
$\theta_0, \hat{\theta}_0, \omega_0, \vartheta_0 \leftarrow \theta^\circ, \hat{\theta}^\circ, \omega^\circ, \vartheta^\circ$;
while termination criteria not met do
    while $t \leq T$ do
        Estimate nonlinear wound state $x_t$;
        if (2.15) is available then
            $u_t \leftarrow -\mathbf{K}h^{\theta_n}(x_t)$;
        else
            Sample $\xi$ uniformly from [0, 1];
            if $\xi \geq \varepsilon$ then
                $u_t \sim \phi^{\vartheta_n}(x_t)$
            else
                Randomly choose $u_t$ from $\mathrm{U}(x_t)$;
            end
        end
        Calculate reward $r(x_t, u_t)$ through (3.5);
        Estimate the next state $x_{t+1}$ with input $u_t$;
    end
    $\theta_{n+1}, \hat{\theta}_{n+1}, \omega_{n+1} \leftarrow \arg\min_{\theta,\hat{\theta}_n,\omega} \frac{1}{T}\sum_{t=1}^{T} L^s(s_t; \theta, \omega, \hat{\theta})$;
    $\vartheta_{n+1} \leftarrow \arg\min_{\vartheta\in\mathbb{R}^d} \frac{1}{T}\sum_{t=1}^{T} [||D_{t+1}(\vartheta)||_2^2]$;
    $\varepsilon_{n+1} = \max(\varepsilon_{\min}, \nu\varepsilon_n)$;
    $n \leftarrow n + 1$;
end
Compute $\vartheta^*, \theta^*, \hat{\theta}^*$, and $\omega^*$ via (3.4);
```

3. Experiment results and discussion

We evaluate the proposed algorithm by applying it to a nonlinear model, as introduced in Zlobina *et al.* [9]. This nonlinear model addresses wound healing by encompassing five key variables: the quantity of debris a, M1 macrophage m_1, M2 macrophage m_2, temporal tissue c and new tissue n. We assume in equations (3.1a)–(3.1e) that the control factor u only impacts the rate of transitions from m_1 to m_2. Consider a circular wound with radius R and denote $x: = [a, m_1, m_2, c, n]^\mathsf{T}$. Each element in x evolves nonlinearly in both the spatial dimension $0 \le \tilde{r} \le R$ and temporal dimension $0 \le t$:

$$\dot{a} = -am_1, \tag{3.1a}$$

$$\dot{m}_1 = \beta a - \dot{a} - \rho \frac{m_1^q}{k^q + m_1^q} - \gamma_1 m_1 + \tilde{D}F(m_1) - um_1, \tag{3.1b}$$

$$\dot{m}_2 = \rho \frac{m_1^q}{k^q + m_1^q} - \gamma_2 m_2 + \tilde{D}F(m_2) + um_1, \tag{3.1c}$$

$$\dot{c} = m_2 - \mu c, \tag{3.1d}$$

$$\dot{n} = c\left[\tilde{\alpha} n(1-n) + \tilde{D}_n F(n)\right], \tag{3.1e}$$

where $\mathscr{F}(x): = \frac{1}{\tilde{r}}\frac{\partial x}{\partial \tilde{r}} + \frac{\partial^2 x}{\partial \tilde{r}^2}$ for any variable x, the wound radius $\tilde{r}$ is directed from the wound centre to the wound edge and u~ϕ with $\phi: \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ a function of space and time that modifies the polarization of m_1 to m_2 and affects the rate of the generation of new tissues. Note that equations (3.1a)–(3.1e) can be written in the form of equation (2.1) with the matrix $\mathbf{B}$ defined as

$$\mathbf{B}: = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$\mathbf{u}: = [0, u, u, 0, 0]^\mathsf{T}$and f capturing the first and second derivatives of variables. We assume that f is unknown to the algorithms and x is measurable through some sensor attached to the wound.

We define the wound size at time t as the smallest radius where the new tissue reaches a value of σ:

$$s(t) = \min_{\tilde{r}} n(t, \tilde{r}) \ge \sigma\,.$$

The wound healing time is defined as the time from injury ($t = 0$) to the moment when the wound radius is zero:

$$\tau = \min_{t \ge 0} s(t) = 0. \tag{3.2}$$

The goal is to find an actuation function ϕ such that τ is minimized. Nevertheless, solving for the optimal $\phi^\star$ directly from equations (3.1a)–(3.1e) is often difficult, particularly when involving second-order derivatives. In the previous work [9], a brute-force search (BFS) method was used. While effective, this approach was not only time-intensive but also heavily reliant on prior knowledge of the structures of the actuation function, which may not cover the true optimal solution. Consequently, there is still untapped potential for expediting wound healing by employing more advanced approaches.

In this section, we first use the method proposed in §2 to learn a linear representation of the system (3.1a)–(3.1e) without any actuation ($u = 0$). Our primary motivation for this is twofold: firstly, to demonstrate the capability of our proposed method in acquiring a meaningful linear representation of the nonlinear system; and secondly, to unveil biologically interpretable insights into the variables embedded within the linear model. We then consider this learned linear representation as prior knowledge to find a linear representation for the nonlinear model with control inputs guided by a DRL agent. The experimental results show that the learned policy $\phi^\star$ is capable of reducing the healing time by 45.56% compared to that without any actuation and 37% compared to that with the method employed in Zlobina *et al.* [9].

We summarize all parameter setups during each experiment in table 1.

Table 1. The values of parameters used in experiments.

parameter	R	L	T	β	ρ	κ	q	γ_1	γ_2	μ	D	$\tilde{D}_n$	$\tilde{\alpha}$	γ	$\varepsilon_{\min}$	σ	ν
value	3 mm	0.03 mm	1/3 day	1	0.1	0.05	5	0.1	0.1	0.2	0.32	3×10^{-4}	1.8	0.995	0.01	0.95	0.99

3.1. Learning the linear representation for a nonlinear model without control input

We assume that there are four variables in the linear representation, and each corresponds to the probability of each stage of wound healing: haemostasis P_h, inflammation P_i, proliferation P_p and maturation P_m. As time goes on, the probability of each stage changes: the wound is initially in the hemostasis stage with probability one and experiences a continuous transition from stage to stage.

During the experiment, we found that the learned linear representation is not unique. This is owing to the fact that we are mapping a nonlinear dynamic to a lower-dimensional linear model. To drive the uniqueness of the output of such mapping, we introduce a four-state ODE model:

$$\frac{\mathrm{d}z}{\mathrm{d}t} = Az + W\mathbf{u}, \quad z = \left[P_h, P_i, P_p, P_m\right]^{\mathrm{T}} \in \mathbb{R}^4, \tag{3.3a}$$

$$A = \begin{bmatrix} -k_h^{\mathrm{nat}} & 0 & 0 & 0 \\ k_h^{\mathrm{nat}} & -k_i^{\mathrm{nat}} & 0 & 0 \\ 0 & k_i^{\mathrm{nat}} & -k_p^{\mathrm{nat}} & 0 \\ 0 & 0 & k_p^{\mathrm{nat}} & 0 \end{bmatrix}, \tag{3.3b}$$

where k_h^{nat}, k_i^{nat} and k_p^{nat} are the constants that control the velocities of transitions without any treatment for the wound, and the matrix $W \in \mathbb{R}^{d_z \times d_u}$ needs to be learned to capture the input effect in the linear model from the nonlinear dynamics.

Data preparation In order to find the linear representation for the nonlinear model without any control inputs, we first solve equations (3.1*a*)–(3.1*e*) numerically by constructing 500 ordinary differential equations on a uniform mesh consisting of 100 spatial cells similar to Zlobina *et al.* [9]. The temporal domain spans from $t \in [0, 20]$ with a sampling interval of 0.5. This yields a dataset composed of 121 data points, each characterized by 500 features.

To enlarge the dataset and promote robustness in our results, we introduce additional variability by adding i.i.d. noise. This noise is sampled from a uniform distribution in the range of $[-0.1, 0.1]^{500}$ and is incorporated into the original dataset. This data augmentation increases the size of the dataset to 12 100 data points, which serves as the training data for solving the optimization problem (equation (2.12)).

The neural network that approximates the function h^{θ} has three fully connected layers with the Softmax function as the output layer. The network approximating function $\hat{h}^{\hat{\theta}}$ has a similar structure but with the output layer replaced by a Sigmoid activation function. The reason for this replacement is that the five variables in the nonlinear dynamics are not necessarily the probabilities, but the four variables in the linear model are. We constrained the parameters for the A^{ω} matrix to have the same structure as that in equation (3.3*b*).

We conduct 100 independent runs of learning those approximations, with parameters in the neural networks randomly initialized by the Kaiming uniform method [44,45], and obtain the learning curves of k_h^{nat}, k_i^{nat} and k_p^{nat} shown in figure 2, where all the three parameters converge after around 6×10^3 epochs. To get a better estimation of these values, we conduct Polyak–Ruppert averaging [47]:

$$\omega_N^{\star} := \frac{1}{N - N_0} \sum_{n=N_0}^{N} \omega_t, \tag{3.4}$$

where N denotes the total number of updates in the parameters, and the interval $[0, N_0]$ with $N_0 < N$ is known as the *burn-in* period; estimates from this period are abandoned to reduce the impact of transients in early stages of the training. In this article, we choose $N_0 = 80\%\, N$ and obtain

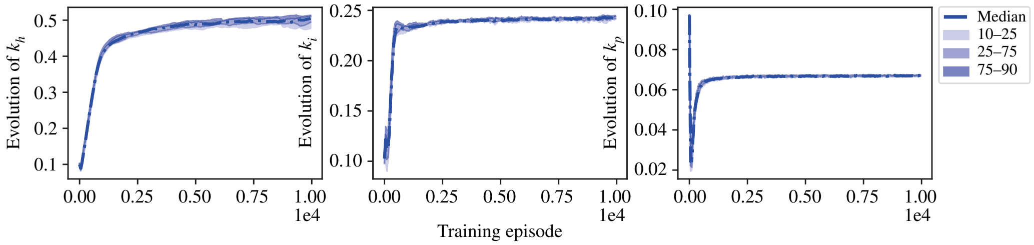

Figure 2. Values of k_h^{nat}, k_i^{nat} and k_p^{nat} during learning, shown by percentile.

$$A^{\omega_T^\star} = \begin{bmatrix} -0.495 & 0 & 0 & 0 \\ 0.495 & -0.247 & 0 & 0 \\ 0 & 0.247 & -0.068 & 0 \\ 0 & 0 & 0.068 & 0 \end{bmatrix}.$$

Figure 3 shows the optimization result for the original trajectory given by equations (3.1a)–(3.1e). The solid red curves represent the exact time derivative calculated from the chain rule, i.e. $\frac{\mathrm{d}z}{\mathrm{d}t} = \mathbf{J}_{\theta^\star}(x)\frac{\mathrm{d}x}{\mathrm{d}t}$. In contrast, the blue dashed curves are the results of the linear approximation, i.e. $\frac{\mathrm{d}z}{\mathrm{d}t} = A^{\omega^\star}z$. These plots show that a linear model has been successfully identified with all the ODEs converging to zero.

In figure 4, it is demonstrated that the decoder $\hat{h}^{\hat{\theta}^\star}$ has effectively mapped the linear variables into the variables of the nonlinear model, underscoring the accuracy and reliability of this mapping process.

Plugging matrix $A^{\omega_T^\star}$ into equation (3.3a), and solving it numerically over a time span from $t \in [0, 20]$, with a sampling interval of 0.5 and $u = 0$, we derive a trajectory of the four variables shown in figure 5. Compared to figure 4, it can be seen that the trajectories of haemostasis, inflammation and proliferation evolve similarly to those of debris, M1 macrophage and M2 macrophage.

3.2. Learning the linear representation for a nonlinear model with deep reinforcement learning control inputs

Subsequently, we proceed to conduct experiments towards acquiring a linear representation of the nonlinear model (equations (3.1a)–(3.1e)) when it involves control inputs ($u \neq 0$) through the DeepMapper. Meanwhile, we would like to simultaneously train a DRL agent to learn an optimal treatment strategy denoted as $\phi^\star$ that minimizes the healing time defined in equation (3.2), which will, in turn, refine the DeepMapper.

In the RL algorithm introduced in §2.2, the Q-network is constructed to have four fully connected layers, and a rectified linear unit (ReLU) activation function follows each layer. The output layer is a Softmax activation function to output the probabilities of each input $u \in U$.

Note that we divide the wound into 100 regions with wound radius evenly spaced along the direction of $\tilde{r}$, and $u \in \mathbb{R}^{100}$ is a vector with each element indicating the amount of actuation at different radiuses of the wound ranging from $[0, R]$ mm. Each element of u takes values from $\{0.1n{:}0 \leq n \leq 10, n \in z\}$. This will result in an input space $U \in \mathbb{R}^{10 \times 100}$.

We took the learned models in §3.1 as a prior for the models with control and updated the Q-network as well as h^{θ} and $\hat{h}^{\hat{\theta}}$ in an online learning way. For each state x_t at time t, the input u_t to the nonlinear model (equations (3.1a)–(3.1e)) is obtained by either equation (2.15) when the solution to the Riccati equation is available or uniformly sampling it from $U(x_t)$. Note that $U(x_t)$ denotes an input space constrained by x_t, so the sampled input will be restricted to the bounds of the wound's biological and physical dynamics.

The reward at time t is defined as

$$r(x_t, u_t) = \begin{cases} e^{-\|\hat{x}_t^\star - x_t\|} - 1, & \text{if (2.15) is available} \\ -2, & \text{otherwise}, \end{cases} \tag{3.5}$$

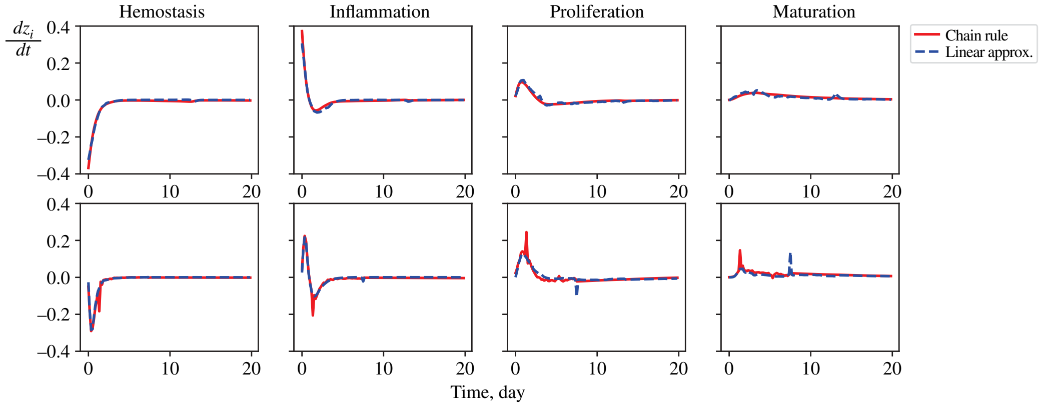

Figure 3. Results of the optimization showing a comparison between the exact time derivative calculated through the chain rule (red curve), i.e. $\frac{\mathrm{d}z}{\mathrm{d}t} = \mathbf{J}_{\theta^\star}(x)\frac{\mathrm{d}x}{\mathrm{d}t}$, and their linear approximation (blue dashed curve), i.e. $\frac{\mathrm{d}z}{\mathrm{d}t} = A^{\omega^\star}z + \mathbf{J}_{\theta^\star}(x)\mathbf{B}u$, with $u = 0$ in the first row and $u \sim \phi^\star$ in the second row.

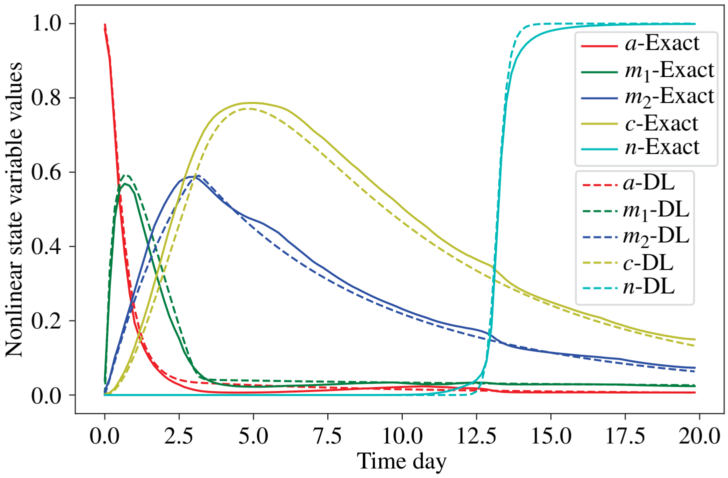

Figure 4. Results of all time-dependent variables in the nonlinear wound healing model (solid curves) and its deep learning decoder approximation (dashed curves) at wound centre ($\tilde{r} = 0$ mm).

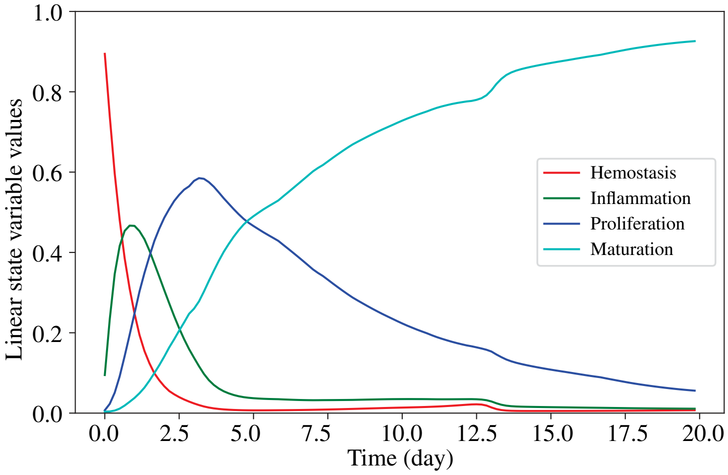

Figure 5. Results of time dependence of all variables in the linear representation.

where x_t is the state from the nonlinear dynamics (3.1*a*)–(3.1*e*) with control input u_t, $\widehat{x}_t^\star$ is the target state decoded by $\widehat{h}^{\widehat{\theta}}$ from the linear state $z_t^\star$ with control input u_t obtained by equation (2.15). If

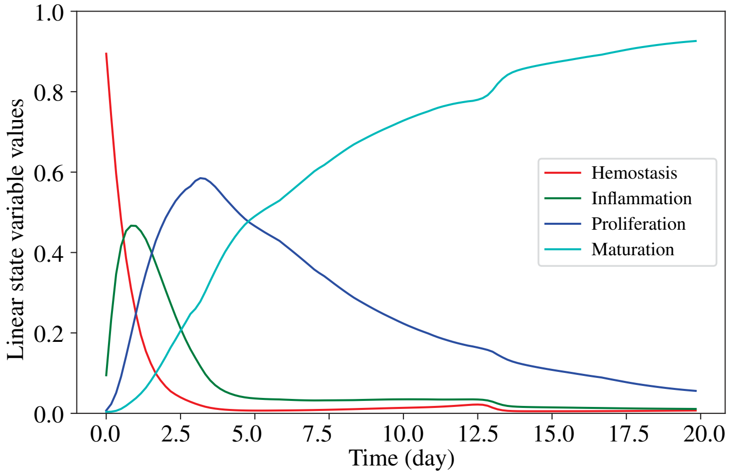

Figure 6. Results of wound size versus time: wound healing time is 7.53 days using $\phi^{\vartheta^\star}$ treatment, 10.67 days with $\tilde{\phi}^{\vartheta^\star}$ treatment, 13.83 days without any treatment and 11.33 days with BFS treatment from Zlobina *et al.* [9].

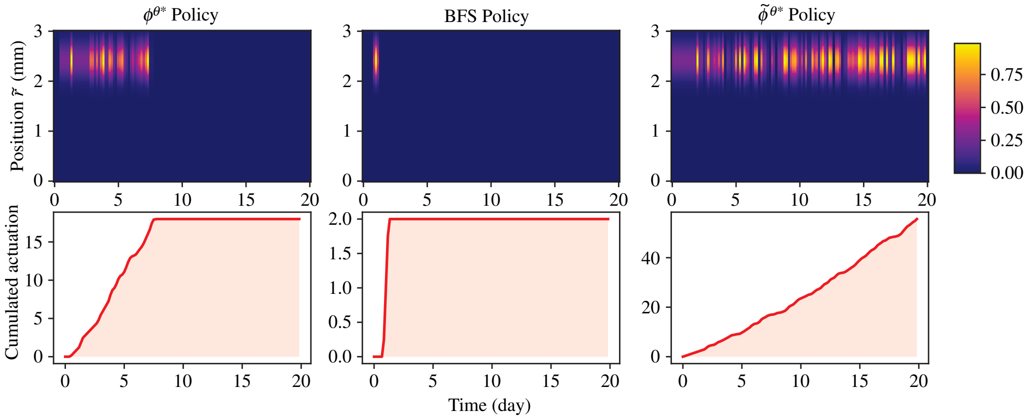

Figure 7. Different policies of wound treatment. The first row gives plots of spatial–temporal actuation given by policies $\phi^{\vartheta^\star}$, BFS and $\tilde{\phi}^{\vartheta^\star}$. The second row gives the cumulated actuation over time.

such target state $\hat{x}_t^\star$ is unavailable, a much smaller reward, i.e. −2, is assigned, which will drive the DRL agent to learn a policy to stabilize the linear representation and track the optimal trajectory with optimal input for this linear system.

One may consider the use of the target state for the DRL agent to track is unnecessary. Instead, a more intuitive definition of reward function could be

$$r^\circ(x_t, u_t) = -\mathbf{1}\{n_t \le 0.95\}. \tag{3.6}$$

We then have two datasets for training the DRL agent:

$$\mathscr{M} = \{x_t, u_t, r(x_t, u_t), x_{t+1} : t \ge 0)\text{and}$$
$$\tilde{\mathscr{M}} = \{x_t, u_t, r^\circ(x_t, u_t), x_{t+1} : t \ge 0\}.$$

We denote the resulting treatment policies as $\phi^{\vartheta^\star}$ and $\tilde{\phi}^{\vartheta^\star}$ respectively.

Using $\tilde{\mathscr{M}}$ will directly minimize the wound's healing time based solely on the feedback from the nonlinear dynamics without tracking any target state as discussed in Lewis *et al.* [48]. However, we show in the experiment that $\tilde{\phi}^{\vartheta^\star}$ is not a safe and economical treatment policy compared with $\phi^{\vartheta^\star}$.

The progressions of wound healing in terms of wound size are shown in figure 6, where we also compared the performances of the treatment strategy given by Zlobina *et al.* [9]. It can be seen from

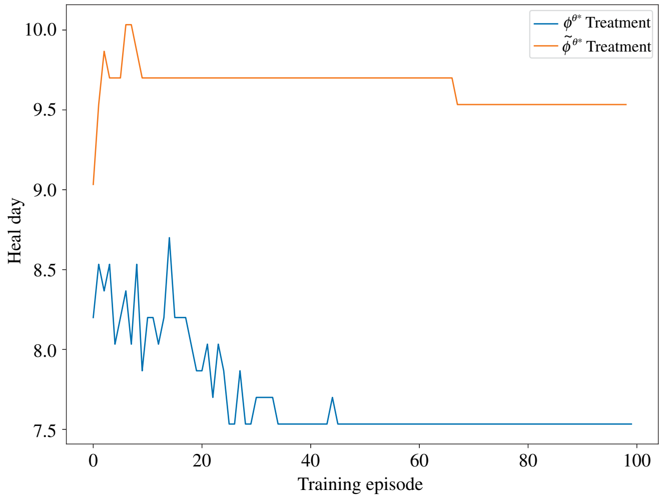

Figure 8. Comparison of expected heal days using different treatment strategies. The DRL algorithm used in both strategies is the A2C.

figure 6 that the treatment strategy $\phi^{\vartheta^\star}$ reduces the healing time the most, around 45.56% time of reduction compared to the one without any treatment and 33.54% time of reduction compared to the one with treatment by BFS [9].

The strategy coming from equation (3.5) also provides a safety advantage over the one from equation (3.6) as can be seen from figure 7. As is discovered in Zlobina *et al.* [9], it would be dangerous to apply any treatment at the early stage of the wound, which corresponds to zero actuation indicated by $\phi^{\vartheta^\star}$ and BFS policy shown in figure 7. However, such avoidance in the danger zone is not captured by $\tilde{\phi}^{\vartheta^\star}$, as is shown in the third column plot of figure 7. Compared with the plots of $\phi^{\vartheta^\star}$ and $\tilde{\phi}^{\vartheta^\star}$, we can also observe that $\phi^{\vartheta^\star}$ will stop actuation when the wound has healed, while $\tilde{\phi}^{\vartheta^\star}$ keeps actuating. This indicates that the proposed closed-loop control framework is also superior in giving treatment at lower doses.

4. Discussion

New biotechnologies have introduced a multitude of sensors for various biological systems, addressing a growing need in medicine to integrate these sensors into closed-loop control systems. However, the complexity of biological processes presents a challenge in formulating accurate mathematical models; thus, there is a demand for control algorithms that do not rely on precise models. While sensors provide valuable insight, their measurements only partially capture the dynamics of real biological systems.

Wound healing serves as an example of a nonlinear process with diverse roles played by different cell types across various stages. In our study, we operate under the assumption that a sensor reflecting wound stages is available [38], and it provides information that can be easily approximated by a linear system of ODEs.

However, owing to discrepancies between measurements and the actual biological components involved in wound healing, linear systems fail to align perfectly with the underlying nonlinear processes.

Nevertheless, we demonstrate the feasibility of monitoring and controlling nonlinear systems through observations derived from linear approximations. We assert that this approach holds promise for a broad spectrum of nonlinear biological processes for which sensors have been developed, but accurate mathematical modelling remains difficult.

Finally, it is worth noting that there remain many other more advanced DRL algorithms that can further improve the control strategy and sample efficiency of the proposed algorithm. For exmaple, when it comes to large state and action spaces, DQN will take much longer time to learn an optimal control strategy and can often fall into local minima. In appendix A, we replaced it with Advantage Actor-Critic (A2C) [49] and showed that it only took around 200 episodes for the proposed algorithm to find a similar treatment strategy and outperformed the one from directly optimizing A2C over the nonlinear system. As our main goal is to propose an adaptive learning structure on how to combine deep learning, optimal control and DRL for accelerating wound healing, we would like to design a better DRL algorithm for future studies.

5. Conclusion

In this article, we propose an adaptive closed-loop control framework for a nonlinear dynamical system. The controller integrates deep learning, optimal control and RL, aiming to accelerate nonlinear biological processes such as wound healing without the need for mathematical modelling. We have demonstrated that the proposed method not only significantly improves wound healing time but also addresses safety concerns and reduces drug usage.

Further development of the controller with more advanced DLR algorithms, as well as its implementation in *in vivo* experiments, will ultimately lead to significant improvements in wound care and broader medical domains leveraging intelligent control algorithms.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. Data and relevant code for this research work are stored in GitHub [50] and have been archived within the Zenodo repository [51].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. F.L.: conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization, writing—original draft, writing—review and editing; K.Z.: conceptualization, formal analysis, methodology, visualization, writing—original draft, writing—review and editing; N.A.R.: data curation, software, writing—review and editing; S.T.: resources; M.G.: conceptualization, formal analysis, funding acquisition, methodology, project administration, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This study was supported by the SciAI Center and funded by the Office of Naval Research (ONR) under Grant Number N00014-23-1-2729 and the DARPA Biotechnologies Office (DARPA/BTO) under Cooperative Agreement Number DC20AC00003.

Disclaimer. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research and the DARPA Biotechnologies Office (DARPA/BTO) or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation herein.

Appendix. A

We further explored replacing the DQN algorithm in the proposed algorithm with the Advantage Actor-Critic algorithm (A2C) [49] and comparing it with directly finding control policies with A2C through interactions with nonlinear dynamics. As a result, we have two types of control policy ϕ^{θ^*} obtained from optimizing A2C agent using $\mathscr{M}$ and $\tilde{\phi}^{\theta^*}$ obtained from optimizing A2C agent using $\tilde{\mathscr{M}}$. Definitions of $\mathscr{M}$ and $\widetilde{\mathscr{M}}$ can be found in §3.2.

Note that when making the comparison, we keep the neural network structures, optimizers and hyperparameters, such as learning rate, discount factor, etc., all the same.

As can be seen from figure 8, while A2C is still learning the strategy when it is directly optimized over the nonlinear system, our proposed algorithm has converged to the treatment strategy that results in much shorter healing days. This further reveals our algorithm's advantages in sample efficiency.

References

1. Zlobina K, Jafari M, Rolandi M, Gomez M. 2022 The role of machine learning in advancing precision medicine with feedback control. *Cell Rep. Phys. Sci.* **3**, 101149. (doi:10.1016/j.xcrp.2022.101149)
2. Dias R, Torkamani A. 2019 Artificial intelligence in clinical and genomic diagnostics. *Genome Med.* **11**, 70. (doi:10.1186/s13073-019-0689-8)
3. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. 2017 Artificial intelligence in precision cardiovascular medicine. *J. Am. Coll. Cardiol.* **69**, 2657–2664. (doi:10.1016/j.jacc.2017.03.571)
4. Roden DM, George AL. 2002 The genetic basis of variability in drug responses. *Nat. Rev. Drug Discov.* **1**, 37–44. (doi:10.1038/nrd705)

5. Bielinski SJ *et al*. 2014 Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time—using genomic data to individualize treatment protocol. *Mayo Clin. Proc.* **89**, 25–33. (doi:10.1016/j.mayocp.2013.10.021)
6. Reinke JM, Sorg H. 2012 Wound repair and regeneration. *Eur. Surg. Res.* **49**, 35–43. (doi:10.1159/000339613)
7. Portou MJ, Baker D, Abraham D, Tsui J. 2015 The innate immune system, toll-like receptors and dermal wound healing: a review. *Vascul. Pharmacol.* **71**, 31–36. (doi:10.1016/j.vph.2015.02.007)
8. Gholami B, Haddad WM, Bailey JM, Tannenbaum AR. 2013 Optimal drug dosing control for intensive care unit sedation by using a hybrid deterministic–stochastic pharmacokinetic and pharmacodynamic model. *Optim. Control Appl. Methods* **34**, 547–561. (doi:10.1002/oca.2038)
9. Zlobina K, Xue J, Gomez M. 2022 Effective spatio-temporal regimes for wound treatment by way of macrophage polarization: a mathematical model. *Front. Appl. Math. Stat.* **8**, 791064. (doi:10.3389/fams.2022.791064)
10. Furutani E, Tsuruoka K, Kusudo S, Shirakami G, Fukuda K. 2010 A hypnosis and analgesia control system using a model predictive controller in total intravenous anesthesia during day-case surgery. In *Proc. SICE Annual Conf. 2010, Taipei, Taiwan, 18–21 August 2010*, pp. 223–226.
11. Soltesz K, Hahn JO, Hägglund T, Dumont GA, Ansermino JM. 2013 Individualized closed-loop control of propofol anesthesia: a preliminary study. *Biomed. Signal Process. Control* **8**, 500–508. (doi:10.1016/j.bspc.2013.04.005)
12. Hahn JO, Dumont GA, Ansermino JM. 2012 Robust closed-loop control of hypnosis with propofol using wavcns index as the controlled variable. *Biomed. Signal Process. Control* **7**, 517–524. (doi:10.1016/j.bspc.2011.09.001)
13. Zhao Y, Zeng D, Socinski MA, Kosorok MR. 2011 Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* **67**, 1422–1433. (doi:10.1111/j.1541-0420.2011.01572.x)
14. Martín-Guerrero JD, Gomez F, Soria-Olivas E, Schmidhuber J, Climente-Martí M, Jiménez-Torres NV. 2009 A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients. *Expert Syst. Appl.* **36**, 9737–9742. (doi:10.1016/j.eswa.2009.02.041)
15. Padmanabhan R, Meskin N, Haddad WM. 2017 Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. *Math. Biosci.* **293**, 11–20. (doi:10.1016/j.mbs.2017.08.004)
16. Daskalaki E, Diem P, Mougiakakou SG. 2013 Personalized tuning of a reinforcement learning control algorithm for glucose regulation. In *2013 35th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013*, pp. 3487–3490. (doi:10.1109/EMBC.2013.6610293)
17. Moore BL, Pyeatt LD, Kulkarni V, Panousis P, Padrez K, Doufas AG. 2014 Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers. *J. Mach. Learn. Res.* **15**, 655–696.
18. Padmanabhan R, Meskin N, Haddad WM. 2015 Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. *Biomed. Signal Process. Control* **22**, 54–64. (doi:10.1016/j.bspc.2015.05.013)
19. Padmanabhan R, Meskin N, Haddad WM. 2019 Optimal adaptive control of drug dosing using integral reinforcement learning. *Math. Biosci.* **309**, 131–142. (doi:10.1016/j.mbs.2019.01.012)
20. Zhou Y, Lu F, Pu G, Ma X, Sun R, Chen HY, Li X. 2019 Adaptive leader-follower formation control and obstacle avoidance via deep reinforcement learning. In *2019 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019*, pp. 4273–4280. (doi:10.1109/IROS40897.2019.8967561)
21. Proctory JL, Bruntonz SL, Kutzx JN. 2018 Generalizing Koopman theory to allow for inputs and control. *SIAM J. Appl. Dyn. Syst.* **17**, 909–930. (doi:10.1137/16M1062296)
22. Šiljak DD, Zečević AI. 2005 Control of large-scale systems: beyond decentralized feedback. *Annu. Rev. Control* **29**, 169–179. (doi:10.1016/j.arcontrol.2005.08.003)
23. Antoulas AC. 2005 An overview of approximation methods for large-scale dynamical systems. *Annu. Rev. Control* **29**, 181–190. (doi:10.1016/j.arcontrol.2005.08.002)
24. Van Overschee P, De Moor B. 2012 *Subspace identification for linear systems: theory—implementation—applications*. Dordrecht, The Netherlands: Springer Science & Business Media.
25. Paraskevopoulos PN. 2017 *Modern control engineering*. Boca Raton, FL: CRC Press. (doi:10.1201/9781315214573)
26. Koopman BO. 1931 Hamiltonian systems and transformation in Hilbert space. *Proc. Natl Acad. Sci. USA* **17**, 315–318. (doi:10.1073/pnas.17.5.315)
27. Mezić I, Banaszuk A. 2004 Comparison of systems with complex behavior. *Physica D* **197**, 101–133. (doi:10.1016/j.physd.2004.06.015)
28. Mezić I. 2005 Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dyn.* **41**, 309–325. (doi:10.1007/s11071-005-2824-x)
29. Lian Y, Wang R, Jones CN. 2021 Koopman based data-driven predictive control. (doi:https://arxiv.org/abs/2102.05122#)
30. Kaiser E, Kutz JN, Brunton SL. 2021 Data-driven discovery of Koopman eigenfunctions for control. *Mach. Learn.* **2**, 035023. (doi:10.1088/2632-2153/abf0f5)
31. Ahmed A, del Rio-Chanona EA, Mercangöz M. 2022 Learning linear representations of nonlinear dynamics using deep learning. *IFAC-PapersOnLine.* **55**, 162–169. (doi:10.1016/j.ifacol.2022.07.305)
32. Melai BB *et al*. 2016 A graphene oxide pH sensor for wound monitoring. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2016**, 1898–1901. (doi:10.1109/EMBC.2016.7591092)
33. Seo HS, Lim H, Lim T, Seo K, Yang J, Kang Y, Han SJ, Ju S, Jeong SM. 2024 Facile and cost-effective fabrication of wearable alpha-naphtholphthalein-based halochromic sensor for wound pH monitoring. *Nanotechnology* **35**, 245502. (doi:10.1088/1361-6528/ad321a)
34. Schreml S, Meier RJ, Weiß KT, Cattani J, Flittner D, Gehmert S, Wolfbeis OS, Landthaler M, Babilas P. 2012 A sprayable luminescent pH sensor and its use for wound imaging in vivo. *Exp. Dermatol.* **21**, 951–953. (doi:10.1111/exd.12042)

35. Mirani B, Hadisi Z, Pagan E, Dabiri SMH, van Rijt A, Almutairi L, Noshadi I, Armstrong DG, Akbari M. 2023 Smart dual-sensor wound dressing for monitoring cutaneous wounds. *Adv. Healthc. Mater.* **12**, e2203233. (doi:10.1002/adhm.202203233)

36. Zheng XT *et al*. 2023 Battery-free and AI-enabled multiplexed sensor patches for wound monitoring. *Sci. Adv.* **9**, eadg6670. (doi:10.1126/sciadv.adg6670)

37. Tang N, Zheng Y, Jiang X, Zhou C, Jin H, Jin K, Wu W, Haick H. 2021 Wearable sensors and systems for wound healing-related pH and temperature detection. *Micromachines* **12**, 430. (doi:10.3390/mi12040430)

38. Carrión H, Jafari M, Yang HY, Isseroff RR, Rolandi M, Gomez M, Norouzi N. 2022 HealNet: self-supervised acute wound heal-stage classification. In *Machine learning in medical imaging*, pp. 446–455. Cham, Switzerland: Springer. (doi:10.1007/978-3-031-21014-3_46)

39. Ogata K. 2010 *Modern control engineering*, 5th edn. Upper Saddle River, NJ: Pearson.

40. Sasane A. 2016 *Optimization in function spaces*. Mineola, NY: Courier Dover Publications.

41. Hinton GE, Salakhutdinov RR. 2006 Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507. (doi:10.1126/science.1127647)

42. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller MA. 2013 Playing Atari with deep reinforcement learning. (doi:https://arxiv.org/abs/1312.5602)

43. Bertsekas D, Tsitsiklis JN. 1996 *Neuro-dynamic programming*. Belmont, MA: Athena Scientific.

44. He K, Zhang X, Ren S, Sun J. 2015 Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In *2015 IEEE Int. Conf. on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015*, pp. 1026–1034. (doi:10.1109/ICCV.2015.123)

45. Paszke A *et al*. 2017 Automatic differentiation in Pytorch. In *Proc. 31st Int. Conf. on Neural Information Processing Systems (NIPS 2017)*, *Long Beach, CA, USA, 4–9 December 2017*.

46. Kingma DP, Ba J. 2014 Adam: a method for stochastic optimization (doi:https://arxiv.org/abs/1412.6980)

47. Polyak BT, Juditsky AB. 1992 Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30**, 838–855. (doi:10.1137/0330046)

48. Lewis FL, Vrabie DL, Syrmos VL. 2012 *Optimal control*. Hoboken, NJ: John Wiley & Sons. (doi:10.1002/9781118122631)

49. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K. 2016 Asynchronous methods for deep reinforcement learning. *Proc. Mach. Learn. Res.* **48**, 1928–1937.

50. GitHub. Enhancingwoundhealingusingdrl. See https://github.com/Fan-Lu/EnhancingWoundHealingUsingDRL.

51. Lu F. 2024 Fan-lu/enhancingwoundhealingusingdrl: enhancing wound healing via deep reinforcement learning for optimal therapeutics (v1.0.0). *Zenodo*. See https://doi.org/10.5281/zenodo.11497170.