Uncertainty-Aware Photoacoustic Oximetry with Conditional Invertible Neural Networks

MRes Mini Project Dissertation

Author:

Michael Doherty

Supervisors:

Janek Gröhl
Sarah Bohndiek

26 August 2022

University of Cambridge

I hereby declare that, except where specifically indicated, the work submitted herein is my own original work.

Michael Doherty
August 26, 2022

Abstract

A novel deep learning model architecture is proposed to enhance the current state-of-the-art image quantification method in photoacoustic oximetry (learned spectral decoloring). A conditional invertible neural network, employing a long short-term memory network as conditioning input, allows greater flexibility in the wavelengths used for data acquisition, while providing a mechanism to construct the posterior distribution of possible blood oxygen saturation (sO_2) values.

The model is trained on and evaluated against three *in silico* datasets, which simulate a blood flow phantom, complex tissue geometries, and the same geometries with melanated skin. The posterior distribution for each test case is reconstructed using 1000 samples and uncertainty statistics calculated to find the median prediction, interquartile range, relative and absolute prediction errors, and calibration error.

The accuracy of predictions is found to increase significantly with the number of spectral components of the initial pressure distribution and is comparable to results from learned spectral decoloring for the flow phantom dataset, with median relative prediction error and IQR of 1.3% (0.4% 5.2%) for measurements at 40 wavelengths. Accuracy for the complex tissue geometries is relatively poor but causes related to the learning rate during training are suggested.

The trained models produce qualitatively reasonable uncertainty estimates, however the maximum calibration errors for models trained on differing numbers of wavelengths are found to be far from optimal (up to 35% deviation). This suggests the models require further training in order to approximate the true posterior distribution of sO_2.

Improvements in the training method are suggested along with further steps to improve model performance, with a view toward re-evaluation to test the model's ability to provide state-of-the-art, uncertainty-aware sO_2 predictions.

List of acronyms

cINN	Conditional Invertible Neural Network
FCN	Fully Connected Network
FOV	Field of View
FrEI	Framework for Easily Invertible Architectures
GAN	Generative Adversarial Network
INN	Invertible Neural Network
IQR	Interquartile Range
LSD	Learned Spectral Decoloring
LSTM	Long Short-Term Memory
LU	Linear Unmixing
ML	Machine Learning
NLL	Negative Log Likelihood
PAI	Photoacoustic Imaging
qPAI	Quantitative Photoacoustic Imaging
ROI	Region of Interest
SASD	Sparsity-Accustomed Spectral Decoloring
sO_2	Oxygen Saturation
VAE	Variational Autoencoder

Contents

Chapter 1

Introduction

Current medical imaging modalities are limited in their ability to provide functional and molecular information from tissue. Photoacoustic imaging (PAI) offers a path to obtain this information in a non-invasive way, without ionising radiation or exogenous imaging agents [1]. The measurement of functional tissue properties has numerous applications in biomedical research and clinical medicine. In particular, *in vivo* measurement of spatially-resolved blood oxygen saturation (sO_2) can be used to monitor therapies, post-operative recovery, gene expression and cancer progression [2].

However, the unique imaging opportunities offered by photoacoustics are currently limited by the challenging problem of inferring the optical properties of tissue from the measured ultrasonic signal. Overcoming this obstacle would allow clinical application of PAI to advance from qualitative assessments of the interrogated tissue to quantitative measurements that inform diagnosis and treatment.

To further the utility of any measurement of tissue properties and aid clinical decision-making, the derived value should have an associated uncertainty [3]. State-of-the-art techniques for inference of optical tissue properties from PA signals rely on machine learning (ML) models, which complicates the process of uncertainty estimation when compared to the standard error propagation formulas available for direct measurements.

This report details the challenges in enabling quantitative photoacoustic imaging (qPAI), from physical principles to the difficulties in generating adequate ML training data, and recent advances in the field. It builds upon recent work that shows promise in predicting sO_2 *in silico*, *in vitro* and *in vivo*, while allowing flexibility in the wavelengths used for illumination [4]. This approach is enhanced by adapting the ML model architecture to be a conditional invertible neural network (cINN), which allows the posterior distribution of possible sO_2 values to be estimated, from which uncertainty statistics

can be calculated.

The goal of this report is to demonstrate a learned technique that is able to produce accurate sO_2 estimates with reliable uncertainty estimates, while retaining flexibility in the wavelengths used for training and evaluation. This is in order to improve usability of the model by other researchers in the field and ultimately translation to clinical practice.

1.1. Photoacoustic Imaging

PAI combines the high spatial resolution of ultrasound in tissue with the superior contrast and specificity of optical imaging through exploitation of the photoacoustic effect. In a typical PAI setup, the tissue region of interest (ROI) is illuminated with a pulsed laser beam. The incident photons propagate through the tissue and undergo optical scattering and absorption. The pulse power is low in order to prevent tissue damage and near-infrared wavelengths are used, resulting in most of the absorbed energy transferred to heat by vibrational relaxation [5]. In the regions where the light is absorbed, the resultant temperature increase causes localised thermoelastic expansion and the generation of an ultrasonic pressure wave.

The conditions for the generation of photoacoustic signal are dictated by the mechanics of the tissue and thermodynamics of the absorbed optical energy. The laser pulse duration must be sufficiently short to ensure the absorbing region is in a state of thermal and stress confinement [6]. The more stringent constraint is the stress confinement time, defined as the ratio of the target spatial resolution to the speed of sound in the tissue. For a target resolution of $15\mu m$ with a typical speed of sound of 1500m/s, the pulse duration must be under 10ns.

Higher ultrasound frequencies enable higher spatial resolution but at the expense of penetration depth, due to increased attenuation. Resolution can be subcellular (200nm) for ultrasonic frequencies $>$50MHz to a depth of 1mm for microscopy applications [7] or 0.5-5mm at depths of 3-7cm using 4-8MHz [8]. The range of depths and resolutions achievable with a single imaging modality, using the same optical contrast, is another unique benefit of PAI [1].

Penetration depth of the optical signal is dictated in large part by the choice of illumination wavelength. Chromophores are the molecular structures that absorb the incident photons. Figure 1.1 shows the absorption spectra of the main tissue chromophores; melanin, oxyhaemoglobin (HbO_2), deoxyhaemoglobin (Hb), water, and fat. The total absorption of the tissue components is lowest in the near-infrared range of 600-1000nm. Illumination within this wavelength range is therefore used as it provides

an 'optical window' for greater penetration, which has been demonstrated up to 6cm *in vivo* [9].

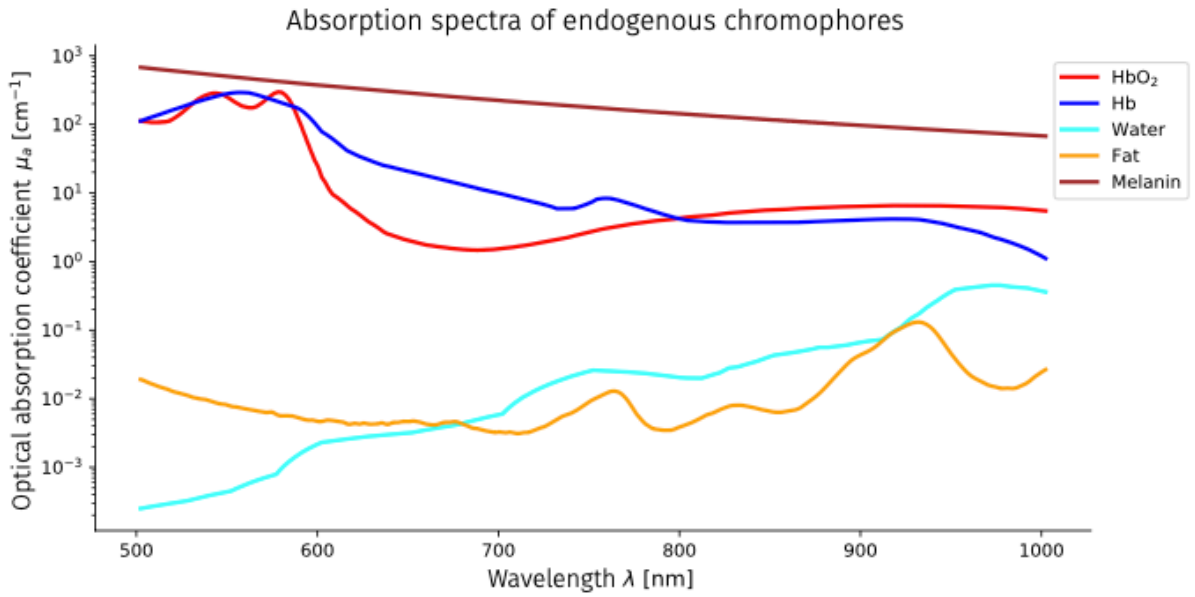


Figure 1.1: The optical absorption spectra of major tissue components across the mid-visible to near-infrared wavelengths. Reproduced from the CC-BY licensed publication [10].

Many different geometries exist for the illumination and detection elements of a PAI system. The detector geometry used to measure the ultrasonic signal may be a linear or curved array of detector elements with a limited field of view (FOV) or a tomographic system capable of measuring ultrasound emitted from the ROI at multiple angles. A tomographic approach is preferable to eliminate FOV artefacts from the image but a handheld PAI system benefits from ease of clinical application. Detection of the ultrasonic signal is often performed with piezoelectric transducers but is challenging due to the broadband nature of signals generated with ultra-short laser pulses, which can span from sub-MHz to hundreds of MHz. All-optical sensing with Fabry-Pérot cavities offer a promising path to broadband ultrasound sensing [11].

The overarching challenge of converting the measured time-series pressure data into spatially-resolved, accurate images of chromophore concentrations can be divided into two inverse problems: (1) the acoustic inverse problem and (2) the optical inverse problem [5].

1.1.1. Acoustic inverse problem

The acoustic inverse problem describes the transformation from the detected pressure data, $p(t)$, back to an image of the initial pressure distribution, p_0. Depending

on the geometry of the imaging setup, this problem is well-posed and has a unique solution, though image resolution is impaired by signal propagation in inhomogeneous and lossy media, finite detector bandwidth and aperture, and reconstruction artefacts, amongst other mechanisms [12].

The chosen reconstruction algorithm for a certain PAI setup depends on the PA imaging modality, and can be categorised into back-projection, series expansions, time reversal, iterative reconstruction and deep learning approaches [12]. A back projection algorithm [13] is often used for its simplicity and computational efficiency.

The reconstructed p_0 provides a spatial map of the tissue and is useful for qualitative analysis [14]. The initial pressure distribution also forms a basis for many techniques to extract functional tissue information, with which the optical inverse problem is concerned.

1.1.2. Optical inverse problem

Knowledge of tissue optical properties is the prize of solving the optical inverse problem. The most useful of these properties is the absorption coefficient, μ_a, as it allows estimates of the underlying chromophore concentrations to be obtained. The statement of the problem can be understood by considering equation 1.1, which shows the p_0 to be a product of μ_a, the light fluence distribution ϕ and the dimensionless Grüneisen parameter [6].

$$p_0 = \mu_a \cdot \phi \cdot \Gamma \tag{1.1}$$

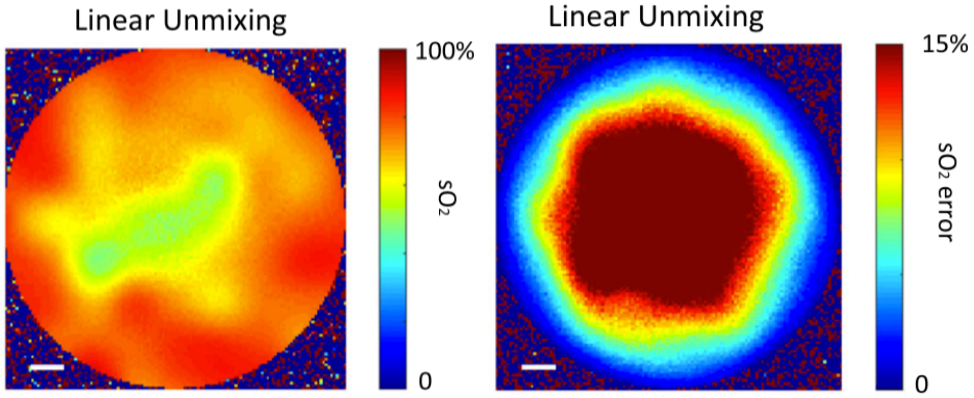
The light fluence distribution is determined by the spatially varying μ_a and μ_s, which also have a chromatic dependency. The resulting chromatic variation in ϕ and hence p_0 is known as spectral coloring. Furthermore, the absorption and scattering increases with depth, thereby amplifying the effect of spectral coloring and causing ϕ to have a non-linear dependency on μ_a, μ_s. To further complicate matters, it is assumed there are multiple possible combinations (μ_a, μ_s) for a given p_0. The optical inverse problem is therefore both non-linear and ill-posed.

The Grüneisen parameter combines several mechanical and thermodynamic properties of the tissue, which vary spatially and with temperature, which can in turn vary across the tissue, leading to yet further non-linearity. Considering these complications, p_0 can be more accurately described by equation 1.2, where x is a spatial variable, λ is optical wavelength, T is temperature and ϵ is a noise term.

$$p_0 = \mu_a(x,\lambda) \cdot \phi(x,\lambda,\mu_a(x,\lambda),\mu_s(x,\lambda)) \cdot \Gamma(x,T(x)) + \epsilon(x,\lambda) \tag{1.2}$$

The most common technique to attempt a solution to the optical inverse problem is linear unmixing (LU), which assumes a linear combination of chromophore concentrations give rise to the observed p_0 [15]. Using *a priori* knowledge of the chromophore absorption spectra (figure 1.1) and multispectral images of the initial pressure distribution p_0, the concentration of chromophores at a spatial location can be disentangled by solving a series of linear equations. From these concentrations, blood oxygenation saturation (sO_2) can be found as the relative ratio of oxyhaemoglobin to total haemoglobin, $sO_2 = [HbO_2]/([HbO_2] + [Hb])$.

This technique can provide reasonable estimates through careful additional modelling to constrain likely sO_2 values but ultimately fails due to the neglected depth-dependent spectral coloring and non-linearities. The effect on the efficacy of LU is best illustrated by the imaging of arterial blood. An even distribution of highly oxygenated blood throughout the artery is expected, but applying LU to the image results in differing rim-core sO_2 due to the depth-dependent signal fluence, as illustrated in figure 1.2.



(a) sO_2 estimates by LU (b) Corresponding absolute error

Figure 1.2: Images of arterial cross-section from [16]. Spectral coloring is not considered by LU, resulting in increasingly erroneous sO_2 estimates at depth.

1.2. Data-driven methods for qPAI

Due to its ill-posed and non-linear nature, an analytical solution to the optical inverse problem is not possible without simplifying assumptions that severely affect the calculated chromophore concentrations. Consequently, current research into PAI signal quantification is trending towards data-driven learned methods that can accurately approximate the non-linear mapping from pressure spectra to tissue properties such as sO_2. While still in their infancy, such methods have the potential to facilitate the clinical translation of qPAI [17].

Data-driven methods, using both classical machine learning and deep learning, have been applied to a range of problems in PAI from the acoustic inverse problem, to image post-processing and semantic image notation [17]. This report concerns methods used to estimate functional tissue properties, which relates to the optical inverse problem.[1]

Difficulty in obtaining accurately labelled ground truths for *in vitro* and *in vivo* data has been a limiting factor in the successful translation of learned methods into clinical application, however. This is due to the domain gap between simulated data and data acquired from real tissue with PAI devices. Ambitious and comprehensive frameworks have been developed to facilitate the task of simulating data [20] and deep learning models such as generative adversarial networks have been applied to the data generation process in attempts to close the domain gap [21].

Nonetheless, *in silico* training and validation of learned models continues and has shown great promise in anticipation of future improvements to training data. The models used for sO_2 estimation so far have included feed-forward neural networks [15] and residual U-nets [16]. Just as important as the model architectures has been the choice of training data, which has invariably been simulated, and has included hand-crafted feature vectors [22] and whole 2D images of initial pressure spectra [23], even 3D volumes [24].

Most notable of the data-driven methods applied thus far to the task of sO_2 estimation has been learned spectral decoloring (LSD) [15], which has provided high accuracy *in silico* and plausible estimates *in vitro* and *in vivo* despite training purely on simulated data. The technique adopts a pixel-wise approach to sO_2 estimation, providing an estimate based purely on the multispectral pressure values for the pixel, thereby rejecting any spatial information which may increase the domain gap. It also eschews any feature engineering, as it is received wisdom in theoretical machine learning that hand-crafted feature vectors are inferior to those selected by the machine learning algorithm as relevant [25]. The choice of training data for LSD is therefore general to PA images and aims to reduce the domain gap.

LSD has been developed further to improve its accuracy, such as by illumination from multiple angles to obtain more information from the imaged region [26], and also to improve its generalisability to training wavelength selection through the integration of a long short-term memory (LSTM) network to its architecture. The enhanced

[1] A distinction is made between solving the optical inverse problem, which yields spatially resolved chromophore concentrations, and estimating functional tissue parameters. Though closely related, the direct estimation of functional tissue parameters such as sO_2, or even blood glucose [18], without the intermediate step of finding chromophore concentrations has the advantage of readily available methods to measure reference values for such parameters. This provides an avenue to experimental validation and even non-simulated training datasets [19].

LSTM-LSD approach is known as sparsity-accustomed spectral decoloring (SASD) and is discussed further in section 2.2. The SASD architecture is adapted in the work presented in this report to conform to an invertible network architecture, thus allowing a posterior distribution of sO_2 values to be estimated (section 1.3).

1.3. Uncertainty estimation for data-driven qPAI

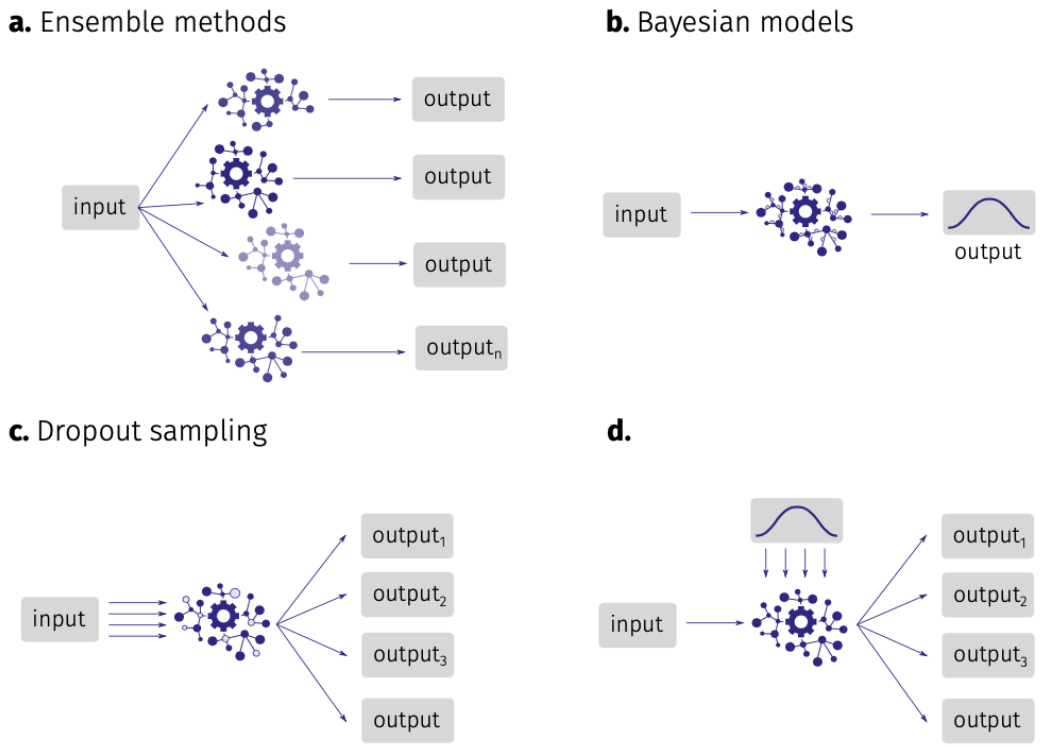


Figure 1.3: Illustration of four possible methods of obtaining uncertainty estimates from a machine learning model. Reproduced from the CC-BY licensed publication [10].

Uncertainty-aware machine learning methods are crucial in medical applications of ML, in order to allow safer deployment and engender trust in the technology with clinicians, who can then factor uncertainty estimates into their decisions [3]. The ability to resolve possible multimodal distributions in the estimated values is another benefit of uncertainty-awareness, especially for ill-posed problems.

Methods to estimate the uncertainty of ML model predictions quantify the combined aleatoric uncertainty from noise in the acquired data and epistemic uncertainty from the imperfect model. Four such methods are illustrated in figure 1.3.

a. **Ensemble methods** use many ML models, trained differently and often on subsets of the total training dataset, to generate multiple predictions. Statistics on the

range of predictions can then be calculated to quantify the uncertainty. Random forest models are an example [27].

b. Bayesian models treat the weights in a deep neural network as probability distributions rather than deterministic variables. The predictions of such models are therefore also probability distributions but this method adds considerable computational complexity, which limits practicality [28].

c. Dropout sampling describes the stochastic deactivation of neurons within the model, which is then sampled to build a distribution of the altered predictions. This technique bas been proposed as an approximation of a probabilistic Bayesian network [29] but the correctness of this view is disputed. Dropout sampling is primarily used as a technique to prevent overfitting but has been used to estimate uncertainty in deep learning approaches to PAI [30].

d. Latent space sampling refers to introducing variation to the feature representation in the network (the so-called latent space), which is then reconstructed to provide a posterior distribution over which statistics can be calculated. The technique was pioneered in the variational autoencoder (VAE) architecture [31], which employs one network (the 'encoder') to learn the mapping from observation space to a latent space, and another (the 'decoder') to map the inverse. The lossy encoding/decoding of VAEs is improved upon by invertible neural networks (INNs), in which a single bijective network performs both mappings [32]. Latent space sampling can only be applied to generative models, such as VAEs and INNs, that learn the forward data generation process, as opposed to solely a discriminator [33]. For PAI, latent space sampling has been performed on a cINN architecture to evaluate uncertainty in the LSD method [34].

The method of uncertainty estimation investigated in this report is informed by prior work [10] that shows latent space sampling from a cINN to provide qualitatively and quantitatively better uncertainty estimates than the other methods presented in figure 1.3, without the computational expense of Bayesian networks. Latent space sampling also outperforms external observing networks trained to estimate aleatoric and epistemic uncertainties separately [35]. The theoretical motivation for selecting an invertible architecture is the ability to build a complete posterior distribution that can identify multimodal probabilities and is proven to converge to the true posterior distribution in the minimum asymptotic limit of training loss, if trained appropriately [32].

Chapter 2

Methods

This chapter provides detail on the investigative methodology used to assess the suitability of the proposed technique in estimating sO_2, based on the accuracy of the median predictions and the posterior distribution of possible sO_2 values. The architecture of the neural networks is explained theoretically and with practical detail, along with the training routine.

The *in silico* datasets used for model training and evaluation are described, as are the processes of spectral partitioning and re-balancing that were applied to the data to explore the effect on model performance. The method to train the model and evaluate it on a flexible range of sparsities is also described. Finally, the relevant metrics used to assess model accuracy and uncertainty estimates are explained and justified.

2.1. Overview

The idea of using pressure spectra from individual pixels to train a deep neural network is the basis of the LSD technique and more recently SASD [4]. Figure 2.1 summarises the processes of data pre-processing, model training, and inference for the investigation presented in this report.

2.1.1. Data pre-processing

Part 1. of figure 2.1 illustrates pixel-specific pressure spectra on which the network is trained and evaluated. For each simulated tissue volume, the optical forward process of photon transport and absorption is simulated. This allows the initial pressure distribution p_0 to be found. This process is repeated at 41 optical wavelengths, from

700nm to 900nm in 5nm steps, to produce 41 initial pressure values at every pixel of the simulated image. These pressure values, in the form of a 41-length vector, form the basis of the training data, along with the corresponding sO_2 label for the pixel.

Part 2. refers to one of the key innovations of the SASD method, which is the random introduction of sparsity into the pressure spectra. For a chosen sparsity level between 2 and 41 wavelengths, values from the pressure spectra are zeroed out until only the chosen number of values remains. This ensures the number of values remains constant but at different wavelengths. This improves the flexibility of the network to the input wavelengths and aids in the method being applied to new datasets that use the same number of spectral images but at different wavelengths (within the 700-900nm range).

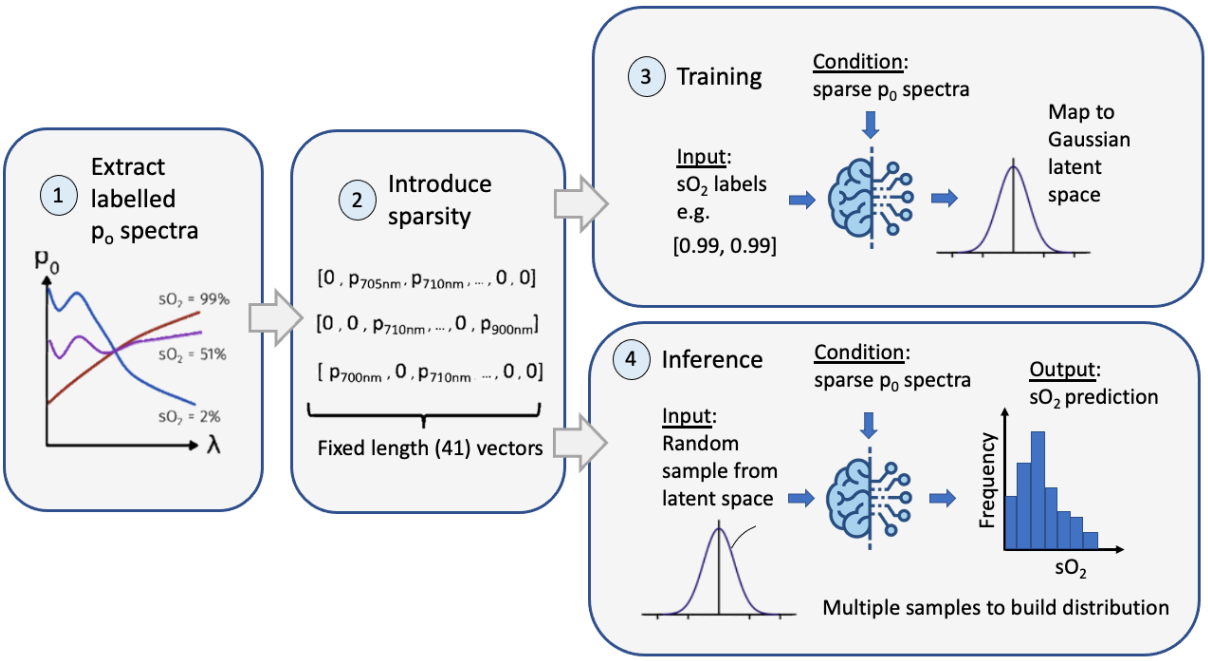


Figure 2.1: Overview of process for data processing, training and evaluation of models. The model is evaluated in opposite directions for training and inference.

Further data preparation steps that must be applied prior to evaluation by the network are as follows:

- **Masking** - An additional binary dimension is added to the condition to indicate if a zero is to be treated as a pressure value or a sparse entry.
- **Normalisation** - After sparsity introduction and masking, each pressure value is normalised by subtracting the mean of the remaining values and dividing by the standard deviation. Normalisation is crucial as it removes any dependency on the incident optical signal intensity, thus improving the generalisability of the network and translation to unseen datasets.

2.1.2. INN fundamentals

To understand the model architecture and steps 3 and 4 of figure 2.1, an appreciation of the theoretical underpinnings for invertible neural networks is required. Neural networks in general can be understood as universal function approximators, capable of modelling high-dimensional non-linear transformations [36]. In the case of INNs, the function f the network approximates can be considered a transformation between two probability distributions; from Z (the latent space) to X (label space) $f : \mathrm{Z} \rightarrow \mathrm{X}$. The probability density p_θ for variable $z \in Z$ can be considered as a standard Gaussian distribution (equation 2.1).

$$z \sim p_\theta(z) = \mathrm{N}(z; 0, 1) \tag{2.1}$$

The label x corresponding to z is obtained by applying f, which can be decomposed into a series of component transformations, f_i.

$$x = f(z) = f_n \cdot f_{n-1} \cdot ... \cdot f_1(z) \tag{2.2}$$

If each of f_i are invertible (bijective), then f is invertible. This sequential application of bijective transformations is called a normalizing flow and came to prominence in machine learning with the non-linear independent components estimation (NICE) architecture [37]. The invertibility of f means the change of variables formula (equation 2.3) can be applied to map between the probability densities of z and x. ∇f indicates the Jacobian of f.

$$p_\theta(x) = p_\theta(z) \cdot |det(\nabla f)| \tag{2.3}$$

For conditional invertible neural networks, the probability density functions can simply be substituted with the conditional probability densities, $p_\theta(x|c)$ and $p_\theta(z|c)$. Appropriate selection of the invertible transformations that comprise f can ensure that the determinant of the Jacobian is easily computable (most easily achieved with a triangular Jacobian for which the determinant is the sum of the main diagonal). Through this mathematical sleight of hand, training can be performed on the forward process mapping observables (x, c) (sO$_2$ and p$_0$ spectra) to the known latent distribution $p_\theta(z|c)$ and get the inverse process 'for free' [32].

Maximum likelihood loss training is convenient for INNs and cINNs with tractable Jacobians. Assuming a Gaussian latent distribution and applying Bayes theorem to the conditional form of equation 2.3, the likelihood of the conditional posterior distribution of sO$_2$ ($p_\theta(x|c)$) is maximised by minimising the negative log likelihood (NLL) (equation

2.4) to find the optimum network weights [38].

$$NLL = \frac{|f(x|c)^2|}{2} - log|det(\nabla f)| \tag{2.4}$$

2.2. Model architecture

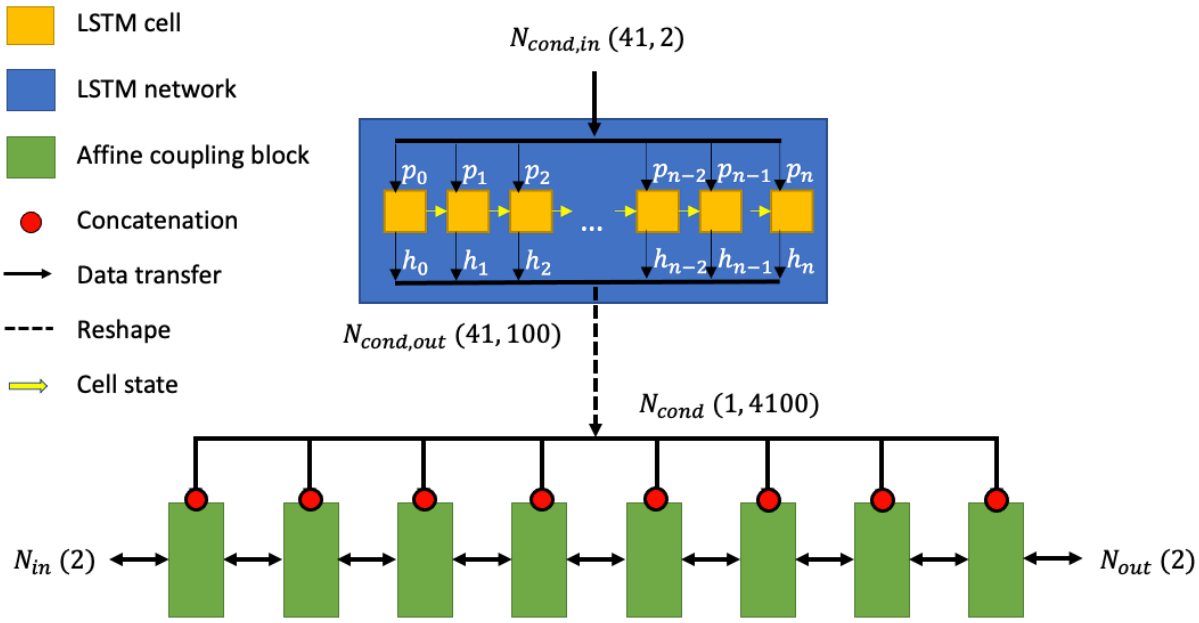


Figure 2.2: Overview of the machine learning model architecture. The LSTM network is represented as a series of LSTM cells, each of which receives a pressure spectra value p_n as input along with the cell state of the previous cell. The output of the LSTM network is the series of LSTM cell hidden states. Dimensions of the relevant tensors at each stage of the network are shown in brackets, without batch dimension.

Figure 2.2 illustrates the ML model architecture, which chiefly comprises a LSTM conditioning network and the INN consisting of series of invertible blocks. Table 2.1 shows the main hyperparameters to dictate the number of parameters in the model. The size of the model was selected as a compromise between ensuring sufficient parameters to capture the complexity of the modelled process and a reasonable training time. cINNs are renowned for their stability during training and are resilient to vanishing gradients and other training instabilities even when the model is larger than strictly required [38].

The model inputs are sO_2 labels during training, random samples from a standard Gaussian distribution during inference, and sparse masked pressure spectra to the LSTM conditioning network for both. The dimensionality of the inputs/outputs are summarised in figure 2.2. The input values to the INN are duplicated to form a two-

element vector. This is to allow the values to be split across the branches of the affine coupling blocks, which require inputs of minimum dimensionality two. During inference, the mean of the two elements in the output vector is taken as the sO_2 value.

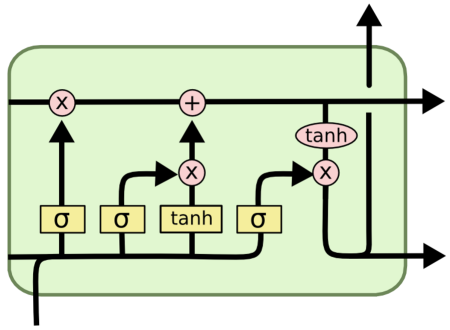


Figure 2.3: Diagram of LSTM cells. The yellow rectangles represent a layer of linear activation units, labelled with their respective activation function.

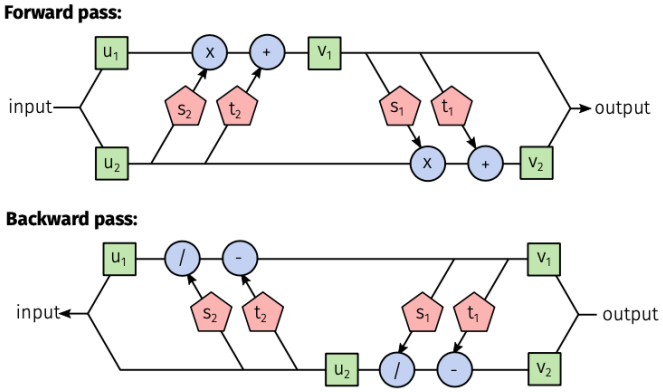


Figure 2.4: Structure of affine coupling block. Reproduced from the CC-BY licensed publication [10]. Green rectangles are data. Blue dots represent element-wise operations. Red pentagons are arbitrary functions (feed-forward neural networks).

An LSTM cell is illustarted in figure 2.3. The subnetworks represented by the yellow rectangles of figure 2.3 comrpise a series of three gates; 'forget', 'input' and 'output'. This gated structure allows the LSTM to have effective internal memory and learn sequential dependencies from data [39]. The recurrent structure structure of the LSTM allows any length sequence to be received as input, thereby allowing a flexible number of optical wavelengths to be used for the multispectral pressure distribution input. The output of the LSTM network is the sequence of 41, 100 dimension hidden states.

Model hyperparameter	**Description**	**Value**
INN no. blocks	Count of affine coupling blocks in INN	8
INN no. layers	Count of linear activation layers in each INN subnet	1
INN hidden size	Count of linear units in each layer of INN subnets	512
LSTM no. layers	Count of LSTM layers	1
LSTM hidden size	Length of hidden state vector returned from each LSTM cell	100

Table 2.1: Summary of model hyperparameters

The INN section of the model was implemented with the open source python project Framework for Easily Invertible Architectures (FrEIA) [40]. FrEIA conveniently

implements an 'AllInOneBlock' class that combines the commonly applied operations of affine coupling, permutation, and global affine transformation (ActNorm) to compose invertible blocks. These blocks were used to construct the INN.

The affine coupling block architecture is illustrated in figure 2.4 and is the same as used in the GLOW architecture [41]. The subnetworks $s_{1,2}$, $t_{1,2}$ are standard feed-forward neural networks. The activation function used throughout the cINN was LeakyReLU with negative slope of 0.01 below 0. The affine transformation includes a soft-clamping mechanism to improve training, first introduced by Real-NVP [42].

Before concatenation to the INN subnetworks, the LSTM output is reshaped into suitable dimensions. A single NLL loss is sufficient to train both the conditioning LSTM network and the INN, as the reshape operation between the two is differentiable.

2.3. Datasets

Three *in silico* datasets were used to test the utility of the method and whether it performs well in different scenarios. Figure 2.5 shows the simulated geometries. The resultant datasets used for training and evaluation only utilised pixels from vessel structures (100% blood volume fraction), from which pressure spectra were then extracted. The pixel-wise approach is advantageous in terms of training data volume, since only a few geometries need be simulated to produce 1000's of training examples. The *in silico* datasets used in this work were generated using the simulation and image processing for photonics and acoustics (SIMPA) Python toolkit [20] as part of the study of SASD [4].

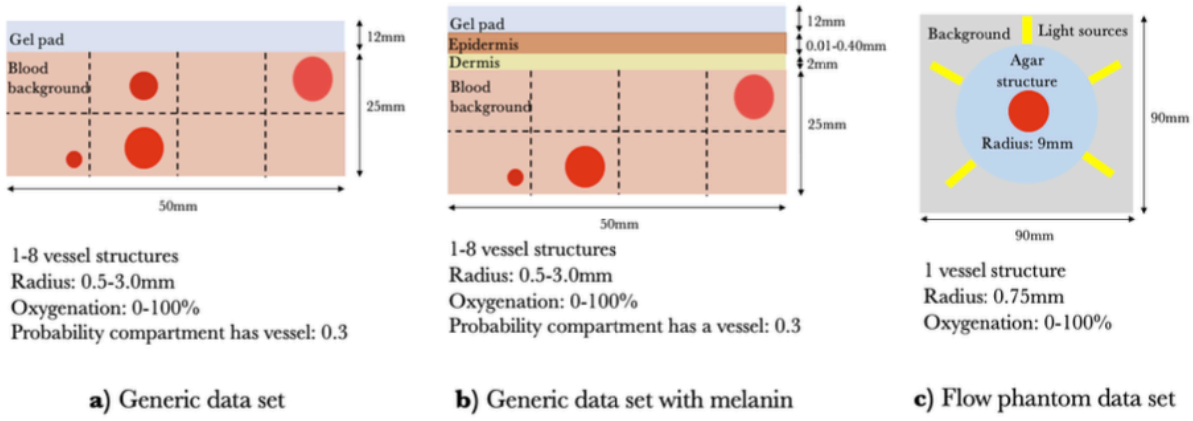


Figure 2.5: The *in silico* datasets investigated in this report. Diagrams from [4].

a) **The generic dataset**: 60,065 samples

A generic tissue model. The volume is split into 8 compartments with max-

imum one vessel structure in each. Vessel radii range from 0.5-3.0 mm with sO_2 from 0%-100%, with all values randomly selected within those ranges. The background medium is homogeneous with a uniform blood volume fraction and oxygenation from 0%-100%.

b) **The melanin dataset**: 60,065 samples
Identical to the generic dataset but with a dermis and epidermis layer, of uniform thickness ranging from 0.01-0.40 mm. The presence of skin leads to more complicated spectral coloring, due to the absorptive properties of melanin. This dataset more closely resembles data acquired from *in vivo* imaging.

c) **The phantom dataset**: 144,245 samples
Comprises an agar structure (μ_s = 5cm$^{-}1$) around a single vessel containing blood oxygenation from 0%-100%. A circular array of five light sources mimics the data acquisition schema for *in vitro* experiments.

2.4. Training and evaluation

The model and training regime were implemented in Python 3.8 using the PyTorch 1.9 framework [43]. A model was trained at sparsities of 3, 5, 10, 25, and 40 wavelengths, with an additional model trained on randomly varying sparsities from 3-41 wavelengths, for each dataset. Models were also trained on either partitioned or unpartitioned data (see section 2.4.2), resulting in a total of 6 models x 3 datasets x 2 partition options = 36 trained models.

Each dataset was split into training, validation and test sets in the ratio 70:10:20. The training regime comprised a minimum of 900 and maximum of 1600 epochs with a batch size of 2046. The training dataset was reinitialised between epochs to ensure new wavelengths were selected and to shuffle the order in which they were received by the model. The mean loss on the validation set was evaluated after every epoch. If the loss was lower than the previous lowest loss, the model was saved. Training was terminated if the maximum epoch count was exceeded or a 50 epoch plateau of no improvement to the maximum training loss was observed. This epoch range was decided to ensure sufficient epochs to saturate the validation loss, based on observations of the validation loss curves of some initially trained models. Training duration per model on a GPU was approximately 8 hours.

The Adam optimiser [44] was used to optimise network weights and therefore the loss function. Adam uses an adaptive learning rate for each weight in the network, which can improve convergence time and is recommended for use in INNs [40]. The

Adam β parameters were (0.9, 0.95), learning rate 10^{-3}, eps 10^{-6}. A global step learning rate scheduler was also used, with the learning rate decaying by a factor of 0.01 every epoch. The Adam optimiser obviates the need for a learning rate scheduler, so the learning rate decay may have negatively impacted results for the models presented in this report, as discussed in chapter 4. Gradient clipping was also applied, as recommended by the FrEIA framework to avoid training instability, with a maximum gradient norm of 10.

2.4.1. Training on flexible sparsities

One model per dataset was trained on a range of sparsities instead of a fixed number of wavelengths, to examine the effect on model performance. The chosen sparsity level was randomised between 3 and 41 wavelengths for every training example, and re-selected for each between epochs.

2.4.2. Spectral partitioning

Previous evaluation of the SASD method [4] indicated that LSD produced more accurate predictions than SASD at very high sparsities ($\sim$ 3 wavelengths). The reason for this is postulated to be the wavelength spacing, which is regular and well-separated for LSD but can be clustered for randomly selected wavelengths. A potential solution to this shortcoming is to partition the pressure spectra into equally-sized sections according to the chosen sparsity level, with one non-zeroed value in each. This has the effect of imposing a minimum spectral range according to sparsity, which may improve the spectral information available and increase accuracy.

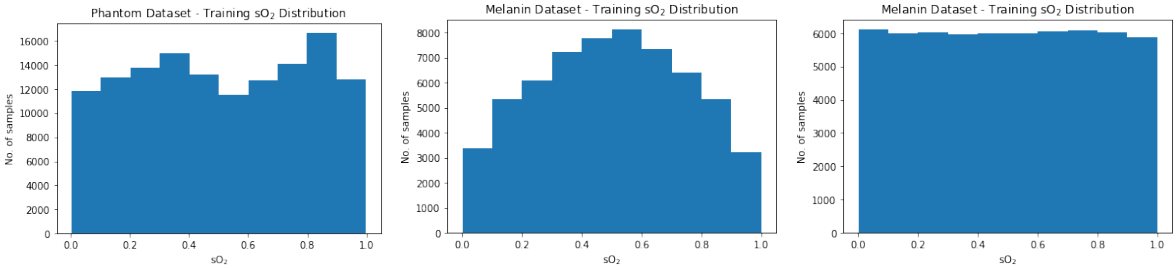
Partitioning occurs before masking and normalisation in the data preparation process, to ensure correct normalisation.

2.4.3. Balanced datasets

The simulated initial pressure distributions are of limited resolution, which affects the sO_2 values of the extracted pixels. This is due to the partial volume effect, which causes a pixel to take the mean sO_2 value of the region that it covers. Therefore, pixels at the edge of vessels, which may partially include background tissue of differing sO_2, are assigned intermediate sO_2 values. Although the tissue geometries and inter-vessel blood oxygen saturations are randomly generated, the partial volume effect causes the sampled pixels to tend towards the median sO_2 value, 50%.

Figure 2.6 (b) shows the resulting sO_2 value frequency distribution for the melanin dataset. Training on an imbalanced dataset is generally accepted to bias ML models to predict values of which it has seen more examples. To correct this, weighted random sampling was applied to oversample under-represented sO_2 values and undersample over-represented values. The result of the re-sampling is shown in figure 2.6 (c), with an approximately equal number of samples in 10% sO_2 bin.

Comparison of the effect of rebalancing on model performance was evaluated for the generic dataset at 10 and 40 wavelengths only, due to time constraints on retraining and evaluating additional models.



(a) (b) (c)

Figure 2.6: Distribution of training sample count with with sO_2 for (a) phantom and (b) melanin datasets. (c) shows the melanin dataset distribution after weighted random sampling. The generic dataset followed the same distribution as the melanin.

2.4.4. Model evaluation

For evaluation of model performance on the test datasets, 1000 random samples were taken from the Gaussian latent space and input to the model with the relevant pressure spectra as a condition, to build the posterior distribution of sO_2 predictions for each test case.

The primary metrics calculated on the model predictions for the test datasets were the median and interquartile range (IQR) of the posterior for each test case. The median and IQR are selected in anticipation of the estimated posteriors being significantly skewed, therefore the median provides a better measure of central tendency than the mean. The median is unable to account for multimodal distributions, which is a shortcoming of the metric.

From those median predictions, the absolute and relative prediction errors, ϵ_{abs} and ϵ_{rel} respectively, are given by equations 2.5-2.6.

$$\epsilon_{abs} = |sO_2^{EST} - sO_2^{GT}| \tag{2.5}$$

$$\epsilon_{rel} = \frac{|sO_2^{EST} - sO_2^{GT}|}{sO_2^{GT}} \tag{2.6}$$

Calibration errors were calculated to determine how well the uncertainty estimates of the model are able to capture the ground truth sO_2 value. The calibration error is determined by calculating confidence intervals of the posterior distribution and the percentage of ground truth inliers that are found in the interval. A well calibrated model should have the same percentage of ground truth inliers as the confidence interval indicates. A positive calibration error therefore suggests the model is underconfident for a given confidence interval, with a negative error indicating overconfidence.

Chapter 3

Results

This chapter presents the data obtained from evaluating the predictions of the models for each test dataset. The experiments and analysis carried out are summarised in the following table:

Summary	Description	Relevant figures
Total median relative prediction error and IQR	The distribution of median predictions for each model is assessed and condensed into a single median relative prediction error and IQR. This is to facilitate comparisons across datasets, sparsities, and partitioning.	Tables 3.1, 3.2, 3.3
Scatter plots of median predictions and IQR	The median and IQR of predictions for each sO_2 value are plotted at 10, 40 wavelengths for each dataset to visualise prediction quality and assess model convergence.	Figure 3.1
Effect of flexible training	The best median relative prediction error and IQR are plotted across sparsities for each dataset for sparsity-specific models and flexibly trained models, to compare performance.	Figure 3.2 (a), (b), (c)
Effect of spectral partitioning	Effect of spectral partitioning is assessed through two examples: the generic dataset at 3 wavelengths and the phantom dataset at 10.	Figures 3.3, 3.4
Calibration of uncertainty estimates	Representative calibration curves for phantom 10 model partitioned/unpartitioned are presented, along with graphical comparison of maximum calibration error for each model across sparsities.	Figures 3.5, 3.6
Effect of dataset re-balancing	Predictions on the generic partitioned dataset at 10 and 40 wavelengths are presented before and after re-balancing with weighted random sampling.	Figures 3.7, 3.8

The maximum absolute calibration error for each model is presented in figure 3.6. The individual calibration curves for each evaluated model, dataset, and sparsity are included in appendix B. Only the maximum errors are considered in the body of

the report as they are considered a reasonable representation of the overall quality of calibration of a model's uncertainty estimates.

Assessment of model performance on sparsities it had not seen during training was not carried out, as the study of SASD suggests the error increases monotonically as the sparsity deviates from the training level [4].

Interpretation of the data is presented in chapter 4.

	Partitioned (Train)		**Unpartitioned (Train)**	
Sparsity	**Part. (Eval) %**	**Unpart. (Eval) %**	**Part. (Eval) %**	**Unpart. (Eval) %**
3	15.2 (6.7, 37.6)	22.5 (9.8, 52.8)	18.1 (8.5, 40.7)	19.3 (8.6, 44.2)
5	11.2 (4.5, 30.9)	16.1 (6.5, 38.9)	10.2 (3.9, 34.3)	11.6 (4.1, 34.5)
10	3.6 (1.4, 15.3)	8.9 (3.5, 26.3)	4.4 (1.7, 17.0)	6.6 (2.2, 21.9)
25	1.6 (0.5, 6.2)	3.0 (1.1, 13.0)	2.0 (0.7, 9.9)	2.4 (0.8, 7.9)
40	1.6 (0.5, 6.7)	1.6 (0.5, 7.6)	1.4 (0.5, 5.7)	1.3 (0.4, 5.2)

Table 3.1: Phantom dataset: Median relative prediction errors and IQR on test set

	Partitioned (Train)		**Unpartitioned (Train)**	
Sparsity	**Part. (Eval) %**	**Unpart. (Eval) %**	**Part. (Eval) %**	**Unpart. (Eval) %**
3	23.3 (11.3, 45.3)	27.8 (14.0, 48.4)	25.5 (13.1, 44.8)	27.9 (14.1, 48.5)
5	22.7 (11.1, 43.9)	24.3 (11.8, 44.7)	22.9 (11.3, 43.6)	24.2 (12.1, 45.0)
10	19.8 (9.8, 41.2)	21.9 (10.8, 42.1)	20.7 (10.0, 41.3)	21.3 (10.4, 42.2)
25	19.6 (8.2, 41.5)	20.5 (9.0, 39.9)	18.7 (8.7, 39.1)	18.8 (9.0, 39.5)
40	18.7 (9.3, 37.6)	19.5 (8.1, 39.8)	18.1 (8.5, 39.3)	17.5 (8.2, 37.7)

Table 3.2: Generic dataset: Median relative prediction errors and IQR on test set

	Partitioned (Train)		**Unpartitioned (Train)**	
Sparsity	**Part. (Eval)**	**Unpart. (Eval)**	**Part. (Eval)**	**Unpart. (Eval)**
3	24.3 (11.9, 46.8)	29.2 (14.5, 50.8)	24.4 (11.9, 47.2)	26.3 (12.9, 48.8)
5	21.2 (10.1, 42.9)	25.2 (12.3, 44.6)	21.7 (10.7, 42.8)	22.8 (11.1, 44.8)
10	23.7 (12.0, 40.1)	24.1 (12.1, 42.4)	22.1 (10.5, 42.3)	22.9 (10.8, 42.8)
25	20.6 (9.7, 40.8)	21.7 (9.8, 41.7)	21.4 (10.6, 39.3)	21.6 (10.2, 42.2)
40	18.7 (9.2, 39.9)	22.4 (10.9, 42.3)	19.2 (8.9, 40.4)	18.9 (8.8, 38.9)

Table 3.3: Melanin dataset: Median relative prediction errors and IQR on test set

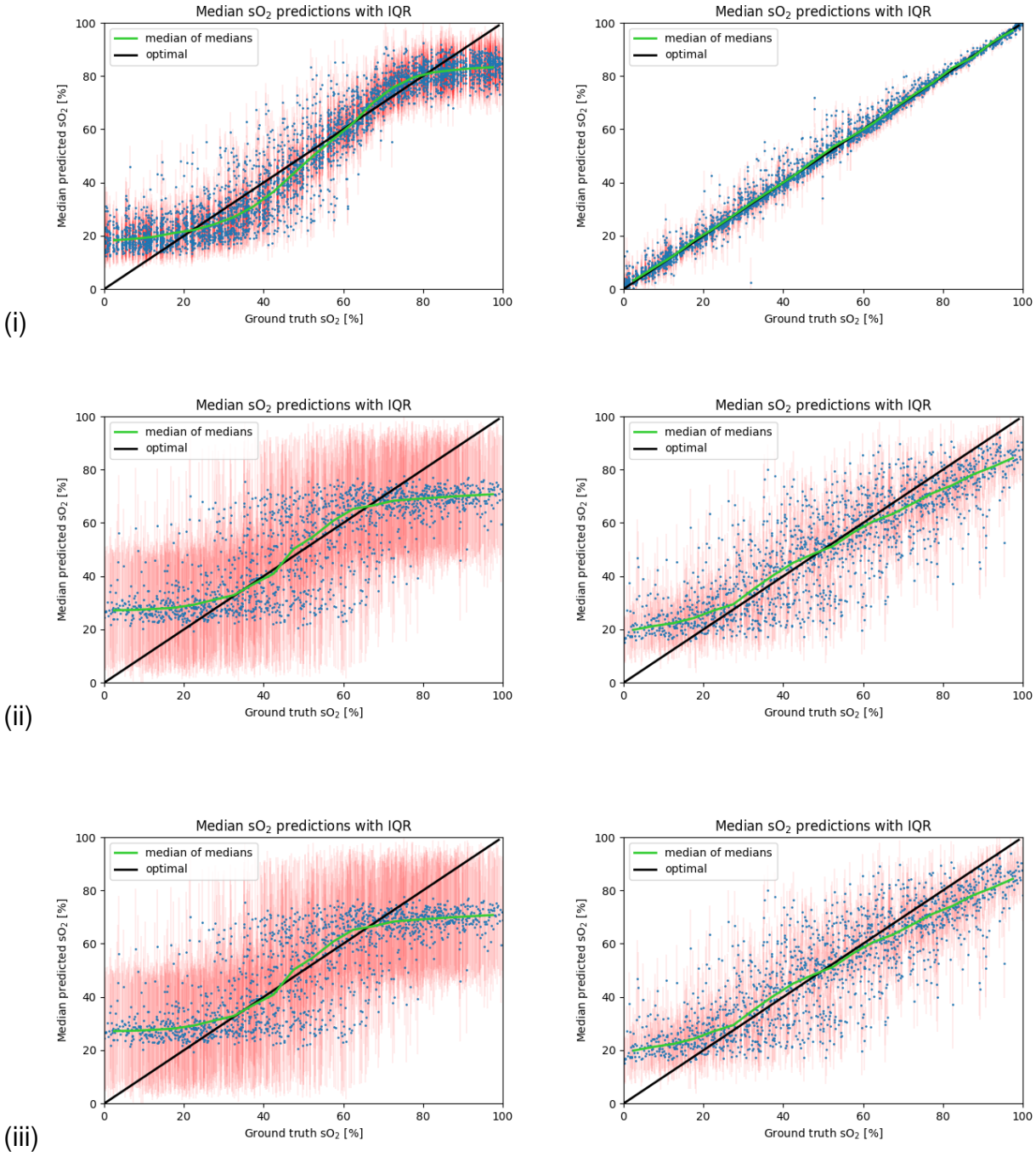
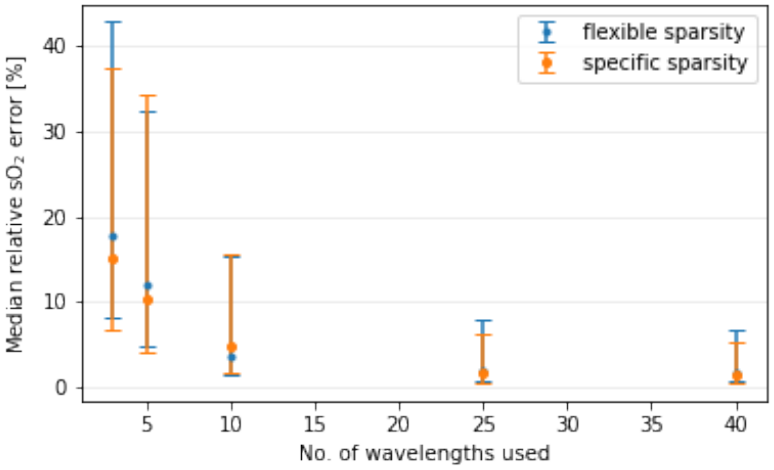


Figure 3.1: sO_2 predictions for (i) phantom, (ii) generic, (iii) melanin datasets (unpartitioned). Each point is the median of 1000 random samples from the Gaussian latent space of the cINN. The error bars are the IQR of the distribution for each median point. The green line shows the median of the estimates in every 5% interval of ground truth sO_2.

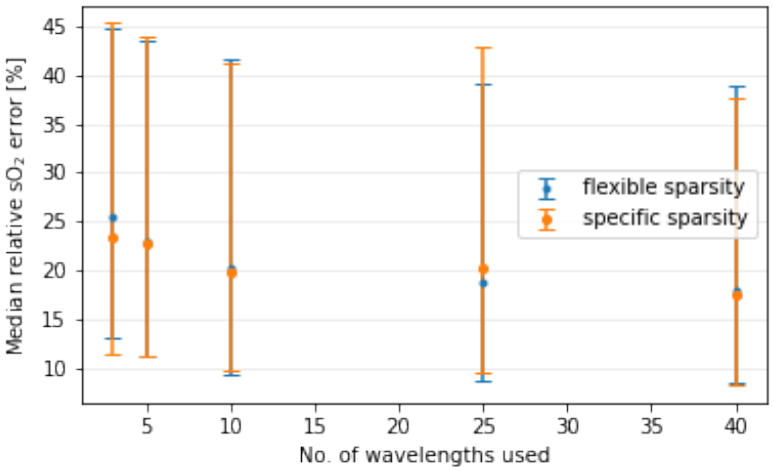
(a)

Phantom dataset



(b)

Generic dataset



(c)

Melanin dataset

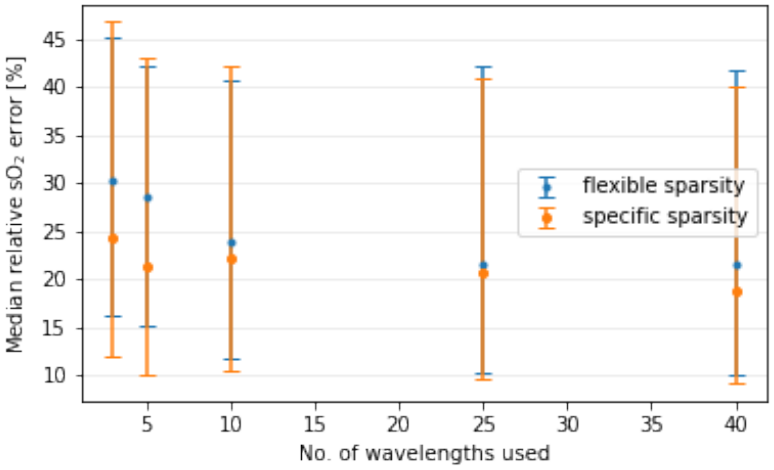
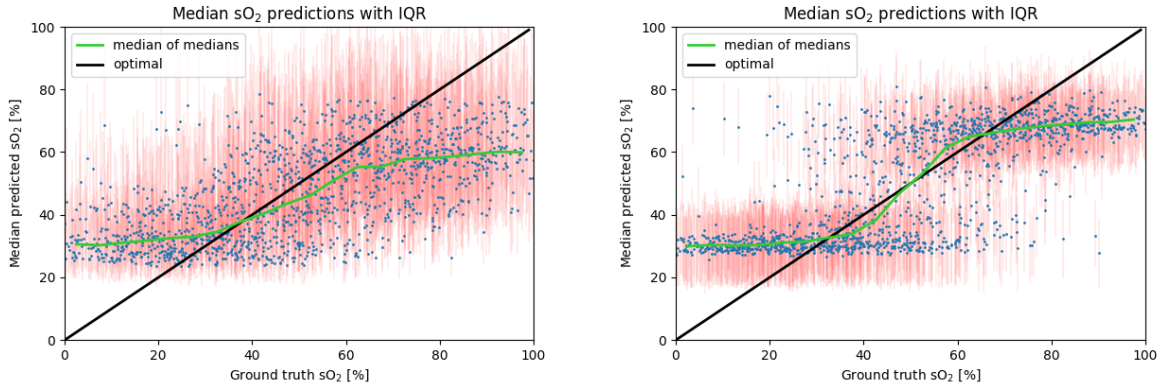


Figure 3.2: Median and IQR of sO_2 relative prediction errors at varying sparsities for flexibly trained model and models trained on specific sparsities. Figures shown are for models trained and evaluated on either partitioned or unpartitioned datasets, whichever provides lower median error at each sparsity.



Unpartitioned Partitioned

Figure 3.3: Median sO_2 predictions and IQRs for model trained on 3 wavelengths for the generic dataset, either (a) unpartitioned or (b) partitioned.

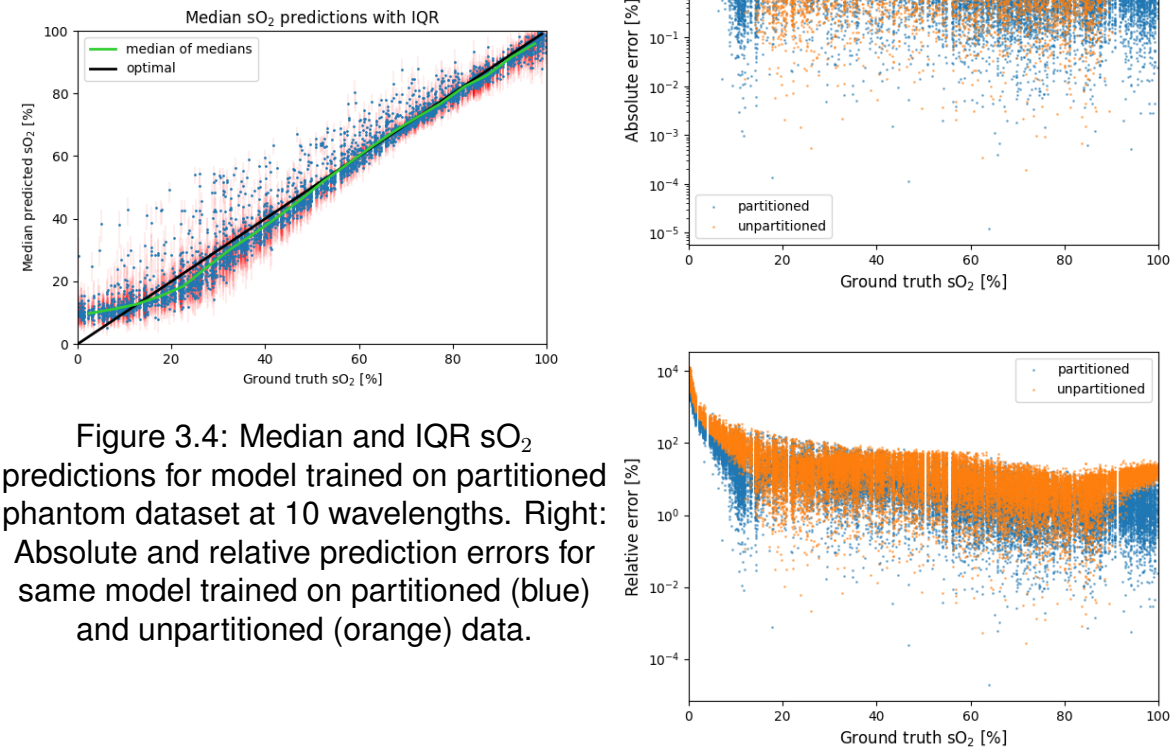
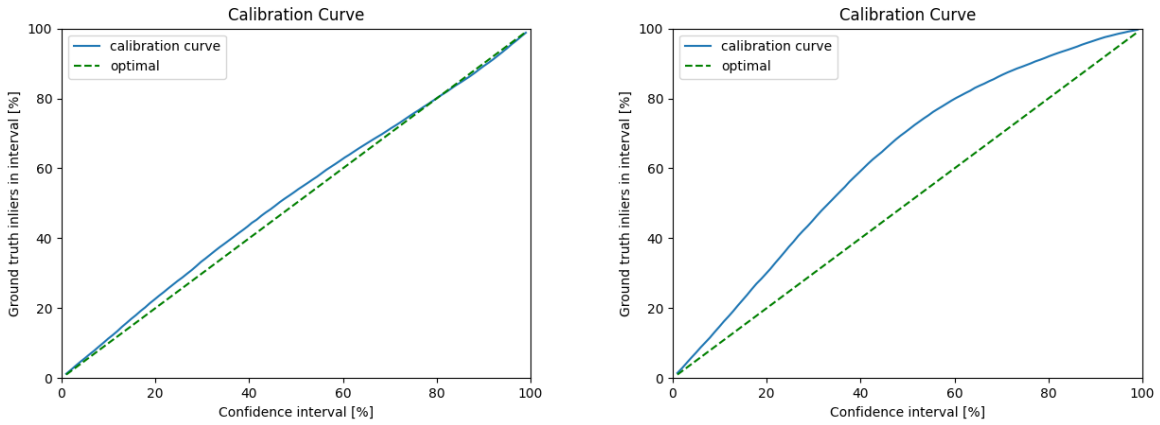


Figure 3.4: Median and IQR sO_2 predictions for model trained on partitioned phantom dataset at 10 wavelengths. Right: Absolute and relative prediction errors for same model trained on partitioned (blue) and unpartitioned (orange) data.



(a) Unpartitioned (b) Partitioned

Figure 3.5: Calibration curves for model trained on 10 wavelengths for the phantom dataset, either (a) unpartitioned or (b) partitioned

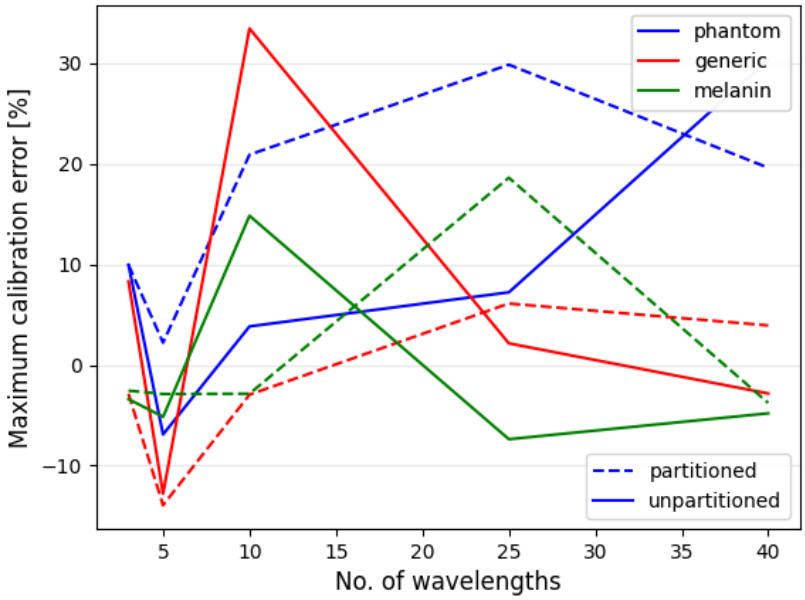
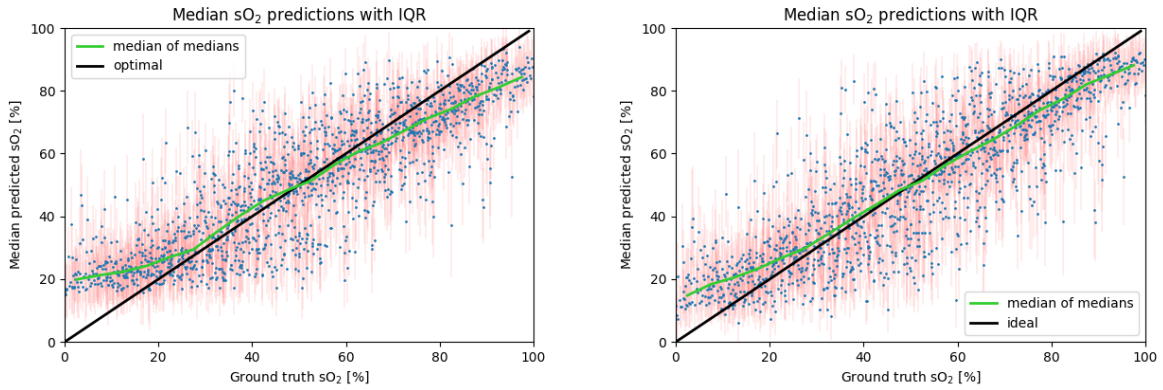


Figure 3.6: Maximum calibration error, based on absolute value, for each dataset at range of sparsities, when trained and evaluated on partitioned or unpartitioned data. Lines are included between points to improve visibility and differentiability of points.

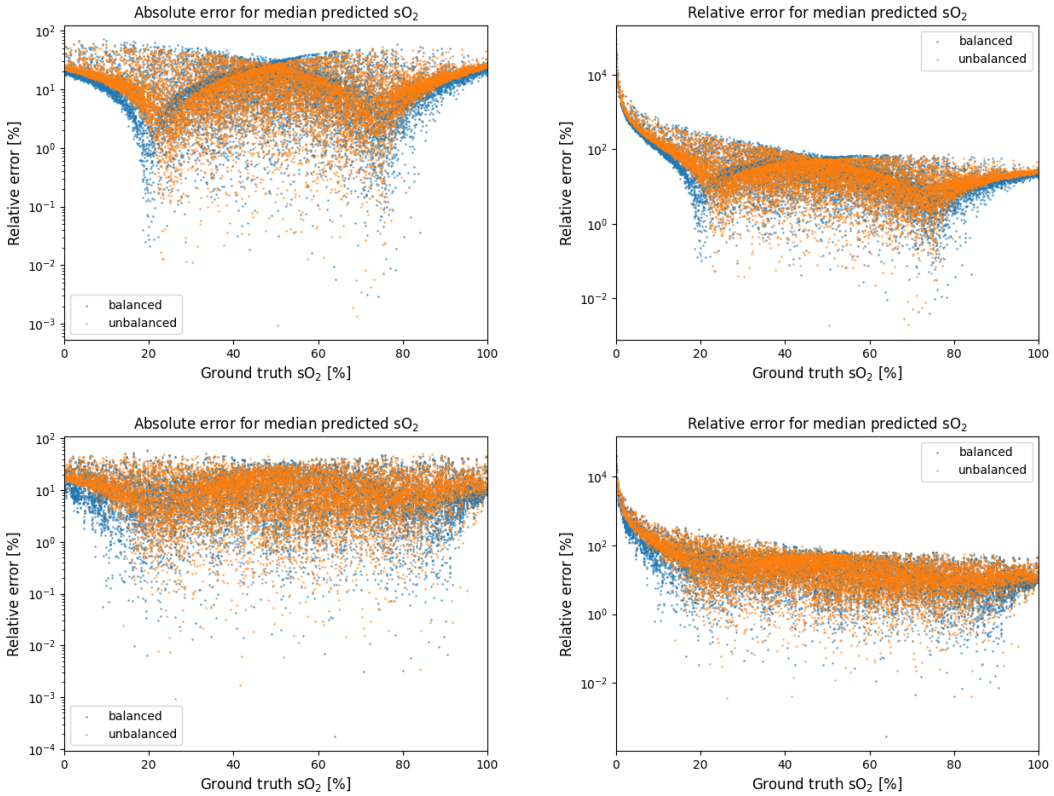


Unbalanced Balanced

Figure 3.7: Median sO_2 predictions and IQRs for model trained on 40 unpartitioned wavelengths for the (a) unbalanced and (b) balanced generic dataset

(i)

(ii)



Absolute error Relative error

Figure 3.8: Absolute and relative errors for median sO_2 predictions for generic dataset at (i) 10 partitioned wavelengths and (ii) 40 unpartitioned wavelengths.

Chapter 4

Discussion

4.1. Comparison to LSD, SASD

The model investigated in this report is designed to combine the flexibility and enhanced performance (compared to LSD) of the SASD method, with the potential for complete reconstruction of a posterior distribution of possible sO_2 values offered by cINNs. The results shown in tables 3.1-3.3 are partly encouraging but also indicative of a problem in model convergence. Models applied to the phantom dataset demonstrate reasonable convergence to an accurate model of the mapping from p_0 to sO_2. This is shown in the median relative prediction error figures of table 3.1 and get close to the SASD accuracy on the same dataset, with median and IQR of 1.3% (0.4% 5.2%) for this method vs. 0.9% (0.3%, 3.5%) for SASD. The shape of the distribution of sO_2 predictions for the phantom dataset at 10 wavelengths, shown in figure 3.1, is distinctively more S-shaped for the cINN than the equivalent plot for SASD, thus revealing a more informative picture of model performance than the aggregate median relative prediction error.

The gap in accuracy on the phantom dataset could likely be closed through tuning of the model hyperparameters. It is worth noting that SASD models were trained and evaluated using the Keras framework. Implementation details between Keras and PyTorch, particularly for LSTM networks, may differ, although an observable performance difference is not expected.

Relative prediction errors on the generic and melanin datasets at 40 wavelengths are close to those produced by LSD but with approximately 2% greater relative prediction error than SASD, with proportionally wider IQR [4]. The comparison breaks down at lower sparsities, however, as the distribution of predictions for melanin and generic at 10 wavelengths resembles a binary classifier (figure 3.1). This binary clas-

sification shows a lack of model convergence or underfitting, in which two cases are identified in the data and no further refinement of the regression occurs. The aggregate median relative prediction error for such a distribution can appear comparable to more accurate distributions but the raw data reveals the shortcomings of the model's performance.

Poorer performance on the generic and melanin datasets is expected as they are the more challenging tissue geometries, closer to *in vivo* data, but they also contain only 42% as many samples as the phantom dataset. A larger training set for the melanin and generic cases could improve model performance.

In compiling this report, a further possible explanation for the poor performance on lower sparsities for the generic and melanin datasets was found in the training routine code. The learning rate scheduler was configured to reduce by a factor of 0.01 every epoch, which is a severe reduction. The Adam optimiser is adaptive and computes a separate learning rate for each model parameter, so the effect of a global learning rate scheduler on the Adam optimiser is not clear at this time and depends on the implementation details, however any effect from the too-low learning rate will likely be to cause slow convergence and/or underfitting. Time constraints have meant this hypothesis has not yet been explored but it remains a hopeful possibility that the architecture can match if not outperform SASD on accuracy with appropriate training and hyperparameter tuning. The idea that a a cINN can match other networks in performance is in agreement with the literature [32] that invertible architectures do not limit the expressivity of a network.[1][2]

The proposed method also retains a key limitation of LSD and SASD currently, namely that the network has to be trained on a dataset that is specific to the target application and that the domain gap, though partly bridged as demonstrated by *in vitro* and *in vivo* studies [15], is still considerable.

4.2. Model calibration and uncertainty estimates

The primary aim of the work presented in this report is to demonstrate that a cINN with, LSTM as conditioning network, is able to produce a well calibrated uncertainty estimate as part of its output. The quality of the uncertainty estimate, and by

[2] cINNs may have greater difficulty than feed-forward networks on some tasks, however, because they are a generative model. From [45]: Generative models *'tend to make stronger assumptions on the data than their purely discriminative counterparts, often leading to higher asymptotic bias when the model is wrong. If one is solely interested in learning to discriminate, and one is in a regime with a sufficiently large amount of data, then purely discriminative models typically will lead to fewer errors in discriminative tasks.'*

extension the similarity of the estimated posterior distribution of sO_2 to the true posterior, was quantified by the calculating the calibration error.

Figure 3.5 show typical calibration curves and figure 3.6 summarises the maximum calibration error (at any confidence interval) for each model. The pattern demonstrated has a few features to comment on. At 5 wavelengths, each model experiences an increase in confidence from 3, verging into overconfidence for most. At 40 wavelengths, the maximum calibration error reduces for the generic and melanin datasets whether partitioned or not. For the phantom dataset, the opposite is observed as the calibration error increases at low sparsity. This can be attributed to the model's high accuracy at this sparsity, causing more ground truth inliers in every confidence interval than their should be for a well calibrated model.

The maximum calibration errors in general can be considered far from optimal and that estimates are generally underconfident. The results suggest improvements in training and hyperparameter tuning are required to improve the model's ability to generate the true posterior.

One major shortcoming of the use of median and IQR, and by extension calibration error, as the primary metrics to quantify uncertainty, is the inability to capture information on multimodal distributions. Devising suitable metrics to quantify the multimodality of the prior distributions produced by random sampling of the latent space would be a significant enhancement to this work. Multimodality can be identified visually by plotting the prior distribution as a histogram, but a quantitative method could identify it automatically [34].

While a high calibration error is concerning as it indicates that the estimated posterior distribution is significantly different from the true distribution, its practical impact on the uncertainty estimates can be mitigated through the use of a correction factor. Prior work by Gróhl [10] has shown that a correction factor optimised on the validation set can converge the calibration error to the optimal case for cINNs.

4.3. Effect of spectral partitioning

The median relative error results in 3.1-3.3 show, unsurprisingly, that models perform better when evaluated against datasets that have the same partitioning (or lack thereof) as they were trained on. Therefore only the first (after sparsity) and last columns of tables 3.1-3.3 contain data relevant to the rest of this discussion.

Two examples are highlighted in figures 3.3 and 3.4 to show the potential positive impact of spectral partitioning on model performance. The model trained on the generic dataset at 3 wavelengths shows very little convergence in the unpartitioned

case but demonstrates a binary classification ability in the partitioned case. This tendency to binary classification is also evident at 10 wavelengths 3.1 for the generic dataset, which suggests that partitioning helped the model to converge at a lower sparsity. Although the resulting model is only marginally better than a random number generator, the improvement compared to the unpartitioned case is considerable (4.6% lower median relative error).

The other positive case is the phantom dataset at 10 wavelengths (figure 3.4). Partitioning results in a near halving (3.6% vs. 6.6% of) the median relative prediction error and a narrower IQR. The benefits of partitioning can be observed qualitatively by comparison of figure 3.4 with the equivalent error bar plot in figure 3.1, or can be more readily observed by the contrasting relative and absolute median prediction errors for the two cases shown in the right plots of figure 3.4. The considerably lowered blue errors at the extremes indicate the partitioning has largely improved the predictions for high and low sO_2 values.

The reason for improved performance may be that certain spectral characteristics of the underlying chromophores cause certain parts of the 700-900nm NIR wavelength range utilised to be more information rich than others. Partitioning may therefore increase the probability of multiple wavelengths falling into informative wavelength ranges for lower sparsities, and that this benefit may be even more pronounced at middling sparsities, i.e. 10 wavelengths, due to the even more consistent spectral separation.

Partitioning degrades performance slightly at 40 wavelengths for the phantom dataset. This may be because, as there are 41 wavelengths total, the each partition will have size 1 except the final partition, containing 2. The final wavelength will therefore be chosen randomly from these 2, whereas in the unpartitioned case, the missing wavelength could be any of the 41. This may help reduce overfitting slightly in comparison to the partitioned case.

Overall, there seem to be benefits to spectral partitioning at lower sparsities, as suggested by the results from the phantom dataset, particularly for 10 wavelengths, and there are plausible reasons to believe so. However, the results are ultimately inconclusive and spectral partitioning may be of limited use when applied to real data acquired with fixed wavelengths that may not be as well separated.

The effect of partitioning on the model calibration is less clear. The relationship between the maximum calibration errors and partitioning across datasets and sparsities, which can be interpreted from figure 3.6, is not easily perceived.

4.4. Effect of training on flexible sparsity

Training on a range of sparsities was conceived as a possible path to greater network generalisability and a way to also reduce the computational burden of training a dedicated model per sparsity. Figure 3.2 compares the median relative prediction error and its IQR for specific-sparsity networks and a flexible network for each dataset. The results show the flexible models to produce higher median errors for nearly every datapoint presented. Qualitative assessment of the underlying distribution of predictions, not presented in this report, also confirm the flexible networks do not converge to the true distribution as strongly as sparsity-specific networks. This confirms previous findings on flexible training for SASD [4] and suggests the idea can be rejected.

4.5. Effect of dataset re-balancing

One unambiguously positive finding of the investigation is the previously unidentified imbalance of the training datasets due to partial volume effects (see section 2.4.3). The effects of rebalancing to include more extreme sO_2 values is shown qualitatively by the distributions of predictions in figure 3.7 for the generic dataset, where the median prediction line of the distribution is seen to bend more towards optimal at the extremes after re-balancing.

This effect is more clearly illustrated by the corresponding absolute and relative prediction errors in figures 3.8. Errors on the balanced dataset are visibly lower at the extremes. The effect on the net median errors is negligible or even a slight increase, but this could be due to the resampling causing the network to see less of the $\sim$50% sO_2 values. Since sO_2 in the 80-100% range is more clinically relevant for in-vivo imaging, this trade-off may be considered worthwhile.

Chapter 5

Conclusion

5.1. Summary of Achievements

To the best of our knowledge, this report presents a novel model architecture in the use of a LSTM as conditioning network to a cINN. The architecture is demonstrated to perform comparably to previous state of the art methods in predicting sO_2 values from single-pixel initial pressure spectra for an *in silico* flow phantom dataset, with median and IQR relative prediction error of 1.3% (0.4%, 5.2%) at 40 wavelengths. The encouraging performance on the phantom dataset suggests that a sub-optimal training routine is at least partly responsible for the poorer-than-expected performance on the generic and melanin datasets, and changes to learning rate are suggested to correct this.

The architecture retains the wavelength flexibility of the SASD method but introduces a mechanism (latent space sampling) to produce theoretically sound uncertainty estimates. The calibration error of the trained models is throughly investigated and found to be sub-optimal, but improved training could also benefit this metric.

Training on a flexible range of sparsities is investigated as a method to improve generalisability to unseen data but is found to produce higher median relative predictive errors, so is rejected in favour of models trained on specific numbers of wavelengths.

Encouraging results are obtained for spectral partitioning, which is shown to improve accuracy at high sparsity for the *in silico* flow phantom dataset, as long as both training and evaluation datasets are partitioned.

The report also highlights the training dataset imbalance caused by partial volume effects, and the utility of weighted random sampling to re-balance datasets and improve extremal sO_2 predictions.

5.2. Future Work

The relatively poor performance on the generic and melanin datasets leaves plenty of reasons to re-examine the encouraging results of the investigation to confirm their suggestions. A retraining of the models with a more suitable learning rate and systematic exploration of the model hyperparameter space was not carried out in this work due to time constraints, but the effects of flexible sparsity training and spectral partitioning should be re-examined with retrained models that display improved accuracy on the generic and melanin datasets. Any re-trained models should also be evaluated against *in vitro* and *in vivo* datasets, to further the comparison with SASD and assess the model's ability to translate into unseen scenarios and clinical practice.

A further possible improvement to the model architecture could be the use of a bi-directional LSTM, to extract sequential dependencies in both spectral directions. Implementation of bidirectional training is an augmentation to the training routine that could theoretically improve convergence [32]. As always in learned methods, greater quantity and quality of training data could aid in closing the domain gap and improving clinical applicability of qPAI.

To improve the model evaluation process, methods to detect bimodal distributions would enhance the assessment of the posterior distribution and resulting uncertainty statistics. A curated dataset featuring pixels that are strongly expected to have bimodal distributions of sO_2 (for example due to shielding vessels containing blood of different sO_2 as in [34]) could even be generated to test the method's resiliency to and ability to detect multimodal distributions.

Overall, the results of this investigation and the theoretical considerations suggest the model architecture has great promise and, with refinement to the training routine, could become the state-of-the-art in learned photoacoustic oximetry.

Appendices

Appendix A - Link to Github repository

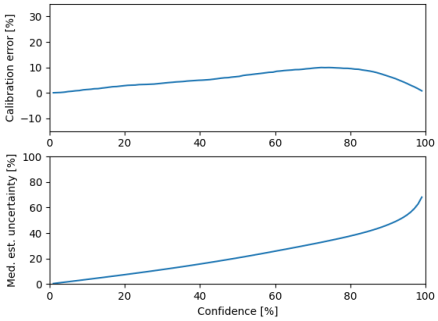
All code used in the work presented in this report is available at the GitHub repository:

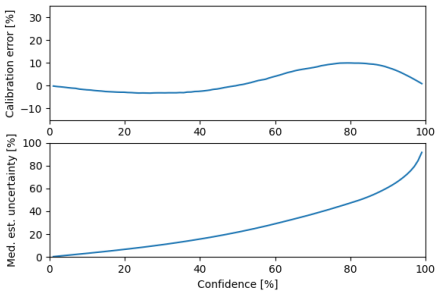
`https://github.com/micdoh/qPAI_cINN_uncertainty_estimation.git`

Appendix B - Model calibration errors

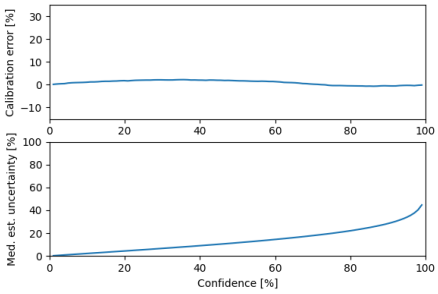
The following pages show the variation of calibration error and median estimated uncertainty with confidence interval for the models trained on each partitioned and unpartitioned dataset and sparsity. The first page shows the plots for the phantom dataset, second for generic and third for melanin.

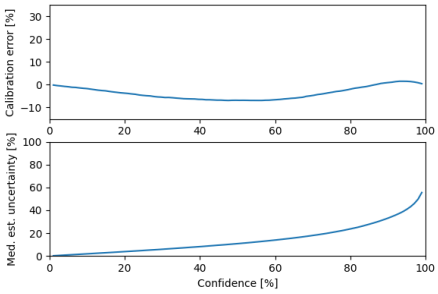
(3)



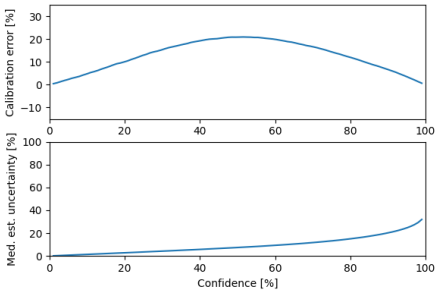


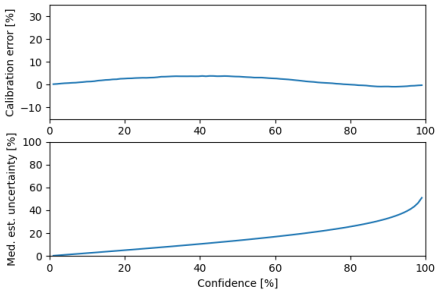
(5)



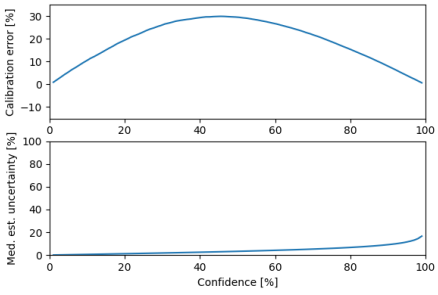


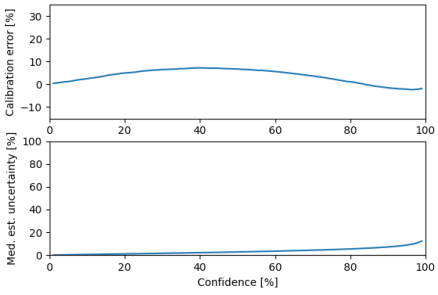
(10)



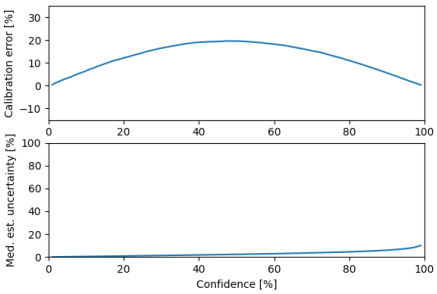


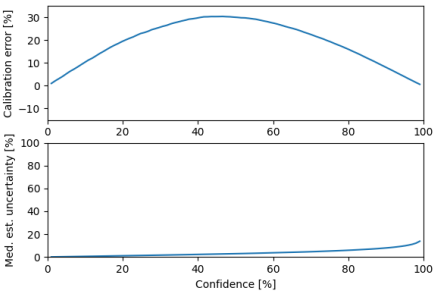
(25)





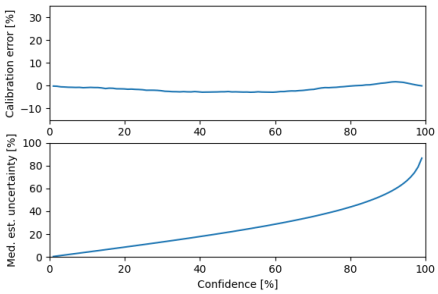
(40)

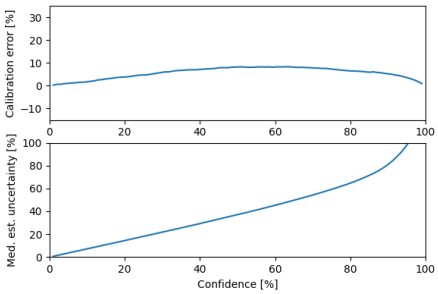




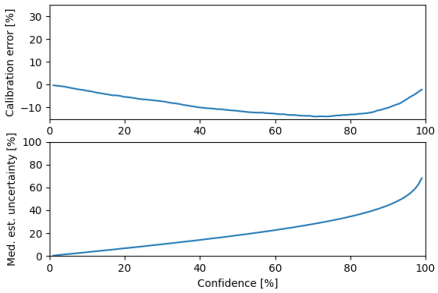
Phantom: Partitioned Unpartitioned

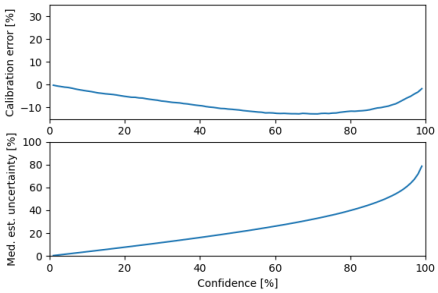
(3)



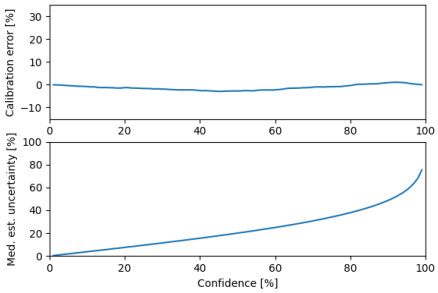


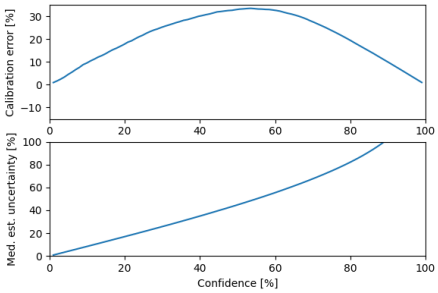
(5)



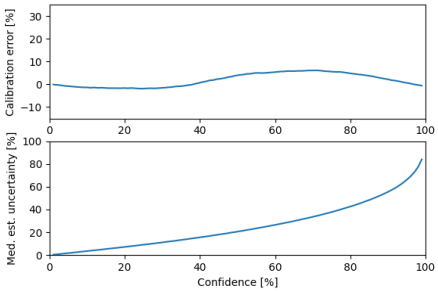


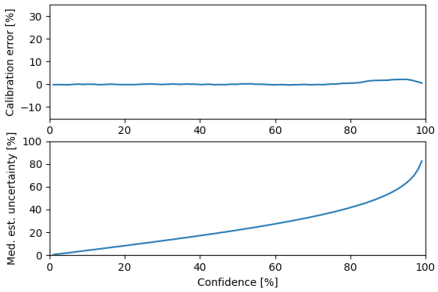
(10)



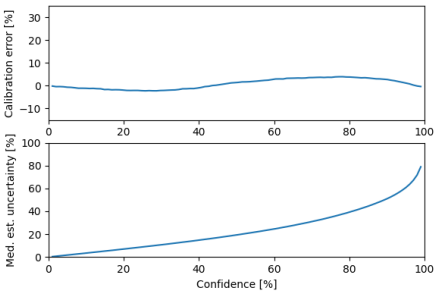


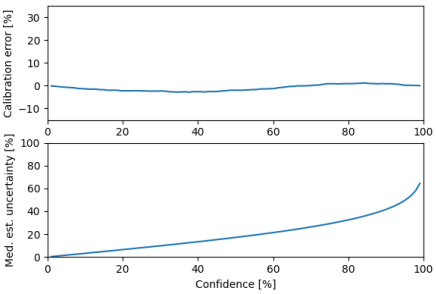
(25)





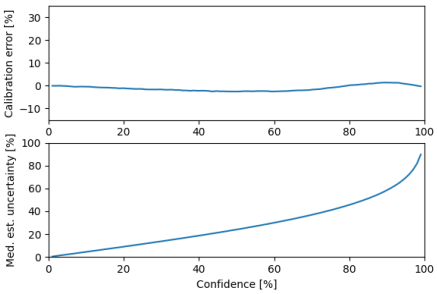
(40)

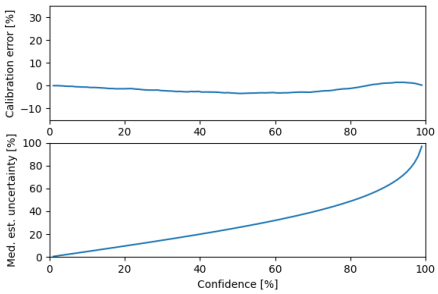




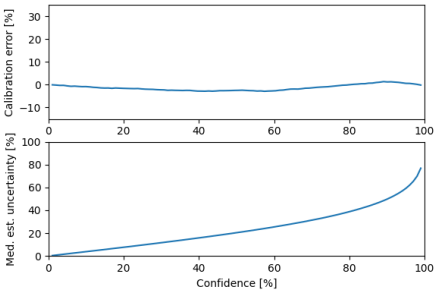
Generic: Partitioned Unpartitioned

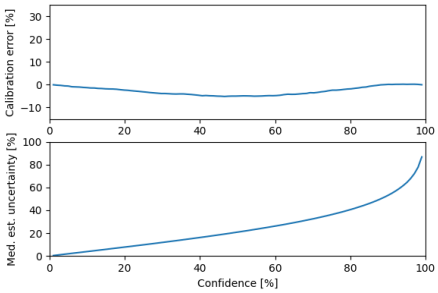
(3)



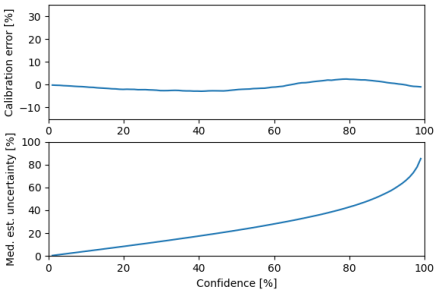


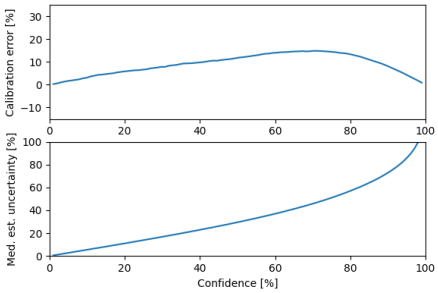
(5)



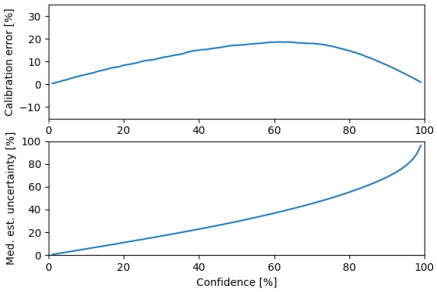


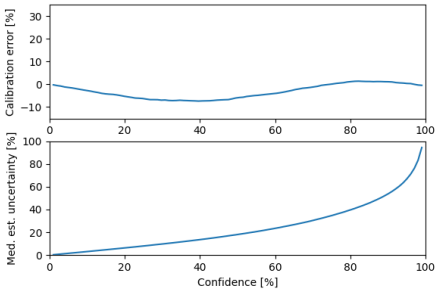
(10)



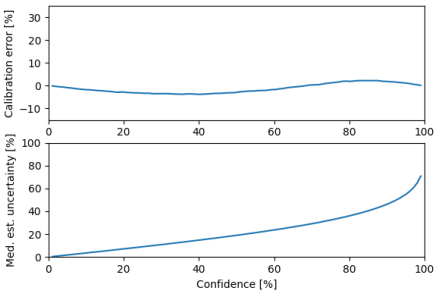


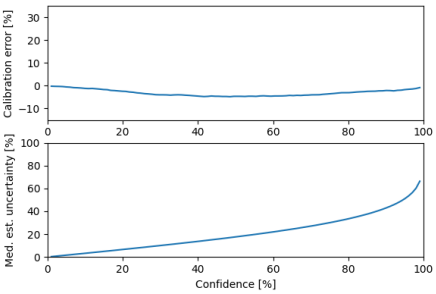
(25)





(40)





Melanin: Partitioned

Unpartitioned

Bibliography

[1] S. Zackrisson, S. van de Ven, and S. Gambhir, "Light In and Sound Out: Emerging Translational Strategies for Photoacoustic Imaging", *Cancer Research*, vol. 74, no. 4, pp. 979–1004, Feb. 2014, ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-13-2387. eprint: https://aacrjournals.org/cancerres/article-pdf/74/4/979/2715151/979.pdf. [Online]. Available: https://doi.org/10.1158/0008-5472.CAN-13-2387.

[2] C. Bench, "Data-driven Quantitative Photoacoustic Tomography", PhD thesis, University College London, 2022.

[3] B. Kompa, J. Snoek, and A. L. Beam, "Second opinion needed: communicating uncertainty in medical machine learning", *npj Digital Medicine*, vol. 4, no. 1, p. 4, 2021. DOI: 10.1038/s41746-020-00367-3. [Online]. Available: https://doi.org/10.1038/s41746-020-00367-3.

[4] K. Gu, "Sparsity accustomed spectral decolouring: A long short-term memory neural network for quantitative photoacoustic oximetry", Part III Thesis, University of Cambridge, 2022.

[5] B. Cox, J. Laufer, and P. Beard, "The challenges for quantitative photoacoustic imaging", vol. 7177, Feb. 2009. DOI: 10.1117/12.806788.

[6] L. V. Wang and H. Wu, *Biomedical optics: principles and imaging*. John Wiley Sons, 2012. DOI: 10.1002/9780470177013.

[7] L. V. Wang and S. Hu, "Photoacoustic Tomography: In Vivo Imaging from Organelles to Organs", *Science*, vol. 335, no. 6075, pp. 1458–1462, 2012. DOI:

10.1126/science.1216210. eprint: https://www.science.org/doi/pdf/10.1126/science.1216210. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1216210.

[8] C. Kim, T. N. Erpelding, L. Jankovic, M. D. Pashley, and L. V. Wang, "Deeply penetrating in vivo photoacoustic imaging using a clinical ultrasound array system", *Biomed. Opt. Express*, vol. 1, no. 1, pp. 278–284, 2010. DOI: 10.1364/BOE.1.000278. [Online]. Available: http://opg.optica.org/boe/abstract.cfm?URI=boe-1-1-278.

[9] S. Manohar, S. E. Vaartjes, J. C. G. van Hespen, J. M. Klaase, F. M. van den Engh, W. Steenbergen, and T. G. van Leeuwen, "Initial results of in vivo non-invasive cancer imaging in the human breast using near-infrared photoacoustics", *Opt. Express*, vol. 15, no. 19, pp. 12 277–12 285, 2007. DOI: 10.1364/OE.15.012277. [Online]. Available: http://opg.optica.org/oe/abstract.cfm?URI=oe-15-19-12277.

[10] J. Gröhl, "Data-driven Quantitative Photoacoustic Imaging", PhD thesis, Ruprecht Karl University of Heidelberg, 2020.

[11] G. Wissmeyer, M. A. Pleitez, A. Rosenthal, and V. Ntziachristos, "Looking at sound: optoacoustics with all-optical ultrasound detection", *Light: Science & Applications*, vol. 7, no. 1, p. 53, 2018. DOI: 10.1038/s41377-018-0036-7. [Online]. Available: https://doi.org/10.1038/s41377-018-0036-7.

[12] C. Tian, C. Zhang, H. Zhang, D. Xie, and Y. Jin, "Spatial Resolution in Photoacoustic Computed Tomography", *Reports on Progress in Physics*, vol. 84, Jan. 2021. DOI: 10.1088/1361-6633/abdab9.

[13] M. Xu and L. V. Wang, "Universal back-projection algorithm for photoacoustic computed tomography", in *Photons Plus Ultrasound: Imaging and Sensing 2005: The Sixth Conference on Biomedical Thermoacoustics, Optoacoustics, and Acousto-optics*, A. A. Oraevsky and L. V. Wang, Eds., International Society

for Optics and Photonics, vol. 5697, SPIE, 2005, pp. 251 –254. DOI: 10.1117/12.589146. [Online]. Available: https://doi.org/10.1117/12.589146.

[14] C. Haisch, "Quantitative analysis in medicine using photoacoustic tomography", *Analytical and Bioanalytical Chemistry*, vol. 393, no. 2, pp. 473–479, 2009. DOI: 10.1007/s00216-008-2479-9. [Online]. Available: https://doi.org/10.1007/s00216-008-2479-9.

[15] J. Gröhl, T. Kirchner, T. J. Adler, L. Hacker, N. Holzwarth, A. Hernández-Aguilera, M. A. Herrera, E. Santos, S. E. Bohndiek, and L. Maier-Hein, "Learned spectral decoloring enables photoacoustic oximetry", *Scientific Reports*, vol. 11, no. 1, p. 6565, 2021. DOI: 10.1038/s41598-021-83405-8. [Online]. Available: https://doi.org/10.1038/s41598-021-83405-8.

[16] C. Cai, K. Deng, C. Ma, and J. Luo, "End-to-end deep neural network for quantitative photoacoustic imaging", *Optics Letters*, vol. 43, pp. 2752–2755, May 2018. DOI: 10.1364/OL.43.002752.

[17] J. Grhl, M. Schellenberg, K. Dreher, and L. Maier-Hein, "Deep learning for biomedical photoacoustic imaging: A review", *Photoacoustics*, vol. 22, p. 100 241, 2021, ISSN: 2213-5979. DOI: https://doi.org/10.1016/j.pacs.2021.100241. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213597921000033.

[18] Z. Ren, G. Liu, and Y. Ding, "Effects of multiple factors on the photoacoustic detection of glucose based on artificial neural network", English, in *Proceedings of SPIE - The International Society for Optical Engineering*, Cited By :2, vol. 10820, 2018. [Online]. Available: www.scopus.com.

[19] I. Olefir, S. Tzoumas, C. Restivo, P. Mohajerani, L. Xing, and V. Ntziachristos, "Deep Learning-Based Spectral Unmixing for Optoacoustic Imaging of Tissue Oxygen Saturation", *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3643–3654, 2020. DOI: 10.1109/TMI.2020.3001750.

[20] J. Gröhl, K. K. Dreher, M. Schellenberg, T. Rix, N. Holzwarth, P. Vieten, L. Ayala, S. E. Bohndiek, A. Seitel, and L. Maier-Hein, "SIMPA: an open-source toolkit for simulation and image processing for photonics and acoustics", *Journal of Biomedical Optics*, vol. 27, no. 8, p. 083 010, 2022. DOI: 10.1117/1.JBO.27.8.083010. [Online]. Available: https://doi.org/10.1117/1.JBO.27.8.083010.

[21] M. Schellenberg, J. Gröhl, K. Dreher, N. Holzwarth, M. D. Tizabi, A. Seitel, and L. Maier-Hein, *Data-driven generation of plausible tissue geometries for realistic photoacoustic image synthesis*, 2021. DOI: 10.48550/ARXIV.2103.15510. [Online]. Available: https://arxiv.org/abs/2103.15510.

[22] T. Kirchner, J. Grhl, and L. Maier-Hein, "Context encoding enables machine learning-based quantitative photoacoustics", *Journal of Biomedical Optics*, vol. 23, no. 05, p. 1, 2018. DOI: 10.1117/1.jbo.23.5.056008. [Online]. Available: https://doi.org/10.1117%2F1.jbo.23.5.056008.

[23] K. Hoffer-Hawlik, A. V. Namen, and G. P. Luke, "Quantitative photoacoustic oximetry using convolutional neural networks (Conference Presentation)", in *Photons Plus Ultrasound: Imaging and Sensing 2020*, A. A. Oraevsky and L. V. Wang, Eds., International Society for Optics and Photonics, vol. 11240, SPIE, 2020, 112402A. DOI: 10.1117/12.2545197. [Online]. Available: https://doi.org/10.1117/12.2545197.

[24] C. Bench, A. Hauptmann, and B. T. Cox, "Toward accurate quantitative photoacoustic imaging: learning vascular blood oxygen saturation in three dimensions", *Journal of Biomedical Optics*, vol. 25, 2020.

[25] S. Liao, Y. Gao, A. Oto, and D. Shen, "Representation Learning: A Unified Deep Learning Framework for Automatic Prostate MR Segmentation", vol. 16, Sep. 2013, pp. 254–61, ISBN: 978-3-642-38708-1. DOI: 10.1007/978-3-642-40763-5_32.

[26] T. Kirchner and M. Frenz, “Multiple illumination learned spectral decoloring for quantitative optoacoustic oximetry imaging”, *Journal of Biomedical Optics*, vol. 26, Aug. 2021. DOI: 10.1117/1.JBO.26.8.085001.

[27] B. Lakshminarayanan, A. Pritzel, and C. Blundell, *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*, 2016. DOI: 10.48550/ARXIV.1612.01474. [Online]. Available: https://arxiv.org/abs/1612.01474.

[28] A. Olivier, M. D. Shields, and L. Graham-Brady, “Bayesian neural networks for uncertainty quantification in data-driven materials modeling”, *Computer Methods in Applied Mechanics and Engineering*, vol. 386, p. 114 079, 2021, ISSN: 0045-7825. DOI: https://doi.org/10.1016/j.cma.2021.114079. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045782521004102.

[29] Y. Gal and Z. Ghahramani, *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, 2015. DOI: 10.48550/ARXIV.1506.02142. [Online]. Available: https://arxiv.org/abs/1506.02142.

[30] G. Godefroy, B. Arnal, and E. Bossy, “Compensating for visibility artefacts in photoacoustic imaging with a deep learning approach providing prediction uncertainties”, *Photoacoustics*, vol. 21, p. 100 218, 2021. DOI: 10.1016/j.pacs.2020.100218. [Online]. Available: https://doi.org/10.1016%2Fj.pacs.2020.100218.

[31] D. P. Kingma and M. Welling, “An Introduction to Variational Autoencoders”, *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019. DOI: 10.1561/2200000056. [Online]. Available: https://doi.org/10.1561%2F2200000056.

[32] L. Ardizzone, J. Kruse, S. Wirkert, D. Rahner, E. W. Pellegrini, R. S. Klessen, L. Maier-Hein, C. Rother, and U. Kthe, *Analyzing Inverse Problems with Invertible Neural Networks*, 2018. DOI: 10.48550/ARXIV.1808.04730. [Online]. Available: https://arxiv.org/abs/1808.04730.

[33] I. Kobyzev, S. J. Prince, and M. A. Brubaker, “Normalizing Flows: An Introduction and Review of Current Methods”, *IEEE Transactions on Pattern Analysis and*

Machine Intelligence, vol. 43, no. 11, pp. 3964–3979, 2021. DOI: 10.1109/tpami.2020.2992934. [Online]. Available: https://doi.org/10.1109%2Ftpami.2020.2992934.

[34] J.-H. Nölke, T. Adler, J. Groehl, T. Kirchner, L. Ardizzone, C. Rother, U. Köthe, and L. Maier-Hein, "Invertible Neural Networks for Uncertainty Quantification in Photoacoustic Imaging", 2020. DOI: 10.48550/ARXIV.2011.05110. [Online]. Available: https://arxiv.org/abs/2011.05110.

[35] J. Gröhl, T. Kirchner, T. Adler, and L. Maier-Hein, "Confidence Estimation for Machine Learning-Based Quantitative Photoacoustics", *Journal of Imaging*, vol. 4, p. 147, Dec. 2018. DOI: 10.3390/jimaging4120147.

[36] P. Cheridito, A. Jentzen, and F. Rossmannek, "Efficient Approximation of High-Dimensional Functions With Neural Networks", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 7, pp. 3079–3093, 2022. DOI: 10.1109/tnnls.2021.3049719. [Online]. Available: https://doi.org/10.1109%2Ftnnls.2021.3049719.

[37] L. Dinh, D. Krueger, and Y. Bengio, *NICE: Non-linear Independent Components Estimation*, 2014. arXiv: 1410.8516 [cs.LG].

[38] L. Ardizzone, C. Lth, J. Kruse, C. Rother, and U. Kthe, *Guided Image Generation with Conditional Invertible Neural Networks*, 2019. DOI: 10.48550/ARXIV.1907.02392. [Online]. Available: https://arxiv.org/abs/1907.02392.

[39] R. C. Staudemeyer and E. R. Morris, *Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks*, 2019. DOI: 10.48550/ARXIV.1909.09586. [Online]. Available: https://arxiv.org/abs/1909.09586.

[40] L. Ardizzone, T. Bungert, F. Draxler, U. Kthe, J. Kruse, R. Schmier, and P. Sorrenson, *Framework for Easily Invertible Architectures (FrEIA)*, 2018-2022. [Online]. Available: https://github.com/VLL-HD/FrEIA.

[41] D. P. Kingma and P. Dhariwal, *Glow: Generative Flow with Invertible 1x1 Convolutions*, 2018. DOI: 10.48550/ARXIV.1807.03039. [Online]. Available: https://arxiv.org/abs/1807.03039.

[42] L. Dinh, J. Sohl-Dickstein, and S. Bengio, *Density estimation using Real NVP*, 2016. DOI: 10.48550/ARXIV.1605.08803. [Online]. Available: https://arxiv.org/abs/1605.08803.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, 2019. DOI: 10.48550/ARXIV.1912.01703. [Online]. Available: https://arxiv.org/abs/1912.01703.

[44] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", *CoRR*, vol. abs/1412.6980, 2015.

[45] A. Banerjee, "An Analysis of Logistic Models: Exponential Family Connections and Online Performance", in *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)*, pp. 204–215. DOI: 10.1137/1.9781611972771.19. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611972771.19. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9781611972771.19.