Chair of Biological Imaging
TUM School of Medicine and Health
Technical University of Munich

Master's Thesis

Assessment of Unmixing Approaches for Multispectral Optoacoustic Tomography

Jérémie Gillet

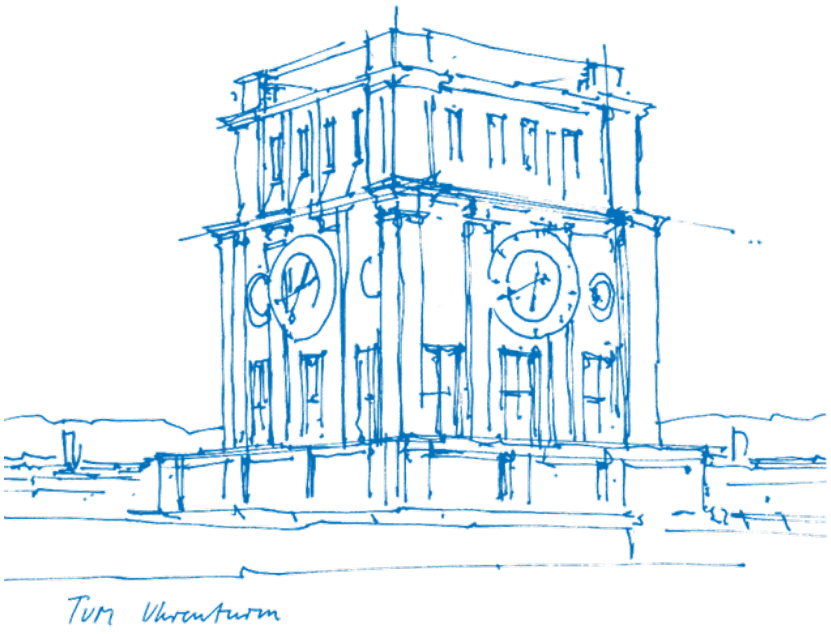
Thesis for the Attainment of the Degree
MSc Mechanical Engineering
at TUM School of Engineering and Design

Examiner
Prof. Vasilis Ntziachristos
Chair of Biological Imaging
TUM School of Medicine and Health

Supervised by
Guillaume Zahnd
iThera Medical GmbH

Submitted by
Jérémie Gillet
Matriculation Number: 03782940

Submitted on
May 19, 2025



Declaration of Academic Integrity

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.

I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted. Further rights of reproduction and usage, however, are not granted here.

This paper was not previously presented to another examination board and has not been published.

Munich, May 19, 2025 Jérémie Gillet

Acknowledgment

I would first and foremost like to thank Dr. Guillaume Zahnd for his trust and support along these few months during which I was able to learn a lot, both from a scientific and professional point of view, as well as for the equilibrium that he found between close supervision and giving me a lot of research liberty and responsibility. I highly value the research discussions that we had and all his advice.

Warm thanks also go to Dr. Christophe Dehner, Dr. Charlène Reichl, Dr. Yi Qiu, Dr. Antonia Longo and Dr. Ledia Ladj for their availability and help in making me better understand and tackle the different aspects of my thesis.

I also would like to thank Moritz Schillinger and Dr. Katharina Breininger from the Department of Artificial Intelligence in Biomedical Engineering (AIBE) at the FAU Erlangen for their advice and the research leads they encouraged me to explore.

A special thanks goes to J.H. Nölke who accepted to share some of his code, which has helped me familiarizing way faster with the Deep Learning techniques that I used in this work.

Thank you to Prof. Vasilis Ntziachristos for raising my interest in biological and medical imaging and accepting to supervise me on the university side.

Thanks also to my colleagues at iThera for their kindness and all the bonds that we created.

Last but not least, thanks to my family and friends for their strong support and their invaluable help in finding my way.

Abstract

Photoacoustic imaging is a non-invasive imaging approach capable of high spatial resolution and centimeter-scale depth penetration into living tissue. This emergent technology provides a combination of anatomical information typical of ultrasound imaging as well as functional information typical of optical imaging. In Multi-Spectral Optoacoustic Tomography (MSOT), we are theoretically able to retrieve information such as blood oxygenation in tissues, that are highly valuable in diagnosing or monitoring diseases such as Peripheral Artery Disease (PAD), where we are mostly interested in assessing muscle oxygenation. To this end, two inverse problems have to be solved: the acoustic inverse problem, which consists in reconstructing a pressure distribution in tissues from the acoustic time series received in the transducer, and the optical inverse problem, solved by "unmixing" acoustic responses of a tissue to incident signals of different wavelengths to retrieve blood oxygenation.

Although some robust and time-wise efficient algorithms already exist in the clinical routine to perform acoustic inversion, the optical one is an active research topic in the scientific community. The most commonly used algorithm, "Linear Unmixing", is not reliable enough, and complex numerical methods are more accurate but not computable in real time. Surrogate deep-learning methods have however shown promising results in this regard for multiple use cases recently. This work aims at implementing and training some of these methods for pixel-wise unmixing, comparing them to Linear Unmixing, and assessing their usability in typical PAD scans.

In order to train these models, synthetic data mimicking photoacoustic measurements with available ground truth tissue oxygenation had to be generated via photoacoustic simulation including optical and acoustic modeling on a variety of model-based generated tissue digital twins. Deep-learning models selected from the state of the art, Learned Spectral Decoloring (LSD) and Conditional Invertible Neural Networks (cINN) were then trained and their performance were compared to Linear Unmixing in silico. Addition of positional information to these models in different way via sinusoidal encoding was tested as well.

Results from this work confirm that Deep Learning methods are interesting to perform spectral unmixing as they outperform Linear Unmixing. Although simple methods (LSD) did not provide us with suitable predictions, more complex ones (cINN) gave interesting results, and the addition of positional encoding to them was proven beneficial for training.

Although important methodological challenges remain for cINNs to become a powerful tool for real time in vivo blood oxygenation assessment, their ability to perform spectral unmixing was confirmed in our use case, at least in silico.

Keywords: *photoacoustic imaging, peripheral artery disease, spectral unmixing, deep-learning*

Contents

List of Figures

List of Tables

1 Introduction

1.1 Context and motivation

Photoacoustic Imaging (PAI) is an emerging biomedical imaging modality in which a beam of optical nature (laser pulses) is sent to tissues that absorb its energy and generate an acoustic signal (ultrasounds). It enables non-invasive detection of both functional tissue properties (owing to the optical part), and structural information as obtained via traditional ultrasound (US) imaging (owing to the acoustic part) at high spatial resolution and centimeter-scale penetration depth in living tissues. In Photoacoustic Tomography (PAT), cross-sectional or volumetric images of tissues are reconstructed by collecting photoacoustic signals from multiple angles. state of the art PAT technologies allow 2D imaging in almost real time and 3D imaging in time scales of seconds to minutes [14]. Quantitative Photoacoustic Tomography (qPAT) aims at retrieving tissue properties thanks to the received signals. It is most of the time performed with multiple optical wavelengths: we call this modality Multispectral Optoacoustic Tomography (MSOT). The received signals are "unmixed" to retrieve most commonly chromophore concentrations in tissues. Here, we will restrict ourselves to the study of Oxyhemoglobin (HbO_2) and Deoxyhemoglobin (Hb) concentrations and derive blood oxygen saturation (sO_2) from these fields.

MSOT has already been translated into multiple clinical use cases [29]. In this study, we will focus on Peripheral Artery Disease (PAD), a common manifestation of atherosclerosis in the peripheral arteries touching approximately 200 million people across the world with a prevalence of more than 20% in people older than 65. Symptoms created by the insufficient blood supply in the lower limbs include intermittent claudication, in rest pain, and, in the most severe cases, tissue losses and skin ulcers. In the diagnostic of this disease, the challenge is its long asymptomatic phase, whereas in the follow-up, clinicians are still looking for a method to determine the muscular damage that it causes. qPAT will, in this case, help evaluating blood perfusion in the calf muscles in a non-invasive way.

1.2 Description of the problem

Whereas meaningful methodological approaches exist for the acoustic inverse problem (reconstructing the sample's pressure distribution from the time series measurements of the acoustic sensors), the optical inverse problem (retrieving concentrations of chromophores from this initial pressure) remains largely unsolved due to physical phenomena generally gathered under the term of "spectral corruption". Solving this problem is usually referred to as "spectral unmixing". The most commonly used method to perform it, so-called Linear Unmixing (LU) [18], is based on simplifying assumptions that contrast with the highly non-linear and ill-posed nature of the optical inverse problem. This makes the method easy to understand and applicable in real time, but not so reliable. Most computational methods from our knowledge [34, 35, 39] have tried to model physical phenomena with increasing complexities, which makes them more accurate but way slower (at the order of 1000 times slower) to compute. Some Deep-Learning (DL) methods [12, 11, 28] have recently been investigated as surrogates and proved to be as or even more accurate than numerical methods with a drastically reduced inference time.

1.3 Research questions and objectives

The general goal of this project was to better understand spectral unmixing in MSOT, and hopefully pave the way for better assessment of spectral coloring phenomena. Given the difficulty of the problem to be solved and the restricted amount of time allocated to solve it, a precise experimental strategy with realistic goals was put in place.

The majority of DL approaches for pixel-wise sO_2 estimation today are not meant for general purposes, meaning that they have to be trained for a specific scanning scenario to be applicable in vivo. To our knowledge, most models have been trained on datasets mimicking forearms or less realistic anatomical structures. This is one of the reasons why this work was first and foremost motivated by the application, namely PAD diagnosis and follow-up by estimating pixel-wise sO_2 in the muscle layers. DL methods from the state of the art were therefore trained on calf-mimicking scans so that they can eventually be applied in vivo. The general pipeline including simulation and training is later schematized in Fig.23. These methods were compared between each other and with LU.

The main objectives defined for this thesis were the following:

- setting-up a well documented photoacoustic simulation pipeline to mimic PAD use cases ;
- implementing, training and testing of DL models for spectral unmixing on these data to reach accuracies comparable to state of the art methods in silico and in vivo ;
- suggesting new structural ideas compared to the evaluated models to better assess spectral coloring.

1.4 Work plan

Part 1: Photoacoustic simulation

An in silico simulation pipeline using the Python framework SIMPA [13] was set up for simulating synthetic PA measurements for human calves with available ground truths with the following steps:

- building of a simple photoacoustic simulation pipeline with SIMPA ;
- adaptation of an in-house model-based acoustic reconstruction algorithm to this framework and addition to the pipeline ;
- definition of a meaningful use case by defining a realistic tissue digital twin and choosing a correct set of simulation parameters ;
- introduction of realistic variability in the simulation for optimal deep learning ;
- automatic generation of a synthetic dataset.

Part 2: Work on existing DL models

In this part, we implemented and trained DL models from the state of the art for pixel-wise sO_2 estimation:

- selection of meaningful models to be tested on our use case [12, 28] ;
- data preprocessing (ROI selection, ...) ;
- training of the models ;
- parameter finetunning.

Part 3: DL model development

This part went beyond the scope of existing literature, as we tried to incorporate new structural changes to the existing models, hoping to raise their awareness to spectral coloring:

- addition of spatial information to the chosen models in different forms including Positional Encoding (PE), an approach inspired by Transformer architectures [36, 8] ;
- training of these new approaches and comparison to our baselines.

Part 4: Validation

Finally, our approach had to be evaluated. Due to time constraints, this validation wasn't extensive:

- in silico evaluation of an in-house LU algorithm as well as our DL models and comparison ;
- same process on in vivo calf scans (this was unfortunately not done in time).

2 Background

2.1 Photoacoustic Imaging and its current challenges

2.1.1 The photoacoustic effect

In 1881, Alexander Graham Bell, in an experiment where he was supposed to investigate long-distance sound transmission, showed that materials exposed to sun light can produce acoustic signals that are dependent on the material type. He later showed that this was also working for UV and IR light, and thereafter invented an instrument, the "spectrophone" (see Fig.1), to identify materials thanks to their spectral signature [2].

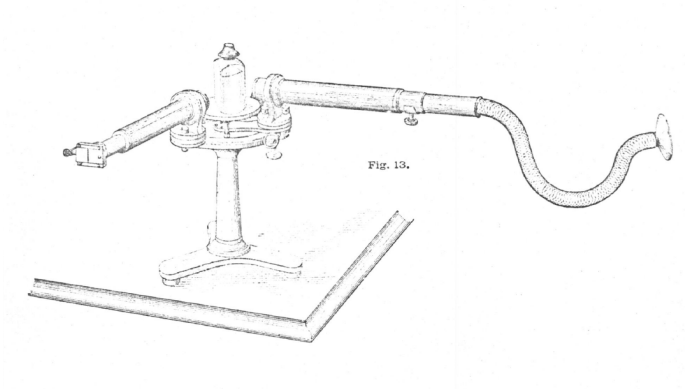


Figure 1. *The spectrophone as described by Bell [2]*

Later, he showed that this effect could also be applied to liquids and gases. Still, genuine applications of the photoacoustic effect had to wait for some development in more intense light sources and sensitive detectors in the mid 20th century to allow for more reliable measurements [37].

2.1.2 Photoacoustic Tomography

PAT is a biomedical imaging modality using this effect: non-ionizing laser pulses (typically in the nanosecond range) in the visible or near-infrared (NIR) parts of the optical spectrum are sent to biological tissues where chromophores (molecules in the tissues) absorb a part of their energy and convert it into heat. Chromophores that are naturally present in the tissues are referred to as endogenous, as opposed to exogenous ones which are artificially added to tissues (injected or ingested by human beings) and play the role of contrast agents to help locating specific structures or molecules. The main endogenous absorbers are oxy- and deoxy-hemoglobin (HbO_2 and Hb), lipids, melanin, collagen, and water [30]. The transient thermoelastic expansion of tissues subsequently generates pressure waves in all directions that can be detected by broadband ultrasonic (MHz) transducers. The physical process of PA signal generation is imaged in Fig.2.

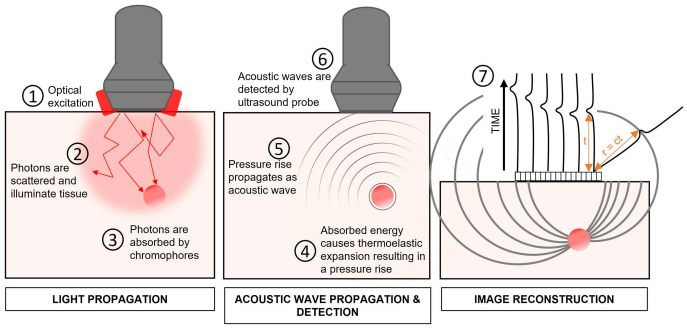


Figure 2. *Schematics representing photoacoustic signal generation [30]*

2.1.3 Quantitative Photoacoustic Tomography

In qPAT, the main goal is to determine the composition of tissues, i.e. chromophore concentrations. In the case of sO_2 estimation, we focus on Hb and HbO_2, and, at a specific point in space, we define blood oxygen saturation as:

$$sO_2 = \frac{C_{HbO_2}}{C_{HbO_2} + C_{Hb}}$$

where C_k refers to the molar concentration of molecule k at this specific point.

In order to be able to estimate sO_2, as well as to increase the information that we can get from the tissue, we usually have to perform pressure measurements at different wavelengths between 650 nm and 950 nm (see 2.2). This window is a trade-off between the reasonably low absorption of light by Hb and HbO_2, which allows a reasonable penetration depth, and their relatively high absorption compared to water and lipids (see their spectroscopy spectrum in Fig.3) which allows to approximate that they are the only absorbers in layers that are not the epidermis (where melanin is present) [14, 31].

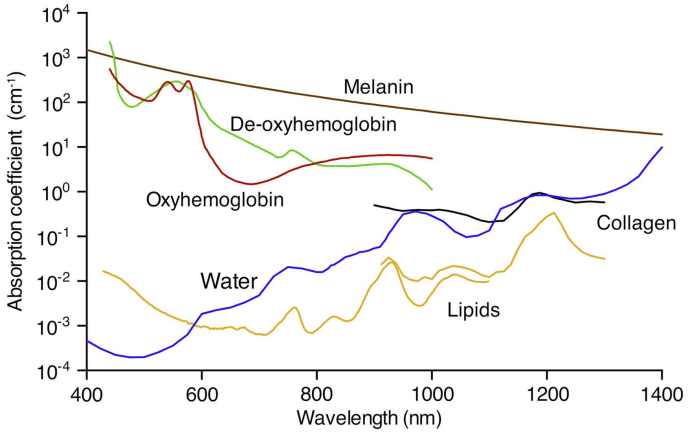


Figure 3. *Absorption spectra of the most commonly studied chromophores [14]*

PAI can therefore be seen as a hybrid modality between US and optical imaging. Compared to traditional optical imaging, PAI allows to break the traditional "optical penetration depth barrier", allowing for imaging living tissues with rich optical contrast and high ultrasonic spatial resolution. The main added value to the

already widespread US imaging (USI) techniques is the possibility to add precise molecular (chromophore concentration) and functional (sO_2 distribution) information to the morphological information (pressure measurement) that we get with regular USI, and that, without the use of exogenous chromophores [29, 30, 32].

2.1.4 Formal description of the problem

In this part, we will study the formal derivations that will be necessary to understand the problem at stake for this particular project. They are represented graphically in Fig.4. For more complete calculations and rigorous description of the assumptions made, see [5].

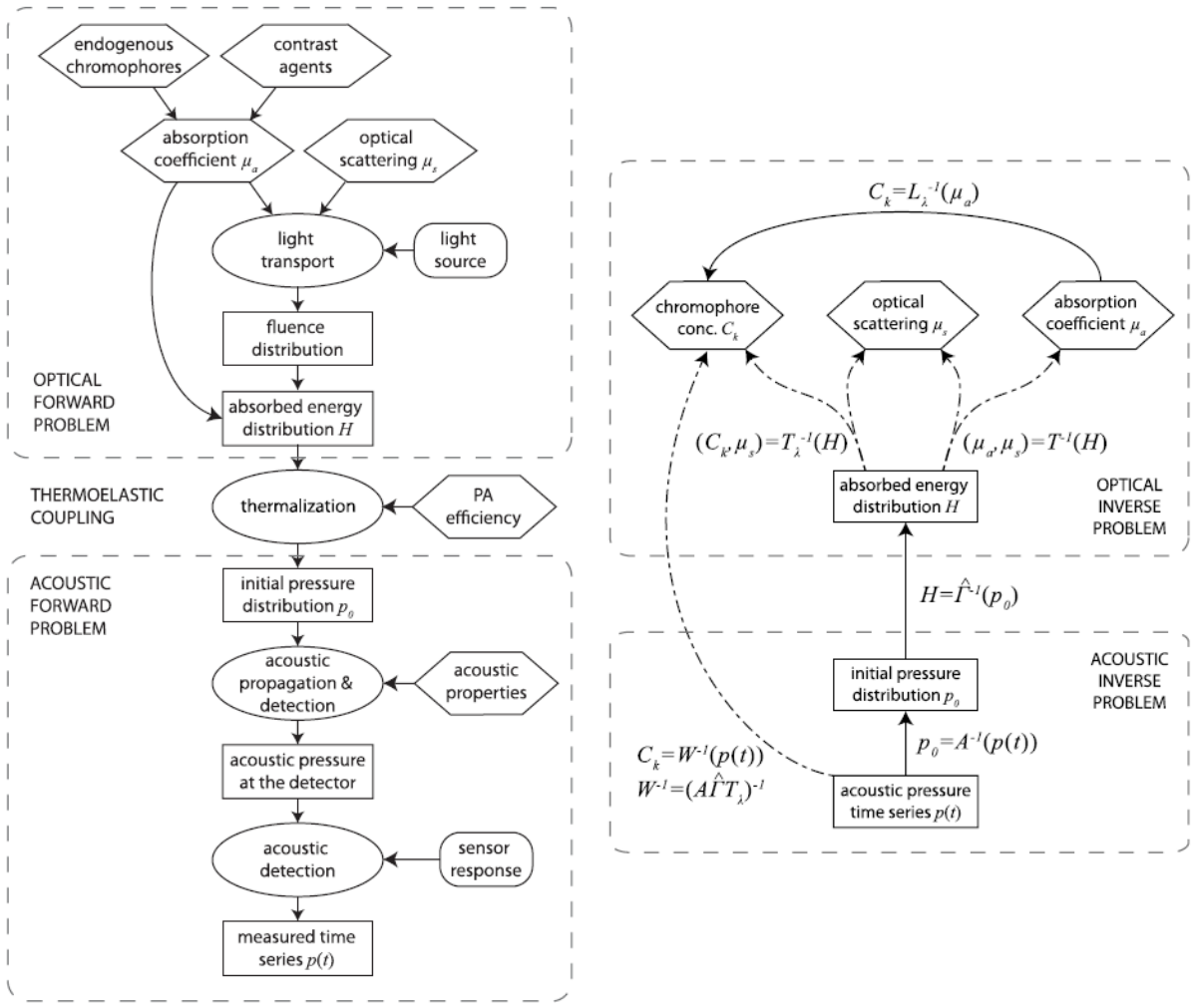


Figure 4. *Forward (left) and inverse (right) problems in optoacoustics [5]*

Optical forward problem and thermoelastic coupling

The photoacoustic effect is composed of three steps:

1. *absorption* of a photon (femtosecond time scale) at a specific wavelength λ

2. *thermalization* of the absorbed energy, creating a local pressure increase (in PA, we always consider that thermalization, i.e. nonradiative decay, prevails over radiative decay)

3. *propagation* of this pressure perturbation due to the elastic nature of the tissue (vibrational relaxation, subnanosecond timescale)

We denote the absorbed energy density (heat per unit volume) as H. The thermalization induces a small local rise of temperature and pressure. This generated pressure is referred to as the initial acoustic pressure p_0:

$$p_0(\vec{x}, \lambda) = \hat{\Gamma}(\vec{x})H(\vec{x}, \lambda)$$

where $\hat{\Gamma}$ is the PA efficiency. In our simplified case, we will consider it to be always equal to the Grüneisen parameter, a thermodynamic parameter dependent on the temperature and the medium. This parameter will always be considered constant (at normal body temperature) and uniform in space and noted $\Gamma = 0.2$.

We also know that the energy density is related to the optical absorption coefficient μ_a and the light fluence ϕ by:

$$H(\vec{x}, \lambda) = \mu_a(\vec{x}, \lambda)\phi(\vec{x}, \lambda; \mu_a, \mu_s)$$

It is highly important to notice that the fluence also itself depends on the absorption and scattering processes of light in the tissue. The scattering is materialized by the scattering coefficient μ_s. In most cases, its influence is solved easily, as we consider it isotropic, contrary to the absorption coefficient.

The initial acoustic pressure is therefore directly related to μ_a and ϕ:

$$p_0(\vec{x}, \lambda) = \Gamma\mu_a(\vec{x}, \lambda)\phi(\vec{x}, \lambda; \mu_a, \mu_s)$$

Acoustic forward problem

We consider that there exists an operator to link $p(t)$ and the initial pressure mapping over space p_0. We refer to it as A:

$$p(t, \vec{x}, \lambda) = Ap_0(\vec{x}, \lambda)$$

It should be noted by the reader that we never actually reconstruct a spatial map $p(\vec{x})$ per se (and therefore never use this operator A): the value that we read on each receptor of the transducer is the sum of the amplitudes of the sound waves coming from all the points in space that were able to reach this specific receptor (one signal per wavelength). The received field is called a *sinogram*, and it is of size $N_{receptors} \times N_{time\ steps}$.

The process of sinogram formation is illustrated in 4.1.5. A signal originated from a point S in space located at (x, y) relatively to a planar transducer and is propagating with a speed of sound c. The left image is in the spatial domain, and the right one in the time domain. The receptor element j will receive the wave originating from S after the delay $\tau(x, y, j)$, so that the curved line that we see on the right is one of the multiple lines that constitute a sinogram.

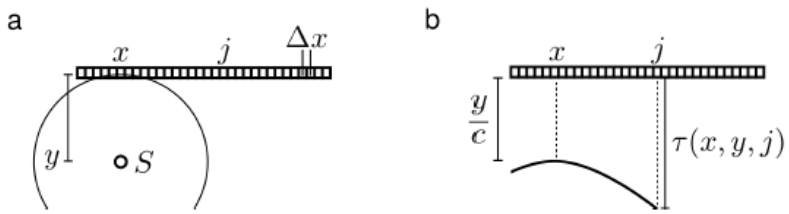


Figure 5. *Illustration of sinogram formation [22]*

Acoustic inverse problem

Acoustic inversion, i.e. retrieving $p_0(\vec{x})$ from the sinogram signal for each wavelength is a better resolved problem than optical inversion. A few classical algorithms perform this inversion well (initially in USI, but then applied to PAI): back-projection, time-reversal, Delay-and-Sum (DAS) [10, 20], Delay-Multiply-and-Sum (DMAS) [26], signed-Delay-Multiply-and-Sum (sDMAS) [22] and finally Model-Based reconstruction (MB rec), which is the one that we chose to use in our simulation (see 4.1.5). Each algorithm has its peculiarities and use cases.

Optical inverse problem

As seen previously, with the assumptions that we make, we can consider the energy deposition $H(\vec{x}, \lambda) = \Gamma^{-1} p_0(\vec{x}, \lambda)$ as the measured data. We also assume that the optical absorption coefficient is linearly linked to the concentrations of chromophores in the tissues, which are the physical quantities that we want to retrieve at the end:

$$\mu_a(\vec{x}, \lambda) = \sum_{k=1}^{n_{chromophores}} C_k(\vec{x}) \alpha_k(\lambda)$$

2.2 Spectral unmixing

2.2.1 Spectral corruption

Using the same notations as in 2.1.4, if we could find a spatial mapping of the absorption coefficient $\mu_a(\vec{x})$ for a given wavelength, we would easily get a unique solution for the wanted constant concentrations $C_k(\vec{x})$ by linear inversion at each point x provided that we scan with enough different wavelengths (see 2.2.4).

Let's therefore consider that we remain at a specific position in space $\vec{x}$. Then we can access an absorbed energy density:

$$H_{\vec{x}}(\lambda) = \mu_{a,\vec{x}}(\lambda) \phi_{\vec{x}}(\lambda; \mu_a, \mu_s)$$

and we are looking for the absorption coefficient $\mu_{a,\vec{x}}(\lambda)$. The biggest issue is that the fluence term cannot be considered as constant. In fact, it highly and non-linearly depends on the absorption and scattering of light that has occurred during the whole travel of light throughout the tissues. In other words, at position $\vec{x}$, $\phi_{\vec{x}}(\lambda; \mu_a, \mu_s)$ does not only depend on the absorption and scattering at position $\vec{x}$, but also on the distribution

of these fields some distance away, which makes a rigorous inversion impossible from a numerical point of view. This phenomenon is known as *spectral coloring* [4].

If we now fix the wavelength and forget the influence of optical absorption and scattering, because the fluence is also a function of position, it will *distort the image structure* [4]:

$$H_\lambda(\vec{x}) = \mu_{a,\lambda}(\vec{x})\phi_\lambda(\vec{x})$$

and $H_\lambda(\vec{x})$ won't be proportional to $\mu_{a,\lambda}(\vec{x})$, which we would have wished in a very simple case.

2.2.2 What do we call spectral unmixing ?

Spectral unmixing is the fact of retrieving properties such as chromophore concentrations (in our case, C_{Hb} and C_{HbO_2}) or oxygenation saturation levels sO_2 from the pressure measurements p_0 or the sinograms $p(t)$ that we obtain in optoacoustic measurements.

This task is made very complicated due to the complex interaction of fluence with tissues as just explained. Different models for spectral unmixing exist [5], but they vary in:

- *accuracy*: how accurately they model the physical phenomena
- *range of validity*: over what range of parameters or variables of interest (C_k or μ_a) they are valid
- *complexity*: how easy they are to solve
- *invertibility*: how easily they can be inverted to compute the variable(s) of interest from the measured data

The ideal model is mostly a balance between those four characteristics.

Some models focus on approximating the physical phenomena (see 2.2.3), and some are rather (or fully) based on mathematical or algorithmic techniques (see 3). This work will use a Monte Carlo model, which can be seen as fluence modelisation, for the simulation of synthetic images, but will focus on algorithmic techniques for the resolution of the optical inverse problem.

2.2.3 Fluence modelisation

Even though light is an electromagnetic wave satisfying the fundamental Maxwell equations, the easiest way to model light propagation in turbid (i.e. highly scattering) media is particle-based methods. The most common modelisation is using the Radiative Transfer Equation (RTE, i.e. Boltzmann's transport equation with low energy, monochromatic photons):

$$\frac{1}{c_{light}}\frac{\partial\phi}{\partial t}(\vec{x},\hat{s},t) = \underbrace{q(\vec{x},\hat{s},t)}_{(1)} - \left[\underbrace{\hat{s}.\nabla}_{(2)} + \underbrace{\mu_a(\vec{x})}_{(3)} + \underbrace{\mu_s(\vec{x})}_{(4)}\right]\phi(\vec{x},\hat{s},t) + \underbrace{\mu_s\int\Theta(\hat{s},\hat{s}')\phi(\vec{x},\hat{s}',t)ds'}_{(5)}$$

with:

- $\phi(\vec{x}, \hat{s}, t)$ the radiance at position $\vec{x}$, time t and in direction $\hat{s}$, which is directly related to the intensity of light and therefore the local number of photons
- $\Theta(\hat{s}, \hat{s}')$ the scattering phase function, which represents the probability that a photon originally traveling in direction $\hat{s}$ ends up in direction $\hat{s}'$ if scattered
- $q(\vec{x}, \hat{s}, t)$ a source of photon
- c_{light} the speed of light in the medium

and that stands for the fact that the rate of change of the radiance is due to: (1) sources, (2) the net outflow of photons due to the radiance gradient, (3) photons absorbed, (4) photons scattered into another direction, and (5) photons scattered into direction $\hat{s}$ from all the others directions $\hat{s}'$.

We will only be interested in solving the stationary form of the RTE because acoustic propagation occurs on a much higher timescale than heat deposition:

$$[\hat{s}.\nabla + \mu_t]\, \phi(\vec{x}, \hat{s}) - \mu_s \int \Theta(\hat{s}, \hat{s}')\phi(\vec{x}, \hat{s}')d\hat{s}' = q(\vec{x}, \hat{s})$$

where $\mu_t = \mu_a + \mu_s$ and the wanted fluence is the integral of the radiance over all angles $\hat{s}'$:

$$\Phi(\vec{x}) = \int \phi(\vec{x}, \hat{s}')ds'$$

Equation 2.2.3 can be written in weak form and solved on a discrete mesh by finite element methods (→ but no invertibility).

Case of a collimated source

In the very specific case of a collimated source propagating in the z direction without scattering, we get a fluence which is only a function of depth:

$$\Phi(z) = \Phi_0 \exp(-\mu_a z)$$

where Φ_0 is the fluence at the source and μ_a is constant. It is an equation known as Beer´s law (→ invertibility and simplicity, but unfaithful assumptions).

Diffusion Approximation

In so-called Diffusion Approximation (DA), we get the following expression for the fluence after harmonic decomposition of the equation and keeping only orders 0 and 1, and further calculation:

$$\Phi(z) \approx k\Phi_0 \exp(-\mu_{eff} z)$$

where:

- $\mu_{eff} = \sqrt{3\mu_a(\mu_a + \mu'_s)}$ is the effective attenuation coefficient
- $\mu'_s = (1-g)\mu_s$ is the reduced scattering coefficient
- g is the anisotropy coefficient

and where the following assumptions are made:

a) $\mu'_s \gg \mu_a$ (scattered fluence is almost isotropic)

b) $z \gg \frac{1}{\mu_a+\mu'_s}$ (fluence everywhere is diffuse)

In typical PA applications, we have the following orders of magnitude: $\mu_a = 0.1$ mm^{-1}, $\mu_s = 10$ mm^{-1}, $g = 0.9$ and $\mu'_s = 1$ mm^{-1}, so a) is fulfilled but b) isn't for distances inferior to approximately 1 mm, which is a zone that we want to study in qPAT and that needs further hypothesis. It must also be noted that these equations are valid only if absorption and scattering coefficients do not vary on the light path, which is never true (→ invertibility and simplicity, but still unfaithful assumptions).

Monte Carlo methods

The gold standard today for simulating fluence propagation numerically is an ensemble of stochastic approaches known as Monte Carlo models, which simulate the random walk of photons distributed by packets of energy in the tissues and the optical phenomena that happen to them: absorption, scattering, transmission and reflection. These interactions are modeled by drawing random variables from underlying probability distributions, which are then used to calculate photon step sizes and directions. By repeating this process a large number of times, the RTE can be approximated numerically. Fig.6 gives an illustration of a simple photon random walk modelisation in tissues, where D_i and WT_i respectively model the probability of interaction and the probabilistic weight of a photon at i^{th} interaction.

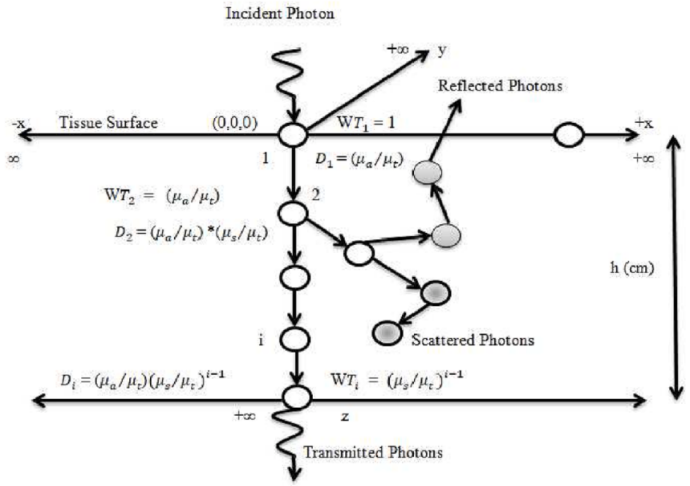


Figure 6. *Illustration of the random walk of photons in tissues modeled by Monte Carlo [3]*

These methods are easily parallelizable because the packets are considered independent. As GPU power is constantly increasing, they are becoming increasingly faster and efficient. What is more, they are *accurate*, *valid* in our use case and *not so complex* because we can compute them. However, they are not *invertible* because stochastic... We will therefore only use them as a good approximation for the forward problem in our implementation, but we are still looking for an efficient unmixing method.

2.2.4 Linear Unmixing

Linear Unmixing (LU) is the simplest way to approximate the optical inverse problem, where we make the assumption that the received optoacoustic signal p_0 for each wavelength and at each point in space is directly proportional to the absorption coefficient μ_a, which corresponds to considering the fluence term as constant in Eq.2.1.4.1:

$$p_0(\vec{x}, \lambda) \propto \sum_{k=1}^{K} C_k(\vec{x}) \alpha_k(\lambda)$$

In our case, this gives:

$$p_0(\vec{x}, \lambda) \propto C_{Hb0_2}(\vec{x}) \times \alpha_{HbO2}(\lambda) + C_{Hb}(\vec{x}) \times \alpha_{Hb}(\lambda)$$

and we use the so-called Linear Mixture Models (LMMs) [18] to retrieve the wanted "concentrations" from the pressure measurement. Namely, we need $n \geq 2$ wavelengths to solve, for each pixel:

$$\underbrace{\begin{bmatrix} p_0(\lambda_1) \\ p_0(\lambda_2) \\ \vdots \\ p_0(\lambda_n) \end{bmatrix}}_{p_m} = \underbrace{\begin{bmatrix} \alpha_{HbO_2}(\lambda_1) & \alpha_{HbO_2}(\lambda_2) & \dots & \alpha_{HbO_2}(\lambda_n) \\ \alpha_{Hb}(\lambda_1) & \alpha_{Hb}(\lambda_2) & \dots & \alpha_{Hb}(\lambda_n) \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} \widetilde{C_{HbO_2}} \\ \widetilde{C_{Hb}} \end{bmatrix}}_{c}$$

The values that we get, noted here $\widetilde{C_k}$, are equal to these concentrations within one factor, but the final goal is to compare them, so the proportionality coefficient does not matter in this case.

This unmixing method was mostly proven successful in applications like remote sensing [18], optical microscopy [40], or spectroscopic PAT in small animals [24, 9] where we work with 2D samples, or samples with a very small depth and therefore do not have to account for the effect of depth on fluence. Because our in vivo use case is subjected to spectral corruption (see 2.2.1), LU cannot be considered reliable here [34].

3 State of science and technology

According to the literature, the subject of finding alternatives to LU has been extensively studied throughout the past 20 years. Alternatives of different complexity, range of validity and accuracy therefore exist today. Methods using *linearization*, *direct inversion*, *fixed-point iteration* and *model-based minimization* are discussed as part of a review in [5]. We will not discuss them here. In this part, we will focus only on some specific methods that go in the same research directions as ours.

3.1 Traditional numerical methods

3.1.1 eMSOT

A method known as eMSOT [35, 34] and that does not fit in the previously quoted categories, will be very shortly summed up here. It uses the decomposition of the fluence in 4 base eigenspectra $\Phi_M(\lambda)$, $\Phi_1(\lambda)$, $\Phi_2(\lambda)$ and $\Phi_3(\lambda)$ by Principal Component Analysis (PCA) on a large number of fluence spectra simulated using the 1D Diffusion Equation (Eq.2.2.3.2) at different depths (from 0 to 1 cm) and values of sO_2 (from 0 to 100%) with fixed physiological optical tissue properties. The problem is therefore seen as:

$$p_0(\vec{x},\lambda) = \Phi'(\vec{x},\lambda)(C'_{Hb0_2}(\vec{x}) \times \alpha_{HbO2}(\lambda) + C'_{Hb}(\vec{x}) \times \alpha_{Hb}(\lambda))$$

with:

$$\Phi'(\vec{x},\lambda) = \Phi_M(\lambda) + m_1(\vec{x})\Phi_1(\lambda) + m_2(\vec{x})\Phi_2(\lambda) + m_3(\vec{x})\Phi_3(\lambda)$$

where $\Phi_M(\lambda)$ is the mean fluence spectrum, and $\Phi_1(\lambda)$, $\Phi_2(\lambda)$, $\Phi_3(\lambda)$ are the three eigenspectra determined as previously explained. The notations $\Phi'(\vec{x},\lambda)$ and $C'_k(\vec{x})$ mean that the fluences and concentrations have been rescaled. The problem therefore becomes a non-linear inversion problem with five unknowns: $C'_{Hb0_2}(\vec{x})$, $C'_{Hb}(\vec{x})$, $m_1(\vec{x})$, $m_2(\vec{x})$ and $m_3(\vec{x})$ and is solved pixel-wise as a constrained optimization problem.

Strengths and weaknesses

This method has proved to be better than LU in vivo, but it is important to highlight that it wouldn't necessarily generalize well to our use case, owing to the very different experimental set up (see the one used for this project in 4.1.1). They use a ring array (MSOT inVision) with a 270° angular coverage and 40.5 mm radius that can only image very small objects (mice of small tissue-like phantoms) and is therefore mostly pre-clinical: these types of arrays are never used for imaging human tissues in vivo. The fluence corruption with depth is much more important in use cases involving a linear or curved array, typical in the MSOT Acuity device (→ limited *range of validity*).

3.1.2 Non segmentation-based iterative method

A family of methods, that we partly mentioned previously are the so-called iterative methods. Among these, the non segmentation-based iterative method suggested by Zhang & al. in 2022 [39] for pixel-wise

reconstruction of μ_a is part of state of the art today. The method build in this article is compared to two previously used families of methods using the measured pressure distribution, p_0 or H, and the light fluence Φ approximated by the transport equations mentioned in 2.2.3 to estimate the μ_a distribution by simply using Eq.2.1.4.1. These two methods require the segmentation of a Region of Interest (ROI) where the μ_a is considered constant:

- Segmentation-Based Direct Correction (SBDC): the μ_a in the segmented ROI is set as an input by the user to an ideal value. The pressure map is then divided by the estimated fluence map to obtain the estimated value for the uniform μ_a.
- Segmentation-Based Iterative Correction (SBIC): the μ_a in the segmented ROI is initially set to 0, and new values are calculated iteratively by minimizing the error between the measured pressure distribution and the estimated one using the current value of μ_a.

The suggested method here points out that tissues actually have non-uniform μ_a distributions and therefore decides not to use any segmentation. The comparison of the three methods is presented in Fig.7.

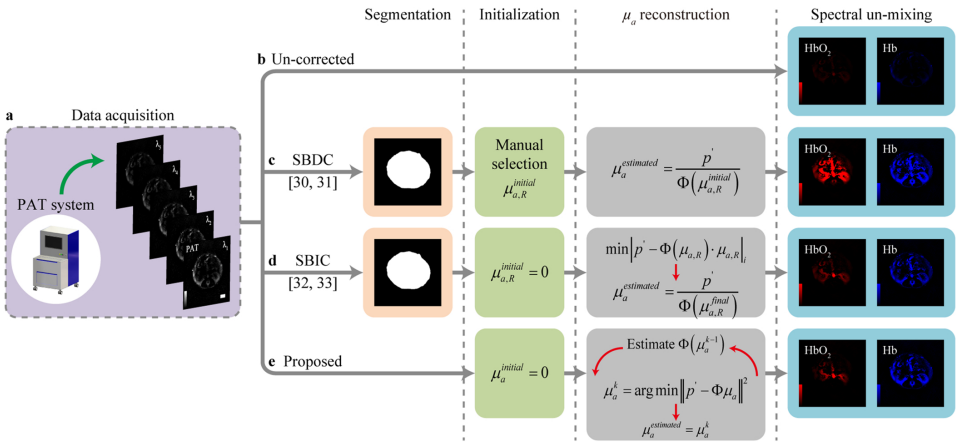


Figure 7. *Comparison of SBDC, SBIC and the non-segmentation iterative method [39]*

The algorithm is iterative, and each iteration is based on two main steps: first, the Φ map is computed by finite elements as a function of the current μ_a map (initially set to a map of zeros) using the DE (Eq.2.2.3.2) and an assumed constant reduced scattering μ_s'; then a new μ_a map is computed via gradient descent. This process is repeated until convergence. The formal algorithm is synthesized in Fig.8. The chromophore concentration maps are subsequently computed from the μ_a map using traditional LU.

Input: $\mu_a^{(0)} = 0$, μ_s', k, *Iter* max, ε.

Repeat

1. Update Φ^k using the diffusion equation (DE) with μ_a^{k-1}, μ_s'.
2. Update μ_a^k by solving $\mu_a^k = \arg\min \left\| p' - \Phi^k \mu_a \right\|^2$ with the constraint that $\mu_a \geq 0$.
3. Calculate the error: $err \leftarrow norm(p' - \Phi^k \mu_a^k)$.
4. Update the iterations: $k \leftarrow k + 1$.

Until convergence where $k > Iter$ max or $err < \varepsilon$.

Output: μ_a^k.

Figure 8. *Algorithm used in the non-segmentation iterative method [39]*

The method was validated in a comprehensive manner via simulation, animal (with mice) and phantom experiments, always showing much better performance than SBDC and SBIC.

Strengths

The main advantages of the method are its accuracy, in the demonstrated cases, due to its pixel-wise approach, the efficiency and the physical sense introduced in the two-step gradient descent algorithm, and the ease of use.

Weaknesses

However, the article claims a few weaknesses that we will be able to overcome in our method:

- The scattering μ_s is fixed to a constant value, which is never the case in practice.
  → Our DL algorithm will be trained on synthetic images obtained via realistic simulation where μ_s varies as a function of anatomical structures.
- Fluence simulation is only carried out in 2D using the DE, which uses strong assumptions.
  → Our simulation used stochastic 3D Monte Carlo simulation, which is much more accurate.
- The computation time is of the order of 90s (longer than SBDC and SBIC), which makes this algorithm impossible to use in real time.
  → One of the main advantages of DL algorithms is their fast inference time once trained.
- Last but not least, although this method could seem appealing, it is important to highlight that we get the same problem as with eMSOT when it comes to comparison, since only pre clinical application was targeted here, and therefore MSOT inVision was used (→ limited *range of validity*).

3.2 Deep Learning methods

Finally, other state of the art methods include DL models. The ones that we will study here have multiple advantages over the aforementioned methods but, just like any other approach, also come with challenges:

Advantages

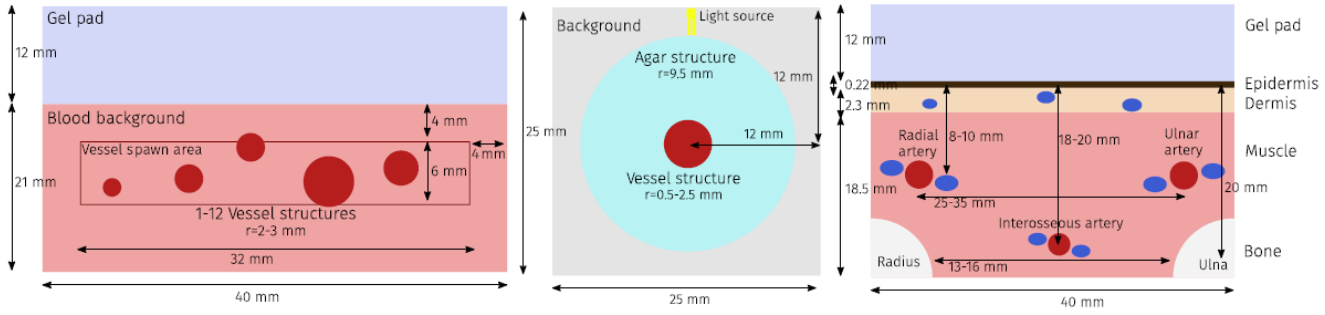
- They directly allow for performing the whole optical inversion, i.e. estimating sO_2 pixel-wise from the measured pressure spectra for all the acquisition wavelengths, bypassing the need for explicit fluence estimation or correction. They therefore allow for a more complex modelisation, which can help capturing part of the non-linearities that generate spectral coloring.
- Although they need to be trained extensively, their inference time is usually several orders of magnitude shorter than other methods, e.g. about 200 ms on a CPU and 2 ms on a GPU for LSD (see 3.2.1) versus about 90 s for the non segmentation-based iterative method (as previously seen in 3.1.2). This allows an application in real time imaging where LU was the only available method because of its simplicity.
- DL methods already applied to problems similar to ours have also simply been more accurate than LU as well as most numerical methods.

Challenges

- Structurally, they are big black boxes that are very hard to understand from the developer's point of view. Although some knowledge begins to exist on what type of architecture works for what kind or problem, the development process tends to be very exploratory and based on chance.
- They need a lot of data to train on. In our case in particular, a complete pipeline, described in 4.1 has to be put in place to generate the training data, which is time-consuming and requires a lot of modeling effort.
- One major concern about these DL methods is the so-called "*domain gap*" that exists between the simulated and real PA data. Although some models can perform well on synthetic data, some may fail in generalizing to experimental cases (i.e. in "bridging this gap").

3.2.1 Learned Spectral Decoloring (LSD)

LSD, introduced in [12], was one of the first DL methods to overcome efficiently this *domain gap*. The synthetic datasets used in this article for training are pictured in Fig.9. They were all simulated with 3D MCX simulations with a similar set up as ours (see 4.1), and each of them corresponds to a clinical or pre-clinical use case of the model: generic data for in vivo application in open surgeries, flow phantom data for in gello experiments, and forearm data for any forearm in vivo scanning.



(a) Generic data set. **(b)** Flow phantom data set. **(c)** Forearm data set.

Figure 9. *Training datasets for LSD [12]*

The DL architecture, depicted in Fig.10, is a Fully-Connected Neural Network (FCNN) that depends on the use case. The author justifies not using the very popular method of convolutions because we do not work at the image scale. The size of the input layer is the number of wavelengths used in the specific use case and the output is the predicted sO_2 value for the studied pixel. Leaky ReLU activations are used, as well as dropout to prevent overfitting.

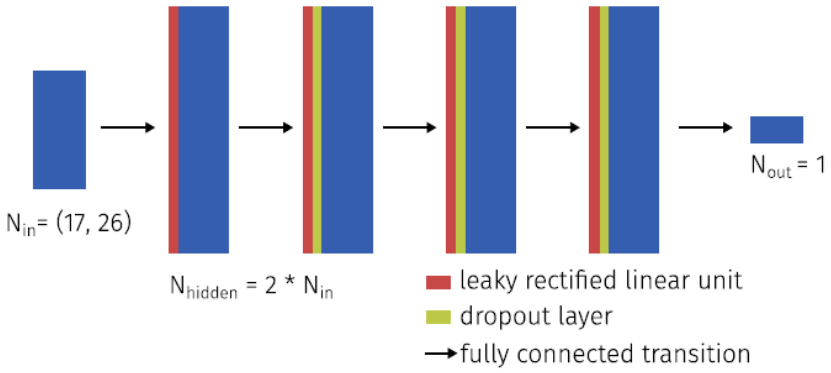


Figure 10. *Schematic of the LSD model architecture [12]*

Strengths

The obtained results show positive outcomes:

- in silico: the median relative sO_2 estimation error ranges from 6 to 15% on the different datasets, with an absolute error below 10% for all the use cases. The author concludes a general feasibility of the study.
- in gello: LSD is able to account for a higher dynamic sO_2 range than LU in the case of a 100% to 0% sweep. Moreover, LSD allows to reduce the rim-core effect (significant difference in the sO_2 estimates at the rim and the core of the vessel).

- in vivo: LSD also provides more plausible sO_2 estimates with higher dynamic range and reduced rim-core effect.

LSD therefore goes one step further in better accounting for spectral corruption for real time sO_2 estimation.

Weaknesses

However, (1) these models are dependent on the application (2) and the number of wavelengths used (→ lack of *flexibility*); (3) the training images are theoretical pressures (output of MCX), so acoustic processes are bypassed, which might add some artifacts to the obtained image (→ incomplete image simulation); (4) they are still subjected to spatial corruption, which increases when the structures become more and more complex (→ *ambiguity* remains).

In this work, we mostly addressed problems (2) and (3). The perspective of having a general purpose model (i.e. solving (1)) still seems a bit far away, and spectral corruption (4) of course remains the biggest problem in any case.

3.2.2 Distribution-informed and wavelength-flexible model - LSTM

A very recent approach [11] goes in the same research direction as the LSD approach but claims much more flexibility:

- *wavelength flexibility*: the model used is a custom network where the backbone is a Long Short Term Memory (LSTM) (see B in Fig.11), a recurrent network that allows for processing sparse data. The input size, 41, is the maximum number of wavelengths with which measurements can be performed. The user can however provide data with missing spectra. The mask layer will inform the backbone about these missing data, and it will ignore them. As often in state of the art models, a Fully-Connected Neural Network (FCNN) is added after the backbone. Here, Leaky ReLU activations are used for every layer except for the last one, which uses sigmoid to provide an sO_2 prediction between 0 and 1.

- *use-case flexibility*: the model is trained on an extensive amount of different synthetic datasets (see A in Fig.11) comprising either initial pressure images or sinograms obtained with a set up similar as the one used in LSD as well as our experiment (see 4.1). The basic dataset is composed of simple pressure images acquired with a linear transducer array and containing a background muscle tissue with vessels varying in their number, diameter and sO_2, but a large amount of alternative datasets are defined (varying the background sO_2, the vessel radii, the illumination geometry, the MCX resolution, adding a skin layer, performing acoustic reconstruction, using MSOT device digital twins).

The method is trained synthetically and validated in gello with a blood flow phantom (similar method to 3.2.1, but this model is not trained on phantom-like samples contrary to LSD) and in vivo on human forearm and mice as illustrated in Fig.11.

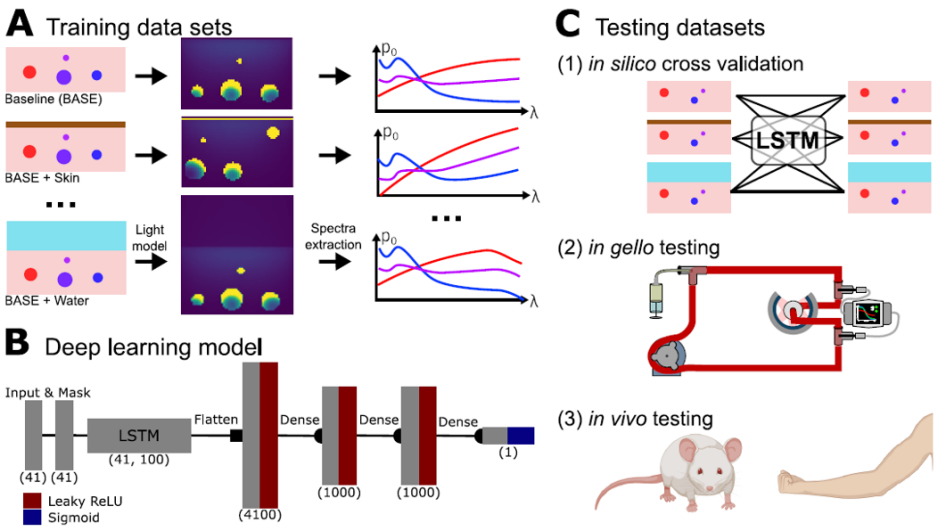


Figure 11. *Visual summary of the approach used in [11]*

Strengths

Noteworthy strengths of this study, mostly due to the high flexibility of the model, are the following:

- They are able to estimate an ideal number of wavelengths to simulate the dataset with (10), as a good trade-off between model performance and reasonable calculation, and to show that it is more convenient to test the model with the same set of wavelength than the one it was trained on.
  → In our project, the number of wavelengths used was fixed to 6 to stick to the experimental set up of PAD scanning with the Acuity device.
- An in silico cross-validation allows them to perform a sensitivity analysis regarding different parameters of the dataset and draw some useful conclusions: variation of background sO_2 has limited effect on the simulation results, finer resolution gives better results, illumination is better with a Gaussian than a point source, inclusion of all the chromophores (including melanin) is important, acoustic modeling is rather detrimental, training on a diverse dataset is better.
  → The penultimate point will be discussed later in this thesis, as the absence of acoustic modeling was rather seen as a weakness in the LSD approach 3.2.1, and we also tend to consider it as such. Apart from this, all the other conclusions were taken into account during this project.
- They introduce a mathematical method based on the calculation of a *Jensen-Shannon divergence* D_{JS} between the training dataset and the testing, or "target" dataset. We denote P and Q the respective probability distributions of predicted sO_2 respectively for the training and the target dataset. The JS divergence is computed as:

$$D_{JS}(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M)$$

where $M = \frac{1}{2}(P+Q)$ and D_{KL} (Kullback-Leiber divergence) is a measure of the relative entropy between two datasets:

$$D_{KL}(P\|Q) = \sum_{x\in X} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

In fact, the divergence has to be computed wavelength-wise and then averaged, so the measure that is used is:

$$\overline{D_{JS}(P\|Q)} = \frac{1}{N_\lambda}\sum_{\lambda\in\Lambda} D_{JS}(P_\lambda\|Q_\lambda)$$

They show that this variable correlates well with the absolute error on predicted sO_2 ϵ_{sO_2} when training on the training dataset and testing on the target one, and can therefore be used to identify the best training dataset for a targeted use case.
→ This would have been a very powerful tool to use in this work, but we unfortunately restricted ourselves to one precise example as a matter of time.

- The suggested model outperforms LSD (and of course LU) in silico and in vivo, and the combined method using the JS divergence as an identifier for the best training dataset for a specific use-case proves applicable in most cases.

Weaknesses

A few limitations and perspectives are however mentioned:

- Using D_{JS} to select a training dataset is not a perfectly robust method, since it remains a summary measure. It is therefore to be used only combined with clinical knowledge.
- Training with combined datasets was shown to be beneficial in silico, but this finding did not generalize in vivo, which is interpreted as overfitting because all the synthetic datasets are seen as subsets of the combined one. It therefore means that the hope for a reference general model that could multi-task is not yet fulfilled.
- The method is, of course, still subjected to *spectral coloring* effects.

3.2.3 Uncertainty-aware Deep Learning methods - cINN

Concept

The solution that we found the most promising to explore in order to tackle *spectral coloring* is the uncertainty-aware oxygenation quantification introduced in [28] using a conditional invertible neural network (cINN), which is, to our knowledge, the first and only to-date DL method to introduce probabilistic considerations directly in the way sO_2 is estimated.

Formally, instead of trying to determine a deterministic mapping $y \mapsto sO_2$, the model provides the posterior $p(sO_2|y)$ as a probability distribution of sO_2 values conditioned on the PA measurement y for each pixel.

This probabilistic representation accounts for the fact that very different tissue configurations can lead to similar PA measurements due to the ill-posedness of the optical inverse problem.

As illustrated in Fig.12, the model can provide different types of outputs. A pixel that is "not ambiguous" would most probably provide a unimodal distribution (case A), and the center of the mode would be close to the ground truth sO_2 value as well as the value predicted by other traditional DL models. However, a pixel that is "ambiguous", would be likely to provide a distribution with multiple modes with, hopefully, one of them centered in the ground truth value (case B). In the latter case, the author suggests that traditional DL models, because they are forced to choose only one value, tend to output a mean of multiple possible values (i.e. multiple modes of the posterior here), or the value of the highest mode.

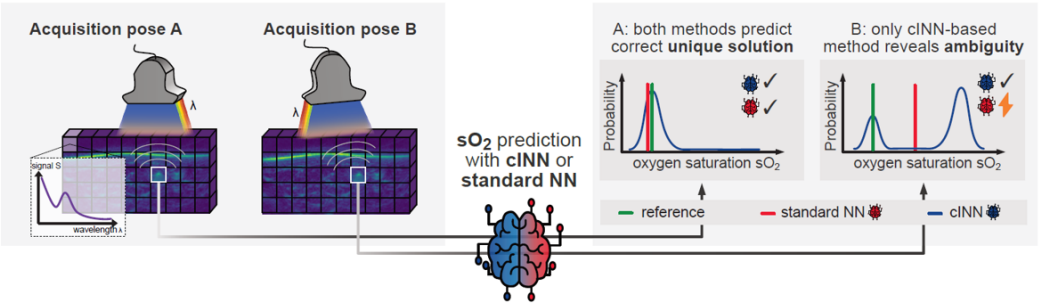


Figure 12. *Illustration of the uncertainty-aware sO_2 estimation principle [28]*

Invertible Neural Networks (INNs)

cINNs are variations of INNs, whose suitability in addressing intractable and ill-posed inverse problems is explained in [1]. Although they are constituted and trained in a slightly different way, it is important to understand them before trying to understand cINNs. The main motivation for using INNs in this application is their invertibility properties: the mapping from the inputs to the outputs is bijective, both forward and inverse mapping are efficiently computable, and both mappings have a tractable Jacobian, which allows explicit computation of posterior probabilities.

Fig.13 compares traditional (Bayesian) Neural Networks and INNs in inverse problem applications. In Bayesian NNs, the input x is mapped to an output y, and the predicted x after the measurement y is compared to a ground truth $\hat{x}$ via a supervised loss (SL). INNs, on the other hand, add a latent variable z to the measurement y. Intuitively, z aims at capturing the information about x that is not contained in y because the mapping $x \mapsto y$ is not bijective. The inverse operation $[y, z] \mapsto x$ (sampling) is not straightforward as opposed to the forward operation $x \mapsto y$. The process $x \mapsto y$ is therefore constrained via the supervised loss $\mathcal{L}_y$, whereas the unsupervised loss $\mathcal{L}_z$ enforces that z follows a normal distribution $p(z)$ and y and z are independent upon convergence, meaning that $p(z|y) = p(z)$ (this result is proven in the paper). Both of them are written as follows:

$$\begin{cases} \mathcal{L}_y = \mathbb{E}\left[\left(y - f_y(x)\right)^2\right] \\ \mathcal{L}_z = D(q(y,z), p(y)p(z)) \end{cases}$$

where:

- $y = s(x)$ is the analytically known or easily computable mapping between the input x and the output y
- $f_y(x)$ is the neural network prediction from output x
- D is the Maximum Mean Discrepancy (MMD)
- $q(y = f_y(x), z = f_z(x))$ is the joint distribution of network outputs
- $p(y = s(x))$ (simulation) and $p(z)$ (normal) are the marginal distributions

The process of sampling is the following: when we get a new observation $\hat{y}$, the latent variable $\hat{z}$ is sampled from a Gaussian distribution and concatenated with $\hat{y}$. It then passes through the network backwards and finally yields a value for $\hat{x}$. By repeating this sampling operation a large number of times, we get an approximation of the desired posterior $p(x|\hat{y})$ as shown mathematically in the paper.

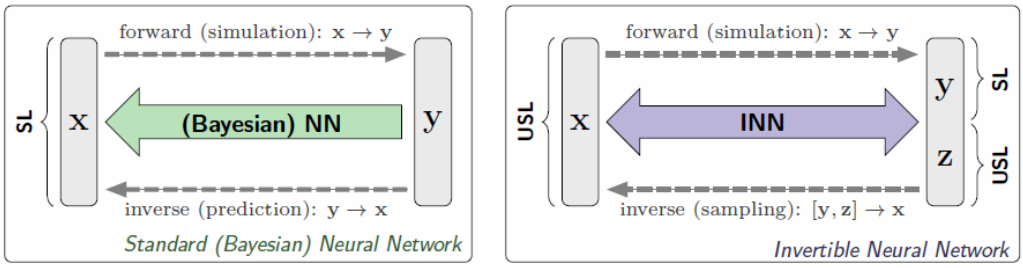


Figure 13. *INN compared to a traditional Bayesian NN [1]*

Conditional Invertible Neural Networks (cINNs)

The cINN used in [28] is pictured in Fig.14. It is an approach that is strongly inspired from INN architectures, but is fundamentally different:

- The approach of cINN is different because the output is a *probability*: y (pressure spectrum as a function of wavelength here), is introduced as a condition in the architecture, so that we are able, for each spectrum, to sample a distribution of x (sO_2 here). An INN approach would have been less appropriate here, as it would have given a unique estimation of sO_2 from the inputted spectrum via optimization in the latent space. INNs are also often considered less good when input values are noisy.
- The *training* is different: if the cINN conditioned by its parameters Θ is expressed as $f_\Theta : (sO_2, y) \mapsto z$ with $z = (z_1, z_2) \in \mathbb{R}^2$ and $y \in \mathbb{R}^{N_\lambda}_{\geq 0}$, then the model is by structure guaranteed to be invertible in the first argument and the searched posterior is supposedly well approximated by sampling the training data

$\{(sO_{2i}, y_i)\}_i$ to a standard normal distribution $\mathcal{N}(\mu, \Sigma)$, where μ is the zero vector and Σ the identity matrix, in the latent space. This has been shown to be performed by minimizing the following loss:

$$\mathcal{L} = \sum_i \left[\frac{1}{2} \| f_\Theta(sO_{2i}, y_i) \|_2^2 - Jf_\Theta(sO_{2i}, y_i) \right]$$

where Jf_Θ is the log-Jacobi determinant of f_Θ. The author of [28] claims that this loss is easier to optimize than those described previously in the INN approach.

- The *scaling* of the model is reduced when introducing the spectrum as a condition rather than as an input.

The model as presented in the paper is constituted of 20 affine coupling blocks with GLOW style affine coupling [7, 21] that are invertible by design. Each block starts with a random channel permutation. Then the scale and shift operators, respectively s and t are encoded by identically-designed but independent FCNNs with a single hidden layer of dimension 1024 followed by ReLU activation and dropout ($p = 0.5$), and having the condition as input. The other parameters of the blocks are fixed. Some values in the blocks are required to be strictly positive due to invertibility constraints and are therefore transformed by an exponential function after application of soft-clamping to 1 by Tanh activation to avoid exploding gradients. Due to the necessity of a permutation at the beginning of the block, the input vector must be two dimensional, and a gaussian auxiliary variable is added to sO_2: $x = (sO_2, av)$ with $sO_2 \in [0, 1]$ and $av \sim \mathcal{N}(0, 1)$. Only sO_2 will be considered during inference.

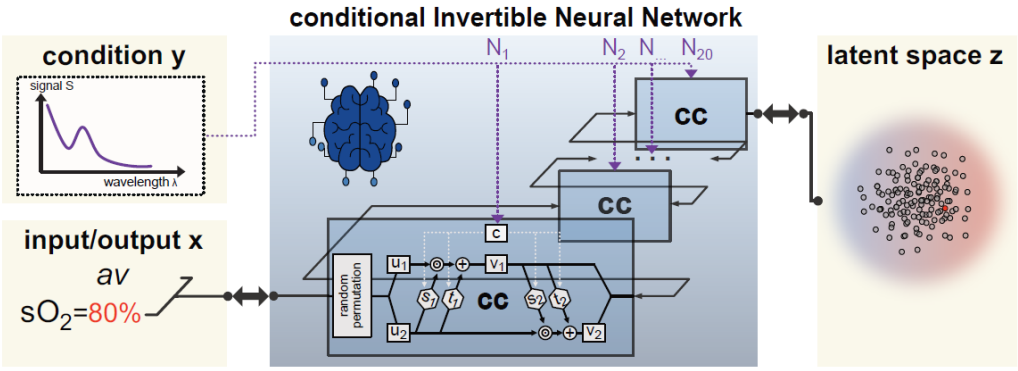


Figure 14. *Structure of the cINN used [28]*

As already explained for INNs, during inference, for every measured spectrum $\hat{y}$, the model will repeatedly sample $\hat{z} = (\hat{z_1}, \hat{z_2})$ with $\hat{z_i} \sim \mathcal{N}(0, 1)$ and compute a corresponding value for $sO_2 = f_\theta^{-1}(\hat{z}, \hat{y})$, which, done a large number of time, yields an approximation for the posterior $p_\Theta(sO_2|\hat{y})$. The method is pictured in Fig.15. Once this sampling is performed, the user needs readable outputs, hence the need for *mode detection*, which is performed by UniDip clustering, which will use the median of each mode as a point estimate. To choose the mode, different strategies can be used:

1. *user-based selection*: if the user (a clinician for example) has knowledge about the expected value. The corresponding estimate is referred to as "Best cluster".

2. *likelihood-based selection*: default strategy of considering the highest mode as the one to choose. The corresponding estimate is referred to as "Dip estimate".

3. *no selection*: if the ambiguity can be removed by different acquisition poses (see Fig.12), or if the user is interested in considering all the possibilities (e.g. presence of a critical value which, even if more likely, could have a dangerous outcome if true)

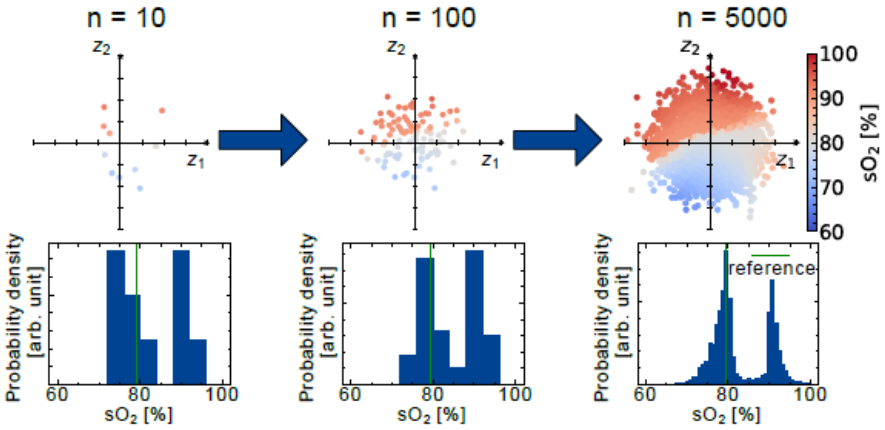


Figure 15. *Sampling process performed in cINNs [28]*

Methods

The experiment that we perform in our project, which is extensively described in 4 is inspired from the one introduced in this paper. Significant differences are the following:

- Additionally to the synthetic dataset, they build a "hybrid dataset" by generating anatomical volumes by segmentation of US measurements and annotating them with relevant physical values. They use is for testing.
  → We didn't do that due to time constraints, and will therefore only test on real and synthetic images.
- They use DAS for reconstruction.
  → We use an in-house Model-Based reconstruction algorithm that gives better image quality and is already implemented in optoacoustic devices.
- They train the models on forearm-like tissues, and evaluate their methods on real forearm measurements, as well as calf and neck measurements as out-of-distribution (o.o.d.) data.
  → We stick to calves both in training and testing.

For performance analysis in the test phase here, the maximum a posteriori probability (MAP) estimate (mode with the highest probability) is used by default and compared to the ground truth with a mean absolute error. However, in order to take into account the *user-based selection* case, the same comparison is also performed

with the closest mode to the ground truth for each pixel. The sensitivity in 5 percentage points (pp) range is computed as well. To account for the non-independence of spectra across each image, these metrics are averaged per image. The model is compared to LU and a version of LSD that is optimized on the simulated datasets.

Results

The main findings of the study are the following:

- Both in silico and in vivo, the ill-posedness of the optical inverse problem is reflected in the significant number of multimodal distributions (respectively 13% and 21%) and the important average number of modes among these (≈ 3 in both), this effect being increased as we increase the depth in tissues.
- Both cINN and LSD outperform LU in all the predictions. cINNs shows comparable performance to LSD in unimodal cases (where no mode selection is performed), and shows its real interest in multimodal cases (where the error is in average bigger) where the "closest mode" version outperforms LSD by far ($\approx 50\%$ relative error reduction and better sensitivity).
- cINN generally makes realistic predictions in vivo, with a tendency, on vessels, to better work on arteries than veins.

Limitations

The main limitations that were pointed out are listed below:

- The predictions on real o.o.d images, including calves, were already quite good, but a specific model trained on synthetic calf measurements might be more precise.
  → We will verify this statement here.
- A domain gap still exists between synthetic and in vivo data, because of the limited realism (simple shapes, lack of tissue heterogeneity, simple reconstruction) of the used simulation tools.
  → The only problem that we will be able to address is the reconstruction which will be more precise.
- real time inference is still not possible.
  → We didn't address this problem.
- The model considered as the baseline here is the "closest mode" version, which assumes that an expert would be provided a method to choose a value for each pixel and, what is more, to make the best choice in each case, which does not really match realistic use. This methodological challenge, which has to be dealt jointly with the clinical and engineering side, remains untreated to our knowledge and is a complex problem that will go way beyond the scope of this project. One of the biggest limitations of this study is therefore to consider this version as the baseline, although the only version that could be used in practice today is the "likelihood-based" one.
  → In this work, we will be aware of this limitation and study both versions separately. We will however only focus on the spectral coloring part and not address any methodological issue.

- The estimations are still made at the pixel and not the image level, and the same signal coming from both a deep and a superficial layer would therefore be treated in the same way by the model although, as discussed before, deeper pixels might have a greater chance to be multimodal. The number of training images were the bottleneck here to proceed with an image-based approach, but the author highlights that adding some spatial context to the input is an interesting research direction.
  → **This is the core of our research contribution to this work.**

4 Materials and methods

4.1 Photoacoustic simulation

As mentioned previously, we want to train a DL model to perform spectral unmixing, which requires a controlled dataset of PA images with ground truth (GT) sO_2 maps. To that end, we use SIMPA (Simulation and Image Processing for Photonics and Acoustics) [13], a framework that simulates PAI from end-to-end via six building blocks:

1. definition of a *digital twin for the PA device* (4.1.1)
2. definition of a *digital twin for the tissue volume* (4.1.2)
3. *optical forward simulation* (4.1.3)
4. *acoustic forward simulation* (4.1.4)
5. *acoustic reconstruction* (4.1.5)

Noise could have been added on the ideal pressure image (multiplicative noise) and the sinogram (additive noise), but we chose not to as a matter of simplicity.

A schematic representation of the simulation process is shown in Fig.16.

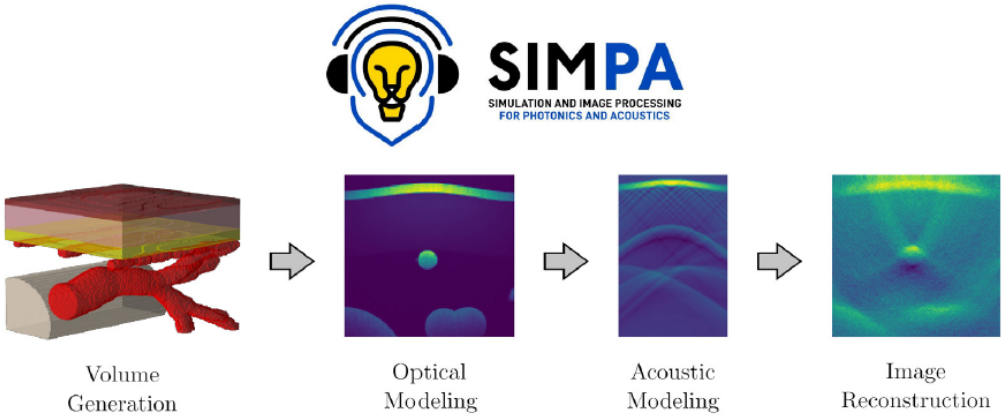


Figure 16. *Process of data generation in SIMPA [13]*

The simulation pipeline that we built for generating training images uses all these steps, and SIMPA stores input and output data of each block in HDF5 files, as well as all the settings.

4.1.1 Device digital twin

An MSOT Acuity Echo (iThera Medical GmbH, Munich, Germany) was used. Verification was performed with the clinical application as well as the R&D teams at iThera Medical to ensure the device is parameterized

correctly regarding our use case. Fig.17 gives a complete and schematized overview of how we deal with the problem from a hardware perspective. The laser illumination and the transducer elements in particular are only grossly schematized.

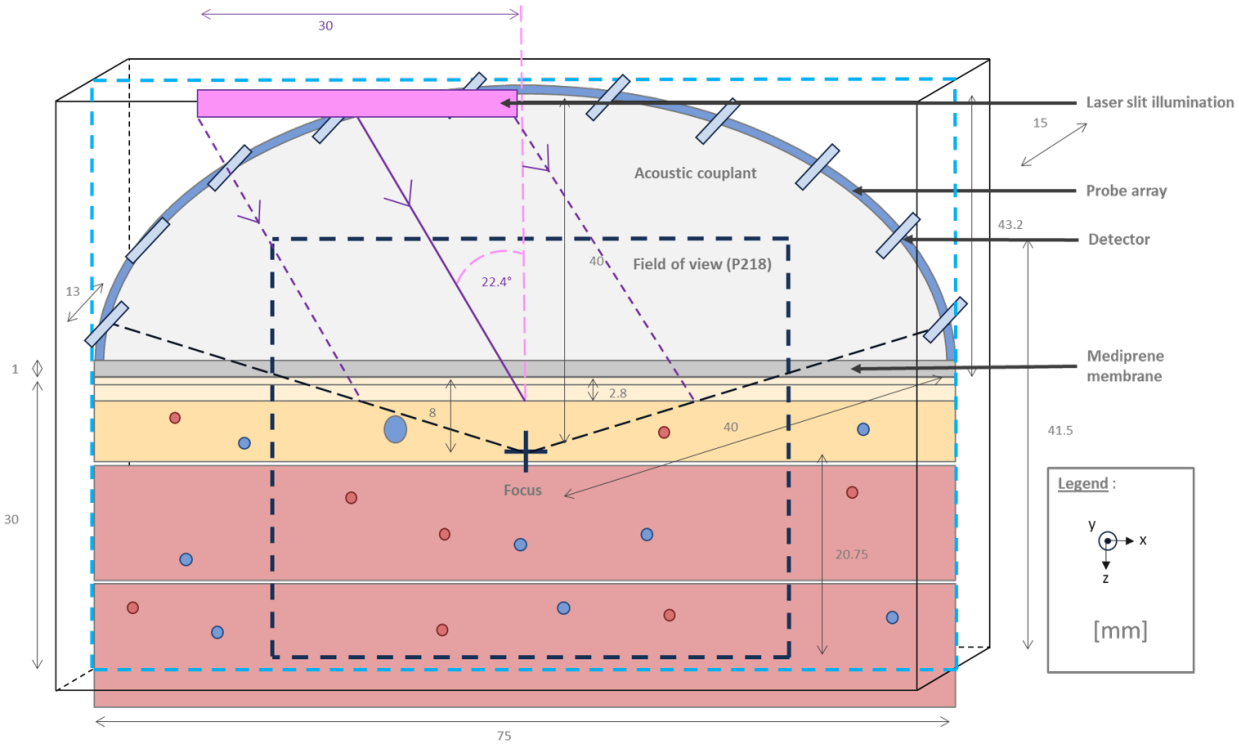


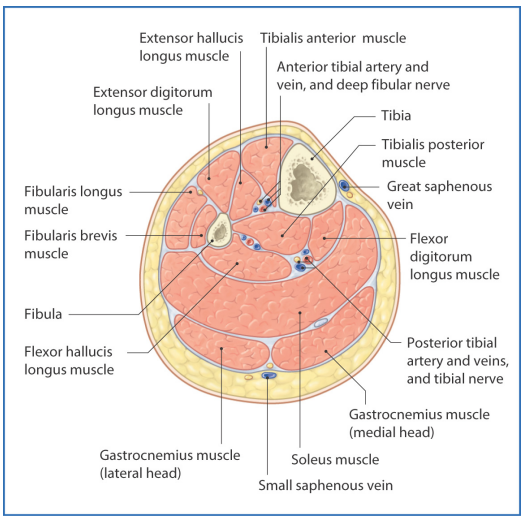
Figure 17. *Detailed scanning scenario*

The detector array is a curved array with 256 transducer elements of width 0.24 mm and length 13 mm that are biconcave (although schematized as linear there), because curved in the transducer plane (x-z) as well as the other vertical plane (y-z). The arc has a radius of 40 mm and a pitch (distance between the center of two consecutive elements) of 0.34 mm. Its center frequency is 3.96 MHz, and its bandwidth 55%. The detector array is located within the probe such that its focus is located 8 mm away from the line formed by the two outermost detectors (where the membrane is located). Between the detector array and a 1 mm thick membrane, there is an acoustic couplant which consists mainly of heavy water. The device has a 30 mm wide slit illumination located 19.0 mm off the imaging plane and 43.2 mm behind the membrane. The light cone has a full width at half maximum of 8.66° and a 22.4° tilt with respect to the imaging plane such that the center of the cone intersects the imaging plane at the focus. The field of view (FOV) that we use is a square of side 41.5 mm centered at the focus of the probe.

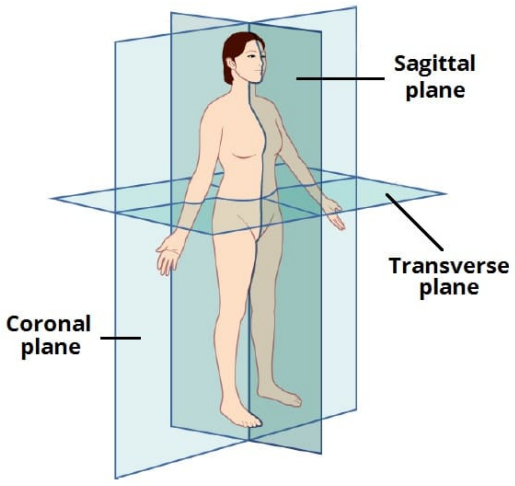
4.1.2 Digital volume definition

In the scope of this work, we exclusively wanted to design anatomical volumes mimicking human calves, and account for the scanning of relevant muscles by physicians in the case of diagnosis or follow-up of PAD. The most common locations for scanning in these cases are the gastrocnemius and the tibialis anterior, which are respectively located in the posterior and the anterior parts of the calf (see Fig.18a). The most common

clinical practice is to perform the scans respectively in the transverse and sagittal planes (see Fig.18b) at mid-calf for the gastrocnemius and the tibialis anterior.



(a) *2D transverse cut at mid-calf [19]*



(b) *Anatomical planes, [https://teachmeanatomy.info]*

Figure 18. *Relevant anatomical schematics*

As a matter of time and to start simple, we chose to study only the scenario of a gastrocnemius scan in the transverse plane. Fig.19 provides a schematic of a 2D cut in the transverse plane of the volume digital twin that was used.

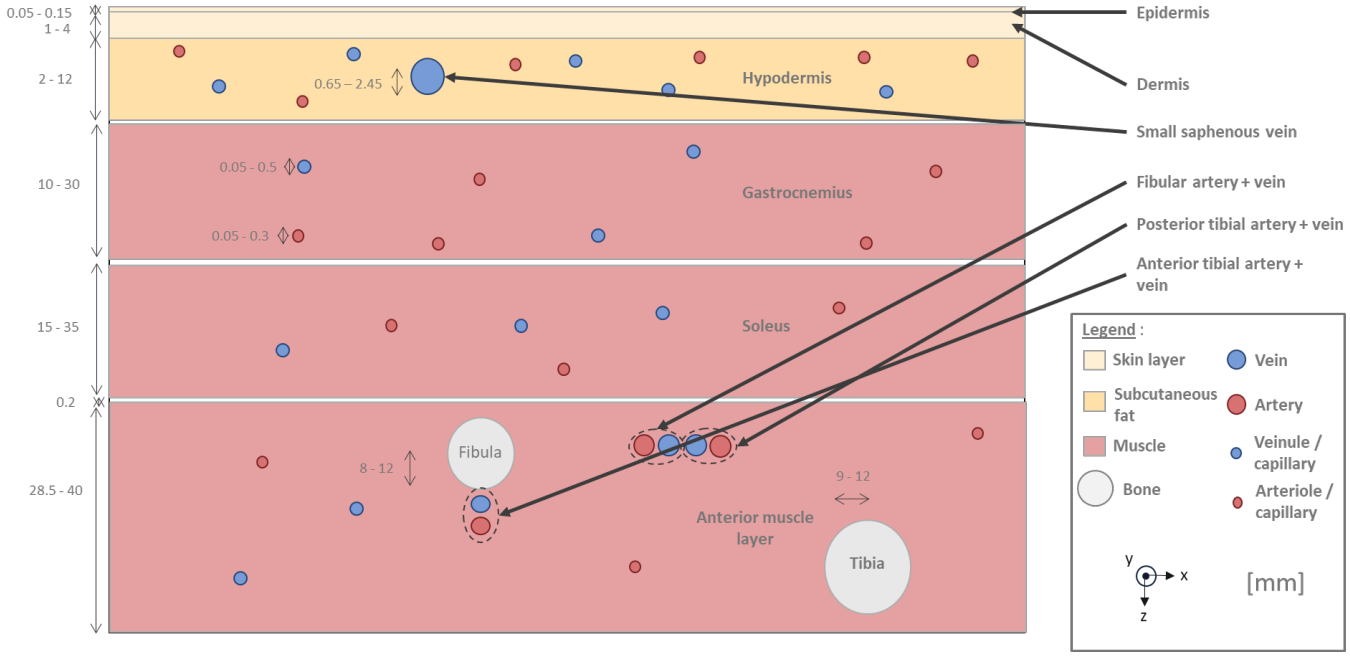


Figure 19. *2D transverse cut of the volume digital twin*

To generate anatomical volumes, SIMPA provides a model-based creator that allows to arrange in space biological tissues of interest such as skin layers, muscles, bones and vessels. These structures are assigned all the relevant parameters for simulating photoacoustic processes:

- anatomical properties: SIMPA can mostly produce simple shapes such as cuboids, cylinders, parallelepipeds and spheres, but has a more sophisticated design for vessel trees (see Fig.20).
- optical properties: absorption coefficient μ_a, scattering coefficient μ_s, anisotropy g and refractive index n (always kept to $n = 1$ here), that influence the way structures interact with light
- acoustic properties: speed of sound c mostly, influences the image formation
- molecular properties: molecular composition of each structure, density ρ

A large majority of these values are gathered in dedicated libraries in the SIMPA software and based on the literature [27, 16].

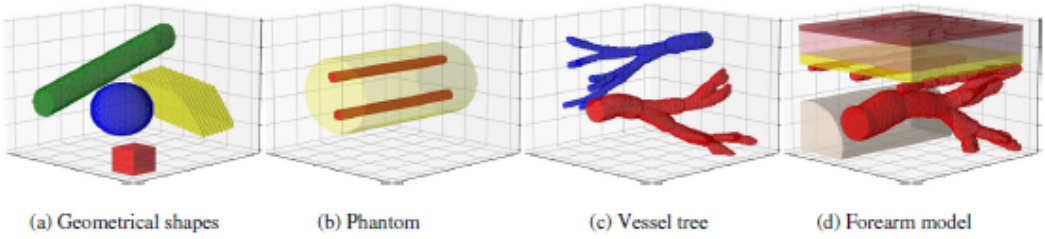


Figure 20. *Examples of volumes generated with SIMPA [13]*

Anatomical, optical, acoustic and molecular parameters of tissues are randomly picked in defined intervals following uniform laws. A majority of these reference values are already documented in SIMPA's libraries [13, 27, 16]. Some modifications and adds-on were made from our side to match our use case. The variability that will be seen by the model is a direct consequence of these choices:

- Average anatomical properties for calves are not part of the SIMPA libraries. We therefore drew them from the literature [25, 33, 15] as well as expertise from the clinical team of iThera. The chosen values are schematically highlighted in Fig.19. The 3D anatomical volume can mostly be seen as an extrusion of the 2D structure in the y direction (skin and muscle layers are parallelepipeds, bones are cylinders). Only vessels are modeled by more sophisticated structures that can have a spatially varying radius, change direction and bifurcate (schematizing all the vessels along the y direction on the drawing was just a matter of simplification).
- A broad spectrum of melanin volume fraction in the epidermis was simulated: $\mathcal{U}([1\%, 16\%])$ based on [17] to account for lighted-skinned and moderately pigmented adults but excluding darkly pigmented adults for the moment because of high absorption of the optical incident light by melanin that results in a too low PA signal with the current technology.

- A similar random choice of sO_2 values in tissues was performed based on [28, 27]: $\mathcal{U}([50\%, 80\%])$ for muscles, $\mathcal{U}([60\%, 80\%])$ for veins, $\mathcal{U}([90\%, 100\%])$ for arteries. Blood volume fraction in muscles was chosen following $\mathcal{U}([10\%, 50\%])$.
- Both the numbers of veins and arteries added in each skin and muscle layer was chosen in $\mathcal{U}([20, 30])$. The position of their origin, their direction across time as well as their curvature is also randomized.
- The small saphenous vein, that is most of the time avoided in calf scans because of the very high signal that it generates (as well as the low interest that we have of estimating sO_2 in vessels), is added in 50% of the images. It is placed along the y direction, but can be oscillate around the axis.

The total volume simulated is of dimensions 75 mm (transducer dimension) × 15 mm (out-of-scanning-plane dimension) × 30 mm (volume height) and the spatial resolution is 0.1 mm. The Grüneisen parameter is set constant to $\Gamma = 0.2$.

4.1.3 Optical forward simulation

It is performed via a 3D Monte Carlo eXtreme (MCX) simulation written in C and NVIDIA CUDA with 10^7 photons and an anisotropy $g = 0.9$. The 6 wavelengths used for the incident laser beam as the ones usually used in scanning experiments performed with MSOT Acuity Echo: 700, 730, 760, 800, 850, and 900 nm.

The laser energies are wavelength dependent and therefore drawn from the calibration spectrum of the system shown in Fig.21. Although the SIMPA software can for the moment only take as input a constant laser energy, we made the necessary modifications locally to allow spectra as inputs. A pull request was made to the SIMPA repository and is currently waiting for approval.

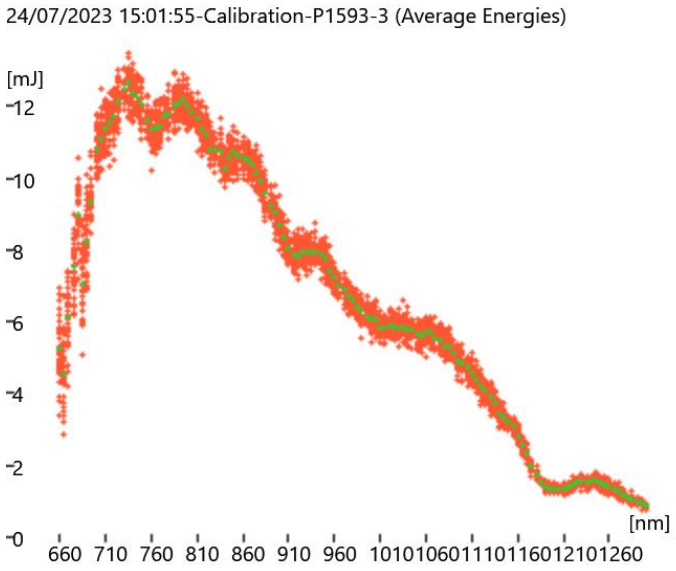


Figure 21. *Calibration spectrum of a MSOT Acuity Echo system*

This module will output a 416 x 416 pixel image that stands for the initial pressure distribution $\mathbf{p_0}$ created by the deposition and interaction of photons with tissues.

4.1.4 Acoustic forward simulation

We used the C++-accelerated kWave simulation toolbox in Matlab, and performed a 3D simulation. Some work had to be done also here since, to our knowledge, the kWave toolbox only allows to add linear elements on an arc transducer. These elements are fine for 2D acoustic forward simulation. However we faced a problem in 3D due to the mismatch between the forward simulation and the reconstruction, as the second one was assuming that signals were focused. We therefore had to bypass some functionalities of kWave to introduce a handcrafted mask with the appropriate shape for elements.

Another point to be noticed is that the sampling frequency had to be virtually adapted for numerical stability reasons due to the small value that we use for the isotropic spatial resolution ($dx = 0.1$). The CFL criterion, accounting for numerical stability, is the following:

$$CFL = \frac{dt}{dx} \times c_{sound} = \frac{c_{sound}}{dx \times f_{sampling}} \overset{!}{\leq} 0.3$$

with dx and dt being respectively the spatial and temporal resolutions. Because this criterion is not fulfilled in our case, the sampling frequency is adapted to account to the limit case of this criterion, while keeping a fixed spatial resolution and speed of sound. While $dx = 0.1$ mm and $c_{\text{sound}} = 1540$ m/s, the frequency used for simulation was therefore $f_{\text{sampling, modified}} = 51.3$ MHz. Each simulated scan lasts $\Delta t_{simu} = 5.1 \times 10^{-5} s$. The generated sinograms are therefore resampled to the ground truth sampling frequency $f_{sampling} = 40$ MHz and $N_{TS} = 2030$ time samples before being fed to the reconstruction block.

4.1.5 Acoustic reconstruction

The conclusions drawn from [11] on the one hand and [28] on the other hand are discordant regarding the interest of training the model on reconstructed images rather than theoretical pressure images (output of MCX). The main argument against reconstructed images is the fact that they introduce artifacts that do not account for in vivo situations. However, this work introduces an in-house developed Model-Based reconstruction (MB rec) algorithm that has shown more robust results than conventional algorithms implemented in SIMPA like DAS, DMAS, sDMAS and time reversal. We therefore formulate the hypothesis that images that will have undergone the whole process as described in the introduction of 4.1 will be closer to in vivo cases than the ones omitting acoustic processes, and therefore more useful for the model to learn.

Model-Based reconstruction (MB rec)

The algorithm used in our pipeline relies on model-based minimization with a least-squares approach. The reconstructed initial pressure field $p_{0,MB}$ (416 x 416, spatially discretized), is the value that minimizes the residuals between the simulated sinogram $\vec{s}$ and the one obtained "model-based":

$$p_{0,MB} = \arg\min_{x \geq 0} \left[\frac{1}{2} \|Ax - \vec{s}\|_2 + \lambda_{reg} |Shearlet(x)|_1 \right]$$

where A is the operator approximating the forward acoustic model, λ_{reg} is a regularization parameter and the Shearlet function is used as a regularization to remove curvilinear singularity artifacts (more details in [23]). The used iterative minimization algorithm is called SpaRSA (see [38]). This reconstruction modality is known as the most accurate (but also complex) one to our knowledge.

Clinical application: Deep Model-Based (Deep MB) reconstruction

Deep MB is a Deep Learning accelerated version of the algorithm described above. It was also developed at iThera Medical and described in [6]. The main motivation behind the development of this algorithm is that, although MB rec gives state of the art image quality, it has a processing time of 30 to 60 s, and is therefore not adapted to clinical use. With Deep MB, a similar reconstructed image quality is obtained approximately 1000 times faster (around 30 ms). Its generalizability to a lot of different use cases is allowed by the possibility of adjusting dynamically the speed of sound (SoS) used in the reconstruction: the image obtained can therefore be in focus for anatomic locations with very different acoustic properties.

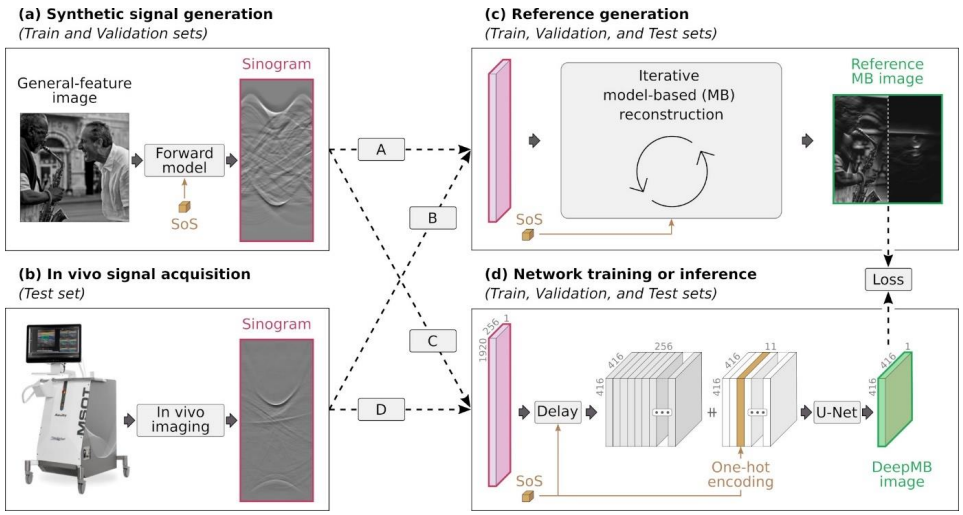


Figure 22. *Deep MB pipeline synthesized [6]*

To go a bit more into detail, the pipeline used for building Deep MB is described schematically in Fig.22. The neural network that will eventually be used as the reconstruction baseline is set against the classical MB rec algorithm. For the generation of a synthetic dataset (a), random real-life images found on publicly available datasets are used to generate sinograms with a known acoustic forward model. In vivo sinograms

are also acquired with the PA system (b). The MB rec model generates reference images from both the synthetic and in vivo datasets (c). Finally, the network is trained and validated using the synthetic data by generating images that are compared to the output of the reference model for the same input, and then tested on in vivo images (d).

The structure of the model is the following: the sinogram is brought to the image space via a domain transformation (by a delay operation); it is then passed through convolutional layers where the SoS is added as a on-hot encoding; a U-Net is subsequently applied to regress the final image and a loss is finally computed with the reference image.

Although we use MB rec for the generation of synthetic images, Deep MB is the reference algorithm to be used in the clinical routine. Because it gives similar results to MB rec's, a combination of Deep MB and our DL-based unmixing algorithm might be appropriate for real time sO_2 assessment.

4.2 Synthetic dataset

4.2.1 Raw dataset

As explained in the introduction, the data that we need for training is GT sO_2 maps (outputted by the volume creation block) and corresponding $p_{0,\,rec}$ maps obtained with the 6 simulation wavelengths (obtained after reconstruction). The models will try to compute an efficient mapping between these images as schematized on Fig.23.

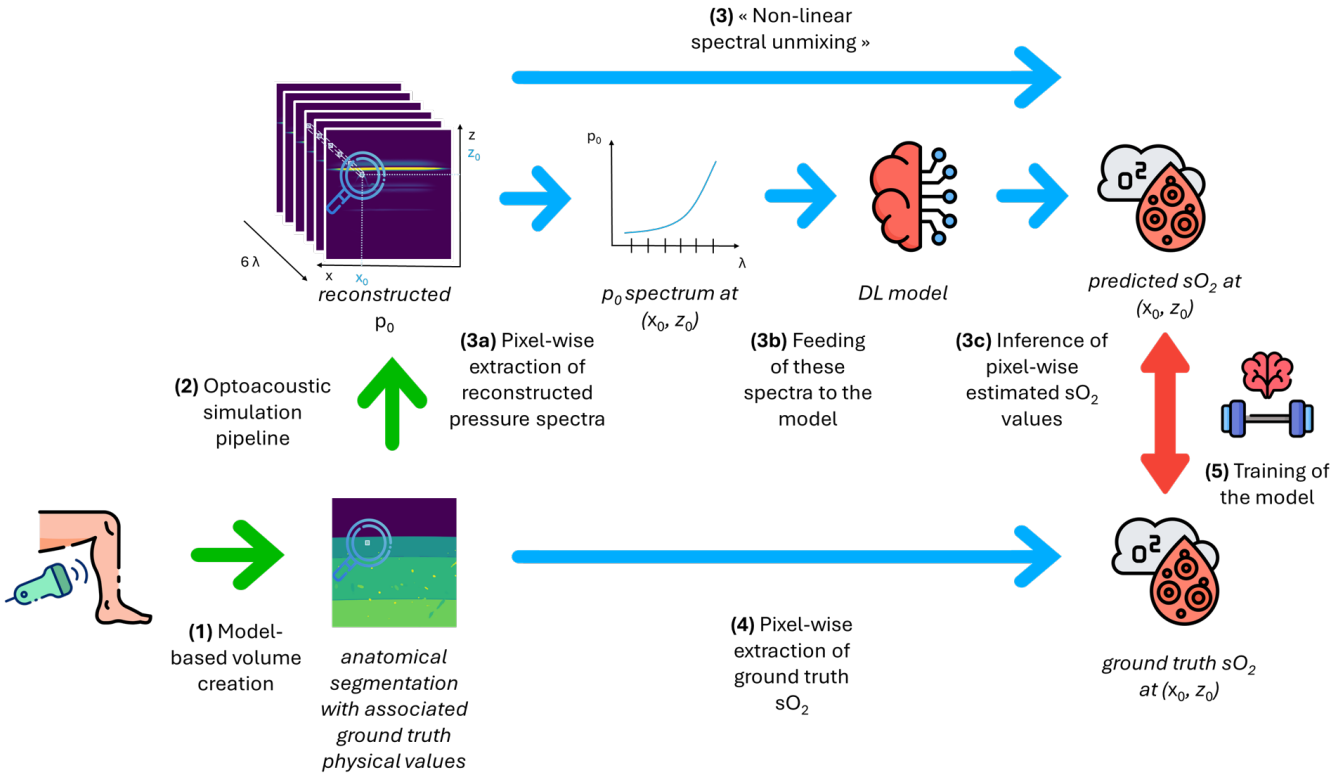


Figure 23. *Schematic of the DL approach for unmixing*

By automating the previously described pipeline that repeatedly launched this multi-block simulation, 220 virtual scans were performed and the fields that are shown in Fig.24 were retrieved. As a matter of completeness of information, certain fields such as blood volume fractions, ideal p_0 and sinograms were also stored, as they might help for analysis. Although most of the 3D fields were cropped to the transducer plane, some slices around it were kept for $p_{0,\,rec}$ and sO_2 to keep a bit of spatial information that might as well be beneficial to the analysis.

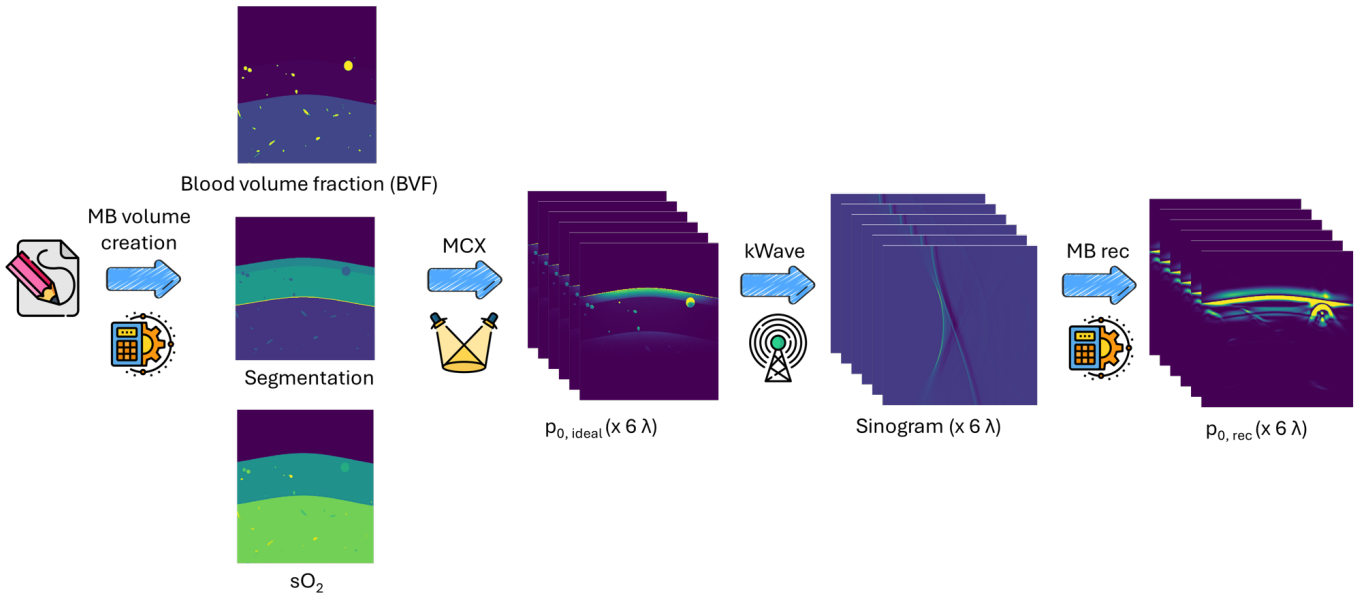


Figure 24. *Data generated by SIMPA during the simulation*

We were particularly careful while looking at the sinograms and reconstructions to ensure that photoacoustic processes were modeled correctly. Fig.25 shows a comparison between an ideal and a reconstructed pressure image at 900 nm. As a matter of optimal visualization, both were saturated to 10% of their maximal intensity.

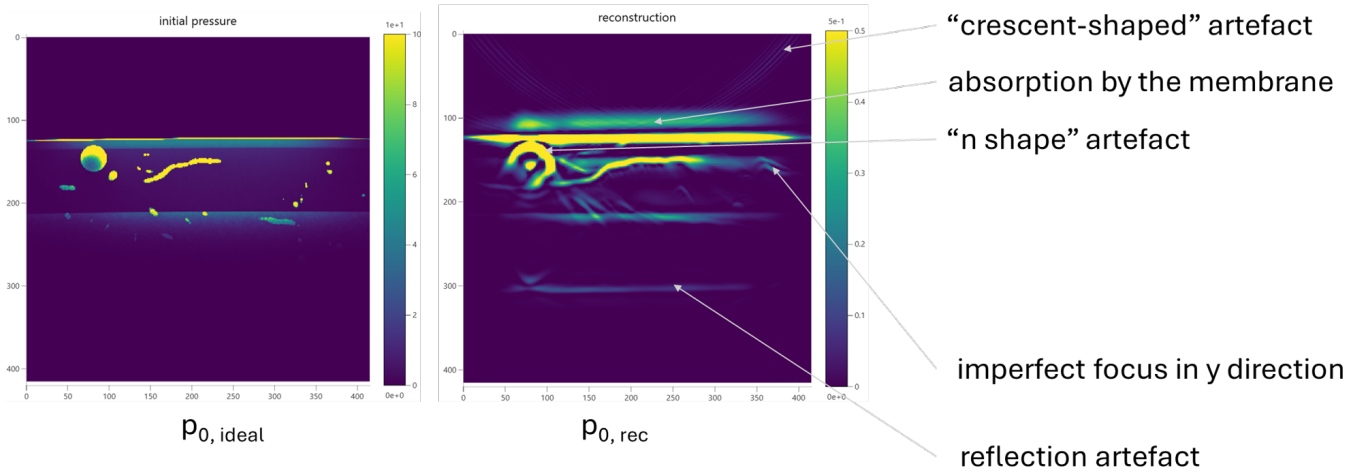


Figure 25. *Comparison between the ideal as the reconstructed pressure map*

Artifacts on the right image are pointed out. They are actually a good thing, as they are also present in vivo and mean that the physical phenomena are accurately reproduced here. In particular:

- "Crescent-shaped" artifacts are due to the limited angular coverage of our transducer. They are almost only seen at low intensities.
- The absorption line above the skin is interpreted as light absorption by the membrane material which is not completely transparent.
- "N-shaped" artifacts appear on structures that have a characteristic speed of sound that is significantly higher than the one used for reconstruction, which is the case for vessels ($c_{s,vessel}$ = 1578 m/s whereas $c_{s,rec}$ = 1540 m/s).
- The imperfect focus of the laser beam on the transducer plane will cause certain structures that are not rigorously in this plane (and thus on the middle slice of the ideal pressure) to appear because they also receive a significant amount of light.
- Reflection artifacts materialize by abnormal boundaries appearing. They are caused by the reflection of acoustic waves on boundaries between media with different acoustic impedances. Here, the deepest straight line is the mirror image of the boundary between the dermis and the hypodermis with regards to the one between the hypodermis and the gastrocnemius muscle.

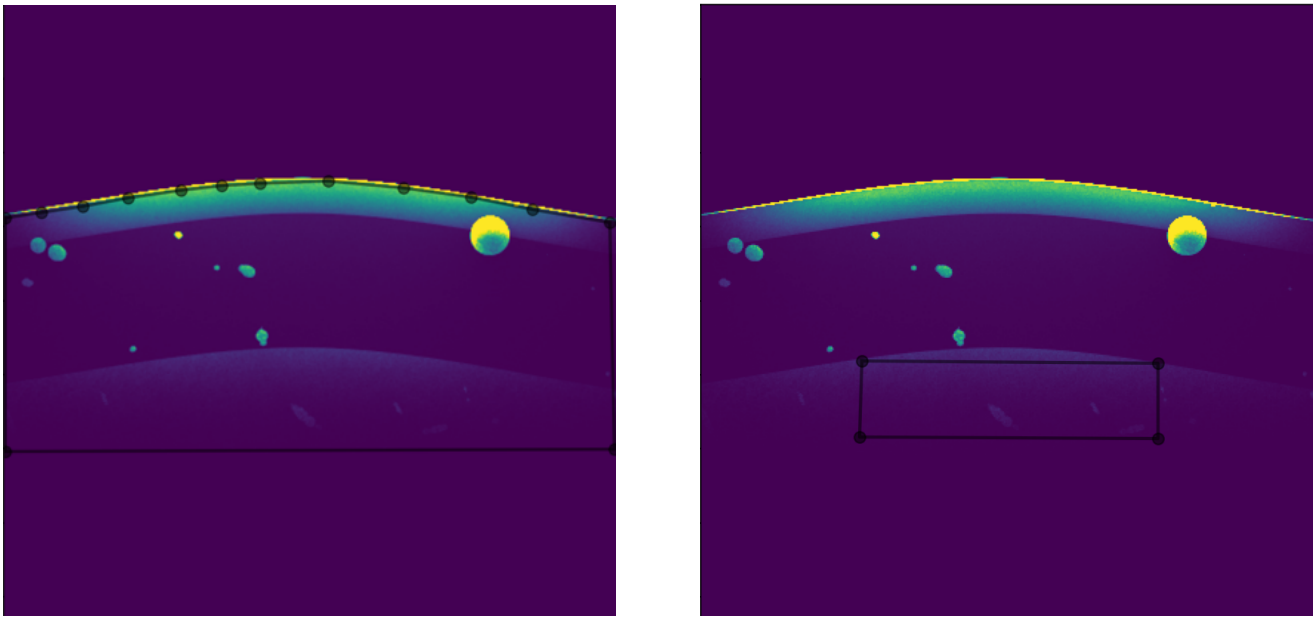
4.2.2 Pixel selection strategy

Although complete images are available, we want our model to be able to predict meaningful values in the gastrocnemius muscle mostly. All the pixels in the image are therefore not useful to our analysis and a ROI had to be defined manually. The following requirements were taken into account:

- Pixels above the skin are not useful so they can be removed.
- Pixels from the epidermis also had to be removed because this layer contains melanin, and therefore Hb and HbO_2 can no longer be considered as the two main absorbers. As we do not inform our model about the melanin content in the epidermis, we assumed that the model would be negatively impacted if it had to predict sO_2 values also for these pixels.
- Locations in the image with a too low Signal-to-Noise Ratio (SNR) are likely to be subjected to unrealistically strong spectral coloring phenomena according to Gröhl & al. [11]. In this paper, they study vessels segmented by hand and only keep pixels that have an intensity superior to 10% of the maximal intensity in their whole dataset. In our work, broader regions are studied, and the choice was made to simply remove the pixels that are located more than approximately 5 mm deep in the gastrocnemius muscle.

The ROI defined for analysis is pictured on an example of ideal p_0 map in Fig.26a. This choice is based on the assumption that the model training would benefit from variability in the inputs and outputs. Although we specifically target precise sO_2 estimation in the muscle layers, it was therefore considered beneficial that

the model also "sees" the layers above to get understanding of the early behavior of photons in the tissues. This was considered all the more relevant as we tried to inform the cINN model about the position of the pixels (as described in 4.3.2), and therefore assumed that it would learn to distinguish between the different layers. Another type of ROI that can also fulfill these requirements but includes less variability in pressure and sO_2 data is the one defined in Fig.26b: a small region of constant width and length (20 mm in x and 5 mm in z here) is defined at the top of the gastrocnemius muscle layer. This case has been studied in the ablations (see 6.4), but might have the advantage of being very specific to our use case and not biased by more superficial layers.



(a) *Baseline ROI* **(b)** *Smaller ROI, studied marginally*

Figure 26. *Example of the defined ROIs on an ideal pressure map*

4.3 Models

4.3.1 Choice of baselines

cINN

First, a choice of DL baselines to study and compare had to be performed. Among the methods reviewed in 3.2, the most promising one appeared to be the uncertainty-aware method using the conditional Invertible Neural Networks (cINN) architecture, widely described in 3.2.3. It is a recent approach using an advanced DL architecture tailored for inverse problems and was the best performing model among the ones that were reviewed for in vivo photoacoustic scanning. It is the only model to date allowing to perform uncertainty quantification and therefore tackle the problem of spectral coloring in a broader way. Although some questions remain about the methodology of use in vivo due to the high time required for inference and analysis of posteriors, this project went in the direction of developing the use of cINNs for this type of applications and hoping that these problems might be solved one day if the method proves to be highly

beneficial compared to more traditional neural networks. Exploratory work, described in 4.3.2 was therefore performed on this model to bring new development perspectives.

LSD

As a matter of comparison, Learned Spectral Decoloring (LSD) was also considered interesting due to its relatively good performances despite its simplicity. In particular, [28] shows that, in pixels where the cINN predicts a unimodal distribution, it outperformed LSD by only 0.1% in terms of Mean Absolute Error (MAE) on synthetic and hybrid test images. A much more significant difference was however observed in the case of multimodal predictions by the cINN (2.2% and 4.3% less in MAEs respectively on synthetic and hybrid test images). This proves again the interest of bringing uncertainty quantification to the problem, and suggests that there is no interest in using cINNs only as a point estimate, as the Dip estimation will give performances similar to well trained FCNNs. This hypothesis will be checked again in the analysis part of this work.

LU

These two methods will also be compared to an in-house-implemented non-negative LU algorithm using the Hb and HbO_2 spectra from the SIMPA [13] libraries.

4.3.2 Model development

One weakness of the cINN that we identified was the absence of spatial awareness of the model, as it will treat pixels identically, regardless of their distance to the transducer. This parameter however has a strong correlation with the ambiguity of this pixel, i.e. the chance that the estimation will have more and/or wider modes (this is however only an assumption and was not rigorously proven here). The only indication that it might have is that shallower pixels are usually brighter, but, illumination kept constant, this is strongly influenced by the concentration of melanin in the epidermis which varies between images. **In this work, we therefore suggest that bringing positional information might be beneficial for the model to learn.**

This information is being added in the condition block of the model together with the spectrum. In order to provide it in a smooth way, we use Positional Encoding (PE). Here we more precisely use Sinusoidal Positional Encoding (SPE), a technique that was widely used in Natural Language Processing (NLP) when the Transformer [36] architecture became gold standard. PE was later translated to image processing [8] in various shapes. In SPE, spatial coordinates are encoded with different spatial frequencies, which would supposedly allow the model to understand the variation of parameters at the corresponding spatial scales. This is mathematically translated as follows:

$$\begin{cases} PE(pos, 2i) = \sin\left(\frac{pos}{n^{2i/d}}\right) \\ PE(pos, 2i+1) = \cos\left(\frac{pos}{n^{2i/d}}\right) \end{cases} \quad \text{with} \quad \begin{cases} pos \in \{x_p, z_p\} \\ d \in \{6, 16\} \\ i \in [\![0, \frac{d}{2} - 1]\!] \\ n = 10\,000 \end{cases}$$

Positional information was to be somehow combined to the pressure spectrum. Two approaches were therefore suggested in this work and are schematized in Fig.27. In the sin cINN architecture, we encode pixels coordinates x_p and z_p into respectively $PE(x_p)$ and $PE(z_p)$ with dimension 6 (3 distinct spatial frequencies), and then $PE(x_p)$, $PE(z_p)$ and the $p_{0,rec}$ spectrum are summed term-by-term. In the sin 2cINN, the encoding is of dimension 16, and $PE(x_p)$, $PE(z_p)$ and $p_{0,rec}$ are stacked together to form a condition of dimension 38. $d = 16$ is a value sometimes used value for SPE in computer vision for light and/or data efficient models which allowed us to keep a reasonably light sin 2cINN model. These two models are purely exploratory and the values of d were chosen based on little experience, so this value might be finetunned in the future. The values of $x_p \in [\![0, 415]\!]$ and $z_p \in [\![0, 415]\!]$ are specific to our set-up and relative to the focus of the probe that is always located at the center of the image. These values might have to be renormalized if the images were to be processed differently.

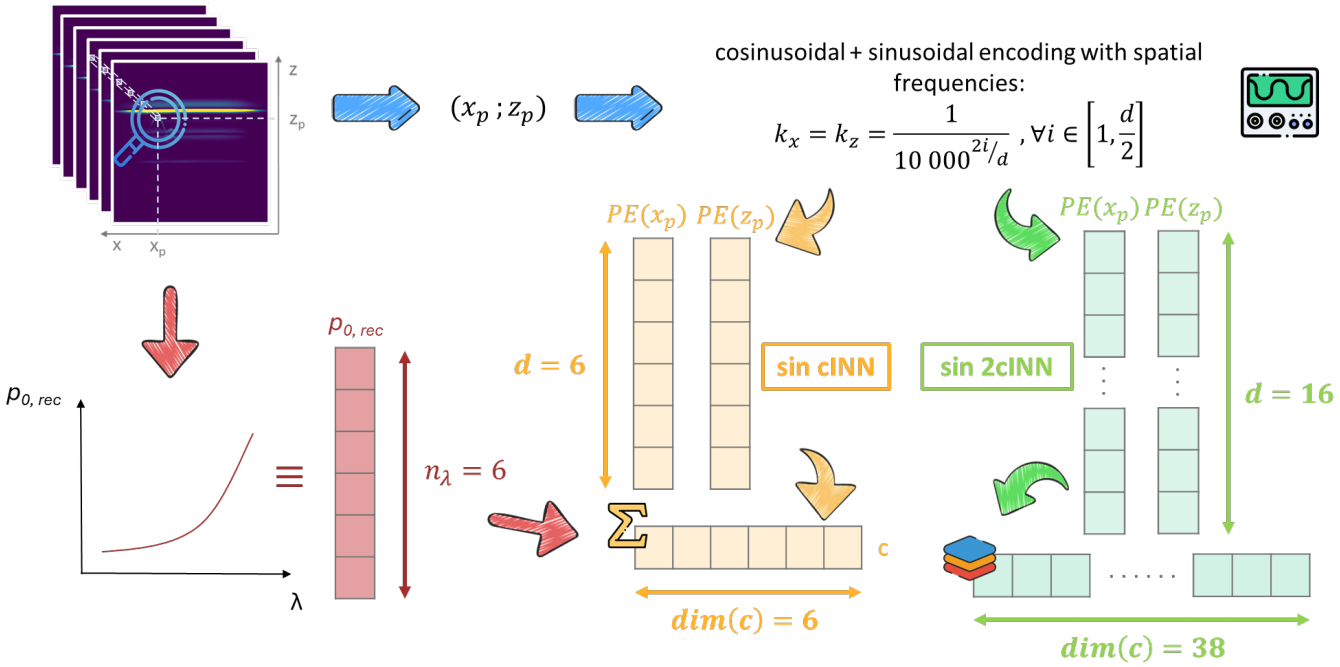


Figure 27. *Exploratory models sin cINN and sin 2cINN*

4.4 Model training and testing

In this work, models were trained on the synthetic dataset presented in 4.2. After training, their performance was assessed in silico. Some gastrocnemius scans in the transversal plane performed with a MSOT Acuity Echo were collected thanks to the application team at iThera Medical to perform in vivo testing. However, as a matter of time, this part of the analysis could not be performed.

5 Experiments

5.1 In silico dataset

The dataset that we used to train our models comprises 220 simulations that were performed on a GPU (NVIDIA GeForce RTX 4090). 215 were kept for the training process, and 10 000 pixels were randomly chosen from the ROI of each image. The ROIs initially contained between 20 000 and 80 000 pixels. This gives a 2 150 000 pixel dataset. The train/val/test sets were respectively weighted 60/20/20%. The 5 remaining images with drawn ROIs were kept apart in order to test the models on complete ROIs and assess their performances visually.

From every pixel, we extract the $p_{0,\,\mathrm{rec}}$ spectrum, the GT sO_2 and the position in the transducer plane (x_p, z_p). Each $p_{0,\,\mathrm{rec}}$ spectrum is L_1-normalized individually:

$$\forall i \in [\![1,6]\!], \quad p_i^{\mathrm{norm}} = \frac{p_i}{\sum_{j=1}^{6} |p_j| + \epsilon}$$

where ϵ is a small constant preventing division by 0 and $p_{0,\,\mathrm{rec}}$ has been simplified to p.

In the dataloader, $p_{0,\,\mathrm{rec}}$, x_p, z_p and sO_2 are z-score standardized on the whole dataset:

$$\forall i \in [\![1,6]\!], \quad x_i^{\mathrm{norm}} = \frac{x_i - \mu}{\sigma}$$

where x replaces any of the just mentioned variables and μ and σ are the mean and standard deviation of this variable on the whole dataset of size n_{DS}:

$$\begin{cases} \mu = \dfrac{1}{n_{DS}} \displaystyle\sum_{i=1}^{n_{DS}} x_i \\ \sigma = \sqrt{\dfrac{1}{n_{DS}} \displaystyle\sum_{i=1}^{n_{DS}} (x_i - \mu)^2} \end{cases}$$

5.2 Model finetunning

5.2.1 cINN

At first, we kept the cINN architecture close to the baseline from [28] that was described in 3.2.3.3. We then assessed the interest of adding PE by comparing cINN, sin cINN and sin 2cINN with identical sets of hyper-parameters. The models with PE soon gave better results (better training losses and metrics on the test set) and the best among the two was almost always sin 2cINN, as shown in 6.2. We therefore decided to focus on optimizing this architecture first.

The first focus was on training parameters, as we didn't know if we would have time for architecture finetunning. Training was performed with the AdamW optimizer, and we tried multiple Learning Rates (LR) in $\{1\times10^{-4}; 2\times10^{-4}; 3\times10^{-4}; 4\times10^{-4}; 5\times10^{-4}; 6\times10^{-4}; 7\times10^{-4}; 8\times10^{-4}; 9\times10^{-4}; 1\times10^{-3}; 1\times10^{-2}\}$, a reduction of the Weight Decay (WD) that was originally 1×10^{-2} to 1×10^{-5}, and multiple Batch Sizes (BS) in $\{64; 512; 1024; 2048\}$ (identical for training and validation). A LR scheduler was set with $milestones = \{80; 90\}$ and $\gamma = 0.1$, meaning that the LR is divided by 10 twice, in the 80^{th} and the 90^{th} epoch as suggested in [28]. Early stopping was implemented with $patience = 50$ and $\delta = 10^{-4}$ on the validation loss, and the maximal number of epochs was set to 500. We therefore let the model train substantially longer than in [28], where they stop at the 100^{th} epoch, as our models still seemed to be learning by this time of training. During training, z sampling was made in a set of size 8 192 with batches of size 128. At inference, posteriors were predicted with size 5 000.

After training was optimized, we focused on architecture finetunning. What was mostly tried was a variation in the number of coupling blocks in $\{10; 20; 40; 60\}$, the number of hidden layers of the FC sub-network (shift and scale) in $\{1; 2\}$, and their dimension in $\{1024; 2048\}$. ReLU activations were also replaced by Gaussian Error Linear Unit (GELU) for more mathematical regularity and smoothness, and dropout was removed because it was considered incompatible with the sought model invertibility.

Finally, when the sin 2cINN finetunning was over, a comparison was established again between the three models with identical sets of parameters.

The models were implemented under the Framework for Easily Invertible Architecture (FrEIA) in PyTorch and trained on a GPU (NVIDIA GeForce RTX 4090, the same that was used for simulations).

5.2.2 LSD

As we stay close to the work performed in [28] and only little time was dedicated to model finetunning and validation, we kept the same architecture as them: a FCNN with two hidden layers of size 256, ReLU activations and dropout (p=0.5). The training was identical to the one of cINNs but patience was reduced to 5, as the model usually stops learning before the 10^{th} epoch. The model was implemented in PyTorch as well.

5.3 Performance assessment

Since the analysis phase for cINNs is time-consuming, models first had to be carefully selected with regards to their behavior during training. Once one model per baseline was selected, further analysis was conducted and comparison between the baselines could be made.

MAE

To compare both cINN and LSD models during training, we monitored the Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|sO_{2,i} - \widehat{sO_{2,i}}\right|$$

where $sO_{2,i}$ is the prediction by the model, $\widehat{sO_{2,i}}$ is the GT sO_2 and n is the size of a batch of data that we study. If we monitor the MAE during training/validation or testing, n was respectively the size of a batch or the the size of the whole test set. When studying cINNs during train/val/test, the point estimate $sO_{2,i}$ was taken as the median of the predicted distribution. However, when further analysis was performed, we computed the "Dip" and "best cluster" estimate (see 3.2.3.3) and considered both cases separately when evaluating the model.

The test MAE was considered a good first indicator of which model to choose over a pool of different trained models. However, our choice also sometimes relied on the metrics mentioned thereafter which are essential for analysis.

Training losses

Between two models of the same kind, we were also using training and validation curves to compare their performances. For LSD, the training loss is a L_1 loss, which is equivalent to a MAE but will give different numerical values, as the MAE is computed on denormalized data, and the L_1 loss on normalized data. As for cINNs, their training loss is:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}\|z_i\|_2^2 - J_i\right]$$

where n is again the size of a batch during train/val and of the test set during test and J_i is the jacobian of the invertible function computed by the neural network. When studying cINNs, and in particular when we compare different architectures, it is important to take a look both at the loss and the MAE, as the jacobian term of the loss is intrinsically linked to the structure of the model. It is therefore possible when comparing model A and B that the train loss of A gets lower than that of B although the MAE is actually higher for A than for B. This can happen for example if A is a less complex model than B, that gets more easily invertible so the jacobian term will be on average higher.

MedAE

When comparing cINNs between each other, other metrics where also taken into account. The Median Absolute Error (MedAE), median value of absolute errors in a batch (train/val) or in the test set (test), is especially important in our case where models tend to make very good predictions on pixels from the top layers but bad ones in deeper layers. This causes that the MAE and MedAE are very different (in this case, the MedAE is often much lower).

mIQR

The Mean Inter-Quartile Range (mIQR), mean value of the inter-quartile range of predictions, is a good indicator of the statistical dispersion of predictions of the model. A low IQR for a pixel indicates that the prediction is confident, and a high IQR means that either the prediction is uncertain, or that multiple modes exist. We therefore do not necessarily want a low mIQR, because we want our model to capture the ambiguity in ambiguous pixels.

MCE

Finally, the Mean Calibration Error (MCE) is defined as follows from the pixel-wise Calibration Error (CE):

$$\text{CE} = \frac{1}{B}\sum_{b=1}^{B} |\hat{p}_b - \hat{a}_b|$$

with B the number of bins (confidence intervals at a given confidence level centered in the median of the predictions for a given input spectrum here), $\hat{p}_b$ the nominal confidence level (e.g. 90% for an interval at 90% confidence) and $\hat{a}_b$ the empirical coverage of the model in the bin b (i.e. the proportion of GT values comprised in the predicted confidence interval). Then:

$$\text{MCE} = \frac{1}{n}\sum_{i=1}^{n} CE_i$$

with n defined as previously depending on the step (train/val/test). This metric allows to check if the model is well calibrated, meaning that it quantifies uncertainty in a coherent way with regards to the actual observed error of the model.

6 Results

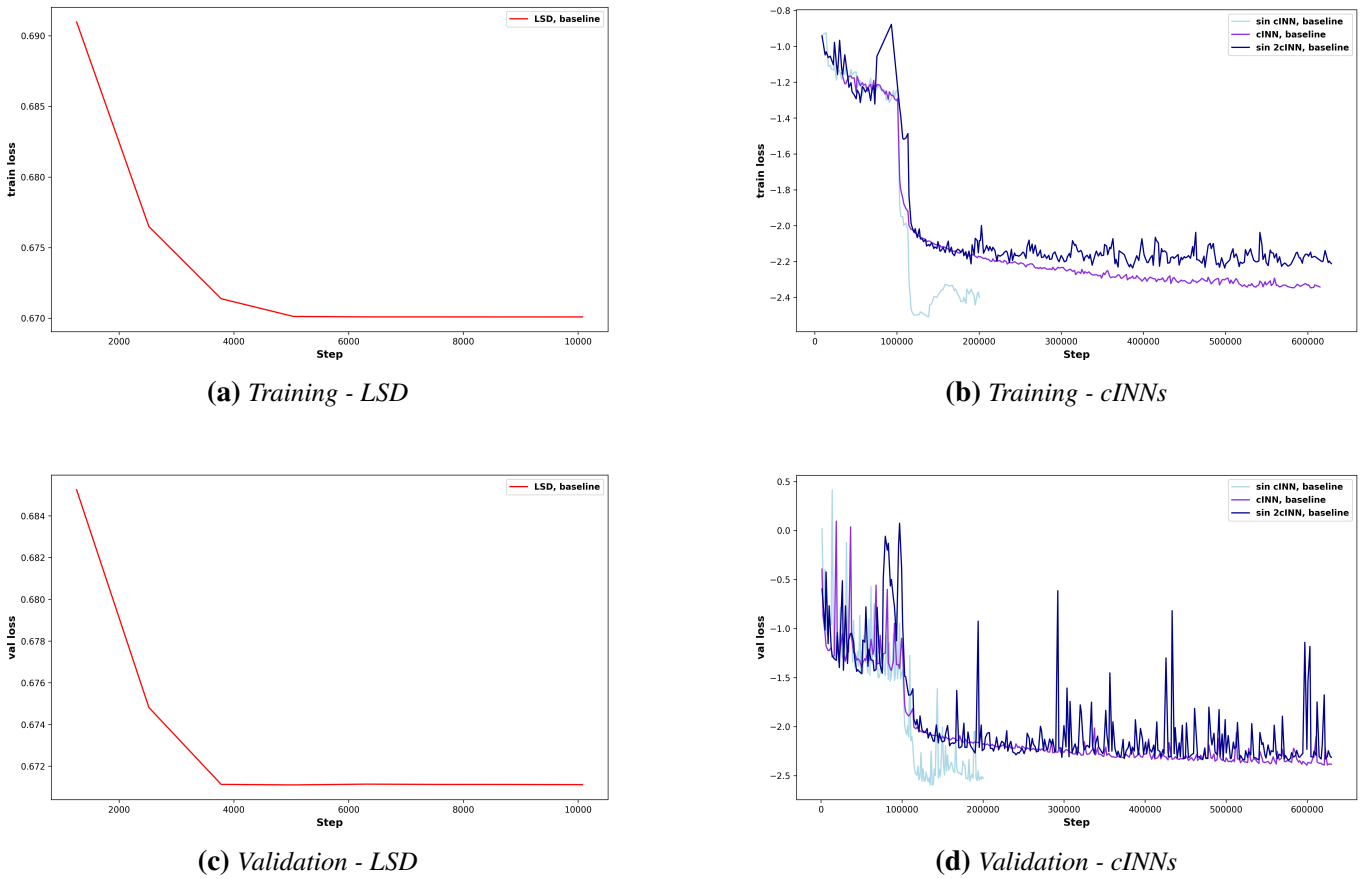
6.1 Finetunning

As already mentioned in 5.2, finetunning was performed specifically on the sin 2cINN architecture. The final optimal parameters for training happened to be: $LR = 5 \times 10^{-4}$, $WD = 10^{-5}$ and $BS = 1024$. As for the model architecture, the retained baseline is a sin 2cINN with 20 blocks and a hidden FC sub-network including one hidden layer of dimension 1024, with GELU and no dropout.

The same parameters were chosen for the sin cINN. The cINN was however showing better performances when it only had 10 blocks instead of 20 (see 6.4.2.1), which is why we chose the first case for the cINN baseline (all other parameters were kept identical).

6.2 Training analysis

Fig.28 shows the behavior of training and validation for our 4 baselines.



(a) *Training - LSD*

(b) *Training - cINNs*

(c) *Validation - LSD*

(d) *Validation - cINNs*

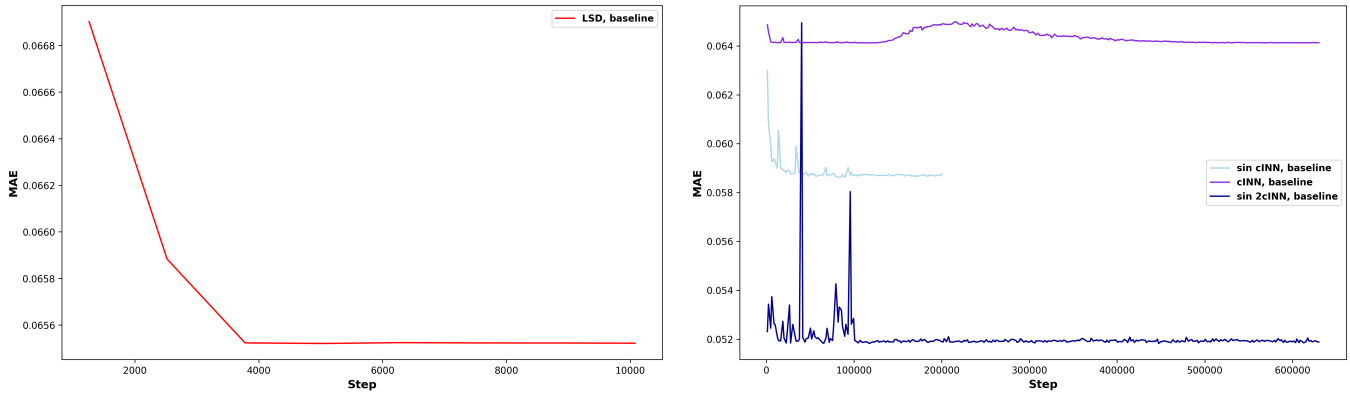
Figure 28. *Training and validation losses for LSD and cINNs.*

LSD and cINNs are separated because the training time scale is of a different order of magnitudes: while training LSD only took 8 epochs and 15 min, cINN and sin 2cINN took 500 epochs and respectively 26h and 34h. sin cINN stopped earlier due to this increase of the val loss around the 150 000$^{\text{th}}$ step so it only

took 159 epochs and 11.5h. For LSD, we are at the same order of magnitude as [12]. As already mentioned, we wanted to train cINNs a bit longer than in [28] to ensure that no training was happening after 100 epochs. It appears that training on 300 epochs might in fact have been sufficient.

The training curves are very noisy, and outliers were removed on the training curve for visualization purposes. The version with outliers is plotted in the appendix (see Fig.43). In particular, the sin 2cINN training curve had only one very high peaks at its beginning. From our knowledge, cINNs are often quite unstable during training due to the jacobian term accounting for invertibility of the model, so we considered it as normal. The drop before the 100 000$^{\text{th}}$ step (the number of the epoch is given by $\frac{100000}{1024} \approx 98$) is most likely due to the scheduled reduction of the LR at the 80$^{\text{th}}$ and 90$^{\text{th}}$ epochs, that were definitely beneficial to training. Further discussion about scheduling will be made in 6.4.1.2.

Fig.29 shows the evolution of the MAE on our 4 baselines during training.



(a) *LSD* **(b)** *cINNs*

Figure 29. *MAE for LSD and cINNs*

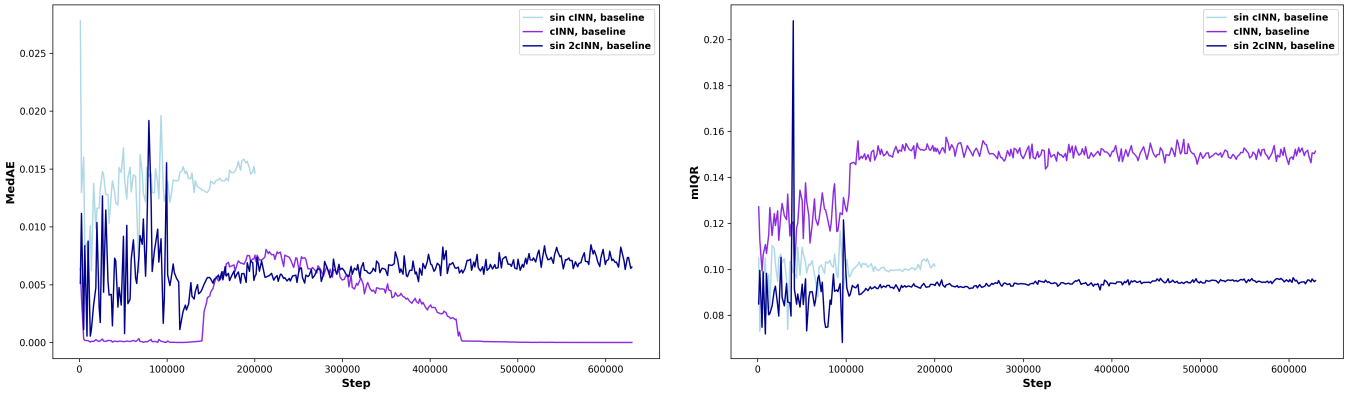
We can see that, from the first step, our methods already have error scores that are close to their final ones. Especially, LSD makes very small progress. Further analysis (see 6.3) will show that it converges to the very simple solution of predicting a constant value on the whole dataset that minimizes its MAE, which is not suitable at all. The test MAEs were listed in the table 1 including the error of an in-house implemented non-negative LU algorithm on the test set.

Model	MAE (%)
LU	44.48
LSD	6.56
cINN	6.56
sin cINN	5.98
sin 2cINN	**5.24**

Table 1. *Comparison of models on the test set*

As a reminder, the considered predicted value for cINNs is the median of the distribution here. The scores might therefore get better after computation of the mode selection algorithm. We show in 6.3 that it does indeed. The score of LU is very far away from the DL models, which highlights that these models are more reliable in this precise use case. A more critical view here is that these methods were trained only on the imaging of tissues having $sO_2 \geq 50\%$, with a very small proportion of values in [80%, 100%] corresponding to vessels and strongly biased towards 50% due to the high proportion of pixels corresponding to skin layers that were uniformly simulated as having $sO_2 = 50\%$. As we now know, LSD predicts a constant value close to 50% and already has a MAE of 6.56%. We therefore conclude that the traditional cINN (outputting the median) is not better than a constant. However, models with positional encoding show better performance, especially sin 2cINN that outperforms LSD by more than 2% in absolute.

Fig.30 shows the evolution of MedAE and mIQR for cINNs during training. Test values are listed in 2 as well.



(a) *Median Absolute Error* **(b)** *Mean Inter-Quartile Range*

Figure 30. *Other relevant metrics for cINNs*

Model	Final val loss	MedAE (%)	MCE (%)	mIQR (%)
cINN	-2.38	**0.00**	-5.36	0.15
sin cINN	**-2.53**	1.55	**1.67**	**0.10**
sin 2cINN	-1.99	0.74	4.04	**0.10**

Table 2. *Comparison specific to cINNs on the test set*

We can first underline that cINN has a quasi-null MedAE because it is strongly affected by the bias of the dataset to 50% sO_2. It has a much higher mIQR than the other methods, which goes along with the fact that it is much less certain in its predictions. On the other hand, sin 2cINN has the lowest MedAE among the two remaining models and a slightly better mIQR (same value on the test set but better training behavior), which suggests that it produces better as well as more confident predictions.

6.3 Further analysis

6.3.1 Test subset analysis

Post-processing the predictions of cINNs takes an important amount of time because the UniDip algorithm is applied to every pixel of the test set. If we wanted to analyze the whole test set (430 000 pixels), we would have needed 15h per cINN model (45h in total). In order to be a bit more flexible, we performed this analysis on a subset of 5 000 pixels of the test set. The main results are highlighted in Tab.3.

Model	MAE (%)	MedAE (%)	Q1 AE (%)	Q3 AE (%)	σ (%)
LU	45.08	50.00	36.82	50.00	17.20
LSD	6.68	**0.00**	**0.00**	13.20	9.87
cINN - Dip	7.78	0.99	**0.00**	16.56	9.76
cINN - Best cluster	6.05	**0.00**	**0.00**	10.07	9.37
sin cINN - Dip	6.79	1.57	**0.00**	13.14	9.02
sin cINN - Best cluster	4.55	0.05	**0.00**	5.69	7.68
sin 2cINN - Dip	6.27	0.36	**0.00**	12.02	8.52
sin 2cINN - Best cluster	**2.34**	**0.00**	**0.00**	**2.72**	**4.98**

Table 3. *Comparison of all the methods on the test subset, including skin layers (5 000 pixels)*

The results are here strongly biased by the predictions in the skin layers, as these pixels represent approximately half of the dataset. The LSD approach chooses the very simple solution of outputting uniformly a value of 50.001% for every pixel that minimizes that error of predicting a constant, and already gets a MAE that is way smaller than that of LU. We also find ourselves with all the methods having their first quartile of absolute error equal to 0.00%. Still, one thing to notice is that sin 2cINN performs very well, which suggests that it has found a "more intelligent solution" than that of LSD. This bias in the dataset is one of the biggest weaknesses of this work.

In order to remove it partly, another analysis was performed by removing the pixels having a GT value of 50%, which represents only 2 160 pixels out of the 5 000 randomly selected. This is shown in Tab.4 and graphically in Fig.31. Although LU keeps similar performances owing to the randomness of its predictions, we can observe that all the other models see their performance strongly decreasing because we evaluate them on pixels that are "more difficult". Almost all the models remain over 10% of absolute error, which is very large, considering that the studied interval is 30% wide. We can however see once again that the addition of positional information significantly increases the performances. A very positive result is the overall performance of sin 2cINN when we select the best cluster, that achieves mean and median absolute errors of respectively 5% and 3% on these muscle layer pixels with the smallest standard deviation among all the models (approximately 6%). These means that, although the highest mode of the distribution does

not help being a lot better than other models, this distribution actually contains the GT in a vast majority of cases, which is what we want to achieve with cINNs.

Model	MAE (%)	MedAE (%)	Q1 AE (%)	Q3 AE (%)	σ (%)
LU	47.28	49.16	30.52	64.29	21.04
LSD	15.47	15.53	7.10	22.83	9.48
cINN - Dip	14.58	13.70	6.61	22.18	9.36
cINN - Best cluster	13.98	13.20	5.45	21.91	9.60
sin cINN - Dip	12.92	11.65	5.25	19.66	8.91
sin cINN - Best cluster	10.06	6.83	2.67	16.49	9.03
sin 2cINN - Dip	11.45	10.25	4.27	17.35	8.46
sin 2cINN - Best cluster	**5.35**	**3.39**	**1.46**	**6.41**	**6.39**

Table 4. *Comparison of all the methods on the test subset, without skin layers (2 160 pixels)*

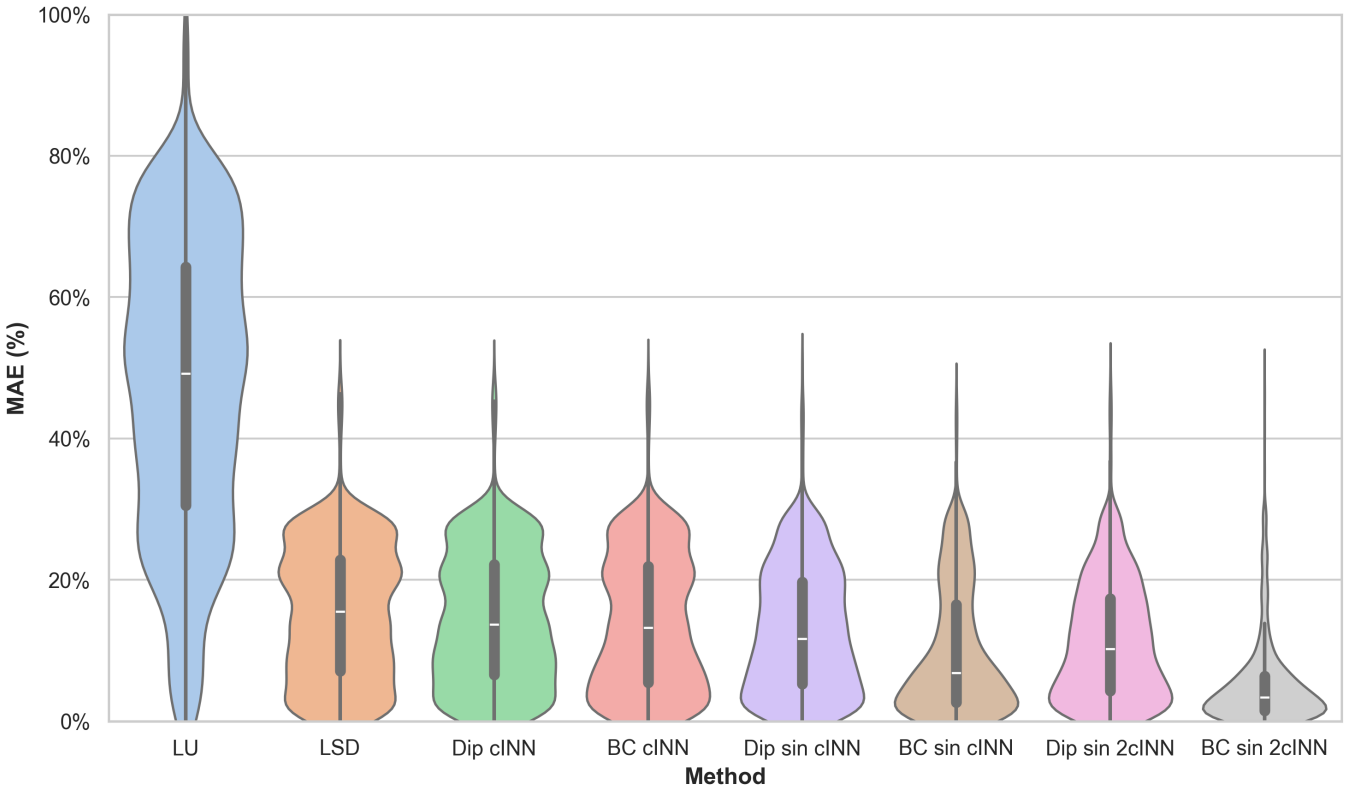


Figure 31. *MAE distributions on 2 160 random pixels from the test set, only muscle layers*

In this reduced dataset, calibration curves were plotted for our models, showing a similar behavior for the cINN baselines. Up to 50 to 60%, the baseline announces a nominal confidence interval that is too high compared to its actual empirical coverage: the model is overconfident for small confidence intervals. On the

second part of the curve, the inverse behavior is observed: the model is underconfident for high confidence intervals. This seems to make sense when we think about the data that was used for training: as the model hasn't seen pixels under 50%, we guess that the predicted samples will most probably place a very small quantity of points under 50%. However, the confidence interval will grow linearly without considering this phenomenon, which will put more empirical values than predicted in every interval crossing the limit of 50%.

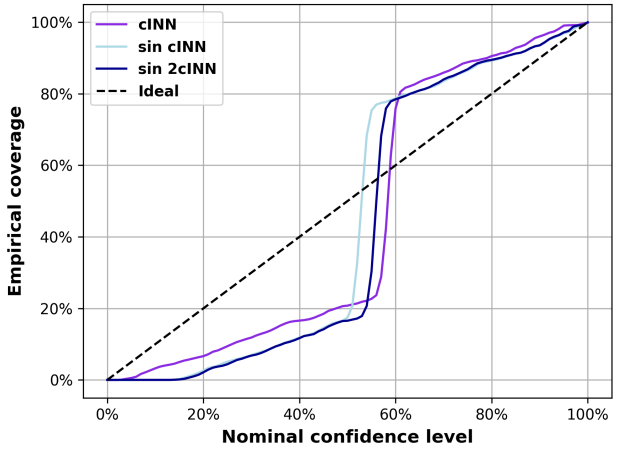
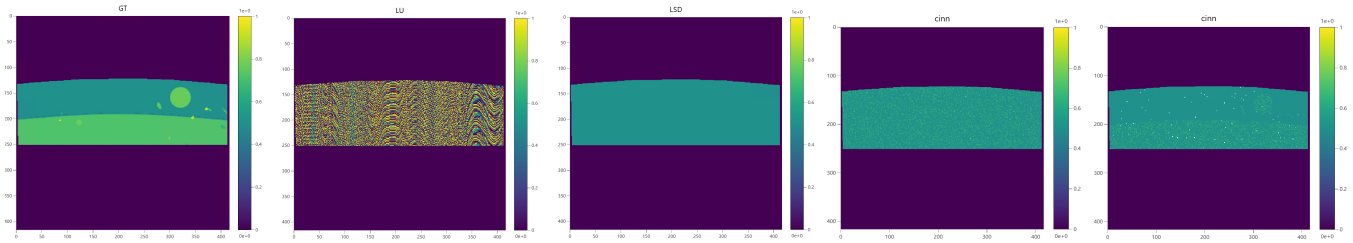


Figure 32. *Calibration curves of cINNs on 5000 pixels of the test set*

6.3.2 Image prediction analysis

As explained in 5.1, 5 simulation results were kept aside for visual analysis of the model predictions. The same analysis was performed as on the subset of pixels studied in 6.3.1. As a matter of time, sin cINN was not evaluated here. The results are shown on the first image for LU, LSD and cINN in Fig.33.



(a) *GT 1* **(b)** *LU 1* **(c)** *LSD 1* **(d)** *cINN - Dip 1* **(e)** *cINN - Best cluster 1*

Figure 33. *Comparison between GT sO_2, LU, LSD and cINN predictions on the first test image*

We mostly observe a bad prediction capacity of these three models. LU gives a lot of extreme values whereas LSD predicts a constant value as already discussed. The Dip estimate of the cINN outputs random values (that however mostly stay in the interval of training data), and only the selection of the best cluster in the cINN makes it able to identify some geometrical borders between anatomical structures. Coupled

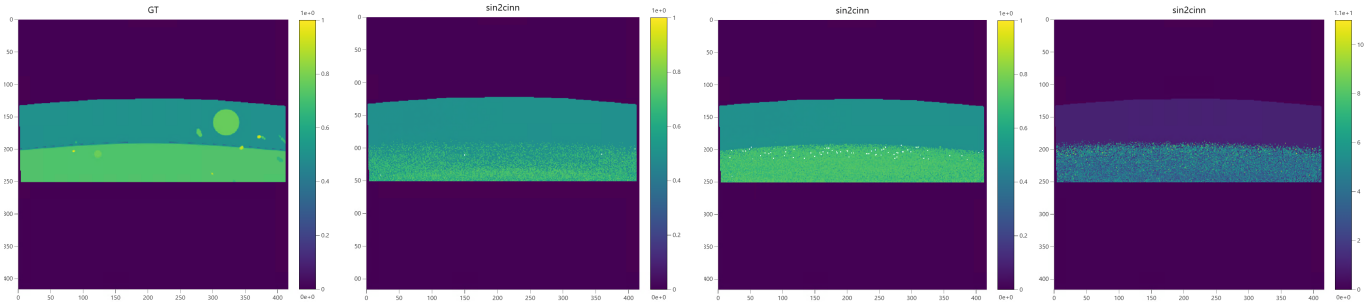
analysis with 6.3.1 however highlights that this does not make the model much better in terms of absolute quantification. The other image predictions are shown in the appendix (see Fig.47).

A more complete analysis was performed for sin 2cINN in Fig.34, and a sum up of the related metrics in shown in Tab.5. The first thing to be highlighted is that the Dip estimation actually seems to output a pattern on each image: the deeper we get, the more the model will predict high values of sO_2, which makes it better than LSD and cINN because it identifies a blurry boundary between skin and muscle layers. This is most probably linked to the positional information that we feed the model. It however appears that this boundary is an average of the boundaries over all the images, which does not provide with an "intelligent" solution. Moreover, the impact of position seems to be overtaking that of the pressure spectrum. Changes in the preprocessing of data or the architecture of the model might be done to increase the weight of spectra. Analyzing the distribution of the number of modes, we can confirm the hypothesis that learning to predict correctly skin layers was "easy", as mostly unimodal predictions are outputted here, and that pixels got more ambiguous while going deeper in the tissues. Finally, the estimation given by the best cluster is much better than the Dip estimation. The border between the skin and muscle layers are accurately identified, and the uniform sO_2 value in the muscle layer is generally well predicted.

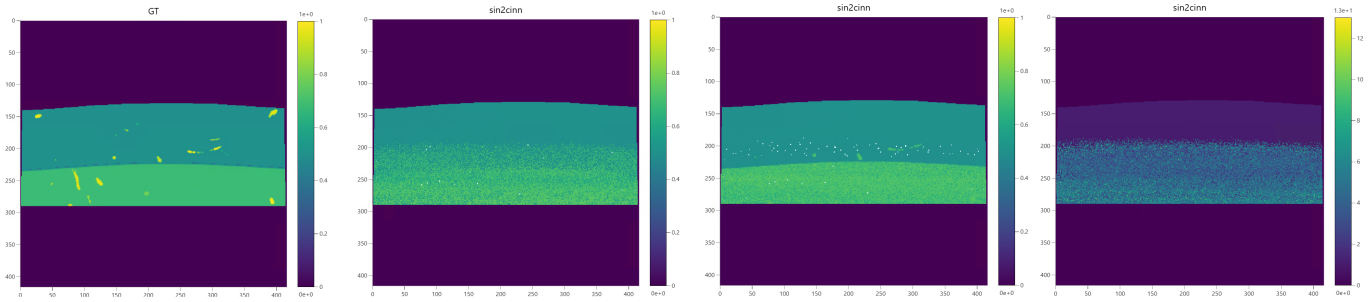
A few leads for improvement can however be made. First, the pixels from the top are strongly biased to 50%, and vessels are therefore not identified here. The model probably only learns that if the pixel is on the top of the image, it has to be skin layer pixel. Then, high sO_2 values in vessels are usually not predicted. This can be linked with the scarcity of pixels having values over 80% in our dataset. Finally, the number of modes can reach very high values in the deeper layers, which raises the question of the difficulty of finding a methodological solution for in vivo application of the method.

Image ID	MAE Dip (%)	MAE best cluster (%)	Average nb modes	Proportion multimodal (%)
1	6.57	2.66	2.41	42.94
2	4.49	1.94	3.06	57.58
3	5.38	3.84	2.37	42.09
4	7.61	0.47	3.11	58.21
5	6.68	1.06	3.08	57.82
Average	**6.15**	**1.99**	**2.81**	**51.73**

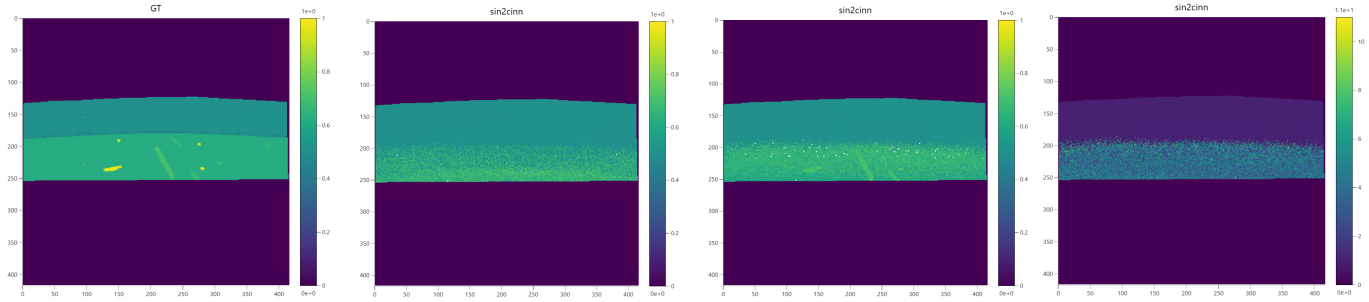
Table 5. *Metrics on the image prediction test set*



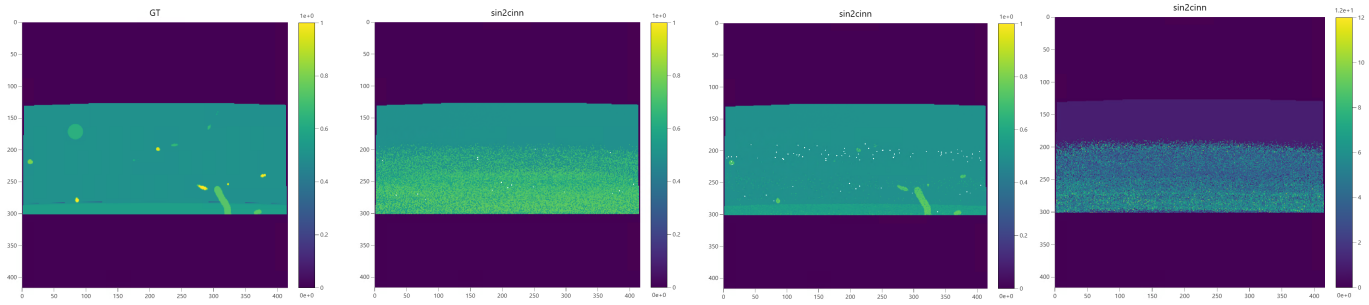
GT 1 *Dip estimate 1* *Best cluster 1* *Number of modes 1*



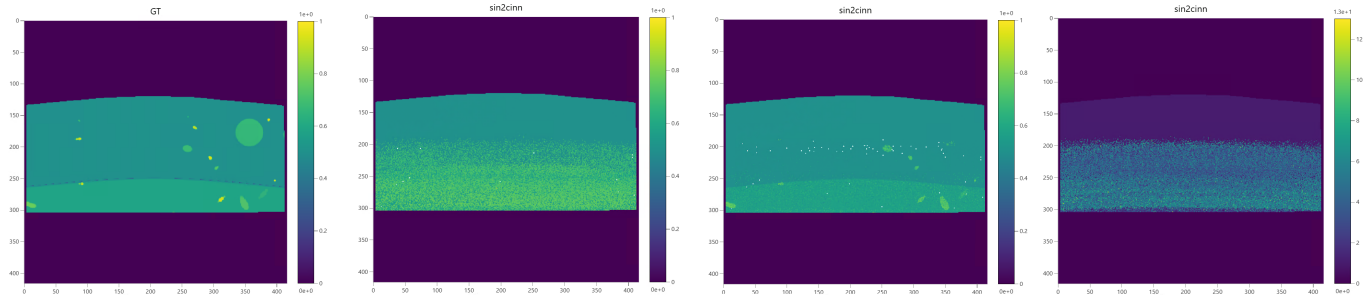
GT 2 *Dip estimate 2* *Best cluster 2* *Number of modes 2*



GT 3 *Dip estimate 3* *Best cluster 3* *Number of modes 3*



GT 4 *Dip estimate 4* *Best cluster 4* *Number of modes 4*



GT 5 *Dip estimate 5* *Best cluster 5* *Number of modes 5*

Figure 34. *Comparison between GT sO_2 and sin 2cINN predictions across the 5 test images*

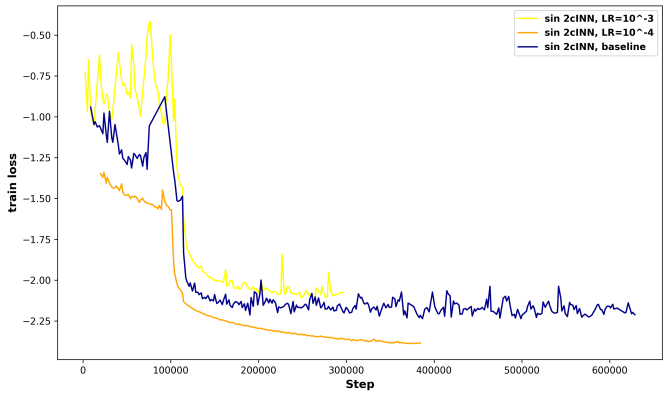
6.4 Ablation studies

In this part, the exploratory work of finding optimal parameters for the sin 2cINN model is presented briefly. All the models shown are variations from the baseline with only the mentioned parameter(s) differing. In most cases, the presented results constitute actual ablations, as they show that the selected baseline is already finetunned to obtain good results. However, some results, that were established after the selection of a baseline for this report, might constitute room for further analysis because they show that variants show performances close to the baseline, and sometimes even a bit better. This inconvenience is due to a lack of time available at the end of this work. The rooms for improvement will however be properly summarized at the end of this report.

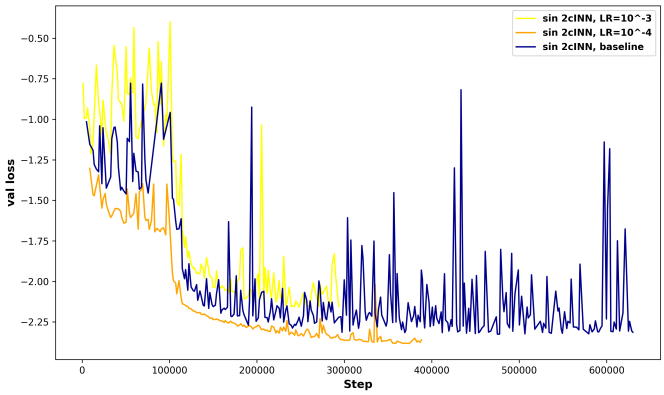
6.4.1 Training parameters

LR

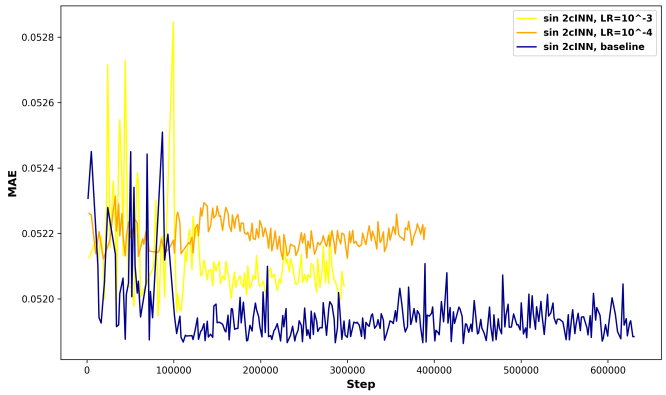
Using a high initial LR like $LR = 10^{-2}$ gave a very unstable training from the first iterations. Such a large value seems inappropriate for training unstable models like cINNs. The optimal value seemed to be in $[10^{-4}, 10^{-3}]$. Fig.35 (outlier were removed for the three curves) shows a comparison of training with different LR within this range.



(a) *Training loss*



(b) *Validation loss*



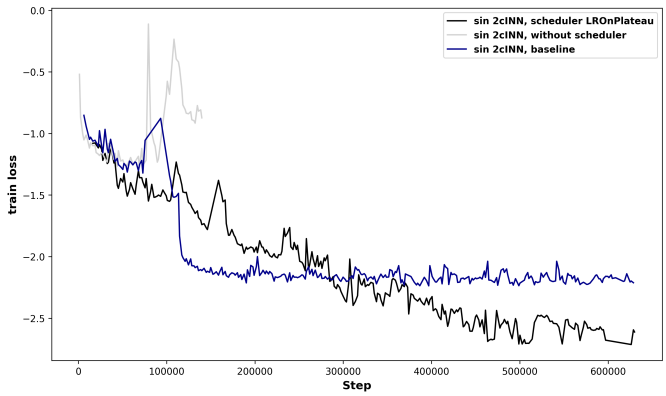
(c) *MAE*

Figure 35. *Metrics during training with different LR*

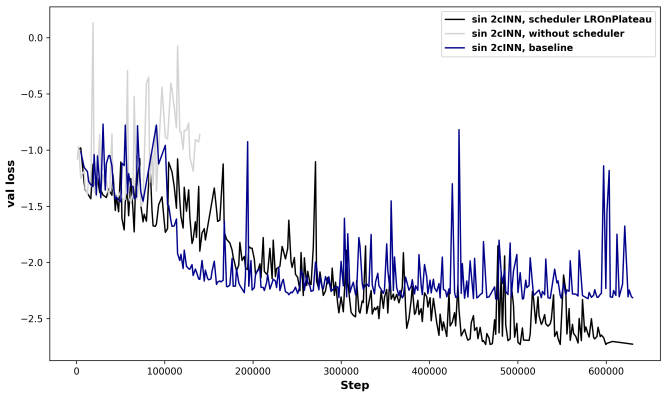
The performances are very close. $LR = 10^{-3}$ shows less good training and MAE, and a more unstable training (mostly visible when we keep the outliers). Between the two remaining cases, we chose the optimum with the appearance of the MAE curve, that was keeping increasing for $LR = 10^{-4}$. This choice is however arbitrary because a slightly better MAE during training does not mean that the Dip or best cluster estimates are going to be more accurate. In 6.4.1.2 we anyway underline that the use of a better scheduling scheme might be beneficial, but would have to be finetunned with LR within this range.

LR scheduler

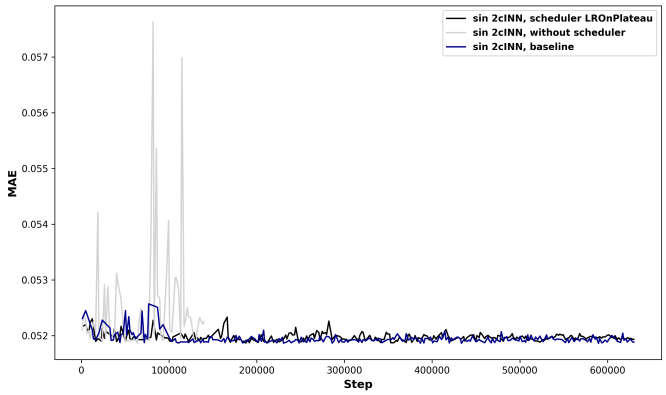
As explained in 5.2, the LR scheduler chosen for the baseline is based on [28]. This ablation was tried at the very end, and constitutes an important area for improvement. Fig.36 shows a comparison between the baseline, a model without LR scheduler, and another model with a more sophisticated scheduler implemented under PyTorch and referred to as ReduceLROnPlateau. In this last case, the LR will be multiplied by 0.5 (factor) when the val loss (monitor) will have spent 20 epochs (patience) without decreasing (mode=min).



(a) *Training loss*



(b) *Validation loss*



(c) *MAE*

Figure 36. *Metrics during training with different LR schedulers*

We can see that without scheduler the model stops learning after about 60 epochs and the training curve even increases. The division of the LR is therefore beneficial to the training. In the baseline, we divide twice by ten in the 80th and the 90th epoch which is efficient and brings to a plateau soon after. In the case

of the new scheduler, the decrease is slower but the training losses get better than the baseline eventually for equivalent MAEs. This suggests that there is work to do on finetunning this scheduler, which might bring even more optimized training.

Model	MAE	Final val loss	MedAE (‰)	MCE (‰)	mIQR (‰)
Baseline	**5.24**	-1.99	**0.74**	**4.02**	9.52
No LR scheduler	5.27	-0.86	0.90	4.62	**8.37**
ReduceLROnPlateau	**5.24**	**-2.66**	0.80	12.10	9.26

Table 6. *Comparison of sin 2cINNs with different schedulers on the test set*

We complete the analysis with final performances that are shown in Tab.6. We can conclude that using a more sophisticated analysis can be interesting: we get a lower final validation loss, but a higher MedAE and MCE. As said before, a higher MedAE or even MAE does not mean that the Dip or best cluster estimates are not going to be better. Besides, MCE was very unstable during training (see appendix, Fig.45), which raises the question of actually using this metric for making such choices. Further analysis therefore must be performed there, and more scheduling schemes with different initial LR, factor and patience would definitely have to be tried.

BS

Using $BS = 2048$ showed unstable training with a training loss increasing from the beginning as shown in Fig.37. It was tried with a quite high LR ($LR = 10^{-3}$) because usually it is advised to increase the LR when the BS is increased.

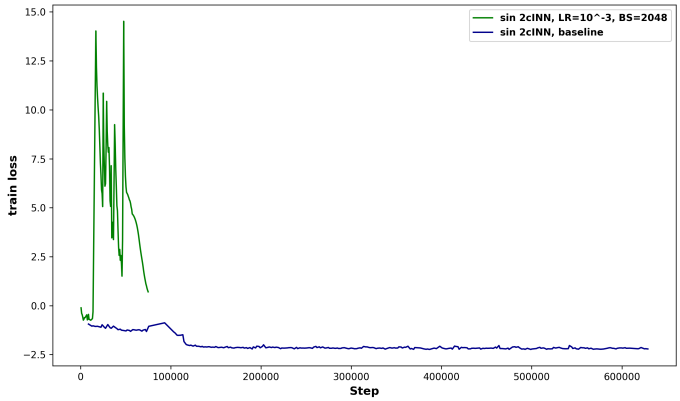


Figure 37. *Comparison of training between baseline sin 2cINN and a variant with higher LR and BS*

Using too small BS (like $BS = 64$ or 512) gave too slow trainings compared to $BS = 1024$, which was therefore considered the optimal value.

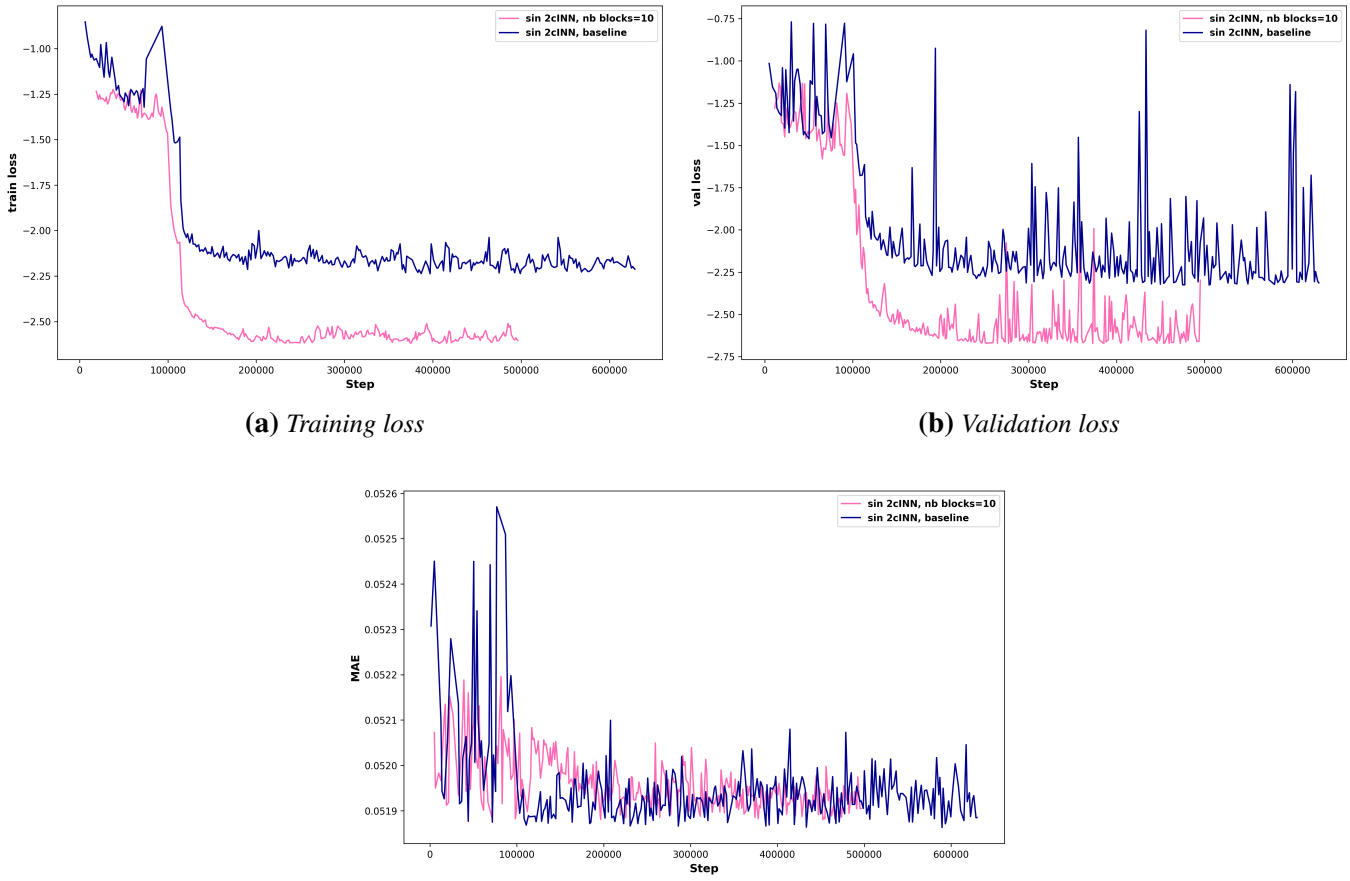
WD

The WD was not extensively finetunned. According to our knowledge on cINNs, it is generally advised to put a small WD to provide a soft regularization and avoid disturbing the invertibility of the model or to underfit by removing its freedom to learn (weights will be forced to be too small). We therefore decided to change $WD = 10^{-2}$ to $WD = 10^{-5}$, which gave similar training curves an a bit better MAE.

6.4.2 Architecture parameters

Amount of blocks

The number of blocks in the sin 2cINN was finetunned because it is closely linked to the model complexity. Analysis showed that having 40 or 60 blocks gave too complex models that were very unstable during training. However, the comparison with the 10 blocks sin 2cINN was more interesting. As shown in Fig.38, we had similar training behaviors.



(a) *Training loss*

(b) *Validation loss*

(c) *MAE*

Figure 38. *Metrics during training with different number of blocks*

The final metrics are shown in Tab.7. We end up in a similar situation as in 6.4.1.2: the model with only 10 blocks has a higher MedAE and MCE, but a better final validation loss. Again, the use of MCE for making choices is questionable regarding its behavior during training (see appendix, Fig.45).

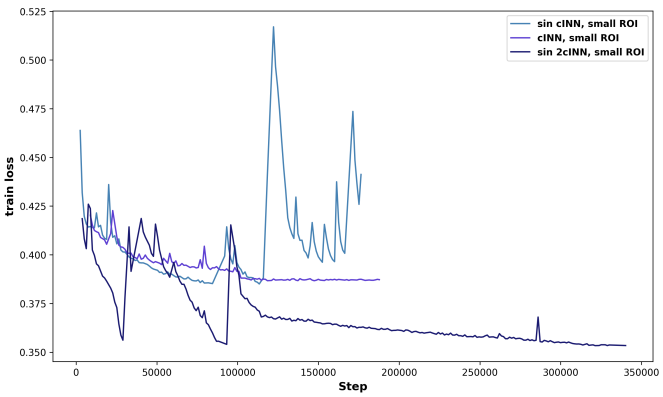
Model	MAE	Final val loss	MedAE (%)	MCE (%)	mIQR (%)
Baseline	**5.24**	-1.99	**0.74**	**4.02**	9.52
10 blocks	**5.24**	**-2.67**	0.82	10.70	**9.02**

Table 7. *Comparison of sin 2cINNs with different number blocks on the test set*

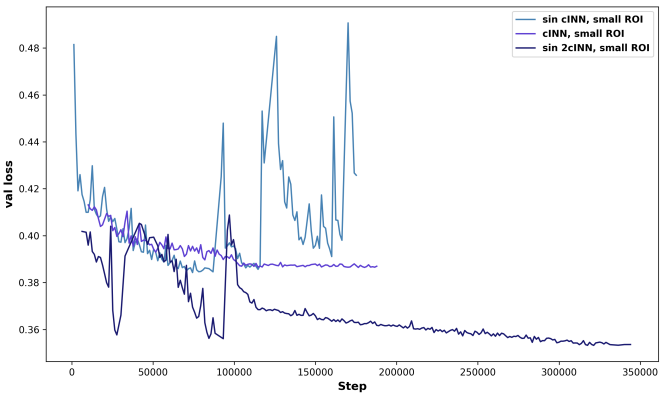
Shift and scale FCNN

We also tried to increase the number of layers in the FCNN sub-network to 2, and, separately, to double the dimension of the single hidden layer to 2048. Both cases gave very unstable training, so the leads were not pursued.

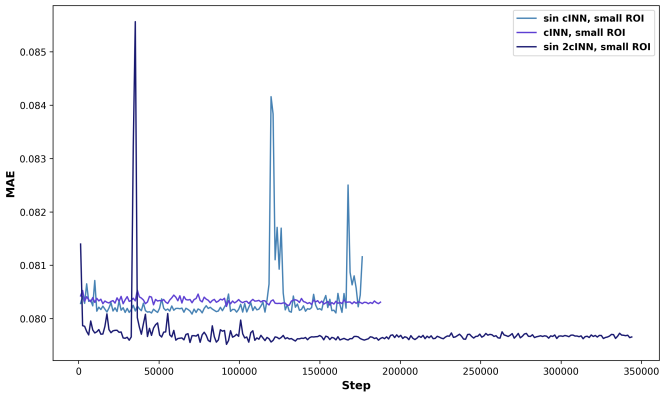
6.4.3 Reduction of the ROI



(a) *Training loss*



(b) *Validation loss*



(c) *MAE*

Figure 39. *Training curves for the cINNs trained on the reduced ROI*

As mentioned in 6.3, one weakness of the synthetic dataset used here is the lack of physiological variability in the skin layers, which has created biases in some models. One way to remove this bias is to study only

the smaller ROI described in 4.2.2. In this approach, we keep exactly the same training process. The only difference is that, instead of picking randomly 10 000 pixels from every image, we actually keep the whole ROI (200×50 pixels). The values of sO_2 seen by the model are therefore randomly drawn in $[50\%, 80\%]$, $[60\%, 80\%]$ and $[90\%, 100\%]$ for muscle, vein and artery pixels respectively. A small analysis was done here by assessing the performances of our three cINN baselines on this reduced ROI. LSD was removed from the analysis because it was again predicting constant values. We got the training curves pictured in Fig.39 without outliers. The trainings are much more unstable than when we train on the larger ROI (see versions with outliers in the appendix, Fig.44), which makes sense because the model is only learning on pixels that are "more difficult" because deeper.

Model	MAE	Final val loss	MedAE (%)	MCE (%)	mIQR (%)
cINN, small ROI	7.99	0.39	7.87	0.09	0.16
sin cINN, small ROI	8.08	0.55	7.99	-3.71	**0.14**
sin 2cINN, small ROI	**7.90**	**0.35**	**7.78**	**0.03**	0.16

Table 8. *Comparison of cINN baselines trained on the smaller ROI on the test set*

The final metrics are shown in Tab.8, and these numbers, coupled to the appearance of the training curves, highlight that, even with a smaller variability in positional parameters, the effect of position is still beneficial to training. The test MAEs for the median of the distribution are around 8% on muscle pixels, which is lower than the minimum 10% that we were getting for Dip estimates when we removed the skin layers in 6.3. If we consider that the performances of the Dip estimate and the median are usually close, this might suggest that these more specific models are better than the ones trained on the large ROI for muscle sO_2 estimation for point-wise estimation. However, the behavior of the best cluster is not known here, and knowing it would require further analysis. The training is much more unstable and the final val losses are less good than the ones for the baselines trained on the large ROI. These leads us to believe that two opposite consequences of the absence of skin pixels are still possible: (1) this makes the training harder because they were "easy pixels", but the model learns the same or even better on the muscle pixels, or (2) this lack of information actually leads the model to learn less than on a broader ROI. Unfortunately, due to a lack of time at the end of this work, this question remains open.

7 Discussion

This Master's Thesis work was written under the supervision of Guillaume Zahnd and with the help of interdisciplinary teams at iThera Medical, mostly R&D and clinical applications, to suit the needs of iThera. It was as well part of the SENDERO project lead by the department of Artificial Intelligence in Biomedical Engineering (AIBE) at FAU Erlangen. Part of this work might be pursued in the scope of research works at AIBE, and in particular in Moritz Schillinger's PhD. Suggestions will be provided here if future work was to be performed based on this one.

7.1 Simulation and dataset

In this work, a complete in silico simulation pipeline for modeling photoacoustic imaging was set up. Its first purpose was the generation of a synthetic dataset to train Deep Learning models selected in the state of the art for spectral unmixing. The generated images specifically mimic scans of the gastrocnemius muscle in the transverse plane, that are used inter alia to detect and follow up Peripheral Artery Disease. This pipeline now includes a state of the art Model-Based reconstruction method, as well as a few adaptations to overcome the current limitations of the SIMPA software. This includes the adaptation of 3D acoustic simulation with kWave to an inputted 3D transducer mask, as well as the possibility to define wavelength-dependent laser energies based on the mimicked system calibration. These modifications will be communicated to the owners of the SIMPA repository and might hopefully help them to develop the tool.

A large effort was put in this work to ensure that our simulated data match reality as much as possible. A digital twin of the calf anatomy was created based on literature values thanks to the model based volume generation algorithm in SIMPA. Another approach that was mentioned in [28] was segmentation-based volume generation, where tissues were segmented from in vivo US measurement and physiological values were subsequently associated with them. The "hybrid data" generated that way were only used for testing in this case, but we think that using them for training might reduce the domain gap between real and synthetic images. As researchers at AIBE were currently developing similar tools for producing hybrid data, this could be a lead that they would like to pursue.

Variability was introduced in the simulated data by varying the geometry as well as physiological parameters of tissues. One weakness of our work is the very low variability of pixels in the skin layers, where a uniform value of 50% blood oxygen saturation was set. These created biases in models that generally learn better from diverse data. Finding a better balance between physiological accuracy and data diversity will have to be found if further simulations have to be done. Increasing the variability of pixels in the skin layers of our tissues seems to be the first thing to try. A more ambitious idea that came up during this work was to use real world images and create an artificial variability in the parameters of the different layers modulated by the intensity of these images, similarly to the approach used in [6]. This would however bring higher uncertainty towards the ability of the models to bridge the *domain gap*.

Synthetic data were preprocessed by selecting a Region of Interest excluding pixels from the epidermis as well as too deep pixels with very low Signal-to-Noise ratio where quantification was considered too difficult. Based on the assumption that Deep Learning models learn better with diverse data, a very large region was selected. As shown in 6.4.3, reducing the region to muscle layers gave mixed results, with unstable models but good accuracy of the median estimate during training. Further work would have to be done to decide the best strategy or try new ones, mostly assessing the performance of sin 2cINN while selecting the Dip estimate or the best cluster.

7.2 Model training and validation

We trained two Deep Learning models from the state of the art to perform spectral unmixing and compared them to Linear Unmixing: a simple Fully-Connected Architecture, Learned Spectral Decoloring (LSD), and a more complex model tailored to handle ill-posed inverse problems and uncertainty quantification, Conditional Invertible Neural Networks (cINNs). We were also able to suggest the addition of positional encoding to state of the art cINNs and therefore came up with two new baselines, sin cINN and sin 2cINN. We showed that sin 2cINN outperformed all the other baselines for sO_2 estimation, with only 5% absolute error on average when the best mode is selected. The values of errors on our use case in silico are higher than in the literature (see [28]). However, the simulation set up was different (we used 3D acoustic simulation while they used 2D, and we performed reconstruction with Model-Based Reconstruction instead of Delay-and-Sum) and all our baselines performed worse, suggesting that our dataset might be more complicated.

Due to the limited time and resources available to conduct this project, training and validation of models was not so extensive. In particular, no time was dedicated to adapt the in-house Linear Unmixing algorithm to our data, which might have been necessary given the very bad predictions that it is making in silico (much worse that in [28]). LSD could as well have been finetunned a bit more to prevent it from overfitting and only outputting a constant value. Trying a larger and more complex model might be a way to go if ever we want it to make finer predictions. However, this model might just be too simple for the complexity of the physical phenomena at stake here compared to new models from the state of the art.

From the perspective of optimizing the sin 2cINN model, although we already have good performances, some room remains for improvement. Using less invertible blocks (10) or using a finer learning rate scheduler seems to make the training better but gives a less good performances for the median estimate. Further analysis have to be performed to assess the potential of these models. Another area for further work would be the optimization of positional encoding and its interaction with spectral data. As shown during the analysis, positional information seems to be what guides the Dip estimate of sin 2cINN, and bringing more weight to spectral data in the architecture might be a way to counterbalance this effect. Besides, the encoding with $d = 16$ was chosen arbitrarily, but usual models in computer vision use larger values that could be tried. The drawback will be longer trainings, that already take about one day each.

7.3 Applicability of the developed methods

It was chosen here to mostly study cINN models because of their ability to perform uncertainty quantification and solve the ill-posedness of the optical inverse problem in a more rigorous and understandable way. However, contrary to LSD, which can infer at the order of 30 ms per image on a GPU, the applicability of these models in real time in the clinical practice is today not tested, and would be possible only by defining really small ROIs on each image due to the time-consuming step of UniDip clustering on predicted posteriors. In [28], the author mentions that these methodological problems might find a solution one day. We assume that optimizing the sampling to a certain size during inference and computational acceleration of the UniDip algorithm might be ways to go. Another problem is the difficulty to exploit the full potential of the method, because "user-based selection" of the correct mode has to be performed pixel-wise for the moment. We hypothesize that a solution might be to use clustering algorithms applied to the image to identify common modes for close pixels that belong to the same tissue. This would allow to propagate the choice made by the clinician on one pixel to any pixel in the neighborhood having a similar value in its modes, and to gradually reconstruct an image without having to select all the pixels one-by-one. This is however just the sketch of an idea without any guaranteed feasibility.

Finally, these methods have to be applied in vivo so that we can assess their actual applicability and ability to actually bridge the gap between simulated and real data. Gastrocnemius transversal scans were already collected on our side, and the remaining work would be the analysis of inference results of our models on these data.

8 Outlook

Although the methods developed here are far from being ready to use, some useful and reusable resources were produced in this work.

First, the simulation pipeline can easily adapted to multiple other purposes. Different scanning scenarios could for example be simulated, either to train other scenario-specific Deep Learning models for spectral unmixing (as general purpose models seem very far away for the moment), or to perform out-of-distribution validation in silico for the already trained methods. For example, PAD scanning is also commonly performed on the tibialis anterior muscle (anterior part of the calf) in the sagittal plane. Using the digital twin from this work and adapting it slightly will give the possibility to simulate this case easily.

The synthetic dataset might also be interesting to train other models for spectral unmixing. As already mentioned, simulation data that was for the moment irrelevant was saved in case new information were to be given to further developed models. In particular, one idea that was discussed with AIBE was the addition of the epidermal melanin content in the information provided to models. As this information strongly influences the propagation of the optical signal, informing models like sin 2cINN about it could benefit them just like positional encoding did. How to quantify melanin content when applying these models in vivo and how to inform the model about this information remain open questions.

Finally, all the work made on models will of course be made available to anyone interested in using them.

Bibliography

[1] Lynton Ardizzone et al. “Analyzing Inverse Problems with Invertible Neural Networks”. In: (2019). eprint: 1808.04730. URL: https://arxiv.org/abs/1808.04730.

[2] Alexander Graham Bell. “The Production of Sound by Radiant Energy”. In: *Science* os-2.48 (1881), pp. 242–253. DOI: 10.1126/science.os-2.48.242. eprint: https://www.science.org/doi/pdf/10.1126/science.os-2.48.242. URL: https://www.science.org/doi/abs/10.1126/science.os-2.48.242.

[3] Jim Elliot Christopherjames et al. “Monte Carlo simulation of handheld probes to detect non-invasive ductal carcinoma from diffuse optical reflectance signals”. In: *Results in Optics* 11 (Mar. 2023), p. 100410. DOI: 10.1016/j.rio.2023.100410.

[4] B. T. Cox, J. G. Laufer, and P. C. Beard. “The challenges for quantitative photoacoustic imaging”. In: 7177 (2009). Ed. by Alexander A. Oraevsky and Lihong V. Wang, p. 717713. DOI: 10.1117/12.806788. URL: https://doi.org/10.1117/12.806788.

[5] Benjamin T. Cox et al. “Quantitative spectroscopic photoacoustic imaging: a review”. In: *Journal of Biomedical Optics* 17.6 (2012), p. 061202. DOI: 10.1117/1.JBO.17.6.061202. URL: https://doi.org/10.1117/1.JBO.17.6.061202.

[6] Christoph Dehner et al. “DeepMB: Deep neural network for real-time optoacoustic image reconstruction with adjustable speed of sound”. In: (2023). DOI: https://doi.org/10.1038/s42256-023-00724-3. eprint: 2206.14485. URL: https://arxiv.org/abs/2206.14485.

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. *Density estimation using Real NVP*. 2017. arXiv: 1605.08803 [cs.LG]. URL: https://arxiv.org/abs/1605.08803.

[8] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[9] Marco Gerling et al. “Real-time assessment of tissue hypoxia in vivo with combined photoacoustics and high-frequency ultrasound”. en. In: *Theranostics* 4.6 (Mar. 2014), pp. 604–613.

[10] L. Griffiths and C. Jim. “An alternative approach to linearly constrained adaptive beamforming”. In: *IEEE Transactions on Antennas and Propagation* 30.1 (1982), pp. 27–34. DOI: 10.1109/TAP.1982.1142739.

[11] Janek Gröhl et al. “Distribution-informed and wavelength-flexible data-driven photoacoustic oximetry”. In: *Journal of Biomedical Optics* 29.S3 (2024), S33303. DOI: 10.1117/1.JBO.29.S3.S33303. URL: https://doi.org/10.1117/1.JBO.29.S3.S33303.

[12] Janek Gröhl et al. “Learned spectral decoloring enables photoacoustic oximetry”. In: *Scientific Reports* 11.1 (Mar. 2021), p. 6565.

[13] Janek Gröhl et al. “SIMPA: an open-source toolkit for simulation and image processing for photonics and acoustics”. In: *Journal of Biomedical Optics* 27.8 (2022), p. 083010. DOI: 10.1117/1.JBO.27.8.083010. URL: https://doi.org/10.1117/1.JBO.27.8.083010.

[14] Lina Hacker et al. “Tutorial on phantoms for photoacoustic imaging applications”. In: *Journal of Biomedical Optics* 29.8 (2024). DOI: 10.1117/1.JBO.29.8.080801. URL: https://doi.org/10.1117/1.JBO.29.8.080801.

[15] Eren İSMAİLOĞLU and Elif İsmailoglu. “Evaluation of Subcutaneous Adipose Tissue in the Thigh and Calf Region for Subcutaneous Injection by Computed Tomography”. In: *Mehmet Akif Ersoy Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi* 9 (Aug. 2021). DOI: 10.24998/maeusabed.971037.

[16] Steven L Jacques. “Optical properties of biological tissues: a review”. en. In: *Phys Med Biol* 58.11 (May 2013), R37–61.

[17] Steven L. Jacques. “Skin optics summary”. In: *Oregon Medical Laser Center News* (1998). URL: https://omlc.org/news/jan98/skinoptics.html.

[18] N. Keshava and J.F. Mustard. “Spectral unmixing”. In: *IEEE Signal Processing Magazine* 19.1 (2002), pp. 44–57. DOI: 10.1109/79.974727.

[19] BasicMedical Key. *Anterolateral Leg*. 2025. URL: https://basicmedicalkey.com/anterolateral leg/ (visited on 01/16/2025).

[20] Jeesu Kim et al. “Programmable Real-time Clinical Photoacoustic and Ultrasound Imaging System”. In: *Scientific Reports* 6.1 (Oct. 2016), p. 35137.

[21] Diederik P. Kingma and Prafulla Dhariwal. *Glow: Generative Flow with Invertible 1x1 Convolutions*. 2018. arXiv: 1807.03039 [stat.ML]. URL: https://arxiv.org/abs/1807.03039.

[22] Thomas Kirchner et al. “Signed Real-Time Delay Multiply and Sum Beamforming for Multispectral Photoacoustic Imaging”. In: *Journal of Imaging* 4.10 (2018). ISSN: 2313-433X. DOI: 10.3390/jimaging4100121. URL: https://www.mdpi.com/2313-433X/4/10/121.

[23] Gitta Kutyniok, Wang-Q Lim, and Rafael Reisenhofer. “ShearLab 3D: Faithful Digital Shearlet Transforms Based on Compactly Supported Shearlets”. In: *ACM Trans. Math. Softw.* 42.1 (Jan. 2016). ISSN: 0098-3500. DOI: 10.1145/2740960. URL: https://doi.org/10.1145/2740960.

[24] Meng-Lin Li et al. “Simultaneous Molecular and Hypoxia Imaging of Brain Tumors In Vivo Using Spectroscopic Photoacoustic Tomography”. In: *Proceedings of the IEEE* 96.3 (2008), pp. 481–489. DOI: 10.1109/JPROC.2007.913515.

[25] Elaine Marieb and Katja Hoehn. *Human Anatomy & Physiology Global Edition*. Pearson Deutschland, 2022, p. 632. ISBN: 9781292421803. URL: https://elibrary.pearson.de/book/99.150005/9781292421780.

[26] Giulia Matrone et al. “The Delay Multiply and Sum Beamforming Algorithm in Ultrasound B-Mode Medical Imaging”. In: *IEEE Transactions on Medical Imaging* 34.4 (2015), pp. 940–949. DOI: 10.1109/TMI.2014.2371235.

[27] Zsolt Molnar and Marton Nemeth. “Monitoring of Tissue Oxygenation: an Everyday Clinical Challenge”. en. In: *Front Med (Lausanne)* 4 (Jan. 2018), p. 247.

[28] Jan-Hinrich Nolke et al. “Photoacoustic Quantification of Tissue Oxygenation Using Conditional Invertible Neural Networks”. en. In: *IEEE Trans Med Imaging* 43.9 (Sept. 2024), pp. 3366–3376.

[29] Jeongwoo Park et al. “Clinical translation of photoacoustic imaging”. In: *Nature Reviews Bioengineering* (Sept. 2024).

[30] Jonas J. M. Riksen, Anton V. Nikolaev, and Gijs van Soest. “Photoacoustic imaging on its way toward clinical utility: a tutorial review focusing on practical application in medicine”. In: *Journal of Biomedical Optics* 28.12 (2023), p. 121205. DOI: 10.1117/1.JBO.28.12.121205. URL: https://doi.org/10.1117/1.JBO.28.12.121205.

[31] Andrew M Smith, Michael C Mancini, and Shuming Nie. “Second window for in vivo imaging”. In: *Nature Nanotechnology* 4.11 (Nov. 2009), pp. 710–711.

[32] Tanja Tarvainen and Ben Cox. “Quantitative photoacoustic tomography: modeling and inverse problems”. In: *Journal of Biomedical Optics* 29.S1 (2023), S11509. DOI: 10.1117/1.JBO.29.S1.S11509. URL: https://doi.org/10.1117/1.JBO.29.S1.S11509.

[33] Kevin Thomas and Jason Peeler. “A Detailed Anatomical Description of the Gastrocnemius Muscle — Is It Anatomically Positioned to Function as an Antagonist to the Anterior Cruciate Ligament?” In: *Anatomia* 3.4 (2024), pp. 244–255. ISSN: 2813-0545. DOI: 10.3390/anatomia3040021. URL: https://www.mdpi.com/2813-0545/3/4/21.

[34] Stratis Tzoumas and Vasilis Ntziachristos. “Spectral unmixing techniques for optoacoustic imaging of tissue pathophysiology”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 375.2107 (2017), p. 20170262. DOI: 10.1098/rsta.2017.0262. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2017.0262. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2017.0262.

[35] Stratis Tzoumas et al. “Eigenspectra optoacoustic tomography achieves quantitative blood oxygenation imaging deep in tissues”. In: *Nature Communications* 7.1 (June 2016), p. 12121.

[36] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[37] M.L. Veingerov. “New methods of gas analysis based on Tyndall-Roentgen Opto-acoustic effect”. In: *Dokl. Akad. Nauk SSSR.* 19 (1938), p. 687.

[38] Stephen J. Wright, Robert D. Nowak, and MÁrio A. T. Figueiredo. “Sparse Reconstruction by Separable Approximation”. In: *IEEE Transactions on Signal Processing* 57.7 (2009), pp. 2479–2493. DOI: 10.1109/TSP.2009.2016892.

[39] Shuangyang Zhang et al. “Pixel-wise reconstruction of tissue absorption coefficients in photoacoustic tomography using a non-segmentation iterative method”. In: *Photoacoustics* 28 (2022), p. 100390. ISSN: 2213-5979. DOI: https://doi.org/10.1016/j.pacs.2022.100390. URL: https://www.sciencedirect.com/science/article/pii/S2213597922000556.

[40] Timo Zimmermann, Jens Rietdorf, and Rainer Pepperkok. “Spectral imaging and its applications in live cell microscopy”. In: *FEBS Letters* 546.1 (2003), pp. 87–92. DOI: https://doi.org/10.1016/S0014-5793(03)00521-0. eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1016/S0014-5793%2803%2900521-0. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1016/S0014-5793%2803%2900521-0.

Nomenclature

Abbreviations

(M)AE	(Mean) Absolute Error
(M)CE	Mean Calibration Error
(m)IQR	(Mean) Inter-Quartile Range
BMI	Body Mass Index
BS	Batch Size
cINN	conditional Invertible Neural Network
DAS	Delay-and-Sum
DA	Diffusion Approximation
Deep MB	Deep Model-Based (reconstruction)
DL	Deep Learning
DMAS	Delay-Multiply-and-Sum
FCNN	Fully-Connected Neural Network
FOV	Field Of View
GPU	Graphical Processing Unit
GT	Ground Truth
HbO_2	Oxyhemoglobin
Hb	Deoxyhemoglobin
INN	Invertible Neural Network
INN	Invertible Neural Network
IR	Infrared
LMM	Linear Mixture Model
LR	Learning Rate
LSD	Learned Spectral Decoloring
LSTM	Long short-term memory
LU	Linear Unmixing

MAP	Maximum A posteriori Probability
MB rec	Model-Based reconstruction
MCX	Monte Carlo eXtreme
MedAE	Median Absolute Error
MSOT	Multispectral Optoacoustic Tomography
NIR	Near infrared
NLP	Natural Language Processing
o.o.d.	Out-of-distribution
PAD	Peripheral Artery Disease
PAI	Photoacoustic Imaging
PAI	Photoacoustic Imaging
PAT	Photoacoustic Tomography
PA	Photoacoustic
PCA	Principal Component Analysis
PE	Positional Encoding
PE	Positional Encoding
pp	Percentage Points
qPAT	Quantitative Photoacoustic Tomography
ROI	Region of Interest
RTE	Radiative Transfer Equation
SBDC	Segmentation-Based Direct Correction
SBIC	Segmentation-Based Iterative Correction
sDMAS	signed Delay-Multiply-and-Sum
SIMPA	Simulation and Image Processing for Photonics and Acoustics
SL	Supervised Loss
SNR	Signal-to-Noise Ratio
sO_2	Blood oxygen saturation
SoS	Speed of Sound

SPE	Sinusoidal Positional Encoding
USI	Ultrasound Imaging
USL	Unsupervised Loss
US	Ultrasounds
WD	Weight Decay

Appendix

A Detailed simulation parameters

A.1 Set-up

Fig.40 lists all the fixed parameters regarding experimental set-up (cf 4.1). Everything was verified compliant to the scenario of a common PAD scan using MSOT Acuity Echo.

Experimental set-up parameters			
Parameter	**Symbol**	**Value(s)**	**Unit**
Spatial resolution	Δx	100 (same for all simulations)	µm
Optical simulation			
Numerical model		Monte Carlo eXtreme (MCX) : MC accelerated on GPU, 3D	
Wavelengths	λ	$[700, 730, 760, 800, 850, 900]$	nm
Number of photons	$N_{photons}$	10^7	
Laser pulse energy	E_{laser}	$f(\lambda) = [11.25, 12.54, 10.59, 11.04, 9.93, 9.75]$	mJ
Post processing noise on initial pressure		Ø	
Acoustic simulation			
Numerical model		kWave, accelerated in C++, 3D	
Sampling frequency	$f_{sampling}$	40	MHz
Post processing noise on sinogram		Ø	
Number of sinogram samples	$N_{samples}$	2030	
Acoustic reconstruction (2D)			
Numerical model		Model-Based reconstruction (MB rec)	
Speed of sound	c_s	1540 (assumed uniform)	m/s
Device (MSOT Acuity Echo)			
US gel thickness	$H_{US\,gel}$	0	mm
Detection			
Probe radius	R_{probe}	40	mm
Probe height (including mediprene)	Δz_{probe}	43.2	mm
Mediprene membrane height	$\Delta z_{mediprene}$	1	mm
Focus of probe in imaging plane	Δz_{focus}	8	mm
Number of detector elements	$N_{detectors}$	256	
Detector element width	$w_{detector}$	0.24	mm
Detector element length	$l_{detector}$	13	mm
Pitch	$\Delta x_{detectors}$	0.34	mm
Central frequency	f	3.96	MHz
Bandwidth	Δf	153	%
Illumination			
Angle of laser to the plane	α_{laser}	22.4	°
Focus of laser in imaging plane	Δz_{laser}	2.8	mm
Width of slit illumination	w_{slit}	30.0	mm
Divergence angle (FWHM)	α_{FWHM}	8.66	°

Legend : input | default | function of input variables

Figure 40. *Simulation parameters*

A.2 Tissue digital twin

Fig.41 lists all the fixed parameters regarding tissue digital twin simulation. The parameters in blue vary as a function of the inputs in green, which are the parameters that create variability between simulations in the dataset.

Tissue digital twin parameters

		Composition (molecules)	Absorption coeff.	Scattering coeff.	Anisotropy	Speed of sound	Blood oxygen saturation	Blood volume fraction	Water volume fraction	Melanin volume fraction
Symbol			μ_a	μ_s	g	c_s	sO_2	x_{blood}	x_{water}	$x_{melanin}$
Unit						m/s	%	%	%	%
Anatomical structures	Background	arbitrary	10^{-4}	10^{-4}	0.9	1540	N/A	N/A	N/A	N/A
	Interstitial (≠ soft) tissue	Hb, HbO_2, muscle scatterer + "custom water"	$f(\mu_{a_{Hb}}; \mu_{a_{HbO_2}})$	$f(\mu_{s_{blood}}; \mu_{s_{muscle}})$	$f(g_{blood} = 0.98; g_{muscle} = 0.9; g_{water\ custom} = 0.895)$	$f(c_{s_{blood}}; c_{s_{muscle}} = 1540; c_{s_{water\ custom}} = 1604.4)$	50	1	68	N/A
	Water	water	$\mu_{a_{water}}$	10^{-10}	1.0	$c_{s_{water}} = 1482.3$	N/A	N/A	100	N/A
	Mediprene	mediprene	1.6×10^{-2}	1.5×10^{-1}	0.9	1540	N/A	N/A	N/A	N/A
	Acoustic couplant	Zerdine® hydrogel	8.0×10^{-4}	10^{-10}	1.0	1540	N/A	N/A	N/A	N/A
	Epidermis	melanin, epidermal scatterer	$f(\mu_{a_{melanin}})$	$f(\mu_{s_{epidermal}})$	$g_{epidermal}$	$c_{s_{skin}} = 1624$	N/A	N/A	N/A	1 - 16
	Dermis	Hb, HbO_2, dermal scatterer	$f(\mu_{a_{Hb}}; \mu_{a_{HbO_2}}; \mu_{a_{skin\ baseline}} \approx 0)$	$f(\mu_{s_{blood}}; \mu_{s_{dermal}})$	$f(g_{blood} = 0.98; g_{dermal} = 0.715)$	$f(c_{s_{blood}}; c_{s_{skin}})$	50	0.2	N/A	N/A
	Hypodermis	Hb, HbO_2, fat, soft tissue scatterer, water	$f(\mu_{a_{Hb}}; \mu_{a_{HbO_2}}; \mu_{a_{fat}} \approx 0; \mu_{a_{water}} \approx 0)$	$f(\mu_{s_{blood}}; \mu_{s_{fat}}; \mu_{s_{soft\ tissue}}; \mu_{s_{water}})$	$f(g_{blood} = 0.98; g_{fat} = 0.9; g_{soft\ tissue} = 0.9; g_{water})$	$f(c_{s_{blood}}; c_{s_{fat}} = 1440.2; c_{s_{water}}; c_{s_{soft\ tissue}} = 1540)$	50	1	68	N/A
	Muscle	Hb, HbO_2, muscle scatterer + "custom water"	$f(\mu_{a_{Hb}}; \mu_{a_{HbO_2}})$	$f(\mu_{s_{blood}}; \mu_{s_{muscle}})$	$f(g_{blood} = 0.98; g_{muscle} = 0.9; g_{water\ custom} = 0.895)$	$f(c_{s_{blood}}; c_{s_{muscle}}; c_{s_{water\ custom}})$	50 - 80	10 - 30	68	N/A
	Bone	bone, water	$f(\mu_{a_{water}} \approx 0; \mu_{a_{bone}} = 1.8)$	$f(\mu_{s_{bone}}; \mu_{s_{water}})$	$f(g_{bone} = 0.9; g_{water})$	$f(c_{s_{water}}; c_{s_{bone}} = 3514.9)$	N/A	N/A	18 - 20	N/A
	Vein	blood	$f(\mu_{a_{Hb}}; \mu_{a_{HbO_2}})$	$f(\mu_{s_{blood}})$	0.98	$c_{s_{blood}} = 1578.2$	60 - 80	100	N/A	N/A
	Artery	blood	$f(\mu_{a_{Hb}}; \mu_{a_{HbO_2}})$	$f(\mu_{s_{blood}})$	0.98	$c_{s_{blood}} = 1578.2$	90 – 100	100	N/A	N/A

Legend : input | default | function of input variables

Figure 41. *Tissue digital twin parameters*

Some μ_a, μ_s and g spectra are wavelength dependent, as shown in Fig.42.

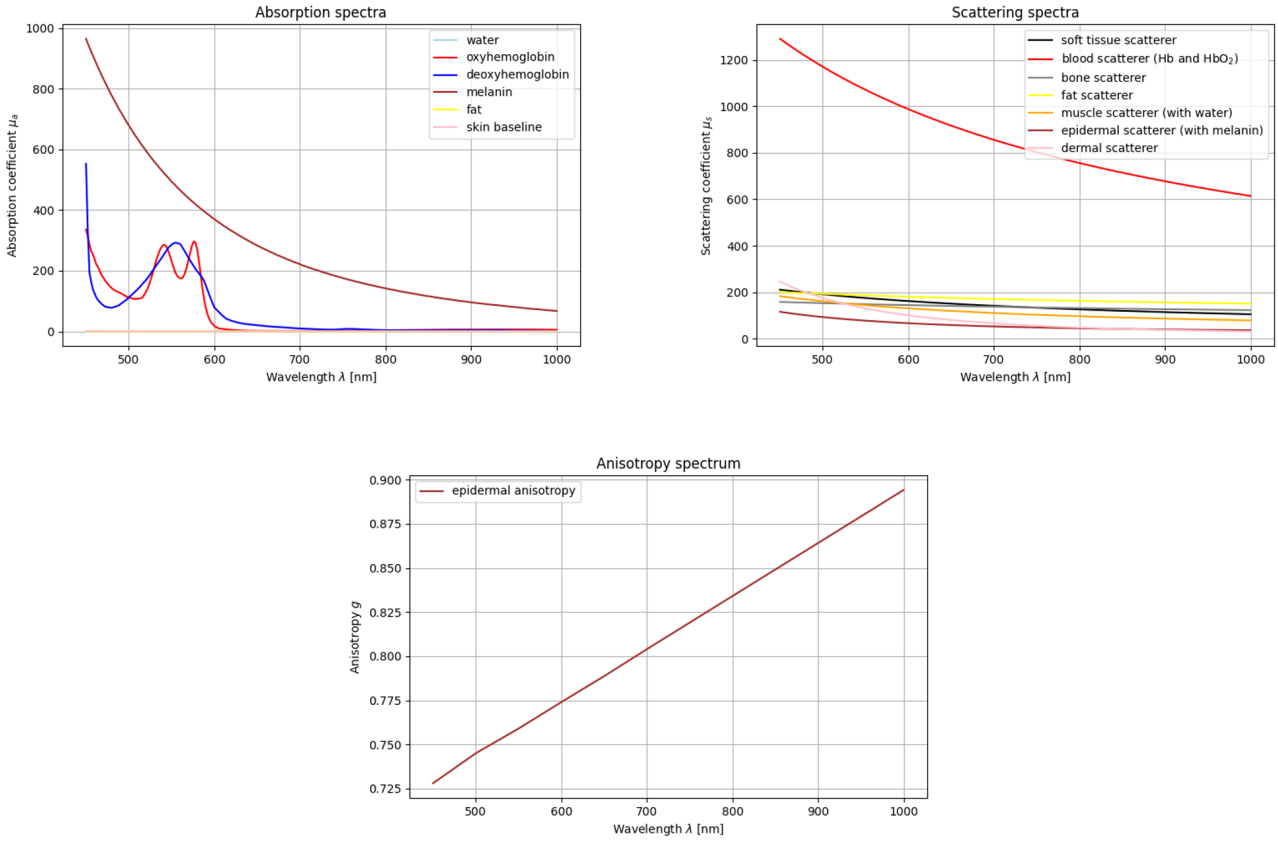


Figure 42. *Tissue absorption, scattering and anisotropy spectra as defined in SIMPA [13]*

B Training

B.1 Training metrics with outliers

The following figures are versions of training curves without using ML Flow's functionality to remove outliers. Fig.43 shows training curves for the three cINN baselines. Only one instability is noticed for the sin 2cINN, which we interpret as normal because cINNs are usually unstable during training.

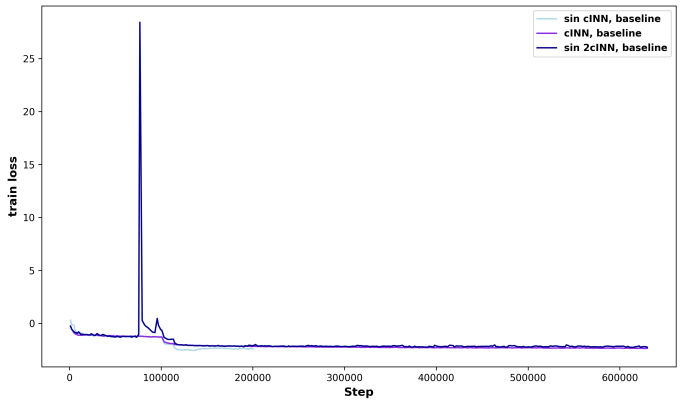


Figure 43. *Training loss for the cINNs trained with outliers*

Fig.44 shows the training and validation losses for cINNs on the smaller ROI. We observe strong instabilities attributed to the "difficulty" of pixels in the smaller ROI.

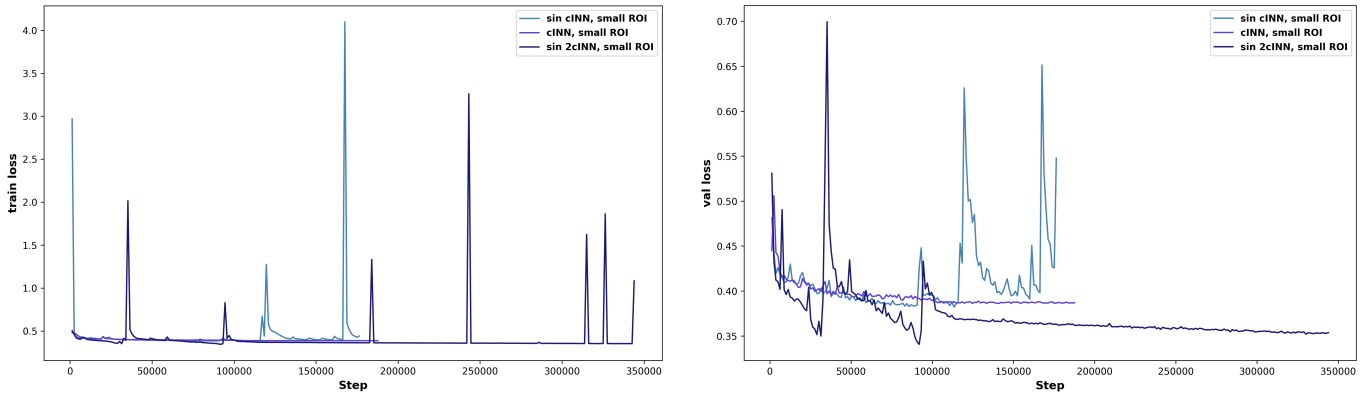


Figure 44. *Training curves for the cINNs trained on the reduced ROI with outliers*

B.2 MCE

In Fig.45 are plotted the MCE for the comparison of the baseline to models with different schedulers (left) and the model with 10 blocks (right). As you can see, the curves are very erratic, which raises the question of using test MCE as a factor for decision making.

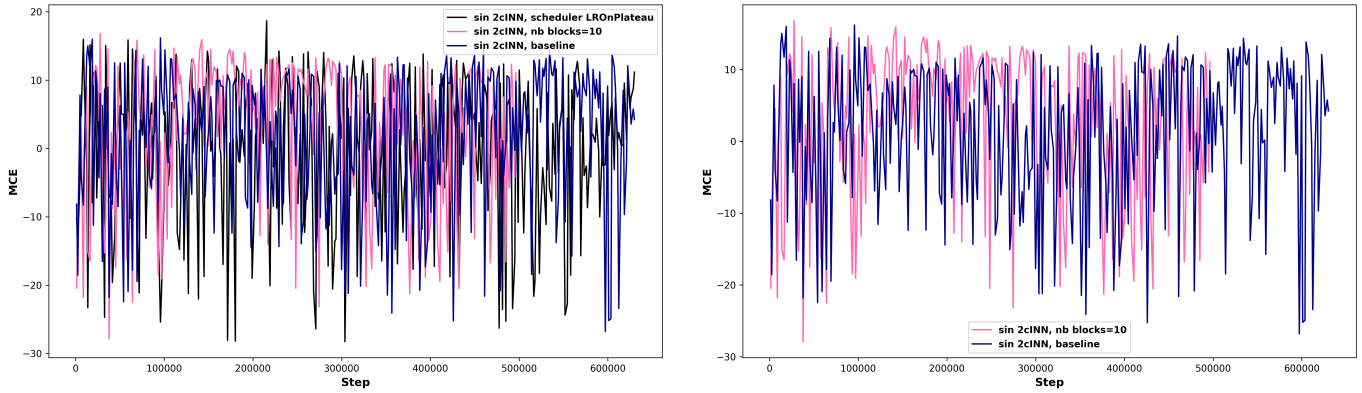


Figure 45. *MCE during training in two different cases*

C Test subset analysis

The distribution of MAEs without filtering the skin layer pixels is shown in Fig.46. The aspect of the distribution was not very useful for analysis.

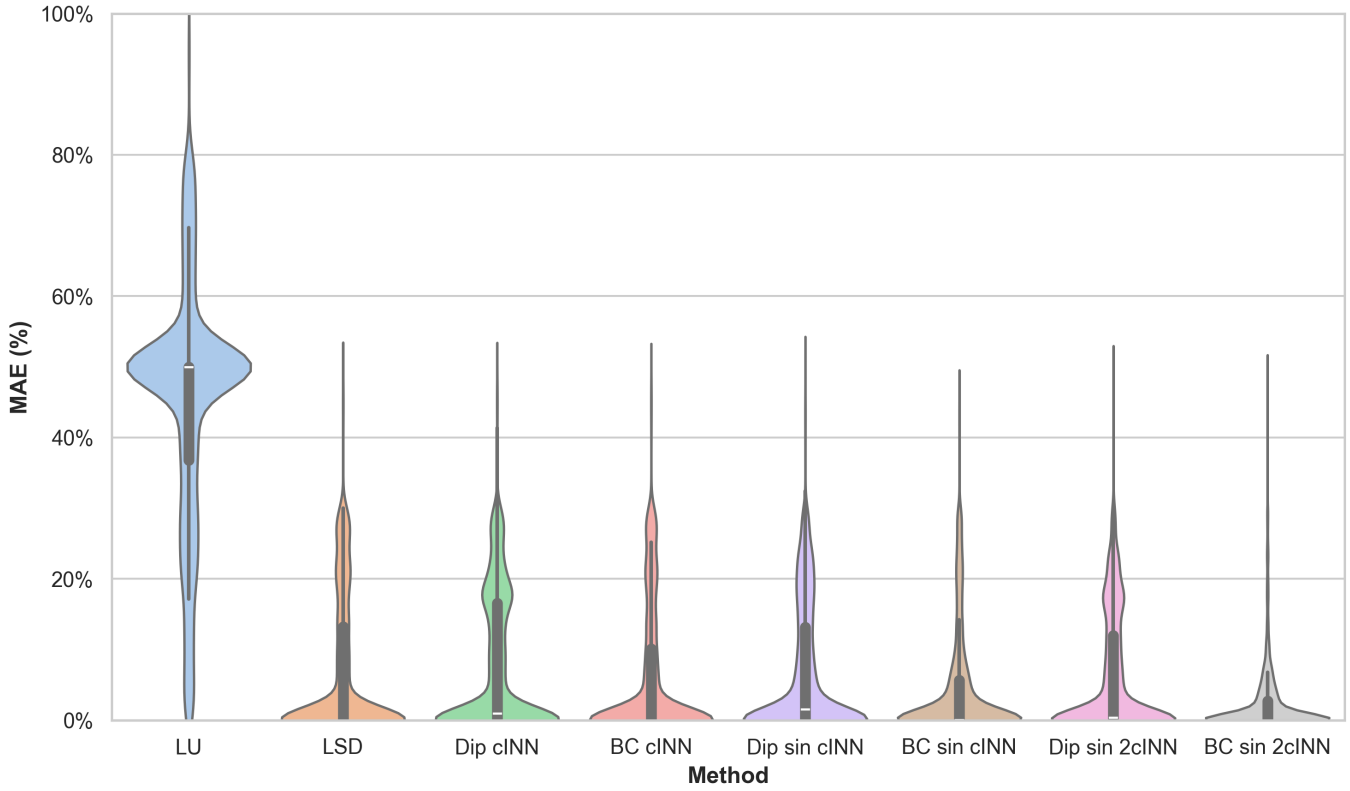


Figure 46. *MAE distributions on the test set on 5000 random pixels from the test set, all the layers*

D Image predictions

The remaining comparisons of LSD, LU and cINN predictions on the test images compared to the ground truth sO_2 is shown in Fig.47. No more analysis can be done on these images.

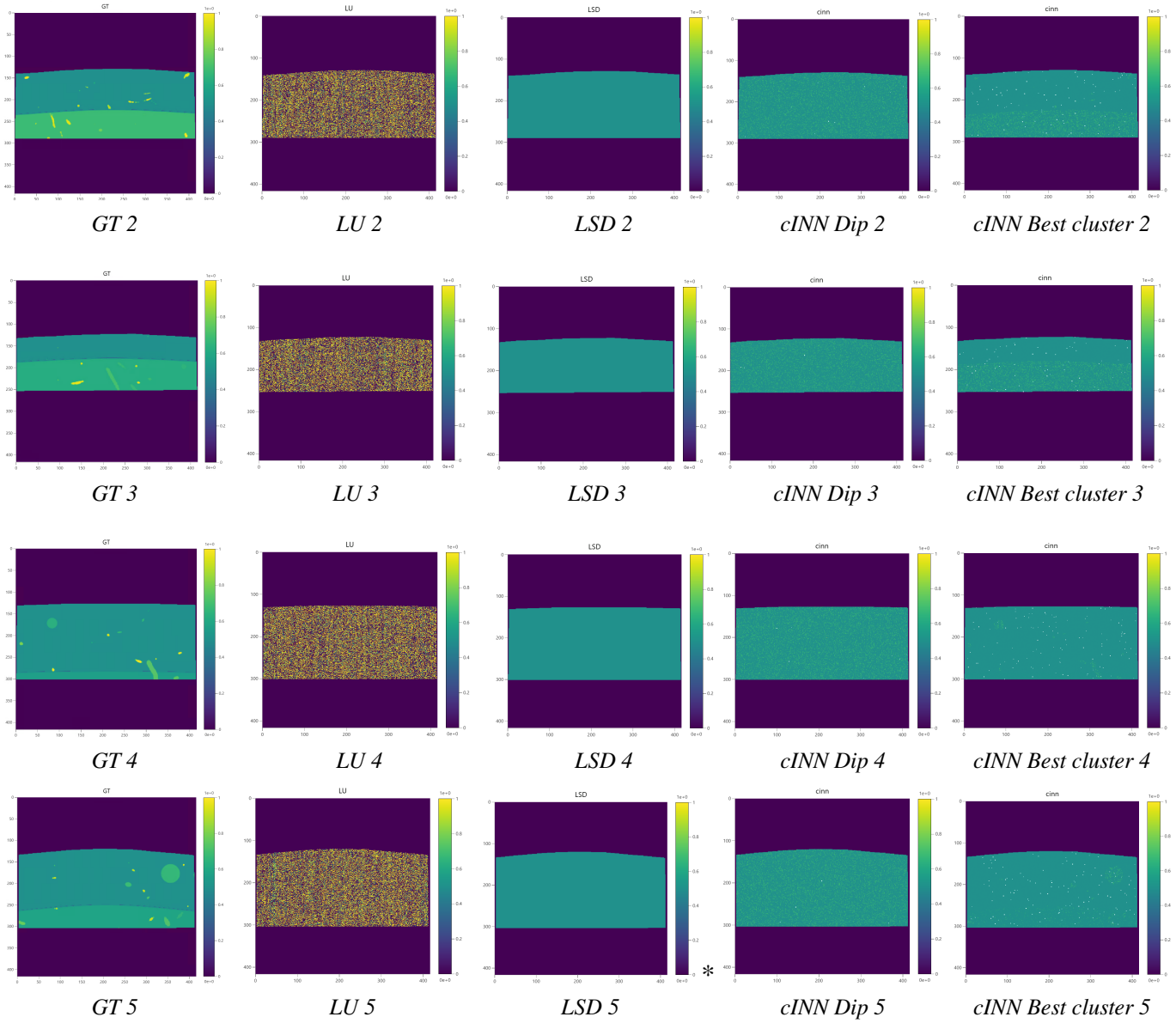


Figure 47. *Comparison between GT sO_2, LU, LSD and cINN on the test images 2 to 5*

Review

Content presentation created by:

Place, date	*Place, date*	*Place, date*
Munich, 19/05/2025		
		
Student	Supervising chair	Center for Key Competencies