TECHNISCHE UNIVERSITÄT
MÜNCHEN

DEPARTMENT OF INFORMATICS

Deep Convolutional Neural Networks for Biomedical Image Analysis

Oliver Schoppe

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Matthias Niessner
Prüfer der Dissertation:	1. Prof. Dr. Bjoern H. Menze
	2. Prof. Dr. Daniel Razansky

Die Dissertation wurde am 28.09.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 21.01.2021 angenommen.

Abstract

Medical imaging plays a steadily increasing role in clinical workflows as well as in pre-clinical biomedical research. The rise of machine learning has enabled a series of breakthroughs in medical image analysis, addressing a long-standing need for higher automation and quality in interpretation of this data. However, the adoption in clinics and laboratories remains slow. A combination of scarcity of labeled training data, limited reliability of those labels, and insufficient generalization and robustness of the models forms a high adoption barrier and causes underwhelming performance in practical settings.

This dissertation aims at addressing these bottlenecks by developing efficient training strategies for models that generalize well and appreciate the imperfection of labels. Chapter A.1 introduces DeepMACT, the first ever whole body analysis of the complete metastatic spread of tumors in mice with the help of deep neural networks. Despite little available training data, a highly efficient 2D approach to solving the 3D detection task enabled a performance en par with a human expert. Trained on one line of breast cancer, DeepMACT generalized well for other tumors and also enabled assessing the efficacy of therapeutic antibodies as a treatment option. In Chapter A.2, the limits of model generalization were tested. Networks were trained on synthetic and real data from highly different biomedical domains (e.g., human blood vasculature in MRI versus microscopy data from the murine nervous system). The cross-prediction performance across these domains revealed the potential of synthetic training data and transfer learning for improved generalization. Further work describes a deep learning pipeline for automated multi-organ segmentation in murine whole-body scans termed AIMOS. The approach generalizes across imaging modalities, exceeds the segmentation performance of the state-of-the-art and is en par with human experts. AIMOS can be trained from scratch with as little as 10 samples and furthermore provides localized metrics of intrinsic ambiguity in the scans. Lastly, additional work explores further applications on (imaged-based) learning models biomedical modeling in the field of pre-clinical neuroscience on the basis of image-like spectral representations of auditory stimuli.

The recent success of machine learning for medical image analysis is still largely confined to highly controlled settings. The work presented here shows that resolving the bottlenecks for adoption *in the wild* poses an underappreciated frontier in our field of science and may be key to turning academic achievements into better health.

Zusammenfassung

Medizinische Bildgebung spielt eine immer wichtigere Rolle im klinischen Betrieb sowie in der vorklinischen, biomedizinischen Forschung. Getrieben vom steten Bedarf an Automatisierung und Qualitätssteigerung in der Auswertung dieser Daten, ermöglichte der Siegeszug des maschinellen Lernens eine Reihe an Durchbrüchen in der medizinischen Bildanalyse. Jedoch hält maschinelles Lernen in den Kliniken und Laboren nur langsam Einzug. Eine Kombination aus Knappheit annotierter Traingingsdaten, begrenzter Verlässlichkeit dieser Annotationen, und einer unzureichenden Allgemeingültigkeit und Robustheit der Modelle stellt eine hohe Hürde für den Praxiseinsatz dar und führt zu enttäuschenden Ergebnissen.

Diese Dissertation setzt sich zum Ziel, jene Hürden durch die Entwicklung effizienter Trainingsstrategien für Modelle mit hohem Grad an Allgemeingültigkeit und unter Berücksichtigung der Fehlerhaftigkeit der Annotationen zu mindern. Das Kapitel A.1 stellt DeepMACT vor, die erste Analyse der kompletten metastatischen Ausbreitung von Tumoren in Ganzkörperaufnahmen von Mäusen auf Basis tiefer neuronaler Netze. Trotz der geringen Trainingsdatenmenge erlaubte ein hocheffizienter 2D Ansatz des 3D Detektionsproblems eine Präzision vergleichbar mit der menschlicher Experten. DeepMACT wurde nur auf eine Art von Brustkrebs trainiert, konnte die Problemlösung jedoch auf andere Tumore verallgemeinern und darüber hinaus auch die Wirksamkeit von therapeutischen Antikörpern beurteilen. In Kapitel A.2 werden die Grenzen der Verallgemeinerungsfähigkeit von neuronalen Netzen ausgelotet. Die Modelle wurden auf synthetische und echte Daten aus unterschiedlichsten biomedizinischen Domänen trainiert (z.B. menschliche Blutgefäße im MRT oder mikroskopische Aufnahmen des Nervensystems von Mäusen). Die Kreuzvorhersage über diese Domänen hinweg legte das Potential synthetischer Trainingsdaten und des Wissenstransfers für die Verallgemeinerungsfähigkeit von Modellen offen. Im Weiteren wird ein lernendes System namens AIMOS für die automatisierte Multi-Organ-Segmentierung in Ganzkörperaufnahmen von Mäusen entwickelt. Der Ansatz lässt sich auf verschiedene Bildgebungsverfahren verallgemeinern, übertrifft die Präzision des bisherigen Standes der Wissenschaft und entspricht der Segmentierungsqualität von Experten. AIMOS kann bereits mit nur etwa 10 Aufnahmen von Grund auf trainiert werden und ermöglicht darüber hinaus eine lokalisierte Einschätzung von Uneindeutigkeiten in der Aufnahme. Letztlich werden in weiteren Arbeiten Anwendungen lernender biomedizinischer Modelle im Feld der vorklinischen Neurowissenschaften untersucht auf Basis bildähnlicher Repräsentationen der Spektren auditiver Stimuli.

Die Erfolgsgeschichte maschinellen Lernens in der medizinischen Bildanalyse spielt sich immer noch in Versuchsbedingungen ab. Die Werke hier zeigen jedoch auf, dass die Überwindung praktischer Hürden ein unterschätztes Forschungsfeld ist, das in der Überführung akademischer Erfolge in eine bessere Gesundheitsversorgung eine Schlüsselrolle einnimmt.

Acknowledgments

I am deeply thankful for all the support I received from my advisors, collaborators, colleagues, and friends (largely overlapping groups) along the way. I stopped counting co-authors at a number of 40-50 for the publications presented in this dissertation - it is obvious that this work is not only mine but the result of a fantastic team effort.

First, I want to express my sincere gratitude to Prof. Dr. Bjoern Menze. Bjoern has provided guidance and inspiration where needed but always gives his PhD candidates maximum freedom. Thank you for your trust and for making this possible! Also for the entire IBBM group: Amir, Anjany, Bran, Caro, Dhriti, Diana, Esther, Fercho, Florian, Giles, Ivan, Jana, Johannes, John, Marie, Markus, Rami, Supro, Yusuf and others - I cannot thank you enough for making this a great place to work and to have fun! I am thankful for the opportunity to supervise many Bachelor and Master students during this time, some of which became colleagues and co-authors: Arno, Dmitrii, Javier, Johannes, Martin, Nimar, Oliver, Prashanth, Rami, Vindhya. Thanks for your work and stimulating discussions!

At the iTERM institute of Dr. Ali Ertürk at the Helmholtz Center Munich, I want to thank the entire team for a great collaboration. Ali, thanks for your drive, your vision, and for your trust to let me build up and lead the AI group of your institute. Very special thanks go to Chenchen and Mike with whom I had the honor of working closely together throughout the entire time. But I am also grateful for having had the opportunity to learn from and work with Marika, Benjamin, Harsh, Ilgin, Marin, Doris, Hongcheng, Zhouyi, Karen, Shan, Madita, Izabela, and Louis. To the AI team: the road is ahead of us - let's keep on rocking and let's make this big!

Last, but surely not least, I want to thank the *Fellow* crew for an amazing ride together, all my friends who do not need to be named to know how much I love them for their constant support, and my loving family!

Publication list

This cumulative dissertation is based on the following work:

Work on biomedical image analysis

Accepted first author publications

1. C Pan*, **O Schoppe***, A Parra-Damas*, R Cai, M Todorov, G Gondi, B v. Neubeck, N Böğürcü-Seidel, S Seidel, K Sleiman, C Veltkamp, B Förstera, H Mai, Z Rong, O Trompak, A Ghasemigharagoz, M Reimer, A Cuesta, J Coronel, I Jeremias, D Saur, A Acker-Palmer, T Acker, B Garvalov, BH Menze, R Zeidler, A Ertürk. *Deep Learning Reveals Cancer Metastasis and Therapeutic Antibody Targeting in the Entire Body*. **Cell** 179, 1661–1676, 2019

2. J Paetzold*, **O Schoppe***, R Al-Maskari, G Tetteh, V Efremov, M Todorov, R Cai, H Mai, Z Rong, A Ertürk, BH Menze. *Transfer learning from synthetic data reduces need for labels to segment brain vasculature and neural pathways in 3D*. **International Conference on Medical Imaging with Deep Learning**, 2019 (peer-reviewed paper)

First author manuscripts in peer-review - not included in this dissertation

3. **O Schoppe**, C Pan, J Coronel, H Mai, Z Rong, M Todorov, A Müskes, F Navarro, A Ertürk, and BH Menze. *Deep learning-enabled organ segmentation with uncertainty quantification of whole-body mouse scans*. **In revision at Nature Communications** 2020

Work on (image-based) biomedical modeling

Accepted first author publications

4. **O Schoppe**, NS Harper, BDB Willmore, AJ King, JWH Schnupp. *Measuring the performance of neural models*. **Frontiers in Computational Neuroscience** 10, 2016

5. BDB Willmore*, **O Schoppe***, AJ King, JWH Schnupp, NS Harper. *Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing*. **Journal of Neuroscience** 36, 280-289, 2016

6. NS Harper*, **O Schoppe***, BDB Willmore, Z Cui, JWH Schnupp, AJ King. *Network receptive field modeling reveals extensive integration and multi-feature selectivity in auditory cortical neurons*. **PLoS Computational Biology** 12, e1005113, 2016

Besides the work listed above, further first- and co-author publications on biomedical image analysis and beyond are listed below but are not included in this dissertation:

Further work (not included in this dissertation)

7. M Todorov*, J Paetzold*, **O Schoppe**, G Tetteh, V Efremov, K Voelgyi, M Duering, M Dichgans, M Piraud, BH Menze, A Ertürk. *Machine learning analysis of whole mouse brain vasculature.* **Nature Methods** 17, 442–449, 2020

8. S Zhao, M Todorov, R Cai, R al-Maskari, H Steinke, E Kemter, H Mai, Z Rong, M Warmer, K Stanic, **O Schoppe**, J Paetzold, B Gesierich, M Wong, T Huber, M Duering, O Bruns, BH Menze, J Lipfert, V Puelles, E Wolf, I Bechmann, A Ertürk. *Cellular and Molecular Probing of Intact Human Organs*. **Cell** 180, Issue 4, 796-812.e19, 2020

9. M Kimm, M Shevtsov, C Werner, W Sievert, Z Wu, **O Schoppe**, BH Menze, E Rummeny, R Proksa, O Bystrova, M Martynova, G Multhoff, S Stangl. *Gold Nanoparticle Mediated Multi-Modal CT Imaging of Hsp70 Membrane Positive Tumors.* **Cancers**, 12, 1331, 2020

10. A Qasim*, I Ezhov*, S Shit, **O Schoppe**, J Paetzold, A Sekuboyina, F Kofler, J Lipkova, H Li, BH Menze. *Red-GAN: Attacking class imbalance via conditioned generation. Yet another medical imaging perspective*. **International Conference on Medical Imaging with Deep Learning**, 2020 (peer-reviewed paper)

11. M Rudnicki, **O Schoppe**, M Isik, F Völk, W Hemmert. *Modeling auditory coding: from sound to spikes.* **Cell and Tissue Research** 361, 159-175, 2015

12. Q Wan*, **O Schoppe***, S Gunasekaran*, D Holland, E Roche, H-C Hur, C Walsh. *Multifunctional Laparoscopic Trocar With Built-in Fascial Closure and Stabilization.* **Journal of Medical Devices** 7, 030912, 2013

**Joint first authorship*

Contents

1. Introduction and methods

1.1. Biomedical imaging

From clinical diagnostics in hospitals to pre-clinical research in biology labs, our society relies on a set of highly developed imaging techniques to obtain visual information not accessible to the eye. Technological advances dramatically increased the informative power of the obtained images while reducing the financial cost, resulting in a ever more central role of biomedical imaging in fundamental research as well as clinical procedures.

1.1.1. Development of and modalities in clinical imaging

The term *clinical imaging* describes a variety of imaging modalities commonly used in clinical settings, for instance in radiology departments of hospitals. The predominant goal of clinical imaging is to provide supporting information to find, specify, or validate the diagnostic assessment of a human patient. However, imaging also plays a role during treatment planning (for instance, for radiation therapies), during surgery (for instance, to monitor the position of surgical instruments), and for assessment of treatment effectiveness.

A variety of imaging modalities have been developed, each optimized for specific procedures and together providing complementary information (see Fig. 1.1). The most commonly used imaging modalities in clinical settings can be primarily grouped by the nature of the acquired signal. While many variants of each modality exist, the variants of each modality follow the same fundamental principles as outlined below.

Radiography and X-ray computed tomography

As one of the oldest of all modalities, radiographs date back to 1895 when Wilhelm Röntgen experimented with cathodes and discovered the *X-ray* band of the electromagnetic spectrum. Exposing his wife to the radiation, Röntgen created the first radiograph visualizing the bones of her hand. Since then, the fundamental principle of radiography (acquiring the X-ray projection with photosensitive material or sensors) has remained unchanged. X-ray computed tomography (CT), while following the same underlying principle, enables the acquisition of volumetric scans. This is

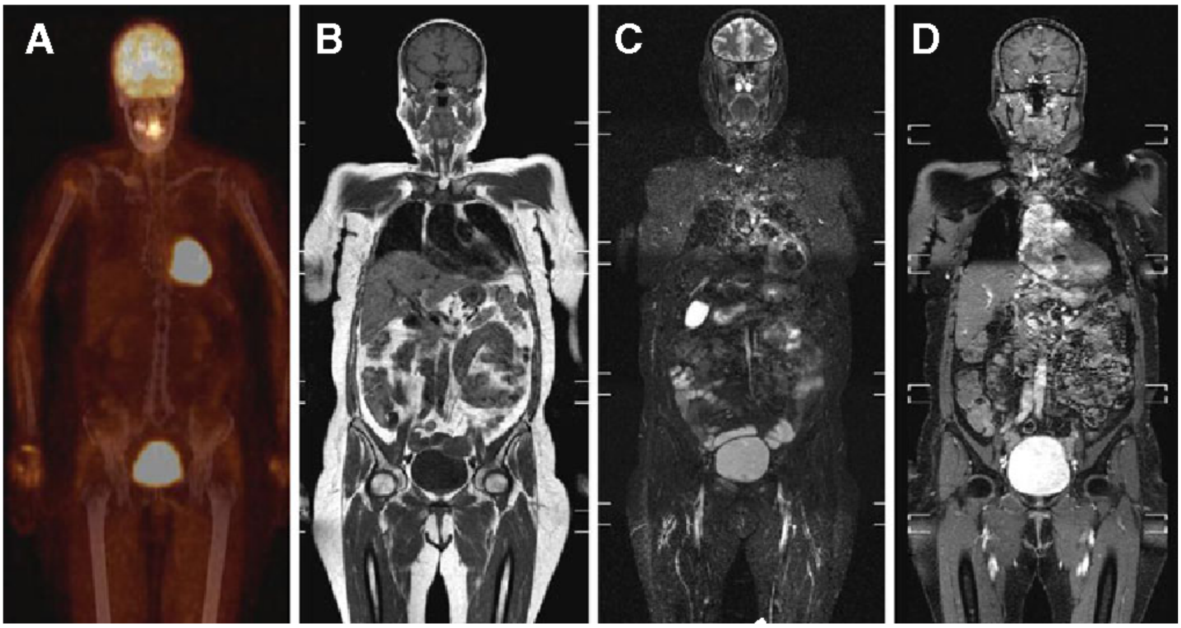


Figure 1.1.: **Clinical imaging modalities provide complementary information**. Whole-body scans of a woman; A) fusion of scans obtained with PET and CT, B) T1 TSE MRI image, C) T2 STIR MRI image with fat saturation, D) contrast-enhanced T1 WATS MRI image. Taken from Derlin et al. 2013

achieved by reconstructing a 3D map of X-ray absorption from a large set of planar X-ray projections acquired at different angles along an axis of rotation. Both, radiography and CT visualize the contrasts of bones (also see Fig. 1.1A) and lungs especially well but provide little visual information about soft organs that mostly consist of water (e.g. in the abdomen). Injection of contrast-enhancing agents, however, allows to partially overcome this limitation. Despite being among the most commonly used imaging modalities, the health risk associated with this ionizing radiation exposure limits the use of X-ray based imaging.

Ultrasonography

Free of any medical risks and available at low costs, ultrasonography is among the most widely used imaging modalities. It exploits the differences in acoustical impedance of different tissues as a basis for visualization. Piezoelectric transducers generate acoustic waves in the megahertz range, which are partially reflected at boundaries of tissues with different impedance. Measuring the time difference between generation and echo detection and the signal strength of the echo allows to compute the a profile of acoustic properties of the tissue. Complementary to X-ray based imaging, ultrasound provides much stronger contrasts for soft abdominal organs, muscles, blood vessels, or tendons.

Magnetic resonance imaging

Magnetic resonance imaging (MRI) is based on a more complex mechanism and is characterized by a large variety of variants and protocols, each optimized for specific imaging requirements. Very strong magnetic fields, up to the range of several teslas, align the magnetic spin of single protons, which form the nucleus of hydrogen atoms. Radiofrequency pulses at the resonance frequency of the protons changes the spin, which subsequently *relax* back to the equilibrium and emit radio waves while doing so. This emission is the acquired signal to form an image. The location of emission can be reconstructed by overlaying the primary magnetic field with *gradient fields* that vary over space and time, which enables highly resolvable spatial encoding of the acquired signals. Since hydrogen atoms mostly occur in watery tissue and fat, MIR is especially well suited to visualize soft tissue such as abdominal organs. Modifying the stimulation pulse patterns allows to tailor the resulting correspondence between tissue characteristics and image contrasts (see Fig. 1.1B-D), proving complementary information.

Nuclear imaging

In contrast the previously described methods, nuclear imaging acquires a signal from within the body that does not rely on external stimulation. The most commonly used variants, positron emission tomography (PET) and single-photon emission computed tomography (SPECT), acquire the signal from radioactive substances. Typically, nuclear imaging is used for *functional* rather than *anatomical* imaging. For instance, it can provide information about metabolic processes by using glucose combined with radioactive fluorine-18 as a tracer. Such a signal will correlate with the metabolic activity of tissue, which is typically especially high in tumors. An example can be seen in Fig. 1.1A, where Derlin et al. 2013 used this tracer to monitor treatment effectiveness in a patient with multiple myeloma.

Beyond these most commonly used modalities, hybrid solutions (such as PET-CT) and further approaches (such as photoacoustic imaging) exist but will not be detailed here.

1.1.2. Imaging in pre-clinical research

Beyond the applications in clinical settings, biomedical imaging plays a central role in pre-clinical and biological research. While some of the underlying image acquisition principles are very similar, it is important to appreciate the differences in its objective and in the boundary conditions. In pre-clinical research settings, the goal of imaging is very much centered around scientific questions on the fundamental working mechanisms in biology and of pathological conditions. This has consequences for what kind of imaging can be used and how it is used.

Importantly, most of biomedical research is not done on living humans but rather on (human or non-human) tissue samples or on laboratory animals, adding a diverse and powerful set of methods that cannot be used in clinical settings due to technical, medical, and moral considerations. For instance, invasive and post mortem procedures allow a more direct observation of the region of interest. But also biotechnological methods such as genetic modification play pivotal roles in opening up what kind of imaging can be performed to answer scientific questions. However, this goes along with a much higher degree of variability in imaging protocols, reducing the comparability of data acquired in different settings.

The range of pre-clinical imaging modalities reflects the breadth of biomedical research questions and spans several orders of magnitude of spatial resolution and field of view, from whole-body imaging techniques also used in clinical settings such as CT or MRI down to sub-cellular resolution imaging in optical microscopy or electron microscopy.

If used for smaller specimens or small laboratory animals such as mice, X-ray based computed tomography is often performed with specialized variants termed *micro-CT*, which provides higher spatial resolution within a smaller field of view as compared to the clinical variant. This enables whole-body scans of small animals down to a resolution on the range of 0.1 mm (Holdsworth and Thornton 2002). However, while such high-resolution, volumetric whole body scans already provide a great level of detail, they are a far cry from resolving cellular or sub-cellular details.

If such level of detail is needed, biomedical researchers can turn to a variety of imaging modalities based on optical microscopy. The simplest of which, bright-field optical microscopy, provides highly magnified views of white light transmitted through thin slices of tissue specimens. Combined with the application of functionally staining agents such as hematoxylin and eosin (H&E), this modality is capable of revealing the sub-cellular structure of tissue down to single cell nuclei (see Fig. 1.2) and is widely used for histological analysis.

A more sophisticated variant of optical microscopy, fluorescence microscopy, is not based on transmission of white light but rather exploits the nature of fluorophores to emit light of a given wavelength after being excited with light of a shorter (more energetic wavelength). This allows highly selective signal acquisition by removing any non-specific light of other wavelengths with optical filters. In combination with naturally occurring fluorescent tissue properties and artificially introduced fluorescent staining, this provides highly resolved images of selective structures of interest (see

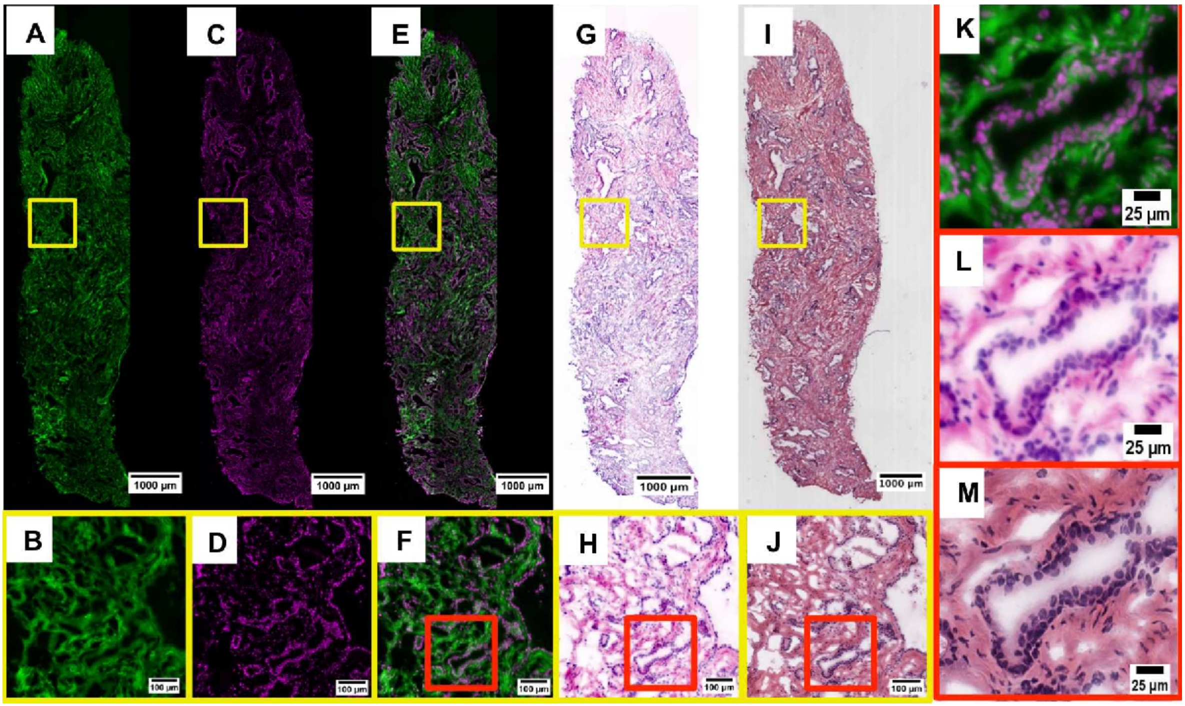


Figure 1.2.: **Preclinical imaging is used in biomedical research**. Functionally stained images of tissue specimens from the prostate; A-F) Fluorescence microscopy imaging, tissue stained with eosin (green) and DRAQ5 (purpple). G-J) Bright-field microscopy imaging, tissue stained with hematoxylin and eosin (H&E). K-M) High-resolution of images allows identification of cell nulcei. Taken from Elfer et al. 2016

Fig. 1.2A-F).

As light is scattered and attenuated when travelling trough tissue, optical microscopy is limited to 2D image acquisition of thin samples of tissue. While variants such as confocal microscopy also allow encoding of depth, such optical microscopy is traditionally not capable of providing volumetric scans at whole body scale as micro-CT, for instance, is capable of. Thus, biomedical researchers are confronted with a trade-off between resolution, field of view and volumetric acquisition, and functional staining. Recent developments around tissue clearing and volumetric light-sheet fluorescent microscopy aim at overcoming this restriction.

1.2. Tissue clearing

One major bottleneck that restricts optical microscopy to 2D images or thin 3D volumes is the scattering and attenuation of light in biological tissue. More specifically, this is caused by the different optical refractory indices (RI) of the fundamental components of biological tissue (e.g., water, proteins, lipid membranes; see Tuchin 2015). The goal of *tissue clearing* is to enable the unrestricted transmission of light through tissue by equalizing the optical refractory index, thereby rendering it transparent.

1.2.1. Overview of tissue clearing approaches

While the origins of research on rendering tissue transparent date back to more than over hundred years ago Spalteholz 1914, only recent work by Erturk et al. 2012 and others (e.g., Kubota et al. 2017) lead to a breakthrough in large-scale tissue clearing. Tissue clearing methods can be grouped along three main approaches (as described by Cai 2019; Richardson and Lichtman 2015), based on (organic) solvents (see Fig. 1.3A), water (see Fig. 1.3B,C), or hydrogels (see Fig. 1.3D). In all approaches, the RI of the tissue is equalized by either replacing or removing components or by changing the RI of a given component.

For instance, in water-based tissue clearing, the comparatively low RI of water is increased by dissolving high RI hydrophilic reagents such as fructose. In organic-solvent based tissue clearing, the main approach used for experiments detailed in Chapters A.1, A.2 and B.1, the tissue is dehydrated with tetrahydrofuran (THF), the lipids are removed with dichloromethane (DCM), and the RI of the tissue is matched with benzyl alcohol/benzyl benzoate (BABB) or dibenzyl-ether (DBE) (see Erturk et al. 2012).

1.2.2. Challenges in whole-body clearing

When applying tissue clearing methods not only to dissected organs but to entire laboratory animals such as mice, further steps need to be considered to achieve results suitable for whole-body imaging (Cai 2019). First, some tissues are especially difficult to clear. For instance, some clearing methods require removal of blood, hair, and, importantly, the skin (Tainaka et al. 2014). Second, some clearing methods enlarge the volume of the tissue, which further complicates whole-body imaging in which the physical size of the imaging chamber is a common bottleneck. Both challenges were addressed by recent work of Pan et al. 2016 and Cai et al. 2018 with the uDISCO and the vDISCO protocols, in which also hard and dense organs like the skin and bones are rendered transparent and in which the overall volume is decreased rather than increased.

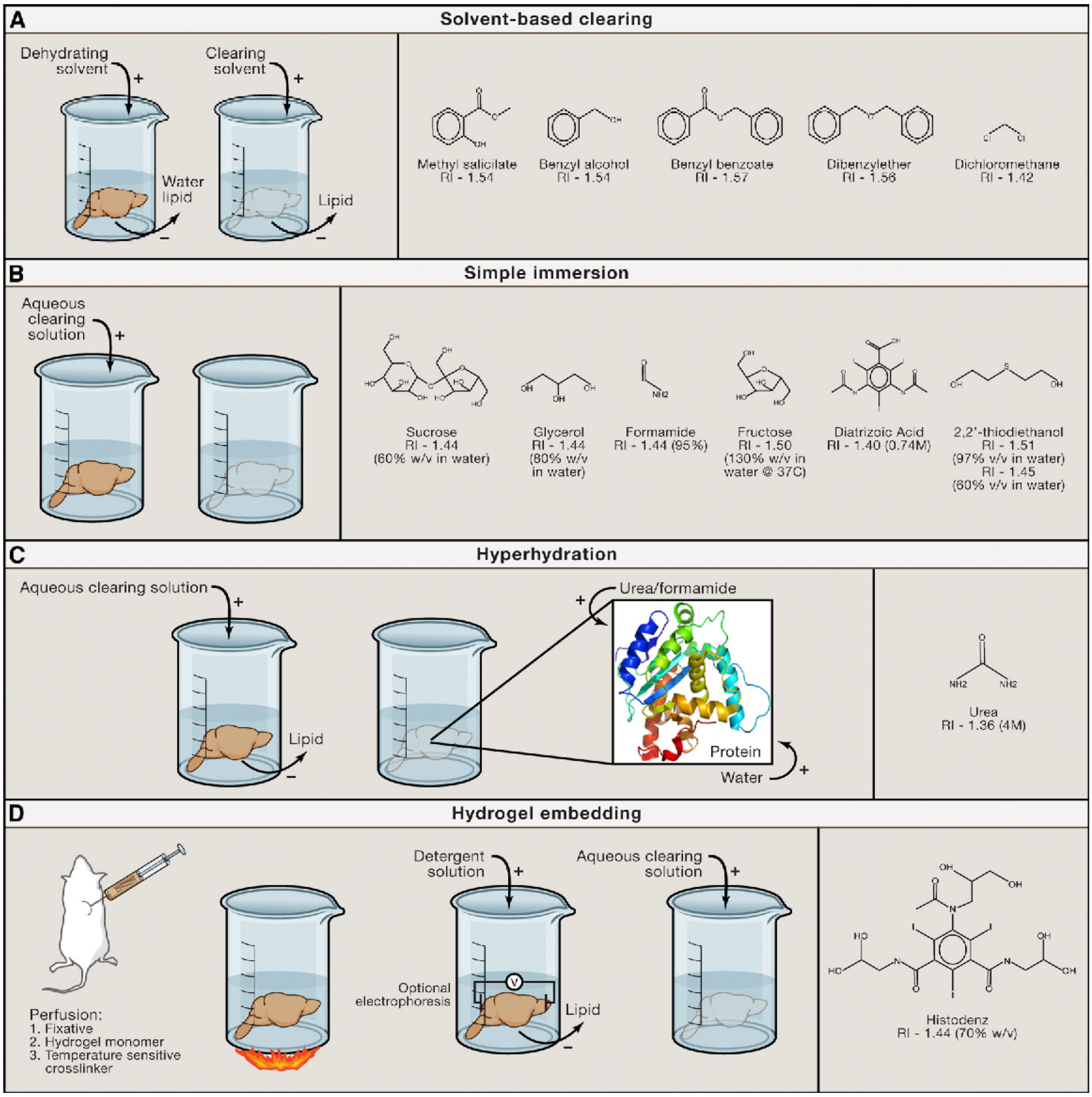

A
Solvent-based clearing
Dehydrating solvent
+
Water lipid
−
Clearing solvent
+
Lipid
−
Methyl salicilate RI - 1.54
Benzyl alcohol RI - 1.54
Benzyl benzoate RI - 1.57
Dibenzylether RI - 1.56
Dichloromethane RI - 1.42
B
Simple immersion
Aqueous clearing solution
+
Sucrose RI - 1.44 (60% w/v in water)
Glycerol RI - 1.44 (80% w/v in water)
Formamide RI - 1.44 (95%)
Fructose RI - 1.50 (130% w/v in water @ 37C)
Diatrizoic Acid RI - 1.40 (0.74M)
2,2'-thiodiethanol RI - 1.51 (97% v/v in water) RI - 1.45 (60% v/v in water)
C
Hyperhydration
Aqueous clearing solution
+
Lipid
−
Urea/formamide
+
Protein
Water
+
Urea RI - 1.36 (4M)
D
Hydrogel embedding
Perfusion:
1. Fixative
2. Hydrogel monomer
3. Temperature sensitive crosslinker
Detergent solution
+
Optional electrophoresis
V
Lipid
−
Aqueous clearing solution
+
Histodenz RI - 1.44 (70% w/v)

Figure 1.3.: **Tissue clearing methods render specimens transparent by equalizing the refractory index (RI).** There are 3 main approaches to tissue clearing. A) Solvent-based tissue clearing. B-C) Water-based tissue clearing. D) Hydrogel-based tissue clearing. Taken from Richardson and Lichtman 2015

1.2.3. Obtaining selective signal contrast

However while clearing the tissue is required to enable large-scale volumetric microscopy, it is not sufficient. Imaging is performed to visualize desired structures of interest and thus, requires a form of (optical) contrast to the background tissue. Hence, the structures of interest (for instance, certain cell types) need to be stained in a way that is compatible with the tissue clearing procedure.

There are at least 4 fundamentally different approaches to achieve this. Laboratory mice can be genetically engineered such that certain cells endogenously express fluorescent proteins. Blood vessels, for instance, can be stained by injecting fluorescent dyes (see our publication in Section B.2; Todorov et al. 2020). Furthermore, so-called immunolabeling makes use of antibodies to selectively target cells of interest with fluorescent components (Renier et al. 2014). Lastly, specialized nano-particles can be used to target desired cells, enabling fluorescent signals with high signal-to-noise ratio (Cai et al. 2018).

1.3. Light-sheet fluorescent microscopy

Once a tissue sample is rendered transparent and structures of interest have been selectively stained with fluorophores, the prerequisites are fulfilled to take optical microscopy from 2D to 3D. Dodt et al. 2007 were the first to achieve this with the help of a light-sheet fluorescence microscope.

1.3.1. Volumetric signal acquisition

The basic working principle is shown in Fig. 1.4. For confocal microscopy, the excitation source illuminates the entire sample from the direction of the objective; the detection region is confined to the confocal plane, which enables volumetric signal acquisition. For light-sheet microscopy, this is achieved by selectively illuminating planes using an excitation source perpendicular to the detection to the objective. Besides enabling a wide-field acquisition of imaging data, light-sheet microscopes also spare out-of-plane regions for illumination, reducing detrimental effects such as photo bleaching of samples.

The combination of tissue clearing with light-sheet fluorescent microscopy enables to combine the large field of view of volumetric imaging modalities such as CT or MRI with the high resolution and selective staining of fluorescent microscopy. This approach allows acquiring whole-body scans of entire animals at cellular resolution, a breakthrough in biomedical imaging.

However, it is noteworthy that some challenges remain. First, the illumination is

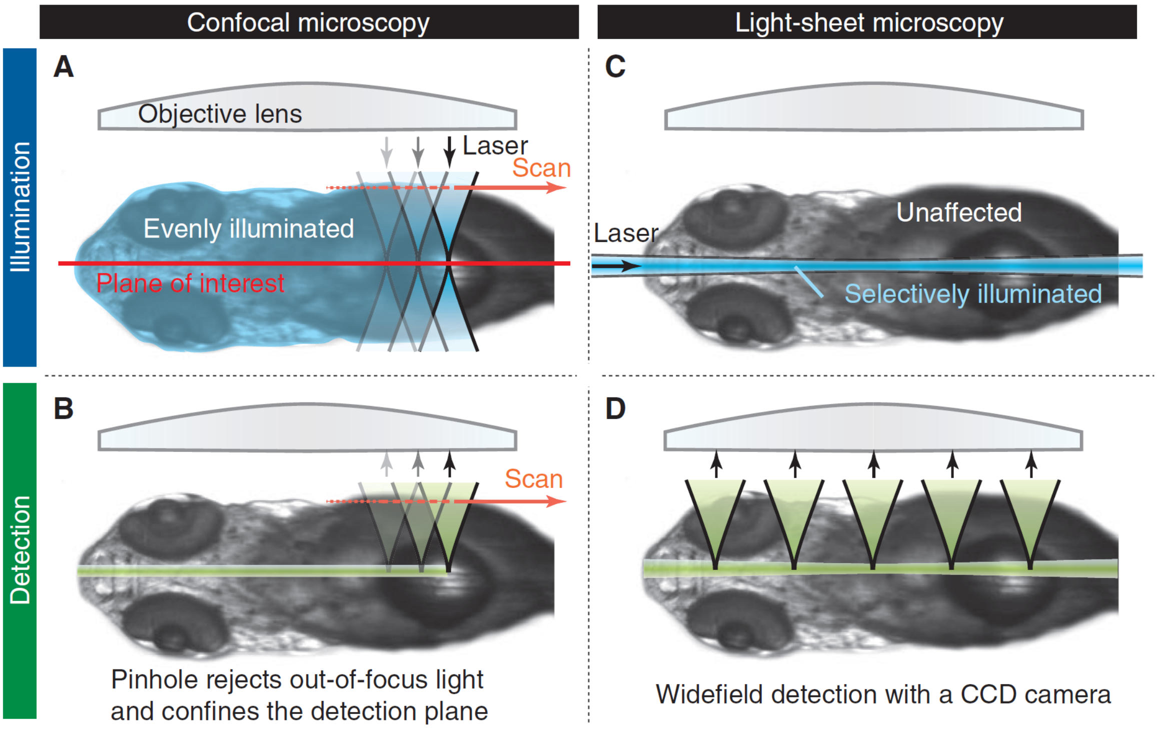


Figure 1.4.: **Volumetric microscopy**. In contrast to confocal microscopy (A,B), in light-sheet microscopy only a selected plane of the specimen is illuminated and emits fluorescent light that is acquired by the sensor. Taken from Huisken and Stainier 2009

not guaranteed to be uniform within the plane of interest, causing spatial gradients in the acquired signal strength. Second, the light sheet is not perfectly planar and not infinitesimally thin. This causes an anisotropy in the resulting resolution and image quality (see Fig. 1.5).

1.3.2. Reconstruction of whole-body scans

Due to technical limitations detailed below, it is not possible to directly acquire volumetric whole-body scans at cellular resolution but they need to be reconstructed from a set of scans from smaller sub-regions.

The first limitation comes from a trade-off between resolution and field of view. Choosing a highly magnifying objective reduces the field of view that can be acquired at once. To cover areas larger than that requires to move the objective laterally to the next region. This process is called *tiling* and the individual tiles are recombined post acquisition. Since the sample is not moved in the chamber and the position of the objective can be precisely monitored, comparatively simple computational

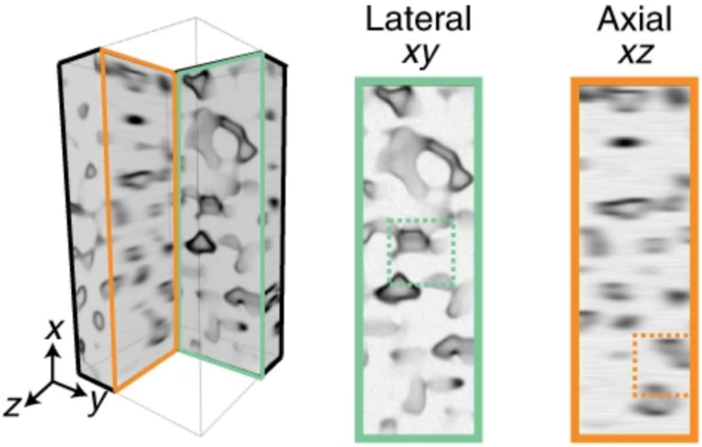


Figure 1.5.: **Anisotropy in light-sheet microscopy**. The imperfection of the *light sheet* reduces spatial resolution and image quality along the axial dimension of the acquired volumetric scan. Taken from Weigert et al. 2018

procedures such as rigid registration are sufficient to recombine the tiles at high precision. However, the tiles are typically unevenly illuminated, causing a signal peak in the center of the tile and lower signal strengths at the borders. After recombination, this can yield a stripe-like pattern in the reconstructed scan. This is an important limitation for further automated image processing, as detailed further below.

The second limitation is associated with the maximum acquisition depth in the tissue. Even after tissue clearing, the tissue is not perfectly transparent and shows residual attenuation and scattering of signals. This limits the maximum acquisition depth. Volumetric scanning of large specimens such as entire mice thus benefits from a two-step approach. First, the upper half of the specimen (facing the objective) is scanned. Then, the sampled is turned upside-down to scan the second half of the specimen. This yields two subsets of volumetric scans. However, the recombination is more challenging than for the tiles since the exact location of the re-positioned sample is difficult to assess. Thus, the recombination requires a more complex *stitching* using potentially non-rigid registration. Not only is this computationally expensive but this step can also introduce artifacts along the interface.

1.4. Biomedical image analysis

Driven by macro-trends of aging population, professionalization of health systems, and technological advances, the use of medical imaging has dramatically increased over the past decades - a development accompanied by an equally dramatic increase in associated cost (see Smith-Bindman et al. 2008). The growing volume of medical images and cost reduction pressures fuel the need for automation in analysis of biomedical images.

1.4.1. Differences to natural image analysis

The scientific field of biomedical image analysis is tightly coupled with the fields of general computer vision and analysis of natural images. Traditionally, much of the progress in medical image analysis has been driven by innovation in natural image analysis, a trend that became even more apparent with the rise of deep neural networks (also see Tajbakhsh, Shin, et al. 2016).

However, it is important to appreciate the fundamental differences between natural images (e.g., photos) and medical images. First, biomedical images may follow different statistics since they are largely visual representations of reconstructions from acquired signals. Second, natural images are typically 2D data with 3 color channels; medical images, in contrast, are typically volumetric (3D) grey-scale data. Third, publicly available data sets and annotations for data sets tend to be orders of magnitude smaller than for natural images. While data sets with 100-200 samples are considered large in the field of biomedical image analysis (e.g., see Rosenhain et al. 2018), popular data sets for natural images are in the range of millions of samples (e.g., ImageNet; see Deng et al. 2009).

As discussed in the sections further below, all these differences have intricate effects on what kind of problems need to be solved and how they need to be solved to enable automated analysis of biomedical images. This caused the field of biomedical image analysis to differentiate from natural images, forming a scientific community of its own.

1.4.2. Fundamental problem classes in image analysis

While the scientific and practical problems addressed in the field of biomedical image analysis are manifold, most work can be associated with one of three fundamental problem classes - for natural as well as for biomedical image analysis.

Classification

Classification describes the task of deriving a categorical, global decision on basis of the presented image sample. Given a pre-defined list of *classes*, the image needs to be allocated to one of these typically mutually exclusive classes. In the field of biomedical image analysis, this could be the decision whether a tumor is benign or malignant, for example.

Detection and localization

Detection and localization are two related tasks of deciding *whether* an object or feature of interest is present (detection) and *where* in the image it is (localization). In biomedical settings, this could be the task of finding the bounding box around a tumor in an image.

Segmentation

Segmentation takes the task of localization further to the level of finding the exact outline of that object, associating each pixel with that object or as background. As a further variant of this, *instance segmentation* would further keep several instances of an object class apart. For instance, by individually segmenting several metastases of a tumor in a single image.

Algorithmic image analysis is as old as digital images themselves. However, the rise of machine learning has brought a fundamental shift in image analysis from rule-based to learning-based approaches - as described below.

1.4.3. Rule-based image analysis algorithms

Traditionally, images have been analyzed by designing a set of custom-tailored rules (explicit algorithms) to solve a specific problem. Without the claim of completeness, such rule-based approaches often rely on *explicit feature detection* and *intensity-based segmentation* as core building blocks to solve a given task. Here, the term *feature* often refers to low-level features such as certain intensity gradients, edges, or simple visual patterns. These can be detected by applying transforms to the image, for example the convolution with a fixed, explicitly defined filter kernel. For instance, edges in images can be detected by convolution with discrete Laplace Gaussian operators. An example can be seen in Fig. 1.6A. Combining and further analyzing sets of low-level features allows to detect or localize more abstract features, for example outlines characteristic of brain tumors (for example, see Sharma et al. 2012).

An example for a rule-based approach to segmentation is the *distance transform watershed*. Here, the image is first binarized and then for each white pixel, the distance to the nearest black pixel is computed (distance transform). Subsequently, a watershed algorithm can segment gray-scale objects even if they were previously overlapping. This method has been successfully used for decades, for example for cell segmentation (see Malpica et al. 1997).

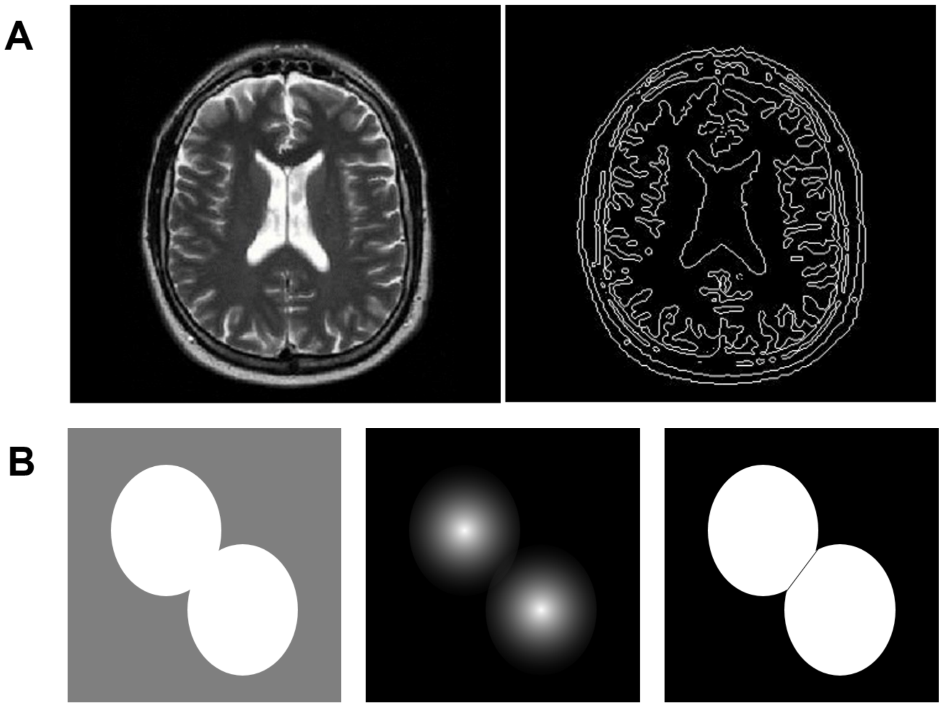


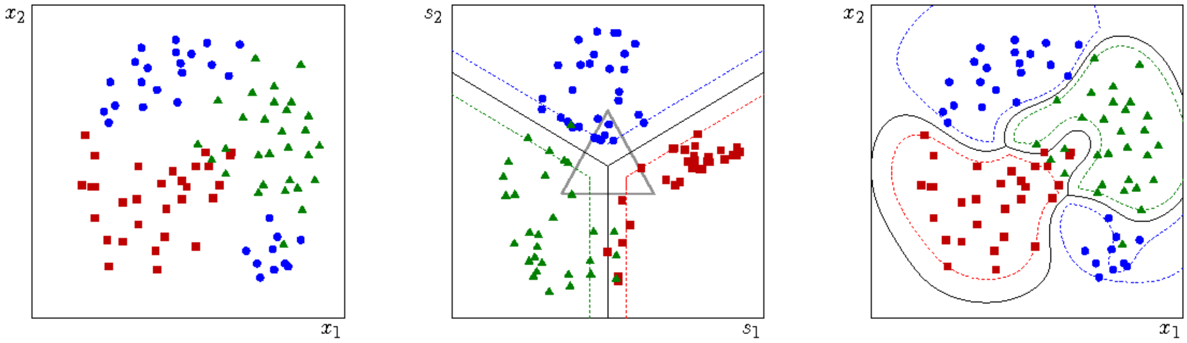
Figure 1.6.: **Rule-based image analysis**. A) Structural analysis of a brain MRI scan using an Laplacian Gaussian filter operator; taken from Gunawan et al. 2017 B) Watershed algorithm (with distance transform) to segment instances of blobs.

1.4.4. Learning-based approaches to image analysis

Learning-based approaches fundamentally differ from rule-based approaches. Instead of defining a specialized set of rules how to solve a problem, a generic algorithm *learns* to solve the problem from the data - typically using pairs of data and corresponding solutions (*supervised learning*). More specifically, *learning* means that the parameters of a generic algorithm are automatically and iteratively fitted so that the algorithmic output converges to the desired solution.

Support vector machines

The first learning-based approaches in image analysis were built on top of rule-based feature extraction methods, for instance by classifying a set of features from an image using support vector machines (SVM) (also see O'Mahony et al. 2019 for a comprehensive overview on the relation of rule-based to learning based methods for image analysis). A SVM divides data samples into subspaces by finding boundaries that optimally separates samples of distinct groups (see Fig. 1.7). This can be achieved with linear or with non-linear boundaries (hyperplanes) by transforming the data into a higher dimensional space in which the samples can be separated linearly. SVMs optimize the hyperplanes to minimize the heterogeneity of samples withing each subspace and to maximize the space around the boundaries that are free of any samples.



(a) Input space

(b) Simplex space

(c) Input space with boundaries

Figure 1.7.: **Support vector machines (SVM)**. The space of data samples with known classes (a) can be subdivided with linear (b) or non-linear (c) boundaries to optimally separate the sample groups. Taken from Van Den Burg and Groenen 2016

Artificial neurons

The concept of an artificial neuron forms the basic building block of neural networks, the most dominant class of machine learning approaches in image analysis. Inspired by the principles of neural information processing, artificial neurons mimic the process of neural excitation via synapses and the propagation of the signal to next neurons (see Fig. 1.8). Each neuron receives input from several other neurons, with weighting that mimics the synaptic strength. The excitation state of the neuron is thus driven by a weighted sum of inputs.

In biological neurons, not every excitation causes the *activation* (discharge of action potential) of the neuron; this only occurs if a certain threshold is passed. This behavior is modeled with a non-linear *activation function* ϕ, which characterizes the mapping of excitation state to the output that is passed to the next signal. Typical non-linear activation functions are the sigmoid function of linear rectification. Mathematically, the activation function expands the computational capabilities of an artificial neuron from linear operations (weighted sum) to non-linear operations.

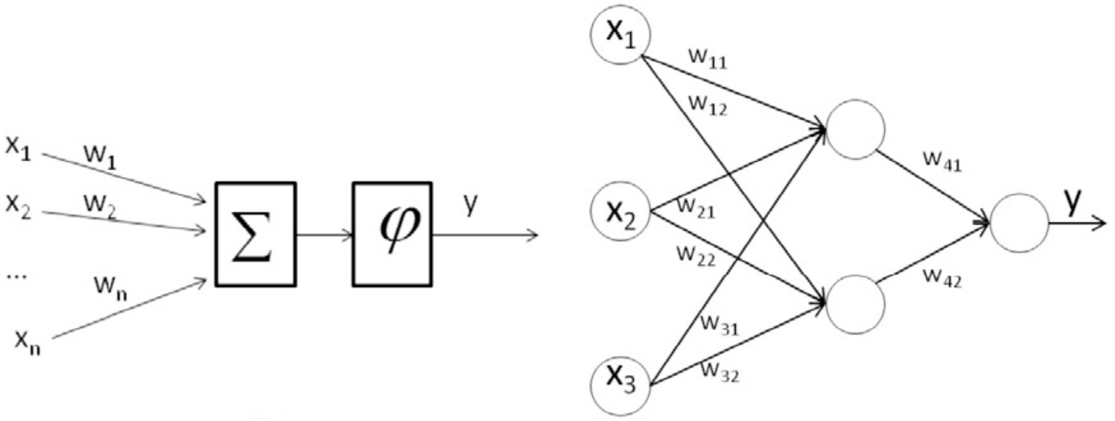


Figure 1.8.: **Artificial neural networks**. The concept of an artificial neuron forms the basic building block (left). A perceptron is the simplest architecture of artificial neural networks (right). Taken from Di Noia et al. 2013

Perceptron

Combining several artificial neurons to a three-layer network forms a perceptron, the simplest architecture of artificial neural networks (see Fig. 1.8). Despite its simple structure, perceptrons are, in theory, already capable of approximating any arbitrary function mapping the input x to the output y (see Irie and Miyake 1988). In such setups, the first layer is termed the *input layer*, which is follow by the *hidden layer*. The number of hidden layers and the number of neurons per layer determine the computational power of the network but also the complexity (number of parameters) of the network - a common trade-off.

If the number of layers increase, these networks are often calle1 *deep networks*. To differentiate them from other architectures (e.g., convolutional neural networks; see below), these architecture are also commonly referred to as *fully connected neural networks*. This means that every neuron of a given layer receives input from all neurons of the previous layer, not only a subset of them. Learning to solve a given task for neural networks requires to fit the network weights in order to derive the desired output y. Achieving this has been and remains the key challenge in *deep learning*.

Backpropagation and gradient descent

Although dating back to the 1960s, the backpropagation algorithm was only established in the late 1980s as the governing principle in training neural networks (Rumelhart et al. 1986). The idea of backpropagation is to assess the error between predicted output $\hat{y}$ and desired output y, and propagate it backwards through the entire network in order to iteratively adjust all network weights into the right direction. To determine whether a given network weight w needs to be increased or decreased in order to reduce the error E, the gradient of the error with respect to the weights is computed for all weights:

$$\nabla E[\vec{w}] \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \cdots \frac{\partial E}{\partial w_n}\right]$$

The training of the network thus occurs by gradually and iteratively updating the weights into the opposite direction of the gradient. This happens layer by layer. The update step size can be controlled with the parameter η, often referred to as the *learning rate*:

$$\Delta\vec{w} = -\eta\nabla E[\vec{w}]$$

$$\Delta w_i = -\eta\frac{\partial E}{\partial w_i}$$

In order to achieve this, the non-linear activation function of the artificial neurons needs to be differentiable. On a global level, the entire procedure follows the concept of *gradient descent* to find the (potentially local) minimum of the differentiable error function. A large variety of optimization approaches have been developed to achieve this efficiently and robustly, e.g. the *Adam Optimizer* (see Kingma and Ba 2014).

Convolutional neural networks

While a sufficiently large *fully connected network* can, in principle, approximate any arbitrary function (as shown before), there are important practical limitations to this approach. The number of network parameters that need to be fitted increases exponentially with the numbers of layers and units per layer. Thus, even moderately deep networks cannot be practically fitted with a typically limited amount of training samples. With a special focus on image analysis, a major breakthrough was achieved with the introduction of *convolutional neural networks* (CNNs, see Mairal et al. 2014).

In contrast to their fully connected counterparts, a unit in a CNN only receives input from a small number of spatially proximate units from the previous layer (see left panel of Fig. 1.9). A single convolutional kernel is moved over the entire spatial range of the input and only the parameters of that comparatively small kernel need to be fitted for that layer. While each given unit only receives direct input from a small number of units from the prior layer, the units in the last layer *indirectly* receive input from a much larger region, if not the complete region, of the input. This enables the network to develop a hierarchical representation of ever more complex, abstract

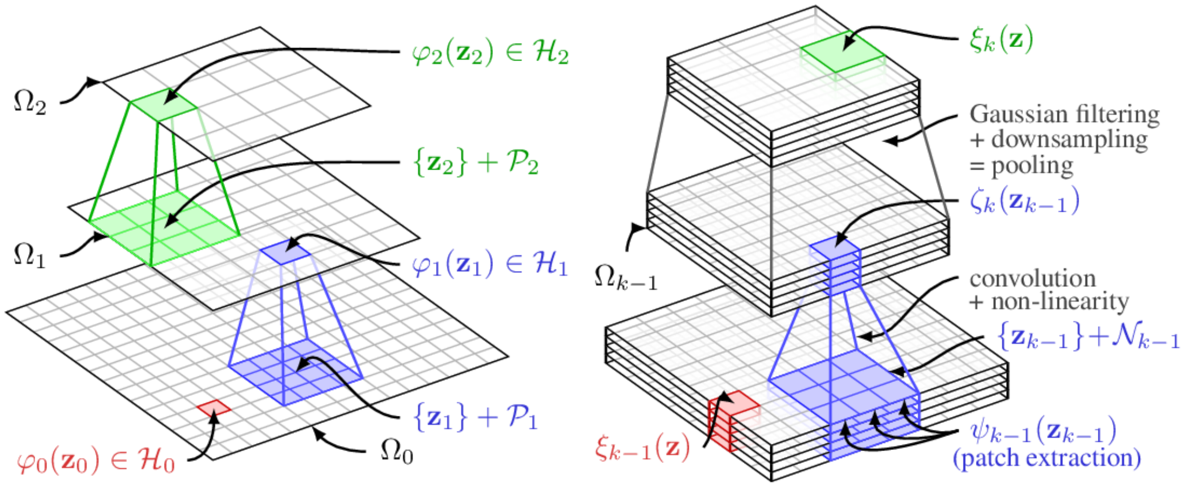

Ω_2
$\varphi_2(\mathbf{z}_2) \in \mathcal{H}_2$
$\{\mathbf{z}_2\} + \mathcal{P}_2$
Ω_1
$\varphi_1(\mathbf{z}_1) \in \mathcal{H}_1$
$\{\mathbf{z}_1\} + \mathcal{P}_1$
$\varphi_0(\mathbf{z}_0) \in \mathcal{H}_0$
Ω_0
$\xi_k(\mathbf{z})$
Gaussian filtering
+ downsampling
= pooling
$\zeta_k(\mathbf{z}_{k-1})$
Ω_{k-1}
convolution
+ non-linearity
$\{\mathbf{z}_{k-1}\}+\mathcal{N}_{k-1}$
$\xi_{k-1}(\mathbf{z})$
$\psi_{k-1}(\mathbf{z}_{k-1})$
(patch extraction)

Figure 1.9.: **Convolutional neural networks**. A convolutional neural network is based on (spatial) convolutions on a selected subset of units from the previous layer (left). This concept can be extended along a feature dimension (right). Taken from Mairal et al. 2014

features that may not be restricted to small subregions of the input data - without the need of connecting every single possible combination of units.

As visualized in the right panel of Fig. 1.9, this concept can be further extended along a *feature dimension*. A unit then not only contains one data point to represent a feature of a given spatial location but a set of data points. In such setups, the convolutional kernel is extended by this dimensionality. In combination with spatial reduction techniques such as pooling operations, CNNs often encode more abstract representations (large number of feature channels) that are less tied to a spatial dimension in deep layers.

The U-net architecture

While a large variety of different and highly specialized CNN architectures have been developed for classification, detection/localization, and segmentation tasks in image analysis, one network architecture stands out for biomedical image segmentation: the U-net. Introduced in 2015, it drastically improved the state-of-the-art for such tasks and has emerged as the standard in its class (see Ronneberger et al. 2015; this article alone has received more than 15,000 citations as of mid 2020). The U-net architecture is depicted in Fig. 1.10. At its core, it consists of convolutional layers, pooling layers, and so-called *skip connections*. To derive fine pixel-wise segmentations of complex objects in an image, it combines non-localized, high-level abstract feature representations with highly localized, low-level feature representations.

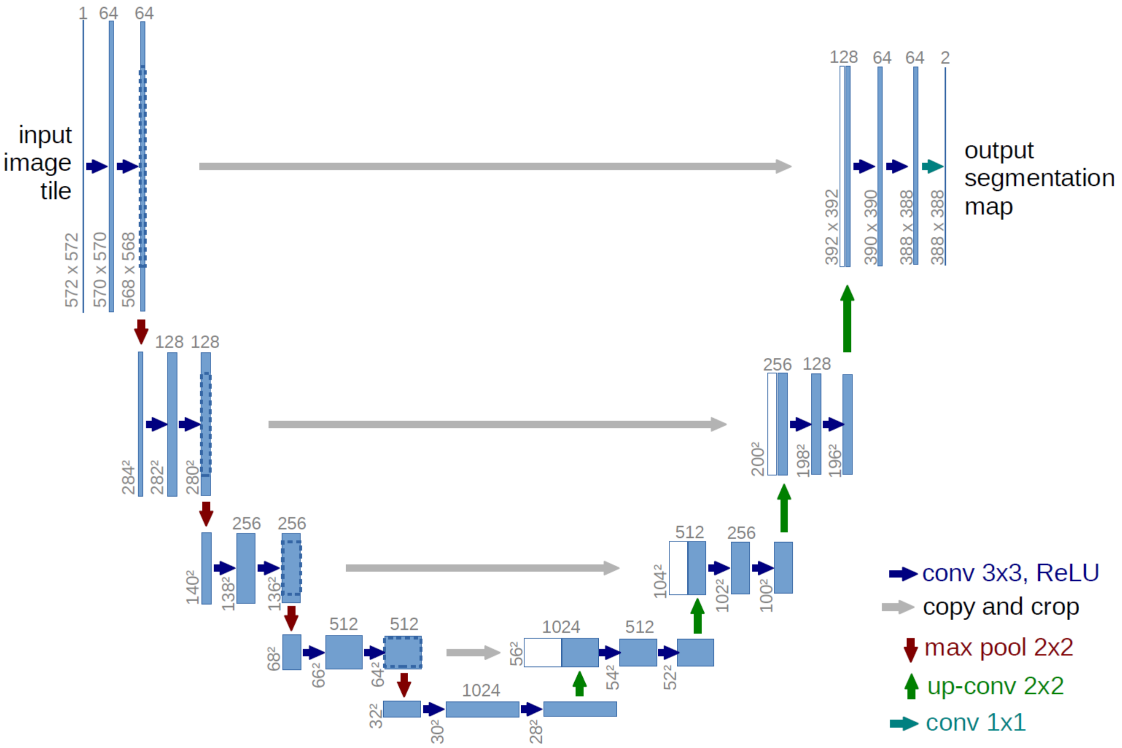


Figure 1.10.: **The U-net architecture**. The U-net architecture is a specialized CNN that has proven especially successful for segmentation tasks in biomedical images. Taken from Ronneberger et al. 2015

This is achieved by using a deep stack of encoding units, which detect characteristic object features, and a corresponding stack of decoding units, which segment each object at pixel-level. Each encoding unit performs two convolutions, extracting information about the environment for each pixel and representing that information in a third dimension — the feature channels. Before being passed on to the next encoding unit, the image is spatially downsampled with a pooling operation. Together, this means that the neural network is steadily increasing the feature channels and steadily decreasing the spatial resolution, enforcing the network to learn even more abstract features in the deeper layers, before mapping the information relevant to the object of interest back to the original resolution in the decoding upward path. This happens by upsampling the abstract, low-resolution information from lower layers and concatenating it with the less abstract, but higher-resolution information from the encoding path via skip connections.

Optimized variants of the U-net architecture form a backbone of many contributions presented in this dissertation. As described in more detail further below, enabling adoption of deep learning methods in clinical and pre-clinical settings is at the core of the motivation of this dissertation and the use of widely used architectures like the U-net is thus a well-considered design choice.

1.4.5. Bottlenecks in adoption of learning-based approaches

It may seem that the fundamental challenges of biomedical image analysis were largely solved, especially with respect to the dramatic breakthroughs achieved with the recent developments in deep learning. However, in practice the adoption of deep learning in biomedical research as well as in clinical settings (such as the workflow of radiology assessments in hospitals) is slow (see Thaler and Menkovski 2019). Analyzing the underlying causes of slow adoption reveals that key problems remain unsolved and represent major bottlenecks in practice (see Tajbakhsh, Jeyaseelan, et al. 2020 for a comprehensive overview).

Availability of training data. Even with the development of CNNs (Mairal et al. 2014), deep neural networks still require large amounts of training data in supervised learning. While research on unsupervised or weakly supervised learning methods aims to reduce this need and progresses in achieving this (Fabiyi 2019), supervised learning remains the most effective training strategy. Thus, large amounts of training data are needed. In practice, this is often not the case for two major reasons. First, the data itself in the medical context is often not readily available due to ethical and legal restrictions on sharing and using this data. While some public data sets exist, they tend to be relatively small (as discussed earlier) and are limited to a few specialized use cases. Second, the data needs to be annotated manually to provide a reference. This is especially time-consuming and costly for medical data as it requires deep medical expertise and as the data is often given as volumetric scans, which requires the annotation of hundreds of images for a single sample.

Quality of training data. The quality of available training data poses a further bottleneck. As also shown in this dissertation, even the annotations of highly trained human experts cannot be assumed to reflect a commonly agreed-upon *ground truth*. On the one hand, annotation tends to be a tiring, repetitive task that is prone to human error due to insufficient attention and diligence. On the other hand, medical images are often intrinsically hard to interpret and not completely conclusive (Warfield et al. 2006; Jungo et al. 2018). Training neural networks on partially flawed data samples not only reduces the overall performance of the final model but also can lead to unexpected behaviour as the network may mimic the errors present in the training data.

Generalization and robustness to variability. Furthermore, neural networks tend to be not very robust to unexpected variability in the data. Even though a trained network may perform well in a given task of a given data set (e.g., segmenting a brain tumor in an MRI scan), small changes in the data may drastically reduce the utility of the trained network. In the given example, it may be enough to use an MRI scanner from a different manufacturer to render the trained algorithm completely useless. Also, small changes in the task may require partial or complete retraining of the network. For example, if the task where not to segment a brain tumor in an MRI scan but a region of ischemic stroke, a large part of the work needs to be repeated. This lack of robustness, scalability, and generalization poses a bottleneck for the adoption in practical settings where such kind of variability in data and tasks are unavoidable.

1.5. Ambition and contribution of this dissertation

Motivated by the observations described in section 1.4.5, the underlying motivation and ambition of this dissertation is to enable a more wide-spread adoption of powerful deep learning approaches in clinical and pre-clinical image analysis. Tackling the bottlenecks listed above can reduce the technical barrier of adopting deep learning solutions and reduce the cost of doing so.

The main contributions of this dissertation are driven by this ambition and centered around finding strategies to overcome scarcity of training data, to deal with imperfect annotations, and to improve generalization and robustness. In chapter A.1, a highly data-efficient approach was developed to solve the 3D task of localizing and segmenting tumor metastases with a 2D network, which requires substantially smaller training data sets. The approach was shown to generalize across different tumor types despite substantially different metastatic characteristics. In chapter A.2, the generalizability of neural network training was assessed across highly different domains: vessel segmentation in human brains in MRI scans, vessel segmentation in mouse brains in 3D light-sheet microscopy, and segmentation of the peripheral nervous system in mouse bodies in 3D light-sheet microscopy. Further, a neural network was designed to segment the main organs in whole-body scans of mice, not only exceeding state-of-the-art in performance but also providing localized measures of contradicting human interpretation of the data. Lastly, additional work explores further applications on (imaged-based) learning models biomedical modeling in the field of pre-clinical neuroscience on the basis of image-like spectral representations of auditory stimuli (chapter A.4 and chapter A.5) and assesses the performance of these models (chapter A.3). Interestingly, this image-based approach for biomedical modeling of neuronal information processing was shown not only to be very effective but also revealing about the underlying biological dynamics.

2. Discussion and outlook

In this chapter, the main contributions of this dissertation will be reflected in the light of the problem statements presented in section 1.4.5 and ambitions presented in section 1.5. Finally, a perspective on future developments and concluding remarks provide a more opinionated view on the scientific context of this dissertation.

2.1. Resolving bottlenecks for deep learning in biomedical image analysis

The introductory chapter described the key breakthroughs that led to the rise of deep learning in biomedical image analysis (also see Maier et al. 2019) and further identified a (non-comprehensive) list of 3 key bottlenecks that slow or hinder the adoption in practice. Examples of these bottlenecks from the work presented in this thesis are discussed below in order to derive prerequisites of general relevance to resolve these bottlenecks.

Availability of training data. The scarcity of large, curated, and annotated biomedical datasets is a widely appreciated problem (Willemink et al. 2020). This problem aggravates for highly resolved volumetric scans with large fields of view such as the whole-body light-sheet microscopy datasets used in this study. A single sample may be in the range of $5,000x5,000x10,000$ pixels (which corresponds to the terabyte scale in file size). Not only is the overall number of samples often limited to the lower 2-digit range but annotation tends to be extremely expensive. Finding all metastases with a diameter of a few pixels in such large volumetric scans took an educated experts more than 2 months for a sample size of 2 (see the work in Chapter A.1). The cost of annotation even increases for curvilinear structures such as blood vessels (see the work in Chapter A.2 and B.2). Thus, a first prerequisite for wide-spread adoption of deep learning in biomedical image analysis is the development of strategies for highly data efficient training of networks.

Generalization and robustness. Once a model has been trained, seemingly small changes in the model requirements, in the biological sample, or in the data acquisition may drastically reduce the utility of the model. While the need for generalization and robustness against these changes is widely appreciated, it remains an active field of research with many unsolved problems (Raghu et al. 2019). Especially in pre-clinical settings of biomedical research, experimental protocols and imaging procedures are inherently non-standardized and evolving. Also for the work presented in this dissertation, changes in the clearing protocol, the staining method, or the microscopic setup affect the characteristics of the acquired image. A second prerequisite to ensure sustainability and scalability of deep learning solutions is a deep understanding of transfer learning and domain adaption.

Quality of training annotations. Supervised machine learning algorithms rely on a reference (here: image annotations) to learn a task. Often, this reference is referred to as *ground truth*, which may cause misunderstandings as the annotations of a single human expert are neither guaranteed to be correct nor to be objective or unbiased (Tajbakhsh, Shin, et al. 2016). As also shown in Chapters A.1 and B.1, human error and bias can be substantial. A third prerequisite for successful adoption of deep learning is thus to appreciate *intrinsic data uncertainty* and *the defectiveness of human annotation*.

2.2. Detection of cancer metastases

The work in Chapter A.1 establishes DeepMACT, an integrated pipeline for analysis of cancer metastases in mice. The pipeline combines the steps of tissue clearing, 3D light-sheet fluorescent microscopy, deep learning based metastasis detection, and subsequent statistical analysis. Resolving several prior limitations along each step, this study marks the first time an animal could be screened for all metastasis throughout the entire body, even detecting single disseminated cancer cells. Beyond that, also the effectiveness of therapeutic antibodies could be assessed by determining which metastases were successfully targeted and which ones were missed. The study received substantial attention from the scientific community (for example, it was featured as a *Research Highlight* in Le Bras 2020) and from the general media (e.g., from the *Federal Ministry of Education and Research* as part of *Wissenschaftsjahr 2019*[1], from *Frankfurter Rundschau*[2], from N-TV[3]). Furthermore, it was awarded the *Rolf Becker-Preis 2020* (EUR 50,000) for the "best original work in the entire field of experimental or clinical medicine as a result of a research project affiliated with Ludwig-Maximilians-University Munich". DeepMACT overcomes one of the central limitations, scarcity of training data, by solving the 3D task of detecting, localizing, and segmenting small metastases in large volumetric scans with a 2D architecture. In short, subsamples

[1] https://www.wissenschaftsjahr.de/2019/neues-aus-der-wissenschaft/dezember-2019/neuer-algorithmus-erkennt-automatisch-krebsmetastasen/

[2] https://www.fr.de/wissen/krebs-metastasen-spur-13369131.html

[3] https://www.n-tv.de/mediathek/videos/wissen/KI-Lupe-spuert-kleinste-Metastasen-auf-article21455498.html

of data are projected along all 3 spatial dimensions; each projection is analyzed and building the outer product of each 2D analysis enables recombination and subsequent 3D reconstruction at high precision. Training a 2D network requires exponentially less training data (also compare Ronneberger et al. 2015 and Çiçek et al. 2016), allowing DeepMACT to be trained effectively with very few annotation samples - reaching a detection performance comparable to that of a human expert (also see Vestjens et al. 2012; Ehteshami Bejnordi et al. 2017). Furthermore, DeepMACT also addresses the need for generalization. The approach of localized detection in subsamples of whole-body scans enables the network to generalize across different cancer lines despite their different organotropic metastatic characteristics (see Nguyen et al. 2009; Hingorani et al. 2003; Schonhuber et al. 2014; Iorns et al. 2012). Here, DeepMACT was trained on data from mice with metastases from human MDA-MB-231 mammary carcinoma but successfully detected metastases from other breast cancer lines as well as prancreatic and lung cancer.

2.3. Transfer learning across biomedical domains

The work in Chapter A.2 analyzes how deep learning solutions can be generalized across biomedical domains on the example of curvilinear structures. Specifically, it analyzes to which degree transfer learning can help for the segmentation of blood vessels (also see F. Zhao et al. 2019), and the peripheral nervous system. This work tests the limits of generalization by training a single network architecture for fundamentally different biomedical data domains and by cross-predicting across these domains: blood vessels in MRI scans of human brains (Tetteh et al. 2018), blood vessels in light-sheet microscopy scans of mouse brains (further work on this was published and is listed in Chapter B.2), and the peripheral nervous system in light-sheet microscopy scans of entire mice. In line with literature (Van Opbroek et al. 2015; Khan et al. 2019), transfer learning was shown to be effective despite the large shift in domains - especially under the constraint of sparsity of training data. For example, a network trained on a large, but distant dataset (synthetic data and human MRI data) matches the performance of the same network trained on small training dataset from the target domain. One important aspect of this work is the transfer from synthetically generated training data (Schneider et al. 2012). Despite its obvious differences to the target domains, exploiting it for transfer learning consistently helped across all domains, suggesting its importance in strengthening generalization. This approach also inspired further work beyond the biomedical community (CVPR 2020: Parshotam and Kilickaya 2020).

2.4. Whole-body organ segmentation

The work in Chapter B.1 introduces AIMOS, a deep learning based pipeline for the automated segmentation of major organs (brain, heart, lungs, liver, kidneys, and spleen) and the skeleton in whole-body scans of mice. The approach generalizes across imaging modalities (here: contrast-enhanced as well as native micro-CT and two variants of light-sheet microscopy). It works orders of magnitude faster than prior, rule-based methods (see H. Wang, Stout, et al. 2011; Van Der Heyden et al. 2018) and outperforms them in terms of segmentation quality (also see Akselrod-Ballin et al. 2016; Yan et al. 2017). The segmentation performance matches or exceeds the quality of human expert annotations for all organs. Importantly, it also follows a data efficient training strategy. The AIMOS network can be trained with as little as 10 samples, at which it already reaches around 90% of its maximum performance. Again, transfer learning was shown to be an effective method to further reduce the need for annotated training data, underlining the generalizability of the approach. For example, if pre-trained on a public contrast-enhanced CT dataset (Rosenhain et al. 2018), AIMOS already reached 84% of its maximum performance on native CT data (which has much lower contrasts) after being trained on a single data sample. Another important contribution from this study focuses on errors and bias in human expert labels and on intrinsic data ambiguity. The variability of human annotations was already discussed by prior literature (e.g., Warfield et al. 2006). Kohl et al. 2018 addressed this by training networks to mimic this behavior and producing equally variable predictions. The work in Chapter B.1 takes a different angle of view and addresses this in two ways: first, it shows that evaluating a deep learning model on a test set annotated by the same individual as the training set, a common practice in the field (e.g., Baiker et al. 2010; Khmelinskii et al. 2011; H. Wang, Han, et al. 2019), overestimates the generalization performance as the models learns the individual bias of the human annotator. Thus, at least a second, independently created test set is needed to determine the true performance. Second, it addresses the variability of human annotations by not only predicting the organ segmentations but also those image regions where human annotators are most likely to disagree.

2.5. Image-based models of neuronal processing

Lastly, additional work explores further applications on image-based learning models biomedical modeling in the field of pre-clinical neuroscience (chapter A.4 and chapter A.5) and assesses the performance of these models (chapter A.3). Here, auditory stimuli are transformed to two-dimensional image-like spectral representations (frequency vs. time). Interestingly, this image-based approach for biomedical modeling of neuronal information processing was shown not only to be very effective but also revealed the underlying dynamics of neuronal processing (see chapter A.4).

2.6. Perspective on future deep learning adoption

For biomedical research in pre-clinical settings, the power of machine learning methods has the potential to truly transform an entire field of science. In combination with novel imaging techniques such as volumetric light-sheet microscopy, which yields unprecedented, cellular detail of large specimens up to entire animals, deep learning based image analysis enables to address biomedical problems at scale and with high precision that would exceed the analytic capabilities of human experts. Towards this vision, efficient training strategies and generalization will be key. Besides strengthening our understanding of *what* and *how* networks learn, the field of biomedical, pre-clinical image analysis would benefit from establishing structures already common in clinical image analysis: public and curated datasets and high-profile challenges. Furthermore, the potential of synthetic training data yields great potential to lower the financial, personal, and technical barrier to adopt deep learning approaches in everyday lab settings. Last but not least, breaking up scientific silos between biology and computer science by fostering deeply integrated collaborations will help bridging the gap between these fields of science. For medical image analysis in clinical settings, the adoption in practice seems to be restricted by different problems. While public challenges, a predominant form of scientific discourse in the field, have fueled technical innovation, they will be less helpful driving *further* improvements. As also shown in this dissertation, the *race* for ever higher performance scores may only seemingly lead to improvements. Critical problems such as the defectiveness of human expert annotations, intrinsic ambiguity in the medical images, and the *long tail of variability* in medical images seem insufficiently appreciated. Algorithms that win challenges may still miserably fail in clinical settings for this reason.

2.7. Concluding remarks

The success of deep learning in biomedical image analysis is astonishing and represents a major step forward for human health. But the research seems too confined to highly controlled settings and the state-of-the-art does not meet the requirements for deployment in practical settings. The bottlenecks for wide-spread adoption in radiological workflows and biomedical research should not be discarded as mere implementation problems. They pose an underappreciated frontier in our field of science and may be key to turning academic achievements into better health.

A. Appendix A: research articles included in this dissertation

A.1. Deep Learning Reveals Cancer Metastasis and Therapeutic Antibody Targeting in the Entire Body

Authors: C Pan*, **O Schoppe***, A Parra-Damas*, R Cai, M Todorov, G Gondi, B v. Neubeck, N Böğürcü-Seidel, S Seidel, K Sleiman, C Veltkamp, B Förstera, H Mai, Z Rong, O Trompak, A Ghasemigharagoz, M Reimer, A Cuesta, J Coronel, I Jeremias, D Saur, A Acker-Palmer, T Acker, B Garvalov, B Menze, R Zeidler, A Ertürk.
**Joint first authorship*

Abstract: Reliable detection of disseminated tumor cells and of the biodistribution of tumor-targeting therapeutic antibodies within the entire body has long been needed to better understand and treat cancer metastasis. Here, we developed an integrated pipeline for automated quantification of cancer metastases and therapeutic antibody targeting, named DeepMACT. First, we enhanced the fluorescent signal of cancer cells more than 100-fold by applying the vDISCO method to image metastasis in transparent mice. Second, we developed deep learning algorithms for automated quantification of metastases with an accuracy matching human expert manual annotation. Deep learning-based quantification in 5 different metastatic cancer models including breast, lung, and pancreatic cancer with distinct organotropisms allowed us to systematically analyze features such as size, shape, spatial distribution, and the degree to which metastases are targeted by a therapeutic monoclonal antibody in entire mice. DeepMACT can thus considerably improve the discovery of effective antibody-based therapeutics at the preclinical stage.

Date of publication: 12.12.2019
Journal name: Cell, Cell Press (Elsevier)

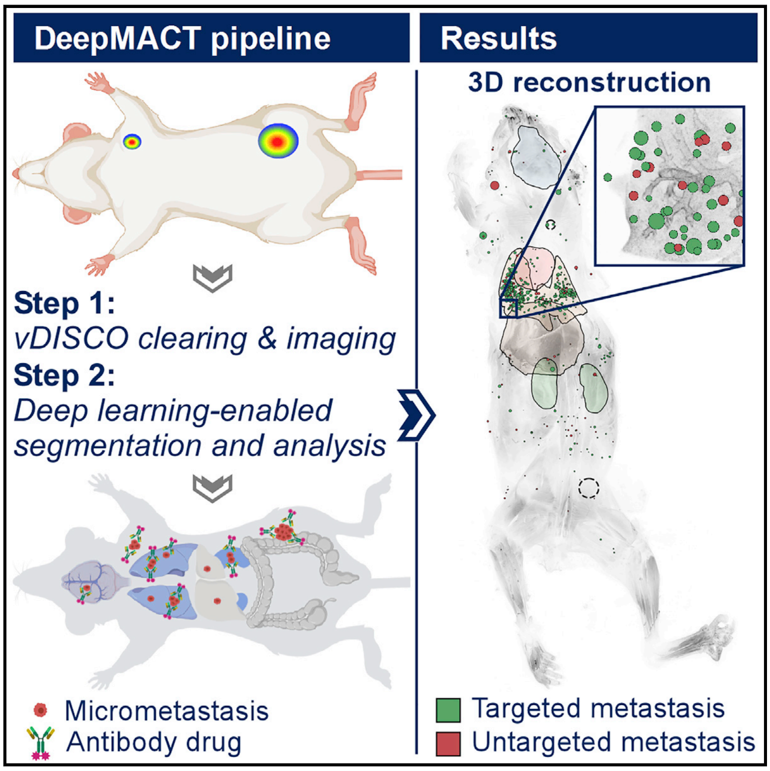
Individual contribution: project coordination, design of quantitative experiments, data processing, conceptual development and implementation of data analysis, co-leading author of manuscript

Resource

Cell

Deep Learning Reveals Cancer Metastasis and Therapeutic Antibody Targeting in the Entire Body

Graphical Abstract



Highlights

- DeepMACT is a deep learning-based pipeline for comprehensive analysis of metastases
- DeepMACT identifies micrometastases and single cancer cells in full-body 3D scans
- DeepMACT reveals the efficacy of antibody-drug targeting in the entire body
- DeepMACT indicates that the tumor microenvironment affects drug targeting efficacy

Authors

Chenchen Pan, Oliver Schoppe, Arnaldo Parra-Damas, ..., Bjoern Menze, Reinhard Zeidler, Ali Ertürk

Correspondence

erturk@helmholtz-muenchen.de

In Brief

Deep learning-based automated detection and quantification of micrometastases and therapeutic antibody targeting down to the level of single disseminated cancer cells provides unbiased analysis of multiple metastatic cancer models at the full-body scale.

Pan et al., 2019, Cell *179*, 1661–1676
December 12, 2019 © 2019 Elsevier Inc.
https://doi.org/10.1016/j.cell.2019.11.013

Resource

Cell

Deep Learning Reveals Cancer Metastasis and Therapeutic Antibody Targeting in the Entire Body

Chenchen Pan,[1,2,18] Oliver Schoppe,[3,4,18] Arnaldo Parra-Damas,[2,18] Ruiyao Cai,[1,2] Mihail Ivilinov Todorov,[1,2,5] Gabor Gondi,[6] Bettina von Neubeck,[6] Nuray Böğürcü-Seidel,[7] Sascha Seidel,[8] Katia Sleiman,[4,9,10] Christian Veltkamp,[4,9,10] Benjamin Förstera,[1,2] Hongcheng Mai,[1,2] Zhouyi Rong,[1,2] Omelyan Trompak,[7] Alireza Ghasemigharagoz,[2] Madita Alice Reimer,[2] Angel M. Cuesta,[8] Javier Coronel,[3] Irmela Jeremias,[11,12,13] Dieter Saur,[4,9,10] Amparo Acker-Palmer,[8] Till Acker,[7] Boyan K. Garvalov,[7,14] Bjoern Menze,[3,4,17] Reinhard Zeidler,[6,15] and Ali Ertürk[1,2,16,19,*]

[1]Institute for Tissue Engineering and Regenerative Medicine (iTERM), Helmholtz Zentrum München, 85764 Neuherberg, Germany
[2]Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig Maximilian University of Munich (LMU), 81377 Munich, Germany
[3]Department of Informatics, Technical University of Munich, 85748 Munich, Germany
[4]Center for Translational Cancer Research (TranslaTUM), Klinikum rechts der Isar, Technical University of Munich, 81675 Munich, Germany
[5]Graduate School of Systemic Neurosciences (GSN), 82152 Munich, Germany
[6]Research Unit Gene Vectors, Helmholtz Zentrum München, 81377 Munich, Germany
[7]Institute of Neuropathology, University of Giessen, 35390 Giessen, Germany
[8]Institute of Cell Biology and Neuroscience and Buchmann Institute for Molecular Life Sciences (BMLS), University of Frankfurt, 60323 Frankfurt, Germany
[9]Division of Translational Cancer Research, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[10]German Cancer Consortium (DKTK), Partner Site Munich, Klinikum rechts der Isar, Technische Universität München, 81675 Munich, Germany
[11]Research Unit Apoptosis in Hematopoietic Stem Cells, Helmholtz Zentrum München, German Center for Environmental Health (HMGU), 81377 Munich, Germany
[12]Department of Pediatrics, Dr. von Hauner Childrens Hospital, Ludwig Maximilian University of Munich (LMU), 81377 Munich, Germany
[13]German Consortium for Translational Cancer Research (DKTK), Partnering Site Munich, 80336 Munich, Germany
[14]Department of Microvascular Biology and Pathobiology, European Center for Angioscience (ECAS), Medical Faculty Mannheim, University of Heidelberg, 68167 Mannheim, Germany
[15]Department for Otorhinolaryngology, Klinikum der Universität München, Ludwig Maximilian University of Munich (LMU), 81377 Munich, Germany
[16]Munich Cluster for Systems Neurology (SyNergy), 81377 Munich, Germany
[17]Munich School of Bioengineering, Technical University of Munich, 85748 Munich, Germany
[18]These authors contributed equally
[19]Lead Contact
*Correspondence: erturk@helmholtz-muenchen.de

https://doi.org/10.1016/j.cell.2019.11.013

SUMMARY

Reliable detection of disseminated tumor cells and of the biodistribution of tumor-targeting therapeutic antibodies within the entire body has long been needed to better understand and treat cancer metastasis. Here, we developed an integrated pipeline for automated quantification of cancer metastases and therapeutic antibody targeting, named DeepMACT. First, we enhanced the fluorescent signal of cancer cells more than 100-fold by applying the vDISCO method to image metastasis in transparent mice. Second, we developed deep learning algorithms for automated quantification of metastases with an accuracy matching human expert manual annotation. Deep learning-based quantification in 5 different metastatic cancer models including breast, lung, and pancreatic cancer with distinct organotropisms allowed us to systematically analyze features such as size, shape, spatial distribution, and the degree to which metastases are targeted by a therapeutic monoclonal antibody in entire mice. DeepMACT can thus considerably improve the discovery of effective antibody-based therapeutics at the pre-clinical stage.

INTRODUCTION

The metastatic process is complex and affects diverse organs (Hanahan and Weinberg, 2011; Lambert et al., 2017; Massagué and Obenauf, 2016). As most cancer patients die of metastases at distant sites developing from disseminated tumor cells with primary or acquired resistance to therapy, a comprehensive and unbiased detection of disseminated tumor cells and tumor targeting drugs within the entire body is crucial (de Jong et al., 2014). Such technology would help to explore mechanisms affecting tumor metastasis and drug targeting in preclinical mouse models much more reliably, hence substantially contributing to the development of improved therapeutics. So far, such efforts have been hampered by the lack of (1) imaging technologies to reliably detect all individual metastases and disseminating tumor cells in mouse bodies, and (2) algorithms to quickly and accurately quantify large-scale imaging data. Here, we developed an analysis pipeline that allows us to efficiently overcome these limitations.

First, we built upon recently developed tissue clearing methods for entire fixed mice (Cai et al., 2019; Pan et al., 2016; Tainaka et al., 2014; Yang et al., 2014) to address the imaging problem. Typically, fluorescent labeling of cancer cells *in vitro* or *in vivo* is achieved by endogenous expression of fluorescent proteins such as GFP, YFP, and mCherry, which emit light in the visible spectrum. However, many tissues in the mouse body show high autofluorescence in this range (Tuchin, 2016; Zipfel et al., 2003), which hinders reliable detection of single cancer cells or small cell clusters in mouse bodies based on their endogenous fluorescent signal. To circumvent this problem, we chose to implement the vDISCO technology (Cai et al., 2019), which enhances the signal of fluorescent proteins of cancer cells more than 100-fold in cleared tissues, enabling reliable imaging not only of large metastases but also micrometastases throughout the entire body.

Second, systematic analysis of metastasis in adult mouse bodies requires quantitative information such as location, size, and shape of all individual metastases. Manual detection and segmentation of numerous metastases in highly resolved full body scans is an extremely laborious task that may take several months per mouse for an expert annotator. In addition, automation by filter-based 3D object detectors is not reliable, as different body tissues have different levels of contrast (Pan et al., 2016), causing a high rate of false-positive and false-negative metastasis detections. Recent studies have demonstrated the high efficacy of deep learning-based analysis of biomedical images, compared to filter-based or manual segmentation methods (Camacho et al., 2018; Christiansen et al., 2018; Esteva et al., 2017; Kermany et al., 2018; Sullivan et al., 2018; Topol, 2019; Wang et al., 2019). To enable automated, robust, and fast mapping of all metastases in transparent mice, we developed an efficient deep learning approach based on convolutional neural networks (CNNs) and optimized it for vDISCO imaging data and metastasis distribution patterns.

Resolving these two bottlenecks allowed us to build an integrated, highly automated pipeline for analysis of metastasis and tumor-targeting therapeutics, which we named DeepMACT (deep learning-enabled metastasis analysis in cleared tissue). Using DeepMACT, we detected cancer metastases and even individual disseminated tumor cells in mouse bodies, including many metastases previously overlooked by human annotators. Furthermore, this enabled analyzing the targeting efficiency of a therapeutic antibody against carbonic anhydrase XII on the level of individual metastases. As a scalable, easily accessible, fast, and cost-efficient method, DeepMACT enables a wide range of studies on cancer metastasis and therapeutic strategies. To facilitate adoption of DeepMACT, a step-by-step handbook (Methods S1), the protocols for clearing and imaging, the deep learning algorithm, the training data, and the trained model are available online to address diverse questions in cancer research.

RESULTS

Focusing on a clinically relevant tumor model, we transplanted human MDA-MB-231 mammary carcinoma cells, expressing mCherry and firefly luciferase, into the mammary fat pad of NOD *scid* gamma (NSG) mice and allowed the tumors to grow and metastasize for 6–10 weeks (Figure 1A; Iorns et al., 2012; von Neubeck et al., 2018). Furthermore, we injected the fluorescently tagged 6A10 therapeutic antibody that has been shown to reduce tumor burden in this model (Gondi et al., 2013; von Neubeck et al., 2018). To comprehensively assess cancer cell dissemination and therapeutic antibody targeting in mouse bodies at the level of individual micrometastases, we developed DeepMACT. In short, we transcardially perfused the animals using standard PFA fixation and applied the vDISCO method to enhance the fluorescent signal of tumor cells. After light-sheet microscopy, the 3D image stacks of entire transparent mouse bodies were analyzed using deep learning algorithms. The DeepMACT pipeline consists of (1) vDISCO panoptic imaging of cancer metastases in transparent mice, and (2) deep learning-based analysis of cancer metastasis and antibody drug targeting (Figure 1B).

DeepMACT Step 1: vDISCO Imaging of Cancer Metastases in Optically Cleared Mice

We previously developed the vDISCO technology to image single fluorescent cells in mouse bodies through intact bones and skin (Cai et al., 2019). The vDISCO method utilizes bright fluorescent dyes conjugated with nanobodies to enhance the fluorescent signal of the cells that is weakened during the fixation and clearing process. Here, we first applied vDISCO to increase the fluorescence signal of mCherry-expressing cancer cells. By enhancing the tumor cell fluorescence with anti-mCherry or anti-GFP nanobodies conjugated to Atto-594 or Atto-647N dyes, we found that nanobodies can increase the signal strength of cancer cells over 100 times compared to imaging the endogenous mCherry signal after clearing (Figure S1). Owing to this significant enhancement in signal contrast, we could readily detect micrometastases buried in centimeters-thick mouse bodies (Figures S1F–S1L) e.g., in deep brain and spinal cord regions through the intact skull and vertebrae (Figures S1F and S1I, yellow arrowheads). To confirm the specificity of vDISCO enhancement of the signal from mCherry expressing cancer cells, we performed the following experiments: (1) we stained control mice without a tumor transplant, thereby lacking mCherry expression, and found no labeling in any of the analyzed organs (Figure S2A); and (2) we analyzed the primary tumors and lung metastases from the mouse bodies by staining them using a specific anti-luciferase antibody, which confirmed that endogenous mCherry fluorescence co-localized with both the signals from nanobodies and from the anti-luciferase antibody (Figures S2B and S2C).

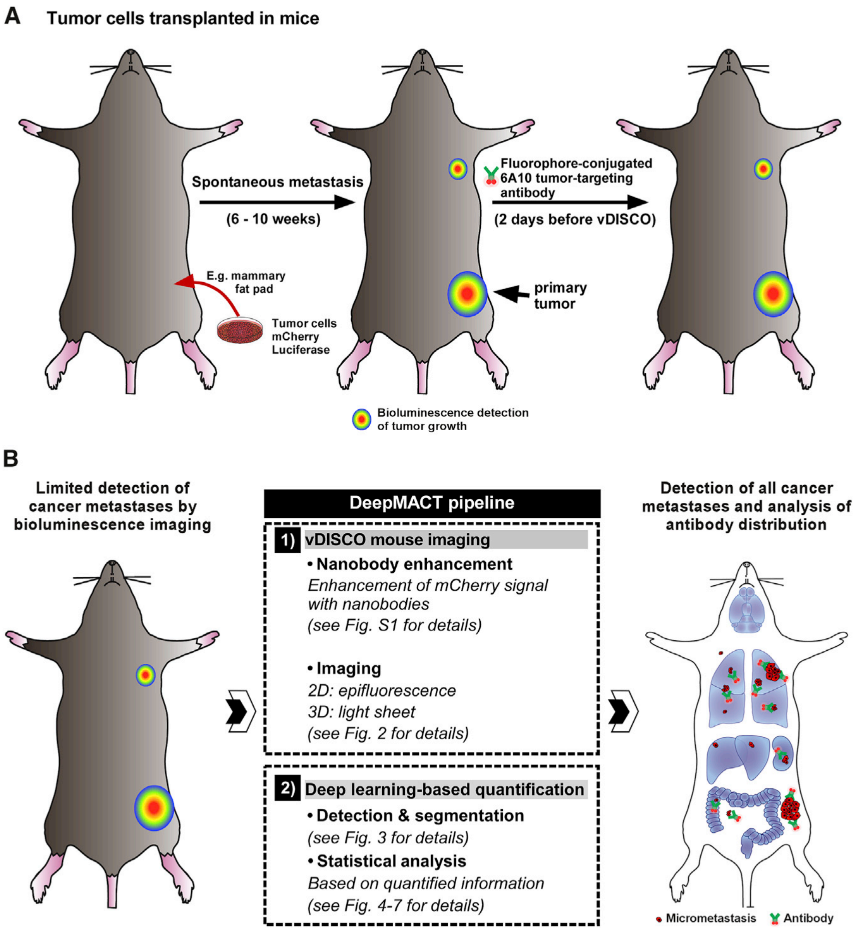

A
Tumor cells transplanted in mice
Spontaneous metastasis
(6 - 10 weeks)
Fluorophore-conjugated 6A10 tumor-targeting antibody
(2 days before vDISCO)
E.g. mammary fat pad
Tumor cells mCherry Luciferase
primary tumor
Bioluminescence detection of tumor growth
B
Limited detection of cancer metastases by bioluminescence imaging
DeepMACT pipeline
1) vDISCO mouse imaging
• Nanobody enhancement
Enhancement of mCherry signal with nanobodies
(see Fig. S1 for details)
• Imaging
2D: epifluorescence
3D: light sheet
(see Fig. 2 for details)
2) Deep learning-based quantification
• Detection & segmentation
(see Fig. 3 for details)
• Statistical analysis
Based on quantified information
(see Fig. 4-7 for details)
Detection of all cancer metastases and analysis of antibody distribution
Micrometastasis
Antibody

Figure 1. Experimental Design and Schematic of the DeepMACT Pipeline for Analysis of Cancer Metastases and Antibody Drug Targeting

(A) Illustration of the experimental workflow for tumor transplantation and antibody application.

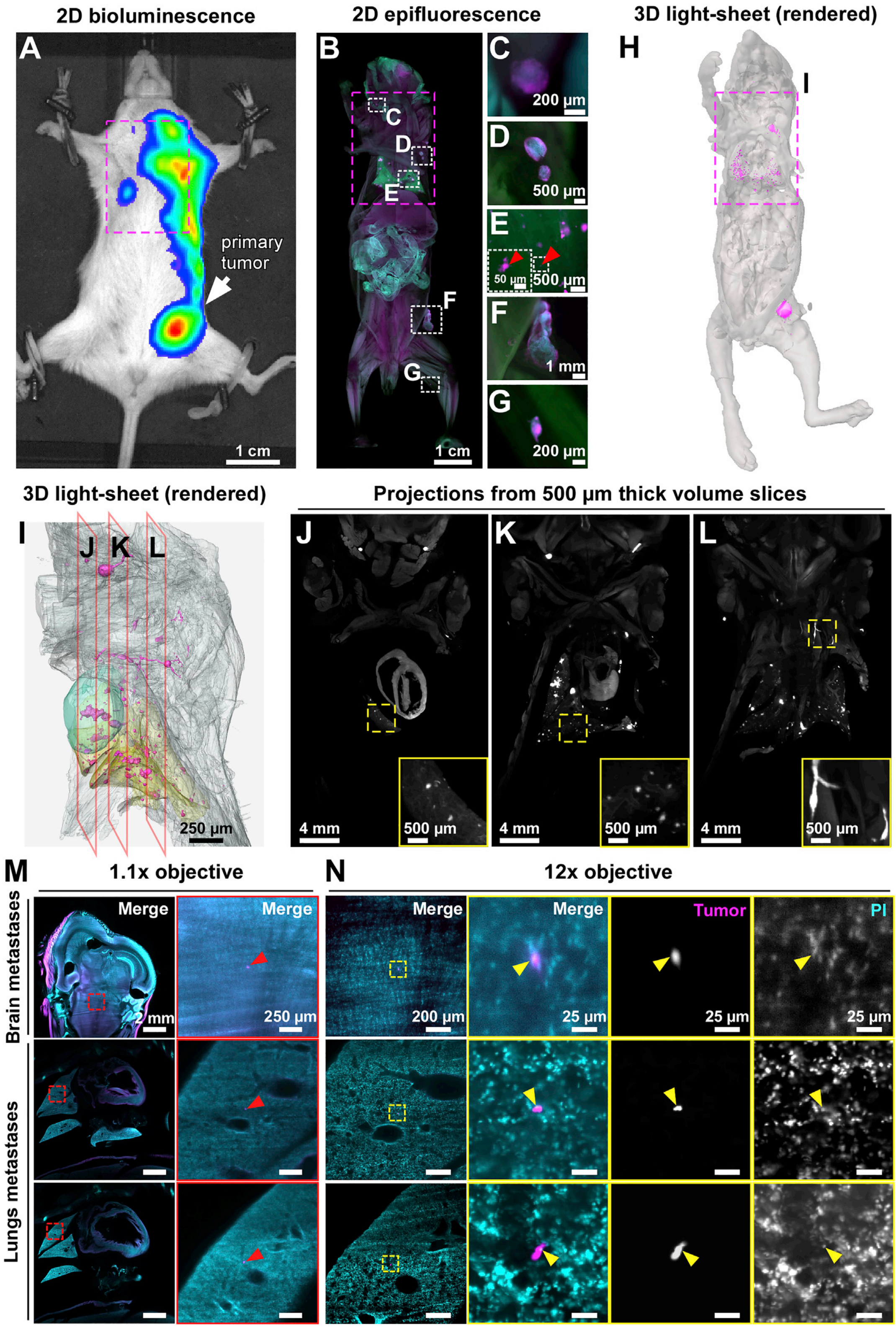
(B) Steps of the DeepMACT pipeline on full-body mouse scans. First, the mice are fixed and processed with the vDISCO protocol to enhance the fluorescent signal of cancer cells. Transparent mice are subsequently imaged from head to toe using light-sheet microscopy, revealing all metastases. Light-sheet images are assembled into a complete 3D image of the mouse. Next, convolutional neural networks are trained to identify and segment all micrometastases in the fluorescence signal. The trained algorithms are then applied to 3D images to detect cancer metastases and an antibody-based drug targeting in full-body mouse scans.

Because the detection of smaller-sized tumor cell clusters, which may represent dormant cancer cells or incipient metastatic nodules, is critical, we next tested if vDISCO allows imaging cancer micrometastases in mouse bodies. In order to compare our approach to conventional methods, we also acquired bioluminescence images of mice before applying DeepMACT. In line with previous findings (Iorns et al., 2012), we detected the earliest large metastasis of transplanted MDA-MB-231 cells at the axillary lymph node of mice by bioluminescence (Figures 2A and S3). However, bioluminescence imaging did not reveal any detailed information such as size or shape and failed to show the presence of micrometastases.

After bioluminescence assessment, we applied vDISCO using anti-mCherry signal enhancing nanobodies conjugated to Atto-647N and imaged the mouse bodies first using epifluorescence in 2D (Figures 2B–2G), then using light-sheet microscopy in 3D (Figures 2H–2L). In epifluorescence, we could readily see both the primary tumor (Figure 2F) and the major metastases at the axillary lymph node (Figure 2D), which were also detected by bioluminescence imaging (Figure 2A), albeit as a bulk signal, lacking information on real size and shape. Importantly, our approach allowed the visualization of several micrometastases in the lungs with conventional epifluorescence imaging, which were not visible in bioluminescence (compare the magenta marked regions in Figures 2A with 2B, and red arrowheads in 2E; more examples shown in Figure S3). Thus, vDISCO followed by epifluorescence imaging, which can be completed within minutes, already provided greater details and sensitivity compared to bioluminescence imaging. Next, we imaged entire fixed transparent mice using a light-sheet microscope (Cai et al., 2019) in 3D to detect individual micrometastases throughout the body (Figure 2H). In the chest area, we could see various metastases not only in the lungs (yellow segmented region in Figure 2I) and lymph nodes, but also at the base of the neck and surrounding tissues (Figures 2J–2L; Video S1). Importantly, light-sheet microscopy scanning allowed us to image even single disseminated tumor cells in the mouse body. Examples of single disseminated tumor cells resolved in full body scans are shown in Figure 2M (see also Video S2), which were further verified by high-magnification light-sheet microscopy imaging showing the colocalization of each single tumor cells with a single nucleus stained by PI (Figure 2N). Thus, our approach allows for the first time to detect micrometastases in full body scans of mice in 3D down to the size of individual cells.

DeepMACT Step 2: Deep Learning for Detection and Quantification of Metastases

We developed an optimized deep learning-based approach to detect and segment all cancer metastases in full-body scans of mice. This framework solves the 3D task of detecting and segmenting metastases in volumetric scans with CNNs that process 2D projections of small sub-volumes (Figure 3A). In brief, we first derived three 2D maximum intensity projections (aligned with the x, y, and z axes) for each sub-volume in order to increase the signal-to-noise ratios (SNRs). We fed the resulting projections to the CNN and obtained 2D probability maps, in which each



(legend on next page)

pixel value represents the estimated probability that this pixel identifies a metastasis under the given projection. We then reconstructed a 3D segmentation from the three projections observing increased reliability in detecting true positive metastases while safely ignoring non-metastatic tissue that would produce false positives in the individual projections. For example, in Figure 3B, the green arrows show successful detection of a real metastasis and the red arrows show successful ignoring of a structure that could be mistaken for a metastasis from a single 2D projection. This approach was highly effective in detecting and segmenting metastases in the imaging data, yielding a binary mask for all metastases in the body.

The core of our architecture makes use of CNNs (Figure 3C), structurally similar to the established U-net (Ronneberger et al., 2015), which learn to distinguish metastases from the background signal. This is achieved by using a deep stack of encoding units, which detect characteristic cancer features, and a corresponding stack of decoding units, which segment each metastasis at pixel-level. Each encoding unit performs two convolutions, extracting information about the environment for each pixel and representing that information in a third dimension—the feature channels. Before being passed on to the next encoding unit, the image is spatially down-sampled. Together, this means that the neural network is steadily increasing the feature channels and steadily decreasing the spatial resolution, enforcing the network to learn even more abstract representations of the data (i.e., features) in the deeper layers, before mapping the information relevant to cancer cells back to the original resolution in the decoding upward path. This happens by up-sampling the abstract, low-resolution information from lower layers and concatenating it with the less abstract, but higher-resolution information from the encoding path via skip connections (some exemplary visualizations of the computational stages are presented in Figures S4A–S4C).

To assess the reliability of our automated deep learning architecture, we applied it to a fresh test set of a full-body scan, which was neither used for training the CNNs nor to optimize hyperparameters. The datasets were manually annotated by human experts and any disagreements between experts were jointly reviewed and discussed in order to derive a refined, commonly agreed reference annotation (see STAR Methods for details). We then systematically compared the performance of our deep learning approach to that of established detection methods as well as the performance of a single human annotator, calculating F1-score (also known as Dice score), a common performance measure based on both the metastasis detection rate (recall) and false positive rate (precision).

As shown in Figure 3D, we found that DeepMACT reached an F1-score of 80%, outperforming existing filter-based detectors such as the ImageJ Object Detector (18%) or a custom-made filter-based detector (36%) by a large margin. The similar performance of 3D CNNs such as a customized 3D U-net (38%) highlights the benefit of the specialized DeepMACT approach for the tumor models we tested. Indeed, the detection performance of DeepMACT comes very close to the level of a single human expert annotator with an F1-score of 83%. The slightly higher F1-score of the human annotator is mainly driven by the high precision. However, the human annotator missed around 29% of all micrometastases (examples are shown in Figures S4D–S4F) and detecting those false negatives would require a repetitive and very laborious re-analysis of the entire animal scans, requiring up to several months of human work time. On the other hand, the F1-score of DeepMACT is a result of a balance between precision and recall, which can be freely adjusted via the model's threshold. For DeepMACT, we can increase detection rate (recall) over 95%. While this also increases the false-positive rate, correcting the false positive data requires only a review of detected signals by a human annotator, which we completed within 1 h per mouse in this study (a typical example for a false positive detection is shown in Figures S4G–S4I). Combining the DeepMACT prediction with this quick review yielded an F1-score of 89%, exceeding the performance of a single human annotator. A more detailed analysis on the trade-off between precision and recall is shown in Figure S4J. Notably, DeepMACT could detect micrometastases ~30 times faster than filter-based detectors and over 300 times faster than a human annotator (Figure 3E). Even taking the time for a manual review of the DeepMACT prediction into account, the total processing speed was still 8 times faster than filter-based detectors and over 60 times faster than a human annotator, who was already supported by a dedicated and interactive software, custom-built for this task and these data; without annotation software, the human manual

Figure 2. DeepMACT Step 1: vDISCO Visualization of Metastases in a Full-Body Scan of a Mouse

(A) Bioluminescence image of a NSG female mouse before vDISCO which was taken 2 months after MDA-MB-231 cancer cell implantation into the mammary fat pad.

(B–G) Epifluorescence images of the same mouse after vDISCO show metastases (magenta) in greater detail compared to bioluminescence. (B) shows the entire mouse, (C-G) shows magnifications of the areas marked with white dashed lines in (B), including small micrometastases that can be readily detected in the lungs (E, red arrowhead) and in the leg (G), in addition to the primary tumor (F) and major metastases (C and D) that are also visible in bioluminescence as bulk signal (A).

(H) 3D visualization of the transparent mouse body imaged by light-sheet microscopy.

(I) Lateral views of the 3D segmentation obtained from the light-sheet imaging data corresponding to the magenta-boxed region indicated in (A), (B), and (H). For simplicity, only a few organs are segmented: the heart (cyan) and the lungs (yellow); the mouse body is shown in transparent gray and the metastases are in magenta.

(J–L) Original light-sheet microscopy data (500 μm projections) showing metastases from the three different sagittal planes indicated in (I) with the corresponding letters.

(M and N) Single cell metastases identified in the brain and in the lungs by full-body light-sheet microscopy scans using a 1.1× objective with 6 μm lateral resolution (tumor cells in magenta and nucleus labeled with propidium iodide [PI] in cyan) (red arrowheads in M). The same metastases were re-imaged by light-sheet microscopy with a 12× objective. Single plane images showed the colocalization of each micrometastasis with a single nucleus (yellow arrowheads in N). Panels in (M) show images acquired with a 1.1x objective, panels in (N) show images acquired with a 12x objective.

See also Figures S1, S2, and S3 and Videos S1 and S2.

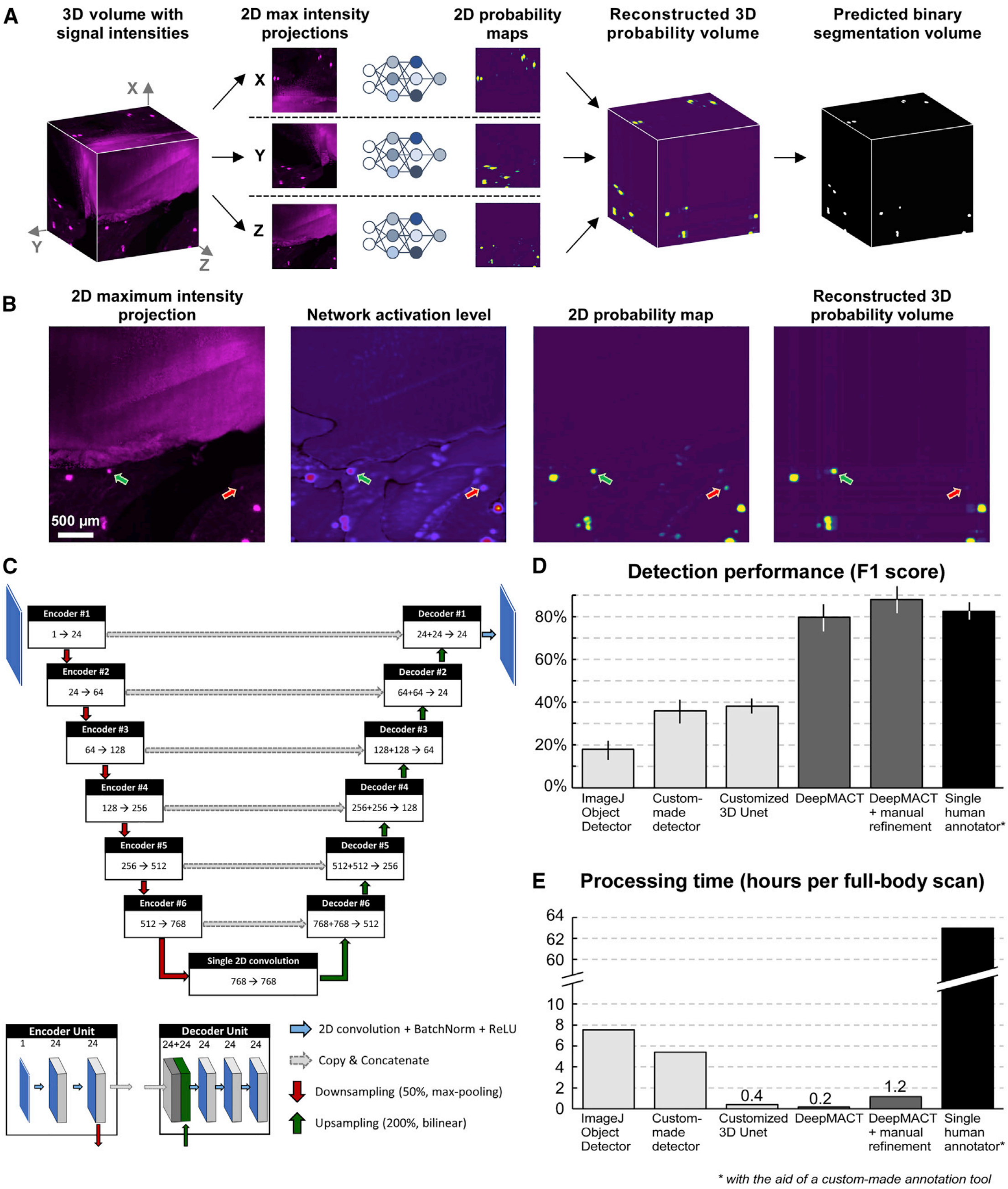


Figure 3. DeepMACT Step 2: Schematic and Performance of the Deep Learning Algorithm

(A) Representation of the deep learning inference workflow to efficiently derive 3D detection and segmentation exploiting three 2D computational operations.

(B) Visualization of the computational stages; the green arrow shows successful detection of a metastasis, the red arrow shows elimination of a false positive detection in the 3D reconstruction stage.

(C) High-level representation of the network architecture with an encoding and a decoding path.

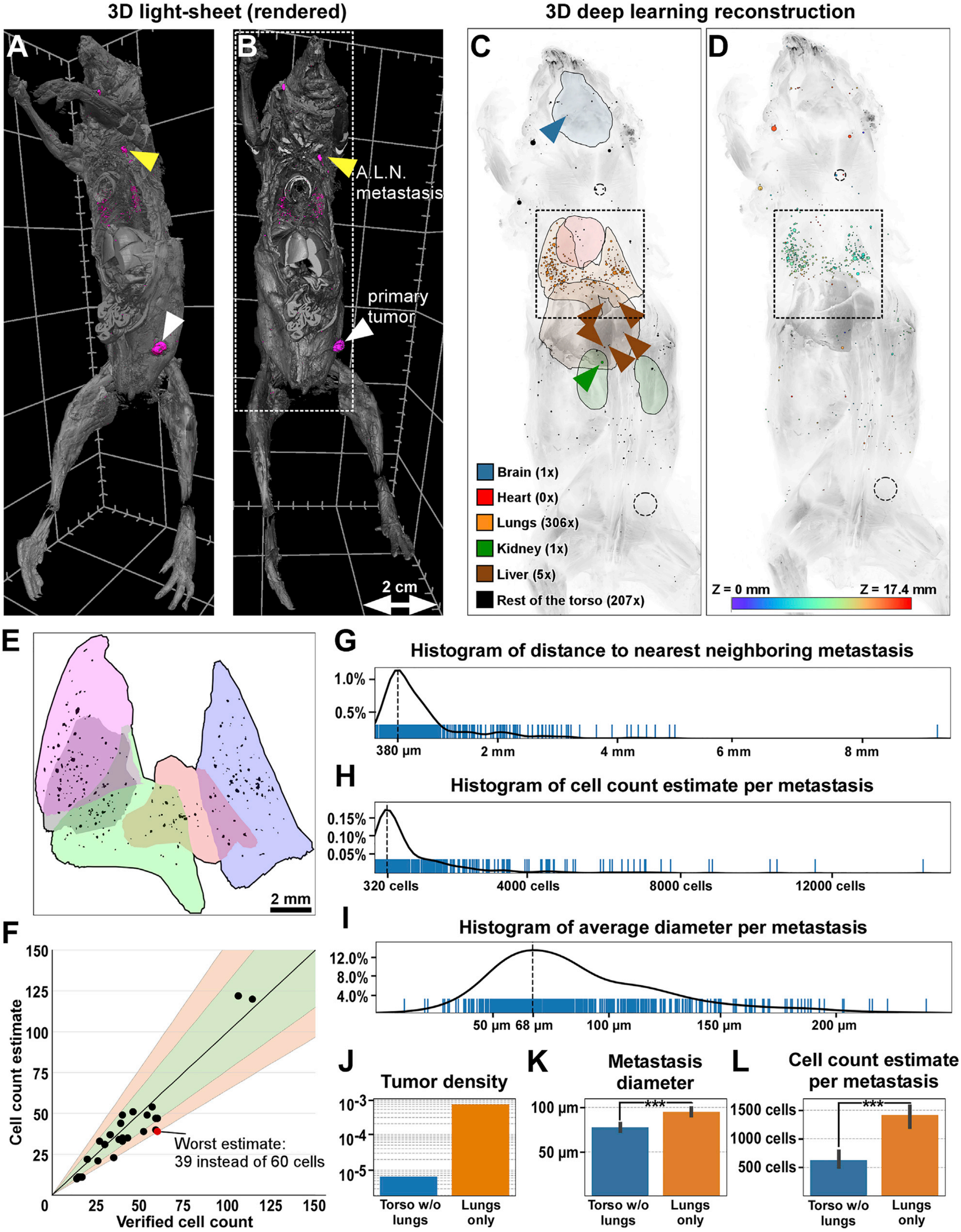
(D and E) Comparison of our deep learning pipeline, DeepMACT, to alternative automated methods and manual segmentation by a human expert in terms of detection performance (D; error bars show SEM) and processing time (E).

See also Figure S4.

annotation would be estimated to take several months for a single mouse. Thus, DeepMACT can complete months to years of human labor within hours without compromising on segmentation quality.

DeepMACT Reliably Detects Micrometastases in Different Tumor Models

After establishing the DeepMACT pipeline, we used it to analyze full mouse bodies. Apart from the primary tumor and the



(legend on next page)

macrometastasis in the axillary lymph node, we could detect hundreds of micrometastases of varying sizes throughout the body, especially in the lungs (Figures 4A and 4B). Overall, DeepMACT identified 520 micrometastases throughout the entire body in this particular mouse, of which there were 306 in the lungs and 214 throughout other organs of the body (Figure 4C). We found that micrometastases are mostly located in the inner tissue layers (~1 cm depth from the surface), as shown by color-coding in Figure 4D, making them particularly difficult to detect by other methods. To analyze the spatial distribution with regard to the lung anatomy, we registered all 306 lung micrometastases to the mouse lung lobes. We found that micrometastases were evenly distributed in all lobes (Figure 4E). Interestingly, the micrometastases were randomly distributed throughout the lungs regardless of their size, suggesting independent colonization at multiple sites. Furthermore, we quantified the size and relative location of all micrometastases in the entire body (Figures 4F–4L). While 79% of micrometastases were within 1 mm to the nearest neighboring micrometastasis, we also found highly isolated micrometastases as distant as 9.3 mm apart from their nearest neighbor (Figure 4G). Importantly, we found a large number of micrometastases with estimated cell counts of a few hundred cells or less (Figures 4F and 4H) and diameters less than 50–100 μm (Figure 4I), which would be very difficult to detect in mice by other methods. Comparing the micrometastases in the lungs with those in the torso, we found that the tumor burden in the lungs was more than a hundred times higher in this tumor model (Figure 4J). Also, micrometastases in the lungs were, on average, 30% larger in diameter (Figure 4K), with a more than 2-fold higher estimated cell number per metastasis, compared to micrometastases in the rest of the torso (Figure 4L).

To verify the robustness and applicability of the DeepMACT pipeline for a wider range of experimental settings, we conducted additional studies. First, we implanted a solid tumor (MDA-MB-231 breast cancer grown in another mouse for 10 weeks) into a healthy mouse and analyzed it right away, leaving no time for metastases to form. As expected, no metastases could be found in this control, indicating that tumor cells do not detach from a solid tumor during the tissue clearing procedure and that no artifacts (such as potential unspecific nanobody accumulations during the staining procedure) would be mistaken for metastases (Figure 5A).

Second, we applied the pipeline to 3 different tumor models with distinct metastatic propensity and organotropism. A nude mouse intracardially injected with human MCF-7 estrogen receptor (ER)-positive breast cancer cells developed metastases throughout the body, with a substantial burden in the lungs (49 metastases), liver (18), and kidneys (11), but also in the bones (2; indicated by yellow arrows) and the brain (1) (Figure 5B). A C57BL/6 mouse transplanted with murine syngeneic R254 pancreatic cancer cells, however, did not develop any metastases in the brain, kidneys or bones, but rather in the lungs (8), the liver (6), and also in distinct tissues such as the peritoneum (Figure 5C; metastasis in peritoneum indicated with a magenta arrow). A further model using the human brain metastatic lung cancer cell line H2030-BrM3 transplanted in nude mice only showed few metastases in the liver (2) or kidneys (1) but many in the brain (31) (Figure 5D). These experiments demonstrate that the DeepMACT pipeline can reliably detect micrometastases in a variety of tumor models with distinct organotropisms, including different immunodeficient or immunocompetent mouse strains, syngeneic tumors and xenotransplants. Furthermore, metastases can be quantified and assessed by DeepMACT in organs in which this is difficult to achieve by other methods, such as bones and the brain.

In a third experiment, we tested the potential of DeepMACT to study the progression of the metastatic process over time. We injected MDA-MB-231 cancer cells intracardially and analyzed the distribution of metastases 2 days, 6 days, and 14 days after injection (Figures 5E–5G). We found metastases in the brain, lungs, liver, kidneys, bones, and other organs at all time points. Moreover, our results showed a substantial increase in the total metastatic burden in the mouse bodies as well as in the lungs as the primary metastatic organ (Figure 5H).

Importantly, neither the increase in overall tumor burden for the time course study nor the differential distribution of metastases across organs for any of the cancer models tested were clearly revealed by bioluminescence images (Figure S5). Thus, our pipeline is the first to enable quantitative analyses of micrometastases in full-body scans, greatly enhancing our ability to assess the metastatic process in a comprehensive manner.

Figure 4. Deep Learning-Based Detection and Segmentation Enables Quantitative Analysis at the Level of Individual Metastases

(A and B) 3D rendering of a mouse transplanted with MDA-MB-231 cells in the mammary fat pad after light-sheet microscopy imaging in lateral and ventral views, respectively. Metastases in the mouse body are shown in magenta. The white arrowhead indicates the primary tumor and the yellow arrowhead indicates metastases in the axillary lymph node (A.L.N.). (A) and (B) show the same mouse at different perspectives.

(C and D) Deep learning reconstructions of all detected metastases (A.L.N. and primary tumor indicated with dashed circles) color-coded by organ (C) and depth along the z axis (D), cropped to the white box in (B) to show higher level of detail.

(E) Detailed view of metastases in the lung region (corresponding to the black box in C) in a projection of 3D deep learning-based detection, with metastases registered to individual lung lobes (shown in different colors).

(F) Validation of cell count estimates by comparing to manual count. 73% of the estimates are within a 20% margin (green region), and all estimates are within a 35% margin (red region) of the manual count (n = 26 randomly selected sample regions).

(G–I) Deep learning-based distributions; blue bars show individual metastases, the black line shows the Gaussian kernel density estimation.

(G) 3D distance to nearest neighboring metastasis.

(H) Estimates of cell counts per metastasis.

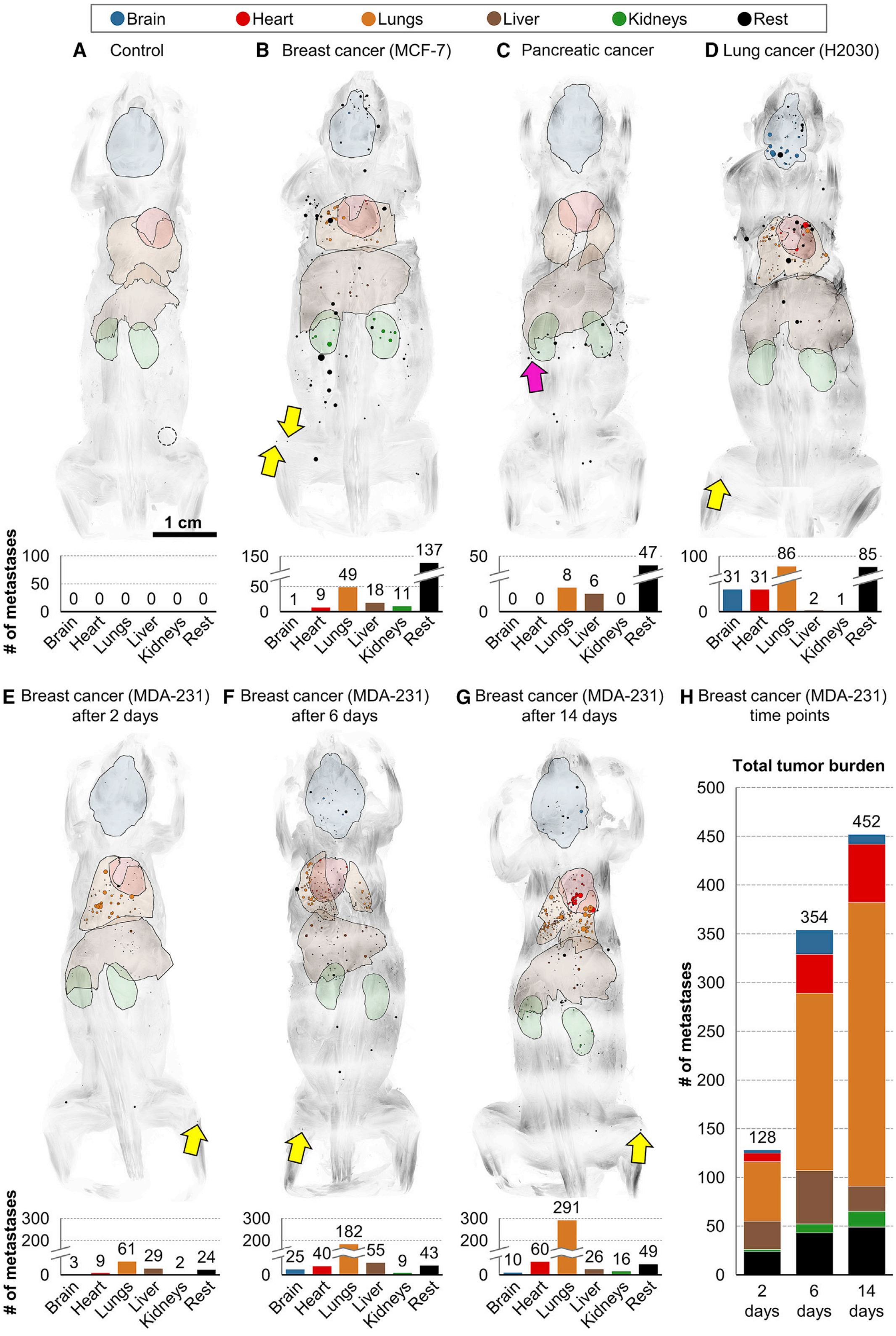
(I) Metastasis diameter averaged in 3D space.

(J–L) Quantitative comparison between metastases in the lungs and the rest of the torso; bars indicate 95% confidence intervals.

(J) Tumor density as share of metastatic tissue of the entire volume is two orders of magnitude higher in lungs versus the rest of the torso.

(K) Metastasis diameter (averaged in 3D space) is significantly higher in lungs ($p < 0.001$; two-sided t test). Error bars show standard deviations.

(L) Cell count estimate per metastasis is significantly higher in lungs ($p < 0.001$; two-sided t test). Error bars show standard deviations.



(legend on next page)

DeepMACT Reveals Therapeutic Antibody Targeting at the Level of Single Metastases

A number of tumor-targeting monoclonal antibodies have become part of the standard treatment for various solid and hematological malignancies and many more are in early or late stages of clinical development (Barker and Clevers, 2006; Pandey and Mahadevan, 2014). However, so far there has been no methodology to determine the distribution of therapeutic antibodies across the entire body, down to the level of single micrometastases. Here, we used DeepMACT to assess the biodistribution of the therapeutic monoclonal antibody 6A10 directed against human carbonic anhydrase XII (CA12) (Battke et al., 2011; Gondi et al., 2013; von Neubeck et al., 2018). CA12 is overexpressed in various types of cancers, and blocking its activity with the antibody 6A10 reduces tumor growth (Gondi et al., 2013) and increases the sensitivity of tumors to chemotherapy (von Neubeck et al., 2018). We intravenously injected 20 μg of 6A10 conjugated to Alexa-568 9 weeks after transplantation of MDA-MB-231 cells and perfused the mouse 2 days after the antibody injection for full-body-scale analysis, enhancing the tumor signal with Atto-647N. Because Alexa-568 excitation/emission spectra overlap with the endogenous mCherry signal of the transplanted cancer cells, we confirmed that the vDISCO pipeline completely eliminates the signal from endogenously expressed mCherry (Cai et al., 2019; Figure S6).

We first acquired 2D images with epifluorescence microscopy and observed an accumulation of the 6A10 antibody at the primary tumor (Figures 6A and 6E; tumor shown in magenta, therapeutic antibody in cyan) and the metastases at the axillary lymph node (Figures 6A and 6B). Focusing on the lungs, we detected micrometastases that were targeted by the 6A10 antibody (Figure 6C, white arrow) and others that were not (Figure 6D, yellow arrow). Acquiring 3D scans with light-sheet microscopy, we assessed the complete biodistribution of the therapeutic antibody and micrometastases throughout the mouse body (Figures 6F–6H; Video S3). The axillary lymph node metastases and the micrometastases in the lungs are shown in Figure 6F. Analyzing the signal of individual micrometastases and the 6A10 antibody by light-sheet microscopy in 3D, we could evaluate the efficiency of antibody drug targeting for all the micrometastases (Figure 6G, white arrows). We also verified the targeting of micrometastases by the 6A10 antibody in different organs such as lungs and kidney, using confocal microscopy (Figure S7).

Next, we used DeepMACT to systematically assess and quantify the efficiency of antibody drug targeting in full body scans at the level of single micrometastases (Figure 6I). While overall 77% of metastases were targeted by the antibody, we found that significantly more micrometastases were targeted in the lungs (85%) as compared to the rest of the body (66%) (Figure 6J; Videos S3 and S4). To further assess the efficiency of drug targeting for micrometastases in the lung versus the rest of the body, we evaluated the antibody concentration by quantifying the antibody signal contrast (relative signal strength versus local surrounding; see STAR Methods for details; Figure 6K). Metastases in the lungs generally tended to have a higher antibody signal ratio, in line with the higher share of targeted metastases. In addition, the antibody signal ratio was much more narrowly distributed compared with micrometastases outside the lungs. The lower average and wider distribution of antibody signal ratio in the micrometastases in the rest of the body indicate that there is a substantially higher variance in the antibody targeting to the cells of those micrometastases. While some are very strongly targeted, many others are not targeted at all. The largest quartile of micrometastases was significantly more likely targeted (88%) than the smallest quartile (67%) (Figure 6L). We also identified various off-target binding sites throughout the body (i.e., binding of the therapeutic antibody to mouse tissues), which is presumably due to unspecific interactions because 6A10 does not bind to murine CA12 (cyan inset in Figure 6H). Overall, these data demonstrate that DeepMACT provides a powerful platform to track the biodistribution of therapeutic antibodies along with micrometastases in mouse bodies. Thus, it represents the first method that allows quantitative analysis of the efficiency of antibody-based drug targeting at the full body scale, with a resolution down to the level of individual micrometastases.

Exploring Potential Mechanisms of Antibody Drug Targeting

The above results demonstrated that antibody-based drugs, which are the basis of many targeted/personalized treatments, may miss as many as 23% of the micrometastases. Next, we aimed to explore potential mechanisms that might explain this failure. We first hypothesized that the efficiency of targeting of micrometastases might depend on the availability of nearby blood supply transporting the therapeutic antibody. To explore if the vascularization of defined tissue regions can have an effect on antibody drug targeting, we performed lectin labeling of vessels in the lungs, where most of the micrometastases are located. Analyzing diverse micrometastases of different sizes, we found that each of them had blood vessels within a distance of 1–6 μm (Figures 7A and 7B). This distance is smaller than even a single cell diameter (~10 μm) suggesting that absence of nearby blood vessels could not be the major reason for the lack of antibody drug targeting (Tabrizi et al., 2010).

Figure 5. DeepMACT Reliably Detects Metastases in All Organs for a Variety of Tumor Models

Metastasis detections in full-body 3D light-sheet microscopy scans; each dot represents a metastasis, color-coded by organ; black metastasis within organ outlines are not inside that organ but rather above or below it.

(A) A control mouse was perfused immediately after implantation of a solid tumor (MDA-MB-231; dashed circle), leaving no time for metastases to form.

(B) MCF-7 breast cancer cells were intracardially injected in a nude mouse.

(C) Pancreatic cancer cells (R254) were transplanted into the pancreas (dashed circle) of a C57BL/6 mouse.

(D) H2030-BrM3 lung cancer cells were intracardially injected in a nude mouse.

(E–G) Three NSG mice were intracardially injected of MDA-MB-231 breast cancer cells and sacrificed after 2 days (E), 6 days (F), and 14 days (G).

(H) DeepMACT analysis shows increase in tumor burden over the three time points.

Yellow arrows indicate metastases in bones; the magenta arrow indicates a metastasis in the peritoneum.

See also Figure S5.

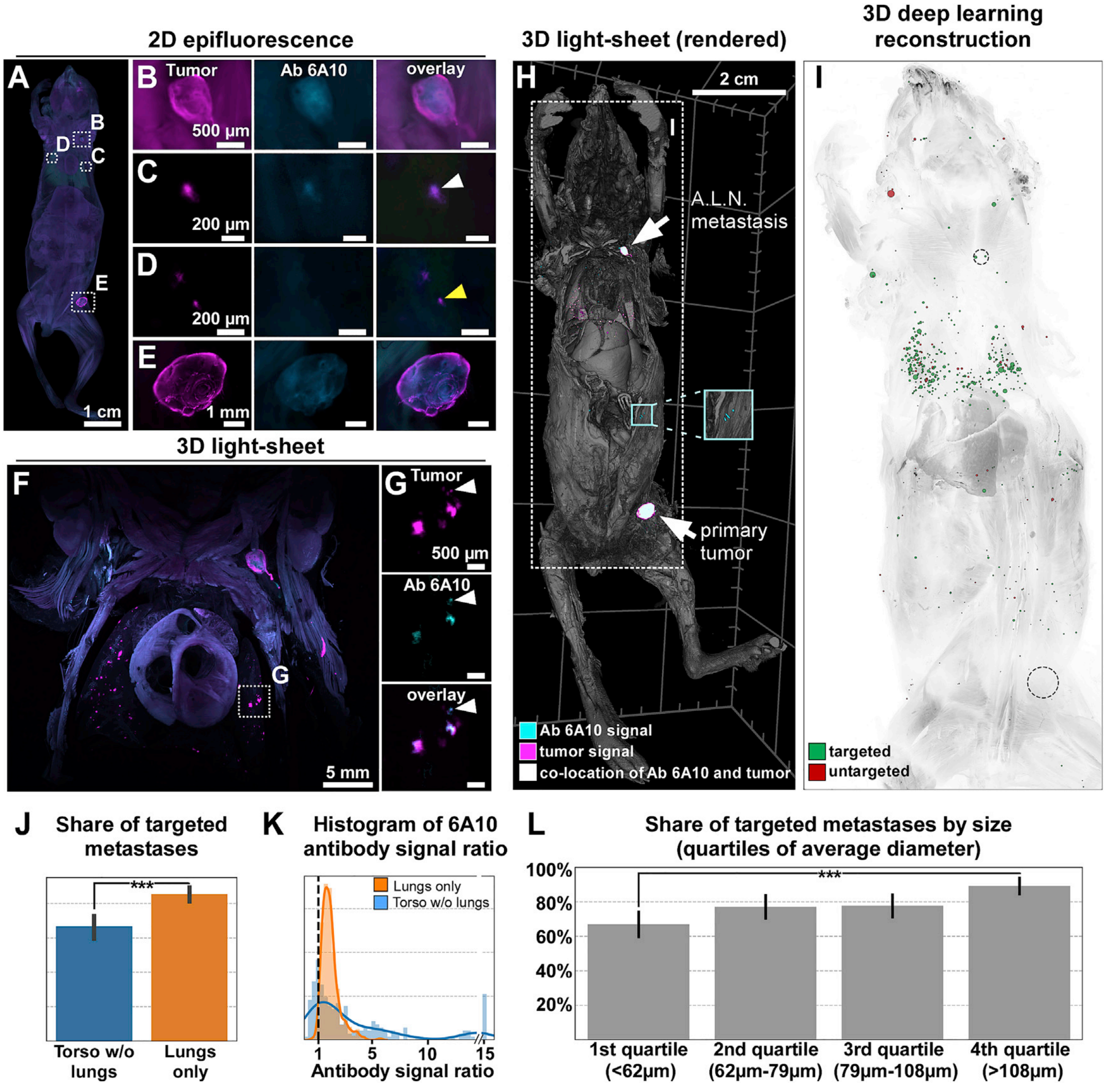


Figure 6. The DeepMACT Pipeline Enables Quantitative Analysis of Drug Delivery Efficacy at the Level of Single Metastases

A mouse transplanted in the mammary fat pad with MDA-MB-231 cells was intravenously injected with 6A10 anti-CA12 antibody 9 weeks later.

(A) Epifluorescence image of a processed mouse.

(B-E) Magnifications of the different areas marked with white dashed lines in (A), showing details of both tumor metastases (enhanced with Alexa647N nanobody, shown in magenta) and 6A10 antibody (conjugated with Alexa568, shown in cyan) distributions and their overlay. While most of the micrometastases are targeted by the antibody (C, white arrowhead), there are some that are not (D, yellow arrowhead).

(F) Full-body 3D light-sheet scan, cropped to the chest region, shows the distributions of metastases (magenta) and antibody (cyan).

(G) Detailed view of the boxed region in (F) showing very small micrometastases targeted by the therapeutic antibody (white arrowheads).

(H) 3D rendering of a mouse body light-sheet scan showing the tumor signal in magenta and the 6A10 antibody signal in cyan (co-localization of the signals is shown in white). The cyan inset shows an example of off-target accumulation of the 6A10 antibody.

(I) Deep learning-based reconstruction of the animal in (H) showing targeted metastases in green and untargeted metastases in red; the dashed circles represent the primary tumor and A.L.N metastases.

(J) A significantly higher share of metastases are targeted in the lungs versus the rest of torso ($p < 0.001$, two-sided t test). Error bars show standard deviations.

(K) Comparison of the distributions of 6A10 antibody signal ratio (signal strength in metastasis versus local surrounding; see the STAR Methods for further details) per metastasis in the lungs versus the rest of torso. The dashed line indicates a ratio of 1 (equal signal strengths).

(L) Share of targeted metastases as a function of their size (split into quartiles of average metastasis diameter; $p < 0.001$, two-sided t test). Error bars show standard deviations.

See also Figures S6 and S7 and Videos S3 and S4.

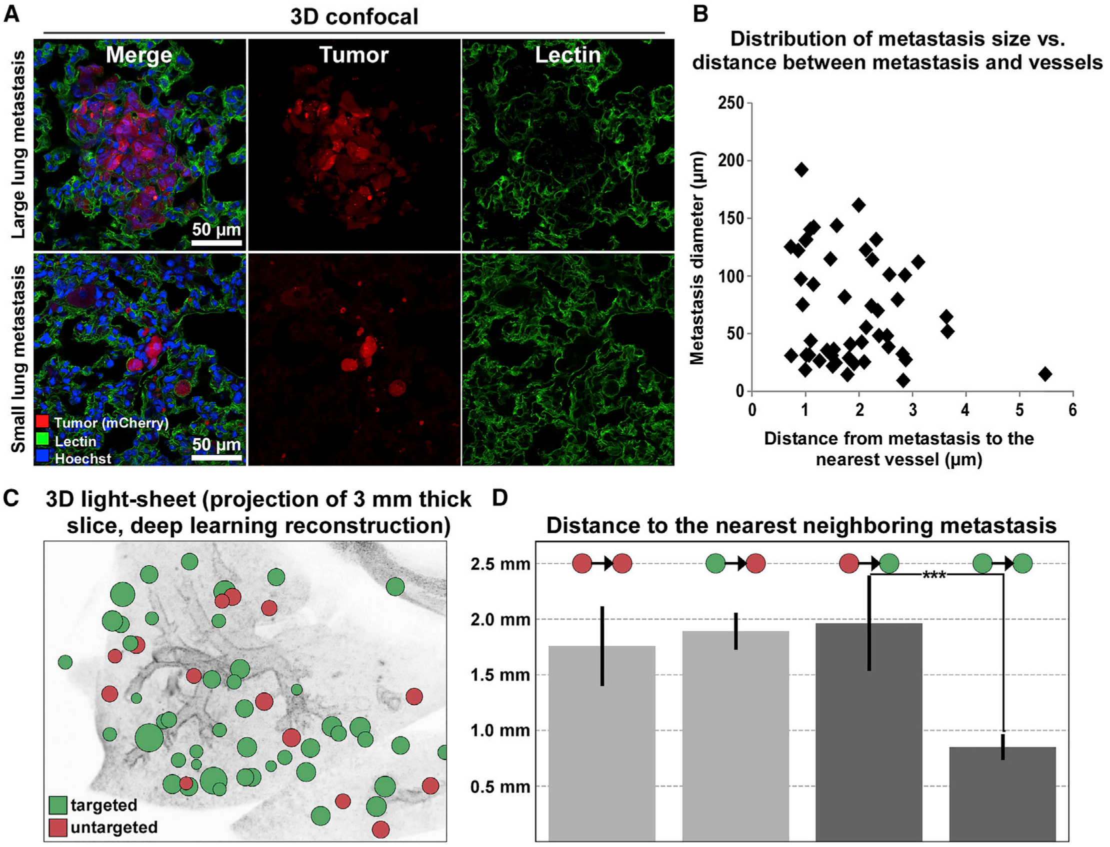


Figure 7. Potential Mechanisms of Metastasis Targeting by Therapeutic Antibody

(A) Confocal images of a large and a small metastasis (<5 cancer cells) in the lungs of a mouse transplanted with MDA-MB-231 cells and intravenously injected with 6A10 anti-CA12 antibody, labeled with lectin (green) and Hoechst (blue).

(B) Distribution of metastasis size and distance to the nearest vessel, showing that most of the metastases are close to vessels (distance <6 μm; n = 50).

(C) Deep learning-based reconstruction of lung metastases with and without 6A10 antibody targeting.

(D) Deep learning-based quantification of distance between metastases and their nearest neighbor. The average distance from an untargeted to the nearest targeted metastasis is significantly ($p < 0.001$; two-sided t test) larger than from a targeted one; this shows local clustering of targeted and untargeted metastases (see the STAR Methods for further details). Error bars show 95% confidence intervalls for the estimation of the mean.

Next, we hypothesized that the tumor microenvironment at the sites of metastases could be related to the efficiency of targeting. If so, we would expect a non-random spatial distribution of targeted and untargeted metastasis on a local scale. To address this, we turned to DeepMACT and assessed the local clustering of micrometastases targeted by the antibody. We quantified the distances between micrometastases and their nearest neighbor for all micrometastases in the entire body, differentiating between targeted and untargeted nearest neighbors. The distance between two neighboring metastases is smaller for two targeted metastases (~0.8 mm) than for two untargeted or a mixed pair of an untargeted and a targeted metastasis (consistently at ~1.7–2.0 mm) (Figures 7C and 7D). Importantly, the average distance from an untargeted to the nearest targeted metastasis is significantly larger than from a targeted one. This would not be expected in a random distribution and indicates a clustering on a local scale. Thus, these analyses suggest the existence of factors in tumor microenvironments influencing the efficiency of antibody drug targeting.

DISCUSSION

Unbiased, comprehensive detection of cancer metastases and the biodistribution of tumor-targeting therapeutics at the level of single micrometastases would substantially accelerate preclinical cancer research. Toward this goal, we capitalized on a powerful tissue clearing and imaging method combined with deep learning-based analysis, enabling us to visualize and analyze cancer metastasis in transparent mouse bodies. The resulting DeepMACT workflow is a straightforward method for systemic analysis of micrometastases and therapeutic antibody drug distribution at the full body scale and with a resolution down to individual micrometastases within days, a task that would otherwise take several months to years of human labor. Thus, DeepMACT-based evaluation of entire transparent mouse bodies instead of selected tissues/organs can foster the development and translation of new therapies from preclinical research much more efficiently than traditional methods.

To further facilitate easy adoption of our technology by diverse labs, we provide (1) a handbook (Methods S1) with detailed step-by-step instructions for carrying out the DeepMACT pipeline; (2) various resource videos and troubleshooting tips; (3) a package including the trained DeepMACT algorithm and annotated data; and (4) an online version of the DeepMACT algorithm that can be executed via any web browser (hosted by the Code Ocean initiative) without downloading any code or installing any software (links to these resources are provided in STAR Methods).

DeepMACT Technology

Here, we set out to make use of recent technologies that can provide scalable and unbiased histological assessment of entire biological specimens. Most full body scale clearing and imaging studies have so far relied on visualization of endogenous fluorescent signal, which is not sufficiently strong to allow imaging and quantification of metastases in transparent mice (Kubota et al., 2017; Pan et al., 2016). To overcome this, we adopted the vDISCO mouse clearing and staining technology, as it can enhance the fluorescent signal in fixed and cleared tissues by more than 100 times (Cai et al., 2019), ensuring reliable detection of micrometastases. Because vDISCO employs nanobody enhancement of the endogenous fluorescent signal, currently up to 21 types of fluorescent proteins can be labeled with available nanobodies. In addition, conjugation of existing nanobodies with fluorescent dyes at diverse spectra, including those in the near infrared range would help to generate more options for multiplex experiments including imaging of more than one type of fluorescently labeled cell along with conjugated therapeutic antibodies.

Second, we developed a highly efficient deep learning architecture based on U-net like CNNs exploiting 2D maximum-intensity projections with high SNR to reliably detect metastases in 3D. Deep learning-based detection not only serves the purpose of automation but also provides a very effective tool in finding metastases that would be easily overlooked by humans. In our data, an expert human annotator missed around 29% of all metastases. This is in line with previous studies where human experts missed 1 in 4 breast cancer metastases in histopathology (Vestjens et al., 2012), a problem that is further exacerbated if humans work under time pressure (Ehteshami Bejnordi et al., 2017). Motivated by this, deep-learning-based approaches for cancer and metastasis detection have recently started gaining substantial momentum for various imaging modalities, also beyond microscopy (Litjens et al., 2016; Liu et al., 2019; Steiner et al., 2018; Wang et al., 2017).

Here, we used an MDA-MB-231 cancer cell-based tumor model to train the algorithms. While training deep networks in general may require large training datasets to diversify their applications, the U-net-like architecture at the core of DeepMACT can be easily adopted to other cancer models (Bhatia et al., 2019; Falk et al., 2019; Wang et al., 2018). Indeed, after learning to detect the characteristic shape and appearance of micrometastases against the background signal, DeepMACT successfully analyzed 3 additional tumor models we used here without further training: MCF-7 estrogen receptor positive breast cancer model, H2030-BrM3 lung cancer model, and R254 syngeneic pancreatic cancer model. Therefore, it would require little effort to apply our algorithms to different types of tumor models. Also, adapting the algorithm to applications in which, for instance, shape and size differ substantially from MDA-MB-231 metastases, would not require training from scratch. Adjusting design parameters such as the size of subvolumes (see the STAR Methods and the detailed handbook [Methods S1] for DeepMACT that we provide) allows the straightforward adaptation of the algorithm to new data with different SNR, metastasis sizes, or spatial resolution of the scan. Furthermore, building upon our pre-trained algorithms, which are freely available online, allows retraining the algorithm with substantially less training data.

To ensure high computational efficiency, our approach solves the three-dimensional task of detecting and segmenting the metastases by exploiting two-dimensional representations of the data. This is important because 2D maximum-intensity-projections increase SNR when there is little background noise owing to the high specificity of the labels in vDISCO clearing. 3D convolutions are exponentially more expensive in terms of model complexity (number of parameters) as well as computational load than 2D convolutions, thus requiring more powerful computing resources and more data annotated in 3D to train the algorithm. Importantly, the increased number of parameters is detrimental to model performance, unless the amount of training data is further increased. In this study, the 3D CNNs we tested failed to reach a high level of detection performance due to limited availability of training data, a common constraint in practice given the cost associated with annotating data (especially in 3D). In addition, the more efficient nature of our approach allows training the entire algorithm on a standard workstation with an ordinary GPU within a few hours; applying the trained algorithm to a new dataset takes in the order of 15 min, highlighting the scalability and cost-efficiency of our pipeline. Thus, the DeepMACT architecture is designed to enable widespread adoption of our approach by minimizing data annotation and computing requirements while allowing for easy adaptation to other experimental setups (such as different tumor models).

DeepMACT Detection of Micrometastases and Tumor-Targeting Drugs

Methods such as magnetic resonance imaging (MRI), computed tomography (CT), and bioluminescence imaging have been widely used to visualize cancer growth at the primary site and distant body regions (Condeelis and Weissleder, 2010; Massoud and Gambhir, 2003, 2007; Ntziachristos, 2010; Pichler et al., 2008; Timpson et al., 2011). While these methods provide crucial longitudinal information on the size of the primary tumor and large metastases, they typically can only resolve structures larger than 75 μm, hence they do not have the resolution to detect smaller micrometastases consisting of fewer cells.

Unbiased high-throughput mapping of tumor micrometastases in full body scans of rodents can be a valuable tool to uncover the biology behind the dissemination of tumor cells. We show here that DeepMACT is a powerful pipeline for detecting and mapping cancer metastases in mouse bodies, allowing identification of the precise locations of even the smallest disseminated tumors. Complex analysis, e.g., of the size, location, and density of micrometastases could be performed in a short time throughout the body, without dissecting any pre-defined region. In addition to detecting the micrometastases in the selected organs such as the lungs and liver, we also identified numerous micrometastases throughout the torso. For the MDA-MB-231 breast cancer line, we could show that metastases are present in deep tissues such as the brain or distant locations such as hindlimb bones as early as 2 days after cardiac injection. DeepMACT also allowed us to assess important differences between distinct cancer models in terms of overall metastatic propensity and organotropism. For example, as expected from previous reports (Nguyen et al., 2009) the H2030-BrM3 lung cancer line developed the highest fraction of brain metastases among all cancer models tested and also metastasized to bones. However, DeepMACT allowed us to comprehensively characterize the distribution of metastases throughout the body, revealing for instance that this model has a high propensity to metastasize to lungs, but produces much fewer liver metastases than any other model tested. The pancreatic tumor, on the other hand, metastasized neither to the brain nor to the kidneys but disseminated for instance into the peritoneum—a pattern observed commonly in human patients as well as in several different mouse models of the disease (Hingorani et al., 2003; Lenk et al., 2017; Ryan et al., 2014; Schönhuber et al., 2014). Overall, we find that our results agree with the existing literature, but while previous studies were structurally limited to selective analysis of micrometastases in small tissue samples, the results shown here represent the first systematic, unbiased, and comprehensive full-body scale screening for micrometastases for these cancer models.

While precise assessment of therapeutic antibody biodistribution is critical for evaluating its specificity and utility for tumor treatment, there has been no method so far that can provide such information down to the level of individual micrometastases on full body scale scans. Here, we applied DeepMACT to study not only the distribution of single metastases but also of a therapeutic monoclonal antibody. We demonstrated that the on-target and off-target binding of antibody drugs throughout the body can readily be assessed by DeepMACT. For example, we observed that not all micrometastases in the lungs were targeted by the anti-CA12 therapeutic antibody 6A10. Understanding why antibody-based therapeutics do not target all metastases would be important for developing more effective treatments. Toward this goal, we studied the potential mechanisms that could contribute to the lack of targeting. Vascular staining demonstrated that blood vessels were present in the immediate vicinity of all examined metastases in the lungs, suggesting that insufficient vascularization is unlikely to be a common cause for the failure of antibody drug targeting in this model. Interestingly, DeepMACT analysis found that micrometastases located in close proximity are more likely to be targeted. This suggests that the local microenvironment within metastatic niches plays an important role in determining the efficiency of antibody targeting, e.g., by altering antibody penetration, binding affinity and clearance. Furthermore, heterogeneity of antigen expression on the surface of tumor cells and internalization and degradation of antigen/antibody complexes might also affect therapeutic antibody targeting efficiency. While our findings are based on one therapeutic antibody, they nevertheless highlight a potential use case for applying the DeepMACT pipeline in pre-clinical studies aimed at understanding and improving the specificity and efficacy of tumor treatments.

In conclusion, DeepMACT is a powerful technology combining unbiased full body scale imaging with automated analysis. It enables visualization, quantification, and analysis of tumor micrometastases and antibody-based therapies in mice with high resolution and an accuracy equivalent to that of human experts but speeding up the workflow by orders of magnitude compared to traditional methods. Because this technology is time- and cost-efficient, scalable, and easily adoptable, it can be used to study metastasis and optimize antibody-based drug targeting in diverse tumor models.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Spontaneous breast cancer metastasis model
  - ○ Estrogen positive breast cancer model and brain metastatic lung cancer model
  - ○ Pancreatic cancer model
  - ○ Injection of therapeutic antibody
- METHOD DETAILS
  - ○ Perfusion and tissue preparation
  - ○ Tissue clearing and staining
  - ○ Image acquisition
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ General data processing
  - ○ Data annotation by human experts
  - ○ Automatic annotation with fixed filter kernel
  - ○ Manual annotation correction by human experts
  - ○ Refinement of annotation to ground truth
  - ○ Deep learning for metastasis detection
  - ○ Analysis of individual metastases
- DATA AND CODE AVAILABILITY
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.11.013.

A video abstract is available at https://doi.org/10.1016/j.cell.2019.11.013#mmc6.

ACKNOWLEDGMENTS

We thank the Monoclonal Antibody Core Facility at the Helmholtz Center Munich for providing the 6A10 antibody, Lisa Sevenich at the Georg Speyer House for help with the intracardial metastasis model, Annette Frank and Stephanie Hoffmann for performing bioluminescence in vivo imaging for MDA-MB-231 cancer model, and Dr. W. Ouyang for constructive comments on the manuscript. M.I.T. is member of Graduate School of Systemic Neurosciences (GSN), Ludwig Maximilian University of Munich. This work was supported by the Vascular Dementia Research Foundation, Synergy Excellence Cluster Munich (SyNergy), the Helmholtz-Center for Environment Health (to A.E. and R.Z.), Fritz Thyssen Stiftung and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) (to A.E.), the German Federal

Ministry of Education and Research via the Software Campus initiative (to O.S.), DFG (SFB 1321 P06 and S01 to D.S.), the European Research Council (ERC CoG 648521 to D.S.), and the Excellence Cluster Cardio-Pulmonary Institute (to T.A.). I.J. was funded by ERC Consolidator Grant 681524, Mildred Scheel Professorship by DFG, Collaborative Research Center 1243, project A05. NVIDIA supported this work with a Titan XP via the GPU Grant Program.

AUTHOR CONTRIBUTIONS

C.P., A.P.-D., and R.C. performed the tissue processing, clearing, and imaging experiments. C.P. and M.I.T. stitched and assembled the full-body scans. O.S. developed the deep learning architecture and performed the quantitative analyses. J.C. helped optimize the network architecture. A.G. performed the image rendering and preliminary data analysis. O.S. developed annotation tools and the custom-made object detector. M.A.R. annotated the data. C.P. and O.S. reviewed the predictions identified by the algorithm. C.P., H.M., Z.R., and M.I.T. manually segmented the internal organs for registration. C.P., G.G., B.v.N., R.Z., N.B.-S., O.T., S.S., T.A., A.M.C., A.A.-P., K.S., C.V., D.S., and I.J. performed tumor transplantation experiments and bioluminescence imaging. B.F. helped to obtain the animal experiments license. B.K.G. helped with data interpretation. B.M. provided guidance in developing the deep learning architecture and helped with data interpretation. A.E., C.P., and O.S. wrote the manuscript. All the authors edited the manuscript. A.E. initiated and led all aspects of the project.

DECLARATION OF INTERESTS

A.E. has filed a patent related to some of the technologies presented in this work.

Received: February 5, 2019
Revised: October 2, 2019
Accepted: November 12, 2019
Published: December 12, 2019

REFERENCES

Barker, N., and Clevers, H. (2006). Mining the Wnt pathway for cancer therapeutics. Nat. Rev. Drug Discov. *5*, 997–1014.

Battke, C., Kremmer, E., Mysliwietz, J., Gondi, G., Dumitru, C., Brandau, S., Lang, S., Vullo, D., Supuran, C., and Zeidler, R. (2011). Generation and characterization of the first inhibitory antibody targeting tumour-associated carbonic anhydrase XII. Cancer Immunol. Immunother. *60*, 649–658.

Bhatia, S., Sinha, Y., and Goel, L. (2019). Lung Cancer Detection: A Deep Learning Approach (Singapore: Springer).

Bolte, S., and Cordelières, F.P. (2006). A guided tour into subcellular colocalization analysis in light microscopy. J. Microsc. *224*, 213–232.

Cai, R., Pan, C., Ghasemigharagoz, A., Todorov, M.I., Forstera, B., Zhao, S., Bhatia, H.S., Parra-Damas, A., Mrowka, L., Theodorou, D., et al. (2019). Panoptic imaging of transparent mice reveals whole-body neuronal projections and skull-meninges connections. Nat. Neurosci. *22*, 317–327.

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., and Collins, J.J. (2018). Next-Generation Machine Learning for Biological Networks. Cell *173*, 1581–1592.

Campbell, J.P., Merkel, A.R., Masood-Campbell, S.K., Elefteriou, F., and Sterling, J.A. (2012). Models of bone metastasis. J. Vis. Exp. *67*, e4260.

Christiansen, E.M., Yang, S.J., Ando, D.M., Javaherian, A., Skibinski, G., Lipnick, S., Mount, E., O'Neil, A., Shah, K., Lee, A.K., et al. (2018). In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. Cell *173*, 792–803.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation (Springer International Publishing).

Condeelis, J., and Weissleder, R. (2010). In vivo imaging in cancer. Cold Spring Harb. Perspect. Biol. *2*, a003848.

de Jong, M., Essers, J., and van Weerden, W.M. (2014). Imaging preclinical tumour models: improving translational power. Nat. Rev. Cancer *14*, 481–493.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.; the CAMELYON16 Consortium (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA *318*, 2199–2210.

Ertürk, A., Becker, K., Jährling, N., Mauch, C.P., Hojer, C.D., Egen, J.G., Hellal, F., Bradke, F., Sheng, M., and Dodt, H.-U. (2012). Three-dimensional imaging of solvent-cleared organs using 3DISCO. Nat. Protoc. *7*, 1983–1995.

Eser, S., Reiff, N., Messer, M., Seidler, B., Gottschalk, K., Dobler, M., Hieber, M., Arbeiter, A., Klein, S., Kong, B., et al. (2013). Selective requirement of PI3K/PDK1 signaling for Kras oncogene-driven pancreatic cell plasticity and cancer. Cancer Cell *23*, 406–420.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al. (2019). U-Net: deep learning for cell counting, detection, and morphometry. Nat. Methods *16*, 67–70.

Gondi, G., Mysliwietz, J., Hulikova, A., Jen, J.P., Swietach, P., Kremmer, E., and Zeidler, R. (2013). Antitumor efficacy of a monoclonal antibody that inhibits the activity of cancer-associated carbonic anhydrase XII. Cancer Res. *73*, 6494–6503.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hingorani, S.R., Petricoin, E.F., Maitra, A., Rajapakse, V., King, C., Jacobetz, M.A., Ross, S., Conrads, T.P., Veenstra, T.D., Hitt, B.A., et al. (2003). Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. Cancer Cell *4*, 437–450.

Iorns, E., Drews-Elger, K., Ward, T.M., Dean, S., Clarke, J., Berry, D., El Ashry, D., and Lippman, M. (2012). A new mouse model for the study of human breast cancer metastasis. PLoS ONE *7*, e47995.

Jones, E., Oliphant, T., and Peterson, P. (2001). SciPy: Open Source Scientific Tools for Python. https://www.scipy.org/.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell *172*, 1122–1131.

Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv, arXiv:1412.6980.

Kubota, S.I., Takahashi, K., Nishida, J., Morishita, Y., Ehata, S., Tainaka, K., Miyazono, K., and Ueda, H.R. (2017). Whole-Body Profiling of Cancer Metastasis with Single-Cell Resolution. Cell Rep. *20*, 236–250.

Lambert, A.W., Pattabiraman, D.R., and Weinberg, R.A. (2017). Emerging Biological Principles of Metastasis. Cell *168*, 670–691.

Lenk, L., Pein, M., Will, O., Gomez, B., Viol, F., Hauser, C., Egberts, J.H., Gundlach, J.P., Helm, O., Tiwari, S., et al. (2017). The hepatic microenvironment essentially determines tumor cell dormancy and metastatic outgrowth of pancreatic ductal adenocarcinoma. OncoImmunology *7*, e1368603.

Litjens, G., Sánchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-van de Kaa, C., Bult, P., van Ginneken, B., and van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci. Rep. *6*, 26286.

Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G.E., Smith, J.L., Mohtashamian, A., Olson, N., Peng, L.H., Hipp, J.D., and Stumpe, M.C. (2019). Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection. Arch. Pathol. Lab. Med. *143*, 859–868.

Massagué, J., and Obenauf, A.C. (2016). Metastatic colonization by circulating tumour cells. Nature *529*, 298–306.

Massoud, T.F., and Gambhir, S.S. (2003). Molecular imaging in living subjects: seeing fundamental biological processes in a new light. Genes Dev. *17*, 545–580.

Massoud, T.F., and Gambhir, S.S. (2007). Integrating noninvasive molecular imaging into molecular medicine: an evolving paradigm. Trends Mol. Med. *13*, 183–191.

Mavandadi, S., Dimitrov, S., Feng, S., Yu, F., Sikora, U., Yaglidere, O., Padmanabhan, S., Nielsen, K., and Ozcan, A. (2012a). Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study. PLoS ONE *7*, e37245.

Mavandadi, S., Feng, S., Yu, F., Dimitrov, S., Nielsen-Saines, K., Prescott, W.R., and Ozcan, A. (2012b). A mathematical framework for combining decisions of multiple experts toward accurate and remote diagnosis of malaria using tele-microscopy. PLoS ONE *7*, e46192.

McKinney, W. (2008). Pandas. https://pandas.pydata.org/.

Nguyen, D.X., Chiang, A.C., Zhang, X.H., Kim, J.Y., Kris, M.G., Ladanyi, M., Gerald, W.L., and Massagué, J. (2009). WNT/TCF signaling through LEF1 and HOXB9 mediates lung adenocarcinoma metastasis. Cell *138*, 51–62.

Ntziachristos, V. (2010). Going deeper than microscopy: the optical imaging frontier in biology. Nat. Methods *7*, 603–614.

Pan, C., Cai, R., Quacquarelli, F.P., Ghasemigharagoz, A., Lourbopoulos, A., Matryba, P., Plesnila, N., Dichgans, M., Hellal, F., and Erturk, A. (2016). Shrinkage-mediated imaging of entire organs and organisms using uDISCO. Nat. Methods *13*, 859–867.

Pandey, M., and Mahadevan, D. (2014). Monoclonal antibodies as therapeutics in human malignancies. Future Oncol. *10*, 609–636.

Paszke, A. (2016). PyTorch. https://pytorch.org/.

Pichler, B.J., Wehrl, H.F., and Judenhofer, M.S. (2008). Latest advances in molecular imaging instrumentation. J. Nucl. Med. *49* (*Suppl 2*), 5S–23S.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation (Springer International Publishing).

Ryan, D.P., Hong, T.S., and Bardeesy, N. (2014). Pancreatic adenocarcinoma. N. Engl. J. Med. *371*, 1039–1049.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nature Methods *9*, 671–675.

Schönhuber, N., Seidler, B., Schuck, K., Veltkamp, C., Schachtler, C., Zukowska, M., Eser, S., Feyerabend, T.B., Paul, M.C., Eser, P., et al. (2014). A next-generation dual-recombinase system for time- and host-specific targeting of pancreatic cancer. Nat. Med. *20*, 1340–1347.

Sevenich, L., Bowman, R.L., Mason, S.D., Quail, D.F., Rapaport, F., Elie, B.T., Brogi, E., Brastianos, P.K., Hahn, W.C., Holsinger, L.J., et al. (2014). Analysis of tumour- and stroma-supplied proteolytic networks reveals a brain-metastasis-promoting role for cathepsin S. Nat. Cell Biol. *16*, 876–888.

Steiner, D.F., MacDonald, R., Liu, Y., Truszkowski, P., Hipp, J.D., Gammage, C., Thng, F., Peng, L., and Stumpe, M.C. (2018). Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. Am. J. Surg. Pathol. *42*, 1636–1646.

Sullivan, D.P., Winsnes, C.F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., Nordgren, A., et al. (2018). Deep learning is combined with massive-scale citizen science to improve large-scale image classification. Nat. Biotechnol. *36*, 820–828.

Susaki, E.A., Tainaka, K., Perrin, D., Kishino, F., Tawara, T., Watanabe, T.M., Yokoyama, C., Onoe, H., Eguchi, M., Yamaguchi, S., et al. (2014). Whole-brain imaging with single-cell resolution using chemical cocktails and computational analysis. Cell *157*, 726–739.

Tabrizi, M., Bornstein, G.G., and Suria, H. (2010). Biodistribution mechanisms of therapeutic monoclonal antibodies in health and disease. AAPS J. *12*, 33–43.

Tainaka, K., Kubota, S.I., Suyama, T.Q., Susaki, E.A., Perrin, D., Ukai-Tadenuma, M., Ukai, H., and Ueda, H.R. (2014). Whole-body imaging with single-cell resolution by tissue decolorization. Cell *159*, 911–924.

Timpson, P., McGhee, E.J., and Anderson, K.I. (2011). Imaging molecular dynamics in vivo–from cell biology to animal models. J. Cell Sci. *124*, 2877–2890.

Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. *25*, 44–56.

Tuchin, V.V. (2016). Editor's Introduction: Optical Methods for Biomedical Diagnosis. In Handbook of Optical Biomedical Diagnostics, Second Edition (Spie Press).

Vestjens, J.H., Pepels, M.J., de Boer, M., Borm, G.F., van Deurzen, C.H., van Diest, P.J., van Dijck, J.A., Adang, E.M., Nortier, J.W., Rutgers, E.J., et al. (2012). Relevant impact of central pathology review on nodal classification in individual breast cancer patients. Ann. Oncol. *23*, 2561–2566.

Vick, B., Rothenberg, M., Sandhöfer, N., Carlet, M., Finkenzeller, C., Krupka, C., Grunert, M., Trumpp, A., Corbacioglu, S., Ebinger, M., et al. (2015). An advanced preclinical mouse model for acute myeloid leukemia using patients' cells of various genetic subgroups and in vivo bioluminescence imaging. PLoS ONE *10*, e0120925.

von Burstin, J., Eser, S., Seidler, B., Meining, A., Bajbouj, M., Mages, J., Lang, R., Kind, A.J., Schnieke, A.E., Schmid, R.M., et al. (2008). Highly sensitive detection of early-stage pancreatic cancer by multimodal near-infrared molecular imaging in living mice. Int. J. Cancer *123*, 2138–2147.

von Burstin, J., Eser, S., Paul, M.C., Seidler, B., Brandl, M., Messer, M., von Werder, A., Schmidt, A., Mages, J., Pagel, P., et al. (2009). E-cadherin regulates metastasis of pancreatic cancer in vivo and is suppressed by a SNAIL/HDAC1/HDAC2 repressor complex. Gastroenterology *137*, 361–371.

von Neubeck, B., Gondi, G., Riganti, C., Pan, C., Parra Damas, A., Scherb, H., Ertürk, A., and Zeidler, R. (2018). An inhibitory antibody targeting carbonic anhydrase XII abrogates chemoresistance and significantly reduces lung metastases in an orthotopic breast cancer model in vivo. Int. J. Cancer *143*, 2065–2075.

Wang, N., Xu, M., Yu, J., Qin, C., Luo, X., Yang, X., Wang, T., Li, A., and Ni, D. (2018). Densely Deep Supervised Networks with Threshold Loss for Cancer Detection in Automated Breast Ultrasound. In Medical Image Computing and Computer Assisted Intervention (MICCAI 2018) 21st International Conference Proceedings, pp. 641–648.

Wang, J., Fang, Z., Lang, N., Yuan, H., Su, M.Y., and Baldi, P. (2017). A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks. Comput. Biol. Med. *84*, 137–146.

Wang, H., Rivenson, Y., Jin, Y., Wei, Z., Gao, R., Günaydın, H., Bentolila, L.A., Kural, C., and Ozcan, A. (2019). Deep learning enables cross-modality super-resolution in fluorescence microscopy. Nat. Methods *16*, 103–110.

Waskom, M. (2012). seaborn: statistical data visualization. https://seaborn.pydata.org/.

Yang, B., Treweek, J.B., Kulkarni, R.P., Deverman, B.E., Chen, C.K., Lubeck, E., Shah, S., Cai, L., and Gradinaru, V. (2014). Single-cell phenotyping within transparent intact tissue through whole-body clearing. Cell *158*, 945–958.

Zipfel, W.R., Williams, R.M., Christie, R., Nikitin, A.Y., Hyman, B.T., and Webb, W.W. (2003). Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation. Proc. Natl. Acad. Sci. USA *100*, 7075–7080.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Human carbonic anhydrase (CA) XII-specific antibody (6A10)	Battke et al., 2011	https://doi.org/10.1007/s00262-011-0980-z
Anti-Firefly Luciferase antibody	Abcam	ab21176; RRID:AB_446076
AlexaFluor 488 goat anti-rabbit IgG antibody	Life Technologies	A11034; RRID:AB_2576217
Chemicals, Peptides, and Recombinant Proteins		
Phosphate Buffer Saline containing Heparin	Ratiopharm GmbH	N68542.03
4% paraformaldehyde (PFA)	Morphisto	11762.01000
CUBIC reagent – urea	Carl Roth	3941.3
CUBIC reagent – Ethylenediamine	Sigma-Aldrich	122262
CUBIC reagent – Triton X-1000	AppliChem	A4975,1000
EDTA	Carl Roth	1702922685
Sodium hydroxide	Sigma-Aldrich	71687
Goat serum	GIBCO	16210072
Bovine Serum Albumin	Sigma-Aldrich	A7906
Methyl-beta-Cyclodextrin	Sigma-Aldrich	332615
trans-1-Acetyl-4-hydroxy-L-proline	Sigma-Aldrich	441562
Sodium azide	Sigma-Aldrich	71290
DISCO solution – *tert*-butanol	Carl Roth	AE16.3
DISCO solution – Tetrahydrofuran	Sigma-Aldrich	186562
DISCO solution – Dichloromethane	Sigma-Aldrich	270997
DISCO solution – Benzyl alcohol	Sigma-Aldrich	24122
DISCO solution – Benzyl benzoate	Sigma-Aldrich	W213802
Diphenyl ether	Alfa Aesar	A15791
Vitamin E (DL-alpha-tocopherol)	Alfa Aesar	A17039
Atto647N conjugated anti-RFP/mCherry nanobody	Chromotek	rba647n-100; RRID:AB_2631440
Atto594 conjugated anti-RFP/mCherry nanobody	Chromotek	rba594-100; RRID:AB_2631390
Atto647N conjugated anti-GFP nanobody	Chromotek	gba647n-100; RRID:AB_2629215
Hoechst 33342	Thermo Fisher Scientific	21492H
Propidium iodide	Sigma-Aldrich	P4864
Gelatin	Sigma-Aldrich	G2500
Alexa 488 conjugated Lectin	Invitrogen	W11261
Fluorescent mounting medium	Dako	10097416
RPMI 1640 medium	GIBCO	11875093
Deposited Data		
Raw data and data labels for DeepMACT	This paper	http://discotechnologies.org/DeepMACT/
Experimental Models: Cell Lines		
Human: MDA-MB-231 breast cancer cells	von Neubeck et al., 2018	https://doi.org/10.1002/ijc.31607
Human: MCF-7	ATCC	ATCC HTB-22
Human: H2030-BrM3	Nguyen et al., 2009	https://doi.org/10.1016/j.cell.2009.04.030
Murine: R254	von Burstin et al., 2008	https://doi.org/10.1002/ijc.23780

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Organisms/Strains		
NOD/SCID/IL2 receptor gamma chain (NSG) knockout mouse line: NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ	Jackson Laboratory	005557
NMRI nude mouse line: Rj:NMRI-*Foxn1*$^{nu/nu}$	Janvier Labs	NMRI-nu
C57BL/6J mouse line	Jackson Laboratory	000664
Software and Algorithms		
ImageJ	Schneider et al., 2012	https://imagej.nih.gov/ij/
AxioZoom EMS3 software	Carl Zeiss AG	https://www.zeiss.com/microscopy/int/products/stereo-zoom-microscopes/axio-zoom-v16.html#downloads
Living Image software 4.2	Caliper Life Sciences	https://www.perkinelmer.com/lab-products-and-services/resources/in-vivo-imaging-software-downloads.html
Photoshop CS6	Adobe	https://www.adobe.com/products/photoshop.html
ImSpector	Aberrior/LaVision	https://www.lavisionbiotec.com/
Amira	FEI Visualization Sciences Group	http://www.vsg3d.com/
Imaris	Bitplane AG	https://imaris.oxinst.com/
Vision4D	Arivis	https://www.arivis.com/de/imaging-science/arivis-vision4d
Python Anaconda distribution	Anaconda	https://www.anaconda.com/distribution/
Scipy package for Python	Jones et al., 2001	https://www.scipy.org
Seaborn package for Python	Waskom, 2012	https://seaborn.pydata.org/
PyTorch deep learning framework for Python	Paszke, 2016	https://pytorch.org/
Cuda	NVIDIA	https://developer.nvidia.com/cuda-downloads
CuDNN	NVIDIA	https://developer.nvidia.com/cudnn
DeepMACT algorithm	This paper	http://discotechnologies.org/DeepMACT/
Other		
Online demonstration (“compute capsule” on CodeOcean.com) of DeepMACT	This paper	https://codeocean.com/capsule/8c13691f-7f9a-4af4-8522-c26f581c9e83/tree?ID=a8ba18d2bf5046b08fafe2d6a42bfd7a
Resource website for DeepMACT	This paper	http://discotechnologies.org/DeepMACT/
Resource website for DISCO clearing	Cai et al., 2019	http://discotechnologies.org/vDISCO/

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Ali Ertürk (erturk@helmholtz-muenchen.de). The lab protocol as well as the algorithms and data for the DeepMACT pipeline are freely available and have been deposited to http://discotechnologies.org/DeepMACT/.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Spontaneous breast cancer metastasis model

Female NSG (NOD/SCID/IL2 receptor *gamma* chain knockout) mice were obtained from Jackson Laboratory and housed at the animal facility of the Helmholtz Center Munich and the Institute of Stroke and Dementia research Munich. All animal experiments were conducted according to institutional guidelines of the Ludwig Maximilian University of Munich and Helmholtz Center Munich after approval of the Ethical Review Board of the Government of Upper Bavaria (Regierung von Oberbayern, Munich, Germany). MDA-MB-231 breast cancer cells transduced with a lentivirus expressing mCherry and enhanced firefly luciferase (Vick et al., 2015) were counted, filtered through a 100 μm filter and resuspended in RPMI 1640 medium (GIBCO, 11875093). 2x10^6 cells per mouse were injected transdermally in a volume of 50 μl into the 4th left mammary fat pad of 3-4 months old female NSG mice. For the intracardial injection model used for the time-course study, 1x10^5 cells per mouse were injected in a volume of 100 μl PBS into the left ventricle of female NSG mice as described before (Campbell et al., 2012). In brief, the mice were anesthetized using an isoflurane vaporizer and placed ventral side up on a heating pad to keep the body temperature around 37°C. Then the chest area was shaved and cleaned by 70% ethanol. The midway point between top of xiphoid process and the sternal notch was marked and the injection point slightly on the left (anatomical) side of sternum was defined. Finally, the injection was conducted by a 0.5 mL insulin syringe (B.Braun, Omnican 50, U100 Insulin 0.5 mL / 50 I.U, 30G x ½'', 9151125) with a bright red blood pulse back in the syringe as a successful sign. After injection, gentle pressure around the injection site was applied to prevent inner bleeding and the mice were kept in a recovery chamber (Mediheat, 34-0516) at 30°C until they fully recovered from the anesthesia.

Tumor growth was monitored by bioluminescence measurement (photons/second) of the full body using an IVIS Lumina II Imaging System (Caliper Life Sciences) as described (Vick et al., 2015). Briefly, mice were anesthetized with isoflurane, fixed in the imaging chamber and imaged 15 minutes after Luciferin injection (150 mg/kg; i.p.). Bioluminescence signal was quantified using the Living Image software 4.2 (Caliper Life Sciences).

Estrogen positive breast cancer model and brain metastatic lung cancer model

Animal experiments were approved by the veterinary department of the regional council in Darmstadt, Hesse, Germany. Xenograft transplantations were performed in athymic 5-6 week old female NMRI nu/nu mice (Janvier Labs) that were kept in a specific pathogen-free animal facility according to the institutional guidelines of the University of Giessen and University of Frankfurt. Intracardial injections were performed as described before (Sevenich et al., 2014). In brief, prior to the tumor cell injection, subconfluent cells, lentivirally transduced with a construct expressing mCherry and enhanced firefly luciferase, were harvested and kept on ice in sterile PBS until the inoculation. Mice were anesthetized with 100 mg/kg ketamine and 10 mg/kg xylazine and the depth of anesthesia was confirmed by the absence of toe reflexes. The chest was sterilized using 70% ethanol and 1x10^{5} MCF-7 cells, or 5x10^{4} H2030-BrM3 in a total volume of 100 μl PBS were injected stepwise into the left cardiac ventricle using a 26G needle. Success of the injections was monitored by pulsating reflux of arterial blood into the syringe. Metastatic growth was monitored by *in vivo* bioluminescence using an IVIS Lumina II Imaging System 5 minutes after an intraperitoneal injection of 150 mg/kg luciferin.

Pancreatic cancer model

Immunocompetent (wild-type C57BL/6) mice were housed at the animal facility of the Klinikum rechts der Isar of TUM. All animal studies were conducted in compliance with European guidelines for the care and use of laboratory animals and were approved by the Institutional Animal Care and Use Committees (IACUC) of Technische Universität München, Regierung von Oberbayern and UK Home Office. The low passaged primary pancreatic cancer cell line R254, derived from a genetically engineered KPC mouse (LSL-KrasG12D/+;LSL-Trp53R172H/+;Ptf1aCre/+) on a C57BL/6 background as described previously (Eser et al., 2013; von Burstin et al., 2009; von Burstin et al., 2008), was transduced with lentiviral particles expressing EGFP and Firefly Luciferase. 2.5x10^{3} cells per mouse in 20 μl Dulbecco's modified Eagle medium were implanted orthotopically into the pancreas of 2-3 months old male mice. Tumor growth was monitored by bioluminescence measurement of the entire body. In brief, mice were anesthetized with midazolam/medetomidine/fentanyl, injected with D-luciferin (Synchem, Kassel, Germany) at 150 mg/kg intraperitoneally (IP) and imaged after 10 minutes using a cooled back-thinned, charge-coupled device camera (OrcaII ER, Hamamatsu, Herrsching, Germany) equipped with an image intensifier for 10-120 s; bin size, 2; gain, 700. A photographic grayscale image was taken, and the bioluminescent signals were displayed in pseudocolors and projected on the grayscale image using SimplePCI software (Hamamatsu).

Injection of therapeutic antibody

9 weeks after tumor cell injections, one mouse was randomly chosen for different experimental procedures including injection of a human carbonic anhydrase (CA) XII-specific antibody (6A10) (Battke et al., 2011). In brief, 20 μg of 6A10 antibody conjugated with Alexa-568 was injected into the tail vein of the mouse. 48 hours later, the mouse was perfused for vDISCO pipeline including enhancing endogenous mCherry fluorescence and clearing as described in the Method Details section.

METHOD DETAILS

Perfusion and tissue preparation

The mice were deeply anesthetized using a combination of midazolam, medetomidine and fentanyl (MMF) (1ml/100 g of body mass for mice; i.p.). Then, the chest cavity of the animals were opened for the standard intracardial perfusion with heparinized 0.01 M PBS (10-25 U/ml of Heparin as final concentration, Ratiopharm, N68542.03; 100-125 mmHg pressure using a Leica Perfusion One system) for 5-10 minutes at room temperature until the blood was washed out, followed by 4% paraformaldehyde (PFA) in 0.01 M PBS (pH 7.4) (Morphisto, 11762.01000) for 10-20 minutes. The skin was carefully removed and the mouse bodies were postfixed in 4% PFA for 1 day at 4°C and transferred to 0.01 M PBS.

Tissue clearing and staining

uDISCO mouse body clearing

The uDISCO protocol to clear bodies of mice was described previously (Pan et al., 2016). In brief, a transcardial-circulatory system was established involving a peristaltic pump (ISMATEC, REGLO Digital MS-4/8 ISM 834; reference tubing, SC0266). Two channels from the pump were set for the circulation through the heart into the vasculature: the first channel pumped the clearing solution into the mouse body and the second channel collected the solution exiting the mouse body and recirculated the solution back to the original bottle. For the outflow tubing of the first channel, which injected the solution into the heart, the tip of a syringe (cut from a 1 mL syringe-Braun, 9166017V) was used to connect the perfusion needle (Leica, 39471024) to the tubing. Meanwhile, the inflow tubing of the second channel, which recirculated the clearing solutions, was fixed to the glass chamber containing the mouse body. The amount of solutions for circulation depended on the capacity of the clearing glass chamber. For example, if the maximum volume of glass chamber is 400 ml, 300 mL of volume of solution was used for circulation.

All clearing steps were performed in a fume hood. First, the mouse body was put in a glass chamber and the perfusion needle was inserted into the heart through the same hole that was used for PFA perfusion. Then, after covering the chamber with aluminum foil the transcardial circulation was started with a pressure of 230 mmHg (60 rpm on the ISMATEC pump). The mouse bodies were perfused for 6 hours with the following gradient of *tert*-butanol (Carl Roth, AE16.3): 30 Vol%, 50 Vol%, 70 Vol%, 90 Vol% (in distilled water),100 Vol% twice, and finally with the refractive index matching solution BABB-D4 containing 4 parts BABB (benzyl alcohol + benzyl benzoate 1:2, Sigma, 24122 and W213802), 1 part diphenyl ether (DPE) (Alfa Aesar, A15791) and 0.4% Vol vitamin E (DL-alpha-tocopherol, Alfa Aesar, A17039), for at least 6 hours until achieving transparency of the bodies. As the melting point of *tert*-butanol is between 23 to 26°C, a heating mat set at 35-40°C was used for the two rounds of 100% *tert*-butanol circulation to prevent the solution from solidifying.

vDISCO mouse body immunostaining and clearing

The detailed protocol of vDISCO was described previously (Cai et al., 2019). The following nanobodies and dyes were used for mouse body immunostaining: Atto647N conjugated anti-RFP/mCherry signal-enhancing nanobodies (Chromotek, rba647n-100), Atto594 conjugated anti-RFP/mCherry signal-enhancing nanobodies (Chromotek, rba594-100), Atto647N conjugated anti-GFP signal-enhancing nanobodies (Chromotek, gba647n-100), Hoechst 33342 (Thermo Fisher Scientific, 21492H), Propidium iodide (PI, Sigma, P4864). Please note that different batch of nanoboosters coming from different companies can have different penetration and stability performances. Please check http://www.discotechnologies.org/vDISCO/ for updates on which nanoboosters to use.

To perform the mouse body immunolabeling, a simplified transcardial-circulatory system using the same type of peristaltic pump was established (ISMATEC, REGLO Digital MS-4/8 ISM 834; reference tubing, SC0266). In short, one reference tubing was connected by two connectors (Omnilab, 5434482) from both ends and extended by additional PVC tubing (Omnilab, 5437920). The head part from a 1 mL syringe (Braun, 9166017V) was cut and inserted into the outflow PVC tubing as a connector for the perfusion needle (Leica, 39471024). Next, a PBS perfused and PFA fixed mouse body was placed into a 250 mL glass chamber (Omnilab, 5163279) and 200 mL of 0.01 M PBS was filled immediately into the chamber. Note that the sample will be kept in the same chamber through the entire immunolabeling and clearing process, it should be always embedded in the respective solutions till the moment of imaging. Then, the inflow tubing of the transcardial-circulatory system was fixed underneath the surface of PBS in the glass chamber using adhesive tape and the pumping circulation was started until the air bubbles were completely removed from the tubing system. The mouse body decolorization, decalcification and immunolabeling steps were conducted subsequently after inserting and fixing the perfusion needle into the heart of the sample through the same pinhole made during sample preparation.

In general, the animals were first perfused with decolorization solution for 2-3 days at room temperature to remove remaining heme and blood before immunostaining. The decolorization solution which is a 1:4 dilution of CUBIC reagent 1 (Susaki et al., 2014) in 0.01 M PBS was refreshed twice during the decolorization step. CUBIC reagent 1 was prepared as a mixture of 25 wt% N,N,N,N'-tetrakis (2-hydroxypropyl) ethylenediamine (Sigma-Aldrich, 122262), 25 wt% urea (Carl Roth, 3941.3) and15 wt% Triton X-100 in 0.01 M PBS, as described in the original publication. Next, after washing with 0.01 M PBS for 3 hours 3 times, the samples were perfused with the decalcification solution (10 wt/vol% EDTA in 0.01 M PBS, pH to 8–9, Carl Roth, 1702922685) for 2 days and for 1 more day with permeabilization solution containing 0.5% Triton X-100, 1.5% goat serum (GIBCO, 16210072), 0.5 mM of Methyl-beta-cyclodextrin (Sigma, 332615), 0.2% trans-1-Acetyl-4-hydroxy-L-proline (Sigma, 441562), 0.05% sodium azide (Sigma, 71290) in 0.01 M PBS. Before the immunostaining step, additional 0,22 μm syringe filters (Sartorius 16532) were attached to the inflow tubing to prevent the potential accumulation of nanobody aggregates and high pressure pumping at 160–230 mmHg (45–60 rpm) was maintained through the entire labeling process. Then the immunostaining solution was prepared as a mixture of permeabilization solution and 35 μl of nanobody (stock concentration 0.5 – 1 mg/ml), 10 μg/ml Hoechst or 300 μl of propidium iodide (stock concentration 1mg/ml) and filtered by the same 0,22 μm syringe filter before use. Subsequently the animals were perfused for 5-6 days with 200 mL of immunostaining solution at room temperature and further passively labeled in the same staining solution with extra 5 μL of signal-enhancing nanobody with gentle shaking for 2 days at 37°C or at room temperature. Then the mice were connected back to the circulation system and perfused with washing solution (1.5% goat serum, 0.5% Triton X-100, 0.05% of sodium azide in 0.01 M PBS) for 12 hours twice at room temperature and at the end with 0.01 M PBS for 3 hours 3 times at room temperature.

After completing the mouse body immunolabeling step, the mouse bodies were passively cleared using 3DISCO (Ertürk et al., 2012) at room temperature with gentle shaking (IKA, 2D digital) under a fume hood. For dehydration, the mouse bodies were incubated in 200 mL of the gradient tetrahydrofuran (THF, Sigma, 186562) in distilled water (6-12 hours for each step): 50 Vol% THF, 70 Vol% THF, 80 Vol% THF, 100 Vol% THF and again 100 Vol% THF; then the mouse bodies were incubated for 3 hours in dichloromethane (Sigma, 270997), and finally in BABB until the tissue were rendered completely transparent. During all clearing steps, the glass chamber was sealed with parafilm and covered by aluminum foil to prevent extra solution evaporation and fluorescence quenching. For details, see also the step-by-step handbook (Methods S1).

Rehydration and immunostaining of cleared tissue

Anti-Firefly Luciferase (dilution 1:2000, Abcam, ab21176) and AlexaFluor 488 goat anti-rabbit IgG (H+L) (dilution 1:400, Life Technologies, A11034) were used to verify the specificity of anti-RFP/mCherry signal-enhancing nanobody labeling. After identification of metastases in the lungs of vDISCO-processed mice, lung tissue was dissected and rehydrated by applying the reverse gradient of tert-butanol used for uDISCO clearing, as follows (6 hours each at 37°C with gentle shaking): 100 Vol% twice, 90 Vol%, 70 Vol%, 50 Vol%, 30 Vol% and 0.01 M PBS twice at room temperature. Rehydrated samples were cut into 1 mm sections using a vibratome (Leica, VT1200S) and were incubated in 0.01 M PBS containing 0.2% gelatin (Sigma, G2500), 0.5% Triton X-100, 0.05% sodium azide and 5% normal goat serum for 1 day at 37°C. The sections were then incubated with the primary antibodies diluted in the same solution overnight at 37°C, washed twice in PBS at room temperature, incubated with secondary antibodies diluted in the same solution for 4 hours at 37°C and at the end washed in PBS three times at room temperature (related to Figure S2B).

Lectin vasculature labeling in lung tissue

The bodies of mice were perfused and collected as described above. After checking with epifluorescence stereomicroscopy (Zeiss AxioZoom EMS3/SyCoP3), the lung lobes with multiple metastases were dissected and sliced into 20 μm thick tissue sections by using a cryostat (Leica, CM3050S). The lung sections were washes 2 times with 0.01 M PBS and then incubated in Alexa 488 conjugated Lectin (4 μg/ml, Invitrogen, W11261) at 4°C overnight. The sections were then stained with Hoechst 33342 (10 μg/ml, Thermo Fisher Scientific, 21492H) for 5 minutes at room temperature to visualize the nucleus. After washing 2 times with PBS, the slides were mounted with fluorescent mounting medium (Dako, 10097416) and were ready for confocal microscopy (related to Figure 7A).

mCherry signal enhancement in lung tissue

20 μm thick lung tissue sections were washed with 0.01 M PBS 2 times before starting the enhancement process. One-hour incubation in blocking solution containing 1% Bovine Serum Albumin (Sigma, A7906), 2% goat serum (GIBCO, 16210-072), 0.1% Triton X-100 and 0.05% Tween 20 (Bio-Rad, 161-0781) in PBS, was performed at room temperature. Then the staining solution was prepared in 1% Bovine Serum Albumin and 0.5% Triton X-100 in PBS. Atto647N conjugated anti-RFP/mCherry signal-enhancing nanobodies was diluted 1:500 in the staining solution and the lungs sections were incubated overnight at 4°C. After the treatment with nanobodies, the lungs sections were washed 3 times with PBS for 5 minutes with gentle shaking. After nuclear staining by Hoechst 33342 (10 μg/ml) and post wash with PBS as described before, the slides were mounted with fluorescence mounting medium and were ready for confocal microscopy (related to Figure S2C).

Image acquisition

Epifluorescence stereomicroscopy imaging

Cleared mouse bodies were fixed in the original clearing chamber and were imaged with Zeiss AxioZoom EMS3/SyCoP3 fluorescence stereomicroscope using a 1x long working distance air objective lens (Plan Z 1x, 0.25 NA, Working distance (WD) = 56 mm). The magnification was set as 7x and imaging areas were selected manually to cover the entire mouse bodies. The images were taken with GFP, RFP and Cy5 filters and files were exported as RGB images in JPEG format. For high resolution imaging of individual metastasis, higher zoom factor can be applied up to 112x.

Light-sheet microscopy imaging

Single plane illumination (light-sheet) image stacks were acquired using an Ultramicroscope II (LaVision BioTec), allowing an axial resolution of 4 μm. For low magnification full-body screening of tumor and antibody signals we used a 1x Olympus air objective (Olympus MV PLAPO 1x/0.25 NA [WD = 65 mm]) coupled to an Olympus MVX10 zoom body, which provides zoom-out and -in ranging from 0.63x up to 6.3x. Using 1x objective, we imaged a field of view of 2 × 2.5 cm, covering the entire width of the mouse body. Tile scans with 60% overlap along the longitudinal y axis of the mouse body were obtained from ventral and dorsal surfaces up to 13 mm in depth, covering the entire volume of the body using a z-step of 10 μm. Exposure time was 150 ms, laser power was 3 to 4 mW (70% to 95% of the power level) and the light-sheet width was kept at maximum. Alternatively, the mouse bodies were scanned with a dipping 1.1x objective (LaVision BioTec MI PLAN 1.1x/0.1 NA [WD = 17 mm]) coupled with an Olympus revolving zoom body unit (U-TVCAC). In brief, 3x8 tile scans with 25% overlap were obtained from both sides to 11 mm in depth, covering the entire volume of the body using a z-step of 6 μm. Light-sheet width was set at 80% and exposure time was 80 ms. The laser power was adjusted depending on the intensity of the fluorescent signal to avoid reaching saturation. For the mouse displayed in Figure 6H, the lower part of the jaw was removed to fit the mouse head into imaging chamber; this was not the case for any other mouse presented in this study.

After low magnification imaging of the full scale mouse body, organs (including lungs, liver, kidneys, brain, spleen, intestines and bones) were imaged individually using high magnification objectives (Olympus XLFLUOR 4x corrected/0.28 NA [WD = 10 mm], LaVision BioTec MI PLAN 12x/0.53 NA [WD = 10 mm] and Zeiss 20x Clr Plan-Neofluar/0.1 NA [WD = 4 mm]) coupled to an Olympus revolving zoom body unit (U-TVCAC) kept at 1x. High magnification tile scans were acquired using 20% overlap and the light-sheet width was reduced to obtain maximum illumination in the field of view. For the data used for the comparison of signal profile plots of lung metastases taken in red and far-red channels and for the analysis of endogenous fluorescence signal depletion after the uDISCO protocol, we used the same MVX10 zoom body, coupled this time with a 2x objective (Olympus MVPLAPO2XC/0.5 NA [WD = 6 mm]) at zoom body magnification 6.3x and 2.5x respectively.

Confocal microscopy imaging

For imaging the thick cleared specimens such as dissected tissues, pieces of organs or whole organs were placed on 35 mm glass bottom Petri dishes (MatTek, P35G-0-14-C), then the samples were covered with one or two drops of the refractive index matching solution such as BABB or BABB-D4. Sealing of this mounting chamber was not necessary. The samples were imaged with an inverted laser-scanning confocal microscopy system (Zeiss, LSM 880) using a 40x oil immersion lens (Zeiss, EC Plan-Neofluar

40x/1.30 Oil DIC M27) and a 25x water immersion long-working distance objective lens (Leica, NA 0.95, WD = 2.5 mm), the latter one was mounted on a custom mounting thread. The z-step size was 1-2.5 μm. For imaging the lung tissue sections with Lectin staining, with nanobodies or with Anti-Firefly Luciferase staining, the slides were imaged with the same inverted laser-scanning confocal microscopy system with a z-step size of 2 μm.

Reconstructions of full-body scans

For epifluorescence microscopy reconstructions (2D montage of entire mouse), the collected images were stitched semi-automatically using Adobe Photoshop photomerge function. After saving the stitched images as TIFF or JPEG files, the signal from individual imaging channels can be extracted by using Split Channels function in ImageJ.

For light-sheet microscopy reconstructions (3D montage of entire mouse), the image stacks were acquired and saved by ImSpector (LaVision BioTec GmbH) as 16-bit grayscale TIFF images for each channel separately. The stacks were first stitched with Fiji (ImageJ2) and fused together with Vision4D (Arivis AG). For details, see also the step-by-step handbook (Methods S1). Further image processing was done mostly in Fiji: first, the autofluorescence channel (imaged in 488 excitation) was equalized for a general outline of the mouse body. The organs were segmented manually by defining the regions of interests (ROIs). Data visualization was done with Amira (FEI Visualization Sciences Group), Imaris (Bitplane AG), Vision4D in both volumetric and maximum intensity projection color mapping.

QUANTIFICATION AND STATISTICAL ANALYSIS

General data processing

All data processing after image volume reconstruction was performed in Python using custom scripts based on publicly available standard packages comprising SciPy (Jones et al., 2001), Seaborn (Waskom, 2012), and Pandas (McKinney, 2008). Deep Learning models were built using the PyTorch framework (Paszke, 2016). Since a single full body scan is in the order of several terabytes due to its high resolution (the data used for training had a voxel size of $(10\mu m)^3$), the volume was divided into 1176 subvolumes of $(350px)^3$ (or $(3.5mm)^3$) to enable efficient processing. Subvolumes were overlapping by 50px to ensure any given metastasis is fully captured by at least one subvolume to avoid artifacts of divided metastases at subvolume interfaces. Please note that the size and overlap of subvolumes are design choices that allow easy adaptation to different datasets, e.g., with different SNR, metastasis sizes, or spatial resolution of the scan. Final analyses were conducted on the re-assembled full volume whereby reconcatenation ruled out any double-counting at previously overlapping subvolumes.

Data annotation by human experts

To provide ground truth in the form of a commonly agreed upon reference annotation for training, as well as for evaluation of the algorithms developed, full body scans of two mice (with MDA-MB-231 tumor cells transplanted in the mammary fat pad) were manually annotated by a group of human experts. This manual process was augmented with a set of tools to reduce the total workload from an estimated total duration of several months down to 150 person-hours net annotation time.

Automatic annotation with fixed filter kernel

To avoid starting from scratch to annotate two volumes of several thousand z-slices, an automatic detection and segmentation method was applied to provide a basis for manual correction. Due to the insufficient performance of established methods (in this case: the 3D Object Detector for ImageJ; Bolte and Cordelières, 2006), we developed a custom-made filter based detector tailored to the specifics of this dataset. In brief, we handcrafted a spatial filter kernel optimized to detect the most common metastases and applied it with 3D convolutions to the dataset; subsequent binarization and connected-component analysis yielded *seed points* collocated with metastases. This allowed for further analyses of the immediate local neighborhood of these candidate regions; a local 3D segmentation was derived by selective region growing around these seed points based on the local signal intensity distribution up to a mean *foreground signal* limited to 4 standard deviations above the mean signal in the local surrounding. Finally, obvious false positives were filtered out. Together, this approach generated a first proposal for the data annotation that at least captured the most obvious metastases while producing an acceptable rate of false positives. As shown in the results section, the quality of this proposal was about twice as good as compared to the 3D Object Detector in ImageJ (35% instead of 18% in F1-score). Importantly, further fine-tuning of filters and parameters and any additional automated pre- or post-processing did not improve the results, indicating that a F1-score of 35% may be close to the performance limit of such approaches with fixed filter kernels and fixed decision rules for such kind of data.

Manual annotation correction by human experts

This first proposal served as a basis for human annotation. In general, three kinds of manual correction were needed to derive a good annotation: removal of false positives, addition of false negatives (previously missed metastases) and adjustment of the 3D segmentation of each metastasis. To avoid the need to perform this task individually for each of the 350 layers of a $(350px)^3$ data subvolume, a custom tool with an interactive graphical user interface was developed. Based on maximum intensity projections along each dimension, the tool allowed to review, adjust, add, and remove each potential metastasis in the subvolume with a few mouse clicks, drastically speeding up the annotation process from hours to minutes per subvolume. Different perspectives (X, Y, Z) and viewing modes (e.g., projections, orthogonal slices, adjusted contrasts, 3D renderings) for each individual metastasis allowed the annotator to take maximally informed decisions even in less obvious cases.

Refinement of annotation to ground truth

A small fraction (3%) of the entire dataset was labeled several times by the annotators without their awareness to assess human labeling consistency. Since the difference in annotation for a given subvolume between two trials of a single annotator was about as big as between two independent annotators and quite substantial (the agreement between two trials of the same annotator or between annotators only reached an F1-score of 80%–85%) we decided to invest additional time to refine the entire dataset. First, all experts (3 graduate students with extensive experience in the field of imaging and tumor biology) jointly discussed examples of annotation differences to build a common understanding. Annotations of subvolumes with the biggest discrepancies were again reviewed and refined. Furthermore, this analysis revealed that the most prevalent source of annotation error was overlooked metastases (false negatives). Here, around 29% of metastases were missed in the human annotation, in line with previous studies (Ehteshami Bejnordi et al., 2017; Vestjens et al., 2012). To effectively identify all missed metastases in the entire dataset, our deep learning algorithm (see next section) was trained on the status quo of the annotations and applied to the dataset with high sensitivity. This yielded a long list of potential candidates. With the help of another custom-built, interactive graphical user interface, all potential candidates were manually reviewed by the annotators and either discarded or manually adjusted and added to the segmentation (this is the manual refinement step referred to in Figures 3D and 3E). A small set of potential metastases, for which human annotators could not take a conclusive decision even after joint discussion, was recorded separately, but not added to the segmentation. These laborious steps ensured the generation of a high-quality *ground truth* for training the algorithm and, importantly, for evaluating its performance in comparison to a single human annotator. Here, this selectively iterative approach of refining annotations based on the input of several human experts was chosen due to the substantial amount of manual work involved with reviewing our high-resolution scans. Since a full review of one person takes about a month of full-time work, repeating this process several times would be desirable but too costly. In applications where several, independent full annotations are available, advanced mathematical frameworks for refining decisions from different experts to a single decision can be applied in order to avoid a bias toward individual decisions (Mavandadi et al., 2012a, 2012b).

Deep learning for metastasis detection

Implementation of DeepMACT model architecture

Inspired by the established U-net architecture (Ronneberger et al., 2015), we designed a deep learning approach that is depicted in Figure 3A and briefly described in the results section. The architecture of the CNN at its core (Figure 3C) is characterized by an encoding downward path and a decoding upward path comprising a total of 7 levels, in which each level also has a lateral skip-connection that bypasses the deeper levels and feeds the output of the encoding unit directly to the corresponding decoding unit. Each encoding unit increases the number of feature channels per pixel with the help of two kernel-based convolutions (kernel size: 3; padding: 1; dilation: 1; stride: 1) followed by batch normalization and a rectifying linear unit (ReLU). While the first convolutional step increases the number of feature channels, this number stays constant for the second convolutional step. Before being passed on to the next encoding stage, the spatial resolution is halved using max-pooling (kernel size: 2, stride: 2). Decoding units take two inputs: the output from the previous layer is spatially upsampled by a factor of two (bilinearly) and concatenated along the feature dimension with the output of the corresponding encoding unit, bypassing the deeper levels. A first convolutional step (same parameters as before) decreases the number of feature channels, which is again kept constant in the two subsequent convolutions. The 24-feature channel output of the last decoder is mapped to logits in the 2D space with a convolutional step without padding, batch normalization, or a rectifying linear unit.

Implementations of customized 3D U-net

To compare the DeepMACT approach with CNNs that operate on volumetric data with 3D convolutions, we implemented several customized versions of 3D U-nets (Çiçek et al., 2016). In this alternative approach, the volumetric data is directly fed to the network (instead of projections) to predict a 3D probability volume (without reconstruction from 2D predictions). While following the overall architectural approach of the DeepMACT implementation, we replaced 2D with 3D convolutional operators. We implemented a variety of derivations by changing the total number of levels of en- and decoding units and thus, the maximum number of feature channels, which both drive model complexity in terms of number of parameters. The best-performing implementation consisted of 3 levels of en- and decoding units with a maximum number of feature channels of 48; the corresponding performance values are reported in Figures 3D and 3E (leaner or more complex models yielded comparable or lower performance). All other parameters and procedures (e.g., for training and testing) are identical to the DeepMACT implementation.

Training, validation, and test sets

Following established standards, model training and evaluation was based on k-fold cross-validation (k = 5). Thus, the annotated dataset was split into mutually exclusive sets for training and validation (80%) and for testing (20%). This process was repeated k times, yielding a total of 5 mutually exclusive test sets that are collectively exhaustive. The network weights and all design choices and hyperparameters (such as batch size, learning rate, etc) were optimized solely with the training and validation set to avoid overfitting on the specifics of the test set. The dataset was confined to subvolumes within the torso of the mouse body as subvolumes containing near-zero values outside the body contain no useful information to train or test on. In contrast to all metastases in the entire body, the tumor tissues of the primary tumor and the auxiliary lymph node are several orders of magnitude larger (i.e., they follow very different statistics than all micrometastases) and were thus excluded. The signal from one subvolume was corrupted by a dirt particle and thus also excluded. In total, these exclusions made up less than 1% of the total scan volume. The split between the three subsets (training, validation, testing) of the data was done on a subvolume level (from which the three projections are created afterward) to avoid information leak between different projections from the same subvolume.

Training procedure

The model training was conducted in two steps. First, a large number of models spanning a broad set of different hyperparameters were trained for 10 epochs using another (nested) k-fold cross-validation (k = 5) within the training and validation set. Second, the model with the best-performing set of hyperparameters (presented here) was trained for the remaining epochs. Thus, any hyperparameter choice was made without looking at the performance on the test set. The model was trained for 40 epochs of the entire training dataset, using random vertical and horizontal flips of the data to augment its variance (further training epochs did not improve the predictive power). We used a batch size B of 4 but found that other batch sizes work similarly well. Each input was normalized by its local subvolume peak value, which was found to work better than normalization to the global volume peak value or non-linear normalizations. To calculate the gradients for network weight optimization (i.e., to train the model), we used weighted binary cross entropy as a loss function for a given prediction $\widehat{Y}$ compared to the ground truth Y, giving more weight w for foreground (FG) pixels p versus background pixels (BG) to account for the *class imbalance* (i.e., that metastases are very sparsely distributed in space):

$$wBCE_{\widehat{Y},Y} = (-1)\sum_{b}^{B}\sum_{p}^{P}\left(w_{FG}y_{b,p}\log\left(y_{b,p}+\varepsilon\right)+w_{BG}\left(1-y_{b,p}\right)\log\left(1-y_{b,p}+\varepsilon\right)\right)$$

A small numerical offset $\varepsilon = 10^{-4}$ was applied for numerical stability. We found equal weights or a slightly stronger bias to foreground to work almost equally well (here, we used $w_{FG} = 2$ and $w_{BG} = 0.5$), larger biases had negative effects. Additionally, we allowed the network to optimize the share of training data that contains at least some foreground by ignoring parts of training data in which no foreground is present. A share of 90% training data with at least some foreground optimized the performance on the validation set and was thus chosen. The network was trained using the Adam optimizer (Kingma and Ba, 2014); the initial learning rate was set to 10^{-4} and was gradually decreased by a factor of 10 to a minimum of 10^{-7} every time the loss function reached a plateau for more than 2 epochs. A single training run over 40 epochs takes only around 20-30 minutes on a normal workstation equipped with a NVIDIA Titan XP GPU.

Testing procedure and inference mode

As mentioned before, we applied k-fold cross-validation. Thus, in each of the $k = 5$ folds the model was tested on fresh data that was not seen by the model during training and validation. Together, all 5 sets span the entire annotated dataset. As depicted in Figure 3A, the trained algorithm in inference mode was used to generate probability masks for each of the three projection perspectives (P_{XY}, P_{YZ}, P_{ZX}), in which the pixel value indicates the network's confidence that this pixel is part of a metastasis in the given sub-volume s. Building the outer product of the three probability masks allows to recombine the three judgements of the network in 3D space:

$$P_{V,s} \in [0,1]^3 = \left(P_{XY,s} \in [0,1]^2 \otimes P_{YZ,s} \in [0,1]^2 \otimes P_{ZX,s} \in [0,1]^2\right)^{1/3}$$

This 3D recombination P_V of the 2D probability maps yields a final predicted segmentation mask after binarization. By default, the confidence threshold was set to 50%; however, changing this parameter allows to manually adjust the trade-off between sensitivity and specificity, if desired (also see Figure S4). Please note that the F1-score for evaluation is not affected by this trade-off (i.e., a better detection rate at the cost of a higher false positive rate would not artificially increase the F1-score and vice versa). Subsequent connected-component analysis converts the output to an explicit list and segmentation of predicted metastases in 3D space.

Importantly, DeepMACT did not have to be re-trained to be applied to the full body scans (n = 7) obtained for other tumor models (intracardially injected MDA-MB-231 breast cancer cells, MCF-7 breast cancer cells, R254 pancreatic cancer cells, H2030 lung cancer cells). Thus, the metastasis distribution for these scans could be readily inferred within minutes without further training data annotation.

Performance evaluation

The same performance evaluation procedure was used for the comparison shown in Figure 3D, including the performance of a single human annotator. A standard test for detection tasks, the F1 score quantifies the accuracy of a model by combining precision (share of true positives among all positive predictions, including false positives) and recall (share of predicted positives among the sum of the true positive and false negative predictions). It is mathematically equivalent to the Sørensen–Dice coefficient ("Dice score"), which is the commonly used name for pixel-wise image segmentation. The F1 score is defined as:

$$F1 = 2\cdot\frac{precision\cdot recall}{precision + recall}$$

For all comparisons of detection and segmentation performance, the ground truth (refined by several human experts as described above) is used as a reference. We quantified the performance of the proposed deep learning algorithm based on its prediction of the test set. For a comparison, the segmentations as provided by established tools like the 3D Object Detector in ImageJ, our custom-made detector as described above, as well as the annotation as provided by a single human annotator (before joint refinement) were quantified in the same manner. Overlapping segmentations for metastases were counted as true positive predictions, non-overlapping predictions as false positives and metastases not detected by the prediction as false negatives. Predictions corresponding to the small set of cases unclear to the group of human experts (see above) were neither counted as true positive predictions nor as false negatives, i.e., they neither increased nor decreased the performance evaluation. All performance evaluations were conducted on the entirety of the test set as a whole. To quantify the inherent variance, the distribution of performance results was estimated with n = 1000 resampled test sets (of same size) using the bootstrapping approach. The correctness of the exact three-dimensional outline of each metastasis (segmentation) was verified by assessing the volumetric overlap with the three-dimensional outline drawn by human annotators. This confirmed an overlap accuracy in 3D of 90% for the worst segmentation; importantly, 90% of all detected metastases were segmented with an accuracy of 97.5% or higher.

Analysis of individual metastases

Organ registration

For the full body scale light-sheet scans (e.g., Figures 4C, 4F, and 5) the outlines of selected organs of interest (all lung lobes, brain, both kidneys, heart, liver) were manually segmented as multi-point polygons in a stack of slices in 3D using Fiji. For each metastasis detected by our deep learning architecture we assessed whether its center of mass falls into the 3D segmentation of one of those organs using a custom Python script. Any metastasis not registered to one of these organs is referred to as located in "the rest of the torso" in this manuscript. The 3D segmentation of the lungs was also used to compute the overall lung volume to assess the tumor density in Figure 4J, which we quantified as the share of the sum of the volume of all metastases registered to an organ of the entire organ volume.

Metastasis characterization

The output of our deep learning architecture is a binary segmentation volume for all metastases. We applied connected component analysis to derive an explicit list of metastases fully characterized in 3D. Based on each metastasis 3D shape and voxel-based volume *V*, we computed its average diameter as

$$d_{avg} = 2\sqrt[3]{\frac{3V}{4\pi}}$$

In order to put the metastasis size into context, we computed an estimate of the number of cells per metastasis. To this end, we measured the diameter of single cells and estimated the number of cells for a given metastasis volume based on volumetric extrapolation. We confirmed the accuracy of the estimations by the number of nuclei (PI or Hoechst labeled) in n = 26 samples. On average, the estimates were off by 3.5 cells (10.3%). For 73% of all samples checked, the estimates were off by less than 20%; the worst estimate was 35% off the actual count (39 instead of 60 cells). The estimated numbers may include other cell types present in the metastases apart from tumor cells, e.g., immune cells, vascular cells and fibroblasts. Given that metastasis sizes in full body scans can vary by orders of magnitude (see Figure 4H) this estimation accuracy was deemed sufficient to derive insightful conclusions from cell count estimates.

The distance of each metastasis *i* to its nearest neighboring metastasis was measured in 3D space as the Euclidian distance between their center of masses *CoM*:

$$distNN_i = \min_j \sqrt{\left(CoM_{i,x} - CoM_{j,x}\right)^2 + \left(CoM_{i,y} - CoM_{j,y}\right)^2 + \left(CoM_{i,z} - CoM_{j,z}\right)^2}$$

Drug targeting analysis

We assessed the 6A10 antibody targeting of a given metastasis by analyzing the distribution of the fluorescent signal strength within the 3D segmentation of each metastasis (ξ_m) versus the distribution in its local surrounding (250 μm around the metastasis) ξ_s. For each signal distribution, the number of voxels within the metastasis segmentation n_m or in its local surrounding n_s can be seen as the number of observations of the underlying true (but unknown) distributions. The degree of targeting was estimated by quantifying the ratio of mean signal strength within the segmentation to the mean signal strength in its surrounding (e.g., in Figure 6K). We refer to this as antibody signal ratio. A ratio larger than 1 means that the antibody signal strength within the 3D segmentation of the metastasis is higher than around it (see dashed line in Figure 6K). Whether or not a metastasis was deemed "targeted" was assessed with a version of the t test to determine whether mean of the observed signal distribution in the metastasis ξ_m was significantly at least $\Delta = 50\%$ *(ratio of 1.5)* above the mean of the observed signal distribution in the local surrounding ξ_s. Importantly, a t test is valid for the signals despite their highly non-normal underlying distribution as the number of observations far exceeds the requirements of the central limit theorem (i.e., while the signals are not normally distributed, the estimation of their means is normally distributed due the high number of observations). This was verified manually. However, due to a typically much larger number of observations in the local surrounding ξ_s

than for the metastasis itself ξ_m, the statistical test was not performed with a *Student's t test* but with the *Welch's t test* that corrects the *degrees of freedom* for an unequal number of observations for both distributions:

$$t = \frac{mean(\xi_m) - (1+\Delta)mean(\xi_s)}{\sqrt{\frac{std(\xi_m)^2}{n_m} + \frac{std(\xi_s)^2}{n_s}}}$$

$$DF_{adjusted} = \left(\frac{std(\xi_m)^2}{n_m} + \frac{std(\xi_s)^2}{n_s}\right)^2 \Bigg/ \frac{\left(\frac{std(\xi_m)^2}{n_m}\right)^2}{n_m - 1} + \frac{\left(\frac{std(\xi_s)^2}{n_s}\right)^2}{n_s - 1}$$

Analysis of fluorescence signal profiles

We considered the fluorescence signal profiles from each channel: excitation 470 nm, 561 nm and 647 nm. These profiles were plotted in the same z stack and normalized as percentage over the maximum peak. To compare the reduction of the background and the improvement of the signal over background ratio (SBR) in far-red and near far-red channels, we analyzed lung metastases expressing mCherry imaged with excitation 545/561 nm after uDISCO clearing, lung metastases labeled with anti-mCherry nanobody conjugated with Atto594 imaged with excitation 590 nm and lung metastases labeled with anti-mCherry nanobody conjugated with Atto647N imaged with excitation 640 nm after vDISCO clearing (n = 3 tumors per each experimental group which consisted of 3 animals per each imaging modality). The signal profile was measured from a defined straight line covering the tumors and surrounding tissue background and all the values of the plot from a representative animal per each experimental group were shown in a representative line chart (Figure S1D). Finally, the normalized plots represented in Figure S1E were calculated by normalizing the plots of lung metastases obtained as described above over the average signal intensity of the respective surrounding background.

To compare the signal-to-background ratio (SBR) in Figures S6C and S6D, the samples were labeled with anti-mCherry nanobody conjugated with Atto647N and primary tumors were imaged with excitation 470 nm, 561 nm and 640 nm respectively after vDISCO clearing. Fluorescence signal intensity profiles and background normalized profiles for each channel were plotted with the same strategy as described above.

Metastasis diameter and vessel distance

Metastasis diameters were verified manually. For quantifying the distance between metastases and vessels, ten points on the border of each metastasis were randomly selected and the shortest distance from these points to the closest vessel wall were measured. The presented distance between each metastasis and nearest vessels was quantified by averaging these ten measurements. In Figure 7B, 50 metastases were quantified to generate the scatterplot; each scatter point represents one single metastasis.

DATA AND CODE AVAILABILITY

The lab protocol as well as the algorithms and data for the DeepMACT pipeline are freely available and have been deposited to http://discotechnologies.org/DeepMACT/. For convenience, this includes a fully functional demonstration script (including data).

ADDITIONAL RESOURCES

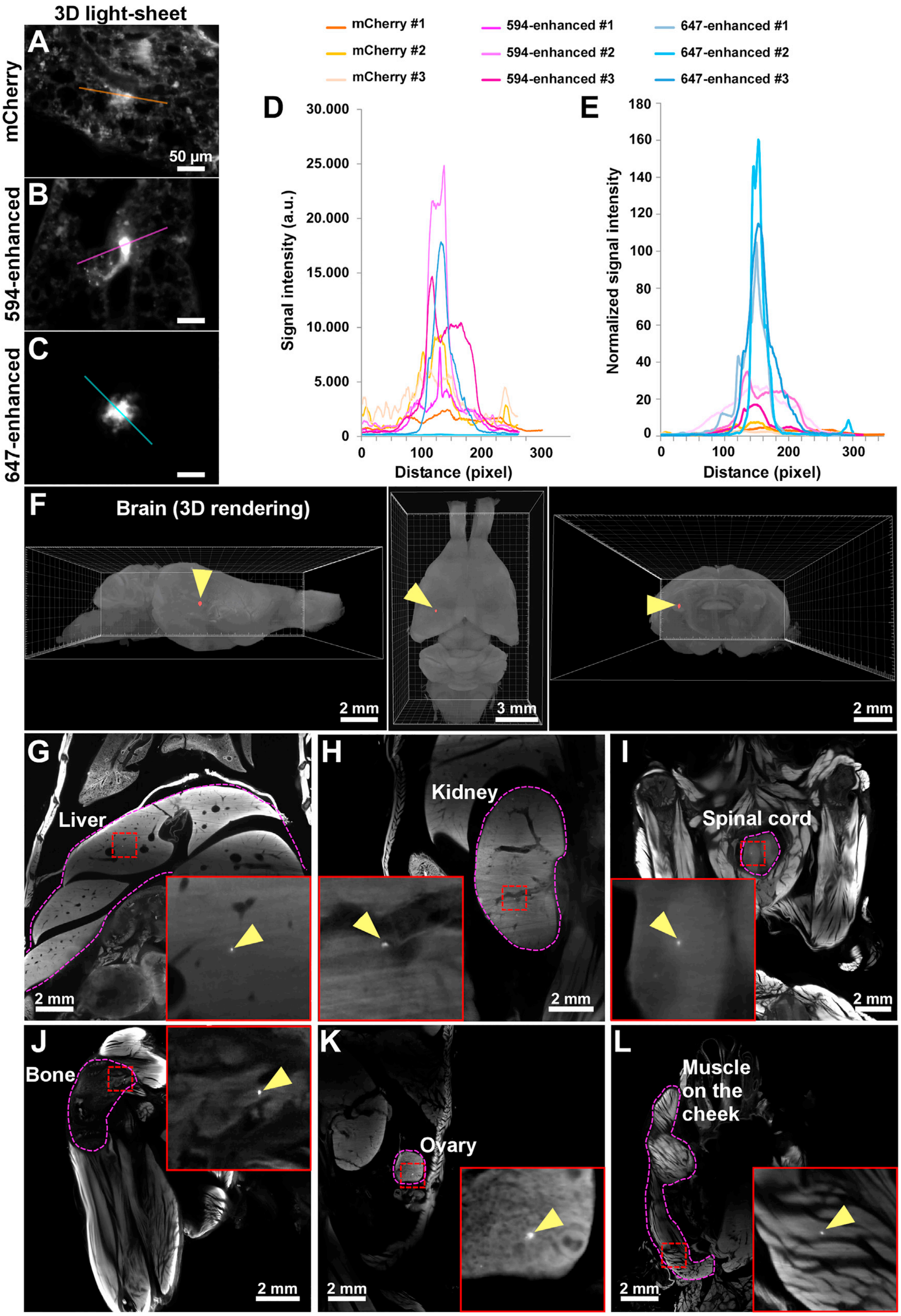
Fully functional online demo of DeepMACT: https://codeocean.com/capsule/8c13691f-7f9a-4af4-8522-c26f581c9e83/tree?ID=a8ba18d2bf5046b08fafe2d6a42bfd7a

Further details on the vDISCO protocol: http://discotechnologies.org/vDISCO/

Registration for in-person workshops: http://discotechnologies.org/workshop/

Videos on tissue clearing: https://www.youtube.com/channel/UCAVXKhQ_ZjEdkAdFR5HwjrQ

Supplemental Figures



(legend on next page)

Figure S1. vDISCO Nanobody Enhancement of the Fluorescent Signal of Cancer Cells, Related to Figure 2

(A-C) Representative light-sheet images of mCherry expressing tumor metastases in the lungs of cleared mice that were not enhanced with nanobodies (A), metastases treated with an anti-mCherry nanobody conjugated to Atto594 (B) or an anti-mCherry nanobody conjugated to Atto647N (C).

(D) Plots of signal intensity profiles along the yellow lines in panels A-C: non-enhanced mCherry (orange), mCherry enhanced with Atto594 (magenta) or mCherry enhanced with Atto647N (cyan) (n = 3 representative metastases).

(E) Intensity profiles of the fluorescence signal in (D) normalized over the background.

(F-I) 3D light-sheet examples of deep-tissue imaging of Atto647N-enhanced tumor metastases in transparent mice after vDISCO. Tumor micrometastases can be detected (yellow arrowheads) which are located several millimeters deep in the brain (F), liver (G), kidney (H) and spinal cord (I) respectively.

(J-L) Examples of micrometastases detected in bone marrow (J), ovary (K) and muscle (L). Note that all the micrometastases shown in G-L were imaged with a 1.1x objective from the MDA-MB-231 tumor model (intracardial injections), except the bone marrow metastasis in (J), which was from the MCF-7 tumor model.

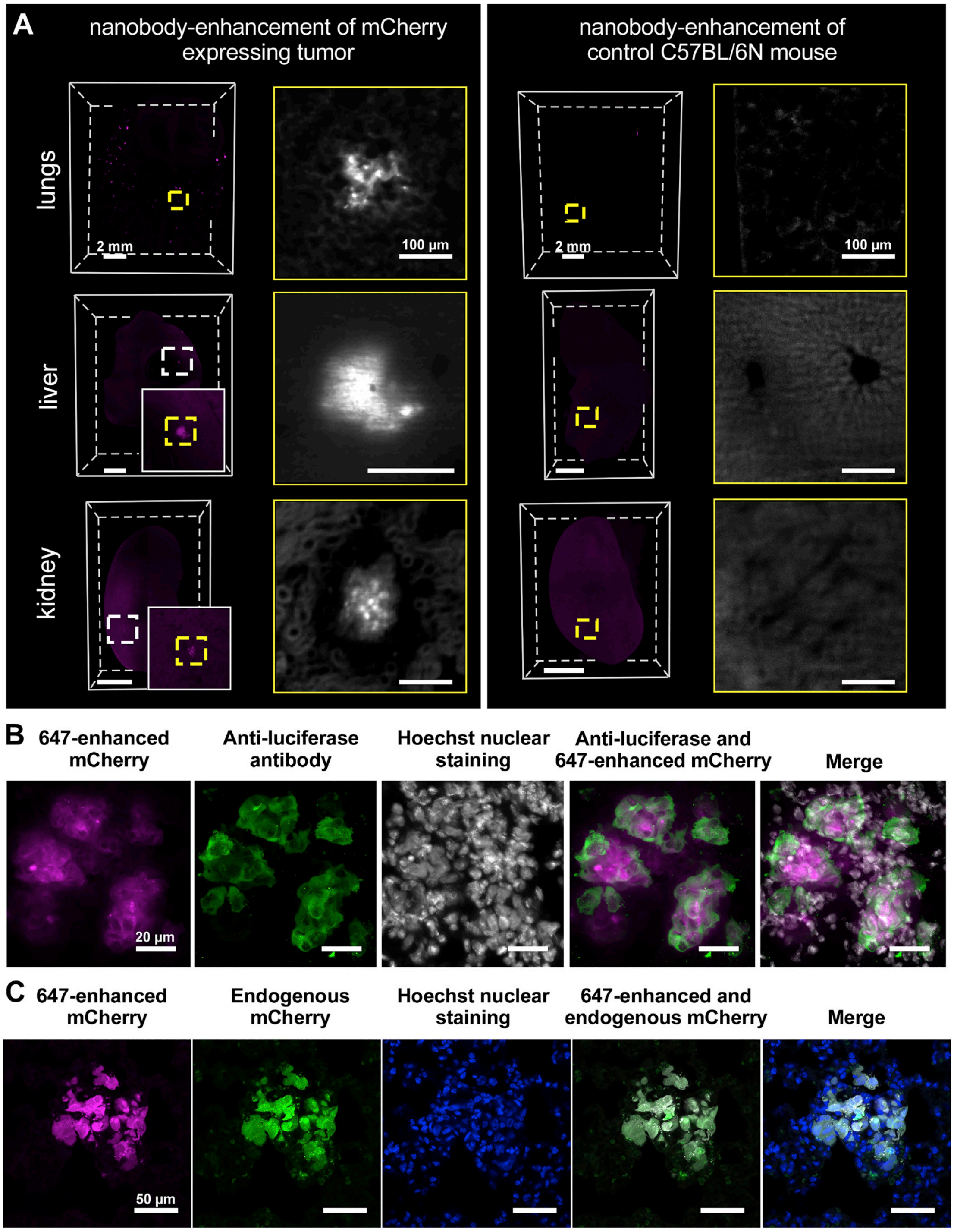


Figure S2. Confirmation of the Specificity of Nanobody-Enhanced Staining in Mice Bearing mCherry-Expressing Tumors, Related to Figure 2

(A) Comparison between an animal bearing an mCherry expressing tumor and a C57BL/6N control animal, which were both enhanced with an anti-mCherry nanobody conjugated to Atto647N (magenta) and imaged by light-sheet microscopy. No signal is detected in organs from the C57BL/6N control. Note that the background is enhanced to demonstrate the absence of signal in the high-magnification images.

(B) Confocal images of metastatic lung tissue immunolabeled with an anti-firefly luciferase antibody (green) after rehydration of the cleared tissue; Atto647N-enhanced cancer cells and cell nuclei are shown in magenta and gray, respectively.

(C) Confocal images of a metastasis in the lung of an animal labeled with an anti-mCherry nanobody conjugated to Atto647N. The enhancing nanobody is shown in magenta, mCherry is shown in green and cell nuclei are shown in blue indicating that nanobody-enhancement specifically detects mCherry.

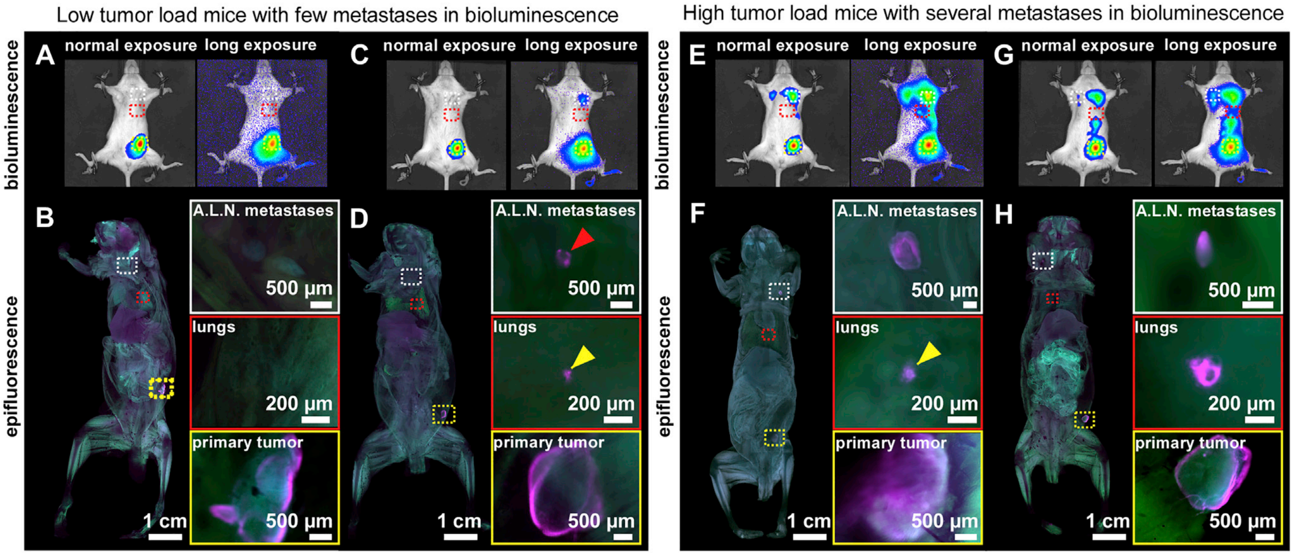


Figure S3. Examples of Tumor Metastasis Detection in Mice Using Bioluminescence Imaging versus vDISCO and Epifluorescence Microscopy, Related to Figure 2

Mice were transplanted in the mammary fat pad with mCherry and firefly luciferase expressing MDA-MB-231 cells and imaged with bioluminescence followed by vDISCO clearing and epifluorescence imaging.

(A-D) We found that bioluminescence imaging with normal exposure is not sufficiently sensitive to detect all the metastases in low tumor load mice. For example, the mice in (A and B) and (C and D) had very similar bioluminescence images with normal exposure. Applying vDISCO to these mice, we found no tumor metastases in one case (A and B) and a large metastasis (red arrowhead) in axillary lymph nodes (A.L.N. metastasis) (C and D) using a fluorescence stereo microscope. Although the signal from the primary tumor is strong in both normal and high exposure bioluminescence images (A and C, yellow box), metastases in lungs (A and C, red boxes) are not visible, but are detected by epifluorescence imaging (D, yellow arrowhead). In epifluorescence images, the tumors (A647N-labeled) are shown in magenta and the background, scanned in 488 nm, is shown in green.

(E-H) In mice with high tumor load, a bulk heatmap of metastatic distribution can be obtained by bioluminescence imaging, without detailed shape and size information. In contrast, vDISCO resolved single micrometastases in whole mouse bodies even with a fluorescence stereo microscope. Especially in the lungs, even micrometastases with a diameter smaller than 100 µm could be resolved in intact mice (F, yellow arrowhead).

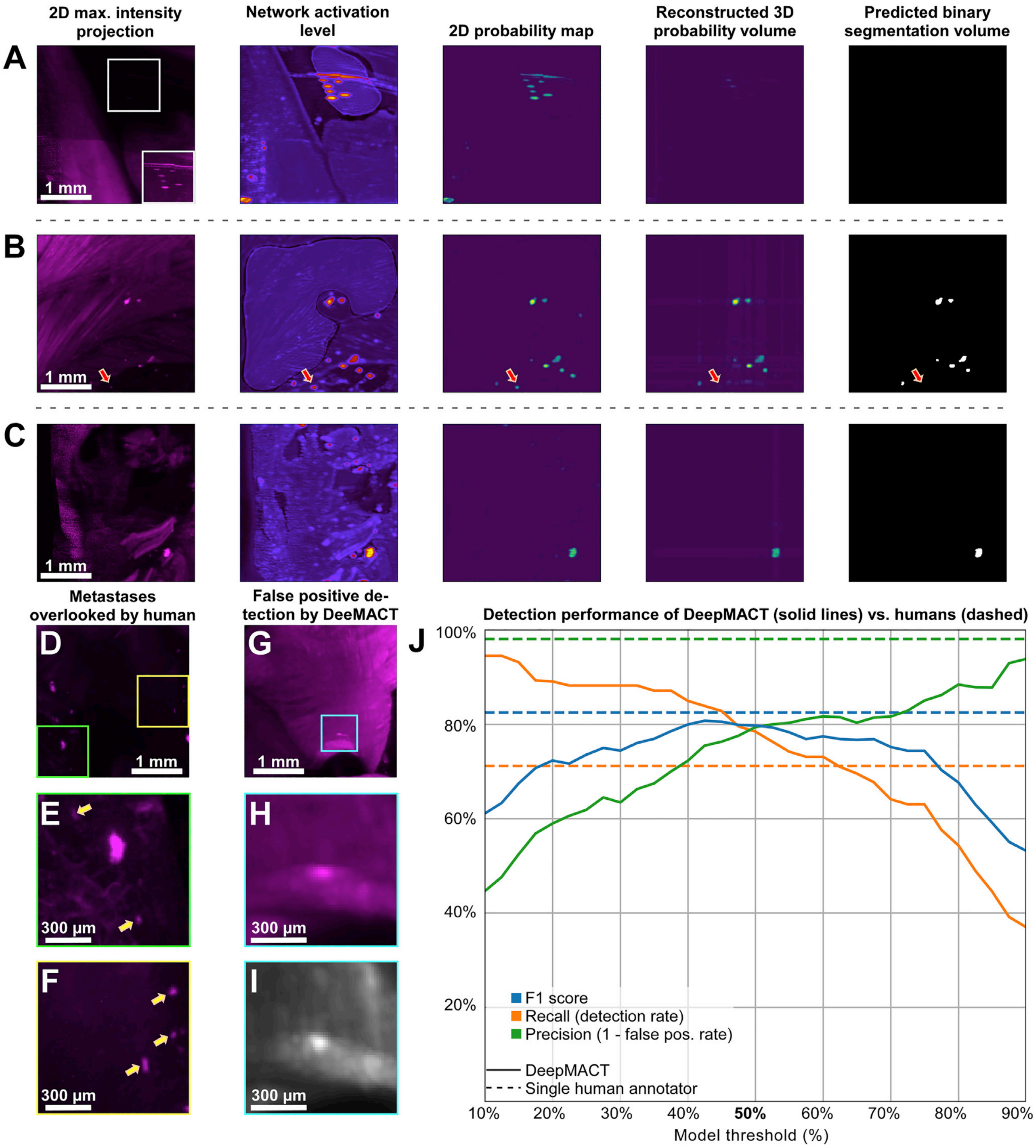

2D max. intensity projection
Network activation level
2D probability map
Reconstructed 3D probability volume
Predicted binary segmentation volume
A
1 mm
B
1 mm
C
1 mm
Metastases overlooked by human
False positive de-tection by DeeMACT
D
1 mm
G
1 mm
E
300 µm
H
300 µm
F
300 µm
I
300 µm
J
Detection performance of DeepMACT (solid lines) vs. humans (dashed)
100%
80%
60%
40%
20%
■ F1 score
■ Recall (detection rate)
■ Precision (1 - false pos. rate)
DeepMACT
Single human annotator
10%
20%
30%
40%
50%
60%
70%
80%
90%
Model threshold (%)

Figure S4. Performance of DeepMACT, Related to Figure 3

(A-C) Visualization of the computational stages of DeepMACT for three different regions. (A) DeepMACT is capable of identifying very low signal peaks but correctly disregards them at the 3D reconstruction stage; the inset shows the region in the white box with a 10-fold increased brightness. (B) While most metastases are correctly identified, few small and dim metastases may be obscured by background structures from some perspectives and consequently be removed at the 3D reconstruction stage (red arrow) (C) In many cases, even a single 2D probability map may already be sufficient for a correct prediction.

(D-F) Examples of metastases that were missed by human annotators but found by DeepMACT (yellow arrows). (E and F) show higher magnifications of the regions in the green and yellow boxes in (D), respectively (regions cropped in 3D); the brightness of (E and F) was increased by 200% compared to (D).

(legend continued on next page)

(G-I) Example of false positive predictions by DeepMACT. The image in (I) shows the same region as in (H) but in the autofluorescence channel (excitation: 488 nm) to confirm that the signal peak in (H) is not caused by metastatic tissue. The region in (H and I) was cropped in 3D.

(J) DeepMACT performance as a function of model confidence threshold (default: 50%) compared to a single human annotator. While the DeepMACT F1 score peaks around a model confidence threshold of 40%–50%, the threshold can be adjusted to increase recall or precision.

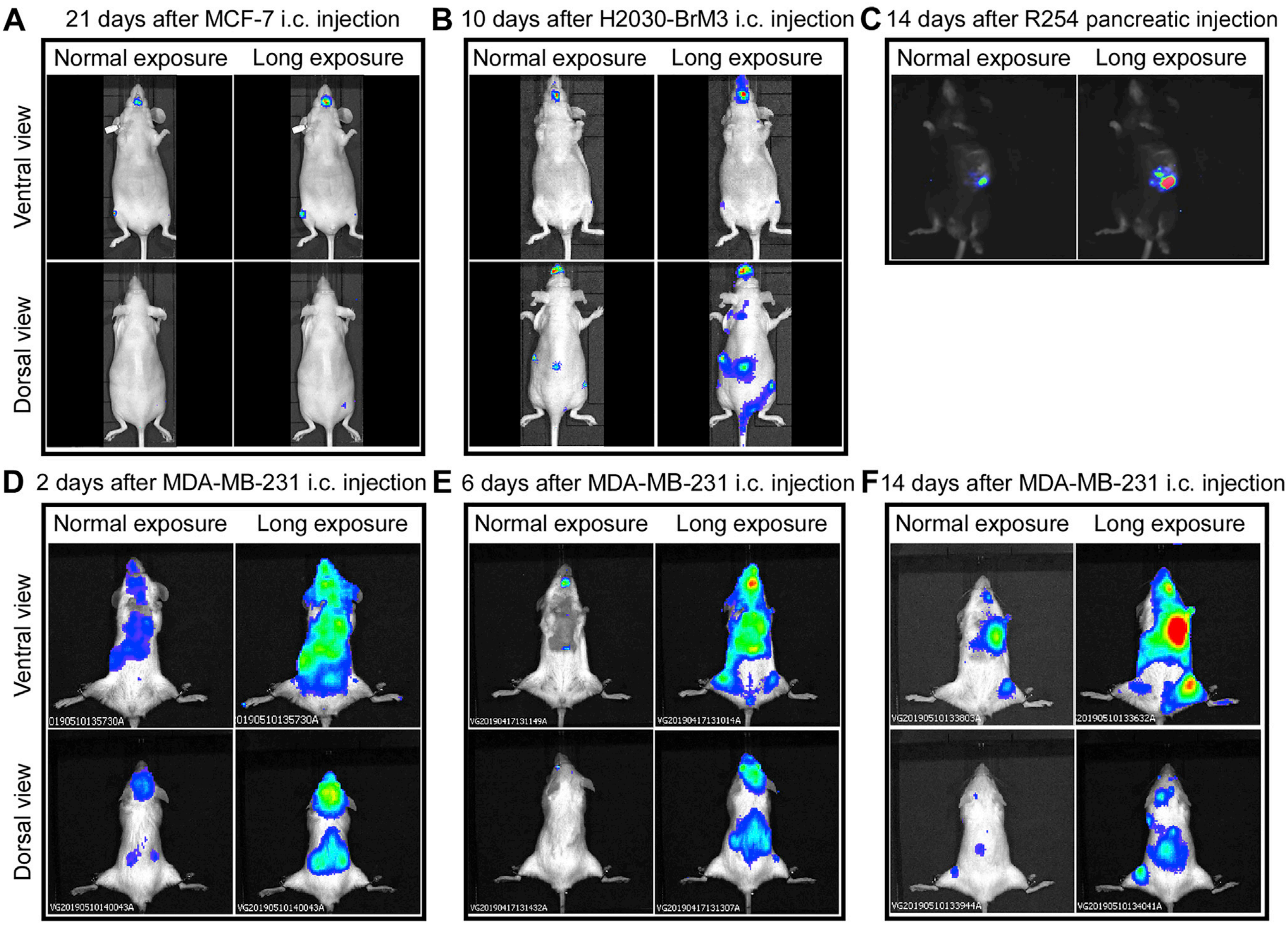

A
21 days after MCF-7 i.c. injection
B
10 days after H2030-BrM3 i.c. injection
C
14 days after R254 pancreatic injection
Normal exposure
Long exposure
Ventral view
Dorsal view
D
2 days after MDA-MB-231 i.c. injection
E
6 days after MDA-MB-231 i.c. injection
F
14 days after MDA-MB-231 i.c. injection

Figure S5. Bioluminescence Imaging of Different Cancer Models, Related to Figure 5

(A) Bioluminescence images with normal and long exposure from ventral and dorsal views of a mouse collected 21 days after intracardial (i.c.) injection of MCF-7 ER positive breast cancer cells.

(B) Bioluminescence images of a mouse collected 10 days after intracardial injection of H2030-BrM3 lung cancer cells.

(C) Bioluminescence images of a mouse collected 14 days after pancreatic injection of R254 cancer cells. Note that the C57BL/6 mouse line used in this model has black fur and therefore a different appearance in overlaid photographic/bioluminescence images compared to the other mouse strains.

(D-F) Bioluminescence images of the time-course experiments shown in Figure 5. The mice were intracardially injected with MDA-MB-231 breast cancer cells and sacrificed 2 days, 6 days and 14 days post injection, respectively.

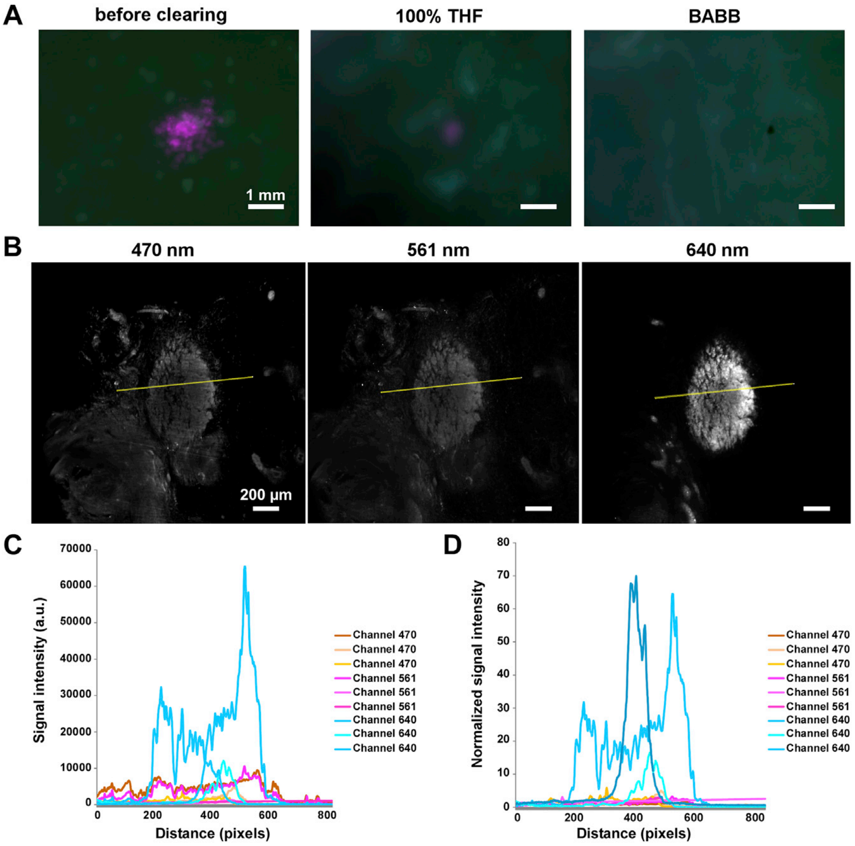


Figure S6. Elimination of Endogenously Expressed mCherry Signal from Tumors after vDISCO, Related to Figure 6

(A) Tumor metastases in lungs from a mouse transplanted with MDA-MB-231 cells in the mammary fat pad were imaged with a fluorescence stereomicroscope before and after vDISCO clearing, showing that the endogenously expressed mCherry signal was eliminated after the THF and BABB incubation steps.

(B) Light-sheet microscopy images of primary tumor with background fluorescence imaged in the green channel (ex: 470 nm, left), mCherry signal in the red channel (ex: 561 nm, middle), and the enhanced mCherry signal (Atto647N) in the far-red channel (ex: 640, right) after vDISCO clearing.

(C) Signal intensity profiles along the yellow lines in panel B were plotted: Channel 470 (orange), Channel 561 (magenta) and Channel 640 (cyan) (n = 3 mice).

(D) Normalized fluorescence signal profiles of the data in (C), showing that the endogenous mCherry signal was depleted to background levels after vDISCO clearing.

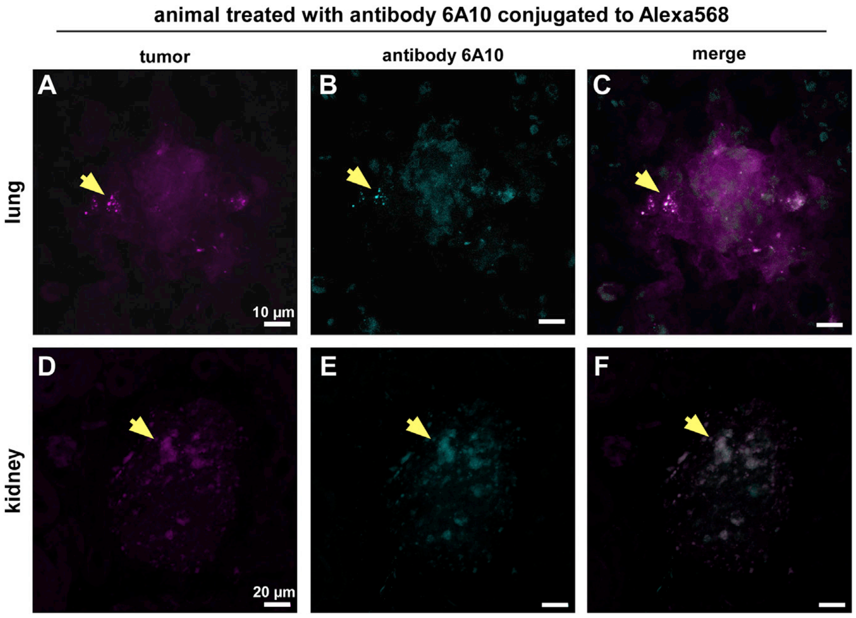


Figure S7. Verification of Antibody Targeting in Different Organs by Confocal Microscopy, Related to Figure 6

(A-F) Confocal images of metastases in the lung (A-C) and kidney (D-F) of a mouse transplanted with MDA-MB-231 cells (labeled with an anti-mCherry nanobody conjugated to Atto647N, magenta) and treated with therapeutic antibody 6A10 conjugated to Alexa568 (cyan). Examples of the colocalization of metastatic cells with the 6A10 antibody are indicated with yellow arrowheads (C and F).

A.2. Transfer learning from synthetic data reduces need for labels to segment brain vasculature and neural pathways in 3D

Authors: J Paetzold*, **O Schoppe***, R Al-Maskari, G Tetteh, V Efremov, M Todorov, R Cai, H Mai, Z Rong, A Ertürk, B Menze
**Joint first authorship*

Abstract: Novel microscopic techniques yield high-resolution volumetric scans of complex anatomical structures such as the blood vasculature or the nervous system. Here, we show how transfer learning and synthetic data generation can be used to train deep neural networks to segment these structures successfully in the absence of or with very limited training data.

Date of publication: 08.07.2019
Conference name: International Conference on Medical Imaging with Deep Learning, 8-10 July 2019, London, United Kingdom (peer-reviewed paper)

Individual contribution: project conception and coordination, experimental design, data analysis, leading author of manuscript

Transfer learning from synthetic data reduces need for labels to segment brain vasculature and neural pathways in 3D

Johannes C. Paetzold[*1,†], **Oliver Schoppe**[*1,†], **Rami Al-Maskari**[1], **Giles Tetteh**[1], **Velizar Efremov**[1], **Mihail I. Todorov**[2], **Ruiyao Cai**[2], **Hongcheng Mai**[2], **Zhouyi Rong**[2], **Ali Ertuerk**[2], **Bjoern H. Menze**[1]

[1] *TranslaTUM and Department of Computer Science, Technical University of Munich, Germany*

[2] *Institute for Stroke and Dementia Research, Ludwig Maximilian University of Munich, Germany*

[†] *Correspondence to johannes.paetzold@tum.de and oliver.schoppe@tum.de*

Abstract

Novel microscopic techniques yield high-resolution volumetric scans of complex anatomical structures such as the blood vasculature or the nervous system. Here, we show how transfer learning and synthetic data generation can be used to train deep neural networks to segment these structures successfully in the absence of or with very limited training data.

Keywords: Deep learning, transfer learning, synthetic data, vasculature, neural pathways.

1. Introduction

Recent advances in tissue-clearing (Ertürk et al., 2012; Chung and Deisseroth, 2013) combined with 3D light-sheet microscopy (*3D LSM*) overcome previous imaging limitations: they enable volumetric acquisition at cellular resolution of entire organisms (Cai et al., 2018; Pan et al., 2019; Stefaniuk et al., 2016; Mano et al., 2018). This yields unprecedented insight into the micro-anatomy at the macro-scale, e.g., to study highly connected structures like the brain vasculature or the peripheral nervous system. Differences in these structures have been associated with a wide range of disorders (Joutel et al., 2010; Li et al., 2010). Thus, segmentation and characterization of these anatomical structures is crucial to study causes and effects of such pathologies. However, manual segmentation of complex structures is very time-consuming, especially in high-resolution volumetric scans. While this motivates the need for deep learning it also implies a high cost of labeling. Here, we substantially reduce the need for manually labeled training data using transfer learning, an approach gaining attention (Van Opbroek et al., 2015; Khan et al., 2019). In short, we show that training deep networks on synthetic data is already sufficient to learn the basic underlying task across different anatomical structures, species, and imaging modalities.

2. Methods

Here, we present results from three widely different applications: human brain vessels (MRI), mouse brain vessels and the mouse peripheral nervous system (both *3D LSM*). The same network was trained either on a small labeled set from the respective application

[*] Joint first authors

(”real data”), on synthetically generated data, or on a combination of both. The synthetic data used is identical for all three applications. We chose DeepVesselNet as our architecture; the schedule for pre-training on synthetic data and refinement on real data match the methods of (Tetteh et al., 2018). The methods for generation of synthetic training data is described in (Schneider et al., 2012). MRI scans from human brain vasculature are taken from (Tetteh et al., 2018) (voxel size: 300μm x 300μm x 600μm). Volumetric scans of the brain vasculature (voxel size: (3μm)^3) and the peripheral nervous system (voxel size: (10μm)^3) were obtained using DISCO tissue clearing and fluorescent light-sheet microscopy as described in (Cai et al., 2018). Representative 2D cross-sections of the synthetic data and segmentations of all three applications are shown in Figure 1.

3. Results

Transfer learning from synthetic data (Table 1, Part 1). For segmenting the human vasculature from MRI scans, training the net on the synthetic data alone yields very good results, 81% in F1-score (note: the synthetic data set had been designed for this application). Training on the real data for this application yields a higher F1-score of 86%. The best result (87%), however, is achieved by a combination of both: pre-training on synthetic data and fine-tuning on real data. Interestingly, the network also converges about 50% faster in this case (data not shown). Motivated by this observation, we repeated this experiment for *3D LSM* scans of the mouse brain vasculature. Again, the same pattern can be observed and the combination of synthetic with real data (F1-score of 76%) outperforms synthetic data (71%) or real data alone (73%). Taking the approach yet further, we applied the approach to *3D LSM* full body scans of the peripheral nervous system of a mouse. While training on synthetic data alone was not very successful (16%) as compared to real data (49%), the gain from combining both was almost completely additive (64%).

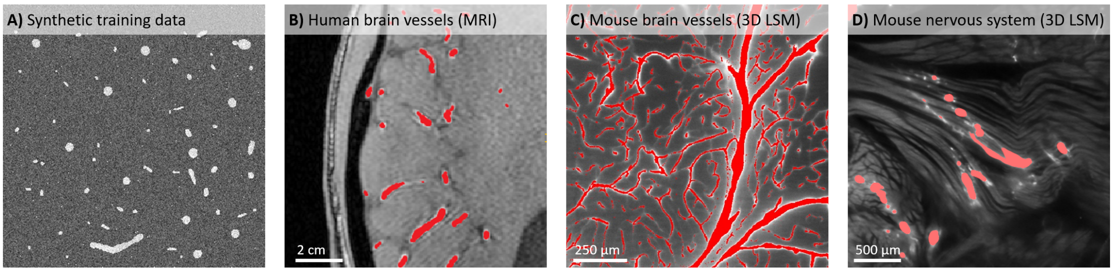
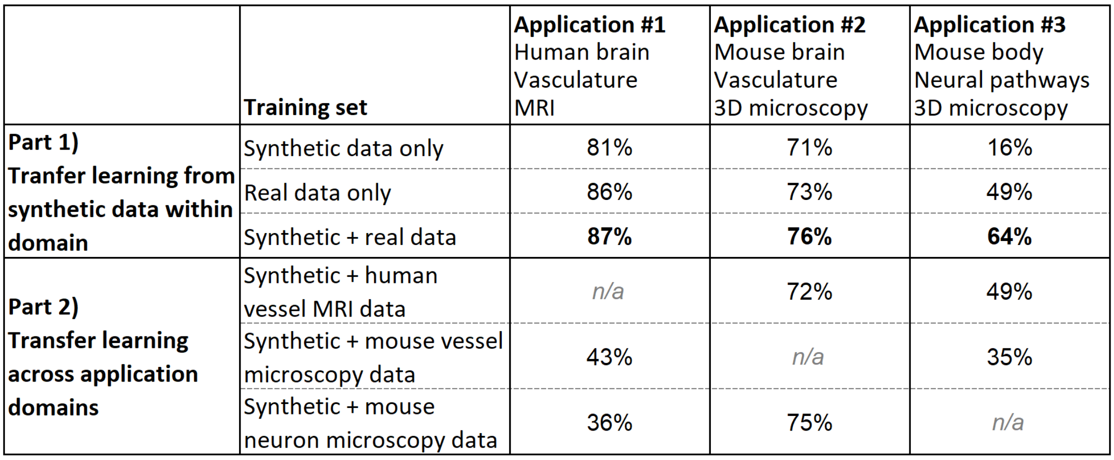

A) Synthetic training data
B) Human brain vessels (MRI)
2 cm
C) Mouse brain vessels (3D LSM)
250 µm
D) Mouse nervous system (3D LSM)
500 µm

Figure 1: A) Synthetic training data was designed to resemble vasculature of human brain in MRI scans. B-D) Predicted segmentations of 3 different applications: MRI scans of human brain vasculature (B), *3D LSM* of mouse brain vasculature (C), and peripheral nervous system (D; shown here: innervated muscle fibres)

Transfer learning across domains (Table 1, Part 2). Here, we trained the network on a combination of synthetic data and the real data from a given application and then predicted on data from another application. When predicting on human vasculatures from MRI scans, the refinement step on real data from another application after pre-training on synthetic data leads to worse results (left column: 43% and 36%) compared to training on synthetic data alone (81%, see Part 1). However, when training the model on synthetic data and real data of human vessels in MRI scans (first row of Part 2), the performance on *3D LSM* scans of mouse brain vessels (72%) or the mouse peripheral nervous system (49%) is about as good as when trained on the respective real data alone. Also, while the domain transfer from mouse vasculature to mouse nervous system only yields mediocre results (35%), it works well the other way around: refining a model trained on synthetic data with real data from the nervous system to segment brain vessels almost works as well (75%) as if it had been refined on data within the same domain (76%, see Part 1).



	Training set	**Application #1** Human brain Vasculature MRI	**Application #2** Mouse brain Vasculature 3D microscopy	**Application #3** Mouse body Neural pathways 3D microscopy
Part 1) **Tranfer learning from synthetic data within domain**	Synthetic data only	81%	71%	16%
	Real data only	86%	73%	49%
	Synthetic + real data	**87%**	**76%**	**64%**
Part 2) **Transfer learning across application domains**	Synthetic + human vessel MRI data	*n/a*	72%	49%
	Synthetic + mouse vessel microscopy data	43%	*n/a*	35%
	Synthetic + mouse neuron microscopy data	36%	75%	*n/a*

Table 1: Quality of predicted segmentations (F1-score) for 3 different applications

4. Discussion

Our results demonstrate how pre-training on synthetically generated data can accelerate model convergence and boost the overall segmentation performance. For a given desired performance, this thus means a reduced need for manually labeled training data, which is very expensive for complex structures in 3D scans. Importantly, a single synthetic data set that was originally designed to represent human vessels also works well for applications from different species, anatomical structures, and imaging modalities. This suggests that the features learned from the synthetic data are of general use for the abstract segmentation tasks, highlighting the generalizability of the approach. Thus, the expensively labeled data for a given application does not have to be used to learn a basic task but rather can be preserved for refining the pre-trained model to the specifics of the application (such as contrast, noise, background structures). Interestingly, this approach may also be of use in cases where no training data is available at all. For instance, we could show that a model trained on synthetic data and real data from another application can match the performance of a model trained from scratch on real data from the application of interest. Together, these results highlight the importance of transfer learning towards the goal of resolving a key bottleneck in adoption of deep learning: the high cost of data annotation.

Acknowledgments

This work was supported the German Federal Ministry of Education and Research via the Software Campus initiative (to O.S.), the Vascular Dementia Research Foundation, Synergy Excellence Cluster Munich (SyNergy), ERA-Net Neuron (01EW1501A to A.E.). Furthermore, NVIDIA supported this work with a Titan XP GPU via the GPU Grant Program.

References

Ruiyao Cai, Chenchen Pan, Alireza Ghasemigharagoz, Mihail Ivilinov Todorov, Benjamin Förstera, Shan Zhao, Harsharan S Bhatia, Arnaldo Parra-Damas, Leander Mrowka, Delphine Theodorou, et al. Panoptic imaging of transparent mice reveals whole-body neuronal projections and skull–meninges connections. Technical report, Nature Publishing Group, 2018.

Kwanghun Chung and Karl Deisseroth. Clarity for mapping the nervous system. *Nature methods*, 10(6):508, 2013.

Ali Ertürk, Klaus Becker, Nina Jährling, Christoph P Mauch, Caroline D Hojer, Jackson G Egen, Farida Hellal, Frank Bradke, Morgan Sheng, and Hans-Ulrich Dodt. Three-dimensional imaging of solvent-cleared organs using 3disco. *Nature protocols*, 7(11):1983, 2012.

Anne Joutel, Marie Monet-Leprêtre, Claudia Gosele, Céline Baron-Menguy, Annette Hammes, Sabine Schmidt, Barbara Lemaire-Carrette, Valérie Domenga, Andreas Schedl, Pierre Lacombe, et al. Cerebrovascular dysfunction and microcirculation rarefaction precede white matter lesions in a mouse genetic model of cerebral ischemic small vessel disease. *The Journal of clinical investigation*, 120(2):433–445, 2010.

SanaUllah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din, and Joel JP C Rodrigues. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, 2019.

Weiguo Li, Roshini Prakash, Aisha I Kelly-Cobbs, Safia Ogbi, Anna Kozak, Azza B El-Remessy, Derek A Schreihofer, Susan C Fagan, and Adviye Ergul. Adaptive cerebral neovascularization in a model of type 2 diabetes: relevance to focal cerebral ischemia. *Diabetes*, 59(1):228–235, 2010.

Tomoyuki Mano, Alexandre Albanese, Hans-Ulrich Dodt, Ali Erturk, Viviana Gradinaru, Jennifer B Treweek, Atsushi Miyawaki, Kwanghun Chung, and Hiroki R Ueda. Whole-brain analysis of cells and circuits by tissue clearing and light-sheet microscopy. *Journal of Neuroscience*, 38(44):9330–9337, 2018.

Chenchen Pan, Oliver Schoppe, Arnaldo Parra-Damas, Ruiyao Cai, Mihail Ivilinov Todorov, Gabor Gondi, Bettina von Neubeck, Alireza Ghasemi, Madita Alice Reimer, Javier Coronel, et al. Deep learning reveals cancer metastasis and therapeutic antibody targeting in whole body. *bioRxiv*, page 541862, 2019.

Matthias Schneider, Johannes Reichold, Bruno Weber, Gábor Székely, and Sven Hirsch. Tissue metabolism driven arterial tree generation. *Medical image analysis*, 16(7):1397–1414, 2012.

Marzena Stefaniuk, Emilio J Gualda, Monika Pawlowska, Diana Legutko, Paweł Matryba, Paulina Koza, Witold Konopka, Dorota Owczarek, Marcin Wawrzyniak, Pablo Loza-Alvarez, et al. Light-sheet microscopy imaging of a whole cleared rat brain with thy1-gfp transgene. *Scientific reports*, 6:28209, 2016.

Giles Tetteh, Velizar Efremov, Nils D Forkert, Matthias Schneider, Jan Kirschke, Bruno Weber, Claus Zimmer, Marie Piraud, and Bjoern H Menze. Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes. *arXiv preprint arXiv:1803.09340*, 2018.

Annegreet Van Opbroek, M Arfan Ikram, Meike W Vernooij, and Marleen De Bruijne. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE transactions on medical imaging*, 34(5):1018–1030, 2015.

A.3. Measuring the Performance of Neural Models

Authors: O Schoppe, N Harper, B Willmore, A King, J Schnupp

Abstract: Good metrics of the performance of a statistical or computational model are essential for model comparison and selection. Here, we address the design of performance metrics for models that aim to predict neural responses to sensory inputs. This is particularly difficult because the responses of sensory neurons are inherently variable, even in response to repeated presentations of identical stimuli. In this situation, standard metrics (such as the correlation coefficient) fail because they do not distinguish between explainable variance (the part of the neural response that is systematically dependent on the stimulus) and response variability (the part of the neural response that is not systematically dependent on the stimulus, and cannot be explained by modeling the stimulus-response relationship). As a result, models which perfectly describe the systematic stimulus-response relationship may appear to perform poorly. Two metrics have previously been proposed which account for this inherent variability: Signal Power Explained (*SPE*, Sahani and Linden, 2003), and the normalized correlation coefficient (CC_{norm}, Hsu et al., 2004). Here, we analyze these metrics, and show that they are intimately related. However, *SPE* has no lower bound, and we show that, even for good models, *SPE* can yield negative values that are difficult to interpret. CC_{norm} is better behaved in that it is effectively bounded between -1 and 1, and values below zero are very rare in practice and easy to interpret. However, it was hitherto not possible to calculate CC_{norm} directly; instead, it was estimated using imprecise and laborious resampling techniques. Here, we identify a new approach that can calculate CC_{norm} quickly and accurately. As a result, we argue that it is now a better choice of metric than *SPE* to accurately evaluate the performance of neural models.

Date of publication: 10.02.2016
Journal name: Frontiers in Computational Neuroscience

Individual contribution: initiated the project; developed methodology; wrote and tested code implementing methods; analyzed method performance both analytically and through experiment; lead author of paper

frontiers
in Computational Neuroscience

METHODS
published: 10 February 2016
doi: 10.3389/fncom.2016.00010

Measuring the Performance of Neural Models

Oliver Schoppe[1,2]*, *Nicol S. Harper*[1], *Ben D. B. Willmore*[1], *Andrew J. King*[1] and *Jan W. H. Schnupp*[1]*

[1] Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford, UK, [2] Bio-Inspired Information Processing, Technische Universität München, Garching, Germany

Good metrics of the performance of a statistical or computational model are essential for model comparison and selection. Here, we address the design of performance metrics for models that aim to predict neural responses to sensory inputs. This is particularly difficult because the responses of sensory neurons are inherently variable, even in response to repeated presentations of identical stimuli. In this situation, standard metrics (such as the correlation coefficient) fail because they do not distinguish between explainable variance (the part of the neural response that is systematically dependent on the stimulus) and response variability (the part of the neural response that is not systematically dependent on the stimulus, and cannot be explained by modeling the stimulus-response relationship). As a result, models which perfectly describe the systematic stimulus-response relationship may appear to perform poorly. Two metrics have previously been proposed which account for this inherent variability: Signal Power Explained (*SPE*, Sahani and Linden, 2003), and the normalized correlation coefficient (CC_{norm}, Hsu et al., 2004). Here, we analyze these metrics, and show that they are intimately related. However, *SPE* has no lower bound, and we show that, even for good models, *SPE* can yield negative values that are difficult to interpret. CC_{norm} is better behaved in that it is effectively bounded between −1 and 1, and values below zero are very rare in practice and easy to interpret. However, it was hitherto not possible to calculate CC_{norm} directly; instead, it was estimated using imprecise and laborious resampling techniques. Here, we identify a new approach that can calculate CC_{norm} quickly and accurately. As a result, we argue that it is now a better choice of metric than *SPE* to accurately evaluate the performance of neural models.

Keywords: sensory neuron, receptive field, signal power, model selection, statistical modeling, neural coding

OPEN ACCESS

Edited by:
Ad Aertsen,
University of Freiburg, Germany

Reviewed by:
Alexander G. Dimitrov,
Washington State University Vancouver, USA
James McFarland,
University of Maryland, College Park, USA
Jannis Hildebrandt,
Carl von Ossietzky University of Oldenburg, Germany

Correspondence:
Oliver Schoppe
oliver.schoppe@tum.de;
Jan W. H. Schnupp
jan.schnupp@dpag.ox.ac.uk

Received: *27 September 2015*
Accepted: *21 January 2016*
Published: *10 February 2016*

Citation:
Schoppe O, Harper NS, Willmore BDB, King AJ and Schnupp JWH (2016) Measuring the Performance of Neural Models. Front. Comput. Neurosci. 10:10. doi: 10.3389/fncom.2016.00010

1. INTRODUCTION

Evaluating the performance of quantitative models of neural information processing is an essential part of their development. Appropriate metrics enable us to compare different models and select those which best describe the data. Here, we are interested in developing improved metrics to assess models of the stimulus-response relationships of sensory neurons, in the challenging (but common) situation where the stimulus-response relationship is complex, and neuronal responses are highly variable. In this case, the development of appropriate performance metrics is not trivial, and so there is a lack of consensus about which metrics are to be used.

The classical way to record and model neural responses has been to repeatedly present an animal with a small, well-defined set of stimuli (such as sinusoidal gratings of different orientations, or sounds of different frequencies). The neural responses to repeated presentations of each stimulus are then averaged. Using a small stimulus set, it may be possible to present the same stimulus enough times that this averaging succeeds in reducing the effect of neuronal response variability (Döerrscheidt, 1981). It may then be possible to produce models which accurately describe the relationship between the stimulus and the averaged responses. These models can then be accurately evaluated by comparing the modeled and actual neuronal responses using standard metrics such as correlation coefficient. Under these circumstances, the correlation coefficient may be appropriate and can easily be interpreted—a poor model will have a correlation coefficient close to 0, a perfect model will have a correlation coefficient close to 1, and the square of the value of the correlation coefficient equals the proportion of the variance in the neural responses that the model is able to account for.

However, recent work in sensory neuroscience has increasingly focused on the responses of neurons to complex stimuli (Atencio and Schreiner, 2013; David and Shamma, 2013), and even natural stimuli (Prenger et al., 2004; Asari and Zador, 2009; Laudanski et al., 2012). For such stimuli, even very sparse sampling of the stimulus space may require the presentation of very large numbers of different stimuli (at least of order 2^d for d stimulus dimensions; also see Shimazaki and Shinomoto, 2007). This makes it difficult to repeatedly present stimuli enough times for response variability to simply average out. Estimating mean responses for a particular stimulus is thus subject to sampling noise, and in addition to that, the neuron under study may also be "intrinsically noisy" in the sense that only a small proportion of the response variability may be attributable to variability of the stimulus. Such situations are very common in sensory neuroscience, and they can render the use of correlation coefficients to evaluate the performance of models that map stimuli to responses very misleading. If only a fraction of the neural response variability is stimulus linked, then even a perfect model of that stimulus linkage will only ever be able to account for some fraction of the variance in the observed neural response data. This places a limit on the maximum correlation coefficient that can be achieved, and the interpretation of the raw correlation coefficients becomes ambiguous: for example, a relatively low correlation coefficient of 0.5 might be due to an excellent model of a noisy dataset, or of a rather poor model of a dataset with very low intrinsic and sampling noise, or something in between.

Different approaches for taking neural variability into account when measuring model performance have been developed. To get an unbiased estimate of *mutual information*, Panzeri and Treves (1996) suggested a method to extrapolate information content to an infinite number of trials (also see Atencio et al., 2012). Roddey et al. (2000) compared the coherence of pairs of neural responses to independent stimulus repetitions to derive a *minimum mean square error (MMSE)* estimator for an optimal model. The difference between the model prediction error and the MMSE of an optimal model allows the quantification of the model performance relative to the best possible performance given the neural variability.

Based not only on pairs, but even larger sets of neural responses to independent stimulus repetitions, Sahani and Linden developed the very insightful decomposition of the recorded signal into *signal power* and *noise power* (Sahani and Linden, 2003). This has lead to the *signal power explained (SPE)*, a measure based on *variance explained* which discounts "unexplainable" neural variability. While the work of Roddey et al. (2000) was already based on the differentiation between explainable and unexplainable neural response components, Sahani and Linden (2003) provided explicit estimations for those components. The *SPE* measure has been widely adopted, albeit under various names such as *predictive power, predicted response power, and relative prediction success* (Sahani and Linden, 2003; Machens et al., 2004; Ahrens et al., 2008; Asari and Zador, 2009; Rabinowitz et al., 2012). Also, it has been used as a basis for specific variants of measures for model performance (Haefner and Cumming, 2009).

Focusing on coherence and the correlation coefficient, Hsu and colleagues developed a method to normalize those measures by their upper bound (CC_{max}), which is given by the inter-trial variability (Hsu et al., 2004). This yields the *normalized correlation coefficient* (CC_{norm}). Following their suggestion, the upper bound can be approximated by looking at the similarity between one half of the trials and the other half of the trials (CC_{half}). This measure has also been used by Gill et al. (2006) and Touryan et al. (2005). Others used the absolute correlation coefficient and controlled for inter-trial variability by comparing the absolute values with CC_{half} (Laudanski et al., 2012).

The two metrics *SPE* and CC_{norm} have been developed independently, but they both attempt—in different ways—to provide a method for assessing model performance independent of neuronal response variability. Here, we here analyze these metrics, show for the first time that they are closely related, and discuss the shortcomings of each. We provide a new, efficient way to directly calculate CC_{norm} and show how it can be used to accurately assess model performance, overcoming previous shortcomings.

2. CRITERIA OF MODEL EVALUATION

Neural responses are often measured as the membrane potential (Machens et al., 2004; Asari and Zador, 2009) or as the time-varying firing rate (Sahani and Linden, 2003; Gill et al., 2006; Ahrens et al., 2008; Rabinowitz et al., 2011; Laudanski et al., 2012; Rabinowitz et al., 2012) (which we will use without loss of generality). Thus, a measure of performance for such models should quantify the similarity of the predicted firing rate $\hat{y}$ and the recorded firing rate y (also known as the peri-stimulus time histogram, PSTH):

$$y(t) = \frac{1}{N}\sum_{n=1}^{N} R_n(t) \tag{1}$$

where R_n is the recorded response of the nth stimulus presentation and N is the total number of stimulus presentations (trials). Both $R_n(t)$ and $y(t)$ are a function of the time bin t, but the argument t will not be shown for rest of the manuscript. Each value of the vector R_n contains the number of spikes that were recorded in the corresponding time bin. Note that, given the trial-to-trial variability of sensory responses, the recorded firing rate y is only an approximation of the true (but unknown) underlying firing rate function that is evoked by the presentation of a stimulus (also see Kass et al., 2003). It is a sample mean which one would expect to asymptote to the true mean as the number of trials increases ($N \to \infty$). As will be discussed in detail at a later point, the difference between the recorded firing rate y and the true underlying firing rate is considered to be noise under the assumption of rate coding. This is the unexplainable variance that reflects the variability of the neuronal response. As the number of trials increases, the difference between y and the true underlying firing rate decreases and so does the non-deterministic and thus unexplainable variance in the signal.

With the recorded firing rate y being the target variable for the prediction $\hat{y}$, a measure of model performance needs to quantify the similarity between both signals, i.e., the prediction accuracy. Note that model performance is not necessarily the same as prediction accuracy (see next section).

3. SIGNAL POWER EXPLAINED

Two somewhat related metrics which are widely applied in statistics are the "coefficient of determination" (*CD*) and the "proportion of variance explained" (*VE*). Both these metrics essentially incorporate the assumption that the quantitative observations under study—in our case the responses of a sensory neuron or neural system—are the sum of an essentially deterministic process which maps sensory stimulus parameters onto neural excitation, plus an additive, stochastic noise process which is independent of the recent stimulus history (Sahani and Linden, 2003). Consequently, if a model is highly successful at predicting the deterministic part, subtracting the predictions from the observations should leave only the noise part, but if its predictions are poor, the residuals left after subtracting predictions from observations will contain both noise and prediction error. Thus, smaller residuals are taken as a sign of better prediction. The *CD* is an index that quantifies the size of the residuals relative to the size of the original observation in a quite direct manner as a sum of squares, and subtracts that unaccounted for proportion from 100% to give an estimate of the proportion of the signal that is accounted for by the model. Thus

$$CD = 1 - \frac{\sum_t (y(t) - \hat{y}(t))^2}{\sum_t y(t)^2} \tag{2}$$

The *VE* quantifies prediction accuracy in a largely analogous manner, but instead of using the "raw" sum of squares of the observations and the residuals, it instead uses the respective sample variances, measured around their respective sample means:

$$VE = 1 - \frac{Var(y - \hat{y})}{Var(y)} \tag{3}$$

This makes the *VE* insensitive to whether the mean of the predicted responses closely corresponds to the mean of the observed responses over all t, which can sometimes be an advantage. Even small errors (biases) in the mean of the prediction can be penalized quite heavily by the *CD* measure as these will accumulate over every sample. The *VE* measure can be thought of as deeming such biases as unimportant, and focusing solely on how well the model predicts the trends in the responses as a function of t.

CD and *VE* have a long established history in statistics, but neither provide an unambiguous measure of model performance because large amounts of residual variance, and therefore low *VE* or *CD* values, could arise either if the model provides a poor approximation to underlying deterministic and predictable aspects of the process under study, or if the model captures the deterministic part of the process perfectly well, but large amounts of fundamentally unpredictable noise in the system nevertheless cause the amount of residual variance to be large. In other words, even a perfect model cannot make perfect predictions, because the neuronal response has a non-deterministic component. Even if the model was completely identical to the neuron in every aspect, it would nevertheless be unable to explain 100% of the variance in the neuronal responses because the PSTHs collected over two separate sets of stimulus presentations cannot be expected to be identical and the first set does not perfectly predict the second. Furthermore, since the number of trials N used to determine any one PSTH is often rather low for practical reasons, observed PSTHs are often somewhat rough, noisy estimators of the underlying neural response function (also see Döerrscheidt, 1981; Kass et al., 2003; Shimazaki and Shinomoto, 2007). A good measure of model performance for sensory neural systems should take these considerations into account and judge model performance relative to achievable, rather than total, prediction accuracy. Such considerations led Sahani and Linden (2003) to introduce metrics which split the variance in an observed PSTH, the *total power* (*TP*), into the *signal power* (*SP*), which depends deterministically on recent stimulus history, and the stochastic *noise power* (*NP*). Only the *SP* is explainable in principle by a model, and the *signal power explained* (*SPE*) thus aims to quantify model performance relative to the best achievable performance. *SPE* is defined as:

$$SPE = \frac{Var(y) - Var(y - \hat{y})}{SP} \tag{4}$$

$$SP = \frac{1}{N-1}\left(N \times Var(y) - TP\right)$$

$$TP = (N-1) \times \sum_{n=1}^{N} Var(R_n) \tag{5}$$

SPE is quantified as the ratio of the *explained* signal power relative to the *explainable* signal power[1]. The *explained* signal power is

[1] Please note that we do not use the notation of Sahani and Linden (2003). However, all definitions are identical. Sahani and Linden define the *power* P of a signal r as the "average squared deviation from the mean: $P(r) = \langle (r_t - \langle r_t \rangle)^2 \rangle$" where $\langle . \rangle$ denotes the mean over time. This is identical to the variance of the signal, which we use.

calculated by subtracting the variance of the residual (the error) from the total variance in the observed firing rate. The *explainable* signal power *SP* is calculated according to formulas developed in Sahani and Linden (2003) and reproduced below (Equation 13). Good models will yield small error variance and thus a large *SPE* - and vice versa. However, this measure lacks an important characteristic: it is not bounded. While a perfect model would yield an SPE of 100%, the measure has no lower bound and can go deeply into negative values when the variance of the error is bigger than the variance of the neural signal. This shortcoming of the *SPE* metric can be exposed by reformulating parts of the equation. First, observe that for two random variables *X* and *Y* the variance of their difference can be expressed as :

$$Var(Y - X) = Var(Y) + Var(X) - 2 \times Cov(X, Y) \qquad (6)$$

Applying this reformulation to Equation 5 reveals that:

$$SPE = \frac{Var(y) - Var(y - \hat{y})}{SP} = \frac{2 \times Cov(y, \hat{y}) - Var(\hat{y})}{SP} \qquad (7)$$

Consider a particularly bad model, which produces predictions that are no better than the output of a random number generator. The covariance between the predictions and the neural responses will then be close to zero, but the variance (i.e., the power of the predicted signal) of the predicted signal may nevertheless be large. The *SPE* for such a model would be a negative number equal to $-Var(\hat{y})/SP$. This is a counterintuitive property of the *SPE* metric: the "proportion of the signal power that is explained" by a list of random numbers should be zero, not negative. Also, two bad models that are equally unable to capture the trends of the signal they are trying to predict and thus have near zero covariance may nevertheless have widely different negative *SPE* values, but how negative their *SPE* values are may have little systematic relationship to how large their prediction errors are on average, which makes small or negative *SPE* values very difficult to interpret.

This can be illustrated with a simple hypothetical example. Imagine a visual cortex simple cell responding to a sinusoidal contrast grating stimulus with a sinusoidal modulation of its firing rate, so its observed response is a sine wave, let's say, of an amplitude of ±1 spikes/s around a mean firing rate of 10 spikes/s at a modulation rate of 1 Hz. Let us further assume that model A predicts sinusoidal firing at a 2 Hz modulation rate with an amplitude of ±2 spikes/s around a mean of 10 spikes/s, and model B predicts a sinusoidal firing at 2 Hz with an amplitude of ±1 spikes/s around a mean of 100 spikes/s. Since neither model A nor B correctly predicted the rate of the sinusoidal firing rate modulations, and because sine waves of different frequencies are orthogonal, both models will have covariance of zero with the observed data. Thus, they have a negative *SPE*, as the signal variance is greater than zero. And because model A predicted larger amplitude fluctuations than model B, and thus has greater variance, the *SPE* of model A will be more negative than that of model B, which one might be tempted to interpret to mean that model A performed worse. However, the discrepancy or prediction error between observed and predicted rates for model A will never be more than 3 spikes/s, while that of model B will never be less than 88 spikes/s, and the more negative *SPE* of model A contrasts sharply with the fact that model A produces a much smaller mean squared prediction error than model B. Furthermore *SPE* can yield negative values even when there is a reasonable amount of covariance between model and prediction, if the variance in the predicted signal is also sizable. This is illustrated in **Figure 1**. Not only is such a measure rather hard to interpret, but it can be misleading. Due to the missing lower bound the values can not only become negative, but the exact value also depends on the variance of the prediction. Consider the prediction with 60% noise in the lower right panel of **Figure 1**. While this prediction is surely not a good one, the fact that data and model prediction co-vary to a fair degree is nevertheless readily apparent, and it would be hard to argue that a model predicting a flat, arbitrary, constant firing rate (say 800 spikes/s) would be a better alternative. Yet the variance of any predicted constant firing rate would be zero and so would be their *SPE*, which may seem indicative of a "better explanatory power" of the constant rate model compared to the "60% noise" added model of **Figure 1** with its $SPE = -39\%$, but the noisy model clearly captures some of the major peaks in the data while constant rate models don't even try.

These examples illustrate that models can be thought of as being wrong in different ways. They can be "biased," predicting an incorrect overall mean response rate, they can be "scaled wrong," predicting fluctuations that are too small or too large, or they can fail to predict the trends and dynamics of the data, leading to near zero covariance between observation and prediction. Different metrics of model performance will differ in how sensitive they are to these different types of error. *SPE* is sensitive both to poor scaling and poor covariance, but not to bias. Some might argue, quite reasonably, that this combined sensitivity to two types of error is a virtue: When *SPE* values are large then we can be confident that the model achieves both good covariance and good scaling. However, the downside of this joint sensitivity is that small or negative *SPE* values have limited diagnostic value because they could be due to small covariance or to overestimated (but not underestimated) predicted variance, or some combination of the two. Consequently, as we will illustrate further in section 6, *SPE* values below about 0.4 become very difficult to interpret, and may be much at odds with other commonly used measures of model performance.

Negative values of the *SPE* have been previously reported (Machens et al., 2004; Ahrens et al., 2008) and have been interpreted as a sign of overfitting of the model. Overfitting usually manifests itself as a decline in covariance between data and predictions in cross-validation tests, and as such would result in small or negative *SPEs*, but because *SPE* will become negative for any prediction which has a residual variance that is larger than the variance of the target signal, negative *SPE* is not a specific diagnostic of overfitting. Also negative *SPEs* do not necessarily imply that a model performs worse than a "null model" which predicts constant responses equal to the mean firing rate. In fact,

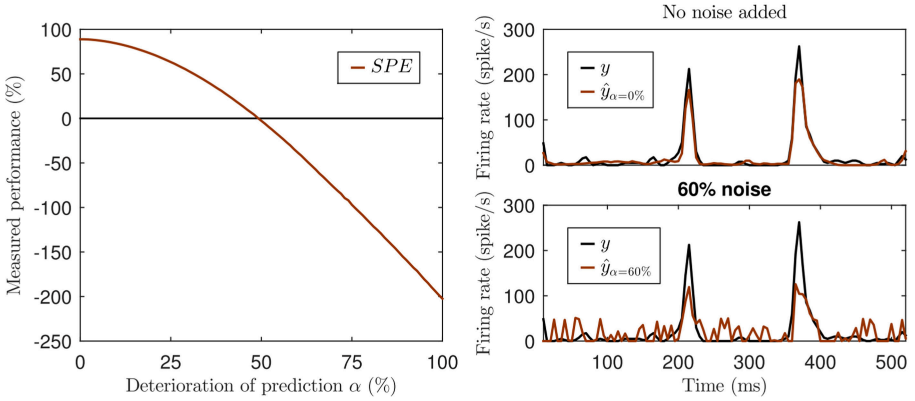


FIGURE 1 | Illustration of the missing lower bound of SPE. Left panel: The simulation was created by adding increasing white noise (w) to an actual prediction $\hat{y}$ generated by an artificial neural network: $\hat{y}_\alpha = \alpha w + (1-\alpha)\hat{y}$ with $0\% \leq \alpha \leq 100\%$, negative values of the deteriorated $\hat{y}_\alpha$ set to 0. Top right panel: The original prediction $\hat{y}$ of the neural network (red) and the actual neural response (black). Lower right panel: The deteriorated prediction at a noise level of 60% ($SPE = -39\%$).

any model predicting any arbitrary constant value (even a "dead neuron model" predicting a constant firing rate of 0 spikes/s) will have an *SPE* of zero and might on that basis be judged to perform better than other models generating noisy but fairly reasonable predictions (see **Figure 1**).

Of the three different types of error just discussed, large bias, poor scaling, small covariance, *SPE* is sensitive to two, covariance and scaling, although it is particularly excessively large, but not excessively small, scaling, that will drive *SPE* values down. Perhaps it is inevitable that single performance measures which are sensitive to multiple different types of error become very difficult to interpret as soon as performance becomes suboptimal. To an extent, whether one deems it preferable to have an error metric that is sensitive to bias, scaling and low covariance all at once, or whether one chooses a metric that is more specific in its sensitivity to only one of type of error is a matter of personal preference as well as of what one is hoping to achieve, but joint sensitivity to multiple different types of error is certainly problematic when the measure is to be used for model comparison, given that the relative weighting of the different types of error in the metric may not be readily apparent and it is unlikely to reflect how problematic the different types of error are in modeling. A constant bias, which would, for example, be heavily penalized by the *CD* metric discussed at the beginning of this section, can be easily fixed by adding or subtracting a constant value from the predictions. Similarly, scaling errors can be easily fixed by multiplication by a scalar. These two types of error pertain only to the relatively uninteresting stationary statistical properties of the data. They are in some sense trivial, and easily remedied through a simple linear adjustment. Low covariance, in contrast, is indicative of a much more profound inability of the model to capture the nature or dynamics of the neural stimulus-response relationships. In our opinion, the assessment of model performance should therefore rely first and foremost measures which are highly sensitive to poor covariance and insensitive to bias or scaling, and we discuss measures which have these properties in the next section. If needed, these could then be supplemented with additional metrics that can diagnose biases or scaling errors.

4. ABSOLUTE AND NORMALIZED CORRELATION COEFFICIENT

Another measure widely used in statistics, Pearson's product-moment correlation coefficient can also be used to assess the similarity of two time-varying signals. The correlation coefficient quantifies the linear correlation and maps it to a value between −1 and +1. To distinguish it from a normalized variant that will be used later in this section, the (absolute) correlation coefficient will from now on be abbreviated as CC_{abs}. It is defined as:

$$CC_{abs} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \tag{8}$$

CC_{abs} satisfies many of the criteria that one might desire in a good measure of model performance. It quantifies the similarity between observation and prediction, it is bounded between −1 and +1, and it can be interpreted easily and unambiguously. The normalization by the square root of the variances makes CC_{abs} insensitive to scaling errors, and the formulae for $Var()$ and $Cov()$ have subtractions of means built in that make CC_{abs} insensitive to bias, so that only the ability of Y to follow trends X is being quantified. However, like *VE*, it does not isolate model performance from prediction accuracy, which is inevitably limited by neural variability. In other words CC_{abs} might be small either because the model predictions Y are poor, or because the measured neural responses X are fundamentally so noisy that even an excellent model cannot be expected to achieve a large CC_{abs}. This was also noted by Hsu and colleagues who went on to develop an approach to quantify and account for the inherent noise in neural data (Hsu et al., 2004). Specifically, they introduced a method for normalizing coherence and correlation to the neural variability, which has later been applied as a performance measure (Touryan et al., 2005; Gill et al., 2006). Hsu and colleagues define the normalized correlation coefficient as follows (Hsu et al., 2004)[2]:

$$
\begin{aligned}
CC_{norm} &= \frac{CC_{abs}}{CC_{max}} \text{ with} \\
CC_{max} &= \sqrt{\frac{2}{1+\sqrt{\frac{1}{CC_{half}^2}}}} \overset{CC_{half}>0}{=} \sqrt{\frac{2}{1+\frac{1}{CC_{half}}}}
\end{aligned} \tag{9}
$$

Where CC_{max} is the maximum correlation coefficient between the recorded firing rate y and the best prediction $\hat{y}$ that a perfect model could theoretically achieve. More specifically, CC_{max} is the correlation coefficient between the recorded firing rate y (which is based on N trials) and the true (but unknown) underlying firing rate function, which could only be determined precisely if the system was completely stationary and an infinite number of trials could be conducted ($N \rightarrow \infty$). Even though the true underlying firing rate function can therefore usually not be determined with high accuracy through experiments, useful estimates of CC_{max} can nevertheless be calculated using the formulae in Equation 9. Following the methods of Hsu et al. (2004), CC_{half} is determined by splitting the data set into halves, and calculating the correlation coefficient between the PSTH constructed from the first half and the PSTH constructed from the second half of the trials. This approach determines CC_{max} by effectively extrapolating from N trials to the value that would be expected for $N \rightarrow \infty$.

Note that there are $\frac{1}{2}\binom{N}{N/2}$ different ways to choose $N/2$ out of N trials, and each such split of the data will yield a slightly different value for CC_{half}. Thus, in theory, the best estimate would average over all possible values of CC_{half} calculated for each possible split. In practice, this resampling technique can be computationally expensive, given the fact that there are already 92, 378 combinations for $N = 20$ trials. Averaging over a smaller number of randomly chosen splits may often be sufficient, but this yields an imprecise estimation of CC_{max}.

In summary, CC_{norm} provides a feasible method for capturing model performance independently of noise in the neural responses. It gives values bounded between -1 and +1 (in practice, they are bounded between 0 and +1, as model predictions are either correlated or not correlated, but typically not anti-correlated to the firing rate). Furthermore, the measure lends itself to unambiguous interpretation, and its limitations are well-known. Finally, it is normalized so that its value does not depend on the variability of a particular data set. Thus, the normalized correlation coefficient CC_{norm} fulfills the criteria for a useful measure of model performance, but its current definition is based in a laborious and potentially imprecise resampling technique.

5. A CONSOLIDATED APPROACH TO QUANTIFYING NEURAL VARIABILITY

As will have become clear in the previous sections, the two measures SPE and CC_{norm} follow the same logic in that both measure prediction accuracy and normalize it by a quantification of the inherent reproducibility of the neural responses that are to be modeled (SP or CC_{max}, respectively). In this section we will show that these two approaches of normalization not only follow the same logic, but are mathematically largely equivalent. This provides a deeper insight into the underlying concept and gives rise to a more elegant and efficient technique to normalize the correlation coefficient.

Following the methods of Sahani and Linden (2003)[3], the signal power SP (i.e., the deterministic part of the recorded firing rate y) can be expressed as:

$$SP = \frac{1}{N-1}\left(N \times Var(y) - TP\right) \tag{10}$$

$$= \frac{1}{N-1}\left(N \times Var\left(\frac{1}{N}\sum_{n=1}^{N} R_n\right) - \frac{1}{N}\sum_{n=1}^{N} Var(R_n)\right) \tag{11}$$

$$= \frac{1}{N-1}\left(N \times \frac{1}{N^2} Var\left(\sum_{n=1}^{N} R_n\right) - \frac{1}{N}\sum_{n=1}^{N} Var(R_n)\right) \tag{12}$$

$$= \frac{1}{N-1}\left(\frac{1}{N} \times Var\left(\sum_{n=1}^{N} R_n\right) - \frac{1}{N}\sum_{n=1}^{N} Var(R_n)\right) \tag{13}$$

Where TP is the total power (i.e., the average variance of a single trial) and R_n is the recorded neural response of the nth trial. Since the normalization factor of SPE is the inverse of SP it will be convenient to express it as:

$$\frac{1}{SP} = \frac{N(N-1)}{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)} \tag{14}$$

Furthermore, using Equation 14 the ratio of the noise power NP over SP can be expressed as:

$$\frac{NP}{SP} = \frac{TP - SP}{SP} = \frac{TP}{SP} - 1 = \frac{(N-1) \times \sum_{n=1}^{N} Var(R_n)}{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)} - 1 \tag{15}$$

For CC_{norm} the normalization factor is the inverse of CC_{max} and, following the methods of Hsu et al. (2004), it is currently determined with an indirect resampling method using Equation

[2] The expression for CC_{max} can be derived from the work of Hsu et al. (2004) in two steps. First, Equations 6 and 8 from Hsu et al. (2004) are combined and solved for $\gamma_{A\bar{R}_M}$. Second, the analogy of the coherence γ^2 and the squared correlation coefficient CC^2 allows to replace $\gamma_{A\bar{R}_M}$ with CC_{max} and $\gamma_{\bar{R}_{1,M/2}\bar{R}_{2,M/2}}$ with CC_{half}. In the notation of Hsu and colleagues $\gamma^2_{A\bar{R}_M}$ denotes the coherence of the mean response over M trials with the true (but unknown) underlying firing rate A, i.e., the maximum achievable coherence of a perfect model.

[3] Again, please note that Sahani and Linden (2003) use $\overline{r^{(n)}}$ to denote the average over trials. In order to facilitate the reformulation of the equation we do not use this abbreviated notation. Despite this difference in notation, this definition of SP is identical to the definition provided by Sahani and Linden (Equations 1 on Page 3).

9. We will now show how CC_{max} can be computed directly by exploiting the relation between SPE and CC_{norm}.

The coherence γ^2_{AB} between a source signal A and a noisy recording B of this signal can be related to the signal-to-noise ratio, i.e., the coherence is just a function of the noise process itself (see Marmarelis, 1978 for details). In the context of neural recordings, Hsu et al. (2004) used this relation to express the coherence of the true (but unknown) underlying firing rate function (the source A) to the observed PSTH (the noisy recording B) as a function of the signal-to-noise ratio of the recording. They quantified this in terms of signal power of the frequency domain signals, but since the power of corresponding time and frequency domain signals is identical, we can rewrite their expression (see formulas 5 and 6 of Hsu et al., 2004) directly in terms of NP and SP to get:

$$\gamma^2_{AB} = \frac{SP}{SP + \frac{1}{N}NP} \tag{16}$$

The derivation of the coherence function between the true underlying firing rate function and the observed neural response is analogous for the squared correlation coefficient between both signals (also see Hsu et al., 2004 for details on this analogy). Thus, we can apply the same principle to express the the inverse of CC_{max} as:

$$\frac{1}{CC_{max}} = \sqrt{1 + \frac{1}{N} \times \frac{NP}{SP}} \tag{17}$$

Combining Equation 17 with Equation 15 now allows us to express the inverse of CC_{max} as:

$$\frac{1}{CC_{max}} = \sqrt{1 + \frac{1}{N}\left(\frac{(N-1) \times \sum_{n=1}^{N} Var(R_n)}{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)} - 1\right)} \tag{18}$$

$$= \sqrt{1 - \frac{1}{N} + \frac{(1-\frac{1}{N}) \times \sum_{n=1}^{N} Var(R_n)}{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)}} \tag{19}$$

Based on Equation 8 and 9 the normalized correlation coefficient CC_{norm} between the recorded firing rate y and the model prediction $\hat{y}$ can now be expressed as:

$$CC_{norm} = \frac{CC_{abs}}{CC_{max}} = \frac{Cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}} \frac{1}{CC_{max}} \tag{20}$$

$$= \frac{Cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}} \sqrt{1 - \frac{1}{N} + \frac{(1-\frac{1}{N}) \times \sum_{n=1}^{N} Var(R_n)}{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)}} \tag{21}$$

$$= \frac{Cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}} \sqrt{1 - \frac{1}{N}} \sqrt{1 + \frac{\sum_{n=1}^{N} Var(R_n)}{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)}} \tag{22}$$

$$= \frac{Cov(y,\hat{y})}{\sqrt{Var(\hat{y})}} \frac{\sqrt{1-\frac{1}{N}}}{\sqrt{\frac{1}{N^2} Var\left(\sum_{n=1}^{N} R_n\right)}} \sqrt{1 + \frac{\sum_{n=1}^{N} Var(R_n)}{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)}} \tag{23}$$

$$= \frac{Cov(y,\hat{y})}{\sqrt{Var(\hat{y})}} \sqrt{\frac{N(N-1)}{Var\left(\sum_{n=1}^{N} R_n\right)}} \sqrt{1 + \frac{\sum_{n=1}^{N} Var(R_n)}{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)}} \tag{24}$$

$$= \frac{Cov(y,\hat{y})}{\sqrt{Var(\hat{y})}} \sqrt{N(N-1)} \sqrt{\frac{1}{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)}} \tag{25}$$

$$= \frac{Cov(y,\hat{y})}{\sqrt{Var(\hat{y})}} \sqrt{\frac{N(N-1)}{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)}} \tag{26}$$

$$= \frac{Cov(y,\hat{y})}{\sqrt{Var(\hat{y})}} \sqrt{\frac{1}{SP}} \tag{27}$$

In other words, we can now express CC_{norm} as a simple function of SP. The previous derivation also shows that both methods, SPE and CC_{norm}, use the covariance to quantify the prediction accuracy and take the neural variability into account by normalizing with the signal power *SP*. This has several implications. First, *SPE* will not reveal more about the prediction accuracy than CC_{norm}, because *SPE* and CC_{norm} quantify the similarity of the prediction and the neural response solely based on the covariance of both signals. It is well known that the (normalized) correlation coefficient is based on covariance, but it has hitherto not been made explicit that this is also the case for *SPE*. Note that *SPE* uses only the covariance to assess prediction accuracy and thus, cannot reveal more information about the similarity of both signals than CC_{norm}. Second, how both measures quantify neural variability is not only related, but mathematically equivalent. Third, in order to calculate CC_{norm} it is not necessary to laboriously compute an approximation to CC_{max} from repeated subsampling of the data to generate computationally inefficient and potentially imprecise estimates of CC_{half}. Instead, the normalization factor can be explicitly calculated with Equation 27, using Equation 13 for *SP* as suggested by Sahani and Linden (2003). The close relationship between both measures can also be visualized by squaring CC_{norm} (left panel of **Figure 2**).

In summary, CC_{norm} as defined in Equation 27 provides an insightful measure of model performance. It quantifies the prediction accuracy using the covariance and isolates model performance by taking the amount of intrinsic variability in the observed neural responses into account. It is in theory bounded between -1 and 1, and in practice values below zero are very rarely observed. If they do occur, their interpretation is unambiguous: negative CC_{norm} implies anticorrelation between prediction and data. CC_{norm} thus behaves uniformly well whether called upon to quantify the performance of good and of poor models, in contrast to *SPE* which behaves well, and very similarly to CC_{norm}, for good models, but becomes increasingly harder to interpret as model performance declines.

6. EXPERIMENTAL VALIDATION

The previous sections show the problems caused by the missing lower bound of *SPE* from a theoretical point of view and illustrate them with a simulation (**Figure 1**). This section demonstrates the implications from a practical point of view by comparing the predictive performance of models for the activity of single neurons in the auditory system in three different experimental settings.

6.1. Neural Recordings

All animal procedures were approved by the local ethical review committee and performed under license from the UK Home Office. Ten adult pigmented ferrets (seven female, three male; all >6 months of age) underwent electrophysiological recordings under anesthesia. Full details are as in the study by Bizley et al. (2010). Briefly, we induced general anesthesia with a single intramuscular dose of medetomidine (0.022 mg/kg/h) and ketamine (5 mg/kg/h), which was maintained with a continuous intravenous infusion of medetomidine and ketamine in saline. Oxygen was supplemented with a ventilator, and we monitored vital signs (body temperature, end-tidal CO2, and the electrocardiogram) throughout the experiment. The temporal muscles were retracted, a head holder was secured to the skull surface, and a craniotomy and a durotomy were made over the auditory cortex. We made extracellular recordings from neurons in primary auditory cortex (A1) and the anterior auditory field (AAF) using silicon probe electrodes (Neuronexus Technologies) with 16 or 32 sites (spaced at 50 or 150 μm) on probes with one, two, or four shanks (spaced at 200 μm). We clustered spikes off-line using klustakwik (Kadir et al., 2014); for subsequent manual sorting, we used either spikemonger (an in-house package) or klustaviewa (Kadir et al., 2014). The time-discrete neuronal firing rate was approximated by binning spikes in 5 ms windows and averaging the spike count in each bin over all trials (compare to Equation 1).

6.2. Acoustic Stimuli

Natural sounds were presented via Panasonic RPHV27 earphones, which were coupled to otoscope specula that were inserted into each ear canal, and driven by Tucker-Davis Technologies System III hardware (48 kHz sample rate). The sounds had root mean square intensities in the range of 75–82 dB SPL. For Experiment 1, we presented 20 sound clips of 5 s duration each, separated by 0.25 s of silence. Sound clips consisted of animal vocalizations (ferrets and birds), environmental sounds (water and wind) and speech. The presentation of these stimuli was repeated in 20 trials. For Experiments 2 and 3, we presented 45 sound clips of 1 s duration, again separated by gaps of silence. The sound clips consisted of animal vocalizations (sheep and birds), environmental sounds (water and wind) and speech. The presentation of these stimuli was repeated in 10 trials. The silent gaps and the first 0.25 s thereafter have been removed from the data set.

6.3. Neuronal Modeling

For Experiment 1, the responses of 119 single neurons were predicted with an LN model, a widely used class of models comprising a linear and a nonlinear stage (Chichilnisky, 2001; Simoncelli et al., 2004). The linear stage fits a spectro-temporal receptive field (STRF), which is a linear filter that links the neuronal response to the stimulus intensities of 31 log-spaced frequency channels (with center frequencies ranging from 1 to 32 kHz) along the preceding 20 time bins (covering a total of 100 ms stimulus history). The linear stage was fitted using GLMnet for Matlab (Qian et al.; see http://web.stanford.edu/~hastie/glmnet_matlab/). The nonlinear stage fits a sigmoidal nonlinearity to further maximize the goodness of fit to the neural response using minFunc by Mark Schmidt (University of British Columbia, British Columbia, Canada; http://www.di.ens.fr/~mschmidt/Software/minFunc.html). For Experiment 2, the same model class was used to predict the response of 77 single neurons. For Experiment 3, the responses of 43 single neurons were model with a standard neural network comprising 620 units in the input layer (31 frequency channels times 20 time bins of stimulus history), 20 hidden units and a single output unit. Hidden units and the output unit comprised a fixed sigmoidal nonlinearity. The connection weights of the

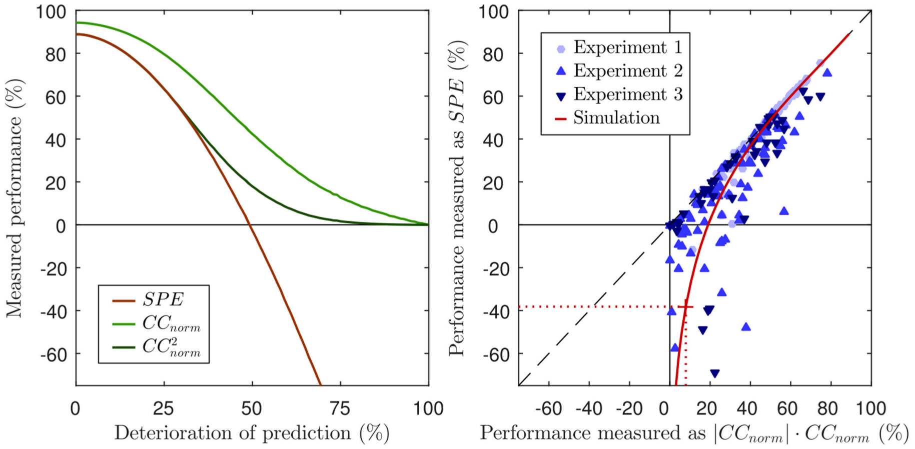


FIGURE 2 | Left panel: Same figure as the left panel of Figure 1, but including CC_{norm} and its squared values for reference. Note that, for good predictions (values above ca 50%), CC^2_{norm} and *SPE* are almost identical, both yielding very similar estimates of the proportion of "explainable variance explained." This is as might be expected given the equality of Equation 7 and 28 when $\hat{y} \rightarrow y$. However, as the prediction performance declines below 50%, $CC^2_{norm}$and *SPE* increasingly and sharply diverge. Right panel: Scatter plot of performance scores for predictions of neuronal responses from three different experiments. Each marker reflects the performance score of the prediction of the response of a single neuron. The black dashed line visualizes where *SPE* equals CC^2_{norm}. The values of CC_{norm} have been multiplied with their absolute value to demonstrate that negative values only occur for *SPE*, but not for CC_{norm}. The solid red line shows the values for the simulation of **Figure 1**, the dotted red line and the red cross mark the performance scores of the 60% noise simulation of the lower right panel in **Figure 1**. Corresponding to the overlap of *SPE* and CC_{norm} for good predictions, the red line approaches the dashed black line.

network were fitted with backpropagation using the Sum-of-Functions Optimizer (Sohl-Dickstein et al., 2013). Both, the STRF weights of the LN models and the connection weights of the neural networks were regularized with a penalty term on the L2-norm in order to avoid overfitting. In all cases, models were trained and tested using a cross-validation procedure. All free model parameters were fitted on a training set comprising 90% of all data. The predictive performance of a model for a given neuron was assessed by measuring *SPE* and CC_{norm} for the model predictions of the neural response to the remaining 10% of the data set. This procedure was repeated 10 times, each time with a distinct 10% of data. The model performance was computed as the mean across all 10 performance measurements.

6.4. Results

We predicted neuronal responses to acoustic stimuli with different model classes in order to address the question how the choice of a performance measure affects the interpretability of the results in a practical setting. To this end, we measured the predictive performance of models with two different methods, *SPE* and CC_{norm}. The right panel of **Figure 2** shows a scatter plot in which each marker indicates the performance scores that the respective measures assign to a given prediction for a given neuron. Instead of raw CC_{norm} values, here we chose to plot the signed square of CC_{norm} as a percentage on the x-axis. This choice is motivated by the fact that the square of the correlation coefficient, also known as the coefficient of determination, quantifies the "proportion of variance explained" by a statistical regression model, and $CC^2_{norm} \times 100$ should thus be interpretable directly as a measure of "percent explainable variance explained" by the model. We plot the signed square to ensure that there are no artificial constraints keeping the *x*-values positive: the fact that there x-range of the data is entirely positive while the y-range extends well into negative territory veridically reflects the way the respective underlying metrics, CC_{norm} and *SPE*, behave in practice. For those cases in which the model predicts the actual neuronal response quite well, one can observe a very tight relation between the SPE value and the signed squared value of CC_{norm}, i.e., both provide very similar, sensible measures of "percent explainable variance explained." However, as expected from the theoretical analysis of both measures in the previous sections, this relation diminishes for cases in which the models poorly predicted the neuronal response. For those cases where there is little or no correspondence between the prediction and the response, the value of CC_{norm} approaches zero (by definition), and for some of those cases, the value of *SPE* also approaches zero, but for many others the *SPE* value becomes a large negative number. Substantially negative *SPEs* are seen even for some cases for which the $|CC_{norm}| \times CC_{norm}$ indicates that the model was able to capture as much as 20–30% of the explainable, stimulus driven variability in the neural firing rate. Thirty percent variance explained may not be a stellar performance for a model, but it certainly does not seem deserving of a negative test score. Indeed, the experimental results are generally in accordance with the simulation in general, shown as a red line in the right panel of **Figure 2**. The simulation is identical to the one in **Figure 1**. To simulate *SPE* and CC_{norm} for a wide range of good and bad predictions, a good prediction was deteriorated by adding an increasing amount of white noise. Just as for the data from the three experiments, *SPE* values match the square of CC_{norm} for good predictions, but go deep into negative values for noisy predictions. For comparison, the *SPE* and CC_{norm} values of the example in the bottom right panel of **Figure 1** (60%

noise added) are marked with dotted lines in the right panel of **Figure 2**. In summary, the analysis of the experimental data from three experiments validate the theoretical analysis of the previous sections.

Figure 2 also visualizes the practical implications of the missing lower bound of *SPE*. *SPE* was from its inception described to be a "quantitative estimate of the fraction of stimulus-related response power captured by a given class of models" (Sahani and Linden, 2003). This interpretation is in conflict with values below zero because a fraction of a signal power cannot be negative. Furthermore, as was discussed in the previous sections, it is even difficult to assign an unambiguous interpretation to small or negative *SPE* values because a variety of poor models which vary widely in the size of their residual error can have similar small or negative *SPE*s, and may have *SPE*s below those of constant mean firing rate models of arbitrary value with an *SPE* of zero (including the "dead neuron model"), even if their residual error is smaller than that of these null models. If researchers are trying to quantify how well a particular class of models can describe the response properties of a sizeable sample population of neurons, a small number of somewhat spurious very negative values can heavily affect the overall population mean. For instance, the mean *SPE* value across the population of 77 neurons in Experiment 2 is just 15%, because a few very negative values drag down the average. But, as we have discussed in section 6, much of the negativity in those *SPE* values simply reflects a large variance in the predictions, which on its own is not very relevant, and constraining the *SPE* to values of zero or above would raise the mean performance by more than a quarter to over 19%.

7. CONCLUSION

Inter-trial variability of neural responses to repeated presentations of stimuli poses a problem for measuring the performance of predictive models. The neural variability inherently limits how similar one can expect the prediction of even a perfect model to be to the observed responses. Thus, when using prediction accuracy as a measure of performance, inherent response variability is a confound, and the need to control for this has been widely acknowledged (e.g., Panzeri and Treves, 1996; Sahani and Linden, 2003; Hsu et al., 2004; David and Gallant, 2005; Laudanski et al., 2012).

Different approaches for taking neural variability into account when measuring model performance have been developed. To get an unbiased estimate of *mutual information*, Panzeri and Treves (1996) have suggested a method to extrapolate information content to an infinite number of trials (also see Atencio et al., 2012). Sahani and Linden have developed the very insightful decomposition of the recorded signal into *signal power* and *noise power* (Sahani and Linden, 2003). This has lead to the *signal power explained (SPE)*, a measure based on *variance explained* which discounts "unexplainable" neural variability. This measure has been widely adopted, albeit under various names such as *predictive power, predicted response power, and relative prediction success* (Sahani and Linden, 2003; Machens et al., 2004; Ahrens et al., 2008; Asari and Zador, 2009; Rabinowitz et al., 2012). Also, it has been used as a basis for specific variants of measures for model performance (Haefner and Cumming, 2009). Focusing on coherence and the correlation, Hsu and colleagues have developed a method to normalize those measures by their upper bound (CC_{max}), which is given by the inter-trial variability (Hsu et al., 2004). This yields the *normalized correlation coefficient* (CC_{norm}). Following their suggestion, the upper bound can be approximated by looking at the similarity between one half of the trials and the other half of the trials (CC_{half}). This measure has also been used by Gill et al. (2006) and Touryan et al. (2005). Others have used the absolute correlation coefficient and controlled for inter-trial variability by comparing the absolute values with CC_{half} (Laudanski et al., 2012).

In this study we have analyzed in detail two measures of model quality that account for neural response variability, *SPE* and CC_{norm}. We have revealed the shortcomings of *SPE*, which has no lower bound and can yield undesirable negative values even for fairly reasonable model predictions. Furthermore, we have uncovered the close mathematical relationship between *SPE* and CC_{norm}, consolidated both approaches and arrived at several insights. First, both measures quantify prediction accuracy using the covariance (and *only* using covariance). Second, both measures quantify neural variability using the *signal power* (*SP*) (and *only* using *SP*). Third, when the variance of the prediction error approaches zero, *SPE* becomes identical to the square of CC_{norm}. And finally, it is not necessary to approximate CC_{max} using computationally expensive and inexact resampling methods because CC_{norm} can be calculated directly via *SP*:

$$CC_{abs} = \frac{Cov(y,\hat{y})}{\sqrt{Var(\hat{y})Var(y)}} \qquad CC_{norm} = \frac{Cov(y,\hat{y})}{\sqrt{Var(\hat{y})SP}} \tag{28}$$

$$SP = \frac{Var\left(\sum_{n=1}^{N} R_n\right) - \sum_{n=1}^{N} Var(R_n)}{N(N-1)} \tag{29}$$

This consolidated definition of CC_{norm} is not only more elegant, precise, and efficient, but it also sheds light on how CC_{norm} can be interpreted. It is almost identical to the well-known Pearson's correlation coefficient CC_{abs}, but the variance (power) of the recorded signal is replaced with the *signal power SP*, i.e., the deterministic and thus predictable part of the signal. As demonstrated, using *SPE* as a measure of model performance can yield misleading results and will limit interpretability of the results. However, CC_{norm} has been shown to fulfill the criteria of Section 2 for insightful measures: it is bounded, interpretable, and comparable across data sets. Thus, CC_{norm} is a well-defined and helpful tool to assess model performance[4].

[4] Matlab code for all measures can be found on GitHub: https://github.com/OSchoppe/CCnorm.

Note, however, that CC_{norm} cannot be estimated accurately if the data are excessively noisy. Equation 28 requires *SP* to be large enough to estimate with reasonable accuracy. For very noisy data or too few trials, observed *SP* values can become dominated by sampling noise, and may then behave as near zero random numbers. This would render CC_{norm} estimates unstable, allowing them to become spuriously large (if *SP* is small and underestimates the true value) or even imaginary (if the *SP* underestimate is severe enough to become negative). Thus, if *SP* or CC_{max} are small or have a very wide confidence interval, CC_{norm} values must be treated with caution.

8. AUTHOR CONTRIBUTIONS

OS: initiated the project; developed methodology; wrote and tested code implementing methods; analyzed method performance both analytically and through experiment; lead author on paper. NH, BW, AK, JS: guided research, co-wrote manuscript.

ACKNOWLEDGMENTS

This work was supported by a Wellcome Trust grant (WT076508AIA) and a BBSRC grant (BB/H008608/1). OS was supported by the German National Academic Foundation.

REFERENCES

Ahrens, M. B., Linden, J. F., and Sahani, M. (2008). Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *J. Neurosci.* 28, 1929–1942. doi: 10.1523/JNEUROSCI.3377-07.2008

Asari, H., and Zador, A. M. (2009). Long-lasting context dependence constrains neural encoding models in rodent auditory cortex. *J. Neurophysiol.* 102, 2638–2656. doi: 10.1152/jn.00577.2009

Atencio, C. A., and Schreiner, C. E. (2013). *Stimulus Choices for Spike-Triggered Receptive Field Analysis*, Chapter 3. New York, NY: Nova Biomedical.

Atencio, C. A., Sharpee, T. O., and Schreiner, C. E. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. *J. Neurophysiol.* 107, 2594–2603. doi: 10.1152/jn.01025.2011

Bizley, J. K., Walker, K. M., King, A. J., and Schnupp, J. W. (2010). Neural ensemble codes for stimulus periodicity in auditory cortex. *J. Neurosci.* 30, 5078–5091. doi: 10.1523/JNEUROSCI.5475-09.2010

Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Network* 12, 199–213. doi: 10.1080/713663221

David, S. V., and Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network* 16, 239–260. doi: 10.1080/09548980500464030

David, S. V., and Shamma, S. A. (2013). Integration over multiple timescales in primary auditory cortex. *J. Neurosci.* 33, 19154–19166. doi: 10.1523/JNEUROSCI.2270-13.2013

Döerrscheidt, G. H. (1981). The statistical significance of the peristimulus time histogram (PSTH). *Brain Res.* 220, 397–401. doi: 10.1016/0006-8993(81)91232-4

Gill, P., Zhang, J., Woolley, S. M. N., Fremouw, T., and Theunissen, F. E. (2006). Sound representation methods for spectro-temporal receptive field estimation. *J. Comput. Neurosci.* 21, 5–20. doi: 10.1007/s10827-006-7059-4

Haefner, R. M., and Cumming, B. G. (2009). "An improved estimator of variance explained in the presence of noise," in *Advances in Neural Information Processing Systems 21*, eds D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Red Hook, NY: Curran Associates, Inc.), 585–592.

Hsu, A., Borst, A., and Theunissen, F. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network* 15, 91–109. doi: 10.1088/0954-898X-15-2-002

Kadir, S. N., Goodman, D. F., and Harris, K. D. (2014). High-dimensional cluster analysis with the masked em algorithm. *Neural Comput.* 26, 2379–2394. doi: 10.1162/NECO-a-00661

Kass, R. E., Ventura, V., and Cai, C. (2003). Statistical smoothing of neuronal data. *Network* 14, 5–16. doi: 10.1088/0954-898X/14/1/301

Laudanski, J., Edeline, J.-M., and Huetz, C. (2012). Differences between spectro-temporal receptive fields derived from artificial and natural stimuli in the auditory cortex. *PLoS ONE* 7:e50539. doi: 10.1371/journal.pone.0050539

Machens, C. K., Wehr, M. S., and Zador, A. M. (2004). Linearity of cortical receptive fields measured with natural sounds. *J. Neurosci.* 24, 1089–1100. doi: 10.1523/JNEUROSCI.4445-03.2004

Marmarelis, P. (1978). *Analysis of Physiological Systems: the White-Noise Approach*. New York, NY: Plenum Press. doi: 10.1007/978-1-4613-3970-0

Panzeri, S., and Treves, A. (1996). Analytical estimates of limited sampling biases in different and information measures. *Network* 7, 87–107. doi: 10.1088/0954-898X/7/1/006

Prenger, R., Wu, M. C.-K., David, S. V., and Gallant, J. L. (2004). Nonlinear V1 responses to natural scenes revealed by neural network analysis. *Neural Netw.* 17, 663–679. doi: 10.1016/j.neunet.2004.03.008

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., and King, A. J. (2011). Contrast gain control in auditory cortex. *Neuron* 70, 1178–1191. doi: 10.1016/j.neuron.2011.04.030

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., and King, A. J. (2012). Spectrotemporal contrast kernels for neurons in primary auditory cortex. *J. Neurosci.* 32, 11271–11284. doi: 10.1523/JNEUROSCI.1715-12.2012

Roddey, J. C., Girish, B., and Miller, J. P. (2000). Assessing the performance of neural encoding models in the presence of noise. *J. Comput. Neurosci.* 8, 95–112. doi: 10.1023/A:1008921114108

Sahani, M., and Linden, J. F. (2003). "How linear are auditory cortical responses?," in *Advances in Neural Information Processing Systems 15*, Vol. 15, eds S. Becker, S. Thrun, and K. Obermayer (MIT Press), 109–116.

Shimazaki, H., and Shinomoto, S. (2007). A method for selecting the bin size of a time histogram. *Neural Comput.* 19, 1503–1527. doi: 10.1162/neco.2007.19.6.1503

Simoncelli, E. P., Paninski, L., Pillow, J., and Schwartz, O. (2004). "Characterization of neural responses with stochastic stimuli," in *The Cognitive Neurosciences, 3rd Edn.*, ed M. Gazzaniga (Cambridge, MA: MIT Press), 327–338.

Sohl-Dickstein, J., Poole, B., and Ganguli, S. (2013). "Fast large-scale optimization by unifying stochastic gradient and quasi-newton methods," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, eds T. Jebara and E. P. Xing (Beijing), 604–612.

Touryan, J., Felsen, G., and Dan, Y. (2005). Spatial structure of complex cell receptive fields measured with natural images. *Neuron* 45, 781–791. doi: 10.1016/j.neuron.2005.01.029

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

A.4. Network Receptive Field Modeling Reveals Extensive Integration and Multi-feature Selectivity in Auditory Cortical Neurons

Authors: N Harper*, **O Schoppe***, B Willmore, Z Cui, J Schnupp, A King.
Joint first authorship

Abstract: Cortical sensory neurons are commonly characterized using the receptive field, the linear dependence of their response on the stimulus. In primary auditory cortex neurons can be characterized by their spectrotemporal receptive fields, the spectral and temporal features of a sound that linearly drive a neuron. However, receptive fields do not capture the fact that the response of a cortical neuron results from the complex nonlinear network in which it is embedded. By fitting a nonlinear feedforward network model (a network receptive field) to cortical responses to natural sounds, we reveal that primary auditory cortical neurons are sensitive over a substantially larger spectrotemporal domain than is seen in their standard spectrotemporal receptive fields. Furthermore, the network receptive field, a parsimonious network consisting of 1-7 sub-receptive fields that interact nonlinearly, consistently better predicts neural responses to auditory stimuli than the standard receptive fields. The network receptive field reveals separate excitatory and inhibitory sub-fields with different nonlinear properties, and interaction of the sub-fields gives rise to important operations such as gain control and conjunctive feature detection. The conjunctive effects, where neurons respond only if several specific features are present together, enable increased selectivity for particular complex spectrotemporal structures, and may constitute an important stage in sound recognition. In conclusion, we demonstrate that fitting auditory cortical neural responses with feedforward network models expands on simple linear receptive field models in a manner that yields substantially improved predictive power and reveals key nonlinear aspects of cortical processing, while remaining easy to interpret in a physiological context.

Date of publication: 11.11.2016
Journal name: PLoS Computational Biology

Individual contribution: formal analysis, methodology, software, validation, visualization, writing of original draft, reviewing and editing

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Network Receptive Field Modeling Reveals Extensive Integration and Multi-feature Selectivity in Auditory Cortical Neurons

Nicol S. Harper[1,2☯]*, Oliver Schoppe[1,3☯], Ben D. B. Willmore[1], Zhanfeng Cui[2], Jan W. H. Schnupp[4,1], Andrew J. King[1]

1 Dept. of Physiology, Anatomy and Genetics (DPAG), Sherrington Building, University of Oxford, United Kingdom, **2** Institute of Biomedical Engineering, Department of Engineering Science, Old Road Campus Research Building, University of Oxford, Headington, United Kingdom, **3** Bio-Inspired Information Processing, Technische Universität München, Germany, **4** Department of Biomedical Science, City University of Hong Kong, Kowloon Tong, Hong Kong

☯ These authors contributed equally to this work.
* nicol.harper@dpag.ox.ac.uk

Abstract

Cortical sensory neurons are commonly characterized using the receptive field, the linear dependence of their response on the stimulus. In primary auditory cortex neurons can be characterized by their spectrotemporal receptive fields, the spectral and temporal features of a sound that linearly drive a neuron. However, receptive fields do not capture the fact that the response of a cortical neuron results from the complex nonlinear network in which it is embedded. By fitting a nonlinear feedforward network model (a network receptive field) to cortical responses to natural sounds, we reveal that primary auditory cortical neurons are sensitive over a substantially larger spectrotemporal domain than is seen in their standard spectrotemporal receptive fields. Furthermore, the network receptive field, a parsimonious network consisting of 1–7 sub-receptive fields that interact nonlinearly, consistently better predicts neural responses to auditory stimuli than the standard receptive fields. The network receptive field reveals separate excitatory and inhibitory sub-fields with different nonlinear properties, and interaction of the sub-fields gives rise to important operations such as gain control and conjunctive feature detection. The conjunctive effects, where neurons respond only if several specific features are present together, enable increased selectivity for particular complex spectrotemporal structures, and may constitute an important stage in sound recognition. In conclusion, we demonstrate that fitting auditory cortical neural responses with feedforward network models expands on simple linear receptive field models in a manner that yields substantially improved predictive power and reveals key nonlinear aspects of cortical processing, while remaining easy to interpret in a physiological context.

OPEN ACCESS

Citation: Harper NS, Schoppe O, Willmore BDB, Cui Z, Schnupp JWH, King AJ (2016) Network Receptive Field Modeling Reveals Extensive Integration and Multi-feature Selectivity in Auditory Cortical Neurons. PLoS Comput Biol 12(11): e1005113. doi:10.1371/journal.pcbi.1005113

Editor: Israel Nelken, Hebrew University of Jerusalem, ISRAEL

Received: December 12, 2015

Accepted: August 22, 2016

Published: November 11, 2016

Data Availability Statement: Data can be found at https://osf.io/ayw2p/.

Funding: NSH was supported by a Sir Henry Wellcome Postdoctoral Fellowship (grant number WT082692) and other Wellcome Trust funding (grant numbers WT076508AIA and WT108369/Z/2015/Z), by the Department of Physiology, Anatomy and Genetics at the University of Oxford, by Action on Hearing Loss (grant number PA07), and by the Biotechnology and Biological Sciences Research Council (grant number BB/H008608/1). OS was supported by the German Academic

Scholarship Foundation. ZC was supported by the Biotechnology and Biological Sciences Research Council (grant number BB/H008608/1). AJK and BDBW were supported by a Wellcome Trust Principal Research Fellowship to AJK (grant number WT076508AIA and WT108369/Z/2015/Z). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Author Summary

Linear filter descriptions of sensory neurons have been with us since the 1970s, and have been enormously influential. But such models, and more recent nonlinear variants, are rather like modeling the entire network as a single neuron, failing to account for the neuron's response being a consequence of a network of many nonlinear units. Here we show how these limitations can be overcome by using recent advances in machine learning to fit "network receptive field models" to neural responses to natural sounds. Feedforward networks of 1–7 nonlinearly-interacting lower-order model neurons are required to model a cortical receptive field. Each lower order neuron is tuned to somewhat different stimulus features, arranged together in complex but interpretable structures, which cover a far wider range of sound frequencies and delays than current receptive field models indicate. The NRF models capture important nonlinear functional characteristics in auditory cortical neurons, including multiplicative gain control and conjunctive feature selectivity, where neurons respond when certain features are present together but not in isolation. This enables NRFs to predict the responses of auditory cortical neurons with considerably greater accuracy than conventional models.

Introduction

Developing models capable of quantitatively predicting neural responses to sensory stimuli is key to understanding the neural computations underlying perception. A widespread model of sensory neurons, including cortical sensory neurons, is the receptive field (RF), which describes the best-fitting linear transformation from the stimulus to the neural response [1–16]. RF models, although simple and useful, are only moderately effective in capturing neural responses since processing by networks of neurons includes highly nonlinear operations. Consequently, they can fail to produce adequate descriptions of neural responses, particularly to natural stimuli [17,18].

While spectrotemporal receptive fields (STRF) of neurons in primary auditory cortex (A1) can be quite broad and complex, many of them are punctate, typically little more than a point in time and frequency, indicating little of the likely complexity of cortical processing [19] (although see [12]). Adding specific nonlinearities to STRF models [17], for example by applying output nonlinearities [19,20] to create linear-nonlinear (LN) models, improves prediction somewhat. However, basic LN models, consisting of just a single STRF and an output nonlinearity, still fail to capture the interactions of sensory filters that are bound to occur naturally in the neural networks of ascending sensory pathways. Recently, more complex and often nonlinear STRF models [20–25] of A1 neurons have achieved improved predictions of experimental data, although sometimes at the expense of being very computationally intensive. These newer models have tended to concentrate on better modeling of features local to the neuron, such as synaptic depression [23] or refractoriness [22]. Other valuable approaches adopted to characterize the feature selectivity of A1 neurons are more phenomenological in nature [20].

Here we take a very different approach, one that embodies the fact that a neuron's response is the result of it being embedded in a network of many neurons, each of which is a nonlinear unit. We take advantage of recent advances in the training of artificial neural networks [26] to produce a new type of RF model, the network receptive field (NRF), which can be rapidly fitted to neural response data. The NRF model is composed of a hierarchical feedforward network of 20 LN units, embodying the fact that cortical neurons integrate the output of many lower order neurons. Although our choice of 20 possible feed-forward connections does not reflect the full range of converging inputs that cortical neurons receive, this approach stands in contrast to the above mentioned recent models of A1 responses [20–24], which use only one, or in some cases two, STRF-like units. In fact, we show here that up to seven effective units are required to model a cortical receptive field.

Receptive field models tend to include large numbers of free parameters, which can lead to problems with "overfitting": the many free parameters of the model may capture unimportant or coincidental details or noise in the training set. This can result in the model appearing to successfully capture the stimulus-response relationships in the training set, but subsequently performing poorly when the model is used to predict neural responses to novel stimuli that were not part of the training set. To prevent the risk of overfitting affecting our results we took the following steps: First, during model fitting, the NRF was regularized by the summed magnitudes of the network's weights (L_1-norm), which automatically prunes away superfluous weights and hidden units (from an initial 20 hidden units). This produces parsimonious and readily interpretable connection patterns that provide insights into the underlying circuitry. Second, we made extensive use of cross-validation during model training (see below) and assessed the performance of all models using a generalization test set which the models had not been exposed to during training.

Together, the regularization, cross-validation and generalization testing adopted here ensure that the improved performance exhibited by our NRF models is not a trivial consequence of the larger number of degrees of freedom that these models can bring to bear, but rather reflect the fact that the structure of these models renders them better able than conventional LN receptive field models to capture aspects of the sensory processing performed by the auditory pathway. Thus, using electrophysiological recordings from ferret auditory cortex, we find that NRF models consistently outperform LN models in predicting the responses of auditory cortical neurons to natural stimuli. The fitted NRF models of auditory cortical neurons reveal sensitivity over substantially wider time and frequency ranges than conventional LN and STRF models, and the NRFs also reveal distinct nonlinear properties, including gain control and conjunctive feature selectivity, features that may be critical to auditory cortex function. Conjunctive feature selectivity, where neurons respond when certain features are present together but not in isolation, allows neurons to show increased selectivity to specific complex spectrotemporal structures and may provide a valuable stage in the sound recognition process.

Results

Network receptive field models of neural responses in auditory cortex

To investigate the ability of NRFs to account for cortical sensory responses, we fitted models to neural responses to clips of natural sounds. Seventy-six single-unit responses were recorded with multi-channel electrodes in the ferret primary auditory cortical areas, A1, and the anterior auditory field (AAF). The stimuli comprised 20 clips of natural acoustic scenes, each of 5 s duration, including ferret vocalizations, speech, and environmental sounds. The model fitting process is shown schematically in Fig 1. and described in detail in the Materials and Methods. The first step in the NRF model was to generate a first order approximation of auditory nerve response patterns to the stimuli, referred to here as the "cochleagram", by measuring the log amplitude of the sound in each of 34 log-spaced frequency channels, spanning 0.5 to 22.6 kHz with ⅙ octave spectral resolution and 5 ms temporal resolution. The task of the model was then to predict the firing patterns recorded from the cortical neurons, also binned with 5 ms time resolution, from the previous 100 ms (20 time bins) of stimulus history.

In accordance with principles of model selection and assessment [27], we divided the neural response data into a cross-validation set and a test set. The cross-validation set was then

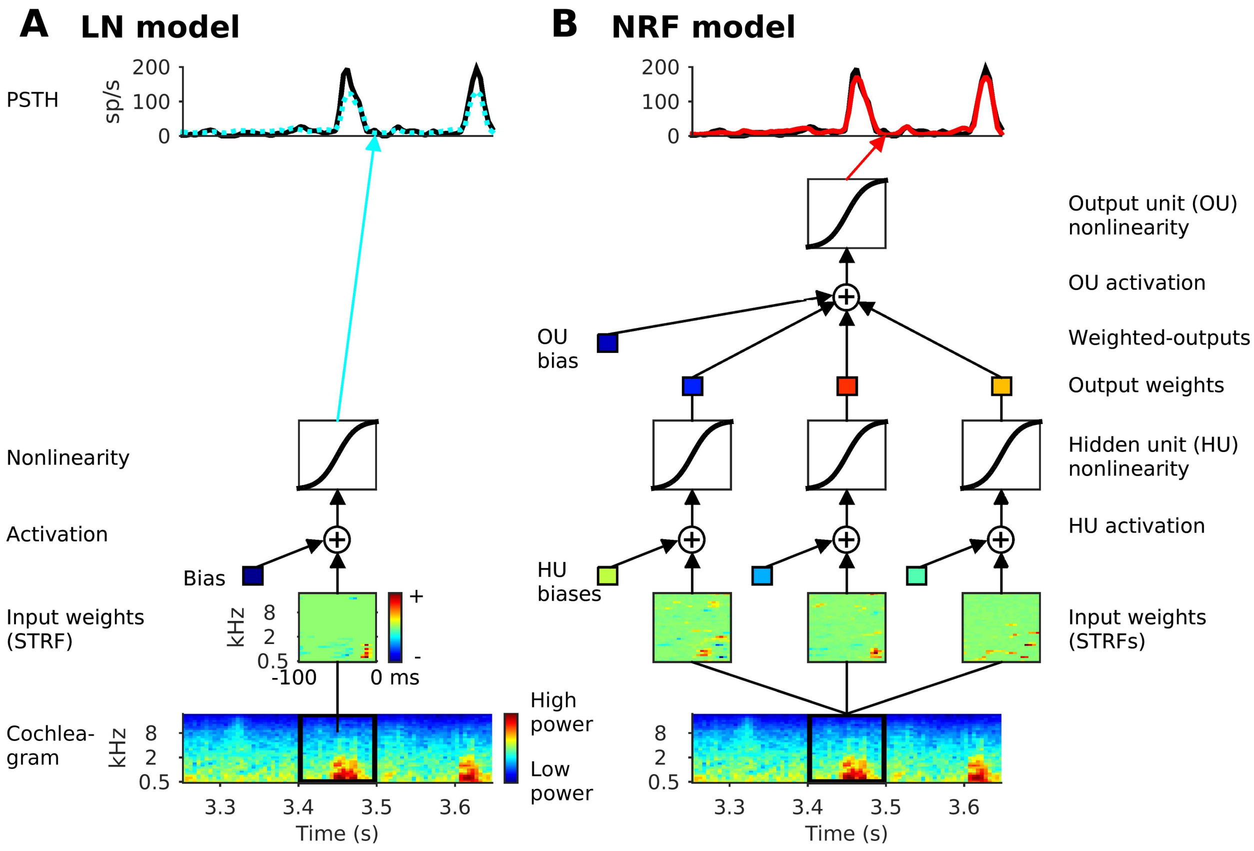


Fig 1. Schematics of the models. (A) The linear-nonlinear (LN) model. (B) The network receptive field (NRF) model, a feedforward neural network

doi:10.1371/journal.pcbi.1005113.g001

divided again into a training set that was used to fit the model parameters, and a validation set on with which the model's capacity to predict neural response was then assessed. By this means, the optimum value for general settings of the model (hyperparameters, such as the degree of regularization) could be determined. This fitting was repeated ten times for ten different ways of dividing the cross-validation set, to ensure a robust assessment of the optimum model hyperparameter values. Note that the model fits, for both the LN and NRF models, tended to differ little in their receptive field forms over these ten fits, despite having slightly different datasets and different randomly chosen weight initializations, indicating the robustness of the fitting procedure (for details and quantification see Materials and Methods). Once the optimum hyperparameters were obtained, the model was re-fitted using the full cross-validation dataset. Finally, the test dataset that was put aside was used to assess the fitted model's capacity to predict responses to sounds not encountered at any stage of the fitting (i.e. to "generalize"). All model performance data reported below refer to results obtained with the test set.

A conventional LN model and an NRF model were fitted for each cortical neuron in our dataset. The LN model comprised a linear STRF, used to calculate the activation of the model neuron, and a sigmoidal output nonlinearity (Fig 1A). The linear STRF on its own also provided a basic linear (L) model. The NRF model was a rate-based, feed-forward neural network (a multilayer perceptron), with units that integrate inputs linearly followed by a nonlinear transformation to produce their output [28]. The NRF effectively computes a weighted sum of several LN models, where each hidden unit (HU) instantiates one LN model, and their outputs are combined linearly as they converge on the output unit (OU). The resulting OU activation passes through a further sigmoidal nonlinear activation function (Fig 1B) to yield the NRF model's prediction of the neural firing rate. The network units have no memory from time point to time point; the model does not use any recurrent or convolutional elements. All models (LN and NRF) were fitted by minimizing the squared error between the model's estimate of the neural response and the actual neural firing rate (see Materials and Methods for details). Importantly, L_1-norm regularization of the connection weights was used to find a parsimonious representation. A recently developed algorithm [26] allowed for good NRF models to be fitted rapidly and efficiently for all 76 cortical neurons in our dataset.

NRF models describe neural responses better than LN models

To assess the models' predictive power, we measured how well they were able to predict responses to a "test set" of stimuli which were not part of the training set used during model fitting. The NRF tends to better predict the amplitude of sharp peaks in the observed neural response than the LN model (Fig 2A and 2B, seconds 3–4 are from the training set, seconds 4–5 are from the test set). We quantified the quality of the response prediction by calculating the normalized correlation coefficient (CC_{norm}) between predicted and observed neural responses [29,30]. A CC_{norm} of 0 would indicate that the model fails to predict the neural responses any better than chance, while CC_{norm} values of 1 indicate predictions that are at the highest achievable accuracy (see Materials and Methods). For the great majority of neurons (70/76), the NRF achieved higher CC_{norm} values than the LN model ($p = 6.3\times10^{-15}$, n = 76, sign test; Fig 2C), with the mean CC_{norm} for the NRFs being 0.73, compared to 0.67 for the LN model. The CC_{norm} for the L-model, the prediction using the STRF but without processing through the fitted sigmoidal output nonlinearity, was 0.60, significantly less than both the NRF (76/76, $p = 2.6\times10^{-23}$, n = 76, sign test) and the LN model (75/76, $p = 2.0\times10^{-21}$, n = 76, sign test). The CC_{norm} value for the NRF model may approach the maximum possible given the duration of the STRFs (100 ms) used by the NRF model [31] (See Discussion). We also report the raw mean correlation coefficient (CC_{raw}), which was 0.61 for the NRF model, and 0.56 for the LN model, to enable comparison with previous publications (but note, that differences in raw CC_{raw} values between different studies are difficult to interpret, as will be discussed further below). As expected, for this measure too, the great majority of neurons (70/76) were significantly better fit by the NRF than the LN model ($p = 6.3\times10^{-15}$, n = 76, sign test). The CC_{raw} for the linear model was 0.50, significantly less than both the NRF (76/76, $p = 2.6\times10^{-23}$, n = 76, sign test) and the LN model (75/76, $p = 2.0\times10^{-21}$, n = 76, sign test).

The capacity of the NRF model to predict better than the LN model is also robust to the exact choice of test set. This is evident from examining the prediction quality for the validation sets, which, in order to require the model to generalize across stimulus types, comprised 2 of the 20 sounds, chosen at random. The mean CC_{norm} for the validation set, averaged over all 10 folds, is greater for the NRF model (0.76) than the LN model (0.71), with significantly more neurons (69/76) showing a greater CC_{norm} for the NRF model than the LN model ($p = 6.4\text{x}10^{-14}$, n = 76, sign test).

To investigate how well the models are able to predict peak responses in the test set, we also measured the "peak activity mean squared error", which was defined as the MSE between the observed firing rates and those predicted by the models during periods where the observed firing rate exceeded two standard deviations above the mean firing rate (the "2σ-threshold", dotted line in Fig 2B). It is readily apparent that the peak activity MSE of the NRF model is smaller than that of the LN model for the great majority of neurons ($p = 5.2\times10^{-16}$, n = 76, sign test; Fig 2D). The peak activity MSE, averaged over all neurons, was 27% smaller for the NRF model

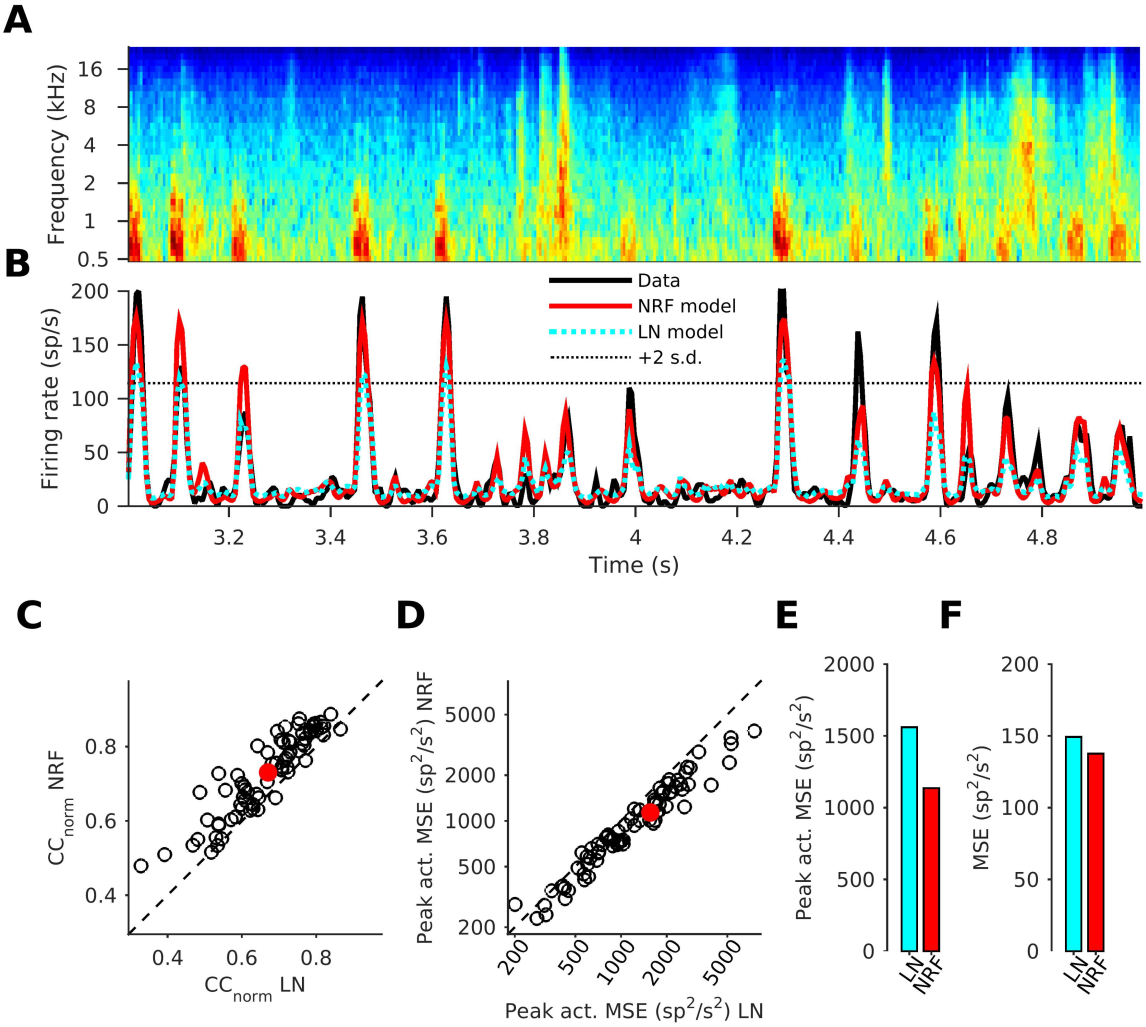


Fig 2. A neural network receptive field model predicts the response of auditory cortical neurons better than the LN model. (A) Cochleagram for a 2 s sound stimulus snippet. (B) The neural response firing rate to the stimulus snippet shown in A (black line) for one example neuron, shown alongside the predicted responses of the LN-model (dotted cyan line) and the NRF model (red line). The thin dotted black line indicates the 2σ-threshold, which was used to identify periods of large response peaks. (C) Prediction quality (normalized correlation coefficient) of the NRF model plotted against that of the LN-model for all 76 neurons in our dataset. (D) Mean squared error (MSE) of the prediction during peak response times of the NRF models plotted against error of the LN-models. (E) Average peak activity MSE (pMSE) over the whole dataset for the NRF and the LN models. (F) Average MSE of the predictions generated by NRF and LN models as in E, but calculated across the whole response to the test stimuli, not just the peak response period.

doi:10.1371/journal.pcbi.1005113.g002

than for the LN model (Fig 2E). This reduction of prediction error during periods of peak excitation appears to drive the improved performance of the NRF model relative to the LN model. This is indicated by much smaller average improvement (8%) for the NRF over the LN model when the MSE was measured over the entire neural response (Fig 2F).

Note that all the model performance data in Fig 2. were calculated exclusively from test sets that the models were not exposed to during training. This is essential to ensure that appropriate model comparisons were made. NRF models have significantly more free parameters than conventional LN models, and, if tested on the training data, might trivially outperform the LN models by overfitting noise in the training data, but such overfitting would become disadvantageous when the models were used to make predictions for novel stimulus sets. The fact that NRF models outperform LN models on previously unseen data indicates that the NRF models mimic aspects of the behavior of the cortical neurons which the structure of LN models cannot account for.

NRFs reveal that cortical neurons are better described by the interaction of multiple, diverse sub-receptive-fields

We first qualitatively examined the fitted characteristics of the two models (Fig 3; each of the 10 numbered rows shows an example neuron; neuron 1 was used in Figs 1 and 2B). In our dataset, as is quite commonly the case, the LN model STRFs are rather "punctate", i.e. the model neuron is driven almost exclusively by stimulus elements clustered narrowly in frequency and recent stimulus history, often an excitatory point with some weak lagging inhibition (Fig 3A, the top panel shows the STRF for each neuron). Moreover, the LN model tends to operate in the near-threshold region of the nonlinear activation function (Fig 3A, lower panel for each neuron), with activations straddling the expansive part of the sigmoidal output function.

The NRF model reveals more complex tuning properties (Fig 3B–3D, for the same 10 example neurons). Each NRF model had 20 HUs, but because the model training incorporated a regularization term that penalizes ineffectual and redundant synaptic weights (see Materials and Methods), HUs could develop substantive synaptic weights only if these were able to "explain" aspects of the firing of the biological neuron that were not already covered by the other HUs in the feedforward network. Any HUs that were redundant would have their input and output weights, and hence their overall contribution to the NRF, shrink to negligibly small values. We found that, of the 20 HUs in each NRF, most turned out to be redundant during the course of model fitting, and each NRF ended up with a relatively small number of "effective" HUs (between 1 and 7), which were the only ones to send strong signals to the output neuron (Fig 3B; for each neuron, each column shows an 'effective' HU with an STRF, top panel, and a nonlinearity, bottom panel). The variance, calculated over the full stimulus set, of an HU's weighted-output (HU output × output weight) provides a measure of the HU's 'effectiveness' (Fig 3D, top panel), with HUs with a variance $\geq$5% (Fig 3D, red line) of the sum of the variances of all 20 HUs being considered effective. The weighted-output of an effective HU varies greatly as it rises and falls to signal the presence or absence of particular stimulus features. In contrast, the weighted-output variance of an "ineffective" HU is close to zero. Thus, an NRF with three effective HUs (Fig 3B, neuron 1) has three weighted-output variances above the 5% threshold (Fig 3D, top panel, neuron 1).

HUs can be classified into excitatory or inhibitory units according to whether their output weight is positive or negative, after adjusting the model to account for HUs where the input weights are predominantly negative and the output weight negative, which is effectively an excitatory HU, and also adjusting for HUs that show the converse (see Materials and Methods). If the plotted line of an HU's nonlinear activation function is red, it is excitatory, whereas if it is blue, the HU is inhibitory (Fig 3B, bottom panel for each HU). The STRFs of HUs are more diverse in form than the STRFs of the LN model, and together appear to cover a wider range of frequencies and times (Fig 3B, top panel for each HU). For display purposes only, the weights of the inhibitory HU STRFs in Fig 3 have their signs inverted so as to show their influence on the OU (see Materials and Methods, "The displayed STRFs").

We can examine how the HUs interact if we "zoom in" on a part of the STRF's frequency and temporal range marked by high levels of sensitivity (Fig 3C, 'zoomed' region identified by the black bars along the axes of the STRF for the first HU of each neuron in Fig 3B). Contours

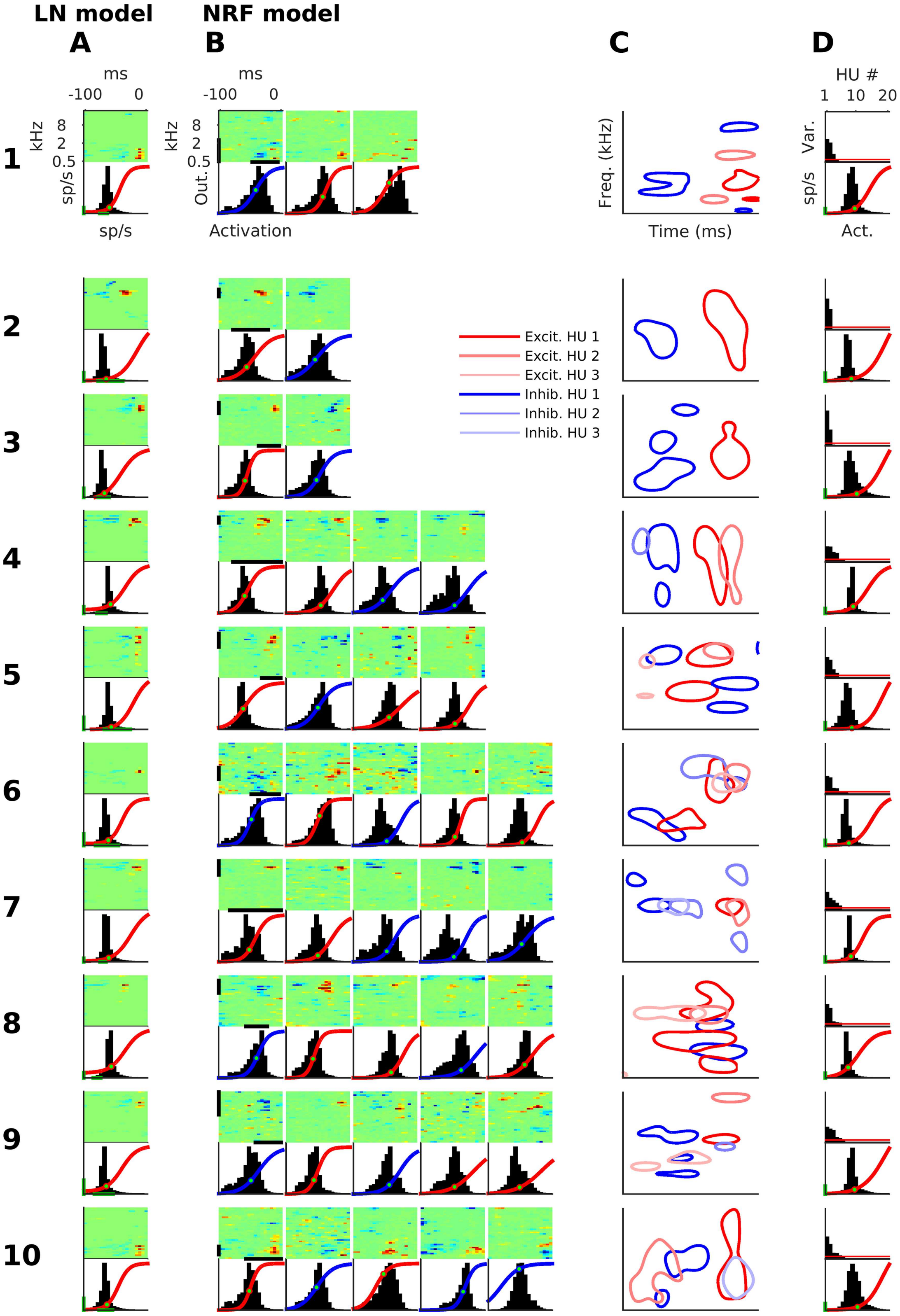

LN model
A
NRF model
B
C
D
ms
-100 0
ms
-100 0
HU #
1 10 20
kHz
8
2
0.5
sp/s
Out.
Freq. (kHz)
Var.
sp/s
Activation
Time (ms)
Act.
1
2
3
4
5
6
7
8
9
10
Excit. HU 1
Excit. HU 2
Excit. HU 3
Inhib. HU 1
Inhib. HU 2
Inhib. HU 3

Fig 3. Example STRFs and nonlinearities for both models. Each numbered row is an example neuron. (A) STRFs (top) and nonlinearities (bottom) for the LN model. Green bars mark 0–20 sp/s. The nonlinearities are superimposed over the distribution of activations. The green dot on the nonlinear activation function marks the mean output value. (B) Hidden unit (HU) 'STRFs' (top) and nonlinearities (bottom) for the NRF model. If the nonlinearity curve is red, it is an excitatory HU, if blue, it is inhibitory. Otherwise format as in A. Note that the STRFs of inhibitory units have been multiplied by -1 for display purposes, to indicate the direction of their influence on the final neural output (See Materials and Methods, The displayed STRFs). One would therefore not necessarily expect to observe extensive inhibitory areas such as those in some of these HU display STRFs in physiological recordings, as such inhibitory fields would most likely manifest in biological networks as excitatory fields that feed on to the next neuron via inhibitory synapses. (C) HU STRFs in B plotted together as contours. The contours are at 50% of an STRF's maximum (if an excitatory HU) or of its minimum (if an inhibitory HU). Each panel in column C only shows a sub-region of the full spectrotemporal range of the STRFs; it is an expansion of an area of interest, whose frequency range and temporal range are shown by the black bars on the edges of the first HU STRF of the neuron. (D) Top: the variance of weighted-output, the input to the output unit (OU), of each HU. The red line marks a variance of 5%, the threshold for distinguishing effective HUs from their "ineffective" counterparts. Bottom: OU nonlinearities for the NRF model for the same 10 cortical neurons. Format as bottom panel in A.

doi:10.1371/journal.pcbi.1005113.g003

delineate the time-frequency regions of high sensitivity for each of the effective HUs, using shades of red for excitatory HUs and shades of blue for inhibitory HUs (Fig 3C). Here, for each excitatory HU, time-frequency regions of high sensitivity were defined as those for which the STRF's weights (as shown in Fig 3C) exceeded half the maximum weight. For the inhibitory effective HUs, high-sensitivity regions were where the STRF weights fell below half the minimum weight. For many neurons, the high-sensitivity regions of the STRFs for different HUs align in close but distinct locations in spectrotemporal space to form apparent structures, suggesting the presence of conjunctive sensitivity to ordered features (see Discussion).

OUs tend to operate "near threshold" (where "threshold" is the lowest possible output value, see Materials and Methods), with activations (Fig 3D, bottom panel per neuron, black histogram) mostly confined to the expansive part of their nonlinear activation function (Fig 3D, bottom panel, red line), just like the LN model. However, the same is not always true for the effective HUs, many of which experience activations (Fig 3B, bottom panels per neuron, black histogram) that sometimes fall in the linear range of their nonlinear activation function (Fig 3B, bottom panels, red or blue line), or even operate over the compressive, upper range of their nonlinear activation function.

NRF models have between 1 and 7 effective hidden units

For the NRF model, the most common number of effective HUs of a neuron was 2; this was the case for 42% (32/76) of neurons (Fig 4A). Such 'bi-feature' neurons always have one excitatory and one inhibitory HU (Fig 4B). A few 'uni-feature' neurons, with only an excitatory effective HU, made up 5% (4/76) of our sample (Fig 4A). The remaining 53% (42/76) were 'multi-feature' neurons with between 3 and 7 (mode = 5) effective HUs (Fig 4A), which tended to have more excitatory HUs than inhibitory HUs ($p = 0.035$, $n = 76$, sign test; Fig 4B).

NRF models reveal wider integration over time and frequency than LN models

To quantify the time and frequency tuning widths of LN model STRFs, we first calculated the "power STRF" for each neuron by squaring the weights in the STRF (see Materials and Methods). The temporal tuning width was then determined by summing the power STRF over the frequency bands and counting the number of time bins with power $\geq$25% of the maximum. Multiplying this count by the temporal bin size gave the temporal tuning width at quarter-height. The frequency tuning width at quarter-height was measured analogously by summing the power STRF over time and multiplying the number of bins exceeding a quarter of the maximal power by the width of each frequency channel.

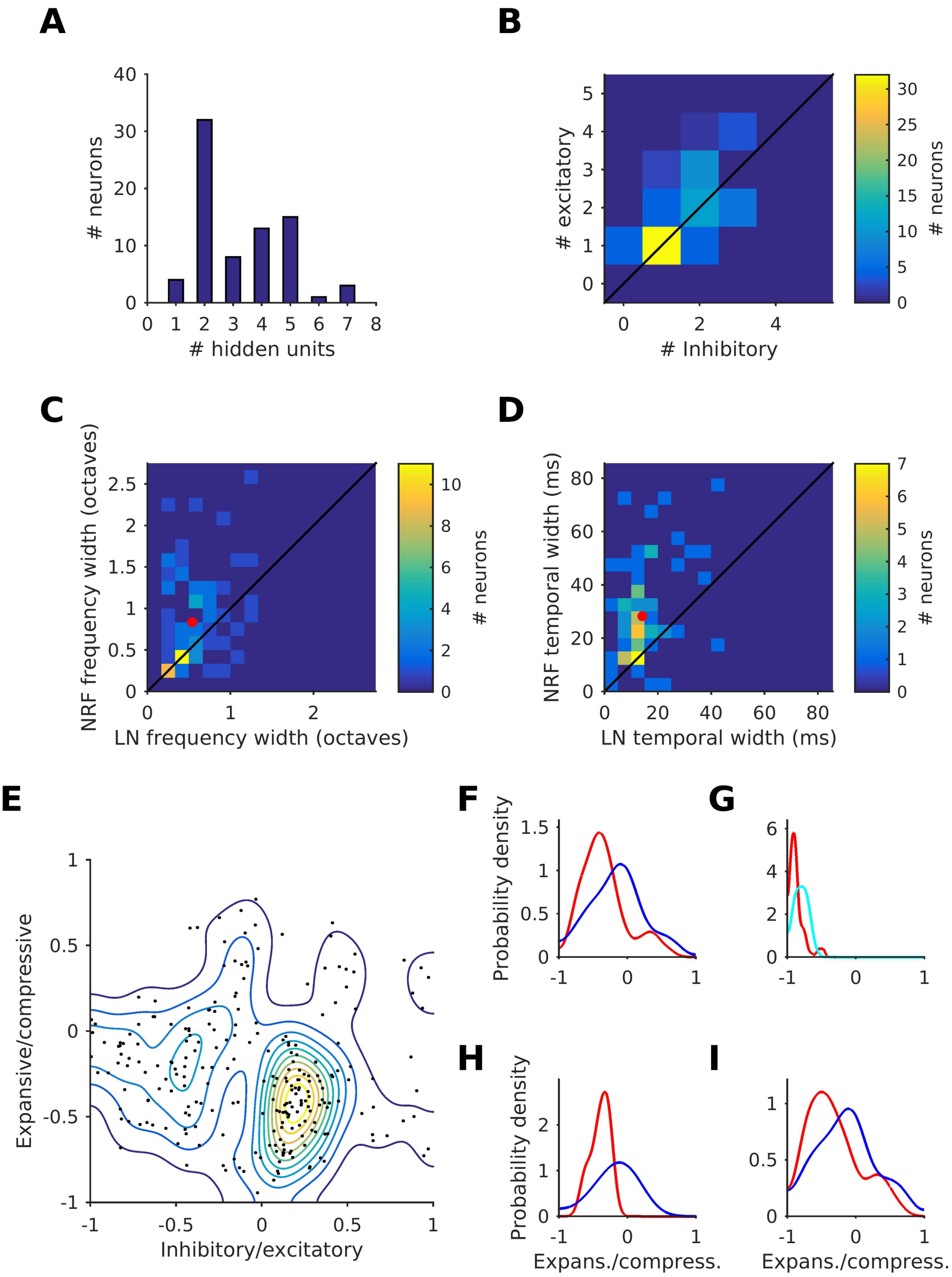


Fig 4. Properties of the feedforward neural network model. (A) Histogram of the number of effective hidden units (HU) of the NRF model fits for each neuron. (B) Distribution of the number of effective excitatory and inhibitory HUs for each neuron. (C) Frequency tuning width of the NRF model versus that of the LN model. The red dot indicates the average frequency tuning width of both models. (D) Temporal tuning width of the power STRF of the NRF model versus that of the LN model. The red dot indicates the average temporal tuning width of both models. (E) A plot of the expansive/compressive (EC) score (which measures how the nonlinear activation function is used) against the excitatory/inhibitory (IE) score for all 246 effective HUs from the 76 neurons. Contour plot shows the density. (F) Distribution of the EC score for excitatory (red) and inhibitory (blue) HUs. (G) Distribution of the EC score for the output unit of the NRF model (red) and for the LN model (cyan). (H) Distribution of the EC score for excitatory (red)

and inhibitory (blue) HUs for the bi-feature neurons. (I) Distribution of the EC score for excitatory (red) and inhibitory (blue) HUs for the multi-feature neurons.

doi:10.1371/journal.pcbi.1005113.g004

To obtain comparable measurements of overall tuning widths for the NRF models, we calculated power STRFs for each of the NRF model's HUs, then summed their power STRFs, weighted by the strength of the signals that they contribute to the OU, which was quantified as the variances of their weighted-outputs, as shown in the bar charts in Fig 3D. Quarter-height frequency and temporal tuning widths were then calculated from the weighted sum power STRF in the same way as for the LN model power STRFs.

The quarter-height frequency and the temporal tuning width for the NRF model was, for most neurons, significantly larger than for the LN model ($n = 76$, $p = 1.6\times10^{-3}$ and $p = 1.8\times10^{-12}$, respectively, sign-test; Fig 4C and 4D respectively). On average, the frequency tuning width for the NRF model was 0.83 octaves, which is 54% larger than the 0.54 octaves for the LN model. The average temporal tuning width for the NRF model was 28.2 ms, which is 99% larger than the 14.2 ms for the LN model. We also carried out analogous analyses for frequency and temporal tuning widths measured at half-height, which produced similar results, showing significantly larger tuning widths in the NRF model for both frequency and time (27% and 46% larger, respectively, $n = 76$, $p = 0.033$ and 1.7×10^{-7}, respectively, sign-test).

Inhibitory and excitatory receptive sub-fields have different nonlinear properties

We next examined whether excitatory or inhibitory HUs differ in the extent to which they operate over the expansive, linear or compressive part of their output nonlinear activation function. For each effective HU in our dataset, we computed an expansive/compressive measure (EC score) and inhibitory/excitatory measure (IE score; see Materials and Methods for details). Both EC and IE scores are bounded between -1 and 1. A unit with a negative EC score operates predominantly in a near-threshold, expansive region of its nonlinearity, while a positive score indicates that it operates in a compressive, saturating region, and a score close to zero indicates operation in a linear region. Negative IE scores mean that a unit is (predominantly) inhibitory and positive scores that it is excitatory. EC scores are plotted against IE scores for all 246 effective HUs from the 76 neurons in Fig 4E. Superimposed on the scatter plot is a contour plot reflecting the density estimate of the scatter in the "EC/IE space". The density estimation used a kernel with its bandwidth optimized to smooth away statistically spurious peaks [32]. The HUs fell into two broad clusters: the first, dense cluster is excitatory ($IE \geq 0$) and expansive ($EC < 0$), whereas the other, broader cluster is inhibitory ($IE < 0$) and more linear ($EC \approx 0$).

To confirm this observation, we divided the neurons into excitatory ($IE \geq 0$) and inhibitory ($IE < 0$), and separately plotted the distribution of the EC score for each, again using a kernel density estimator with optimally chosen kernel bandwidth (Fig 4F). The EC scores of the 115 inhibitory HUs (blue) were more or less symmetrically distributed around 0, indicating that these HUs mostly operate in a linear region, whereas the great majority of the 131 excitatory HUs have EC scores <0, indicating that they tend to operate in the near threshold, expansive region of their output nonlinear activation function. The difference in the median EC value of the two distributions was significant ($p = 6.1\times10^{-6}$, rank sum test), confirming the EC/IC space clustering observations (Fig 4E). The distribution of EC scores for the OUs of the NRF model for all 76 neurons (Fig 4G, red), shows that the OUs operate largely near threshold, in the expansive region ($EC < 0$). This is similar to the case for the EC scores in the LN model (Fig 4G, cyan), where all the neurons also operate in the expansive region ($EC < 0$).

If we restrict the above analysis to only the HUs of uni-feature and bi-feature neurons (Fig 4H, for density estimation the bandwidths from Fig 4E were used), we observe that the median EC values of the 36 excitatory and 32 inhibitory HUs differ significantly ($p = 4.8\times10^{-5}$, rank sum test), as is the case for all HUs (Fig 4F). If we restrict the EC distribution analysis to multi-feature neurons alone we again observe the same pattern ($p = 2.4\times10^{-3}$, rank sum test) as for all HUs (Fig 4I, again the Fig 4E bandwidths were used). However, the 36 excitatory HUs for the uni-feature and bi-feature neurons are significantly ($p = 1.7\times10^{-6}$, Levene's test) more tightly clustered than the 95 excitatory HUs of multi-feature neurons, indicating that the uni- and bi-feature neurons show less diversity in their use of the nonlinear activation function. The decrease in diversity for the 32 inhibitory HUs of uni- and bi-feature neurons relative to the 83 inhibitory HUs of the multi-feature neurons is also significant ($p = 0.032$, Levene's test).

Potential functional role of nonlinear characteristics

We have seen that NRF models capture more of the response properties of auditory cortical neurons than conventional LN models, and that they achieve this through the interplay of modest numbers of excitatory and inhibitory HUs. In this section, we consider which functional properties of cortical neurons might be captured by the NRF models. We identify two such properties: gain control and multi-feature sensitivity.

Gain control. We can use simulations derived from our modeling to show that the interplay of excitation and inhibition seen in the NRF models enables them to exhibit gain control. For an example bi-feature neuron (neuron 2 of Fig 3), we plotted the OU firing rate as a function of the activation of the excitatory HU (Fig 5A). Using different hues from red to magenta, this dependence of OU firing rate on excitatory HU activation is shown for 7 different levels of inhibitory HU activation, which were also chosen to span the range of the inhibitory HU activation. Observe that increasing the inhibitory drive reduces the slope of the relationship between excitatory drive and OU firing rate. In other words, the inhibitory input reduces the gain of the excitatory drive on the OU: the effect of the inhibition is more "divisive" than "subtractive". This form of gain control is not observed if we feed the inhibition into the excitatory HU instead of the OU; instead a threshold shift, i.e. a "subtractive inhibition", is seen (Fig 5B). Gain control is also not observed if the HU output functions are linear rather than sigmoidal. We measure the gain as the steepest slope of the OU-firing-rate vs. excitatory-HU-activation curves (Fig 5A). For the bi-feature neurons the gain tends to decrease with increasing inhibitory HU activation (Fig 5C), with the gain for 31/32 neurons being lower for the highest inhibitory HU activation than it is for the lowest inhibitory HU activation.

Multi-feature selectivity. We have seen that NRFs can reveal multiple excitatory fields of several HUs arranged closely together, but distinct, in time-frequency space (Fig 3C). This raises the question of what advantage parsing these excitatory regions out over several HUs rather than just combining them in a single STRF, as in an LN model, would bring. To address this, we have plotted the activity of the components of the NRF of one multi-feature neuron (neuron 5 in Fig 3) during 800 ms of test-set auditory stimulation. Plotted are the OU output (Fig 5D) and effective HU weighted-outputs (excitatory HUs, Fig 5E, inhibitory HUs, Fig 5F). Note that the weighted-outputs of the excitatory HUs are often correlated, and that the OU tends to give a substantial response only when the weighted-output of more than one excitatory HU peaks at the same time. For example, at 4.2 s (Fig 5E, left black box), one of the excitatory HUs is highly active, but the other two are not, and the OU gives little response, while at 4.75 s (Fig 5E, right black box) two HUs are active and the OU response is ~6 times higher.

Given that the output and excitatory HUs tend to operate over the expansive, near-threshold range of the nonlinear activation function, one might expect a conjunctive effect to be

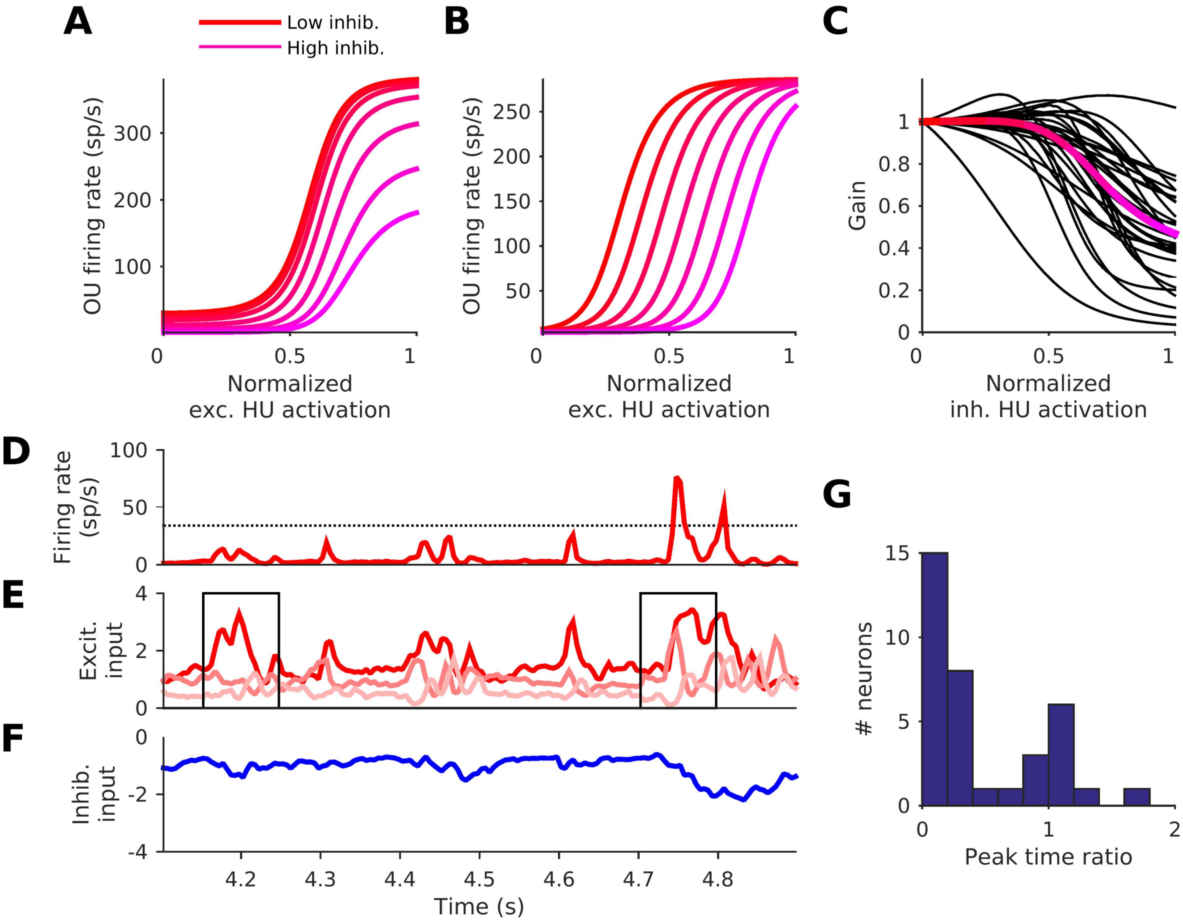


Fig 5. Functional implications of the neural network model fits. (A) The effect of activation of the inhibitory hidden unit (HU) on the relationship between output unit (OU) firing rate and excitatory HU activation, for an example bi-feature neuron's NRF. The steepest slope of each curve is its gain. Activation in all plots is normalized to span 0–1, where 0 is the 1st centile and 1 the 99th centile of the distribution of activations over the stimulus set (Fig 3C). (B) The effect of inhibiting the excitatory HU (instead of the OU) on the same relationship for the same example NRF. (C) The gain as a function of inhibitory HU activation for all 32 bi-feature neurons (black lines). For each neuron the gain is normalized to be 1 when the normalized inhibitory HU activation is 0. Red/magenta line: the mean. Note that the full range of excitatory HU activations, from threshold to saturation of the HU, was examined to find the steepest slopes and hence the gain (i.e. beyond the 1st and 99th centiles). (D) The OU firing rate of an NRF model fit for an example multi-feature neuron (red line). The dotted line is the 2σ-threshold. (E) The weighted-output (the input to the OU, HU output × output weight) from the 3 effective excitatory HUs. (F) The weighted-output from the single effective inhibitory HU for this neuron. (G) The distribution of peak time ratio for all multi-feature neurons. The peak time ratio is the number of times the sum of the outputs of all reduced-NRF models exceeded the 2σ-threshold, relative to the number of times the output of the full NRF did so. The reduced-NRF models of a neuron each retain just one of the excitatory HUs.

doi:10.1371/journal.pcbi.1005113.g005

common, whereby the OU "goes substantially above threshold" only when several features of the excitatory HUs occur together. To examine how single excitatory HUs on their own can drive the OU, we ran the natural sounds through the NRF model with all but one of its excitatory HUs disabled (set below 'threshold'). We did this for each excitatory HU in turn to obtain the response of the OU if it only had that one excitatory HU. Then, to provide a conservative comparison with the original model, we summed the OU response for all of those single excitatory HU reduced-NRF models (for example, a 3 excitatory HU NRF would produce 3 single excitatory HU reduced-NRF models, whose responses were summed). We then determined the number of time bins for which this summed response was above the 2σ-threshold (Fig 5D, as in Fig 2B), and called this number the summed-reduced-NRF peak time. We compared this to the number of time bins during which the response of the original model exceeded the 2σ-threshold; we called this number the NRF peak time. We did this for all 36 neurons with more than one excitatory HU. Across the 36 neurons, the NRF peak time typically (for 78%, 28/36, of the neurons) exceeded the summed-reduced-NRF peak time ($p = 1.2\times10^{-3}$, sign test). Fig 5G shows the distribution over the neurons of the peak time ratio, the neuron's summed-reduced-NRF peak time (the number of times the summed single excitatory HU response was above the 2σ-threshold) divided by its NRF peak time (the number of times the 2σ-threshold was surpassed by the unmodified model). Here we can see the strength of the conjunctive effect, the summed single excitatory HU response exceeded the 2σ-threshold less than half as often as the unmodified model alone (peak time ratio < 0.5) for 67% (24/36) of the multi-feature neurons. This implies that, for many of the multi-feature neurons, spectrotemporal features often interact in a conjunctive, supra-additive manner. Such neurons require the simultaneous presence of multiple particular features to produce a substantial response, and respond very little to just one such feature alone. This need not have been the case, as it could have been that each feature alone could substantially drive the neuron, as is seen for 22% (8/36) of the neurons, which have a peak time ratio $\geq$1.

Discussion

The network models we developed here represent a substantial improvement over conventional LN models in that they are able to produce more accurate predictions of the responses of cortical neurons to natural sounds, while remaining sufficiently parsimonious that they it can be quickly fitted using limited data and interpreted in a manner relevant to the known physiology of the auditory pathway. The NRF models are significantly more complex than LN models–in fact, their number of degrees of freedom is greater in proportion to the number of HUs in the network. Nevertheless, compared to the enormous complexity of the lemniscal auditory pathway, in which individual neurons receive potentially thousands of converging inputs, the model complexity remains very modest. Accordingly, and as with all models of cortical processing, we cannot expect the artificial neural network to replicate the biological network in any strict anatomical detail. Instead, given their capacity to predict the responses of auditory cortical neurons to natural sounds, we propose that the NRF models capture important aspects of the general signal processing performed by the neural circuitry driving the recorded neurons, and that this is likely to apply to other areas of the brain too.

Primary auditory cortical neurons are nonlinearly sensitive to a broad spectrotemporal domain

Our results indicate that auditory cortical neurons likely integrate more widely over time and frequency than linear STRF or LN models would suggest (Fig 4C and 4D). That this integration is highly nonlinear may be the reason why linear STRFs do not effectively measure this broad tuning. Although they can be quite complex [12], LN model cortical STRFs tend to be relatively simple in structure [13]. Given the extensive network that constitutes the central auditory pathway, it seems likely that more sophisticated processing is being carried out than implied by linear STRFs. NRF models may help to shed light on the nature of this processing, as they reveal a diversity, complexity and breadth of spectrotemporal integration well beyond that which can be described by conventional LN models.

One consequence of this finding relates to models of sparse representation of natural sounds [33], a hypothesized method by which the brain may perform unsupervised learning of the statistical structure of the environment. These sparse models result in projective fields (for many parameter settings) that are often broad in frequency and particularly in time. Our results suggest that many neurons with punctate STRFs may be better described by NRF models with broader tuning, which is more consistent with these sparse representational models.

The neurons are well characterized by 1–7 features that segregate into inhibitory and excitatory features

Although we found that many of our cortical neurons can be characterized as bi-feature (one inhibitory and one excitatory HU), we also found many multi-feature neurons, with 3–7 effective HUs, typically with slightly more excitatory HUs than inhibitory (Fig 4A and 4B). It is interesting to speculate that the bi-feature neurons may mostly be located in granular cortical layers and the multi-feature neurons in the supra/infragranular layers, since the former receive most of the thalamic inputs and are known to show simpler tuning properties than neurons in the supra/infragranular layers [34].

The features (HU STRFs) naturally segregate into those that inhibit the neuron and those that excite it (Fig 4E). The excitatory features tend to operate in the expansive part of the NRF model's nonlinear activation function, 'near threshold', whereas the inhibitory features tend to operate in the more linear part of the nonlinear activation function (Fig 4F). We speculate that the excitatory and inhibitory HUs may reflect the massed effects on the recorded neuron of directly connected excitatory neurons and inhibitory neurons, respectively. Should this be the case, the difference in nonlinear characteristics may reflect the observation that inhibitory neurons tend to have higher evoked and spontaneous firing rates than excitatory neurons [35], thus placing inhibitory inputs further above threshold than excitatory inputs and perhaps providing them with a more linear dependence on input.

Bi-feature neurons

Many (42%) of the neurons showed NRF fits with just two effective HUs, one excitatory and one inhibitory. The excitatory HU operates near threshold (expansive), and the inhibitory HU is more linear. The OU is also expansive. Under this arrangement, the inhibition acts on the output in a manner that appears to decrease the gain (Fig 5A–5C). A number of possible mechanisms, which might work in isolation or together, have been proposed for gain control, including synaptic depression [36], shunting inhibition [37] and recurrent connectivity [38]. The NRF model illustrates another possible mechanism—feedforward expansive excitation and feedforward linear inhibition acting together on a neuron with an expansive nonlinearity. In vivo patching approaches may provide a method to explore this possibility, since it may be possible to measure the inhibition and excitation separately, as well as assess the output nonlinearity. The above discussion prompts two modifications to the model to be examined in future work. The first is to include some explicit gain control mechanism, for example, HUs with a divisive effect as a functional model of shunting inhibition. The second is to add an additional layer, which will allow for HUs to depend more directly on nonlinear measures like the standard deviation of the stimulus and perhaps capture the use of gain control to normalize for contrast, as has been observed for auditory neurons with artificial stimuli [19,39–41].

Multi-feature neurons

The multi-feature neurons are quite a diverse group, and substantially larger population samples would therefore be needed to look for trends in their properties and investigate whether they form identifiable groups that may serve distinct purposes. However, we can make a number of observations. Although the HUs of multi-feature neurons show more diverse nonlinearity characteristics than bi-feature neurons, they still tend towards having expansive-range excitatory HUs and linear-range inhibitory HUs. The set of STRFs of multi-feature neurons can be quite complex and varied (Fig 3B), and can show distinctly structured relationships between these STRFs (e.g. neurons 7 and 8, Fig 3C). Often the spectrotemporal regions of high sensitivity (half-height tuning area) of HU STRFs do not substantially overlap (e.g. neurons 1, 4 and 9, Fig 3C). However, some overlap of high sensitivity regions in the STRFs can occur, between excitatory HUs (e.g. neurons 5–8, Fig 3C), between inhibitory HUs (neuron 7, Fig 3C), and between excitatory and inhibitory HUs (neurons 6, 7, and 10, Fig 3C).

Given that the model fitting process penalizes redundant STRF weights, the presence of spectrotemporal overlap in STRFs of different HUs may indicate that the NRF is using multiple HUs to alter the nonlinearity of the input-output mapping in order to achieve a better fit to the true output nonlinearity of the biological neuron. However, for the most part, the high sensitivity regions of HU STRFs are non-overlapping, suggesting that additional factors drive the diversity of multi-feature neuron STRFs. In a number of cases, excitatory fields of different HUs align consecutively along the time axis, sometimes with some overlap (e. g. neurons 4, 6, 7 and 10, Fig 3C). This may to some extent capture the relationship between sound intensity and response latency found in both the auditory nerve and the cortex [42], as in some cases the shorter latency HU also has a higher 'threshold' (i.e. has a lower EC value, e.g. neuron 4, Fig 3C). However, this is unlikely to be the whole story, because in other cases different excitatory HUs exhibit distinct well-separated regions of temporal tuning (e.g. neuron 10, Fig 3C). In addition, STRF excitatory fields may also align over the frequency axis (e.g. neurons 1, 5 and 8, Fig 3C) or align diagonally over time and frequency (e.g. neurons 6 and 9, Fig 3C).

For many multi-feature neurons (although far from all), the NRF requires that multiple excitatory HUs are activated simultaneously to produce a substantial response (Fig 4D–4G). That the NRF model can capture such supra-additive sensitivity to particular conjunctions of multiple spectrotemporal features, while the LN model cannot, may explain why the NRF model is better able to predict the peak amplitudes of the responses of cortical neurons. This conjunctive feature selectivity allows for increased selectivity for particular complex spectrotemporal patterns consisting of a number of more basic features, a characteristic with an obvious potential role in sound recognition.

Related work

A number of methods have been used previously to examine the spectrotemporal sensitivity of auditory cortical neurons. Previous studies have attempted to extend the application of the LN model to auditory cortical data, mostly using maximum-likelihood methods. Indeed, several studies have used approaches that have fundamental similarities to the one we explore here, in that they combine or cascade several linear filters in a nonlinear manner. One such body of work that improved predictions over the LN model is based on finding the maximally-informative dimensions (MID) [20,21,34,43–46] that drove the response of auditory cortical neurons. This method involves finding usually one or two maximally informative linear features that interact through a flexible 1D or 2D nonlinearity, and is equivalent to fitting a form of LN model under assumptions of a Poisson model of spiking variability [46–48]. When this method was applied to neurons in primary auditory cortex it was found that the neurons' response properties are typically better described using two features rather than one [20,34], in contrast to midbrain neurons which are well fitted using a single feature [43]. That result thus seems consistent with ours, in that we found NRFs fitted to cortical responses most commonly evolved to have two effective HUs (or input features). Another approach, that has been found to improve predictions of auditory cortical responses, is to apply a multi-linear model over the dimensions of frequency, sound level, and time lag, and for the extended multi-linear model also over dimensions involved in multiplicative contextual effects [21]. However, the above studies in auditory cortex [20,21,34,43] did not use natural stimuli, and hence might not have been in the right stimulus space to observe some complexities, as STRFs measured with natural stimuli can be quite different than when measured with artificial stimuli [49]. An advantage of the NRF model is that its architecture is entirely that of traditional feedforward models of sensory pathways in which activations of lower level features simply converge onto model neurons with sigmoidal input-firing rate functions. NRFs can therefore be interpreted in a context that is perhaps simpler and more familiar than that of, for example, maximally informative dimension models [20,44].

Other developments on the standard LN model have included model components that can be interpreted as intraneuronal rather than network properties, such as including a post-spike filter [22] or synaptic depression [23], and have also been shown to improve predictions. Pillow and colleagues [50,51] applied a generalized linear model (GLM) to the problem of receptive field modelling. Their approach is similar to the basic LN model in that it involves a linear function of stimulus history combined with an output nonlinearity. However, unlike in LN models, the response of their GLM also depends on the spike history (using a post-spike filter). This post-spike filter may reflect intrinsic refractory characteristics of neurons, but could also represent network filter effects. A GLM model has been applied to avian forebrain neurons [22], where it has been shown to significantly improve predictions of neural responses over a linear model, but not over an LN model.

Although they haven't yet been applied to auditory cortical responses, it is worth mentioning two extensions to GLMs. First, GLMs can be extended so that model responses depend on the history of many recorded neurons [50], representing interconnections between recorded neurons. While this approach is thus also aimed at modeling network properties, it is quite different from our NRF model, where we infer the characteristics of hidden units. Second, the extension of the GLM approach investigated by Park and colleagues [52] included sensitivity to more than one stimulus feature. Thus, like our NRF or the multi-feature MID approach, this "generalized quadratic model" (GQM) has an input stage comprising several filters which are nonlinearly combined, in this case using a quadratic function. One might argue that our choice for the HUs of a sigmoidal nonlinearity following a linear filter stage, and the same form for the OU, is perhaps more similar to what occurs in the brain, where dendritic currents might be thought of as combining linearly according to Kirchhoff's laws as they converge on neurons that often have sigmoidal current-firing rate functions. However, we do not wish to overstate either the physiological realism of our model (which is very rudimentary compared to the known complexity of real neurons) or the conceptual difference with GQMs or multi-feature MIDs. A summation of sigmoidal unit outputs may perhaps be better motivated physiologically than a quadratic function, but given the diversity of nonlinearity in the brain this is a debatable point.

Another extension to GLMs, a generalized nonlinear model (GNM), does, however, employ input units with monotonically-increasing nonlinearities, and unlike multi-neuron GLMs or GQMs, GNMs have been applied to auditory neurons by Schinkel-Bielefeld and colleagues [24]. Their GNM comprises a very simple feedforward network based on the weighted sum of an excitatory and an inhibitory unit, along with a post-spike filter. The architecture of that model is thus not dissimilar from our NRFs, except that the number of HUs is fixed at two, and their inhibitory and excitatory influences are fixed in advance. It has been applied to mammalian (ferret) cortical neural responses, uncovering non-monotonic sound intensity tuning and onset/offset selectivity.

For neurons in the avian auditory forebrain, although not for mammalian auditory cortex, GNMs have also been extended by McFarland and colleagues to include the sum of more than two input units with monotonically-increasing nonlinearities [53]. Of the previously described models, this cascaded LN-LN 'Nonlinear Input Model (NIM)' model bears perhaps the greatest similarity with our NRF model. Just like our NRF, it comprises a collection of nonlinear units feeding into a nonlinear unit. The main differences between their model and ours thus pertain not to model architecture, but to the methods of fitting the models and the extent to which the models have been characterized. The NIM has been applied to a single zebra finch auditory forebrain neuron, separating out its excitatory and inhibitory receptive fields in a manner similar to what we observe in the bi-feature neurons described above.

One advantage of the NRF over the NIM is that the fitting algorithm automatically determines the number of features that parsimoniously explain each neuron's response, obviating the need to laboriously compare the cross-validated model performance for each possible number of hidden units. Another difference is that the NRF is simpler while still maintaining the capacity to capture complex nonlinear network properties of neural responses; for example, the NIM [53] had potentially large numbers of hyperparameters (four for each hidden unit or "feature") that were manually turned, something that would be very difficult to do if the model needed to be fitted to datasets comprising large numbers of neurons. In contrast, the NRF has only one hyperparameter for the entire network, which can easily be tuned in an automated parameter search with cross-validation. Consequently, we have been able to use the NRF to characterize a sizeable population of recorded neurons, but so far no systematic examination of the capacity of the NIM to explain the responses of many neurons has been performed.

Another recent avian forebrain study [54] used a maximum noise entropy (MNE) approach to uncover multiple receptive fields sensitive to second-order aspects of the stimulus. Unlike the above two GNM [24,53] approaches, this model does not have hidden units with sigmoidal nonlinearities, but finds multiple quadratic features. The MNE predicted neural responses better than a linear model, although still poorly, with an average CC_{raw} of 0.24, and it was not determined whether it could out-predict an LN model. Note, however, that the CC_{raw} values reported in that study do not distinguish stimulus-driven response variability from neural "noise". Consequently, it is unclear whether the relatively modest CC_{raw} values reported there might reflect shortcomings of the model or whether they are a consequence of differences in the species, brain regions and stimuli under study. Finally, perhaps the most relevant study in the avian forebrain used a time delay feedforward neural network to predict responses of zebra finch nucleus ovoidalis neurons to birdsong [55]. These authors reported that the network predicted neural responses better than a linear model, but performed no quantitative comparisons to support this.

Advances on the LN model have also been applied in other brain regions. Various advances on the LN model have also been made in studies of primary visual cortex, and of particular relevance are the few cases where neural networks have been used to predict neural responses. Visual cortical responses to certain artificial stimuli (randomly varying bar patterns and related stimuli) have been fitted using a single hidden layer neural network, resulting in improvements in prediction over linear models for complex but not simple cells in one study [56] and over LN-like models in another study [57]. However, the challenge we tackle here is to predict the responses to natural stimuli. In this respect we are aware of only one similar study by Prenger and colleagues [58] which used a single hidden layer neural network to predict responses to series of still images of natural scenes. The network model in this study gave better predictions than an LN model with a simple rectifying nonlinearity. However, the improvements had limited consistency, predicting significantly better in only 16/34 neurons, and it did worse than an LN model applied to the power spectra of the images. Additionally, the CC_{raw} of the model predictions with the neural data were somewhat small (0.24). This appears to contrast with the seemingly better performance we obtained with our NRF model.

These apparent differences in model performance may, however, not all be attributable to differences in model design or fitting. In addition to the fact we already noted that low CC_{raw}

values might be diagnostic of very noisy neurons rather than shortcomings of the model, we also need to be cognizant of the differences in the types of data that are being modeled: we applied our model responses of auditory cortical neurons to natural auditory sound recordings, whereas Prenger and colleagues [58] applied theirs to visual cortical neuron responses to random sequences of photographs of natural scenes. Furthermore, the neural responses to our stimuli were averaged over several repeats, whereas the above study did not use repeated stimuli, which may limit how predictable their neural responses may be. However, there are also notable structural differences between their model and ours. For example, the activation function on the OU in the Prenger et al. study [58] was linear (as with [56] but not [57]), whereas the OU of our NRF has a nonlinear activation function, which enables our NRF to model observed neuronal thresholds explicitly. Furthermore, we used a notably powerful optimization algorithm, the sum-of-function optimizer [26], which has been shown to find substantially lower values of neural network cost function than the forms of gradient descent used in the above neural network studies. Finally, the L_1-norm regularization that we used has the advantage of finding a parsimonious network quickly and simply, as compared with the more laborious and often more complex methods of the above three studies: L_2-norm-based regularization methods and hidden unit pruning [58], early stopping and post-fit pruning [56] or no regularization and comparing different numbers of hidden units [57].

Predictive capacity and possible model improvements

The predictions of the NRF models correlate with the observed neural responses with a CC_{norm} of 0.73 on average. Asari and Zador [31] estimated an upper limit on the performance that any model of A1 neurons might be able to achieve in predicting responses from a given duration of stimulus history. Our models predict responses from the last 100 ms of stimulus history, for which Asari and Zador [31] give an upper performance limit of 0.5–0.55 “signal power explained” (SPE). For SPE values in this range, SPE is approximately equal to the square of CC_{norm} [59], so that an upper limit SPE of 0.5–0.55 corresponds to an upper limit CC_{norm} of 0.71–0.74. This suggests that the NRF may possibly be capturing the majority of the neural response that is dependent on the stimulus, given the duration of stimulus history provided (100 ms).

The performance upper bound reaches its maximal plateau when about 3 s of stimulus history are provided [31]. This suggests that the most important way of advancing neural network models of auditory cortex might be to include a substantially longer stimulus history in the analysis. However, simply extending the number of time bins in the current model some 30 fold further into the past would likely lead to far too many free parameters. A better option might be to extend the approach presented here in the direction of convolutional or recurrent neural networks. Artificial recurrent neural networks have been applied successfully to sound recognition problems [60], and it is well known that feedback projections are common features of the auditory pathway. Developing recurrent versions of the NRFs introduced here is therefore likely to be important, particularly if we hope to develop successful models of higher order auditory cortical neurons.

Conclusions

In summary, we have shown that fitting feedforward network models (with regularization of the weights to be sparse) to single neuron activity in primary cortical areas (A1/AAF) allows for better predictions of their responses to natural sounds, and has the potential to unmask some of the nonlinear signal processing strategies used by the auditory brain. This approach reveals more of the underlying richness and nonlinearity of cortical processing in an easily interpretable form. Neural responses to natural sounds in A1/AAF appear to be dependent on multiple features in the stimulus space that often interact in structured nonlinear ways, and depend upon a substantially larger spectrotemporal domain than is suggested by linear models with a simple output nonlinearity.

Materials and Methods

Electrophysiological recording

To assess the capacity of NRFs to account for cortical sensory responses, we fitted models to neural responses to clips of natural sounds. Single-unit responses were recorded with multi-channel electrodes in the ferret primary auditory cortex (A1) and the anterior auditory field (AAF), which are both considered to be primary cortical areas [61]. All animal procedures were performed under license from the United Kingdom Home Office and were approved by the local ethical review committee. For full details of the recording procedures see [62]. In brief, electrophysiological recordings were made from 6 adult pigmented ferrets under ketamine (5 mg/kg/h) and medetomidine (0.022 mg/kg/h) anesthesia. Bilateral extracellular recordings were made in A1/AAF using either 16 or 32 channel silicon probe electrodes (Neuronexus Technologies). Because these primary cortical fields share a common tonotopic gradient [61,63], we did not attempt to assign our sample of 76 units to one or other of these regions.

Stimuli

In this study we modeled the responses of neurons to 20 clips of natural sound recordings. Each clip was 5 s long, and presented at a sampling rate of 48,828.125 Hz, using earphones as described by [19]. The clips were presented in random order, with a ~1 s silent interval between clips, and were repeated 20 times. The natural sound recordings included animal sounds (e.g. ferret vocalization and birdsong), environmental sounds (e.g. water and wind), and speech. The RMS intensity of clips ranged from 75 to 82 dB SPL. Data recorded during the first 250 ms after the onset of each stimulus were discarded, leaving an effective set of neural responses to 20 repeats of 20 sounds of 4.75 s duration each.

Preprocessing of neural data and stimuli

NRF and LN models were fitted to the relationship between the neural data and the sound stimuli, after appropriate preprocessing as described below.

Neural data. Recorded spikes were sorted offline using Spikemonger, in-house software built around Klustakwik [64], to isolate single units. For each neuron, for each clip, peri-stimulus time histograms (PSTHs) were constructed, counting spikes in 5 ms bins, averaging over all 20 repeats, and subsequently smoothing with a 21 ms wide Hanning window [29] to estimate the spike count PSTH $y_n(t_n)$ for each neuron at time t_n, where t_n is the time since the start of clip n (n goes from 1 to $N = 20$, t_n from 1 to $T_n = 949$). For fitting the NRF model the spike counts were also linearly rescaled to span the standard network nonlinear activation function (see below), so spike count 0 mapped to $-\sigma_1$ and spike count 1 to $+\sigma_1$, where $\sigma_1 = 1.7159$. For model comparison, all spike counts were rescaled back, and for display all spike counts were rescaled to spike rates. To identify those neurons that were driven by the stimuli, we calculated a "noise ratio" (NR) statistic for each neuron [19,65] and excluded from further analysis any neurons with a NR>40.

Cochleagram. To transform the sound stimuli into a simple approximation of the activity pattern received by the auditory pathway, we processed the sound waveforms to calculate log-scaled spectrograms ('cochleagrams'). For each sound, the power spectrogram was taken using 10 ms Hamming windows, overlapping by 5 ms. The power across neighboring Fourier frequency components was then aggregated using overlapping triangular windows comprising 34 frequency channels with center frequencies ranging from 500 Hz to 22,627 Hz (⅙ octave spacing). Next, the log was taken of the power in each time-frequency bin, and finally any values below a low threshold were set to that threshold. These calculations were performed using code adapted from melbank.m (http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html). Both the LN and the NRF models were trained to predict the firing rate $y_n(t_n)$ at time t_n from a snippet of the cochleagram extending 100 ms (20 time bins) back in time from t_n. The input to the models at time t_n is thus a 34×20 matrix (F = 34 frequency channels by H = 20 stimulus history time bins) of log sound power values preceding time t_n. We denote this as $x_{nf\tau}(t_n)$, where n is the index of the presented clip, f indexes the frequency bands, and τ indexes time history bins preceding time t_n. For fitting the NRF model, $x_{nf\tau}(t_n)$ was also normalized so the whole dataset had zero mean ($<x_{nf\tau}(t_n)>_{nf\tau} = 0$) and unit variance. To simplify notation we define t as all the times t_n of all the sound clips n, where t goes from 1 to $T_n \times N$. This gives $y(t)$ and $x_{f\tau}(t)$.

LN model

Linear stage. The LN model (Fig 1A) consists of two stages: a linear STRF followed by a sigmoidal output nonlinearity. The linear part of the model is:

$$\hat{a}(t) = \sum_{f,\tau} w_{f\tau} x_{f\tau}(t) + b$$

where $\hat{a}(t)$ is the model neuron's "activation", and $w_{f\tau}$ is the synaptic weight for frequency band f and history bin τ (all the weights compose the STRF). The bias b represents the neuron's background activity level. $w_{f\tau}$ and b are the free parameters of the model, and were estimated by regressing $y(t)$ against $xf\tau(t)$ using 'glmnet' [66]. Thus $\hat{a}(t)$ can be seen as the best linear prediction from $x_{f\tau}(t)$ of $y(t)$. To avoid overfitting and to find a parsimonious model, the regression was regularized by penalizing the L_1-norm of $w_{f\tau}$ (LASSO regression). The strength of the regularization was controlled with a hyperparameter λ. The optimum value of λ was found using k-fold cross-validation for a set of log-spaced values and for each neuron, and the λ that gave the best prediction was chosen (see *Training, validation, and testing of models* below). The resulting $\hat{a}(t)$ serves as the input to the nonlinear stage for our LN-model, and as the linear prediction (output) of the purely linear L-model used for the model comparisons which are described in Results.

Nonlinear stage. The second stage involved fitting a logistic sigmoid nonlinear activation function,

$$\hat{y}(t) = \frac{\rho_1}{1 + \exp(-(\hat{a}(t) - \rho_3)/\rho_2)} + \rho_4$$

which mapped the linear activation $\hat{a}(t)$ to the predicted PSTH $\hat{y}(t)$ so as to minimize the error between the predicted PSTH and the observed PSTH $y(t)$. Recent work [67] indicates that choosing different nonlinear output functions from a wide range of plausible candidates has only modest effects on the ability of LN models to capture neural response properties. We therefore did not attempt to systematically explore different types of output nonlinearity or to make the choice of nonlinearity as physiological as possible, but rather focused on an output nonlinearity that is simple, well characterized and widely used in the artificial network literature. The four parameters ρ_i of the function were fitted by minimizing the squared error

$$E = \sum_t (\hat{y}(t) - y(t))^2$$

using a quasi-Newton iterative numerical method (http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html).

NRF model

Model description. NRFs (Fig 1B) model cortical responses using a rate based feedforward artificial neural network (multilayer perceptron) with one hidden layer of $J = 20$ hidden units (HU) converging onto a single output unit (OU). Each unit in the network operates in a fashion similar to an LN model—each unit integrates inputs through a set of linear weights, and this linear activation is passed through a nonlinear activation function to compute its output. The activation of the j-th HU $a_j(t)$ is,

$$a_j(t) = \sum_{f,\tau} w_{jf\tau} x_{f\tau}(t) + b_j$$

where $w_{jf\tau}$ is the weight from frequency band f and time delay τ to HU j, and b_j is the bias on the HU. The output of the HU is $z_j(t)$, given by,

$$z_j(t) = \mathrm{g}(a_j(t))$$

where $\mathrm{g}(\zeta)$ is a nonlinear function. The OU then provides the prediction $\hat{y}(t)$, of the firing rate $y(t)$, as a weighted sum of the HU outputs, also passed through the nonlinear activation function. The activation $a_o(t)$ of the OU is,

$$a_o(t) = \sum_j w_j z_j(t) + b_o$$

where w_j is the weight from HU j to the OU, and b_o is the bias on the OU. The output $\hat{y}(t)$ of the OU is;

$$\hat{y}(t) = \mathrm{g}(a_o(t))$$

which is the model's prediction of the rescaled firing rate (see *Preprocessing of neural data and stimuli*). For both HUs and OUs, the nonlinear activation function $\mathrm{g}(\zeta)$ was a hyperbolic tangent function:

$$\mathrm{g}(\zeta) = \rho_1 \tanh(\zeta/\rho_2)$$

In the LN model, the parameters σ_i of the nonlinear activation function were optimized for each neuron, but in the NRF model these parameters were fixed to $\rho_1 = 1/\tanh(2/3) \approx 1.7159$ and $\rho_2 = 3/2$, which ensures that $\mathrm{g}(\pm 1) = \pm 1$. Using this particular form of nonlinear activation function [68] facilitates efficient learning with error backpropagation by maintaining statistical properties of the input distribution. Furthermore, for a given network with tanh activation functions, there is an equivalent network with logistic activation functions (for which units have non-negative outputs), which can be found with a simple linear rescaling of the weights and biases (see page 109 of [69]). This rescaling does not affect the structure of the STRFs. We use this equivalent network for display in the Results (for details see *The adjusted network* below). Note that the nonlinearities $\mathrm{g}(\zeta)$ employed by the NRF and the LN models are equivalent except for a scaling and shifting.

Learning. The free parameters of the NRF, $w_{jf\tau+}$, b_j, w_j and b_o, were optimized by minimizing the following objective function:

$$E = \frac{1}{2}\sum_t (\hat{y}(t) - y(t))^2 + \lambda\left(\sum_{j,f,\tau} |w_{jf\tau}| + \sum_j |w_j|\right)$$

This objective function is the sum of two terms: The first term quantifies total square error between the observed PSTH $y(t)$ and the PSTH $\hat{y}(t)$ predicted by the model. The second term, proportional to the sum of the absolute values of all the weights in the network (the L_1-norm of the weight vectors), serves to regularize the weights. That is, it puts a "cost" on non-zero synaptic weights and will tend to drive most weights to close to zero, except for a few, and thereby encourages parsimonious models and prevents overfitting. The regularization was therefore similar to the LASSO regression used to fit the LN model, which also incorporates an L_1-norm regularization term.

For both the NRF and the LN models, the constant λ is the hyperparameter that determines the strength of the regularization. Its optimum value was determined using k-fold cross-validation (k = 10) over a log-spaced range, and for each model and neuron the value of λ that gave the best prediction for each neuron was chosen (see *Training, validation, and testing of models*). The NRF was initialized with the weights and biases independently drawn from a uniform distribution between $\pm 1/\sqrt{M}$ where M is the number of incoming connection weights and biases to a given unit of the network. The objective function of the NRF model was minimized using the Sum-of-Functions Optimizer, a recently developed algorithm which combines a Newton method with batch stochastic gradient descent, and which is substantially faster and finds lower minima than other optimization algorithms for multilayer feedforward networks [26]. The optimizer was run for 40 iterations, but usually settled within 20. On a desktop PC (Intel Xeon 8-core 3.1GHz CPU) it took on the order of hours to fit all 76 neurons, including the 10-fold cross validation.

Training, validation, and testing of the models

For both the LN models and the neural networks, the model parameters (weights w and biases b, and for the LN models also the parameters of the nonlinear activation function ρ_i) were found through the model fitting steps just described, but the models can only perform effectively if the model hyperparameters (regularization strength λ and, for the NRFs, also the number of HUs J) are appropriately chosen. We therefore conducted a parameter search which systematically explored the behavior of the models for a range of hyperparameters in a cross-validation test. To this end, the entire data set of 95 s duration (20 natural sound clips of 4.75 s duration each) was first split into a cross-validation set (80%) and a test set (20%). The test set was the last 20% (0.95 s) of each of the 20 sounds. The test set was put aside. The cross-validation set was then used to determine the hyperparameters by using k-fold cross-validation (k = 10). The cross-validation set was split into a training set (90% of the cross-validation set, that is the first 3.8s of 18 of the sounds) and a validation set (the remaining 10% of the cross-validation set, that is first 3.8s of 2 of the sounds).

The following steps were performed for each neuron and for each model. For a given λ, the model was first fitted on the training set, then the fitted model was used to predict the PSTH of the reserved validation set, and the prediction performance quantified by the normalized correlation coefficient (see *Performance measures*). This process was repeated $k = 10$ times, each time using a different non-overlapping 10% of the data as a validation set. The above process was performed for a log spaced set of λ values. Then the λ was chosen that maximized the mean prediction performance of the 10 validation sets. Optimum λ differed across neurons (for both LN and NRF models), and over the two models, and was thus set separately for each model and neuron. A similar process was also performed over J, the number of HUs, for a number of reasonable λ values. However, as NRF prediction performance varied little as a function of J, this was simply set to 20 for all neurons.

Then for each neuron, both models were re-fitted to the full cross-validation set, using the optimum λ values, and each model was used to predict the PSTH of the test set. The prediction performance of two fitted models was compared using the performance measures described below. These model fits are the ones used throughout the results section.

To verify that the model fits (at the best λ for each neuron) were consistent across the ten different cross-validation fits, we quantitatively compared the STRF of the effective HUs obtained for each validation set. For a given neuron, the effective HU from a given fit that was most correlated with the effective HU from a different fit was found on average to share a correlation coefficient of 0.82, while the second most correlated pair of HUs across fits shared a correlation coefficient of 0.69. These high correlation coefficients are indicative of a high degree of consistency. We verified that, in the absence of repeatable fits, one would expect these correlation coefficients to be close to zero by randomly permuting the weights within every effective HU STRF matrix. This randomization caused the correlation coefficients to drop to 0.06 and 0.02 respectively.

Performance measures

Model performance was quantified using three different performance measures: the normalized correlation coefficient CC_{norm}, the mean squared error MSE, and the peak activity mean square error $pMSE$. While the MSE is a well known quantifier of "goodness of fit", the other two require further explanation.

Normalized correlation coefficient. CC_{norm} quantifies model performance relative to a theoretically achievable maximum and independently of physiological noise. We use it as our standard performance measure in this paper, as it has a number of desirable properties [59], including the fact that it discounts the intrinsic noise of neural responses and quantifies the proportion of the stimulus driven response variability that is captured by the model. If the (Pearson's) correlation coefficient CC_{raw} between observed and predicted responses is low, then this could either indicate that the model is poor, or that the firing of the neuron under study is poorly stimulus driven and thus fundamentally quite unpredictable by a model that relies on stimulus history as the only explanatory variable. CC_{norm} does not have that shortcoming, and thus provides a more objective measure of model performance. We calculated the CC_{norm} [29,30] as the ratio of the CC_{raw} between the model's predictions $\hat{y}(t)$ and the real PSTH $y(t)$, over the maximum correlation coefficient CC_{max} that is achievable by a perfect model, given the inherent variability of a particular set of neural responses:

$$CC_{norm} = \frac{CC_{raw}}{CC_{max}}$$

CC_{max} is defined as the correlation coefficient between the PSTH of the recorded dataset constructed from the R repeats of the stimulus (here: $R = 20$) and the PSTH for an infinite number of repeats. CC_{max} cannot be measured directly, but, one can compute good estimates [29,30] of CC_{max} using the formula:

$$CC_{max} = \sqrt{\frac{2}{1 + \frac{1}{CC_{half}}}}$$

Here CC_{half} is the correlation coefficient of the mean PSTH of $R/2$ repeats with the mean PSTH of the remaining $R/2$ repeats. CC_{half} depends on the particular split of the R observations, and in order to minimize error the splitting is repeated many times and the values of CC_{half} are averaged. We took the average CC_{half} over a randomly chosen 126 combinations.

Peak activity mean square error. The $pMSE$ is the MSE between the predicted and observed PSTH for those parts of the signal where the observed firing rate is above the "2σ-

threshold", defined as two standard deviations above the mean firing rate to the sound. That is:

$$pMSE = \frac{1}{\sum_t p(t)} \sum_t p(t)(\hat{y}(t) - y(t))^2$$

where $p(t)$ is a binary window function consisting of all the binary window functions $p_n(t_n)$ for each clip n. The binary window functions isolate the peaks of the signal:

$$p_n(t_n) = \left\{ \begin{array}{l} 1 \text{ if } y_n(t_n) \geq \mu_n + 2\sigma_n \\ 0 \text{ if } y_n(t_n) < \mu_n + 2\sigma_n \end{array} \right\}$$

where μ_n is the average firing rate and σ_n its standard deviation for sound clip n. See *Preprocessing of neural data and stimuli* for how t relates to t_n, put briefly, we define t as all the times t_n of all the sound clips n.

Quantifying model properties

The adjusted network. While a biological neuron can only produce positive firing rate outputs and its 'synaptic output weights' are either only excitatory or only inhibitory, the non-linear activation function of the NRF model allows both positive and negative outputs and is symmetric around zero. Likewise, with the NRF the same weight can be either positive or negative. While these aspects of the NRF model thus lack biological realism, they do offer distinct practical advantages. First, there is a considerable literature on how to train this type of multi-layer perceptron model efficiently [28,68]. Second, it gives a network the freedom to discover during training how many excitatory or inhibitory neurons it requires, obviating the need to stipulate a fixed set of excitatory HUs and a fixed set of inhibitory HUs from the outset. However, the fact that the output and weights of a NRF unit can span both positive and negative values makes the distinction between excitatory and inhibitory neurons less categorical and somewhat ambiguous. Nevertheless, this can be resolved because an equivalent 'adjusted' network of excitatory and inhibitory units with logistic activation functions (and hence non-negative outputs) can be found, thus overcoming this problem [69]. Hence we can take advantage of the ease of training tanh networks while preserving the interpretability of logistic networks. This 'adjusted network' was used for all results.

In making the adjusted network, we first consider whether a HU is excitatory or inhibitory. The NRF unit output nonlinearities preserve the sign of the unit activation, so whether the j-th HU of an NRF has an overall inhibitory or excitatory effect on the OU will not only depend on the sign of the synaptic weight w_j that connects these two units, but also on the sign of the "expected activation" of the HU, which in turn depends on whether the HU's STRF is composed mostly of negative or positive weights. To give an extreme example, a HU with all negative STRF weights and a negative w_j would have a positive, excitatory influence on the output neuron. Whether a HU should be considered excitatory or inhibitory thus depends on the product of w_j and the sum of the weights of its STRF $\sum_{f,\tau} w_{jf\tau}$. If both are positive or both negative, the HU is excitatory, otherwise it is inhibitory. This is potentially confusing when readers are used to the idea that, whether a HU is inhibitory or not, can simply be determined from the sign of its output synaptic weight. Note, however, that the equation governing the NRF model unit's nonlinearity is symmetric and odd, so that for HU j, multiplying w_j, b_j, and $w_{jf\tau}$ by -1 leaves the influence of that HU on the OU completely unchanged. Consequently, we can ensure that all our inhibitory HUs do indeed have negative values for w_j and all excitatory HUs have positive ones by switching the signs of w_j, b_j, and $w_{jf\tau}$ in all those HUs for which $\sum_{f,\tau} w_{jf\tau}$ was found to be negative after training. In the interest of easier interpretation, that is what we did, producing the 'partially adjusted' NRF model.

Next let us consider how to further adjust the network to have non-negative outputs. For the partially adjusted NRF model, threshold is simply the most negative output value -ρ_1 (where weighted-output ranges from -$\rho_1 w_j$ to +$\rho_1 w_j$). The explanation for this is as follows: An excitatory HU (positive w_j) can equivalently be seen as a neuron with a positive-only output, whose weighted-output goes from 0 to +2$\rho_1 w_j$, acting on an OU whose resting state is less by -$\rho_1 w_j$. An inhibitory HU (negative w_j) can equivalently be seen as a neuron with a positive-only output, whose weighted-output goes from 0 to -2 $\rho_1 w_j$, acting on an OU whose resting state is greater by +$\rho_1 w_j$. Making these adjustments thus produces the final 'adjusted' network, which was used for all the results.

The inhibitory/excitatory score. As discussed, the HUs of the NRF model are not 'hard-wired' as excitatory or inhibitory, but each after fitting may nevertheless be “predominantly” excitatory or inhibitory in its influence on the OU. The sign of their output weight and the balance of positive and negative weights in the STRF will determine whether the HU is predominantly excitatory or inhibitory. To quantify the extent to which a HU is inhibitory or excitatory we calculated an inhibitory/excitatory (IE) score for each HU *j*:

$$IE = \mathrm{sign}(w_j)\frac{\sum_{f,\tau} w_{jf\tau}}{\sum_{f,\tau} |w_{jf\tau}|}$$

This *IE* score is bounded between -1 and 1. For a positive w_j, if all elements $w_{jf\tau}$ are non-negative (and at least one is not 0), IE = 1, if all elements are non-positive (and at least one is not 0), *IE* = -1. For a negative w_j, the opposite is true. If the sum of the negative elements equals the sum of the positive elements, *IE* = 0. This measure was used to investigate whether excitatory or inhibitory HUs play different functional roles in the NRF (Fig 4E).

The expansive/compressive score. The distributions of HU activation in the nonlinearity plots (Fig 3B) are calculated with the adjusted model. The expansive/compressive (EC) score measures where the unit tends to operate along the nonlinear activation function:

$$EC = \frac{\rho_6}{\rho_1}\sum_t z_j(t) - \rho_4 - \rho_5$$

The EC score is the average output of the unit over all the stimuli, scaled to be between -1 at threshold and +1 at saturation. For the NRF model ρ_1 = 1.7159, ρ_4 = 0, ρ_5 = 0, ρ_6 = 1 and for the OU we replace $z_j(t)$ with $\hat{y}(t)$. For the LN model, $\hat{y}(t)$ replaces $z_j(t)$, ρ_1 and ρ_4 are the fitted values from the nonlinearity, and ρ_5 = 1 and ρ_6 = 2. The average output (unscaled EC) for each unit is shown as the green dot on its nonlinear activation function (Fig 3A, 3B and 3D).

The displayed STRFs. The HU SRTF weights (Fig 3B, top panels of each HU) are shown with each element sign-reversed for the inhibitory HUs. Plotting HU STRFs in this manner thus ensures that the STRF plots always show the direction of effect of an STRF weight on the model output, rather than on the HU. This was done to facilitate the comparison between NRF HU STRFs and LN model STRFs.

Measuring the contours, and temporal and spectral tuning width of the STRFs. To obtain smooth contours for high sensitivity regions of the HU STRFs (Fig 3C), we spline interpolated the HU’s display STRF onto an evenly spaced grid at 8 times the resolution, with 7 additional values between each frequency, and 7 between each time. For the excitatory HUs the contours at half the maximum value of this matrix were plotted. For the inhibitory HUs, the contours at half the minimum value of this matrix were plotted.

To get a measure of the tuning width (Fig 4C and 4D), we calculated the power STRF by taking the square of each element of the interpolated STRF. Next, to measure the frequency tuning width, we summed the power STRF over all the time bins and then determined the maximum value of the resulting vector. The half-height frequency tuning width was defined as the number of elements of this vector ≥50% of this maximum value, multiplied by the frequency range covered by each bin in the interpolated STRF (⅙ × ⅛ = 1/48 octaves). Quarter-height frequency tuning widths were defined analogously at ≥25%. The half- and quarter-height temporal tuning widths were calculated analogously.

Acknowledgments

We thank Neil Rabinowitz, Kerry Walker, Astrid Klinge-Strahl and Fernando Nodal for help with some of the experiments, and Joseph Nour for implementing the nonlinearity fitting code for the LN model.

Author Contributions

Conceptualization: NSH.

Data curation: BDBW NSH.

Formal analysis: OS NSH BDBW.

Funding acquisition: AJK JWHS ZC.

Investigation: BDBW NSH.

Methodology: NSH OS BDBW.

Project administration: AJK ZC JWHS.

Resources: AJK ZC JWHS.

Software: OS NSH BDBW.

Supervision: NSH BDBW JWHS AJK ZC.

Validation: OS NSH.

Visualization: NSH OS.

Writing – original draft: NSH OS.

Writing – review & editing: NSH JWHS BDBW AJK OS.

References

1. Adelson EH, Bergen JR. Spatiotemporal energy models for the perception of motion. J Opt Soc Am A. 1985; 2: 284–299. PMID: 3973762

2. Aertsen A, Johannesma PIM, Hermes DJ. Spectro-temporal receptive fields of auditory neurons in the grassfrog. Biol Cybern. 1980; 38: 235–248.

3. Aertsen AMHJ Johannesma PIM. A comparison of the Spectro-Temporal sensitivity of auditory neurons to tonal and natural stimuli. Biol Cybern. 1981; 42: 145–156. doi: 10.1007/BF00336732 PMID: 6976799

4. Christianson GB, Sahani M, Linden JF. The Consequences of Response Nonlinearities for Interpretation of Spectrotemporal Receptive Fields. J Neurosci. 2008; 28: 446–455. doi: 10.1523/JNEUROSCI.1775-07.2007 PMID: 18184787

5. David SV, Mesgarani N, Fritz JB, Shamma SA. Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J Neurosci. 2009; 29: 3374–3386. doi: 10.1523/JNEUROSCI.5249-08.2009 PMID: 19295144

6. deCharms RC, Blake DT, Merzenich MM. Optimizing Sound Features for Cortical Neurons. Science. 1998; 280: 1439–1444. doi: 10.1126/science.280.5368.1439 PMID: 9603734

7. Escabı MA, Schreiner CE. Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. J Neurosci. 2002; 22: 4114–4131. PMID: 12019330

8. Fritz J, Shamma S, Elhilali M, Klein D. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci. 2003; 6: 1216–1223. doi: 10.1038/nn1141 PMID: 14583754

9. Gill P, Zhang J, Woolley SMN, Fremouw T, Theunissen FE. Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci. 2006; 21: 5–20. doi: 10.1007/s10827-006-7059-4 PMID: 16633939

10. Gourévitch B, Noreña A, Shaw G, Eggermont JJ. Spectrotemporal receptive fields in anesthetized cat primary auditory cortex are context dependent. Cereb Cortex N Y N 1991. 2009; 19: 1448–1461. doi: 10.1093/cercor/bhn184 PMID: 18854580

11. Klein DJ, Depireux DA, Simon JZ, Shamma SA. Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. J Comput Neurosci. 2000; 9: 85–111. PMID: 10946994

12. Linden JF, Liu RC, Sahani M, Schreiner CE, Merzenich MM. Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. J Neurophysiol. 2003; 90: 2660–2675. doi: 10.1152/jn.00751.2002 PMID: 12815016

13. Miller LM, Escabí MA, Read HL, Schreiner CE. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol. 2002; 87: 516–527. PMID: 11784767

14. Reid RC, Soodak RE, Shapley RM. Linear mechanisms of directional selectivity in simple cells of cat striate cortex. Proc Natl Acad Sci U S A. 1987; 84: 8740–8744. PMID: 3479811

15. Schnupp JW, Mrsic-Flogel TD, King AJ. Linear processing of spatial cues in primary auditory cortex. Nature. 2001; 414: 200–204. doi: 10.1038/35102568 PMID: 11700557

16. Theunissen FE, Sen K, Doupe AJ. Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds. J Neurosci. 2000; 20: 2315–2331. PMID: 10704507

17. Machens CK, Wehr MS, Zador AM. Linearity of cortical receptive fields measured with natural sounds. J Neurosci. 2004; 24: 1089–1100. doi: 10.1523/JNEUROSCI.4445-03.2004 PMID: 14762127

18. Olshausen BA, Field DJ. How close are we to understanding v1? Neural Comput. 2005; 17: 1665–1699. doi: 10.1162/0899766054026639 PMID: 15969914

19. Rabinowitz NC, Willmore BDB, Schnupp JWH, King AJ. Contrast Gain Control in Auditory Cortex. Neuron. 2011; 70: 1178–1191. doi: 10.1016/j.neuron.2011.04.030 PMID: 21689603

20. Atencio CA, Sharpee TO, Schreiner CE. Cooperative nonlinearities in auditory cortical neurons. Neuron. 2008; 58: 956–966. doi: 10.1016/j.neuron.2008.04.026 PMID: 18579084

21. Ahrens MB, Linden JF, Sahani M. Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. J Neurosci. 2008; 28: 1929–1942. doi: 10.1523/JNEUROSCI.3377-07.2008 PMID: 18287509

22. Calabrese A, Schumacher JW, Schneider DM, Paninski L, Woolley SMN. A Generalized Linear Model for Estimating Spectrotemporal Receptive Fields from Responses to Natural Sounds. PLoS ONE. 2011; 6: e16104. doi: 10.1371/journal.pone.0016104 PMID: 21264310

23. David SV, Shamma SA. Integration over multiple timescales in primary auditory cortex. J Neurosci. 2013; 33: 19154–19166. doi: 10.1523/JNEUROSCI.2270-13.2013 PMID: 24305812

24. Schinkel-Bielefeld N, David SV, Shamma SA, Butts DA. Inferring the role of inhibition in auditory processing of complex natural stimuli. J Neurophysiol. 2012; 107: 3296–3307. doi: 10.1152/jn.01173.2011 PMID: 22457454

25. Willmore BD, Schoppe O, King AJ, Schnupp JWH, Harper NS. Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing. J Neurosci. in press;

26. Sohl-Dickstein J, Poole B, Ganguli S. Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods. ArXiv13112115 Cs. 2013; Available: http://arxiv.org/abs/1311.2115

27. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. 2nd ed. 2009. Corr. 7th printing 2013 edition. New York, NY: Springer; 2011.

28. Bishop CM. Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc.; 2006.

29. Hsu A, Borst A, Theunissen FE. Quantifying variability in neural responses and its application for the validation of model predictions. Netw Bristol Engl. 2004; 15: 91–109.

30. Touryan J, Felsen G, Dan Y. Spatial structure of complex cell receptive fields measured with natural images. Neuron. 2005; 45: 781–791. doi: 10.1016/j.neuron.2005.01.029 PMID: 15748852

31. Asari H, Zador AM. Long-lasting context dependence constrains neural encoding models in rodent auditory cortex. J Neurophysiol. 2009; 102: 2638–2656. doi: 10.1152/jn.00577.2009 PMID: 19675288

32. Botev ZI, Grotowski JF, Kroese DP. Kernel density estimation via diffusion. Ann Stat. 2010; 38: 2916–2957. doi: 10.1214/10-AOS799

33. Carlson NL, Ming VL, DeWeese MR. Sparse Codes for Speech Predict Spectrotemporal Receptive Fields in the Inferior Colliculus. PLoS Comput Biol. 2012; 8: e1002594. doi: 10.1371/journal.pcbi.1002594 PMID: 22807665

34. Atencio CA, Sharpee TO, Schreiner CE. Hierarchical computation in the canonical auditory cortical circuit. Proc Natl Acad Sci U S A. 2009; 106: 21894–21899. doi: 10.1073/pnas.0908383106 PMID: 19918079

35. Moore AK, Wehr M. Parvalbumin-Expressing Inhibitory Interneurons in Auditory Cortex Are Well-Tuned for Frequency. J Neurosci. 2013; 33: 13713–13723. doi: 10.1523/JNEUROSCI.0663-13.2013 PMID: 23966693

36. Carandini M, Heeger DJ, Senn W. A synaptic explanation of suppression in visual cortex. J Neurosci. 2002; 22: 10053–10065. PMID: 12427863

37. Carandini M, Heeger DJ, Movshon JA. Linearity and Normalization in Simple Cells of the Macaque Primary Visual Cortex. J Neurosci. 1997; 17: 8621–8644. PMID: 9334433

38. Abbott LF, Chance FS. Drivers and modulators from push-pull and balanced synaptic input. Prog Brain Res. 2005; 149: 147–155. doi: 10.1016/S0079-6123(05)49011-1 PMID: 16226582

39. Dean I, Harper NS, McAlpine D. Neural population coding of sound level adapts to stimulus statistics. Nat Neurosci. 2005; 8: 1684–1689. doi: 10.1038/nn1541 PMID: 16286934

40. Rabinowitz NC, Willmore BDB, King AJ, Schnupp JWH. Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. PLoS Biol. 2013; 11: e1001710. doi: 10.1371/journal.pbio.1001710 PMID: 24265596

41. Rabinowitz NC, Willmore BDB, Schnupp JWH, King AJ. Spectrotemporal Contrast Kernels for Neurons in Primary Auditory Cortex. J Neurosci. 2012; 32: 11271–11284. doi: 10.1523/JNEUROSCI.1715-12.2012 PMID: 22895711

42. Heil P, Irvine DR. First-spike timing of auditory-nerve fibers and comparison with auditory cortex. J Neurophysiol. 1997; 78: 2438–2454. PMID: 9356395

43. Atencio CA, Sharpee TO, Schreiner CE. Receptive field dimensionality increases from the auditory midbrain to cortex. J Neurophysiol. 2012; 107: 2594–2603. doi: 10.1152/jn.01025.2011 PMID: 22323634

44. Sharpee T, Rust NC, Bialek W. Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput. 2004; 16: 223–250. doi: 10.1162/089976604322742010 PMID: 15006095

45. Sharpee TO, Atencio CA, Schreiner CE. Hierarchical representations in the auditory cortex. Curr Opin Neurobiol. 2011; 21: 761–767. doi: 10.1016/j.conb.2011.05.027 PMID: 21704508

46. Sharpee TO. Computational Identification of Receptive Fields. Annu Rev Neurosci. 2013; 36: 103–120. doi: 10.1146/annurev-neuro-062012-170253 PMID: 23841838

47. Kouh M, Sharpee TO. Estimating linear-nonlinear models using Renyi divergences. Netw Bristol Engl. 2009; 20: 49–68. doi: 10.1080/09548980902950891 PMID: 19568981

48. Williamson RS, Sahani M, Pillow JW. The Equivalence of Information-Theoretic and Likelihood-Based Methods for Neural Dimensionality Reduction. Bethge M, editor. PLOS Comput Biol. 2015; 11: e1004141. doi: 10.1371/journal.pcbi.1004141 PMID: 25831448

49. Laudanski J, Edeline J-M, Huetz C. Differences between Spectro-Temporal Receptive Fields Derived from Artificial and Natural Stimuli in the Auditory Cortex. PLoS ONE. 2012; 7: e50539. doi: 10.1371/journal.pone.0050539 PMID: 23209771

50. Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, et al. Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature. 2008; 454: 995–999. doi: 10.1038/nature07140 PMID: 18650810

51. Paninski L. Maximum likelihood estimation of cascade point-process neural encoding models. Netw Comput Neural Syst. 2004; 15: 243–262.

52. Park IM, Archer EW, Priebe N, Pillow JW. Spectral methods for neural characterization using generalized quadratic models. Advances in neural information processing systems. 2013. pp. 2454–2462. Available: http://papers.nips.cc/paper/4993-spectra

53. McFarland JM, Cui Y, Butts DA. Inferring Nonlinear Neuronal Computation Based on Physiologically Plausible Inputs. PLoS Comput Biol. 2013; 9: e1003143. doi: 10.1371/journal.pcbi.1003143 PMID: 23874185

54. Kozlov AS, Gentner TQ. Central auditory neurons have composite receptive fields. Proc Natl Acad Sci. 2016; 113: 1441–1446. doi: 10.1073/pnas.1506903113 PMID: 26787894

55. Margoliash D, Bankes SC. Computations in the Ascending Auditory Pathway in Songbirds Related to Song Learning. Am Zool. 1993; 33: 94–103.

56. Lau B, Stanley GB, Dan Y. Computational subunits of visual cortical neurons revealed by artificial neural networks. Proc Natl Acad Sci. 2002; 99: 8974–8979. doi: 10.1073/pnas.122173799 PMID: 12060706

57. Lehky SR, Sejnowski TJ, Desimone R. Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. J Neurosci. 1992; 12: 3568–3581. PMID: 1527596

58. Prenger R, Wu MC-K, David SV, Gallant JL. Nonlinear V1 responses to natural scenes revealed by neural network analysis. Neural Netw. 2004; 17: 663–679. doi: 10.1016/j.neunet.2004.03.008 PMID: 15288891

59. Schoppe O, Harper NS, Willmore BDB, King AJ, Schnupp JWH. Measuring the Performance of Neural Models. Front Comput Neurosci. 2016; 10. doi: 10.3389/fncom.2016.00010 PMID: 26903851

60. De Mulder W, Bethard S, Moens M-F. A survey on the application of recurrent neural networks to statistical language modeling. Comput Speech Lang. 2015; 30: 61–98. doi: 10.1016/j.csl.2014.09.005

61. Bizley JK, Nodal FR, Nelken I, King AJ. Functional organization of ferret auditory cortex. Cereb Cortex N Y N 1991. 2005; 15: 1637–1653. doi: 10.1093/cercor/bhi042 PMID: 15703254

62. Bizley JK, Walker KMM, King AJ, Schnupp JWH. Neural Ensemble Codes for Stimulus Periodicity in Auditory Cortex. J Neurosci. 2010; 30: 5078–5091. doi: 10.1523/JNEUROSCI.5475-09.2010 PMID: 20371828

63. Nelken I, Bizley JK, Nodal FR, Ahmed B, Schnupp JWH, King AJ. Large-Scale Organization of Ferret Auditory Cortex Revealed Using Continuous Acquisition of Intrinsic Optical Signals. J Neurophysiol. 2004; 92: 2574–2588. doi: 10.1152/jn.00276.2004 PMID: 15152018

64. Kadir SN, Goodman DFM, Harris KD. High-dimensional cluster analysis with the Masked EM Algorithm. ArXiv13092848 Cs Q-Bio Stat. 2013; Available: http://arxiv.org/abs/1309.2848

65. Sahani M, Linden JF. How linear are auditory cortical responses? In: Advances in Neural Information Processing Systems [Internet]. 2003 [cited 9 May 2015]. Available: http://discovery.ucl.ac.uk/8281/

66. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33: 1–22. PMID: 20808728

67. Thorson IL, Liénard J, David SV. The Essential Complexity of Auditory Receptive Fields. Theunissen FE, editor. PLOS Comput Biol. 2015; 11: e1004628. doi: 10.1371/journal.pcbi.1004628 PMID: 26683490

68. LeCun Y, Bottou L, Orr GB, Müller K-R. Efficient BackProp. In: Orr GB, Müller K-R, editors. Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg; 1998. pp. 9–50. Available: http://link.springer.com/chapter/10.1007/3-540-49430-8_2

69. Príncipe JC, Euliano NR, Lefebvre WC. Neural and adaptive systems: fundamentals through simulations. Wiley; 2000.

A.5. Incorporating Midbrain Adaptation to Mean Sound Level Improves Models of Auditory Cortical Processing

Authors: B Willmore, **O Schoppe***, A King, J Schnupp, N Harper.
**Joint first authorship*

Abstract: Adaptation to stimulus statistics, such as the mean level and contrast of recently heard sounds, has been demonstrated at various levels of the auditory pathway. It allows the nervous system to operate over the wide range of intensities and contrasts found in the natural world. Yet current standard models of the response properties of auditory neurons do not incorporate such adaptation. Here we present a model of neural responses in the ferret auditory cortex (the IC Adaptation model), which takes into account adaptation to mean sound level at a lower level of processing: the inferior colliculus (IC). The model performs high-pass filtering with frequency-dependent time constants on the sound spectrogram, followed by half-wave rectification, and passes the output to a standard linear–nonlinear (LN) model. We find that the IC Adaptation model consistently predicts cortical responses better than the standard LN model for a range of synthetic and natural stimuli. The IC Adaptation model introduces no extra free parameters, so it improves predictions without sacrificing parsimony. Furthermore, the time constants of adaptation in the IC appear to be matched to the statistics of natural sounds, suggesting that neurons in the auditory midbrain predict the mean level of future sounds and adapt their responses appropriately.

Date of publication: 13.01.2016
Journal name: The Journal of Neuroscience

Individual contribution: performing the research, analyzing the data, writing the paper

Systems/Circuits

Incorporating Midbrain Adaptation to Mean Sound Level Improves Models of Auditory Cortical Processing

Ben D.B. Willmore,[1]* Oliver Schoppe,[1,2]* Andrew J. King,[1] Jan W.H. Schnupp,[1]** and Nicol S. Harper[1]**

[1]Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford OX1 3PT, United Kingdom, and [2]Bio-Inspired Information Processing, Technische Universität München, 85748 Garching, Germany

Adaptation to stimulus statistics, such as the mean level and contrast of recently heard sounds, has been demonstrated at various levels of the auditory pathway. It allows the nervous system to operate over the wide range of intensities and contrasts found in the natural world. Yet current standard models of the response properties of auditory neurons do not incorporate such adaptation. Here we present a model of neural responses in the ferret auditory cortex (the IC Adaptation model), which takes into account adaptation to mean sound level at a lower level of processing: the inferior colliculus (IC). The model performs high-pass filtering with frequency-dependent time constants on the sound spectrogram, followed by half-wave rectification, and passes the output to a standard linear–nonlinear (LN) model. We find that the IC Adaptation model consistently predicts cortical responses better than the standard LN model for a range of synthetic and natural stimuli. The IC Adaptation model introduces no extra free parameters, so it improves predictions without sacrificing parsimony. Furthermore, the time constants of adaptation in the IC appear to be matched to the statistics of natural sounds, suggesting that neurons in the auditory midbrain predict the mean level of future sounds and adapt their responses appropriately.

Key words: adaptation; auditory cortex; inferior colliculus; mean sound level; model; spectrotemporal receptive field

Significance Statement

An ability to accurately predict how sensory neurons respond to novel stimuli is critical if we are to fully characterize their response properties. Attempts to model these responses have had a distinguished history, but it has proven difficult to improve their predictive power significantly beyond that of simple, mostly linear receptive field models. Here we show that auditory cortex receptive field models benefit from a nonlinear preprocessing stage that replicates known adaptation properties of the auditory midbrain. This improves their predictive power across a wide range of stimuli but keeps model complexity low as it introduces no new free parameters. Incorporating the adaptive coding properties of neurons will likely improve receptive field models in other sensory modalities too.

Introduction

Adaptation to stimulus statistics is an important process in sensory coding, whereby neurons adjust their sensitivity in response to the statistics of recently presented stimuli (Fairhall et al., 2001; Wark et al., 2007; Carandini and Heeger, 2012). For example, neurons in the auditory nerve (Wen et al., 2009), inferior colliculus (IC; Dean et al., 2008), and cortex (Watkins and Barbour, 2008; Rabinowitz et al., 2013) shift their dynamic ranges to compensate for changes in the mean level of recent sound stimulation. Neurons in the auditory periphery (Joris and Yin, 1992), midbrain (Rees and Møller, 1983; Kvale and Schreiner, 2004; Dean et al., 2005; Nelson and Carney, 2007; Dahmen et al., 2010; Rabinowitz et al., 2013), and higher auditory pathways (Nagel

Received June 26, 2015; revised Nov. 3, 2015; accepted Nov. 10, 2015.

Author contributions: B.D.B.W. and N.S.H. designed research; B.D.B.W., O.S., and N.S.H. performed research; A.J.K. contributed unpublished reagents/analytic tools; B.D.B.W., O.S., and N.S.H. analyzed data; B.D.B.W., O.S., A.J.K., J.W.H.S., and N.S.H. wrote the paper.

This work was supported by the Wellcome Trust through a Principal Research Fellowship to A.J.K. (WT076508AIA). O.S. was supported by the German National Academic Foundation. N.S.H. was supported by a Sir Henry Wellcome Postdoctoral Fellowship (WT082692) and other Wellcome Trust funding (WT076508AIA); by the Department of Physiology, Anatomy, and Genetics at the University of Oxford; by Action on Hearing Loss Grant PA07; and by Biotechnology and Biological Sciences Research Council Grant BB/H008608/1. We are grateful to Josh McDermott, Neil Rabinowitz, Sandra Tolnai, Kerry Walker, Fernando Nodal, Astrid Klinge-Strahl, Vani Rajendran, and James Cooke for help with these experiments and to Stephen David for providing stimulus generation code. We used sounds from the freesound.org database contributed by users juskiddink, felix-blume, duckduckpony, lebcraftlp, and spoonbender.

The authors declare no competing financial interests.

*B.D.B.W. and O.S. are joint first authors.

**J.W.H.S. and N.S.H. joint last authors.

This article is freely available online through the *J Neurosci* Author Open Choice option.

Correspondence should be addressed to Ben D.B. Willmore, Department of Physiology, Anatomy, and Genetics, University of Oxford, Sherrington Building, Parks Road, Oxford OX1 3PT, UK. E-mail: **benjamin.willmore@dpag.ox.ac.uk**.

DOI:10.1523/JNEUROSCI.2441-15.2016

and Doupe, 2006; Malone et al., 2010; Rabinowitz et al., 2011) also adapt to the variance of recently presented stimuli. Similar processes operate in the visual (Mante et al., 2005) and somatosensory (Garcia-Lazaro et al., 2007) systems, and it has been proposed that these forms of adaptation allow the nervous system to efficiently represent stimuli across the wide range of intensities and contrasts found in the natural world (Fairhall et al., 2001).

Functional models aimed at predicting responses of sensory neurons generally do not incorporate adaptation to stimulus statistics. In the auditory system, such models typically involve variations of the spectrotemporal receptive field (STRF), the standard computational model of neuronal responses (Aertsen et al., 1980; Aertsen and Johannesma, 1981; deCharms et al., 1998; Klein et al., 2000; Theunissen et al., 2000; Escabí and Schreiner, 2002; Miller et al., 2002; Fritz et al., 2003; Linden et al., 2003; Gill et al., 2006; Christianson et al., 2008; David et al., 2009; Gourévitch et al., 2009). Each STRF is a set of coefficients that describe the best linear approximation to the relationship between the spiking responses of a neuron and the power in the spectrogram of the sounds heard by the animal.

In principle, STRFs are powerful computational tools because they provide both a way to characterize neurons, by quantifying their sensitivity to different sound frequencies, and to predict responses to arbitrary new stimuli (deCharms et al., 1998; Schnupp et al., 2001; Escabí and Schreiner, 2002). In practice, STRFs are only moderately successful in achieving this (Linden et al., 2003; Machens et al., 2004). To improve the predictive power of STRFs, nonlinear extensions have been proposed, including output nonlinearities (Atencio et al., 2008; Rabinowitz et al., 2011), feedback kernels (Calabrese et al., 2011), second-order interactions, and input nonlinearities (Ahrens et al., 2008; David et al., 2009; David and Shamma, 2013). However, prediction accuracy remains far from perfect. Also, some of these approaches add complexity to the model and can be difficult to interpret in biological terms. Here we take an alternative approach that seeks to improve the prediction accuracy of STRF-like models by incorporating a simple, adaptive, nonlinear preprocessing step that mimics the physiological properties of neurons in the auditory midbrain.

Adaptation to mean sound level in the IC has been characterized by Dean et al. (2008), who measured how the time constants of adaptation in guinea pigs vary with frequency. This information can be used to build a model of adaptation to stimulus statistics in the IC, which can then be incorporated into an STRF model of neural responses. We recorded the responses of neurons in ferret auditory cortex to a range of sounds and constructed STRF models relating the responses to the sound spectrograms. We then augmented these models by incorporating a nonlinear transform of the spectrogram, which captures adaptation to mean sound level in the IC. Since the IC provides an obligatory relay for ascending inputs to the auditory cortex, this transform was incorporated at the input stage of the model, forming a nonlinear–linear–nonlinear (NLN) cascade. This NLN model provides a substantial improvement over the standard STRF models in describing and predicting the responses of cortical neurons.

Materials and Methods

Experimental procedures

All animal procedures were approved by the local ethical review committee and performed under license from the UK Home Office. Ten adult pigmented ferrets (seven female, three male; all >6 months of age) underwent electrophysiological recordings under ketamine–medetomidine anesthesia. Full details are as in the study by Bizley et al. (2009). Briefly, we induced general anesthesia with a single intramuscular dose of medetomidine (0.022 mg $\cdot$ kg $^{-1}$ $\cdot$ h $^{-1}$) and ketamine (5 mg $\cdot$ kg $^{-1}$ $\cdot$ h $^{-1}$), which was then maintained with a continuous intravenous infusion of medetomidine and ketamine in saline. Oxygen was supplemented with a ventilator, and we monitored vital signs (body temperature, end-tidal CO_2, and the electrocardiogram) throughout the experiment. The temporal muscles were retracted, a head holder was secured to the skull surface, and a craniotomy and a durotomy were made over the auditory cortex. We made extracellular recordings from neurons in primary auditory cortex (A1) and the anterior auditory field (AAF) using silicon probe electrodes (Neuronexus Technologies) with 16 or 32 sites (spaced at 50 or 150 μm) on probes with one, two, or four shanks (spaced at 200 μm). Stimuli were presented via Panasonic RPHV27 earphones, which were coupled to otoscope specula that were inserted into each ear canal, and driven by Tucker-Davis Technologies System III hardware (48 kHz sample rate). We clustered spikes off-line using klustakwik (Kadir et al., 2014); for subsequent manual sorting, we used either spikemonger (an in-house package) or klustaviewa (Kadir et al., 2014).

Stimuli

We used several stimulus classes: two types of dynamic random chords (DRCs), temporally orthogonal ripple combinations (TORCs), modulated noise, and natural sounds.

DRCs (deCharms et al., 1998; Schnupp et al., 2001; Rutkowski et al., 2002; Linden et al., 2003) consist of sequences of superposed pure tones whose levels are chosen pseudorandomly. Each chord contained 31 pure tones whose frequencies were log-spaced between 1 kHz and 32 kHz at 1/6 octave intervals. Each chord lasted 62.5 ms with 5 ms linear ramps between chords. The levels were chosen from a uniform distribution between 30 and 70 dB sound pressure level (SPL). We also included variable-rate DRCs, a novel stimulus designed to have a richer modulation structure, while retaining the other advantages of DRCs. In this case, each chord lasted 10.4 ms, but the level of each tone was kept constant for between 1 and 12 chords (lengths were chosen from a uniform distribution, independently for each frequency), rather than changing on every chord.

TORCs (Klein et al., 2000) consist of superposed noise stimuli with spectrograms modulated by superpositions of sinusoids. We used a set of 30 TORCs (each 3 s long) covering frequency space from 1 to 32 kHz, with temporal modulations from 4 to 48 Hz and frequency modulations up to 1.4 cycles/octave.

The modulated noise stimulus was generated using the sound texture synthesis algorithm developed by McDermott and Simoncelli (2011). The modulated noise had a pink power spectrum between 1 and 32 kHz and a white modulation spectrum between 13.3 and 160 Hz. This stimulus has a somewhat naturalistic structure, but without the complex higher-order statistical relationships of real, natural sounds.

In the first series of experiments (BigNat), we presented natural sounds only. We made recordings from 535 units in six ferrets (five female, one male). There were 20 sound clips of 5 s duration each, separated from each other by ~0.25 s silence. We recorded responses to the clips, presented in random order and repeated this 20 times. The sound clips included recordings of animal vocalizations (e.g., ferrets and birds), environmental sounds (e.g., water and wind), and speech. The sequences had root mean square intensities in the range 75–82 dB SPL. We presented the sounds at a sampling rate of 48,828.125 Hz. We discarded data recorded in the first 250 ms after the onset of each stimulus, leaving an effective data size of 20 × 95 s (20 repeats of 20 sounds with a duration of 4.75 s each).

In the second series of experiments (Comparison), we presented natural sounds, DRCs, TORCs, and modulated noise in an interleaved fashion, to enable comparison between different stimulus types. We recorded responses to the clips, presented in random order, and repeated this 10 times. Recordings were made from 220 units in four ferrets (two female, two male). The stimulus sampling rate was 97,656.25 Hz. Again, we discarded the first 250 ms after the onset of each stimulus, leaving an effective data size of 5 × 10 × 45 s (five stimulus types with 10 repeats of 45 s each).

In the Comparison dataset, the natural sounds were 1 s snippets of vocalizations (human, bird, sheep) and environmental sounds. These

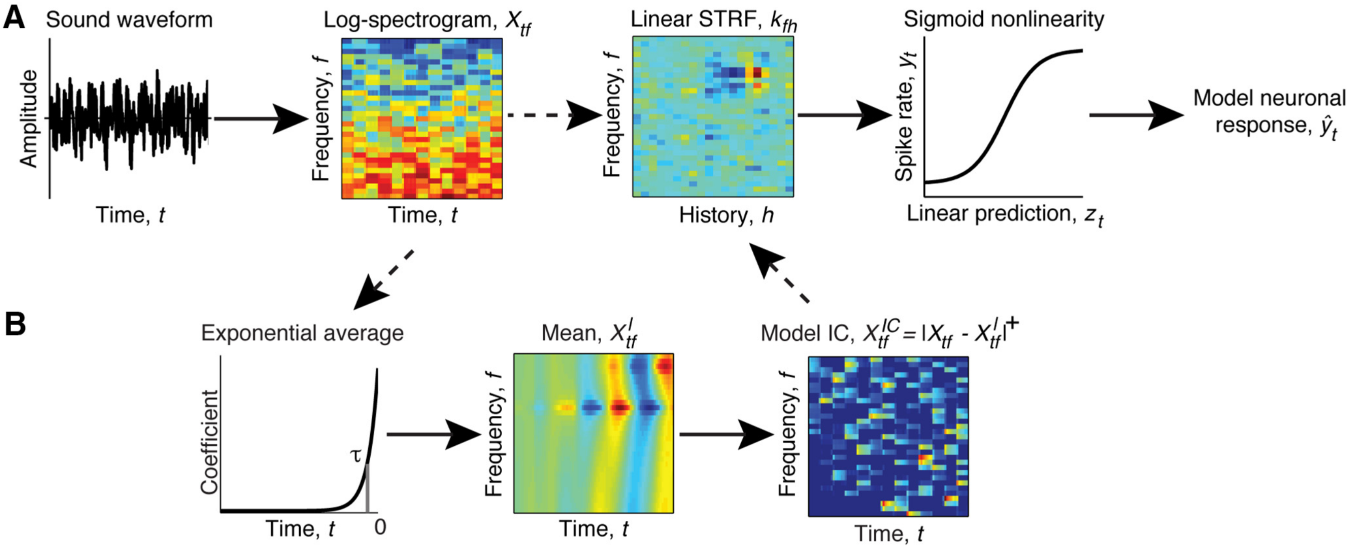


Figure 1. Two models of the stimulus–response relationship for auditory neurons. ***A***, Standard LN model. The log-spectrogram of the sound waveform, X_{tf}, is operated on by a linear kernel, k_{fh}, and sigmoid output nonlinearity to produce a model, $\hat{y}_t$, of the neuronal response. ***B***, The IC Adaptation model augments the LN model by adding a nonlinear transform of the spectrogram. The dashed arrows indicate the alternative processing paths for the Standard LN and IC Adaptation models. The nonlinear transform consists of high-pass filtering each frequency band of the spectrogram by subtracting the convolution of that frequency band with an exponential filter with time constant τ (shown by the vertical line), followed by half-wave rectification. The resulting modified spectrogram, X^{IC}_{tf}, is then used as an alternative input to the standard LN model.

were separated by silent gaps, and the silent periods along with the first 250 ms of neural responses after each silent period were removed.

Neural responses

For each unit, we counted spikes in 5 ms time bins and averaged these counts over all trials to compute the peristimulus time histogram (PSTH). We smoothed the PSTH with a 21 ms Hanning window (Hsu et al., 2004) to estimate each neuron's evoked firing rate. We denote the (trial-averaged) neuronal response as y_t.

Unit selection criterion

Only units whose firing rate was modulated in response to the stimuli in a reliable, repeatable manner were included for analysis. We measured this using the noise ratio (NR; Sahani and Linden, 2003; Rabinowitz et al., 2011) for the PSTH of each unit:

$$\text{noise ratio} = \frac{\text{noise power}}{\text{signal power}} = \frac{\text{total variance} - \text{explainable variance}}{\text{explainable variance}}.$$

Each unit was included in our analyses if it had a noise ratio of <200 across the entire dataset (i.e., across all natural stimuli for the BigNat set or across all stimulus classes for the Comparison set). Three hundred of 535 units were included from the BigNat set and 77 of 220 from the Comparison set.

Log-spectrograms

We characterized the power in each stimulus using a log-spaced, log-valued spectrogram. We first calculated the spectrogram of each sound using 10 ms Hanning windows, overlapping by 5 ms (giving 5 ms temporal resolution). We then aggregated across frequency using overlapping triangular windows with log-spaced characteristic frequencies to compute the signal power in each frequency band, using code modified from melbank.m by Mike Brookes (Imperial College London, London, UK; http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html). For the BigNat stimulus, we used 34 log-spaced frequencies from 500 to 22.627 Hz (1/6 octave spacing). For the Comparison stimuli, we used 31 log-spaced frequencies from 1 to 32 kHz (also 1/6 octave spacing). Finally, we took the logarithm of the resulting values, and values lower than a threshold (approximately equivalent to the mean activity caused by a 0 dB SPL flat spectrum noise) were set to that threshold, giving the log-spectrogram, X_{tf}, at time t and frequency f.

As input to the models, we reorganized X_{tf} as a three-tensor X_{tfh}, where X_{tfh} gives the sound intensity (elements of X_{tf}) for the recent stimulus history, $h = 0$ to H-1 time bins in the past, from time t, at frequency f, i.e., $X_{tfh} = X_{(t-h),f}$. For STRF estimation, we used a history length of 20 bins of 5 ms duration (100 ms total).

Model testing and comparison

To fit and test our models, we used a k-fold testing procedure ($k = 10$) for both datasets. Thus, each dataset was split into 10 segments consisting of a contiguous 10% of the data. One of the 10 segments was set aside as a test set, and the model was trained on the remaining 90% of the data (the training set) to fit the STRF and the parameters of the nonlinearity. Model performance was then measured with the unused test set, i.e., the model was used to predict the neural response to the test set stimulus. We repeated this process 10 times, each time using a different segment as a test set, and averaged the performance measure over the 10 segments.

Linear–nonlinear STRF model

We described the responses of cortical neurons using two models. The first was a standard linear–nonlinear (LN) model (Chichilnisky, 2001; Simoncelli et al., 2004; Fig. 1*A*) relating neural responses, y_t, to the log-spectrogram, X_{tf}, of the stimuli. To do this, we first found the STRF k_{fh}, the linear approximation to the mapping between the PSTH, y_t, and the log-spectrogram, X_{tf}. We estimated k_{fh} by minimizing (subject to regularization; see below) the mean squared error between the PSTH, y_t, and its linear estimate from X_{tf}. This linear estimate, z_t, is given by the following:

$$z_t = \sum_{f,h} X_{tfh} k_{fh}. \quad (1)$$

Previous studies have used separable kernel estimation (Linden et al., 2003; Rabinowitz et al., 2011), which sometimes provides better descriptions of auditory neurons than inseparable approaches (particularly using DRC stimuli). Here, however, we used inseparable kernels to allow for the possibility of inseparable kernel structure with the TORCs and modulated noise stimuli. To estimate inseparable kernels, we used glmnet for Matlab (J. Qian, T. Hastie, J. Friedman, R. Tibshirani, and N. Simon, Stanford University, Stanford, CA; see http://web.stanford.edu/~hastie/glmnet_matlab/), which uses elastic net regularization. This technique can optimize k_{fh} using a linear combination of L1 (Willmore et al., 2010) and L2 (Willmore and Smyth, 2003) penalties we have used in the past, with a parameter, α, that determines the relative strength of each penalty. We explored three approaches: using an L1 penalty ($\alpha = 1$), using an L2 penalty ($\alpha = 0$), and optimizing α for each unit. Here we present the results obtained using L2 regularization, but the results are similar for the other forms of regularization. The regularization parameter, λ, determines the strength of regularization. To determine the optimal choice of λ, we reserved a randomly chosen 10% of each training set for cross-validation. STRFs were estimated using the remaining 90% of the training set, using a wide range of choices of λ. We then selected the STRF that provided the best prediction (minimum mean square error) on the cross-validation set. The use of three separate subsets (where STRFs are fitted using one set, the regularization parameter is chosen using a second set, and prediction scores are measured using a third set), minimizes overfitting in both model fitting and assessment.

We then fitted a sigmoid (logistic) nonlinearity to relate the output of the linear model, z_t, to the neural responses by minimizing the mean squared error between the PSTH, y_t, and the nonlinear estimate of the PSTH, $\hat{y}_t$:

$$\hat{y}_t = a + \frac{b}{1 + \exp\left(-(z_t - c)/d\right)}, \qquad (2)$$

where a is the minimum firing rate, b is the output dynamic range, c is the input inflection point, and d is the reciprocal of the gain (Rabinowitz et al., 2011, 2012). All parameters a, b, c, and d were fitted to the whole training set, using minFunc by Mark Schmidt (University of British Columbia, British Columbia, Canada; http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html).

To ensure that this sequential fitting procedure did not adversely affect our results, we also fitted the STRF and the sigmoid output nonlinearity using two other fitting methods (for the Comparison dataset only). The first was an iterative procedure where we estimated the STRF, then estimated the sigmoid, then inverted the sigmoid and refitted the STRF, and repeated for 10 iterations. The second was a neural network with linear input units, one logistic hidden unit, and a final linear output unit; this model was fitted using backpropagation. Both of these models are identical in mathematical form to the original model and only differ in the fitting procedure. We found that they provided very similar results (data not shown) to the sequential fitting procedure.

Nonlinear–linear–nonlinear STRF models

To extend the LN model to incorporate our knowledge about adaptation to mean sound level in the IC, we introduced a nonlinear transformation of the log-spectrogram, producing the IC Adaptation model (Fig. 1*B*). To test the importance of different aspects of this model, we also used several variations as controls.

IC Adaptation model. We convolved every frequency band in the log-spectrogram of the stimulus with an exponential filter, E_{fh}:

$$X^{l}_{tf} = \sum_{h} X_{(t-h),f}\, E_{fh}, \text{ where } E_{fh} = \frac{1}{N_f}\exp(-h/\tau_f). \qquad (3)$$

N_f was a normalization constant, chosen so that the exponential filter for each frequency band summed to 1. Here, the number of time bins, H, is 499, giving 2.5 s of history. The time constants, τ_f, of the filters varied with sound frequency, following the frequency dependence of the time constant found by Dean et al. (2008). The relationship we used was a linear regression (Fig. 2) relating τ_f (in milliseconds) to the logarithm of the units' characteristic frequency (in hertz):

$$\tau_f = 500 - 105\log_{10}(f), \qquad (4)$$

so that τ_f depends on the logarithm of frequency, between $\tau_{f\,=\,500\text{ Hz}} = 217$ ms and $\tau_{f\,=\,32\text{ kHz}} = 27$ ms.

The time-varying response of each exponential filter, X^{l}_{tf}, was then subtracted from the corresponding frequency band in the log-spectrogram, X_{tf}, giving a high-pass-filtered version, $X^{h}_{tf} = X_{tf} - X^{l}_{tf}$, which was then half-wave rectified to give $X^{IC}_{tf} = |X^{h}_{tf}|^{+}$. We then used X^{IC}_{tf} in place of X_{tf} for STRF analysis.

No-half-wave-rectification model. This model is the same as the IC Adaptation model, but without half-wave rectification; i.e., X^{h}_{tf} was used for STRF analysis.

Median-τ model (τ^{med}). This model is the same as the IC Adaptation model, but a fixed time constant [$\tau^{\text{med}} = 160$ ms; equal to the median of the time constants measured by Dean et al. (2008)] was used for all frequency channels instead of the frequency-dependent τ_f.

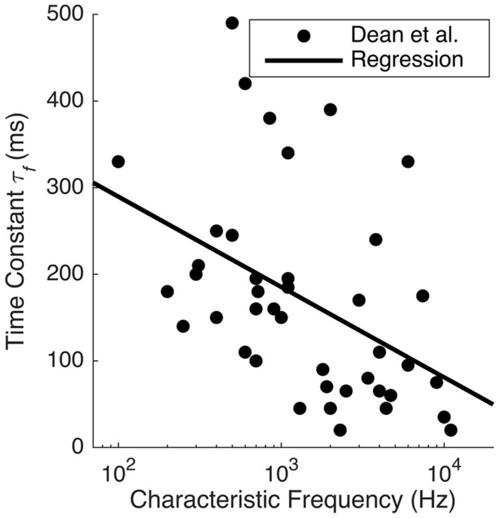


Figure 2. Time constants of adaptation to stimulus mean observed by Dean et al. (2008) in the guinea pig IC. The *x*-axis shows the characteristic frequency of each IC unit, and the *y*-axis shows the time constant of an exponential fit to the adaptation curve for the corresponding unit. The line is our regression fit to these data (Eq. 4).

Minimum-τ model (τ^{min}). This model is the same as the IC Adaptation model, but with $\tau^{\text{min}} = 27$ ms.

Maximum-τ model (τ^{max}). This model is the same as the IC Adaptation model, but with $\tau^{\text{max}} = 217$ ms.

Performance measures

Prediction performance was primarily assessed using the normalized correlation coefficient (CC_{norm}), as introduced for coherence by Hsu et al. (2004), and used for the correlation coefficient by Touryan et al. (2005). The prediction accuracy, as quantified by the raw correlation coefficient, CC_{raw}, is affected both by model performance and by the variability of neural responses to the stimulus. To correct for the contribution of response variability, and measure only model performance, we use CC_{norm}, defined as the ratio of the CC_{raw} to the theoretical maximum CC_{max}:

$$CC_{norm} = \frac{CC_{raw}}{CC_{max}}. \qquad (5)$$

CC_{max} is the correlation coefficient between the recorded mean firing rate across all repeats of the stimulus and the (unknown) true mean firing rate (measured over infinite repeats) and is an upper limit on model performance. Following Hsu et al. (2004) and Touryan et al. (2005), we estimated CC_{max} using the following:

$$CC_{max} = \sqrt{\frac{2}{1 + 1/CC_{half}}}, \qquad (6)$$

where CC_{half} is the correlation coefficient of the mean PSTH for one-half of the trials with the mean PSTH for the other half of the trials. We took the mean CC_{half} over all 126 possible combinations for the Comparison dataset (10 trials) and over 126 randomly chosen sets of half of the trials for the BigNat dataset (20 trials).

Natural sound analysis

In addition to modeling the neural responses, we performed an analysis of a database of natural sound recordings with the aim of asking how the distribution of IC adaptation time constants reported in the literature across frequencies (Dean et al., 2008) might relate to the properties of the natural acoustic environment. Most sounds in the database were recorded by us under various conditions (ranging from open air to an anechoic chamber). An additional seven sound recordings were taken from the freesound.org database. We arranged the sounds into seven broad categories: breaking wood sounds, crackling fire, rustling foliage, vocalizations (human, frog, bird, sheep), walking footsteps on numerous surfaces, ocean and river water ("water"), and rain and thunderstorms ("weather"). We calculated the log-spectrogram, X_{tf}, of each sound (as for the STRF analysis) and estimated the time-varying mean level in each frequency band, μ_{tf}, by convolution with a boxcar filter of length, T_{av} (for eight log-spaced values of T_{av} between 15 and 1000 ms), arranged so that μ_{tf} at time t contained the mean sound level between t and $t + \mathrm{T}_{av}$. Thus, μ_{tf} is an estimate of the mean sound level in a given frequency band in the immediate future.

We concatenated all sounds in a given category (including, at most, 10 s from any single sound). We then estimated a set of linear filters, E_{fh}^{nat} (one per frequency band, sound category, and value of T_{av}), which operated over 2.5 s of sound history and were optimized to produce an estimate, $\hat{\mu}_{tf}$, of the time-varying mean level, μ_{tf}:

$$\hat{\mu}_{tf} = \sum_{h} X_{(t-h),f} E_{fh}^{nat}. \quad (7)$$

The kernels, E_{fh}^{nat}, were constrained to be exponential in shape as follows:

$$E_{fh}^{nat} = A_f \exp(-h/\tau_f^{nat}). \quad (8)$$

Estimating the kernels therefore consisted of fitting two parameters: A_f (amplitude) and τ_f^{nat} (time constant). We optimized these parameters to minimize the mean squared error between $\hat{\mu}_{tf}$ and μ_{tf} using fminsearch in Matlab. Both parameters were allowed to vary freely for different frequencies.

Results

Using the methods described above, we presented a range of natural and synthetic sounds to anesthetized ferrets and recorded responses of neurons in the primary cortical areas A1 and AAF. We modeled these responses using a classical LN model of spectrotemporal tuning as well as using a novel model that included a nonlinear input stage that incorporates adaptation of IC neurons to stimulus statistics (the IC Adaptation model). We compared the predictive power of the two models by measuring the accuracy of their prediction of neural responses to a reserved test set of sounds.

The IC Adaptation model predicts responses to natural sounds more accurately than conventional LN models

We first evaluated the performance of the IC Adaptation model on the responses of each unit to a set of natural sounds (BigNat dataset). Natural sounds provide the ultimate test of models of neural responses because of their ecological relevance. A good model should be able to predict responses to natural sounds, but this is often challenging because of the variety and statistical complexity of sounds that are encountered in daily life.

We first compared the performance of the IC Adaptation model and standard LN model using a correlation coefficient ($\mathrm{CC_{raw}}$). To accurately assess performance, we measured predictions using a 10-fold testing procedure: for each stimulus type, we selected 10 nonoverlapping subsets of the data to be our test dataset. For each test set, we fitted an LN model using the rest of the data and used the LN model to predict responses to the test set. We measured the mean $\mathrm{CC_{raw}}$ (over all 10 test sets) between the LN model predictions and the actual neural responses. Using this measure suggests that there is an advantage for the IC Adaptation model over the LN model (Fig. 3*A*).

However, $\mathrm{CC_{raw}}$ is affected by neuronal response variability as well as by model performance, as can be seen from the relationship between the colors of the points in Figure 3*A* and the model performance. The red points show data from units with a high noise ratio; these have highly variable responses and consequently have low values of $\mathrm{CC_{raw}}$. The blue points show data from units with a low noise ratio; these have relatively reliable responses and so have high values of $\mathrm{CC_{raw}}$. To reduce this confound between model performance and neuronal response variability, we used the normalized correlation coefficient

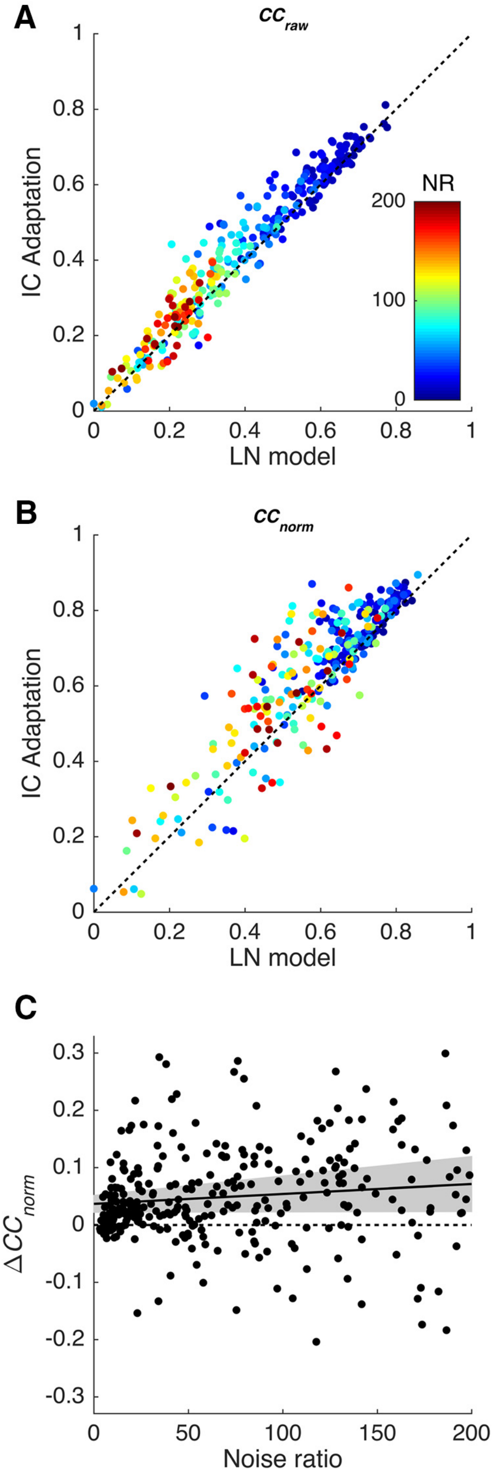


Figure 3. Comparison of the ability of the standard LN model and the IC Adaptation model to predict neural responses to natural sounds. ***A***, ***B***, Scatterplots showing the correlation coefficients between model predictions and actual neural responses (BigNat dataset). The *x*-axis shows performance of the standard LN model, the *y*-axis shows performance of the IC Adaptation model, and colors indicate the NR of each unit. ***A***, Raw correlation coefficient $\mathrm{CC_{raw}}$. ***B***, Normalized correlation coefficient $\mathrm{CC_{norm}}$. ***C***, Scatterplot showing how the difference in $\mathrm{CC_{norm}}$ between the two models varies with NR. The solid line is a linear regression, and the shaded area shows the 95% confidence intervals on the regression.

(CC_{norm}; see Materials and Methods) as our primary measure of model performance. CC_{norm} is the ratio of CC_{raw} to the estimated maximum possible correlation coefficient given the level of response variability in the data, CC_{max}. Since CC_{max} is a constant for each unit, using CC_{max} does not affect the relative performance of two models for that unit (i.e., for models 1 and 2, $CC^{(1)}_{norm}/CC^{(2)}_{norm} = CC^{(1)}_{raw}/CC^{(2)}_{raw}$). However, it gives a more accurate picture of the performance of the models across the whole dataset.

For both the CC_{norm} and the CC_{raw} measures, the IC Adaptation model provided better predictions than the LN model in 77% of neurons (Fig. 3*A*,*B*). The mean CC_{norm} for the IC Adaptation model was 0.64 compared with 0.59 for the LN model. This improvement in performance is highly significant ($p \ll 0.0001$, paired *t* test; df = 299).

It is conceivable that the advantage of the IC Adaptation model could, at least in part, be an artifact of data quality. For example, half-wave rectification of the stimulus spectrogram removes parts of the sound whose level is lower than the mean. This reduces the effective dimensionality of the stimulus set and may also reduce the effective number of STRF parameters that must be estimated. Because simple models require less data to constrain them than complex models, it is possible that this might give the IC Adaptation model an artificial advantage over the LN model for noisy neurons.

To rule out this possibility, we investigated the relationship between CC_{norm} and the NR. The NR quantifies the relative contributions of unpredictable and stimulus-driven variability in the neuronal responses (see Materials and Methods); a low NR indicates that a neuron was reliably driven by the stimulus. A scatterplot of the difference in CC_{norm} for the two models (Fig. 3*C*) shows that the advantage of the IC Adaptation model is only weakly dependent on the NR, indicating that the IC Adaptation model is generally superior to the LN model, regardless of the NR.

The IC Adaptation model also outperforms conventional LN models when tested with commonly used synthetic stimuli

An important aspect of any model is its generality. If the IC Adaptation model is a better model of cortical neurons than the standard LN model, it should provide better predictions of cortical responses to a wide range of stimuli. In one sense, testing the model on natural sounds is a good test of generality, because an appropriate collection of natural sounds will sample the space of ecologically relevant stimuli. However, it is also important to test the model using synthetic stimuli that have been widely used in neurophysiology experiments because they have been designed to exhibit well defined statistics that may provide particularly stringent tests of the model.

We therefore tested the IC Adaptation model on a second dataset (Comparison) in which several stimulus classes were presented to each unit, randomly interleaved. The classes were DRCs, TORCs, modulated noise, and natural sounds (see Materials and Methods for details). We fitted the IC Adaptation model and the LN model to each stimulus class in turn and measured predictions using CC_{norm} for reserved test data from the same class (see Materials and Methods). For every stimulus class, we found that the IC Adaptation model performs better than the LN model (Fig. 4*A*). This difference is significant at $p < 0.0001$ or better (paired *t* test) for every stimulus class except TORCs (for which $p = 0.07$). Across all stimulus classes, the mean CC_{norm} for the IC Adaptation model was 0.53 compared with 0.47 for the LN model.

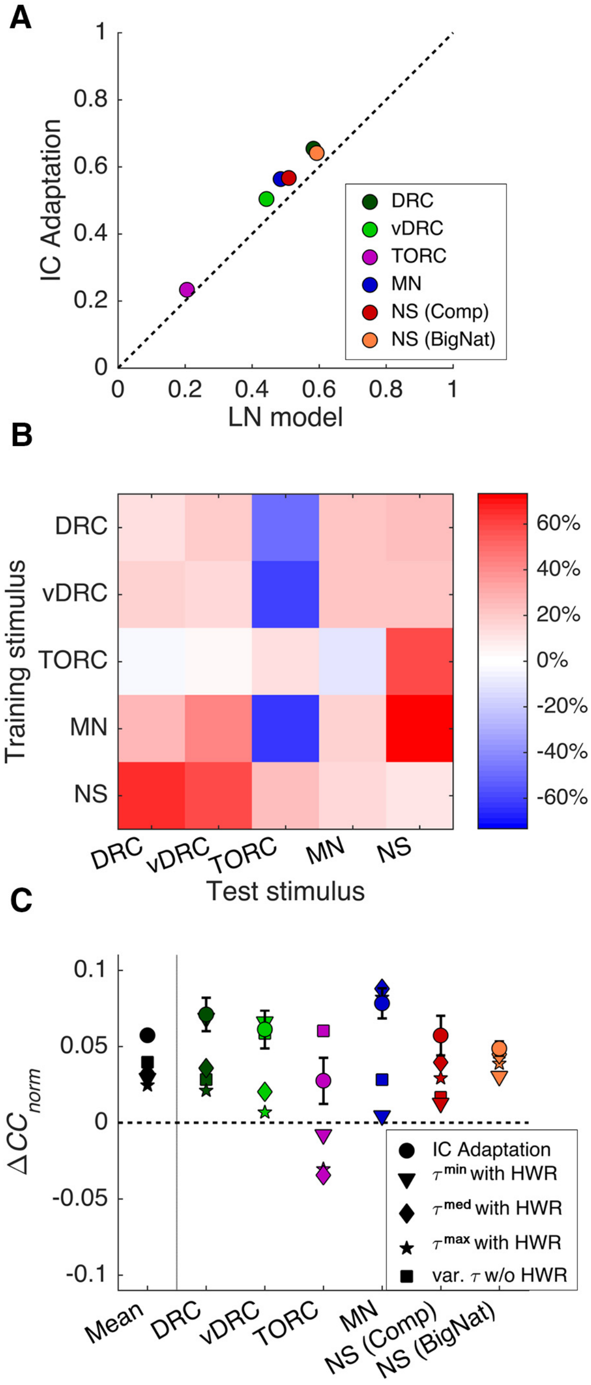


Figure 4. ***A***, Comparison of the LN and IC Adaptation models for several stimulus classes, when models are trained and tested on the same stimulus class (dots show the mean of the within-class predictions for all units). ***B***, Percentage improvement in mean model performance between the LN and IC Adaptation models, $\overline{\Delta CC'_{norm}}$, when models are trained (rows) on one stimulus class and tested (columns) on another (cross-class predictions; Comparison dataset only). ***C***, Difference between prediction performance of control models with fixed time constants (τ^{med}, τ^{min}, and τ^{max}) and without half-wave rectification, compared with the LN model for each stimulus class (colors as in Fig. 4*A*).

How well do models fare when fitted with one class of sound stimuli and tested with another?

Another important test of generality is that a model should be able to generate accurate predictions across stimulus classes. For example, if the model is trained on DRCs, it should be able to generate accurate predictions of responses to natural sounds. This has been a problem for STRF models of sensory neurons, which often perform much worse for cross-class prediction than for within-class prediction (Olshausen and Field, 2005).

To test cross-class predictions, we fitted the IC Adaptation model and the LN model to each stimulus type in the Comparison dataset in turn and measured predictions using $\mathrm{CC}_{\mathrm{norm}}$ for reserved test sets of data from other sound classes. We tested all combinations of within- and cross-class predictions by training and testing on every combination of stimulus classes and plotted the percentage difference in mean performance between the IC Adaptation model and the LN model, $\overline{\Delta \mathrm{CC}}'_{\mathrm{norm}} = 100 \times (\overline{\mathrm{CC}}^{\mathrm{IC}}_{\mathrm{norm}} - \overline{\mathrm{CC}}^{\mathrm{LN}}_{\mathrm{norm}})/\overline{\mathrm{CC}}^{\mathrm{LN}}_{\mathrm{norm}}$ (Fig. 4*B*). In 20 of 25 cases, the IC Adaptation model performed better than the LN model. The five cases where the LN model performed better than the IC Adaptation model all involve TORC stimuli. It appears that TORCs are an unusual case where the IC Adaptation model performs particularly poorly. This may be explained by the fact that the TORCs are regularly interleaved with periods of silence. Our log-spectrograms thresholded all stimuli (see Materials and Methods) to prevent model predictions from being disproportionately affected by large negative sound levels. Normally, this has little effect on the IC Adaptation model, but during periods of silence, the IC Adaptation will adapt to the threshold value. As a result, the precise value of this threshold will significantly affect model predictions.

When trained on synthetic stimuli and tested on natural sounds (NS, right-hand column), the IC Adaptation model always outperformed the LN model, suggesting that the IC Adaptation model provides superior generalization from synthetic to natural stimuli.

It is also notable that the percentage improvements are slightly higher (though not significantly so) for cross-class predictions (off-diagonal elements; median = 22%) than for within-class predictions (main diagonal; mean = 13%; difference not significant according to a rank-sum test). This indicates that the IC Adaptation model has no negative effect on the generality of STRF models; in fact, it may improve cross-class generalization relative to the LN model.

Both frequency-dependent time constants and half-wave rectification contribute to the success of the IC adaptation model

The IC Adaptation model adds two main components to the standard LN model: high-pass modulation filtering by subtraction using exponential filters with time constants derived from Dean et al. (2008) and half-wave rectification of the resulting filtered log-spectrogram. To test whether both of these components are essential to the model, or whether either component on its own is sufficient, we investigated the effects of manipulating the IC Adaptation model in two ways.

In one set of manipulations, we produced control models where the frequency dependence of the time constants, τ_f, was removed. We replaced the frequency-dependent time constants with three fixed time constants: $\tau^{\mathrm{med}} = 160$ ms, $\tau^{\mathrm{min}} = 27$ ms, and $\tau^{\mathrm{max}} = 217$ ms (corresponding to the median, minimum, and maximum of the time constants observed by Dean et al. (2008) across the range of frequencies in our log-spectrogram). In the second set, we kept the frequency-dependent time constants but removed the half-wave rectification from the IC Adaptation model. We compared the performance of these control models with the LN and IC Adaptation models for within-class predictions on all stimulus types (Fig. 4*C*). All of the control models perform better than the LN model when performance is averaged across stimulus type, and the IC Adaptation model outperforms all of the controls (Fig. 4*C*, left-hand column).

For each individual stimulus class (Fig. 4*C*, other columns), the IC Adaptation model performs better than or equivalently to the controls in almost all cases. The only exception is the model without half-wave rectification (no-HWR), which performs better than the IC Adaptation model for a single stimulus class, TORCs. However, the no-HWR model performs significantly worse for DRCs, modulated noise, and natural sounds (Comparison dataset). Overall, these results suggest that both frequency dependence of the time constants and half-wave rectification are important components of the IC Adaptation model.

Why are the adaptation time constants in IC distributed as they are?

To investigate why the time constants of adaptation in the IC exhibit the characteristic frequency-dependent distribution that has been described previously, we asked whether sound levels in natural sounds exhibit a similar frequency-dependent distribution, which would imply that the IC time constants are optimized to match the statistical structure of natural sounds.

Specifically, we assumed that mean sound level is a nuisance variable and that the role of adaptation to mean sound level in the IC is to subtract this nuisance variable from the neural representation of sound. To perform this subtraction, the IC must estimate the mean level in each frequency band. If the mean level is stable over time, an adaptation process with a short time constant should be able to reliably estimate the mean. If the mean level is unstable, a longer time constant will be required to produce a reliable estimate. We therefore measured the stability of the mean level using the autocorrelation of the spectrogram of natural sounds. A narrow autocorrelation function indicates that sound level is poorly correlated over time (unstable), whereas a wide autocorrelation function indicates that sound level is well correlated over time (stable). Given the results of Dean et al. (2008), we might expect that low frequencies will have narrower autocorrelation functions than high frequencies.

For a large set of natural sounds, we took the spectrogram of each sound and measured the autocorrelation of the spectrogram for each frequency band (see Materials and Methods). The mean (across all sounds) of the resulting autocorrelation functions is shown in Figure 5*A*. The width of the autocorrelation function (Fig. 5, black lines) varies with frequency, as predicted from the data of Dean et al. (2008). The widths range from 10 ms at the lowest frequencies (500 Hz) to 260 ms at the highest frequencies (32 kHz).

We also investigated what time constants are required to optimally estimate the mean sound level in different frequency bands. We took a large set of natural sounds and divided them into seven broad categories (see Materials and Methods). For each category, we estimated a set of linear kernels (one for each frequency channel) that optimally predict the mean level of the next T_{av} ms of sound, based on the past 2.5 s of sound. Each kernel was constrained to have an exponential shape. The time constants of the exponential kernels were optimized separately for each frequency channel, and for 8 log-spaced values of T_{av}, (between 15 ms and 1000 ms).

The time constants for the lowest and the highest frequency bands (centered at 70 Hz and 20 kHz, respectively) are plotted as a function of T_{av} in Figure 5*B*. In all cases, the mean time constants for low frequencies (Fig. 5*B*, red line; error bars show SE) are consistently higher than those for high frequencies (Fig. 5*B*, blue line). This confirms that longer time constants are required

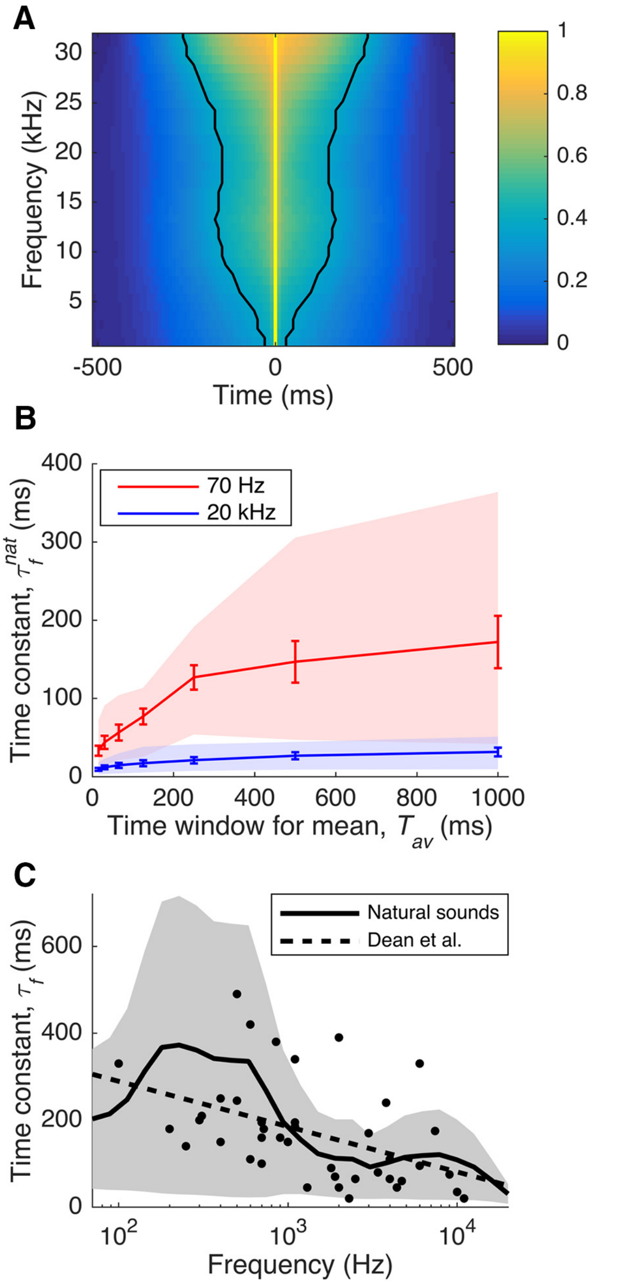


Figure 5. ***A***, Autocorrelation of the spectrogram of an ensemble of natural sounds. The black lines indicate the point where the autocorrelation has decreased to 1/*e*. ***B***, Optimal time constants for predicting the mean sound level of natural sounds in a window of length T_{av} ms into the future. This is plotted as a function of the window size, for frequency bands centered on a low value (70 Hz; red line) and a high value (20 kHz; blue line). Lines indicate mean, error bars show SEM, and shaded regions indicate entire range. ***C***, Optimal time constants for prediction of the mean level of natural stimuli in each frequency band. The shaded region indicates the range of time constants in that frequency band; the solid line indicates the middle of the range. The dots indicate time constants observed in guinea pig IC by Dean et al. (2008), and the dashed line is the linear regression of those time constants against log(frequency).

to reliably estimate the mean for low frequencies. We find that as the length of the averaging window, T_{av}, increases, the optimal time constants increase. This effect begins to saturate above 250 ms. We also find that there is substantial variation between sound categories (the range from the minimum to maximum value is shown by the red and blue shaded regions). However, for every sound category, the optimal time constants are typically larger for low-frequency sounds than for high-frequency sounds, suggesting that this is a consistent property of natural sounds.

To compare the time constants measured by Dean et al. (2008) with the time constant range that is optimal for natural sounds, we took the results of the above analysis and measured the whole range of time constants obtained (across all sound categories and values of T_{av} from 250 ms upward) for each frequency band. The results, plotted in Figure 5*C* (solid line indicates center; shaded region indicates range), are in good agreement with the time constants for IC neurons from Dean et al. (2008) (dots and regression line). The two datasets show very similar inverse relationships between time constant and frequency. Most of the IC data fall within the range of time constants that might be considered optimal. Nevertheless, there is significant variation, both within the data of Dean et al. (2008) and among the time constants for different sound categories. This may reflect optimization of subpopulations of neurons in IC for different subsets of natural sounds with different temporal autocorrelations. Overall, these results support our hypothesis that the time constants of adaptation to mean level in the IC are indeed optimized to enable neurons in the auditory midbrain to estimate and then subtract the mean sound level in each frequency band.

Discussion

Developing predictive, quantitative models of the response properties of individual neurons is fundamental to our ability to describe and understand information processing in the brain. Although STRF-based models have their limitations, they remain valuable and provide simple models for describing the behavior of sensory neurons. Many of the more sophisticated models (Sharpee et al., 2004, 2011; Ahrens et al., 2008; Atencio et al., 2009; Calabrese et al., 2011; Schinkel-Bielefeld et al., 2012) require many more parameters and may also be difficult to interpret biologically. Thus, a key challenge is to extend STRF models so that they more accurately describe neuronal behavior, while remaining simple and biologically relevant. Here we have shown that we can significantly improve STRF models of neurons in the auditory cortex by introducing a simple nonlinear input transform that reflects adaptation to stimulus statistics in the midbrain.

Advantages and power of nonlinear input stages

It is widely accepted that it can be useful to add a nonlinear output stage to the basic STRF model, resulting in an LN model (Chichilnisky, 2001). A nonlinear output stage has numerous advantages over the purely linear model. Sigmoid (or similar) functions can model threshold and saturation effects that are present in real neurons. However, the output of an LN model is only a simply transformed version of the output of the linear model. Introducing a nonlinear input stage is potentially far more powerful, because it allows the model to perform potentially quite complex nonlinear computations. In principle, unlimited nonlinear processing of the input can be performed before the linear summation stage, so that the full NLN model can perform complex computations. However, in practice it is difficult to harness this power because models that involve complex input transformations typically involve large numbers of free parameters, which are difficult to estimate given that neurophysiological datasets are always limited and noisy. We therefore need a way to introduce appropriate nonlinear transformations of the input without introducing many free parameters.

Here we have circumvented this problem by constraining the model nonlinear input transformations to replicate known operations of early stages of the sensory pathway. To the extent that the sensory systems are hierarchically organized, we can characterize the input by recording the response properties of neurons at earlier stages of processing. In the present case, we have used a model of the characteristics of neurons in the IC as the input to a model of the behavior of cortical neurons. Because the parameters of this input transformation can be determined by recording neuronal responses in the IC, it is possible to incorporate a well characterized transformation without introducing any free parameters whatsoever, which is what we have done here.

Relationship with other models

The IC Adaptation model includes a nonlinear input transformation with three components: adaptation to the stimulus mean, half-wave rectification, and frequency-dependent time constants of adaptation. Other models in the auditory literature have introduced nonlinear input stages to STRF models but have not examined this particular combination. For example, the synaptic depression model by David and colleagues (David et al., 2009; David and Shamma, 2013) contains similar adaptation to the stimulus level, but without half-wave rectification or frequency-dependent time constants. It is therefore similar to one of the controls used here. We find that both the half-wave rectification and the frequency-dependent time constants introduced by the IC Adaptation model significantly improve the power of the model to predict neural responses to most new stimuli.

The context model described by Ahrens et al. (2008) has a considerably greater model complexity than our IC Adaptation model and is therefore, in principle, capable of far more powerful nonlinear transformations, including nonmonotonicity. However, it also introduces many additional free parameters, which are difficult to fit reliably to neural data. This limits the improvements in predictive power that can be achieved in practice. The IC Adaptation model, in contrast, has no more free parameters than the LN model and therefore provides some of the benefits of the context model without introducing extra model complexity.

Time constants of adaptation

This study builds on the results of Dean et al. (2008), who found that neurons in the guinea pig IC adapt to the mean level of recent sound stimulation and that this adaptation has frequency-dependent time constants. We found that the specific time constants measured in that study were valuable in improving our models of ferret cortical neurons.

We show (Fig. 5C) that the time constants are optimized for a specific representation of the sound waveform. We assumed that the time-varying mean sound level in each frequency band, μ_{tf}, is a nuisance variable, which does not need to be included in the neuronal representation of sound. If this is the case, then a plausible role for adaptation to mean sound level in the IC is to subtract an estimate, $\hat{\mu}_{tf}$, of the time-varying mean (in a time window, T_{av}) from the sound level, X_{tf}, so that the responses of IC neurons are functions of $X_{tf} - \hat{\mu}_{tf}$ rather than μ_{tf} itself. We also assumed that the adaptation process estimates future values of $\hat{\mu}_{tf}$ by exponential averaging over recent values of X_{tf}. Finally, we assumed that the time constants of this adaptation process are optimized so that, for each frequency band, $\hat{\mu}_{tf}$ is as close as possible to the true mean, μ_{tf}. Using only these assumptions, we were able to derive optimal time constants for each frequency band in the spectrogram and found that these time constants are a good match for the real time constants measured in IC. This supports our hypothesis that adaptation in IC is optimized to subtract the time-varying mean in each frequency band of natural sounds.

It may initially seem surprising that time constants measured for neurons in the guinea pig IC should be relevant for a different species. However, since our natural sound analysis makes no assumptions that are specific to guinea pigs, the time constants should be similar for a range of species. While our assumption that mean sound level is a nuisance variable is a good general principle, we expect that there is some behaviorally valuable information in the absolute sound level that will also need to be transmitted. It is notable that in the auditory system adaptation is not complete, and even neurons which adapt optimally will not perfectly estimate and subtract μ_{tf}. As a result, some residual information about mean level will still be transmitted to the auditory cortex.

Future directions

The IC Adaptation model is a simple, easily implemented extension of classical STRF models of auditory neurons. It improves predictions of the behavior of neurons in the auditory cortex by incorporating a model of adaptation to stimulus statistics at an earlier stage of processing. Because it introduces no free parameters, it neither increases the complexity of the model nor the amount of data required to fit it.

In the present study, we have modeled adaptation to mean sound level in the IC and applied this to the responses of neurons in A1/AAF. It is likely that this work can be generalized to other structures in the auditory system. For example, further adaptation to stimulus mean is present in the responses of cortical neurons (Rabinowitz et al., 2013) and may have a thalamic or cortical origin. Future studies could experimentally characterize adaptation properties in these structures and use this to improve models of neurons in primary and higher auditory cortices.

In the visual system, adaptation to stimulus mean luminance is also present (Dawis et al., 1984; Rodieck, 1998; Mante et al., 2005) but is not yet routinely incorporated into receptive-field models of visual neurons (for review, see Sharpee, 2013). Nishimoto and Gallant (2011) found that incorporating luminance (and contrast) normalization did not significantly improve models of processing in MT; however, their model of normalization was a general one that did not closely match the characteristics of any particular adaptation process, and their stimuli contained a relatively narrow range of luminances. An approach similar to the one used here, using measured time constants of adaptation, may be better able to improve predictions of neural responses to motion and other visual stimuli.

Finally, neurons at multiple levels of the visual (Fairhall et al., 2001; Carandini and Heeger, 2012), somatosensory (Garcia-Lazaro et al., 2007), and auditory (Rabinowitz et al., 2011) systems show adaptation to higher-order statistics such as stimulus variance. Using an approach very similar to the one presented here, it should be possible to incorporate adaptation to higher-order stimulus statistics into neural models at multiple levels of different sensory pathways.

References

Aertsen AM, Johannesma PI (1981) The spectro-temporal receptive field. Biol Cybern 42:133–143. CrossRef Medline

Aertsen AMHJ, Johannesma PIM, Hermes DJ (1980) Spectro-temporal receptive fields of auditory neurons in the grassfrog. Biol Cybern 38: 235–248. CrossRef

Ahrens MB, Linden JF, Sahani M (2008) Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. J Neurosci 28:1929 –1942. CrossRef Medline

Atencio CA, Sharpee TO, Schreiner CE (2008) Cooperative nonlinearities in auditory cortical neurons. Neuron 58:956 –966. CrossRef Medline

Atencio CA, Sharpee TO, Schreiner CE (2009) Hierarchical computation in the canonical auditory cortical circuit. Proc Natl Acad Sci U S A 106: 21894 –21899. CrossRef Medline

Bizley JK, Walker KM, Silverman BW, King AJ, Schnupp JW (2009) Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. J Neurosci 29:2064 –2075. CrossRef Medline

Calabrese A, Schumacher JW, Schneider DM, Paninski L, Woolley SM (2011) A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. PLoS One 6:e16104. CrossRef Medline

Carandini M, Heeger DJ (2012) Normalization as a canonical neural computation. Nat Rev Neurosci 13:51– 62. CrossRef Medline

Chichilnisky EJ (2001) A simple white noise analysis of neuronal light responses. Network 12:199 –213. CrossRef Medline

Christianson GB, Sahani M, Linden JF (2008) The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. J Neurosci 28:446 – 455. CrossRef Medline

Dahmen JC, Keating P, Nodal FR, Schulz AL, King AJ (2010) Adaptation to stimulus statistics in the perception and neural representation of auditory space. Neuron 66:937–948. CrossRef Medline

David SV, Shamma SA (2013) Integration over multiple timescales in primary auditory cortex. J Neurosci 33:19154 –19166. CrossRef Medline

David SV, Mesgarani N, Fritz JB, Shamma SA (2009) Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J Neurosci 29:3374 –3386. CrossRef Medline

Dawis S, Shapley R, Kaplan E, Tranchina D (1984) The receptive field organization of X-cells in the cat: spatiotemporal coupling and asymmetry. Vision Res 24:549 –564. CrossRef Medline

Dean I, Harper NS, McAlpine D (2005) Neural population coding of sound level adapts to stimulus statistics. Nat Neurosci 8:1684 –1689. CrossRef Medline

Dean I, Robinson BL, Harper NS, McAlpine D (2008) Rapid neural adaptation to sound level statistics. J Neurosci 28:6430 – 6438. CrossRef Medline

deCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. Science 280:1439 –1443. CrossRef Medline

Escabí MA, Schreiner CE (2002) Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. J Neurosci 22:4114 – 4131. Medline

Fairhall AL, Lewen GD, Bialek W, de Ruyter Van Steveninck RR (2001) Efficiency and ambiguity in an adaptive neural code. Nature 412:787–792. CrossRef Medline

Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci 6:1216 –1223. CrossRef Medline

Garcia-Lazaro JA, Ho SS, Nair A, Schnupp JW (2007) Shifting and scaling adaptation to dynamic stimuli in somatosensory cortex. Eur J Neurosci 26:2359 –2368. CrossRef Medline

Gill P, Zhang J, Woolley SM, Fremouw T, Theunissen FE (2006) Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci 21:5–20. CrossRef Medline

Gourévitch B, Noreña A, Shaw G, Eggermont JJ (2009) Spectrotemporal receptive fields in anesthetized cat primary auditory cortex are context dependent. Cereb Cortex 19:1448 –1461. CrossRef Medline

Hsu A, Borst A, Theunissen FE (2004) Quantifying variability in neural responses and its application for the validation of model predictions. Network 15:91–109. CrossRef Medline

Joris PX, Yin TC (1992) Responses to amplitude-modulated tones in the auditory nerve of the cat. J Acoust Soc Am 91:215–232. CrossRef Medline

Kadir SN, Goodman DF, Harris KD (2014) High-dimensional cluster analysis with the masked EM algorithm. Neural Comput 26:2379 –2394. CrossRef Medline

Klein DJ, Depireux DA, Simon JZ, Shamma SA (2000) Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. J Comput Neurosci 9:85–111. CrossRef Medline

Kvale MN, Schreiner CE (2004) Short-term adaptation of auditory receptive fields to dynamic stimuli. J Neurophysiol 91:604 – 612. Medline

Linden JF, Liu RC, Sahani M, Schreiner CE, Merzenich MM (2003) Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. J Neurophysiol 90:2660 –2675. CrossRef Medline

Machens CK, Wehr MS, Zador AM (2004) Linearity of cortical receptive fields measured with natural sounds. J Neurosci 24:1089 –1100. CrossRef Medline

Malone BJ, Scott BH, Semple MN (2010) Temporal codes for amplitude contrast in auditory cortex. J Neurosci 30:767–784. CrossRef Medline

Mante V, Frazor RA, Bonin V, Geisler WS, Carandini M (2005) Independence of luminance and contrast in natural scenes and in the early visual system. Nat Neurosci 8:1690 –1697. CrossRef Medline

McDermott JH, Simoncelli EP (2011) Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron 71:926 –940. CrossRef Medline

Miller LM, Escabí MA, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87:516 –527. Medline

Nagel KI, Doupe AJ (2006) Temporal processing and adaptation in the songbird auditory forebrain. Neuron 51:845– 859. CrossRef Medline

Nelson PC, Carney LH (2007) Neural rate and timing cues for detection and discrimination of amplitude-modulated tones in the awake rabbit inferior colliculus. J Neurophysiol 97:522–539. CrossRef Medline

Nishimoto S, Gallant JL (2011) A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. J Neurosci 31:14551–14564. CrossRef Medline

Olshausen BA, Field DJ (2005) How close are we to understanding V1? Neural Comput 17:1665–1699. CrossRef Medline

Rabinowitz NC, Willmore BD, Schnupp JW, King AJ (2011) Contrast gain control in auditory cortex. Neuron 70:1178 –1191. CrossRef Medline

Rabinowitz NC, Willmore BD, Schnupp JW, King AJ (2012) Spectrotemporal contrast kernels for neurons in primary auditory cortex. J Neurosci 32:11271–11284. CrossRef Medline

Rabinowitz NC, Willmore BD, King AJ, Schnupp JW (2013) Constructing noise-invariant representations of sound in the auditory pathway. PLoS Biol 11:e1001710. CrossRef Medline

Rees A, Møller AR (1983) Responses of neurons in the inferior colliculus of the rat to AM and FM tones. Hear Res 10:301–330. CrossRef Medline

Rodieck RW (1998) The first steps in seeing. Sunderland, MA: Sinauer.

Rutkowski RG, Shackleton TM, Schnupp JW, Wallace MN, Palmer AR (2002) Spectrotemporal receptive field properties of single units in the primary, dorsocaudal and ventrorostral auditory cortex of the guinea pig. Audiol Neurootol 7:214 –227. CrossRef Medline

Sahani M, Linden JF (2003) How linear are auditory cortical responses? Adv Neural Inform Process Syst 15:125–132.

Schinkel-Bielefeld N, David SV, Shamma SA, Butts DA (2012) Inferring the role of inhibition in auditory processing of complex natural stimuli. J Neurophysiol 107:3296 –3307. CrossRef Medline

Schnupp JW, Mrsic-Flogel TD, King AJ (2001) Linear processing of spatial cues in primary auditory cortex. Nature 414:200 –204. CrossRef Medline

Sharpee T, Rust NC, Bialek W (2004) Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput 16:223–250. CrossRef Medline

Sharpee TO (2013) Computational identification of receptive fields. Annu Rev Neurosci 26:103–120. CrossRef Medline

Sharpee TO, Atencio CA, Schreiner CE (2011) Hierarchical representations in the auditory cortex. Curr Opin Neurobiol 21:761–767. CrossRef Medline

Simoncelli EP, Paninski L, Pillow J, Schwartz O (2004) Characterization of neural responses with stochastic stimuli. In: The cognitive neurosciences III (Gazzaniga MS, ed), pp 327–338. Cambridge, MA: MIT.

Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci 20:2315–2331. Medline

Touryan J, Felsen G, Dan Y (2005) Spatial structure of complex cell receptive fields measured with natural images. Neuron 45:781–791. CrossRef Medline

Wark B, Lundstrom BN, Fairhall A (2007) Sensory adaptation. Curr Opin Neurobiol 17:423– 429. CrossRef Medline

Watkins PV, Barbour DL (2008) Specialized neuronal adaptation for preserving input sensitivity. Nat Neurosci 11:1259 –1261. CrossRef Medline

Wen B, Wang GI, Dean I, Delgutte B (2009) Dynamic range adaptation to sound level statistics in the auditory nerve. J Neurosci 29:13797–13808. CrossRef Medline

Willmore B, Smyth D (2003) Methods for first-order kernel estimation: simple-cell receptive fields from responses to natural scenes. Network 14:553–577. CrossRef Medline

Willmore BD, Prenger RJ, Gallant JL (2010) Neural representation of natural images in visual area V2. J Neurosci 30:2102–2114. CrossRef Medline

B. Appendix B: overview of further work not included in this dissertation

B.1. Deep learning-enabled organ segmentation with uncertainty quantification of whole-body mouse scans

Authors: O Schoppe, C Pan, J Coronel, H Mai, Z Rong, M Todorov, A Müskes, F Navarro, A Ertürk, and B H Menze

Abstract: Whole-body imaging of mice is a key source of information for research. Segmentation of major organs in such scans is a prerequisite for quantitative analysis but is a tedious and error-prone task if done manually. Here, we present a deep learning solution called AIMOS that automatically segments major organs (brain, lungs, heart, liver, kidneys, spleen) and the skeleton with unrivalled accuracy. AIMOS segments a whole-body scan in less than a second, orders of magnitude faster than prior algorithms, and matches or exceeds the segmentation quality of state-of-the-art approaches and the segmentation quality of human experts. We demonstrate direct applicability for biomedical research with an exemplary analysis of the biodistribution of cancer metastases. Furthermore, we show that expert annotations are subject to human error and bias. Importantly, AIMOS addresses this issue by identifying the regions where humans are most likely to disagree, and thereby localises and quantifies this uncertainty for improved downstream analysis. In summary, AIMOS is a powerful open-source tool to increase scalability, reduce bias, and foster reproducibility in many areas of biomedical research.

Date of submission: 30.04.2020
Journal name: Nature Communications (Springer Nature)
Publication status: In revision

Individual contribution: project conception and coordination, experimental design, data analysis, leading author of manuscript

B.2. Machine learning analysis of whole mouse brain vasculature

Authors: M Todorov*, J Paetzold*, **O Schoppe**, G Tetteh, V Efremov, K Voelgyi, M Duering, M Dichgans, M Piraud, B Menze, A Ertürk
Joint first authorship

Abstract: Tissue clearing methods enable the imaging of biological specimens without sectioning. However, reliable and scalable analysis of large imaging datasets in three dimensions remains a challenge. Here we developed a deep learning-based framework to quantify and analyze brain vasculature, named Vessel Segmentation & Analysis Pipeline (VesSAP). Our pipeline uses a convolutional neural network (CNN) with a transfer learning approach for segmentation and achieves human-level accuracy. By using VesSAP, we analyzed the vascular features of whole C57BL/6J, CD1 and BALB/c mouse brains at the micrometer scale after registering them to the Allen mouse brain atlas. We report evidence of secondary intracranial collateral vascularization in CD1 mice and find reduced vascularization of the brainstem in comparison to the cerebrum. Thus, VesSAP enables unbiased and scalable quantifications of the angioarchitecture of cleared mouse brains and yields biological insights into the vascular function of the brain.

Date of publication: 11.03.2020
Journal name: Nature Methods (Springer Nature)

Individual contribution: support with project coordination, support with experimental design, support with data analysis, support for creation and revision of manuscript

B.3. Cellular and Molecular Probing of Intact Human Organs

Authors: S Zhao, M Todorov, R Cai, R al-Maskari, H Steinke, E Kemter, H Mai, Z Rong, M Warmer, K Stanic, **O Schoppe**, J Paetzold, B Gesierich, M Wong, T Huber, M Duering, O Bruns, B Menze, J Lipfert, V Puelles, E Wolf, I Bechmann, A Ertürk.

Abstract: Optical tissue transparency permits scalable cellular and molecular investigation of complex tissues in 3D. Adult human organs are particularly challenging to render transparent because of the accumulation of dense and sturdy molecules in decades-aged tissues. To overcome these challenges, we developed SHANEL, a method based on a new tissue permeabilization approach to clear and label stiff human organs. We used SHANEL to render the intact adult human brain and kidney transparent and perform 3D histology with antibodies and dyes in centimeters-depth. Thereby, we revealed structural details of the intact human eye, human thyroid, human kidney, and transgenic pig pancreas at the cellular resolution. Furthermore, we developed a deep learning pipeline to analyze millions of cells in cleared human brain tissues within hours with standard lab computers. Overall, SHANEL is a robust and unbiased technology to chart the cellular and molecular architecture of large intact mammalian organs.

Date of publication: 20.02.2020
Journal name: Cell, Cell Press (Elsevier)

Individual contribution: coordination of machine learning work (data processing, quantitative experiments), support for manuscript revision

B.4. Gold Nanoparticle Mediated Multi-Modal CT Imaging of Hsp70 Membrane Positive Tumors

Authors: M Kimm, M Shevtsov, C Werner, W Sievert, Z Wu, **O Schoppe**, B Menze, E Rummeny, R Proksa, O Bystrova, M Martynova, G Multhoff, S Stangl

Abstract: Imaging techniques such as computed tomographies (CT) play a major role in clinical imaging and diagnosis of malignant lesions. In recent years, metal nanoparticle platforms enabled effective payload delivery for several imaging techniques. Due to the possibility of surface modification, metal nanoparticles are predestined to facilitate molecular tumor targeting. In this work, we demonstrate the feasibility of anti-plasma membrane Heat shock protein 70 (Hsp70) antibody functionalized gold nanoparticles (cmHsp70.1-AuNPs) for tumor-specific multimodal imaging. Membrane-associated Hsp70 is exclusively presented on the plasma membrane of malignant cells of multiple tumor entities but not on corresponding normal cells, predestining this target for a tumor-selective in vivo imaging. In vitro microscopic analysis revealed the presence of cmHsp70.1-AuNPs in the cytosol of tumor cell lines after internalization via the endo-lysosomal pathway. In preclinical models, the biodistribution as well as the intratumoral enrichment of AuNPs were examined 24 h after i.v. injection in tumor-bearing mice. In parallel to spectral CT analysis, histological analysis confirmed the presence of AuNPs within tumor cells. In contrast to control AuNPs, a significant enrichment of cmHsp70.1-AuNPs has been detected selectively inside tumor cells in different tumor mouse models. Furthermore, a machine-learning approach was developed to analyze AuNP accumulations in tumor tissues and organs. In summary, utilizing mHsp70 on tumor cells as a target for the guidance of cmHsp70.1-AuNPs facilitates an enrichment and uniform distribution of nanoparticles in mHsp70-expressing tumor cells that enables various microscopic imaging techniques and spectral-CT-based tumor delineation in vivo.

Date of publication: 22.05.2020
Journal name: Cancers, MDPI

Individual contribution: coordination and implementation of machine learning work (data processing, quantitative experiments), support with writing the manuscript

B.5. Red-GAN: Attacking class imbalance via conditioned generation. Yet another medical imaging perspective

Authors: A Qasim*, I Ezhov*, S Shit, **O Schoppe**, J Paetzold, A Sekuboyina, F Kofler, J Lipkova, H Li, B Menze
**Joint first authorship*

Abstract: Exploiting learning algorithms under scarce data regimes is a limitation and a reality of the medical imaging field. In an attempt to mitigate the problem, we propose a data augmentation protocol based on generative adversarial networks. We condition the networks at a pixel-level (segmentation mask) and at a global-level information (acquisition environment or lesion type). Such conditioning provides immediate access to the image-label pairs while controlling global class specific appearance of the synthesized images. To stimulate synthesis of the features relevant for the segmentation task, an additional passive player in a form of segmentor is introduced into the adversarial game. We validate the approach on two medical datasets: BraTS, ISIC. By controlling the class distribution through injection of synthetic images into the training set we achieve control over the accuracy levels of the datasets' classes.

Date of publication: 06.07.2020
Conference name: International Conference on Medical Imaging with Deep Learning, 6-9 July 2020, Montréal, Canada (peer-reviewed paper)

Individual contribution: support for creation and revision of manuscript

Bibliography

[1] A. Akselrod-Ballin, H. Dafni, Y. Addadi, I. Biton, R. Avni, Y. Brenner, and M. Neeman. "Multimodal correlative preclinical whole body imaging and segmentation". In: *Scientific reports* 6 (2016), p. 27940.

[2] M. Baiker, J. Milles, J. Dijkstra, T. D. Henning, A. W. Weber, I. Que, E. L. Kaijzel, C. W. Lowik, J. H. Reiber, and B. P. Lelieveldt. "Atlas-based whole-body segmentation of mice from low-contrast Micro-CT data". In: *Medical image analysis* 14.6 (2010), pp. 723–737.

[3] N. Barker and H. Clevers. "Mining the Wnt pathway for cancer therapeutics". In: *Nat Rev Drug Discov* 5.12 (Dec. 2006), pp. 997–1014. ISSN: 1474-1776 (Print) 1474-1776 (Linking). DOI: 10.1038/nrd2154.

[4] C. Battke, E. Kremmer, J. Mysliwietz, G. Gondi, C. Dumitru, S. Brandau, S. Lang, D. Vullo, C. Supuran, and R. Zeidler. "Generation and characterization of the first inhibitory antibody targeting tumour-associated carbonic anhydrase XII". In: *Cancer Immunol Immunother* 60.5 (May 2011), pp. 649–58. ISSN: 1432-0851 (Electronic) 0340-7004 (Linking). DOI: 10.1007/s00262-011-0980-z.

[5] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hotker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu. "Phiseg: Capturing uncertainty in medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 119–127.

[6] N. Beckmann, R. Kneuer, H.-U. Gremlich, H. Karmouty-Quintana, F.-X. Ble, and M. Muller. "In vivo mouse imaging and spectroscopy in drug discovery". In: *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo* 20.3 (2007), pp. 154–185.

[7] S. Bhatia, Y. Sinha, and L. Goel. "Lung Cancer Detection: A Deep Learning Approach". In: ed. by J. C. Bansal, K. N. Das, A. Nagar, K. Deep, and A. K. Ojha. Soft Computing for Problem Solving. Springer Singapore, 2019, pp. 699–705.

[8] S. Bolte and F. Cordelieres. "A guided tour into subcellular colocalization analysis in light microscopy". In: *Journal of Microscopy* 224.3 (2006), pp. 213–232. DOI: doi:10.1111/j.1365-2818.2006.01706.x.

[9] Y. Boykov and M.-P. Jolly. "Interactive organ segmentation using graph cuts". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2000, pp. 276–286.

[10] M. Brett, M. Hanke, B. Cipollini, M.-A. Cote, C. Markiewicz, S. Gerhard, E. Larson, G. R. Lee, Y. Halchenko, E. Kastman, et al. "nibabel: 2.1. 0". In: *Zenodo* (2016).

[11] J. von Burstin, S. Eser, M. C. Paul, B. Seidler, M. Brandl, M. Messer, A. von Werder, A. Schmidt, J. Mages, P. Pagel, A. Schnieke, R. M. Schmid, G. Schneider, and D. Saur. "E-cadherin regulates metastasis of pancreatic cancer in vivo and is suppressed by a SNAIL/HDAC1/HDAC2 repressor complex". eng. In: *Gastroenterology* 137.1 (July 2009), 361–71, 371 e1–5. ISSN: 1528-0012 (Electronic) 0016-5085 (Linking). DOI: `10.1053/j.gastro.2009.04.004`.

[12] J. von Burstin, S. Eser, B. Seidler, A. Meining, M. Bajbouj, J. Mages, R. Lang, A. J. Kind, A. E. Schnieke, R. M. Schmid, G. Schneider, and D. Saur. "Highly sensitive detection of early-stage pancreatic cancer by multimodal near-infrared molecular imaging in living mice". eng. In: *Int J Cancer* 123.9 (Nov. 2008), pp. 2138–47. ISSN: 1097-0215 (Electronic) 0020-7136 (Linking). DOI: `10.1002/ijc.23780`.

[13] R. Cai, C. Pan, A. Ghasemigharagoz, M. I. Todorov, B. Forstera, S. Zhao, H. S. Bhatia, A. Parra-Damas, L. Mrowka, D. Theodorou, M. Rempfler, A. L. R. Xavier, B. T. Kress, C. Benakis, H. Steinke, S. Liebscher, I. Bechmann, A. Liesz, B. Menze, M. Kerschensteiner, M. Nedergaard, and A. Erturk. "Panoptic imaging of transparent mice reveals whole-body neuronal projections and skull-meninges connections". In: *Nat Neurosci* (Dec. 2018). ISSN: 1546-1726 (Electronic) 1097-6256 (Linking). DOI: `10.1038/s41593-018-0301-3`.

[14] R. Cai. "DISCO whole body clearing and imaging to study systemic changes in neuronal pathologies". PhD thesis. lmu, 2019.

[15] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, and J. J. Collins. "Next-Generation Machine Learning for Biological Networks". In: *Cell* 173.7 (June 2018), pp. 1581–1592. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking). DOI: `10.1016/j.cell.2018.05.015`.

[16] J. P. Campbell, A. R. Merkel, S. K. Masood-Campbell, F. Elefteriou, and J. A. Sterling. "Models of bone metastasis". eng. In: *J Vis Exp* 67 (Sept. 2012), e4260. ISSN: 1940-087X (Electronic) 1940-087X (Linking). DOI: `10.3791/4260`.

[17] S. Carregal-Romero, S. Plaza-Garcıa, R. Pinol, J. L. Murillo, J. Ruiz-Cabello, D. Padro, A. Millan, and P. Ramos-Cabrer. "MRI Study of the Influence of Surface Coating Aging on the In Vivo Biodistribution of Iron Oxide Nanoparticles". In: *Biosensors* 8.4 (2018), p. 127.

[18] E. M. Christiansen, S. J. Yang, D. M. Ando, A. Javaherian, G. Skibinski, S. Lipnick, E. Mount, A. ONeil, K. Shah, A. K. Lee, P. Goyal, W. Fedus, R. Poplin, A. Esteva, M. Berndl, L. L. Rubin, P. Nelson, and S. Finkbeiner. "In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images". In: *Cell* 173.3 (Apr. 2018), 792–803 e19. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking). DOI: 10.1016/j.cell.2018.03.040.

[19] K. Chung and K. Deisseroth. "CLARITY for mapping the nervous system". In: *Nature methods* 10.6 (2013), p. 508.

[20] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation". In: ed. by S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Springer International Publishing, 2016, pp. 424–432. ISBN: 978-3-319-46723-8.

[21] L. Clarke, M. Silbiger, C. Naylor, and K. Brown. "Artificial neural net system for interactive tissue classification with MR imaging and image segmentation". In: *Seventy sixth scientific assembly and annual meeting of the Radiological Society of North America*. 1990.

[22] J. Condeelis and R. Weissleder. "In vivo imaging in cancer". In: *Cold Spring Harb Perspect Biol* 2.12 (Dec. 2010), a003848. ISSN: 1943-0264 (Electronic) 1943-0264 (Linking). DOI: 10.1101/cshperspect.a003848.

[23] N. Davoudi, X. L. Dean-Ben, and D. Razansky. "Deep learning optoacoustic tomography with sparse data". In: *Nature Machine Intelligence* 1.10 (2019), pp. 453–460.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[25] T. Derlin, K. Peldschus, S. Munster, P. Bannas, J. Herrmann, T. Stubig, C. R. Habermann, G. Adam, N. Kroger, and C. Weber. "Comparative diagnostic performance of 18 F-FDG PET/CT versus whole-body MRI for determination of remission status in multiple myeloma after stem cell transplantation". In: *European radiology* 23.2 (2013), pp. 570–578.

[26] A. dEsposito, P. W. Sweeney, M. Ali, M. Saleh, R. Ramasawmy, T. A. Roberts, G. Agliardi, A. Desjardins, M. F. Lythgoe, R. B. Pedley, et al. "Computational fluid dynamics with imaging of cleared tissue and of in vivo perfusion predicts drug uptake and treatment responses in tumours". In: *Nature biomedical engineering* 2.10 (2018), pp. 773–787.

[27] T. Di Noia, V. C. Ostuni, F. Pesce, G. Binetti, D. Naso, F. P. Schena, and E. Di Sciascio. "An end stage kidney disease predictor based on an artificial neural networks ensemble". In: *Expert systems with applications* 40.11 (2013), pp. 4438–4445.

[28] H.-U. Dodt, U. Leischner, A. Schierloh, N. Jährling, C. P. Mauch, K. Deininger, J. M. Deussing, M. Eder, W. Zieglgänsberger, and K. Becker. "Ultramicroscopy: three-dimensional visualization of neuronal networks in the whole mouse brain". In: *Nature methods* 4.4 (2007), pp. 331–336.

[29] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. van der Laak, C. C. the, M. Hermsen, Q. F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. C. van Dijk, P. Bult, F. Beca, A. H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H. J. Lin, P. A. Heng, C. Hass, E. Bruni, Q. Wong, U. Halici, M. U. Oner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y. W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. Ahmady Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, and R. Venancio. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer". In: *JAMA* 318.22 (Dec. 2017), pp. 2199–2210. ISSN: 1538-3598 (Electronic) 0098-7484 (Linking). DOI: `10.1001/jama.2017.14585`.

[30] K. N. Elfer, A. B. Sholl, M. Wang, D. B. Tulman, S. H. Mandava, B. R. Lee, and J. Q. Brown. "DRAQ5 and eosin (D&E) as an analog to hematoxylin and eosin for rapid fluorescence histology of fresh tissues". In: *PLoS One* 11.10 (2016), e0165530.

[31] A. Erturk, K. Becker, N. Jahrling, C. P. Mauch, C. D. Hojer, J. G. Egen, F. Hellal, F. Bradke, M. Sheng, and H.-U. Dodt. "Three-dimensional imaging of solvent-cleared organs using 3DISCO". In: *Nature protocols* 7.11 (2012), p. 1983.

[32] S. Eser, N. Reiff, M. Messer, B. Seidler, K. Gottschalk, M. Dobler, M. Hieber, A. Arbeiter, S. Klein, B. Kong, C. W. Michalski, A. M. Schlitter, I. Esposito, A. J. Kind, L. Rad, A. E. Schnieke, M. Baccarini, D. R. Alessi, R. Rad, R. M. Schmid,

G. Schneider, and D. Saur. "Selective requirement of PI3K/PDK1 signaling for Kras oncogene-driven pancreatic cell plasticity and cancer". eng. In: *Cancer Cell* 23.3 (Mar. 2013), pp. 406–20. ISSN: 1878-3686 (Electronic) 1535-6108 (Linking). DOI: 10.1016/j.ccr.2013.01.023.

[33] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (Feb. 2017), pp. 115–118. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/nature21056.

[34] S. D. Fabiyi. "A review of unsupervised Artificial Neural Networks with applications". In: *International Journal of Computer Applications* 181.40 (2019), pp. 22–26.

[35] T. Falk, D. Mai, R. Bensch, O. Cicek, A. Abdulkadir, Y. Marrakchi, A. Bohm, J. Deubner, Z. Jackel, K. Seiwald, A. Dovzhenko, O. Tietz, C. Dal Bosco, S. Walsh, D. Saltukoglu, T. L. Tay, M. Prinz, K. Palme, M. Simons, I. Diester, T. Brox, and O. Ronneberger. "U-Net: deep learning for cell counting, detection, and morphometry". In: *Nat Methods* 16.1 (Jan. 2019), pp. 67–70. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking). DOI: 10.1038/s41592-018-0261-2.

[36] G. Gondi, J. Mysliwietz, A. Hulikova, J. P. Jen, P. Swietach, E. Kremmer, and R. Zeidler. "Antitumor efficacy of a monoclonal antibody that inhibits the activity of cancer-associated carbonic anhydrase XII". eng. In: *Cancer Res* 73.21 (Nov. 2013), pp. 6494–503. ISSN: 1538-7445 (Electronic) 0008-5472 (Linking). DOI: 10.1158/0008-5472.CAN-13-1110.

[37] T. S. Gunawan, I. Z. Yaacob, M. Kartiwi, N. Ismail, N. F. Za'bah, and H. Mansor. "Artificial neural network based fast edge detection algorithm for mri medical images". In: *Indonesian Journal of Electrical Engineering and Computer Science* 7.1 (2017), pp. 123–130.

[38] D. Hanahan and R. A. Weinberg. "Hallmarks of cancer: the next generation". In: *Cell* 144.5 (Mar. 2011), pp. 646–74. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking). DOI: 10.1016/j.cell.2011.02.013.

[39] M. P. Heinrich, O. Oktay, and N. Bouteldja. "OBELISK-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions". In: *Medical image analysis* 54 (2019), pp. 1–9.

[40] T. Hemalatha, P. Prabu, D. N. Gunadharini, N. R. Kamini, and M. K. Gowthaman. "Dual acting methotrexate conjugated nanocomposite for MR and CT imaging: Perspectives on therapeutic efficacy and in vivo biodistribution". In: *Materials Letters* 255 (2019), p. 126583.

[41] S. R. Hingorani, E. F. Petricoin, A. Maitra, V. Rajapakse, C. King, M. A. Jacobetz, S. Ross, T. P. Conrads, T. D. Veenstra, B. A. Hitt, Y. Kawaguchi, D. Johann, L. A. Liotta, H. C. Crawford, M. E. Putt, T. Jacks, C. V. Wright, R. H. Hruban, A. M. Lowy, and D. A. Tuveson. "Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse". eng. In: *Cancer Cell* 4.6 (Dec. 2003), pp. 437–50. ISSN: 1535-6108 (Print) 1535-6108 (Linking).

[42] S. Hirsch, J. Reichold, M. Schneider, G. Szekely, and B. Weber. "Topology and hemodynamics of the cortical cerebrovascular system". In: *Journal of Cerebral Blood Flow & Metabolism* 32.6 (2012), pp. 952–967.

[43] D. W. Holdsworth and M. M. Thornton. "Micro-CT in small animal and specimen imaging". In: *Trends in Biotechnology* 20.8 (2002), S34–S39.

[44] G. Huang, T. Zhao, C. Wang, K. Nham, Y. Xiong, X. Gao, Y. Wang, G. Hao, W.-P. Ge, X. Sun, et al. "PET imaging of occult tumours by temporal integration of tumour-acidosis signals from pH-sensitive 64 Cu-labelled polymers". In: *Nature biomedical engineering* 4.3 (2020), pp. 314–324.

[45] J. Huisken and D. Y. Stainier. "Selective plane illumination microscopy techniques in developmental biology". In: *Development* 136.12 (2009), pp. 1963–1975.

[46] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in science & engineering* 9.3 (2007), pp. 90–95.

[47] D. Inderbitzin, M. Gass, G. Beldi, E. Ayouni, A. Nordin, D. Sidler, B. Gloor, D. Candinas, and C. Stoupis. "Magnetic resonance imaging provides accurate and precise volume determination of the regenerating mouse liver". In: *Journal of gastrointestinal surgery* 8.7 (2004), pp. 806–811.

[48] E. Iorns, K. Drews-Elger, T. M. Ward, S. Dean, J. Clarke, D. Berry, D. El Ashry, and M. Lippman. "A new mouse model for the study of human breast cancer metastasis". In: *PLoS One* 7.10 (2012), e47995. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: `10.1371/journal.pone.0047995`.

[49] B. Irie and S. Miyake. "Capabilities of three-layered perceptrons." In: *ICNN*. 1988, pp. 641–648.

[50] E. Jones, T. Oliphant, and P. Peterson. "SciPy: Open Source Scientific Tools for Python". In: URL http://www.scipy.org/ (2001).

[51] T. N. Jones and D. N. Metaxas. "Automated 3D segmentation using deformable models and fuzzy affinity". In: *Biennial International Conference on Information Processing in Medical Imaging*. Springer. 1997, pp. 113–126.

[52] M. de Jong, J. Essers, and W. M. van Weerden. "Imaging preclinical tumour models: improving translational power". In: *Nat Rev Cancer* 14.7 (July 2014), pp. 481–93. ISSN: 1474-1768 (Electronic) 1474-175X (Linking). DOI: `10.1038/nrc3751`.

[53] A. A. Joshi, A. J. Chaudhari, C. Li, D. W. Shattuck, J. Dutta, R. M. Leahy, and A. W. Toga. "Posture matching and elastic registration of a mouse atlas to surface topography range data". In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2009, pp. 366–369.

[54] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna. "Inter-observer variability of manual contour delineation of structures in CT". In: *European radiology* 29.3 (2019), pp. 1391–1399.

[55] A. Joutel, M. Monet-Lepretre, C. Gosele, C. Baron-Menguy, A. Hammes, S. Schmidt, B. Lemaire-Carrette, V. Domenga, A. Schedl, P. Lacombe, et al. "Cerebrovascular dysfunction and microcirculation rarefaction precede white matter lesions in a mouse genetic model of cerebral ischemic small vessel disease". In: *The Journal of clinical investigation* 120.2 (2010), pp. 433–445.

[56] A. Jungo, R. Meier, E. Ermis, M. Blatti-Moreno, E. Herrmann, R. Wiest, and M. Reyes. "On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 682–690.

[57] G. C. Kagadis, G. Loudos, K. Katsanos, S. G. Langer, and G. C. Nikiforidis. "In vivo small animal imaging: current status and future prospects". In: *Medical physics* 37.12 (2010), pp. 6421–6442.

[58] E. L. Kaijzel, G. van Der Pluijm, and C. W. Lowik. "Whole-body optical imaging in animal models to assess cancer development and progression". In: *Clinical Cancer Research* 13.12 (2007), pp. 3490–3497.

[59] H. Kantamneni, M. Zevon, M. J. Donzanti, X. Zhao, Y. Sheng, S. R. Barkund, L. H. McCabe, W. Banach-Petrosky, L. M. Higgins, S. Ganesan, et al. "Surveillance nanotechnology for multi-organ cancer metastases". In: *Nature biomedical engineering* 1.12 (2017), pp. 993–1003.

[60] N. Karssemeijer. "Three-dimensional stochastic organ-models for segmentation in CT-scans". In: *Biostereometrics 88*. Vol. 1030. International Society for Optics and Photonics. 1989, pp. 177–184.

[61] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. N. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, and K. Zhang. "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning". In: *Cell* 172.5 (Feb. 2018), 1122–1131 e9. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking). DOI: `10.1016/j.cell.2018.02.010`.

[62] S. Khan, N. Islam, Z. Jan, I. U. Din, and J. J. C. Rodrigues. "A Novel Deep Learning based Framework for the Detection and Classification of Breast Cancer Using Transfer Learning". In: *Pattern Recognition Letters* (2019).

[63] A. Khmelinskii, M. Baiker, E. L. Kaijzel, J. Chen, J. H. Reiber, and B. P. Lelieveldt. "Articulated whole-body atlases for small animal image analysis: construction and applications". In: *Molecular imaging and biology* 13.5 (2011), pp. 898–910.

[64] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv e-prints* (Dec. 2014).

[65] G. Knittel, T. Rehkaemper, D. Korovkina, P. Liedgens, C. Fritz, A. Torgovnick, Y. Al-Baldawi, M. Al-Maarri, Y. Cun, O. Fedorchenko, et al. "Two mouse models reveal an actionable PARP1 dependence in aggressive chronic lymphocytic leukemia". In: *Nature communications* 8.1 (2017), pp. 1–13.

[66] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. A. Eslami, D. J. Rezende, and O. Ronneberger. "A probabilistic U-Net for segmentation of ambiguous images". In: *Advances in Neural Information Processing Systems*. 2018, pp. 6965–6975.

[67] S. I. Kubota, K. Takahashi, J. Nishida, Y. Morishita, S. Ehata, K. Tainaka, K. Miyazono, and H. R. Ueda. "Whole-Body Profiling of Cancer Metastasis with Single-Cell Resolution". eng. In: *Cell Rep* 20.1 (July 2017), pp. 236–250. ISSN: 2211-1247 (Electronic). DOI: `10.1016/j.celrep.2017.06.010`.

[68] A. W. Lambert, D. R. Pattabiraman, and R. A. Weinberg. "Emerging Biological Principles of Metastasis". eng. In: *Cell* 168.4 (Feb. 2017), pp. 670–691. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking). DOI: `10.1016/j.cell.2016.11.037`.

[69] D. T. Lauber, A. Fulop, T. Kovacs, K. Szigeti, D. Mathe, and A. Szijarto. "State of the art in vivo imaging techniques for laboratory animals". In: *Laboratory animals* 51.5 (2017), pp. 465–478.

[70] A. Le Bras. "An automated pipeline for metastasis detection". In: *Lab Animal* 49.2 (2020), pp. 48–48.

[71] F. Leblond, S. C. Davis, P. A. Valdes, and B. W. Pogue. "Pre-clinical whole-body fluorescence imaging: Review of instruments, methods and applications". In: *Journal of photochemistry and photobiology B: Biology* 98.1 (2010), pp. 77–94.

[72] L. Lenk, M. Pein, O. Will, B. Gomez, F. Viol, C. Hauser, J. H. Egberts, J. P. Gundlach, O. Helm, S. Tiwari, R. Weiskirchen, S. Rose-John, C. Rocken, W. Mikulits, P. Wenzel, G. Schneider, D. Saur, H. Schafer, and S. Sebens. "The hepatic microenvironment essentially determines tumor cell dormancy and metastatic outgrowth of pancreatic ductal adenocarcinoma". eng. In: *Oncoimmunology* 7.1 (2017), e1368603. ISSN: 2162-4011 (Print) 2162-4011 (Linking). DOI: 10.1080/2162402X.2017.1368603.

[73] L. Li, L. Zhu, C. Ma, L. Lin, J. Yao, L. Wang, K. Maslov, R. Zhang, W. Chen, J. Shi, et al. "Single-impulse panoramic photoacoustic computed tomography of small-animal whole-body dynamics at high spatiotemporal resolution". In: *Nature biomedical engineering* 1.5 (2017), pp. 1–11.

[74] W. Li, R. Prakash, A. I. Kelly-Cobbs, S. Ogbi, A. Kozak, A. B. El-Remessy, D. A. Schreihofer, S. C. Fagan, and A. Ergul. "Adaptive cerebral neovascularization in a model of type 2 diabetes: relevance to focal cerebral ischemia". In: *Diabetes* 59.1 (2010), pp. 228–235.

[75] O. Liba and A. de la Zerda. "Photoacoustic tomography: breathtaking whole-body imaging". In: *Nature Biomedical Engineering* 1.5 (2017), pp. 1–3.

[76] G. Litjens, C. I. Sanchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-van de Kaa, P. Bult, B. van Ginneken, and J. van der Laak. "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis". eng. In: *Sci Rep* 6 (May 2016), p. 26286. ISSN: 2045-2322 (Electronic) 2045-2322 (Linking). DOI: 10.1038/srep26286.

[77] Y. Liu, T. Kohlberger, M. Norouzi, G. E. Dahl, J. L. Smith, A. Mohtashamian, N. Olson, L. H. Peng, J. D. Hipp, and M. C. Stumpe. "Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection". In: *Arch Pathol Lab Med* (Oct. 2018). ISSN: 1543-2165 (Electronic) 0003-9985 (Linking). DOI: 10.5858/arpa.2018-0147-OA.

[78] Y. Liu, M. Gargesha, M. Qutaish, Z. Zhou, B. Scott, H. Yousefi, Z. Lu, and D. L. Wilson. "Deep learning based multi-organ segmentation and metastases segmentation in whole mouse body and the cryo-imaging cancer imaging and therapy analysis platform (CITAP)". In: *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*. Vol. 11317. International Society for Optics and Photonics. 2020, p. 113170V.

[79] L. Louhivuori, S. Kanatani, and P. Uhlen. “Predicting a tumours drug uptake”. In: *Nature biomedical engineering* 2.10 (2018), pp. 717–718.

[80] A. Maier, C. Syben, T. Lasser, and C. Riess. “A gentle introduction to deep learning in medical image processing”. In: *Zeitschrift für Medizinische Physik* 29.2 (2019), pp. 86–101.

[81] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid. “Convolutional kernel networks”. In: *Advances in neural information processing systems*. 2014, pp. 2627–2635.

[82] N. Malpica, C. O. De Solorzano, J. J. Vaquero, A. Santos, I. Vallcorba, J. M. Garcia-Sagredo, and F. Del Pozo. “Applying watershed algorithms to the segmentation of clustered nuclei”. In: *Cytometry: The Journal of the International Society for Analytical Cytology* 28.4 (1997), pp. 289–297.

[83] T. Mano, A. Albanese, H.-U. Dodt, A. Erturk, V. Gradinaru, J. B. Treweek, A. Miyawaki, K. Chung, and H. R. Ueda. “Whole-Brain Analysis of Cells and Circuits by Tissue Clearing and Light-Sheet Microscopy”. In: *Journal of Neuroscience* 38.44 (2018), pp. 9330–9337.

[84] B. Masi, T.-A. Perles-Barbacaru, C. Laprie, H. Dessein, M. Bernard, A. Dessein, and A. Viola. “In vivo MRI assessment of hepatic and splenic disease in a murine model of schistosmiasis”. In: *PLoS neglected tropical diseases* 9.9 (2015).

[85] J. Massague and A. C. Obenauf. “Metastatic colonization by circulating tumour cells”. In: *Nature* 529.7586 (Jan. 2016), pp. 298–306. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: `10.1038/nature17038`.

[86] T. F. Massoud and S. S. Gambhir. “Integrating noninvasive molecular imaging into molecular medicine: an evolving paradigm”. In: *Trends Mol Med* 13.5 (May 2007), pp. 183–91. ISSN: 1471-4914 (Print) 1471-4914 (Linking). DOI: `10.1016/j.molmed.2007.03.003`.

[87] T. F. Massoud and S. S. Gambhir. “Molecular imaging in living subjects: seeing fundamental biological processes in a new light”. In: *Genes Dev* 17.5 (Mar. 2003), pp. 545–80. ISSN: 0890-9369 (Print) 0890-9369 (Linking). DOI: `10.1101/gad.1047403`.

[88] S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, U. Sikora, O. Yaglidere, S. Padmanabhan, K. Nielsen, and A. Ozcan. “Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study”. In: *PLoS One* 7.5 (2012), e37245. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: `10.1371/journal.pone.0037245`.

[89] S. Mavandadi, S. Feng, F. Yu, S. Dimitrov, K. Nielsen-Saines, W. R. Prescott, and A. Ozcan. "A mathematical framework for combining decisions of multiple experts toward accurate and remote diagnosis of malaria using tele-microscopy". In: *PLoS One* 7.10 (2012), e46192. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: `10.1371/journal.pone.0046192`.

[90] W. McKinney. "Pandas". In: URL https://pandas.pydata.org/ (2008).

[91] B. von Neubeck, G. Gondi, C. Riganti, C. Pan, A. Parra Damas, H. Scherb, A. Erturk, and R. Zeidler. "An inhibitory antibody targeting carbonic anhydrase XII abrogates chemoresistance and significantly reduces lung metastases in an orthotopic breast cancer model in vivo". In: *Int J Cancer* 143.8 (Oct. 2018), pp. 2065–2075. ISSN: 1097-0215 (Electronic) 0020-7136 (Linking). DOI: `10.1002/ijc.31607`.

[92] D. X. Nguyen, A. C. Chiang, X. H. Zhang, J. Y. Kim, M. G. Kris, M. Ladanyi, W. L. Gerald, and J. Massague. "WNT/TCF signaling through LEF1 and HOXB9 mediates lung adenocarcinoma metastasis". eng. In: *Cell* 138.1 (July 2009), pp. 51–62. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking). DOI: `10.1016/j.cell.2009.04.030`.

[93] V. Ntziachristos. "Going deeper than microscopy: the optical imaging frontier in biology". In: *Nat Methods* 7.8 (Aug. 2010), pp. 603–14. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking). DOI: `10.1038/nmeth.1483`.

[94] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh. "Deep learning vs. traditional computer vision". In: *Science and Information Conference*. Springer. 2019, pp. 128–144.

[95] M. F. Osuchowski, D. G. Remick, J. A. Lederer, C. H. Lang, A. O. Aasen, M. Aibiki, L. C. Azevedo, S. Bahrami, M. Boros, R. Cooney, et al. "Abandon the mouse research ship? Not just yet!" In: *Shock (Augusta, Ga.)* 41.6 (2014), p. 463.

[96] C. Pan, R. Cai, F. P. Quacquarelli, A. Ghasemigharagoz, A. Lourbopoulos, P. Matryba, N. Plesnila, M. Dichgans, F. Hellal, and A. Erturk. "Shrinkage-mediated imaging of entire organs and organisms using uDISCO". In: *Nat Methods* (Aug. 2016). ISSN: 1548-7105 (Electronic) 1548-7091 (Linking). DOI: `10.1038/nmeth.3964`.

[97] M. Pandey and D. Mahadevan. "Monoclonal antibodies as therapeutics in human malignancies". In: *Future Oncol* 10.4 (Mar. 2014), pp. 609–36. ISSN: 1744-8301 (Electronic) 1479-6694 (Linking). DOI: `10.2217/fon.13.197`.

[98] K. Parshotam and M. Kilickaya. "Continual Learning of Object Instances". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 224–225.

[99] A. Paszke. "PyTorch". In: URL https://pytorch.org/ (2016).

[100] S. Perrin. "Preclinical research: Make mouse studies work". In: *Nature News* 507.7493 (2014), p. 423.

[101] B. J. Pichler, H. F. Wehrl, and M. S. Judenhofer. "Latest advances in molecular imaging instrumentation". In: *J Nucl Med* 49 Suppl 2 (June 2008), 5S–23S. ISSN: 0161-5505 (Print) 0161-5505 (Linking). DOI: `10.2967/jnumed.108.045880`.

[102] J. Provost, A. Garofalakis, J. Sourdon, D. Bouda, B. Berthon, T. Viel, M. Perez-Liva, C. Lussey-Lepoutre, J. Favier, M. Correia, et al. "Simultaneous positron emission tomography and ultrafast ultrasound for hybrid molecular, anatomical and functional imaging". In: *Nature biomedical engineering* 2.2 (2018), pp. 85–94.

[103] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. "Transfusion: Understanding transfer learning for medical imaging". In: *Advances in neural information processing systems*. 2019, pp. 3347–3357.

[104] N. Renier, Z. Wu, D. J. Simon, J. Yang, P. Ariel, and M. Tessier-Lavigne. "iDISCO: a simple, rapid method to immunolabel large tissue samples for volume imaging". In: *Cell* 159.4 (2014), pp. 896–910.

[105] D. S. Richardson and J. W. Lichtman. "Clarifying tissue clearing". In: *Cell* 162.2 (2015), pp. 246–257.

[106] S. Rojas, J. D. Gispert, R. Martin, S. Abad, C. Menchon, D. Pareto, V. M. Victor, M. Alvaro, H. Garcia, and J. R. Herance. "Biodistribution of amino-functionalized diamond nanoparticles. in vivo studies based on 18F radionuclide emission". In: *ACS nano* 5.7 (2011), pp. 5552–5559.

[107] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[108] S. Rosenhain, Z. A. Magnuska, G. G. Yamoah, F. Kiessling, F. Gremse, et al. "A preclinical micro-computed tomography database including 3D whole body organ segmentations". In: *Scientific data* 5.1 (2018), pp. 1–9.

[109] N. Rosenthal and S. Brown. "The mouse ascending: perspectives for human-disease models". In: *Nature cell biology* 9.9 (2007), pp. 993–999.

[110] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[111] D. P. Ryan, T. S. Hong, and N. Bardeesy. "Pancreatic adenocarcinoma". eng. In: *N Engl J Med* 371.11 (Sept. 2014), pp. 1039–49. ISSN: 1533-4406 (Electronic) 0028-4793 (Linking). DOI: `10.1056/NEJMra1404198`.

[112] K. Saatchi and U. O. Hafeli. "Radiolabeling of biodegradable polymeric microspheres with [99mTc (CO) 3]+ and in vivo biodistribution evaluation using microSPECT/CT imaging". In: *Bioconjugate chemistry* 20.6 (2009), pp. 1209–1217.

[113] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, et al. "Fiji: an open-source platform for biological-image analysis". In: *Nature methods* 9.7 (2012), pp. 676–682.

[114] M. Schneider, J. Reichold, B. Weber, G. Szekely, and S. Hirsch. "Tissue metabolism driven arterial tree generation". In: *Medical image analysis* 16.7 (2012), pp. 1397–1414.

[115] N. Schonhuber, B. Seidler, K. Schuck, C. Veltkamp, C. Schachtler, M. Zukowska, S. Eser, T. B. Feyerabend, M. C. Paul, P. Eser, S. Klein, A. M. Lowy, R. Banerjee, F. Yang, C. L. Lee, E. J. Moding, D. G. Kirsch, A. Scheideler, D. R. Alessi, I. Varela, A. Bradley, A. Kind, A. E. Schnieke, H. R. Rodewald, R. Rad, R. M. Schmid, G. Schneider, and D. Saur. "A next-generation dual-recombinase system for time- and host-specific targeting of pancreatic cancer". eng. In: *Nat Med* 20.11 (Nov. 2014), pp. 1340–1347. ISSN: 1546-170X (Electronic) 1078-8956 (Linking). DOI: `10.1038/nm.3646`.

[116] O. Schoppe. *AIMOS - light-sheet microscopy dataset*. Version V1. 2020. DOI: `https://doi.org/10.7910/DVN/LL3C1R`. URL: `https://doi.org/10.7910/DVN/LL3C1R`.

[117] O. Schoppe. *AIMOS - pre-trained models*. Version V1. 2020. DOI: `https://doi.org/10.7910/DVN/G6VLZN`. URL: `https://doi.org/10.7910/DVN/G6VLZN`.

[118] L. Sevenich, R. L. Bowman, S. D. Mason, D. F. Quail, F. Rapaport, B. T. Elie, E. Brogi, P. K. Brastianos, W. C. Hahn, L. J. Holsinger, J. Massague, C. S. Leslie, and J. A. Joyce. "Analysis of tumour- and stroma-supplied proteolytic networks reveals a brain-metastasis-promoting role for cathepsin S". eng. In: *Nat Cell Biol* 16.9 (Sept. 2014), pp. 876–88. ISSN: 1476-4679 (Electronic) 1465-7392 (Linking). DOI: `10.1038/ncb3011`.

[119] P. Sharma, M. Diwakar, and S. Choudhary. "Application of edge detection for brain tumor detection". In: *International Journal of Computer Applications* 58.16 (2012).

[120] R. Smith-Bindman, D. L. Miglioretti, and E. B. Larson. “Rising use of diagnostic medical imaging in a large integrated health system”. In: *Health affairs* 27.6 (2008), pp. 1491–1502.

[121] W. Spalteholz. *Ueber das Durchsichtigmachen von menschlichen und tierischen Praeparaten und seine theoretischen Bedingungen*. S. Hirzel, 1914.

[122] M. Stefaniuk, E. J. Gualda, M. Pawlowska, D. Legutko, P. Matryba, P. Koza, W. Konopka, D. Owczarek, M. Wawrzyniak, P. Loza-Alvarez, et al. “Light-sheet microscopy imaging of a whole cleared rat brain with Thy1-GFP transgene”. In: *Scientific reports* 6 (2016), p. 28209.

[123] D. F. Steiner, R. MacDonald, Y. Liu, P. Truszkowski, J. D. Hipp, C. Gammage, F. Thng, L. Peng, and M. C. Stumpe. “Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer”. In: *Am J Surg Pathol* 42.12 (Dec. 2018), pp. 1636–1646. ISSN: 1532-0979 (Electronic) 0147-5185 (Linking). DOI: 10.1097/PAS.0000000000001151.

[124] D. P. Sullivan, C. F. Winsnes, L. Akesson, M. Hjelmare, M. Wiking, R. Schutten, L. Campbell, H. Leifsson, S. Rhodes, A. Nordgren, K. Smith, B. Revaz, B. Finnbogason, A. Szantner, and E. Lundberg. “Deep learning is combined with massive-scale citizen science to improve large-scale image classification”. In: *Nat Biotechnol* 36.9 (Oct. 2018), pp. 820–828. ISSN: 1546-1696 (Electronic) 1087-0156 (Linking). DOI: 10.1038/nbt.4225.

[125] M. Tabrizi, G. G. Bornstein, and H. Suria. “Biodistribution mechanisms of therapeutic monoclonal antibodies in health and disease”. eng. In: *AAPS J* 12.1 (Mar. 2010), pp. 33–43. ISSN: 1550-7416 (Electronic) 1550-7416 (Linking). DOI: 10.1208/s12248-009-9157-5.

[126] K. Tainaka, S. I. Kubota, T. Q. Suyama, E. A. Susaki, D. Perrin, M. Ukai-Tadenuma, H. Ukai, and H. R. Ueda. “Whole-body imaging with single-cell resolution by tissue decolorization”. eng. In: *Cell* 159.4 (Nov. 2014), pp. 911–24. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking). DOI: 10.1016/j.cell.2014.10.034.

[127] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding. “Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation”. In: *Medical Image Analysis* (2020), p. 101693.

[128] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. “Convolutional neural networks for medical image analysis: Full training or fine tuning?” In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1299–1312.

[129] G. Tetteh, V. Efremov, N. D. Forkert, M. Schneider, J. Kirschke, B. Weber, C. Zimmer, M. Piraud, and B. H. Menze. "Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes". In: *arXiv preprint arXiv:1803.09340* (2018).

[130] S. Thaler and V. Menkovski. "The role of deep learning in improving healthcare". In: *Data Science for Healthcare*. Springer, 2019, pp. 75–116.

[131] P. Timpson, E. J. McGhee, and K. I. Anderson. "Imaging molecular dynamics in vivo–from cell biology to animal models". In: *J Cell Sci* 124.Pt 17 (Sept. 2011), pp. 2877–90. ISSN: 1477-9137 (Electronic) 0021-9533 (Linking). DOI: `10.1242/jcs.085191`.

[132] M. I. Todorov, J. C. Paetzold, O. Schoppe, G. Tetteh, S. Shit, V. Efremov, K. Todorov-Völgyi, M. Düring, M. Dichgans, M. Piraud, et al. "Machine learning analysis of whole mouse brain vasculature". In: *Nature Methods* 17.4 (2020), pp. 442–449.

[133] E. J. Topol. "High-performance medicine: the convergence of human and artificial intelligence". In: *Nat Med* 25.1 (Jan. 2019), pp. 44–56. ISSN: 1546-170X (Electronic) 1078-8956 (Linking). DOI: `10.1038/s41591-018-0300-7`.

[134] N. Tran, N. Bye, B. A. Moffat, D. K. Wright, A. Cuddihy, T. M. Hinton, A. M. Hawley, N. P. Reynolds, L. J. Waddington, X. Mulet, et al. "Dual-modality NIRF-MRI cubosomes and hexosomes: High throughput formulation and in vivo biodistribution". In: *Materials Science and Engineering: C* 71 (2017), pp. 584–593.

[135] V. V. Tuchin. "Tissue optics and photonics: light-tissue interaction". In: *Journal of Biomedical Photonics & Engineering* 1.2 (2015).

[136] V. V. Tuchin. "Editors Introduction: Optical Methods for Biomedical Diagnosis". In: (2016).

[137] G. J. Van Den Burg and P. J. Groenen. "GenSVM: A generalized multiclass support vector machine". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 7964–8005.

[138] B. Van Der Heyden, M. Podesta, D. B. Eekers, A. Vaniqui, I. P. Almeida, L. E. Schyns, S. J. Van Hoof, and F. Verhaegen. "Automatic multiatlas based organ at risk segmentation in mice". In: *The British journal of radiology* 92.1095 (2018), p. 20180364.

[139] A. Van Opbroek, M. A. Ikram, M. W. Vernooij, and M. De Bruijne. "Transfer learning improves supervised image segmentation across imaging protocols". In: *IEEE transactions on medical imaging* 34.5 (2015), pp. 1018–1030.

[140] J. H. Vestjens, M. J. Pepels, M. de Boer, G. F. Borm, C. H. van Deurzen, P. J. van Diest, J. A. van Dijck, E. M. Adang, J. W. Nortier, E. J. Rutgers, C. Seynaeve, M. B. Menke-Pluymers, P. Bult, and V. C. Tjan-Heijnen. “Relevant impact of central pathology review on nodal classification in individual breast cancer patients”. In: *Ann Oncol* 23.10 (Oct. 2012), pp. 2561–6. ISSN: 1569-8041 (Electronic) 0923-7534 (Linking). DOI: `10.1093/annonc/mds072`.

[141] B. Vick, M. Rothenberg, N. Sandhofer, M. Carlet, C. Finkenzeller, C. Krupka, M. Grunert, A. Trumpp, S. Corbacioglu, M. Ebinger, M. C. Andre, W. Hiddemann, S. Schneider, M. Subklewe, K. H. Metzeler, K. Spiekermann, and I. Jeremias. “An advanced preclinical mouse model for acute myeloid leukemia using patients cells of various genetic subgroups and in vivo bioluminescence imaging”. eng. In: *PLoS One* 10.3 (2015), e0120925. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: `10.1371/journal.pone.0120925`.

[142] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature methods* (2020), pp. 1–12.

[143] N. Vogt. “Imaging the mouse as a whole”. In: *Nature methods* 16.3 (2019), pp. 213–213.

[144] S. v. d. Walt, S. C. Colbert, and G. Varoquaux. “The NumPy array: a structure for efficient numerical computation”. In: *Computing in Science & Engineering* 13.2 (2011), pp. 22–30.

[145] C. B. Wang N. “Densely Deep Supervised Networks with Threshold Loss for Cancer Detection in Automated Breast Ultrasound”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 - 21st International Conference, 2018, Proceedings* (2018), pp. 641–648.

[146] H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Gunaydin, L. A. Bentolila, C. Kural, and A. Ozcan. “Deep learning enables cross-modality super-resolution in fluorescence microscopy”. In: *Nat Methods* 16.1 (Jan. 2019), pp. 103–110. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking). DOI: `10.1038/s41592-018-0239-0`.

[147] H. Wang, Y. Han, Z. Chen, R. Hu, A. F. Chatziioannou, and B. Zhang. “Prediction of major torso organs in low-contrast micro-CT images of mice using a two-stage deeply supervised fully convolutional network”. In: *Physics in Medicine & Biology* 64.24 (2019), p. 245014.

[148] H. Wang, D. B. Stout, and A. F. Chatziioannou. “Estimation of mouse organ locations through registration of a statistical mouse atlas with micro-CT images”. In: *IEEE transactions on medical imaging* 31.1 (2011), pp. 88–102.

[149] J. Wang, Z. Fang, N. Lang, H. Yuan, M. Y. Su, and P. Baldi. “A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks”. In: *Comput Biol Med* 84 (May 2017), pp. 137–146. ISSN: 1879-0534 (Electronic) 0010-4825 (Linking). DOI: 10.1016/j.compbiomed.2017.03.024.

[150] S. K. Warfield, K. H. Zou, and W. M. Wells. “Validation of image segmentation by estimating rater bias and variance”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2006, pp. 839–847.

[151] M. Waskom. “seaborn: statistical data visualization”. In: URL https://seaborn.pydata.org/ (2012).

[152] M. Weigert, U. Schmidt, T. Boothe, A. Muller, A. Dibrov, A. Jain, B. Wilhelm, D. Schmidt, C. Broaddus, S. Culley, et al. “Content-aware image restoration: pushing the limits of fluorescence microscopy”. In: *Nature methods* 15.12 (2018), pp. 1090–1097.

[153] D. Welch, A. Harken, G. Randers-Pehrson, and D. Brenner. “Construction of mouse phantoms from segmented CT scan data for radiation dosimetry studies”. In: *Physics in Medicine & Biology* 60.9 (2015), p. 3589.

[154] J. Wen, D. Wu, M. Qin, C. Liu, L. Wang, D. Xu, H. V. Vinters, Y. Liu, E. Kranz, X. Guan, et al. “Sustained delivery and molecular targeting of a therapeutic monoclonal antibody to metastases in the central nervous system of mice”. In: *Nature biomedical engineering* 3.9 (2019), pp. 706–716.

[155] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren. “Preparing medical imaging data for machine learning”. In: *Radiology* 295.1 (2020), pp. 4–15.

[156] D. Yan, Z. Zhang, Q. Luo, and X. Yang. “A novel mouse segmentation method based on dynamic contrast enhanced micro-CT images”. In: *PloS one* 12.1 (2017).

[157] B. Yang, J. B. Treweek, R. P. Kulkarni, B. E. Deverman, C. K. Chen, E. Lubeck, S. Shah, L. Cai, and V. Gradinaru. “Single-cell phenotyping within transparent intact tissue through whole-body clearing”. eng. In: *Cell* 158.4 (Aug. 2014), pp. 945–958. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking). DOI: 10.1016/j.cell.2014.07.017.

[158] V. Yeghiazaryan and I. Voiculescu. “An overview of current evaluation methods used in medical image segmentation”. In: *Department of Computer Science, University of Oxford* (2015).

[159] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer. "A fourier perspective on model robustness in computer vision". In: *Advances in Neural Information Processing Systems*. 2019, pp. 13276–13286.

[160] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig. "User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability". In: *Neuroimage* 31.3 (2006), pp. 1116–1128.

[161] F. Zhao, Y. Chen, Y. Hou, and X. He. "Segmentation of blood vessels using rule-based and machine-learning-based methods: a review". In: *Multimedia Systems* 25.2 (2019), pp. 109–118.

[162] W. R. Zipfel, R. M. Williams, R. Christie, A. Y. Nikitin, B. T. Hyman, and W. W. Webb. "Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation". In: *Proc Natl Acad Sci U S A* 100.12 (June 2003), pp. 7075–80. ISSN: 0027-8424 (Print) 0027-8424 (Linking). DOI: `10.1073/pnas.0832308100`.