Machine Learning: Science and Technology

Extending the relative seriality formalism for interpretable deep learning of normal tissue complication probability models

To cite this article: Tahir I Yusufaly 2022 *Mach. Learn.: Sci. Technol.* **3** 024001

MACHINE
LEARNING
Science and Technology

CrossMark

OPEN ACCESS

RECEIVED
20 January 2022

REVISED
23 March 2022

ACCEPTED FOR PUBLICATION
21 April 2022

PUBLISHED
13 May 2022

PAPER

Extending the relative seriality formalism for interpretable deep learning of normal tissue complication probability models

Tahir I Yusufaly

Johns Hopkins Department of Radiology and Radiological Sciences, Baltimore, MD, 21287, United States of America

E-mail: tyusufa2@jhmi.edu

Keywords: deep learning, NTCP, relative seriality, interpretability

Abstract

We formally demonstrate that the relative seriality (RS) model of normal tissue complication probability (NTCP) can be recast as a simple neural network with one convolutional and one pooling layer. This approach enables us to systematically construct deep relative seriality networks (DRSNs), a new class of mechanistic generalizations of the RS model with radiobiologically interpretable parameters amenable to deep learning. To demonstrate the utility of this formulation, we analyze a simplified example of xerostomia due to irradiation of the parotid gland during alpha radiopharmaceutical therapy. Using a combination of analytical calculations and numerical simulations, we show for both the RS and DRSN cases that the ability of the neural network to generalize without overfitting is tied to 'stiff' and 'sloppy' directions in the parameter space of the mechanistic model. These results serve as proof-of-concept for radiobiologically interpretable deep learning of NTCP, while simultaneously yielding insight into how such techniques can robustly generalize beyond the training set despite uncertainty in individual parameters.

1. Introduction

In radiotherapy, the risk of toxic side effects, as quantified by the normal tissue complication probability (NTCP), is a common dose-limiting factor during treatment planning [1]. The ability to predict the risk of such side-effects is critical in enabling clinicians to rationally and systematically sculpt dose distributions to avoid critical organs at risk (OARs). Traditional approaches to such treatment planning typically utilize simplified, easily interpretable radiobiological models, if only indirectly (e.g. through dose-volume histogram metrics that are based on an underlying radiobiological model). Examples of such approaches that are used in clinical practice include the Lyman–Kutcher–Burman model [2, 3] and the relative seriality (RS) model of Kallman *et al* [4].

While having the advantage of parsimony and interpretability, such stylized models inevitably involve simplifying approximations that neglect the underlying complexity of the radiobiological response. In recent years, the rise of big data approaches has therefore motivated significant interest in machine learning (ML) more complex dose-toxicity models with enhanced predictive power [5], using the entire 3D voxel-by-voxel dose distribution. Of these approaches, artificial neural networks (ANNs) have shown especially promising success [6–8].

However, a common criticism of ML methods, especially ANNs, is that they are often black boxes, with model structures and parameters that have unclear clinical meaning [9, 10]. This opacity is a major cause for concern among clinicians, and frequently results in hesitation applying ML models to clinical decision making and treatment planning. A framework for interpreting ML dose-toxicity models, and particularly for relating them to more conventional and intuitive radiobiological models, would therefore be useful in persuading clinicians to adopt ML more routinely during patient treatment. In this paper, such a framework is presented.

We start in section 2 with a theoretical overview of the basics of ANNs, and show that the RS model can be recast as a particularly simple ANN. We then demonstrate that this provides a natural way of understanding the radiobiology of various additional architectural complexities. This enables us to define

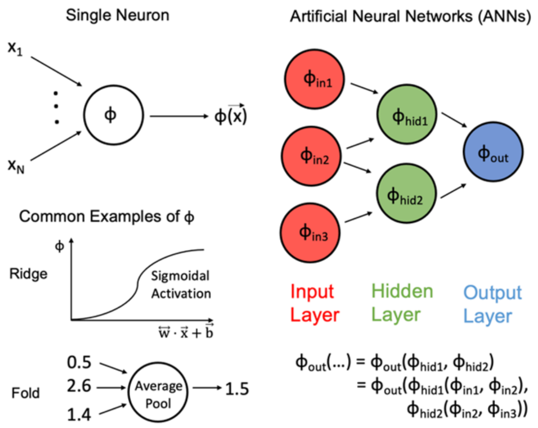

Figure 1. Diagrammatic illustrations of a single neuron (top left), along with some common activation functions (bottom left), and an example (right) of how more general multilayer ANN architectures are formed through function composition, including input, hidden and output layers. The parameters w and b denote weights and biases, respectively, which form a linear input combination that goes into a ridge activation.

deep relative seriality networks (DRSNs), a set of mechanistically interpretable ANNs capable of modeling more complex patterns of radiation response.

In section 3, we apply our framework to an example problem of xerostomia in alpha radiopharmaceutical therapy (αRPT) due to irradiation of the parotid gland, which we model with both the RS and DRSN networks. The results are analyzed in section 4, where we show that the training of an interpretable neural network is a nonconvex optimization problem with a broad range of local minima, all of which are nearly equivalent to the global minimum in their ability to generalize without overfitting. To reconcile mechanistic interpretability of the parameters with this robustness to precise parameter values, we draw on the concept of 'sloppiness' introduced by theoretical physicists studying the information geometry of multiparameter models. We characterize 'stiff' and 'sloppy' directions in parameter space, corresponding to parameter changes that a model is sensitive and insensitive to, respectively. We demonstrate, for both the RS and DRSN networks, that the nonconvexity of the loss landscape is confined to the sloppy directions, such that the variability in parameters learned across different training replicates is inconsequential for performance.

Finally, in section 5 we summarize and conclude the manuscript, after briefly discussing some open questions and extensions.

2. Background and formal analysis

2.1. Brief overview of neural networks

The fundamental building block of all ANNs is the artificial neuron [11], which is a mathematical function mapping a multivariate input $\vec{x}$ onto a univariate output $\phi(\vec{x})$. ϕ is labeled the activation function of the neuron. Although this function can in principle be any nonlinear mapping, in practice the most frequently used activation functions usually fall into two categories: (a) Ridge functions, which act on a linear combination of the input variables, (b) Fold functions, which perform some sort of aggregate operation over the inputs. An example of the former is a sigmoidal activation function (e.g. tanh) and an example of the latter is a pooling operation, such as taking the maximum or average value from a given set of inputs.

ANN are constructed by 'connecting' individual neurons together through the operation of function composition. The number of neurons, along with their layout and connectivity to other neurons, define the architecture of an ANN. An example of how single neurons can be used to construct an architecture, along with the corresponding symbolic form of the composite activation function, is illustrated in figure 1. From the ANN architecture, neurons can be broadly classified as either input neurons (which are the initial network input), output neurons (which compute final network output) or hidden neurons (which serve as intermediates between inputs and outputs).

2.2. Mapping the RS model to an ANN

In the RS model [4], an OAR is composed of N functional subunits (FSUs), which in practice are often defined by the voxels of the images and dose maps. In the RS model, if the ith subunit, with relative volume v_i, is irradiated at dose D_i, then the NTCP is calculated as

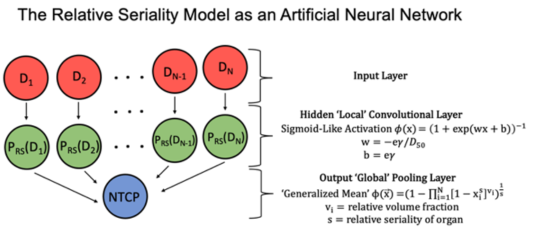
The Relative Seriality Model as an Artificial Neural Network
D_1
D_2
D_{N-1}
D_N
$P_{RS}(D_1)$
$P_{RS}(D_2)$
$P_{RS}(D_{N-1})$
$P_{RS}(D_N)$
NTCP
Input Layer
Hidden 'Local' Convolutional Layer
Sigmoid-Like Activation $\phi(x) = (1 + \exp(wx + b))^{-1}$
$w = -e\gamma/D_{50}$
$b = e\gamma$
Output 'Global' Pooling Layer
'Generalized Mean' $\phi(\vec{x}) = (1 - \prod_{i=1}^{N}[1 - x_i^s]^{v_i})^{\frac{1}{s}}$
v_i = relative volume fraction
s = relative seriality of organ

Figure 2. Schematic illustration of how the RS model maps onto an ANN with radiobiologically interpretable parameters.

$$\mathrm{NTCP} = \left\{1 - \prod_{i=1}^{N}\left[1 - P_{\mathrm{RS}}(D_{\mathrm{i}})^{s}\right]^{v_i}\right\}^{\frac{1}{s}} \tag{1}$$

where $P_{\mathrm{RS}}(D_i)$ is the probability of damage to the ith subunit, typically approximated as a sigmoid-like function

$$P_{\mathrm{RS}}(D_i) = \frac{1}{1 + \exp\left(-e\gamma\left(\frac{D_i}{\mathrm{D}_{50}} - 1\right)\right)}. \tag{2}$$

Comparing equations (1) and (2) with the examples and notation of figure 1, we see that the RS model maps onto the ANN architecture illustrated in figure 2. This architecture consists of a hidden convolutional layer acting on the input layer of doses to estimate survival probability, followed by an output pooling layer that compresses the entire dataset into a single NTCP value. We note that, in line with the strict definition of a convolutional layer, the FSUs must be identical, such that the weights and biases are translationally invariant across the layer.

Notably, the parameters in this ANN now all have clear radiobiological interpretations. The intrinsic radiosensitivity of the FSU is determined by the two model parameters D_{50} and γ, which determine the threshold and sharpness of $P_{\mathrm{RS}}(D_i)$. These are directly related to the weights and biases in the hidden layer. Meanwhile, tissue architecture and dose-volume effects are encoded in the pooling layer parameter s, which quantifies the ratio of serial FSUs to total FSUs in the OAR. Large values ($\approx$1) of s indicate a serial structure, such as the spinal cord, in which damage to a single FSU is sufficient to damage the entire OAR. On the other hand, small values ($\ll$1) indicate a parallel structure, such as the kidneys, where the individual FSUs act independently, and a critical number of FSUs must be destroyed to disrupt function.

2.3. DRSNs: radiobiologically interpretable RS extensions

A conscientious and skeptical reader might reasonably point out that, while the mathematical reformulation in the previous section might be interesting, its practical utility remains uncertain. In particular, the deep networks used in state-of-the-art machine learning of NTCP often have significantly more complex architectures than the simplified setup of the RS model. Thus, one might justifiably wonder whether thinking in terms of the RS model will yield any meaningful insight into practical deep learning models.

As a first step towards addressing such objections, in this section we will describe how starting from the RS-based ANN provides an intuitive and useful way of interpreting more sophisticated extensions of this architecture. Concretely, we will demonstrate this using a specific class of architectures, which we name DRSNs.

The DRSN architecture is illustrated in figure 3. It is characterized by two critical differences from the RS model: (a) the inclusion of denser feedforward connections in the convolutional layer, and (b) the addition of a second hidden pooling layer in between the first pooling layer and the final NTCP output. Radiobiologically, the first difference can be interpreted as describing off-target damage, whether from bystander signaling or larger-scale processes such as inflammation [12], and the second difference can be interpreted as describing the hierarchical, modular organization of FSUs in the OAR, a concept that has been previously described [13] as 'meta-FSUs'. In the following subsections, we make these connections quantitative.

2.3.1. Off-targeted effects lead to denser connectivity in convolutional layer

To demonstrate the interpretation of extra feedforward connections in the convolutional layer, it will be convenient to start with a particularly simple example, consisting of two neurons each in the input layer and the convolutional layer. This setup is illustrated in figure 4, where we note that, in addition to the original

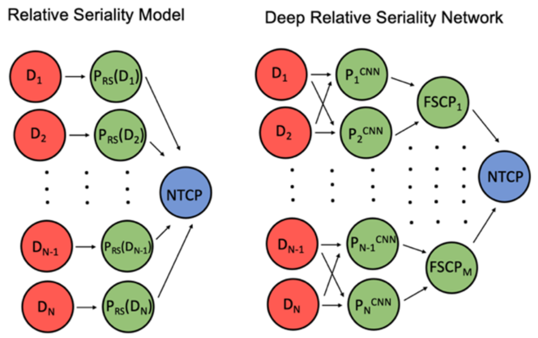

Figure 3. Writing the RS model as an ANN (left) makes it easier to construct a more sophisticated yet interpretable class of architectures (right), named deep relative seriality networks (DRSNs). The definitions of the various new notation terms in the DRSN will be described in the main text.

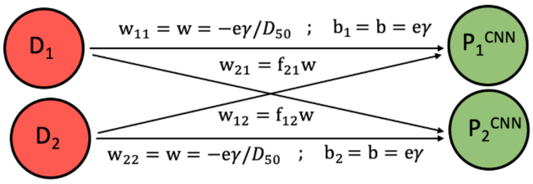
D1
D2
w11 = w = −eγ/D50 ; b1 = b = eγ
w21 = f21w
w12 = f12w
w22 = w = −eγ/D50 ; b2 = b = eγ
P1CNN
P2CNN

Figure 4. A simple convolutional layer with two neurons and two cross-weights. In addition to the local weight w and bias b from the RS model, we now have cross-weight terms, which are conveniently parametrized in terms of their ratios to the local weight.

local RS interpretable weight w and bias b that we already had in figure 2, we now also have two new 'cross-weights' connecting input neuron 1 to convolutional neuron 2, and vice versa. For reasons that will soon become apparent, it will be convenient to parameterize these cross-weights in terms of their ratio to w, let us call it f.

Let us take the RS activation function from equation (2), and generalize it to a 'CNN' function that operates on a linear combination of the input doses

$$P_1^{\mathrm{CNN}} = \frac{1}{1+\exp\left(w_{11}D_1 + w_{12}D_2 + b_1\right)} = \frac{1}{1+\exp\left(-e\gamma\left(\frac{(D_1+f_{12}D_2)}{D_{50}} - 1\right)\right)} \tag{3}$$

$$P_2^{\mathrm{CNN}} = \frac{1}{1+\exp\left(w_{21}D_1 + w_{22}D_2 + b_2\right)} = \frac{1}{1+\exp\left(-e\gamma\left(\frac{(f_{21}D_1+D_2)}{D_{50}} - 1\right)\right)}. \tag{4}$$

Comparing these with equation (2), we see that the effects of the new cross-weights can be conveniently described by imagining that there is an 'effective local' dose that is input into the RS activation function,

$$P_1^{\mathrm{CNN}} = P_{\mathrm{RS}}\left(D_1^{\mathrm{eff}} = D_1 + f_{12}D_2\right) \tag{5}$$

$$P_2^{\mathrm{CNN}} = P_{\mathrm{RS}}\left(D_2^{\mathrm{eff}} = f_{21}D_1 + D_2\right). \tag{6}$$

This can be radiobiologically understood as off-target doses resulting in indirect DNA damage to neighboring FSUs, via bystander signaling, inflammation or other biological processes [12]. The strength of the interconnection weight relative to the local RS weight tells us the relative contribution of such off-target damage relative to local damage. Generalizing beyond the simple example presented here to an arbitrary number of neurons, increasing the density of feedforward connections, such that each input neuron connects to more neurons in the hidden convolutional layer, can be interpreted as increasing the spatial range of off-targeted damage.

2.3.2. Hierarchical tissue organization leads to multiple hidden pooling layers

Our analysis up to this point has implicitly assumed that there is only one 'level' of organization between the base FSUs and the aggregate OAR. However, in general, tissues and organs are characterized by hierarchical structure, with multiple intermediate levels of organization. In other words, it is usually more accurate to think of the FSUs themselves as being composed of an even more fine-grained set of 'meta-FSUs' [13]. This

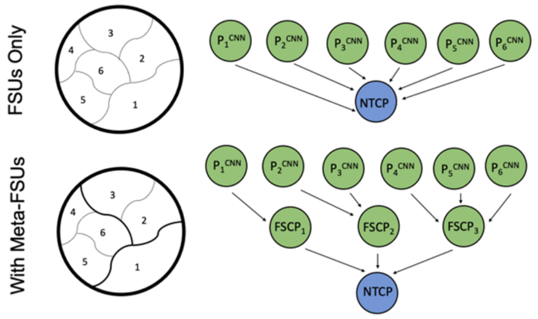

Figure 5. Generalization of a single pooling layer (top) when only allowing for FSUs leads to hidden pooling layers (bottom) describing meta-FSUs.

argument can be straightforwardly extended to allow for more than one organization level, but for the sake of simplicity, we restrict ourselves to just one intermediate layer here.

To take a concrete example, consider an OAR built out of six FSUs, as shown in figure 5. If we assume no meta-FSUs, then the pooling layer for this is a straightforward analog of the architecture shown in figure 2.

However, suppose that now, we allowed for an intermediate level of organization, such that the six FSUs now become six meta-FSUs. The first meta-FSU would then by itself form an FSU, the second and third meta-FSUs would aggregate into a second FSU, and the fourth, fifth and sixth meta-FSUs would aggregate into a third FSU. Then, to calculate the NTCP in a way analogous to equation (1), the relevant set of probabilities that need to be pooled over is the set of functional subunit complication probabilities (FSCPs):

$$\mathrm{NTCP} = \left\{1 - \prod_{i=1}^{3}\left[1 - \mathrm{FSCP}_i^{s}\right]^{v_i}\right\}^{\frac{1}{s}}. \tag{7}$$

To calculate these FSCPs, in turn, we pool over the relevant base meta-FSUs:

$$\mathrm{FSCP}_1 = \mathrm{P}_1^{\mathrm{CNN}} \tag{8}$$

$$\mathrm{FSCP}_2 = \left(1 - \left(1 - \mathrm{P}_2^{\mathrm{CNN}\ \mathrm{s_{F2}}}\right)^{\mathrm{v}_{22}}\left(1 - \mathrm{P}_3^{\mathrm{CNN}\ \mathrm{s_{F2}}}\right)^{\mathrm{v}_{23}}\right)^{1/\mathrm{s_{F2}}} \tag{9}$$

$$\mathrm{FSCP}_3 = \left(1 - \left(1 - \mathrm{P}_4^{\mathrm{CNN}\,\mathrm{s_{F3}}}\right)^{\mathrm{v}_{34}}\left(1 - \mathrm{P}_5^{\mathrm{CNN}\,\mathrm{s_{F3}}}\right)^{\mathrm{v}_{35}}\left(1 - \mathrm{P}_5^{\mathrm{CNN}\,\mathrm{s_{F3}}}\right)^{\mathrm{v}_{36}}\right)^{1/\mathrm{s_{F3}}}. \tag{10}$$

While equation (8) is trivial since there is only one FSU in the meta-FSU, more generally we see that for each FSCP_i, we must specify a distinct exponent $\mathrm{s_{Fi}}$ and a distinct set of relative volume fractions v_{ij}, for each of the j meta-FSUs that make up this ith FSU.

While the symbolic formalism in equations (7)–(10) can very rapidly become cumbersome, we see that the corresponding graphical representation in figure 5 is much easier to manipulate and work with. In this way, when deep learning different architectures with different numbers and organizations of intermediate layers, we can more flexibly adapt our models to represent more general kinds of functional organization than are usually described with the RS model.

It is worth noting that this approach can be straightforward generalized to 'multi-channel' inputs and outputs. For instance, rather than assuming that each input neuron comes with a known FSU label, we can instead take the raw image data, perform autosegmentation on it to generate N channels each corresponding to a separate segmentation mask, and combine those channels with a 3D dose map to generate our DRSN input dose layer from scratch. Additionally, instead of just learning a single toxicity outcome, we can equip our intermediate and output neurons with multiple channels, each corresponding to a different clinical endpoint that we wish to predict. Although we will not address these generalizations here, they are important avenues for future work.

3. Case study: xerostomia in αRPT

In this section, we will apply our developed formalisms to analyze deep learning, for both the RS and DRSN models, in the context of xerostomia during αRPT. We start with a brief clinical overview motivating both RS

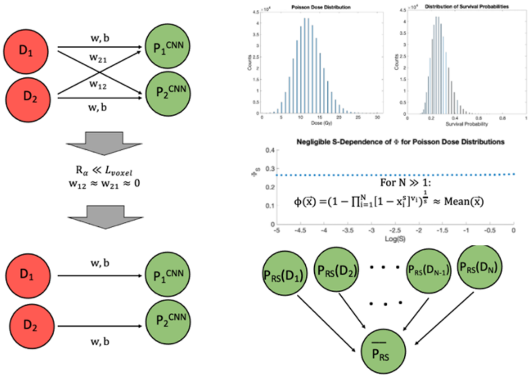

Figure 6. RPT-specific approximations, including neglect of off-target weights due to short range (left) and replacement of generalized mean pooling with global average pooling due to the nature of the Poisson distribution (right).

and DRSN models of salivary gland toxicity. We follow this with a mathematical model of radionuclide dosimetry, and show how it leads to useful approximations, specific to the context of αRPT, in both the RS and DRSN network architectures. We then show that information geometric reasoning, based on the concept of sloppiness when fitting multiparameter models, yields quantitative intuition into why neural networks generalize without overfitting despite significant redundancy in network parameters.

3.1. Clinical background and significance of salivary gland radiobiology

Xerostomia arises from irradiation of the salivary glands, most notably the parotid gland, and is a common dose-limiting toxicity for several αRPT agents currently under development [14, 15]. Foremost among these is 225Ac-PSMA-617, which has demonstrated promise in treating metastatic castration-resistant prostate cancer. However, the mechanisms of xerostomia in αRPT remain incompletely understood.

Conventional radiobiology, based on experience from external beam radiation therapy (EBRT), suggests that parotid glands can be modeled as RS-like parallel organs [16], with each voxel of a parotid gland contour constituting an individual FSU. The NTCP is then predominantly determined by the average dose. However, recent evidence from αRPT is inconsistent with this [17], with patients reporting xerostomia despite average biologically effective doses well below expected EBRT toxicity thresholds.

A proposed resolution to this paradox has been to consider a more refined, DRSN-like model of parotid gland radiobiology [18]. Specifically, rather than assume that all voxels in the parotid gland are identical and organized in parallel, a more reliable approach may be to segment the parotid gland contour into two sub-contours: (a) an acinar region composed of parallel organized voxels responsible for the production and secretion of saliva, and (b) a ductal region composed of serially organized voxels responsible for excreting the saliva into the mouth. The acinar and ductal regions are then serially linked to form the aggregate parotid gland architecture. Initial estimates based on this model suggest it as a plausible mechanism for the unexpectedly high toxicities observed in early αRPT treatments.

3.2. Radionuclide dosimetry results in further simplified neural network architectures

As a reasonable first approximation to αRPT dosimetry, we can take the dose to the N parotid gland voxels to be Poisson-distributed [19] with an average dose $\langle \mathrm{D} \rangle$,

$$p(D_i) = \frac{\langle D \rangle^{D_i} \mathrm{e}^{-\langle D \rangle}}{D_i!}, i = 1, \ldots, N. \tag{11}$$

For the case of αRPT, we may make two further simplifying approximations, as illustrated in figure 6. First, note that the characteristic length scale of a voxel L_{voxel} is typically on the order of a few millimeters, while R_α, the characteristic range of emitted alpha particles R_α is typically around tens of microns [20], orders of magnitude smaller. As a result, for now we can ignore off-target effects. Additionally, we note that for the case of a Poisson distribution of dose and survival probabilities (based on QUANTEC parameters), and for large numbers of voxels $N \gg 1$, the generalized pooling function Φ shows negligible variation with the seriality parameter s. Thus, we can safely replace it with a fixed global average pooling, with no learnable parameters.

It is worth emphasizing that approximating Φ with the global average is only justified if $N \gg 1$. In particular, for the DRSN model, the final pooling of the serially-linked acinar and ductal neurons still requires an explicit evaluation of Φ. However, we note that, since s is defined as the fraction of serial subunits

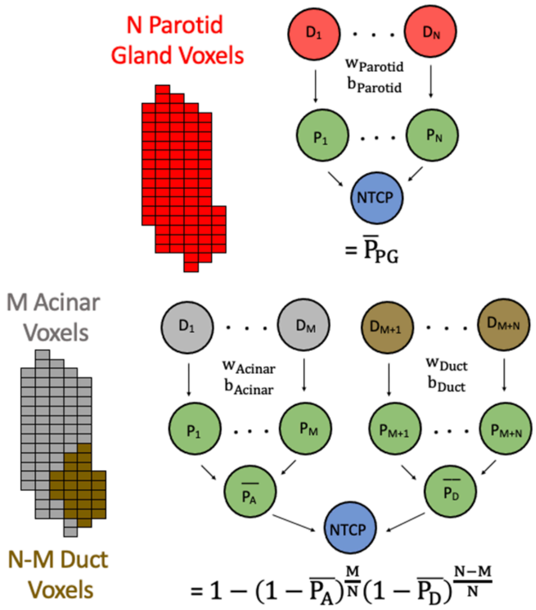

Figure 7. RS (top) and DRSN (bottom) network architectures for xerostomia during αRPT.

in a structure, for the case of two neurons the only two physically sensible values it can take are $s = 0$ or 1, corresponding to two parallel or serial linked subunits respectively. For this specific problem, we can thus assign s a fixed value of 1, as is appropriate for serially linked acinar and duct subunits.

With these approximations in place, we arrive at the simplified set of RS and DRSN networks illustrated in figure 7. The RS model assigns each voxel in the parotid gland two learnable activation parameters, a weight and a bias, and the NTCP results from average pooling across all activations. Meanwhile the DRSN model allows for two distinct sets of learnable activation parameters, one for the M acinar voxels and the other for the N-M duct voxels, resulting in four learnable parameters in total. After separately averaging both the acinar and duct activations, the two aggregate failure rates can then be combined using the NTCP pooling function for the case of two serially linked subunits.

3.3. Simulation details and parameters

For the RS model, we take from [21] the parotid gland parameter values $(D_{50}, \gamma_{50})_{\text{PG}} = (2.22\text{ Gy}, 0.83)$, which translates to corresponding neural network parameters $(w_{\text{parotid,}}\ b_{\text{parotid}}) = (0.102\text{ Gy}^{-1}, -2.26)$. For the DRSN model, although we do not have acinar-specific or duct-specific parameters, for the purpose of demonstration we may make the ansatz that ductal cells have a somewhat higher dose threshold and a somewhat steeper dose-response. Thus, we can take $(D_{50}, \gamma_{50})_{\text{acinar}} = (2\text{ Gy}, 0.5)$ and $(D_{50}, \gamma_{50})_{\text{duct}} = (3\text{ Gy}, 0.9)$, which translates to $(w_{\text{acinar,}}\ b_{\text{acinar}}) = (0.068\text{ Gy}^{-1}, -1.36)$ and $(w_{\text{duct}}, b_{\text{duct}}) = (0.122\text{ Gy}^{-1}, -2.45)$. We also take $\langle D \rangle = 12$ Gy, which is a good qualitative estimate of a typical dose in αRPT. Additionally, for convenience, we take the parotid gland to have $N = 400$ voxels, equally split between $M = 200$ acinar voxels and 200 remaining duct voxels. As we will show later, the qualitative principles gleaned from our analysis are rather general and independent of specific parameter values, although we do comment on the influence of parameter variations where appropriate.

For both the RS and DRSN models, we generate synthetic datasets using the above-mentioned parameters. For each model, we generate 50 training samples and 50 test samples. The RS-generated datasets are then used as inputs for training an RS neural network, and likewise for the DRSN-generated datasets and a DRSN network. In all cases, we run 50 different replicates, with random initialization of parameters. The networks are trained using stochastic gradient descent with momentum 0.1, and with mean-squared-error (MSE) loss function without regularization. We used a fixed learning rate of 10^{-4} and 200 epochs of batch size 1 for the RS network, and corresponding values of 10^{-3} and 100 epochs of batch size 1 for the DRSN network. To evaluate performance and generalization, we calculate the MSE of the training and test sets for each replicate.

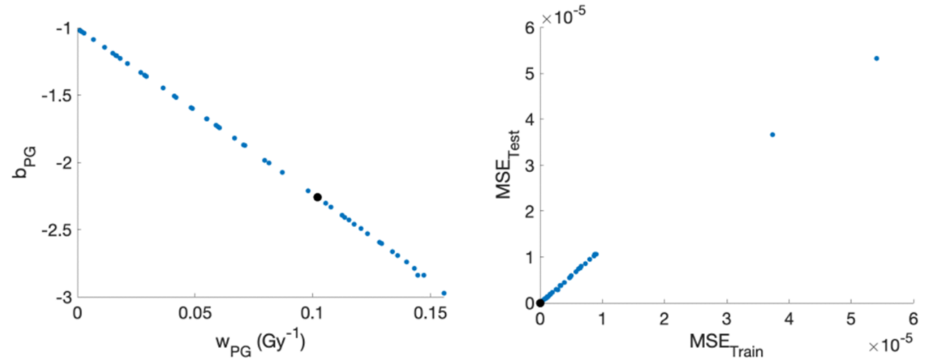

Figure 8. (Left) Distribution of learned parameters over the 50 training replicates for the RS model. The black data point indicates the true parameter value used to generate the data. (Right) Train and test MSE plots for the corresponding parameters, including the global minimum at the origin.

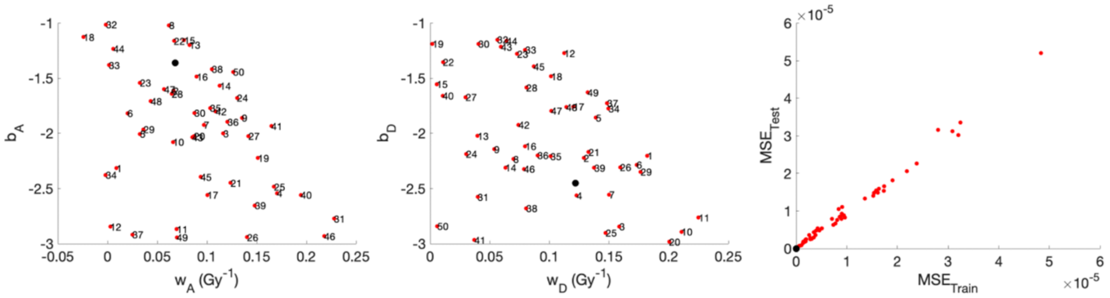

Figure 9. (Left and middle) Distribution of learned parameters, for both the acinar (left) and duct (middle) over the 50 training replicates for the DRSN model. The black data point indicates the true parameter value used to generate the data. The data points are annotated by replicate number. (Right) Train and test MSE plots for the corresponding parameters, including the global minimum at the origin.

4. Results and discussion

4.1. Results: robust generalization despite variability of trained parameters

The results of our simulations for the RS and DRSN simulations are shown in figures 8 and 9 respectively. For both cases, there is a variable distribution of learned parameters across replicates. It is worth highlighting two salient features of this variability. Firstly, the distribution of learned parameters is not random, but falls along a structured manifold in parameter space. This is evident by inspection for the case of the RS networks, where the distribution of the learned RS weight and bias, falls on a straight line. Furthermore, although it is not obvious from the initial plots, the distributions of the acinar and duct parameters when training DRSN networks also has a characteristic structure, which we will describe momentarily.

Secondly, the ability of the neural network to generalize without overfitting, such that the test MSE is not significantly greater than the train MSE, is robust to the specific replicate parameter values. It is true that the absolute values of the MSEs for both the train and test would be globally minimized to 0 by training networks that converged onto the correct parameter values. However, the ability to generalize beyond the training set without overfitting holds even when converging onto an 'incorrect' set of parameters. In addition, it is worth pointing that even though the different replicates, strictly speaking, have higher MSEs than the true global minimum value of 0, these MSEs are still extremely low ($<10^{-4}$), and any difference in performance among replicates is likely to be negligible from a practical standpoint.

4.2. 'Sloppiness' in mechanistic multiparameter models: a way of reconciling radiobiological interpretability with neural network parameter robustness

The results of these calculations are not particularly surprising to experienced practitioners of deep learning—indeed, it is well-known that many different parameter values can perform comparably on a given dataset [22]. On the other hand, the results do raise philosophical concerns about the validity of the radiobiological interpretation. When fitting a mechanistic model to data, one might think the ability of the fit to extrapolate and make predictions would depend strongly on fitting to the 'correct' parameters. However, our results here indicate that many different choices of parameters are practically equallypredictive, and the fitted model can accurately extrapolate even in the presence of substantial uncertainty of individual parameters.

Remarkably, this is precisely what has been found in recent decades by theoretical physicists studying the information geometry of fitting large multiparameter models, in areas as diverse as biology, physics, economics and engineering [23, 24]. Explicitly, the relative importance of a fitting parameter for predictions can be characterized, to a first approximation, by the Hessian of the loss function, evaluated at the true parameter values associated with the global minimum. By diagonalizing the Hessian and finding the corresponding eigenvectors and eigenvalues, we can classify parameter combinations as 'stiff' or 'sloppy', depending on whether the eigenvalues are large or small, respectively. A surprisingly universal finding for models across a range of disciplines has been that most parameter combinations are sloppy, with only a few stiff combinations being important for predictions for practical purposes.

In the following subsection, we will apply these concepts to characterize the stiff and sloppy directions in both the RS and DRSN models, but in a somewhat idiosyncratic way. Instead of explicitly calculating the complete Hessian, we start by estimating it to first order with a 'mean-field' approximation. We then show that in this approximation, by defining suitable reparameterizations of the original weights and biases, the cost function can be recast in a form that makes the calculation of the Hessian, and its eigenvalues and eigenvectors, almost trivial.

4.2.1. Identifying stiff and sloppy directions in parameter space

The mean-field approximation works by decomposing the Poisson distribution from equation (1) into two parts: a 'delta-function' perfectly localized at the mean value, and higher order corrections accounting for sample-to-sample fluctuations around the mean. The mean-field approximation is equivalent to ignoring these corrections

$$p(D_i) = \frac{\langle D\rangle^{\mathrm{D_i}} \mathrm{e}^{-\langle D\rangle}}{D_i!} = \delta(D_i - \langle D\rangle) + \text{h.o.c.} \approx \delta(D_i - \langle D\rangle),\ i = 1, \ldots, N \text{ voxels.} \tag{12}$$

Then, the average probability of FSU failure $\bar{P}$, as a function of w and b, simplifies to

$$\bar{P}(w,b) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1+\exp(w^* D_i + b)} \approx \frac{1}{1+\exp(w^*\langle D\rangle + b)}. \tag{13}$$

With the mean-field approximation in place, we now write down the loss function, which is the MSE loss for $G = 50$ data points. For the RS model, we have

$$\begin{aligned} L_{\mathrm{RS}}(w_{\mathrm{PG}}, b_{\mathrm{PG}}) &= \frac{\sum_{i=1}^{G}\left(\bar{P}(w_{\mathrm{PG}}, b_{\mathrm{PG}}) - \bar{P}\left(w_{\mathrm{parotid}}, b_{\mathrm{Parotid}}\right)\right)^2}{G} \\ &= \left(\bar{P}(w_{\mathrm{PG}}, b_{\mathrm{PG}}) - \bar{P}\left(w_{\mathrm{parotid}}, b_{\mathrm{parotid}}\right)\right)^2 \end{aligned} \tag{14}$$

where in the second equality we are using the fact that the quantity in the summation is independent of the replicate data point. Meanwhile, for the DRSN model, using analogous arguments we find that

$$L_{\mathrm{DRSN}}(w_{\mathrm{A}}, b_{\mathrm{A}}, w_{\mathrm{D}}, b_{\mathrm{D}}) = \left(\begin{array}{c} \sqrt{(1-\bar{P}(w_{\mathrm{A}}, b_{\mathrm{A}}))(1-\bar{P}(w_{\mathrm{D}}, b_{\mathrm{D}}))} - \\ \sqrt{(1-\bar{P}(w_{\mathrm{Acinar}}, b_{\mathrm{Acinar}}))(1-\bar{P}(w_{\mathrm{Duct}}, b_{\mathrm{Duct}}))} \end{array} \right)^2. \tag{15}$$

Inspecting equations (1) and (15), what we in fact observe is that in both cases, the weights and biases only influence the loss function through their influence on the NTCP evaluated for a uniform dose $\langle D\rangle$, let us define it as $\overline{\mathrm{NTCP}}$

$$\overline{\mathrm{NTCP}} = \mathrm{NTCP}(\langle D\rangle) = \left\{ \begin{array}{cc} \bar{P}(w_P, b_P) & \mathrm{RS} \\ (1-\bar{P}(w_A, b_A))(1-\bar{P}(w_D, b_D)) & \mathrm{DRSN} \end{array} \right\}. \tag{16}$$

Therefore, it will be convenient to do a reparameterization in term of $\overline{\mathrm{NTCP}}$ and any parameter variation that leaves it unchanged. In the case of the RS model, there is only one remaining degree of freedom if we fix $\overline{\mathrm{NTCP}}$, let us call it θ_{RS}, while for the DRSN model, there are three remaining degrees of freedom $\vec{\theta}_{\mathrm{DRSN}} = (\theta_{\mathrm{DRSN1}}, \theta_{\mathrm{DRSN2}}, \theta_{\mathrm{DRSN3}})$. We will explain the geometric forms of the θ degrees of freedom in a moment. For now, it suffices to point out that with this reparameterization the loss functions take very simple forms

$$L_{\mathrm{RS}}(w_{\mathrm{P}}, b_{\mathrm{P}}) \to L_{\mathrm{RS}}(\overline{\mathrm{NTCP}}, \theta_{\mathrm{RS}}) \ = \ (\overline{\mathrm{NTCP}} - \overline{\mathrm{NTCP}}_0)_2 \tag{17}$$

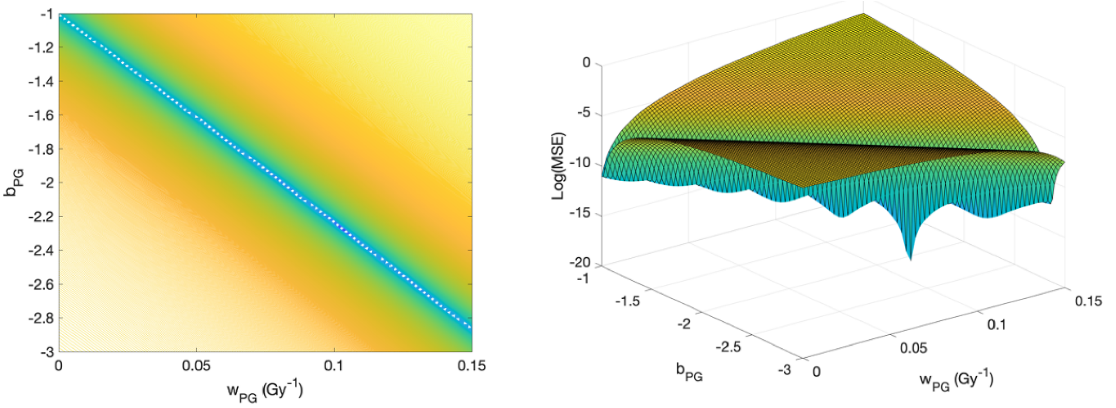

Figure 10. (Left) 2D and (right) 3D contour plots of the log of the MSE for the RS model, with the global minimum at the true parameter value, along with local minima spanning the sloppy direction of parameter variation but localized in the stiff direction.

$$L_{\mathrm{DRSN}}\left(w_{\mathrm{A}},b_{\mathrm{A}},w_{\mathrm{D}},b_{\mathrm{D}}\right)\to L_{\mathrm{DRSN}}\left(\overline{\mathrm{NTCP}},\vec{\theta}_{\mathrm{DRSN}}\right) \;=\; \left(\overline{\mathrm{NTCP}}-\overline{\mathrm{NTCP}}_0\right)_2 \tag{18}$$

where $\overline{\mathrm{NTCP}}_0$ is $\overline{\mathrm{NTCP}}$ evaluated at the true parameter values (w_{Parotid} and b_{Parotid} for the RS model ; $w_{\mathrm{Acinar}}, b_{\mathrm{Acinar}}, w_{\mathrm{Duct}}$ and b_{Duct} for the DRSN model).

With this setup, we can now calculate the Hessian. Consider a model with a vector of J fitting parameters $\vec{X}=(X_1,\ldots,X_J)$, with the fit based on minimizing a loss function $L\left(\vec{X}\right)$. The Hessian, defined at a given point in parameter space $H\left(\vec{X}=\vec{X}_0\right)$, is

$$H_{jk}\left(\vec{X}=\vec{X}_0\right)=\left(\frac{\partial L}{\partial X_j}\cdot\frac{\partial L}{\partial X_k}\right)_{\vec{X}=\vec{X}_0}\quad j,k=1,\ldots,J \tag{19}$$

where the partial derivatives are evaluated at $\vec{X}=\vec{X}_0$. If we define $\vec{X}_{\mathrm{RS}}=\left(\overline{\mathrm{NTCP}},\theta_{\mathrm{RS}}\right)$ and $\vec{X}_{\mathrm{DRSN}}=\left(\overline{\mathrm{NTCP}},\vec{\theta}_{\mathrm{DRSN}}\right)$, then it is straightforward to see from equations (17)–(19) that H_{jk} is 2 if j and k are both 1, and is 0 otherwise. Thus, by construction, H is already diagonalized, and has a single stiff eigenvector, corresponding to $\overline{\mathrm{NTCP}}$, with eigenvalue 2. The remaining eigenvector directions $\vec{\theta}$ are sloppy, with eigenvalues 0. Summarizing, in the mean-field approximation only the aggregate $\overline{\mathrm{NTCP}}$ is meaningful for predictions.

4.2.2. The geometry of sloppy parameter variations corresponds to the geometry of local minima in the neural network loss function

We now turn to the geometric interpretations of the sloppy directions $\vec{\theta}$. The case of the RS model is easier since it is clear from equations (1), (16) and (17) that $(w_{\mathrm{PG}},b_{\mathrm{PG}})$ only enter $\overline{\mathrm{NTCP}}$ via the linear combination $w_{\mathrm{PG}}\langle D\rangle+b_{\mathrm{PG}}$. Thus, sloppy parameter variation corresponds to any variation that leaves this linear combination invariant, or the line

$$w_{\mathrm{PG}}\langle D\rangle+b_{\mathrm{PG}}=w_{\mathrm{Parotid}}\langle D\rangle+b_{\mathrm{Parotid}}. \tag{20}$$

If we inspect the form of the MSE loss function for the RS model, as shown in figure 10, and compare it with the results in figure 8, we observe that the local minima in parameter space all lie along this sloppy manifold. In other words, sloppiness explains why the ability of the neural network to generalize without overfitting is robust to variability in the learned parameters across replicates. Namely, the different local minima that each of the replicates converge to during training are not randomly distributed but restricted to vary only in the directions that do not result in overfitting.

We now turn to the interpretation of $\vec{\theta}$ for the DRSN case, where there is even more freedom. To start, note that just like for the parotid gland parameters in the RS model, the individual acinar and duct parameters are free to vary without changing the aggregate probabilities $\bar{P}(w_{\mathrm{A}},b_{\mathrm{A}})$ and $\bar{P}(w_{\mathrm{D}},b_{\mathrm{D}})$

$$w_{\mathrm{A}}\langle D\rangle+b_{\mathrm{A}}=\ln\left(\frac{1-\bar{P}(w_{\mathrm{A}},b_{\mathrm{A}})}{\bar{P}(w_{\mathrm{A}},b_{\mathrm{A}})}\right) \tag{21}$$

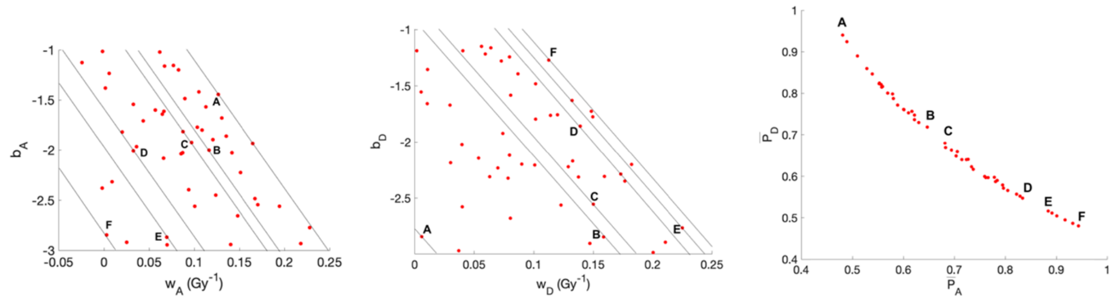

Figure 11. (Left and middle) Replots of the distributions of learned parameters, for both the acinar (left) and duct (middle), but with select annotated lines of constant failure probability that intersect with six selected points. (Right) Plot showing the anticorrelation of the acinar and duct probabilities of failure that encapsulates the hidden geometric structure of the sloppiness in the DRSN model.

$$w_{\mathrm{D}}\langle D\rangle + b_{\mathrm{D}} = \ln\left(\frac{1-\bar{P}(w_{\mathrm{D}},b_{\mathrm{D}})}{\bar{P}(w_{\mathrm{D}},b_{\mathrm{D}})}\right). \tag{22}$$

Note that, in contrast to equation (2), $\bar{P}(w_{\mathrm{A}},b_{\mathrm{A}})$and $\bar{P}(w_{\mathrm{D}},b_{\mathrm{D}})$are not fixed, but still have some freedom to vary while keeping $\overline{\mathrm{NTCP}} = (1-\bar{P}(w_{\mathrm{A}},b_{\mathrm{A}}))(1-\bar{P}(w_{\mathrm{D}},b_{\mathrm{D}}))$fixed. A geometric interpretation of this is that the sloppy variation of the acinar and duct parameters still amounts to straight lines in individual w-b space, with the same slope as in the RS model, but with variable intercepts constrained to be anti-correlated with one another. A pictorial demonstration of this geometry is shown in figure 11, where just as for the RS model, we see that the distribution of`local minima’ among different training replicates lies along the expected parameter subspace.

5. Outlook and conclusions

The architectures and simulations described here are just the tip of the iceberg, and there is much more to explore relating to the connection between mechanistic radiobiology and deep learning. Here, we will briefly comment on one particularly relevant and promising direction.

5.1. Robustness of sloppy and stiff parameter spaces as a guide to regularization

In our simulations, we considered somewhat idealized scenarios, with clean data and more training data than network parameters. Surprisingly, we found that, even in this regime, the trained neural networks can robustly generalize without having to converge to the global minimum, so long as the train and test cases were generated with the same distribution.

Very often, however, the number of network parameters is greater than the number of data points, outcome measurements are noisy and probabilistic, and there is some amount of ‘mismatch’ between the training and test set distributions. It is well established that if such effects become too great, at some point the ability of neural networks to generalize without overfitting breaks down, despite the robustness to parameter variation that we have discussed. As a result, regularization of model complexity, through techniques such as adding explicit penalties to the loss function or including dropout layers, is usually still necessary to prevent overfitting.

Although we have not directly addressed regularization in this work, the lessons we have learned from analyzing sloppy and stiff directions of parameter space are directly relevant to it. A subtle point that we glossed over is that the structure of the stiff and sloppy directions in parameter space is implicitly dependent on the distribution of the training data. This concept provides intuition as to when and why regularization may be needed.

For instance, if the training data is too noisy, or if the model is sufficiently complex that the data is sparse, the sample distribution is biased away from the true distribution of training data. As a result of this sampling bias, the corresponding sloppy regions in parameter space are also biased, increasing the susceptibility of the training to ‘spuriously sloppy’ local minima that will overfit.

As another example, if the test distribution differs from the training distribution, then the sloppy region of parameter space for the training data will be different from that of the test data. Although the global minimum, associated with the true parameter values, will always lie in the sloppy space regardless of data distribution, the local minima in general will be deformed away from their locations in the training distribution. Thus they will not typically generalize robustly to the mismatched test data set.

With this insight in mind, our work here suggests a new design principle for regularization, which in practice is often treated as an art form. Specifically, when increasing model complexity, the distribution of local minima in the loss landscape of the neural network might provide clues into when the model is starting to overfit, based on deviations from expected stiff and sloppy directions. Furthermore, when modeling a test set that is mismatched from the training set, analysis of how the stiff and sloppy directions are deformed could be a guide as to how close to the global minimum the training needs to converge to in order to stay in the test set's sloppy region.

5.2. Conclusion

In conclusion, we have shown that the RS model is equivalent to a simple ANN, and that generalizing this to allow for both hierarchical tissue organization and off-target effects leads to an interpretable class of ANN architectures for dose-toxicity mapping, which we name DSRNs. Using simulations on a test case of xerostomia in αRPT, we have highlighted how thinking about the information geometric concept of sloppiness, as it relates to curve-fitting a mechanistic multiparameter model, provides intuition into the ability of the ANN to generalize without overfitting. We anticipate that our work will more generally open up many new avenues for interdisciplinary collaboration between practitioners in the fields of radiobiology and machine learning.

Data availability statement

No new data were created or analyzed in this study.

Acknowledgments

I thank George Sgouros, Colin Paulbeck, Remco Bastiaanet, Rob Hobbs and Mark Transtrum for helpful comments and suggestions. I also thank my two anonymous reviewers, whose insightful comments and questions have inspired and greatly improved the manuscript.

ORCID iD

Tahir I Yusufaly https://orcid.org/0000-0002-8755-3614

References

[1] Ebert M A, Gulliford S, Acosta O, de Crevoisier R, McNutt T, Heemsbergen W D, Witte M, Palma G, Rancati T and Fiorino C 2021 Spatial descriptions of radiotherapy dose: normal tissue complication models and statistical associations *Phys. Med. Biol.* **66** 12TR01

[2] Lyman J T 1985 Complication probability as assessed from dose-volume histograms *Radiat. Res.* **104** S13

[3] Kutcher G J and Burman C 1989 Calculation of complication probability factors for non-uniform normal tissue irradiation: the effective volume method gerald *Int. J. Radiat. Oncol. Biol. Phys.* **16** 1623–30

[4] Källman P, Ågren A and Brahme A 1992 Tumour and normal tissue responses to fractionated non-uniform dose delivery *Int. J. Radiat. Biol.* **62** 249–62

[5] Tseng H-H, Wei L, Cui S, Luo Y, Haken R K T and Naqa I E 2020 Machine learning and imaging informatics in oncology *Oncology* **98** 344–62

[6] Zhen X, Chen J, Zhong Z, Hrycushko B, Zhou L, Jiang S, Albuquerque K and Gu X 2017 Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study *Phys. Med. Biol.* **62** 8246–63

[7] Men K, Geng H, Zhong H, Fan Y, Lin A and Xiao Y 2019 A deep learning model for predicting xerostomia due to radiation therapy for head and neck squamous cell carcinoma in the RTOG 0522 clinical trial *Int. J. Radiat. Oncol. Biol. Phys.* **105** 440–7

[8] Liang B, Tian Y, Chen X, Yan H, Yan L, Zhang T, Zhou Z, Wang L and Dai J 2020 Prediction of radiation pneumonitis with dose distribution: a convolutional neural network (CNN) based model *Front. Oncol.* **9** 1500

[9] Valdes G and Interian Y 2018 Comment on 'Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study' *Phys. Med. Biol.* **63** 068001

[10] El Naqa I and Das S 2020 The role of machine and deep learning in modern medical physics *Med. Phys.* **47** e125–6

[11] Mehta P, Wang C-H, Day A G R, Richardson C, Bukov M, Fisher C K and Schwab D J 2019 A high-bias, low-variance introduction to machine learning for physicists *Phys. Rep.* **810** 1–124

[12] Prise K M and O'Sullivan J M 2009 Radiation-induced bystander signalling in cancer therapy *Nat. Rev. Cancer* **9** 351–60

[13] D'Andrea M, Benassi M and Strigari L 2016 Modeling radiotherapy induced normal tissue complications: an overview beyond phenomenological models *Comput. Math. Methods Med.* **2016** 2796186

[14] Belli M L *et al* 2020 Targeted alpha therapy in mCRPC (metastatic castration-resistant prostate cancer) patients: predictive dosimetry and toxicity modeling of 225Ac-PSMA (prostate-specific membrane antigen) *Front. Oncol.* **10** 531660

[15] Wahl R L, Sgouros G, Iravani A, Jacene H, Pryma D, Saboury B, Capala J and Graves S A 2021 Normal-tissue tolerance to radiopharmaceutical therapies, the knowns and the unknowns *J. Nucl. Med.* **62** 23S–35S

[16] Dijkema T, Raaijmakers C P J, Ten Haken R K, Roesink J M, Braam P M, Houweling A C, Moerland M A, Eisbruch A and Terhaard C H J 2010 Parotid gland function after radiotherapy: the combined michigan and utrecht experience *Int. J. Radiat. Oncol. Biol. Phys.* **78** 449–53

[17] Jentzen W, Hobbs R F, Stahl A, Knust J, Sgouros G and Bockisch A 2010 Pre-therapeutic 124I PET(/CT) dosimetry confirms low average absorbed doses per administered 131I activity to the salivary glands in radioiodine therapy of differentiated thyroid cancer *Eur. J. Nucl. Med. Mol. Imaging* **37** 884–95

[18] Hobbs R, Plyku D, Jentzen W, Bockisch A and Sgouros G 2018 Small scale dosimetry and modeling for salivary gland toxicity in thyroid cancer patients treated with 131I *J. Nucl. Med.* **59** 470

[19] Rzeszotarski M S 1999 The AAPM/RSNA physics tutorial for residents: counting statistics *RadioGraphics* **19** 765–82

[20] Sgouros G and Hobbs R F 2014 Dosimetry for radiopharmaceutical therapy *Semin. Nucl. Med.* **44** 172–8

[21] Moiseenko V, Wu J, Hovan A, Saleh Z, Apte A, Deasy J O, Harrow S, Rabuka C, Muggli A and Thompson A 2012 Treatment planning constraints to avoid xerostomia in head-and-neck radiotherapy: an independent test of QUANTEC criteria using a prospectively collected dataset *Int. J. Radiat. Oncol. Biol. Phys.* **82** 1108–14

[22] Neyshabur B, Bhojanapalli S, McAllester D and Srebro N 2017 Exploring generalization in deep learning (arXiv:1706.08947 [cs.LG])

[23] Quinn K N, Abbott M C, Transtrum M K, Machta B B and Sethna J P 2021 Information geometry for multiparameter models: new perspectives on the origin of simplicity (arXiv:2111.07176 [cond-mat.stat-mech])

[24] Mannakee B K, Ragsdale A P, Transtrum M K and Gutenkunst R N 2016 Sloppiness and the geometry of parameter space *Uncertainty in Biology* vol 17, ed L Geris and D Gomez-Cabrero (Cham: Springer International Publishing) pp 271–99