DeepMB: Deep neural network for real-time model-based optoacoustic image reconstruction with adjustable speed of sound

Christoph Dehner1,2,*, Guillaume Zahnd1,3,*, Vasilis Ntziachristos1,2,4,*, Dominik Jüstel1,2,5,+

1Institute of Biological and Medical Imaging, Helmholtz Zentrum München, Neuherberg, Germany
2Chair of Biological Imaging at the Central Institute for Translational Cancer Research (TranslaTUM), School of Medicine, Technical University of Munich, Germany
3iThera Medical GmbH, Munich, Germany
4Munich Institute of Robotics and Machine Intelligence, Technical University of Munich, Germany
5Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany

*These authors contributed equally to this work.
+These authors share the corresponding authorship.

Abstract: Multispectral optoacoustic tomography (MSOT) is a high-resolution functional imaging modality that can non-invasively access a broad range of pathophysiological phenomena by quantifying the contrast of endogenous chromophores in tissue. Real-time imaging is imperative to translate MSOT into clinical imaging, visualize dynamic pathophysiological changes associated with disease progression, and enable in situ diagnoses. Model-based reconstruction affords state-of-the-art optoacoustic images; however, the advanced image quality provided by model-based reconstruction remains inaccessible during real-time imaging because the algorithm is iterative and computationally demanding. Deep-learning may afford faster reconstructions for real-time optoacoustic imaging, but existing approaches only support oversimplified imaging settings and fail to generalize to in vivo data. In this work, we introduce a novel deep-learning framework, termed DeepMB, to learn the model-based reconstruction operator and infer optoacoustic images with state-of-the-art quality in less than 10 ms per image. DeepMB accurately generalizes to in vivo data after training on synthesized sinograms that are derived from real-world images. The framework affords in-focus images for a broad range of anatomical locations with various acoustic properties because it supports dynamic adjustment of the reconstruction speed of sound during imaging. Furthermore, DeepMB is compatible with the data rates and image sizes of modern multispectral optoacoustic tomography scanners, thus enabling straightforward adoption into clinical routine. We evaluate DeepMB both qualitatively and quantitatively on a diverse dataset of in vivo images and demonstrate that the framework reconstructs images 3000 times faster than the iterative model-based reference method while affording near-identical image qualities. Accurate and real-time image reconstructions with DeepMB can enable full access to the high-resolution and multispectral contrast of handheld optoacoustic tomography in deep tissue, thus facilitating advanced dynamic imaging applications.

1. Introduction

Multispectral optoacoustic tomography (MSOT) is a ground-breaking functional imaging modality because it uniquely enables non-invasive detection of optical contrast at high spatial resolution and centimeter-scale penetration depth in living tissue1–6. Accessing the multispectral contrast of endogenous chromophores, MSOT can quantify a broad range of pathophysiological surrogate biomarkers such as tissue fibrosis, inflammation, vascularization, and oxygenation, and provide unmatched clinical information for multifarious diseases such as breast cancer2, 6, Duchenne muscular dystrophy7, or inflammatory bowel disease3.

In order to fully translate and integrate MSOT into clinical imaging, real-time application is imperative8–10. Handheld MSOT imaging requires — similar to ultrasound imaging — live image feedback at sufficiently high frame-rates (at least 24 fps for full-video rendering) to avoid hindering visuo-tactile coordination, identify and localize relevant tissue structures using anatomical landmarks in their surroundings, and find the optimal transducer pose for the target region. Furthermore, real-time optoacoustic imaging is necessary to visualize dynamic pathophysiological changes associated with disease progression and enable in situ guidance and diagnosis during intra-operative and endoscopy imaging11,12. In practice, real-time reconstruction of optoacoustic images (i.e., recovery of the initial pressure distribution in the imaged tissue) is generally conducted via the backprojection algorithm13. However, the backprojection formula is based on over-simplified modelling assumptions of the imaging
process and cannot compensate for the ill-posedness of the underlying inverse problem arising from limited-angle acquisition, measurement noise, and finite transducer bandwidth. Consequently, backprojection images systematically suffer from low spatial resolution and contrast, as well as negative pixel values that invalidate a physical interpretation of the image as initial pressure distribution. In contrast, iterative model-based reconstruction\cite{14,15} can provide accurate optoacoustic images with state-of-the-art quality by incorporating a physical model of the imaging device into the reconstruction process, constraining the reconstructed image to be non-negative, and introducing regularization to mitigate the ill-posedness of the inversion problem. However, model-based reconstruction is computationally demanding because of the iterative (and thus sequential) nature of the algorithm, which is prohibitive for real-time imaging. Therefore, the full imaging potential of MSOT is only available offline after considerable computational time and currently remains inaccessible for clinical applications that require live image feedback.

Real-time model-based reconstruction has been demonstrated for a pre-clinical MSOT system by computing the reconstruction on a graphics processing unit (GPU)\cite{16}. However, a similar acceleration is infeasible for state-of-the-art model-based reconstruction of data from modern clinical systems because these reconstructions are substantially more computationally demanding (larger images, more complex regularization functionals, inclusion of the total impulse response of the system in the model, necessity of a higher number of iterations until convergence\cite{14,15}). Recently, various deep-learning methods demonstrated a partial proof-of-concept for real-time optoacoustic image reconstruction by utilizing the computational efficiency of deep neural networks to directly infer high-quality optoacoustic images from recorded signals\cite{17-22} or speed up iterative model-based reconstruction\cite{23-25}. However, the aforementioned methods are unable to comprehensively generalize from training data (which can be in vivo, synthetic, or a combination of both) to experimental input data and have either not been evaluated on in vivo data\cite{19,21} or have afforded only rudimentary image quality when applied to in vivo scans\cite{17,18,23}. Furthermore, all previously reported deep-learning-based reconstruction methods have inherent limitations that impede their usage in clinical applications. First, they cannot dynamically adjust the speed of sound (SoS) during imaging — a necessity when confronted with different tissue types. Second, they build upon physically-oversimplified and sub-optimal imaging setups (scanners with only 30-128 transducers instead of at least 256\cite{17-22}, image sizes of only 128×128 pixels instead of approximately 400×400 pixels\cite{18-22}). These simplifications alleviate the learning process but result in a low image quality ceiling that severely reduces the applicability to state-of-the-art clinical MSOT systems.

We introduce a novel deep-learning model-based optoacoustic reconstruction framework (DeepMB), which affords image quality nearly-indistinguishable from state-of-the-art iterative model-based reconstructions at speeds enabling live imaging (100 fps or <10 ms/image, versus 30-60 s/image for iterative model-based reconstruction). DeepMB overcomes the shortcomings of previous deep-learning reconstruction approaches to generalize from synthetic training data to experimental test data by training on optoacoustic signals that are synthesized from a publicly available dataset of real-world images\cite{26}, while using as ground-truth the optoacoustic images generated via model-based reconstruction of the corresponding signals. This training scheme enables DeepMB to learn an accurate and universally applicable model-based optoacoustic reconstruction operator. Critically, DeepMB supports dynamic adjustments of the SoS parameter during imaging, which enables the reconstruction of in-focus images for arbitrary tissue types. In contrast to other attempts at applying deep-learning to optoacoustic reconstruction\cite{17-22}, DeepMB is directly compatible with state-of-the-art clinical MSOT scanners because it supports high throughput data acquisition (sampling rate: 40 MHz; number of transducers: 256) and large image sizes (416×416 pixels). We demonstrate the performance of DeepMB both quantitatively and qualitatively on a diverse dataset of in vivo images (4814 images, 6 participants, 25-29 scanned locations per participant). With DeepMB, clinical MSOT could provide high quality feedback during live imaging and thus facilitate advanced dynamic imaging applications.

2. Results

To validate the capability of DeepMB to reconstruct images in real-time and with adjustable SoS, the framework was applied to a modern handheld optoacoustic scanner (MSOT Acuity Echo, iThera
Medical GmbH, Munich, Germany) with SoS values ranging from 1475 m/s to 1525 m/s in steps of 5 m/s.

**DeepMB pipeline**

Fig. 1 illustrates the overall training and evaluation pipeline. DeepMB was trained using input sinograms synthesized from general-feature images to facilitate the learning of an unbiased and universally applicable reconstruction operator. These sinograms were generated by employing a diverse collection of publicly available real-world images as initial pressure distributions and simulating thereof the signals recorded by the acoustic transducers with an accurate physical model of the considered scanner \(^1\) (Fig. 1a and section “Methods / Synthesis of sinograms for training and validation”). The SoS values for the forward simulations were drawn uniformly at random from the considered range for each image. Ground-truth images for the synthesized sinograms were computed via model-based reconstruction (Fig. 1c). Fig. 1d shows the deep neural network architecture of DeepMB, which inputs a sinogram (either synthetic or in vivo) and a SoS value and outputs the final reconstructed image. The underlying design is based on the U-Net architecture augmented with two extensions that promote the network to learn and express the effects of the different input SoS values onto the reconstructed images: (1) all signals were mapped from the input sinogram to the image domain with a linear delay operator based on the given input SoS value (no trainable weights), and (2) the input SoS value (one-hot encoded and concatenated as additional channels) was passed to the trainable convolutional layers of the network. A detailed description of the network training is given in the section “Methods / Network training”. After training, the applicability of DeepMB to clinical data was tested with a diverse dataset of in vivo sinograms acquired by scanning six participants at up to eight anatomical locations each (Fig. 1b). The corresponding ground-truth images of the acquired in vivo test sinograms were obtained analogously to the training data via model-based reconstruction. The inference time of DeepMB was less than 10 ms per sample on a modern GPU (GeForce RTX 3090).

---

Figure 1: DeepMB pipeline. (a) Real-world images, obtained from a publicly available dataset, are used to generate synthetic sinograms by applying an accurate physical forward model of the scanner. SoS, speed-of-sound. (b) In vivo sinograms are acquired from diverse anatomical locations in six participants. (c) Optoacoustic images are reconstructed via iterative model-based reconstruction for the purpose of generating reference images for either the synthetic dataset (A) or the in vivo dataset (B). (d) Network training is conducted by using the synthetic data as sets for training (n=8000) and validation (n=2000) (C), while the in vivo data constitutes the test set (n=4814) (D). A domain transformation is first applied to the input sinograms via a delay operation to map the time samples values into the image space. The SoS is then one-hot encoded and concatenated as additional channels. A U-Net convolutional neural network is subsequently applied to the channel stack to regress the final image. The loss is calculated between the network output and the corresponding reference image (see section “Methods / Network training” for further details about the network training).
Qualitative evaluation

DeepMB successfully reconstructed high-quality optoacoustic images. To qualitatively evaluate DeepMB, all DeepMB images from the in vivo dataset (Fig. 1b) were thoroughly compared to their corresponding model-based reference images (Fig. 1c). Fig. 2 shows four representative reconstructed images, corresponding to scans of the carotid artery, biceps, breast, and abdomen. DeepMB reconstructions (Fig. 2a–d) are systematically nearly-indistinguishable from the model-based references (Fig. 2e–h), with no noticeable failures, outliers, or artifacts for any of the participants, anatomies, probe orientations, SoS values, or laser wavelengths. The similarity between the DeepMB and model-based images is also confirmed by their negligible pixel-wise absolute differences (Fig. 2i–l). The zoomed region D in Fig. 2j depicts one of the largest observed discrepancies between the DeepMB and model-based reconstructions, which manifests as minor blurring, showing that the DeepMB image is only marginally affected by these errors. In comparison, backprojection images (Fig. 2m–p) exhibit notable differences from the reference model-based images and suffer from reduced spatial resolution and physically-nonsensical negative initial pressure values. Finally, to facilitate relating the reconstructed optoacoustic images to the scanned anatomies, Fig. 2q–t depict schematic sketches of the anatomical context for all scans, while Fig. 2u–x, depict the interleaved-acquired ultrasound images overlayed with the temporally-corresponding DeepMB reconstructions.

Extended Data Videos 1-2 further illustrate the real-time optoacoustic imaging capabilities of DeepMB. Extended Data Video 1 shows a carotid artery continuously imaged in the transversal view at 800 nm, which demonstrates that DeepMB can be used to visualize motion at 25 Hz with state-of-the-art image quality. Extended Data Video 2 shows the optoacoustic image of a biceps in the transversal view at 800 nm while the SoS is gradually adjusted via a series of DeepMB reconstructions, which illustrates the importance of on-the-fly SoS tuning for optimal image quality.
Figure 2: Representative examples from the in vivo test dataset for different anatomical locations (carotid artery: a,e,i,m,q,u; biceps: b,f,j,n,r,v; breast: c,g,k,o,s,w; abdomen: d,h,l,p,t,x). The first four rows show deep model-based (DeepMB) reconstructions, model-based (MB) reconstructions, the pixel-wise absolute difference between DeepMB and MB reconstructions, and backprojection (BP) reconstructions, respectively. The last two rows display sketches of the anatomical context of the scans and the interleaved-acquired ultrasound (US) images overlayed with DeepMB reconstructions, respectively. All images were slightly cropped to a field of view of 4.16×2.80 cm² to disregard the space above the skin. Each enlarged region is 0.41×0.41 cm² and displays various anatomical details. All displayed scans were acquired at 800 nm. Mb: probe membrane, Sk: skin, Mu: muscle, Fa: fascia, Ca: common carotid artery, Ju: jugular vein, Th: thyroid, Tr: trachea, Ve: blood vessel, Ne: nerve, Ft: fat, Gl: glandular tissues, Co: colon.
Quantitative evaluation

The ability of DeepMB to reconstruct images with equivalent fidelity to those afforded by model-based reconstruction was then confirmed by quantitative comparison. To quantify the image fidelity of DeepMB reconstructions, the data residual norm was calculated for all in vivo test images (see section “Methods / Data residual norm” for the precise definition). The data residual norm measures the fidelity of a reconstructed image by computing the mismatch between the image and the corresponding recorded acoustic signals with regard to the accurate physical forward model of the used scanner, and is provably minimal for model-based reconstruction\(^\text{27}\). The data residual norm was also calculated for all model-based and backprojection reconstructions, for comparison purposes.

First, data residual norms were calculated with in-focus images (that is, reconstructed with optimal SoS values) to evaluate the fidelity of DeepMB images with the best possible quality (Fig. 3a). Data residual norms of DeepMB images (green, mean±std = 0.156±0.088) are almost as low as the provably minimal data residual norms of model-based images (blue, mean±std = 0.139±0.095). The close agreement between data residual norms of DeepMB and model-based images confirms that both reconstruction approaches afford equivalent image qualities. In contrast, the data residual norms of backprojection images are markedly higher (gray, mean±std = 0.369±0.098), which reaffirms the shortcomings of backprojection to accurately model the imaging process, and explains the lower image quality observed in Fig. 2d,h,l,p. Table 1 summarizes the data residual norms of all reconstruction approaches evaluated in this paper.

Second, data residual norms were calculated with out-of-focus images (that is, reconstructed with sub-optimal SoS values) to evaluate the fidelity of DeepMB images during imaging applications with a priori unknown SoS (Fig. 3b, also see Table 1). Data residual norms of DeepMB images remain close to those of model-based images for all considered levels of mismatch between the optimal and the employed SoS, thus confirming that DeepMB and model-based images are similarly trustworthy independent of the selected SoS.

Advantages of one-hot-encoded SoS values

The most accurate DeepMB models were obtained when passing the one-hot-encoded input SoS value to the trainable layers of the network (as shown in Fig. 1d). To evaluate the benefits of this strategy, two other models with alternative SoS encoding schemes were trained and assessed: the first without providing the SoS to the U-Net (referred to as DeepMB\textsuperscript{no-sos}), and the second with the SoS encoded as a scalar value into one additional input channel for the U-Net (referred to as DeepMB\textsuperscript{scalar-sos}). The SoS

---

Figure 3: Data residual norms of optoacoustic images from deep model-based (DeepMB), model-based (MB), and backprojection (BP) reconstruction. (a) Data residual norms of in-focus images reconstructed with optimal speed of sound (SoS) values, on all 4814 samples from the in vivo test set. (b) Data residual norms of out-of-focus images reconstructed with sub-optimal SoS values, on a subset of 638 samples. The five sub-panels depict the effect of SoS mismatch via gradual increase of the offset \(\Delta\)SoS in steps of 10 m/s. The inner bars indicate the 25th, 50th, and 75th percentiles.
was in both models used to apply the delay operator before the trainable U-Net layers, analogously to the standard DeepMB model (see Fig. 1d). Not providing the SoS as input to the U-Net was found to be a marginally inferior alternative to the standard one-hot-based SoS encoding: DeepMB$_\text{so-sos}$ inferred high-quality and artifact-free images, however with slightly higher data residual norms than the standard DeepMB model (0.164±0.083, also see Table 1). Providing the SoS as scalar value to the U-Net was found to be a disadvantageous encoding scheme that impedes the ability of the neural network to learn an accurate reconstruction operator, because the overall brightness of images reconstructed with DeepMB$_\text{scalar-sos}$ was associated with the used SoS values. More specifically, inferring DeepMB$_\text{scalar-sos}$ onto the same sinogram with different input SoS values obtained images of higher average intensities for higher input SoS values. These intensity differences were visually imperceptible with default colormaps but resulted in notably higher data residual norms for the obtained images in comparison to DeepMB$_\text{so-sos}$ or the standard DeepMB model (0.169±0.081, also see Table 1).

**Advantages of synthesized training data**

Synthesized training data is essential for DeepMB to learn an accurate and general reconstruction operator. To demonstrate the advantages of synthesized training data for DeepMB, alternative DeepMB models were trained on in vivo data instead of real-world images (section “Methods / Network training”). These models, referred to as DeepMB$_\text{in-vivo}$ inferred images with similar data residual norms (0.155±0.093, also see Table 1) as the standard DeepMB model. However, DeepMB$_\text{in-vivo}$ images frequently contained visible artifacts, either at the left or right image borders, or in regions showing strong absorption at the skin surface. Extended Data Fig. 1 shows representative examples of such artifacts. No artifacts were observed with the standard training strategy of DeepMB, even when reducing the size of the synthetic training set from 8000 to 3500 to match the reduced amount of available in vivo training data.

**Advantages of model-based reference images**

Using model-based reference images as ground-truth targets during training is essential to learn a generalizable model-based reconstruction operator. To compare the training strategy of DeepMB to the training methodology reported in previous deep-learning-based reconstruction methods\(^\text{1722}\), another alternative model, referred to as DeepMB$_\text{initial-images}$, was trained using as ground truth targets the true synthetic initial pressure images (left side of Fig. 1a) instead of model-based reconstructions (right side of Fig. 1c). The reconstruction operator learnt by DeepMB$_\text{initial-images}$ was inferior in comparison to the standard DeepMB model: In vivo images reconstructed with DeepMB$_\text{initial-images}$ suffer from low resolution and contrast (see Extended Data Fig. 2) and have notably worse data residual norms (0.267±0.094, also see Table 1) than the standard DeepMB model.

|                         | Our method | Reference method | Traditional method | Alternative DeepMB training strategies |
|-------------------------|------------|------------------|--------------------|---------------------------------------|
| DeepMB                  |            | BP               |                    | **DeepMB$_\text{in-vivo}$**           |
| In-focus images         | 0.156±0.088| 0.139±0.095      | 0.369±0.098        | 0.164±0.083                           |
|                         |            |                  |                    | 0.169±0.081                           |
|                         |            |                  |                    | 0.155±0.093                           |
|                         |            |                  |                    | 0.267±0.094                           |
| Out-of-focus images     | 0.166±0.096| 0.149±0.104      | 0.365±0.103        | 0.176±0.090                           |
|                         |            |                  |                    | 0.181±0.087                           |
|                         |            |                  |                    | 0.164±0.103                           |
|                         |            |                  |                    | 0.275±0.109                           |

*All DeepMB$_\text{scalar-sos}$ images systematically have their overall brightness associated with the input SoS. **Some DeepMB$_\text{in-vivo}$ images suffer from visible reconstruction artifacts that manifest as coffee-stain-like structures (see Extended Data Fig. 1). ***All DeepMB$_\text{initial-images}$ images suffer from strong reconstruction artifacts that manifest as intensity saturation (see Extended Data Fig. 2).*

Table 1. Data residual norms (mean±std) of all seven reconstruction methods assessed in this paper for in-focus images (4814 in vivo sinograms reconstructed with each one optimal SoS values) and out-of-focus images (638 in vivo sinograms reconstructed each with all 11 available SoS values). SoS: speed of sound. DeepMB: deep model-
based. MB: model-based. BP: backprojection. DeepMB_{so-sos}: training conducted without providing the SoS as additional input to the U-Net. DeepMB_{scalar-sos}: training conducted with encoding the SoS value into one additional input channel for the U-Net. DeepMB_{initial-images}: training conducted on the true synthetic initial pressure images instead of the corresponding MB reconstructions. DeepMB_{in-vivo}: training conducted on in vivo data instead of synthetic data.

3. Discussion

MSOT is a high-resolution functional imaging modality that can non-invasively quantify a broad range of pathophysiological phenomena by accessing the endogenous contrast of chromophores in tissue. Model-based reconstruction affords optoacoustic images with state-of-the-art quality, but remains inaccessible during real-time imaging because of the iterative and computationally demanding nature of the algorithm. In this work, we introduce a deep neural network to learn the model-based reconstruction operator and infer images with qualities nearly-identical to model-based reconstruction in less than 10 ms per image. Our reconstruction framework, termed DeepMB, achieves three seminal features compared to previous approaches: Accurate generalization to in vivo measurements after training on synthesized sinograms that were derived from real-world images, dynamic adjustment of the reconstruction SoS during imaging, and compatibility with the data rates and image sizes of modern MSOT scanners. DeepMB will therefore enable dynamic-imaging applications of optoacoustic tomography and deliver high-quality images to clinicians in real-time during examinations, furthering the clinical translation of this technology and leading to more accurate diagnoses and surgical guidance.

We trained DeepMB on synthesized sinograms from real-world images instead of in vivo images because these synthesized sinograms afford a large training dataset with a versatile set of image features, allowing DeepMB to accurately reconstruct images with diverse features. In particular, such general-feature training datasets reduce the risk of encountering out-of-distribution samples (test data with features that are not contained in the training dataset) when applying the trained model to in vivo scans. In contrast, training a model on in vivo scans systematically introduces the risk of overfitting to specific characteristics of the training samples and can lead to decreased image quality for never-seen-before scans that may involve different anatomical views or disease states. We indeed observed that alternative DeepMB_{in-vivo} models trained on in vivo data failed to adequately generalize to some in vivo test scans and introduced artifacts within the reconstructed images (see Extended Data Fig. 1). Furthermore, using synthesized data instead of in vivo data alleviates the training of new DeepMB models because it obviates the need for recruiting and scanning a cohort of participants. Instead, training data can be automatically generated and used to straightforwardly obtain specifically-trained DeepMB models for new scanners or different reconstruction parameters.

Accurate generalization from synthesized training to in vivo test data is possible with DeepMB because the underlying inverse problem to solve (that is, regularized model-based reconstruction) is well-posed; for each input sinogram there is a unique and stable solution (i.e., the reconstructed image). Therefore, the network can learn a data transform that is agnostic to specific characteristics of the ground-truth images during training and generalizes to images with any content (be it synthesized or in vivo). In contrast, previous deep-learning-based optoacoustic reconstruction approaches were limited in their ability to generalize from synthesized training data to in vivo test data because the underlying inverse problems were ill-posed. More precisely, the targets used during the training of these deep neural networks were the true synthetic initial pressure images (left side in Fig. 1a), containing image information not available in the input sinogram due to limited angle acquisition, measurement noise, or finite transducer bandwidth. To restore the missing information, these deep neural network models were required to incorporate information from the training data manifold, which limited their generalization and provides a likely explanation for the rudimentary image quality previously reported for in vivo measurements. The experiments conducted with the alternative model DeepMB_{initial-images} confirm that models trained on true synthetic initial pressures images cannot adequately generalize to in vivo data (see Table 1 and Extended Data Fig. 2).

The presented methodology to increase the speed of iterative model-based reconstruction is also applicable to other optoacoustic reconstruction approaches. For instance, frequency-band model-based reconstruction or Bayesian optoacoustic reconstruction can disentangle structures of different physical scales and quantifying reconstruction uncertainty, respectively, but their long reconstruction
times currently hinder their use in real-time applications. The underlying methodology of DeepMB could also be exploited to accelerate parametrized (iterative) inversion approaches for other imaging modalities, such as ultrasound, X-ray computed tomography, magnetic resonance imaging, or, more generally, for any parametric partial differential equation. In conclusion, we introduced DeepMB as a fully operational software-based prototype for real-time model-based optoacoustic image reconstruction. We are currently working on embedding DeepMB into the hardware of a modern MSOT scanner, to use DeepMB for real-time imaging in clinical applications.

4. Methods

Handheld MSOT imaging system

We evaluated DeepMB with a modern handheld MSOT scanner (MSOT Acuity Echo, iThera Medical GmbH, Munich, Germany). The system is equipped with a multi-wavelength laser that illuminates tissues with short laser pulses (<10 ns) at a repetition rate of 25 Hz. The scanner features a custom-made ultrasound detector (IMASONIC SAS, Voray-sur-l'Ognon, France) with the following characteristics: Number of piezoelectric elements: 256; Concavity radius: 4 cm; Angular coverage: 125°; Central frequency: 4 MHz. Parasitic noise generated by light-transducer interference is reduced via optical shielding of the matching layer, yielding an extended 153% frequency bandwidth. The raw channel data for each optoacoustic scan is recorded with a sampling frequency of 40 MHz during 50.75 µs, yielding a sinogram of size 2030×256 samples. Co-registered B-mode ultrasound images are also acquired and interleaved at approximately 6 Hz for live guidance and navigation. During imaging, optoacoustic backprojection images well as B-mode ultrasound images are displayed in real time on the scanner monitor for guidance.

Acquisition of in vivo test sinograms

To collect in vivo data for DeepMB evaluation, we scanned six healthy volunteers. The involved participants were three females and three males, aged from 20 to 36 years (mean age: 28.3±5.7). Self-assessed skin color according to the Fitzpatrick scale was type II (2 participants), type III (3 p.), and type IV (1 p.). Self-assessed body type was ectomorph (2 p.), mesomorph (3 p.), and endomorph (1 p.). We have complied with all relevant ethical regulations following the guidelines provided by local ethics committee of the Technical University of Munich. All participants gave written informed consent upon recruitment.

For each participant, we scanned between 25 and 29 different combinations of anatomical locations and probe orientations: biceps, thyroid, carotid, calf (each left/right and transversal/longitudinal), elbow, neck, colon (each left/right), and breast (each left/right and top/bottom, female participants only). For each combination of anatomical location and probe orientation, we conducted between one and four acquisitions. During each acquisition, we recorded sinograms for approximately 10 s at wavelengths cyclically iterating from 700 to 980 nm in steps of 10 nm. We then selected, per acquisition, the 29 consecutively acquired sinograms for which we observed minimal motion in the interleaved ultrasound images, amounting to a total of 4814 in vivo test sinograms. Finally, we band-pass filtered all selected in vivo sinograms between 100 kHz and 12 MHz to remove frequency components beyond the transducer bandwidth and cropped the first 110 time samples to remove device-specific noise present at the beginning of the sinograms.

Determination of the SoS values

To evaluate DeepMB reconstructions under both in-focus and out-of-focus conditions, we manually tuned the SoS value of all in vivo test scans. We used a SoS step size of 5 m/s to enable SoS adjustments slightly below the system spatial resolution (approximatively 200 µm). We found that the range of optimal SoS values was (1475–1525 m/s) for the in vivo dataset, and we therefore used the same range to define the supported input SoS values of the DeepMB network.

For each scan, we manually selected the SoS value that resulted in the most well-focused reconstructed image. To speed up tuning, we selected the optimal SoS values based on approximate and high-frequency-dominated reconstructions that we computed by applying the transpose model of the system to the recorded sinograms. Furthermore, we tuned the SoS only for scans at 800 nm and adopted the
values for all scans at other wavelengths acquired at the time exploiting their spatial co-registration due to the absence of motion (see previous sections for details).

Synthesis of sinograms for training and validation

For network training and validation, optoacoustic sinograms were synthesized with an accurate physical forward model of imaging process that incorporates the total impulse response of the system, parametrized by a SoS value drawn uniformly at random from the range (1475–1525 m/s) with step size 5 m/s. Real-world images serving as initial pressure distributions for the forward simulations were randomly selected from the publicly available PASCAL Visual Object Classes Challenge 2012 (VOC2012) dataset, converted to mono-channel grayscale, and resized to 416×416 pixels. After the application of the forward model, each synthesized sinogram was scaled by a factor drawn uniformly at random from the range (0–450) to match the dynamic range of in vivo sinograms.

Image reconstruction

To generate ground-truth optoacoustic images, we reconstructed all sinograms (synthetic as well as in vivo) via iterative-model-based. We used Shearlet L1 to tackle the ill-posedness of the inverse problem. Shearlet L1 regularization is a convex relaxation of Shearlet sparsity, which can reduce limited-view artifacts in reconstructed images, because Shearlets provide a maximally-sparse approximation of a larger class of images (known as cartoon-like functions) with a provably optimal encoding rate. The optimal pressure field to find is characterized as

$$p_0 = \arg \min_{p \geq 0} \| M_{\text{SoS}} p - s \|_2^2 + \lambda \| \text{SH}(p) \|_1,$$

where $p_0$ is the reconstructed image, $M_{\text{SoS}}$ is the forward model of the imaging process for the selected reconstruction SoS, $s$ is the input sinogram, $\lambda$ is the regularization parameter tuned via an L-curve, $\text{SH}$ is the Shearlet transform, and $\| \cdot \|_n$ is the n-norm. The minimization problem was solved via bound-constrained sparse reconstruction by separable approximation. All images were reconstructed with a size of 416×416 pixels and a field of view of 4.16×4.16 cm². For comparison purposes, we also reconstructed all images using the backprojection formula.

Network training

DeepMB was trained — either on synthetic or in vivo data — for 300 epochs using stochastic gradient descent with batch size=4, learning rate=0.01, momentum=0.99, and per epoch learning rate decay factor=0.99. The network loss was calculated as the mean square error between the output image and the reference image. The final model was selected based on the minimal loss on the validation dataset.

To facilitate training, all input sinograms were scaled by $K=450$ to ensure that their values never exceed the range (-1; 1). The same scaling factor was also applied to all target images. Furthermore, the square root was applied to all target reference images used during training and validation to reduce the network output values and limit the influence of high intensity pixels during loss calculation. When applying the trained network on in vivo test data, inferred images were first squared then scaled by $K^{-1}$, to revert the preprocessing operation.

When training on synthetic data to build the standard DeepMB model, we used 8000 sinograms as train split and 2000 sinograms as validation split. The alternative scenario involving training on in vivo data to build the DeepMB in-vivo models was carried out as described hereafter: six different permutations were conducted, with a 4/1/1 participants division between the train, validation, and test splits, respectively, each participant being once and only once part of the validation and test splits.

The DeepMB network is based upon the U-Net architecture, with a depth of 5 layers and a width of 64 features. To gradually reduce the total number of data channels from 267 (that is, 256 transducer elements, and one-hot encoding of 11 possible SoS values) down to 64, three 2D convolutional layers with 208, 160, and 112 features, respectively, were added prior to the U-Net. All kernel and padding size were (3, 3) and (1, 1), respectively. Biases were accounted for, and the final activation was the absolute value function.

Data residual norm
To quantify the image fidelity of reconstructions from DeepMB, model-based, or backprojection, we evaluated the data residual norm $R$, defined as

$$R := \frac{\|M_{\text{SoS}} p_0 - s\|^2_2}{\|s\|^2_2},$$

where $p_0$ is the reconstructed image, $M_{\text{SoS}}$ is the forward model from model-based reconstruction, $s$ is the input sinogram, and $\|\cdot\|_2$ is the 2-norm. To constrain the solutions space for all reconstruction methods in a similar way and enable a meaningful comparison between backprojection on one hand, versus non-negative model-based and DeepMB on the other hand, negative pixel values were set to zero prior to residual calculation for backprojection images. All images were individually scaled using the linear degree of freedom in reconstructed optoacoustic image so that their data residual norms are minimal.

For the evaluation of in-focus images, data residual norms were calculated for model-based, DeepMB, and backprojection reconstructions with the optimal SoS values, for all 4814 samples from the in vivo test set. For the evaluation of out-of-focus images, data residuals were calculated for model-based and DeepMB reconstructions with all 11 SoS values, for a subset of 638 randomly selected in vivo samples.

**Code availability.** The entirety of the DeepMB source code is publicly available on GitHub\(^1\).

**Data availability.** The data that support the findings of this study are available on request from the corresponding author Dominik Jüstel. The data are not publicly available due to them containing information that could compromise research participant privacy/consent.

**Acknowledgements.** The authors would like to thank Antonia Longo for her precious contribution during in vivo image acquisition and the conception of Fig. 2, and Robert Wilson for his attentive reading and improvements of the manuscript.

**Funding.** This project has received funding from the Bavarian Ministry of Economic Affairs, Energy and Technology (StMWi) (DIE-2106-0005//DIE0161/02, DeepOpus) and from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme under grant agreement No 694968 (PREMSOT).

5. **References**

1. Ntziachristos, V. & Razansky, D. Molecular imaging by means of multispectral optoacoustic tomography (MSOT). *Chem Rev* 110, 2783-2794 (2010).
2. Diot, G., Metz, S., Noske, A., Liapis, E., Schroeder, B., Ovsepian, S.V., Meier, R., Rummeny, E. & Ntziachristos, V. Multispectral Optoacoustic Tomography (MSOT) of Human Breast Cancer. *Clin Cancer Res* 23, 6912-6922 (2017).
3. Kniefel, F., Neufert, C., Hartmann, A., Claussen, J., Urich, A., Egger, C., Vetter, M., Fischer, S., Pfeifer, L., Hagel, A., Kielisch, C., Gortz, R.S., Wildner, D., Engel, M., Rother, J., Uter, W., Siebler, J., Attiya, R., Rascher, W., Strobel, D., Neurath, M.F. & Waldner, M.J. Multispectral Optoacoustic Tomography for Assessment of Crohn’s Disease Activity. *N Engl J Med* 376, 1292-1294 (2017).
4. Karlas, A., Kallmayer, M., Fasoula, N.-A., Liapis, E., Bariotakis, M., Krönke, M., Anastasopoulou, M., Reber, J., Eckstein, H.-H. & Ntziachristos, V. Multispectral optoacoustic tomography of muscle perfusion and oxygenation under arterial and venous occlusion: A human pilot study. *Journal of Biophotonics* 13, e201960169 (2020).
5. Dehner, C., Olefir, I., Basak Chowdhury, K., Jüstel, D. & Ntziachristos, V. Deep learning based electrical noise removal enables high spectral optoacoustic contrast in deep tissue. *arXiv:2102.12960*, 2021.
6. Kukačka, J., Metz, S., Dehner, C., Muckenhuber, A., Paul-Yuan, K., Karlas, A., Fallenberg, E.M., Rummeny, E., Jüstel, D. & Ntziachristos, V. Image processing improvements afford second-generation handheld optoacoustic imaging of breast cancer patients. *Photoacoustics*, 100343 (2022).

\(^1\) [https://github.com/juestellab/deepmb](https://github.com/juestellab/deepmb)
7. Regensburger, A.P., Fonteyne, L.M., Jungert, J., Wagner, A.L., Gerhalter, T., Nagel, A.M., Heiss, R., Flenkenthaler, F., Qurashi, M., Neurath, M.F., Klymiuk, N., Kemter, E., Frohlich, T., Uder, M., Woelfle, J., Rascher, W., Trollmann, R., Wolf, E., Waldner, M.J. & Knieling, F. Detection of collagens by multispectral optoacoustic tomography as an imaging biomarker for Duchenne muscular dystrophy. *Nat Med* 25, 1905-1915 (2019).
8. Dima, A. & Ntziachristos, V. Non-invasive carotid imaging using optoacoustic tomography. *Opt Express* 20, 25044-25057 (2012).
9. Taruttis, A. & Ntziachristos, V. Advances in real-time multispectral optoacoustic imaging and its applications. *Nature Photonics* 9, 219-227 (2015).
10. Ivankovic, I., Mercep, E., Schmedt, C.G., Dean-Ben, X.L. & Razansky, D. Real-time Volumetric Assessment of the Human Carotid Artery: Handheld Multispectral Optoacoustic Tomography. *Radiology* 291, 45-50 (2019).
11. Sethuraman, S., Aglyamov, S.R., Amirian, J.H., Smalling, R.W. & Emelianov, S.Y. Intravascular photoacoustic imaging using an IVUS imaging catheter. *IEEE Trans Ultrason Ferroelectr Freq Control* 54, 978-986 (2007).
12. Yang, J.M., Maslov, K., Yang, H.C., Zhou, Q., Shung, K.K. & Wang, L.V. Photoacoustic endoscopy. *Opt. Lett.* 34, 1591-1593 (2009).
13. Xu, M. & Wang, L.V. Universal back-projection algorithm for computed photoacoustic tomography. *Phys Rev E Stat Nonlin Soft Matter Phys* 71, 016706 (2005).
14. Chowdhury, K.B., Prakash, J., Karlas, A., Jüstel, D. & Ntziachristos, V. A Synthetic Total Impulse Response Characterization Method for Correction of Hand-held Optoacoustic Images. *IEEE Transactions on Medical Imaging* 39, 3218-3230 (2020).
15. Chowdhury, K.B., Bader, M., Dehner, C., Justel, D. & Ntziachristos, V. Individual transducer impulse response characterization method to improve image quality of array-based handheld optoacoustic tomography. *Opt. Lett.* 46, 1-4 (2021).
16. Ding, L., Dean-Ben, X.L. & Razansky, D. Real-Time Model-Based Inversion in Cross-Sectional Optoacoustic Tomography. *IEEE Trans Med Imaging* 35, 1883-1891 (2016).
17. Kim, M., Jeng, G.S., Pelivanov, I. & O’Donnell, M. Deep-Learning Image Reconstruction for Real-Time Photoacoustic System. *IEEE Trans Med Imaging* 39, 3379-3390 (2020).
18. Lan, H., Jiang, D., Yang, C., Gao, F. & Gao, F. Y-Net: Hybrid deep learning image reconstruction for photoacoustic tomography in vivo. *Photoacoustics* 20, 100197 (2020).
19. Waibel, D., Gröhl, J., Isensee, F., Kirchner, T., Maier-Hein, K. & Maier-Hein, L. Reconstruction of initial pressure from limited view photoacoustic images using deep learning. Photons Plus Ultrasound: Imaging and Sensing 2018.
20. Feng, J., Deng, J., Li, Z., Sun, Z., Dou, H. & Jia, K. End-to-end Res-Unet based reconstruction algorithm for photoacoustic imaging. *Biomed. Opt. Express* 11, 5321-5340 (2020).
21. Tong, T., Huang, W., Wang, K., He, Z., Yin, L., Yang, X., Zhang, S. & Tian, J. Domain Transform Network for Photoacoustic Tomography from Limited-view and Sparsely Sampled Data. *Photoacoustics* 19, 100190 (2020).
22. Guan, S., Khan, A.A., Sikdar, S. & Chitnis, P.V. Limited-View and Sparse Photoacoustic Tomography for Neuroimaging with Deep Learning. *Sci Rep* 10, 8510 (2020).
23. Hauptmann, A., Lucka, F., Betcke, M., Huynh, N., Adler, J., Cox, B., Beard, P., Ourselin, S. & Arridge, S. Model-Based Learning for Accelerated, Limited-View 3-D Photoacoustic Tomography. *IEEE Trans Med Imaging* 37, 1382-1393 (2018).
24. Grohl, J., Schellenberg, M., Dreher, K. & Maier-Hein, L. Deep learning for biomedical photoacoustic imaging: A review. *Photoacoustics* 22, 100241 (2021).
25. Hauptmann, A. & Cox, B. Deep learning in photoacoustic tomography: current approaches and future directions. *Journal of Biomedical Optics* 25, 112903 (2020).
26. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. & Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 303-338 (2010).
27. Rosenthal, A., Ntziachristos, V. & Razansky, D. Acoustic Inversion in Optoacoustic Tomography: A Review. *Curr Med Imaging Rev* 9, 318-336 (2013).
28. Zhu, B., Liu, J.Z., Cauley, S.F., Rosen, B.R. & Rosen, M.S. Image reconstruction by domain-transform manifold learning. *Nature* 555, 487-492 (2018).
29. Longo, A., Jüstel, D. & Ntziachristos, V. Disentangling the frequency content in optoacoustics. arXiv:2104.14966, 2021.
6. Extended Data Figures

Extended Data Figure 1: Representative examples of reconstruction artifacts (red arrows) from alternative models DeepMB<sub>in-vivo</sub> (i.e., trained on in vivo data instead of synthesized data). The three rows depict different anatomies (biceps: a–e, breast: f–j, thyroid: k–o). The three leftmost columns correspond to images reconstructed via model-based (MB), alternative DeepMB trained on in vivo data (DeepMB<sub>in-vivo</sub>), and standard DeepMB (DeepMB). The two rightmost columns show the absolute differences between the reference model-based image and the image inferred from DeepMB<sub>in-vivo</sub> and DeepMB, respectively. The field of view is 4.16×4.16 cm², the enlarged region is 0.61×0.61 cm².
Extended Data Figure 2: Representative examples showing the inaptitude of the alternative model DeepMB\textsubscript{initial-images} (i.e., trained on true initial pressure images) to reconstruct in vivo images. The three rows depict different anatomies (elbow: a–e, abdomen: f–j, calf: k–o). The three leftmost columns correspond to images reconstructed via model-based (MB), alternative DeepMB\textsubscript{initial-images}, and standard DeepMB. The two rightmost columns show the absolute differences between the reference model-based image and the image inferred from DeepMB\textsubscript{initial-images} and DeepMB, respectively. The field of view is 4.16×4.16 cm\textsuperscript{2}, the enlarged region is 0.61×0.61 cm\textsuperscript{2}. 