BMC Neuroscience

RESEARCH

Open Access

Objective hearing threshold identification from auditory brainstem response measurements using supervised and self-supervised approaches

Dominik Thalmeier[1,3†], Gregor Miller[2†], Elida Schneltzer[2], Anja Hurt[2], Martin Hrabě deAngelis[2,6,7*], Lore Becker[2], Christian L. Müller[1,3,4,5*] and Holger Maier[2]

Abstract

Hearing loss is a major health problem and psychological burden in humans. Mouse models offer a possibility to elucidate genes involved in the underlying developmental and pathophysiological mechanisms of hearing impairment. To this end, large-scale mouse phenotyping programs include auditory phenotyping of single-gene knockout mouse lines. Using the auditory brainstem response (ABR) procedure, the German Mouse Clinic and similar facilities worldwide have produced large, uniform data sets of averaged ABR raw data of mutant and wildtype mice. In the course of standard ABR analysis, hearing thresholds are assessed visually by trained staff from series of signal curves of increasing sound pressure level. This is time-consuming and prone to be biased by the reader as well as the graphical display quality and scale.In an attempt to reduce workload and improve quality and reproducibility, we developed and compared two methods for automated hearing threshold identification from averaged ABR raw data: a supervised approach involving two combined neural networks trained on human-generated labels and a self-supervised approach, which exploits the signal power spectrum and combines random forest sound level estimation with a piece-wise curve fitting algorithm for threshold finding.We show that both models work well and are suitable for fast, reliable, and unbiased hearing threshold detection and quality control. In a high-throughput mouse phenotyping environment, both methods perform well as part of an automated end-to-end screening pipeline to detect candidate genes for hearing involvement. Code for both models as well as data used for this work are freely available.

Keywords: Automation, Auditory brainstem response, Evoked potentials, High-throughput hearing screening, Objective hearing threshold detection

Introduction

Impaired hearing has a high impact on quality of life and age-related hearing loss is a common health burden in an aging society [1, 2]. Disease models of hearing loss using mutant mouse lines can be useful for research of the underlying pathophysiological and molecular mechanisms. Using auditory brainstem response (ABR), a large-scale screen of 1211 single-gene knock-out mouse lines has recently identified dozens of candidate genes associated with hearing threshold impairment [3]. Earlier, an

†D. Thalmeier and G. Miller are the co-first authors

*Correspondence: hrabe@helmholtz-muenchen.de; christian.mueller@helmholtz-muenchen.de

[2] Institute of Experimental Genetics, Helmholtz Zentrum München, München, Germany
[5] Center for Computational Mathematics, Flatiron Institute, New York, USA
Full list of author information is available at the end of the article

even larger study [4] revealed 52 novel candidate genes with hearing loss involvement from analysis of ABR data measured on 3006 mutant mouse strains within the International Mouse Phenotyping Consortium (IMPC) [5, 6] effort.

The ABR is a form of electroencephalography (EEG), in which electrical potentials are recorded from the scalp of clinical patients or laboratory animals as evoked responses to auditory stimulation. Response signals following rapidly repeated stimulus sequences are averaged and produce typical ABR waveforms that are characterised by specific peaks and troughs, their amplitudes and latencies. As the ABR results from neurological and involuntary processing of sound signals by the different regions of the auditory brainstem, it is an easy-to-perform diagnostic method for hearing assessment of unconscious patients, infants, or animals. In large-scale mouse phenotyping and—more generally—in basic auditory research, it is established as standardised method for measuring hearing function for many years [7].

When using ABR for hearing threshold identification, this involves a series of measurements at increasing sound pressure levels (SPL) in 5 dB steps at different pure-tone frequencies ("tone pips") at 6, 12, 18, 24, and 30 kHz as well as a broadband frequency stimulus ("click"). For each tone pip and click stimulus, the ABR waveforms are displayed in a stacked diagram ordered by ascending SPL. In this audiogram, the hearing threshold (HT) for each frequency is then determined as the lowest SPL where a trained human reader can still detect a signal during a visual assessment of the stacked curves diagram. This signal has to be consistent with higher SPL signals, i.e. exhibiting the same, however weaker and shifted peaks. A plot of hearing threshold SPLs vs. stimulus frequency ("hearing curve") allows rapid overall characterisation of hearing sensitivity.

It is well-established that threshold determination by human readers is prone to reader bias [8] as well as intra- and inter-reader variability [9, 10, 11]. This might depend on different visualisation tools, reader concentration, experience, training, and personal visual skills. In particular in high-throughput environments, maintaining the same conditions over hours is difficult. Another challenge is to achieve and maintain low inter-reader variability in teams with different readers.

Accordingly, since early on in ABR application, there have been attempts to automate and develop objective methods to determine hearing thresholds from ABR measurements. Over the years, ABR has been discussed in literature as *Auditory Evoked Potentials* (AEP) [12], *Cortical Auditory Evoked Potentials* (CAEP) [13], *Brainstem auditory evoked potential* (BAEP) [14, 15], *Brainstem Evoked Response Audiometry* (BERA) [16], and *Auditory Evoked Potential* (EAP) [17]. Many approaches applied and combined methods from different fields of statistics [9, 13, 17–30], often involving feature extraction from the time and/or the frequency domain. Some approaches also involved bootstrapping [31], comparison to templates [15, 23], or deep learning [12, 14, 32–34].

While most of the published methods for automated threshold identification use averaged response data, a recently published method [30] processes individual sweep responses with good results. Unfortunately, although always generated during ABR, individual sweep response time curves are not always easily accessible. Instead, readers are usually only provided with the averaged curves.

Despite all these published efforts, automated approaches seem to have not yet replaced the visual threshold identification by experienced human readers in research practice. This is unfortunate since determining hearing thresholds in thousands of mice is not only laborious and subjective, as discussed above. In addition, long-term structured phenotyping efforts as performed in the German Mouse Clinic (GMC) [35, 36], or in the IMPC generate a huge corpus of ABR data. When it comes to big data analysis, ensuring objective, accurate, and same-standard threshold reading across the whole data set is hardly feasible with human readers.

In this work, we present our efforts and results towards developing a solution for objective and automated high-throughput identification of hearing thresholds from averaged ABR raw data in large-scale research environments. It is intended to reduce human workload, generate accurate, objective, and reproducible results, re-evaluate legacy data, and establish automated quality control processes.

Using a data set generated at the German Mouse Clinic within the IMPC effort as well as an independent external data set provided by the Wellcome Sanger Institute, we developed both a supervised and a self-supervised automated threshold detection method that work on the averaged data available to the researcher. Using two independent data sets, performance and quality of both methods are compared to the standard method for hearing threshold finding—the subjective estimate by experts, using visual detection. Furthermore, we developed an evaluation method that allows relative comparison of threshold detection methods without requiring any kind of ground truth.

In addition, we developed and evaluated data processing and visualisation methods that allow rapid identification of hearing involvement candidate genes using comparative manual and automated threshold finding.

Materials and methods

Data generation

In this work, averaged ABR raw data from measurements conducted in the German Mouse Clinic on mice from both sexes at 14 weeks of age was used. The ABR measurements were performed as part of a large-scale, primary comprehensive phenotyping effort within the IMPC. Accordingly, the data set comprised mutant mice, representing hundreds of different single-gene knockouts, as well as control wildtype mice. All mice were either on a C57BL/6NTac or C57BL/6NCrl genetic background and measured between 2013 and 2020. Mice were group-housed in standard individually ventilated cages under a 12h light/dark schedule in controlled environmental conditions of $22 \pm 2\,^{\circ}\mathrm{C}$ and $50 \pm 10\%$ relative humidity and fed a normal chow diet and water ad libitum. Measurements were performed mainly in the morning.

Mice were anaesthetised with ketamine/xylazine and transferred onto a heating blanket in a sound-attenuating booth. Subcutaneous needle electrodes were inserted in the skin on the vertex (active) and overlying the ventral region of the left (reference) and right (ground) bullae. Stimuli were presented as free-field sounds from a loudspeaker in front of the interaural axis. The sound delivery system was calibrated using a microphone (PCB Piezotronics). For threshold determination, custom software (kindly provided by the Wellcome Sanger Institute) and Tucker Davis Technologies hardware were used to deliver click (0.01 ms duration) and tone pip (6, 12, 18, 24, and 30 kHz of 5 ms duration, 1 ms rise/fall time) stimuli over a range of sound pressure levels (SPL) in 5 dB steps (Click: 0–85 dB, 6 kHz: 20–85 dB, 12–24 kHz: 0–85 dB, 30 kHz: 20–85 dB). Averaged responses to 256 stimuli, presented at 42.6/s, were analysed. For manual threshold detection, the lowest sound intensity giving a visually detectable ABR response was determined. For further reference, this data set is addressed as the *GMC* data set.

To test the methods with external data, a large, published resource of ABR raw data from the Wellcome Sanger Institute [37] measured on 9000+ mice from 1211 single-gene mutant lines and respective control (wildtype) mice on largely C57BL/6N but also other genetic backgrounds was used. We thank the authors for kindly making this invaluable resource publicly available. This data set is addressed as the *ING* data set.

Data pre-processing

All ABR data used was pre-processed to create a single csv file containing the ABR time series (columns t0–t999), an individual mouse identifier, stimulus frequency, stimulus SPL, and a manually determined hearing threshold. For each mouse, there are different ABR time series corresponding to six different sound stimuli: broadband click, 6, 12, 18, 24, and 30 kHz, each of which was measured for a range of sound pressure levels. The exact range of sound levels can vary between the different mice and stimuli, as described above. Mice not having a complete set of data for all six stimuli were excluded during pre-processing of the GMC data set.

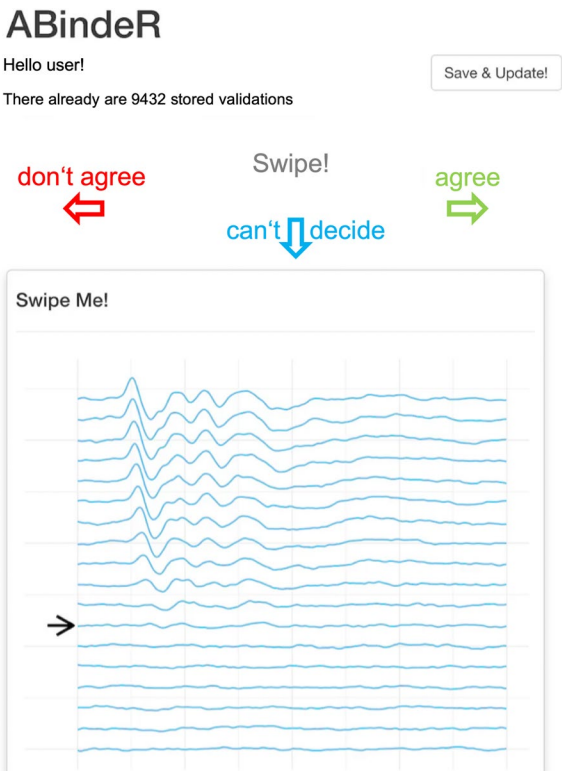


Fig. 1 ABR threshold validation app used on standard tablet computers. Users are presented a randomly selected ABR measurement along with the original human-assigned threshold, visually indicated by an arrow. Users can enter their evaluation by swiping left: "don't agree", right: "agree", down: "can't decide". No mouse or stimulus information is provided in order to allow unbiased evaluation

Data validation

In order to obtain the best-possible label quality in the supervised approach, the hearing thresholds of roughly one-seventh of the GMC data set were re-validated using a simple R/shiny app on standard tablet computers, as shown in Fig. 1. In the app, ABR-trained users had to state their agreement with the original human-assigned threshold for randomly presented hearing curves. Measurements with an "agree" validation result were subsequently weighted higher in the supervised neural

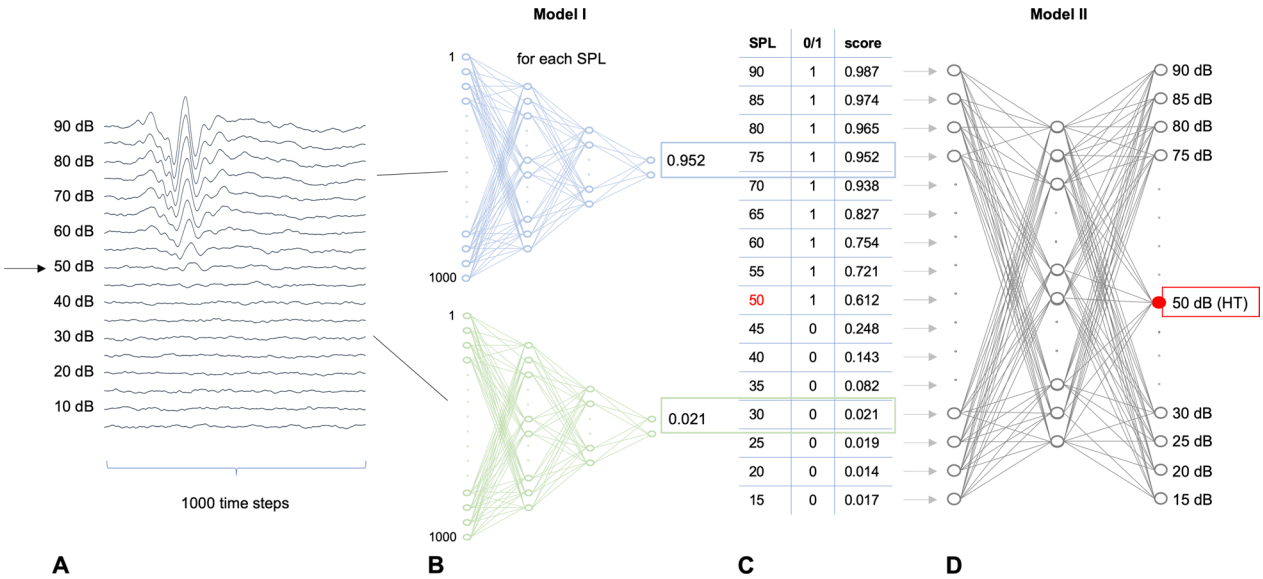


Fig. 2 Scheme illustrating the two-stage method. **A** Example ABR plot for one stimulus. Stacked curves correspond to evoked response time signal (1000 time steps) for increasing sound pressure levels (SPL). The arrow indicates the human-assigned hearing threshold. **B** For each curve of all sound pressure level (SPL), a trained neural network (Model I) classifier predicts if a response is present (1, blue example) or not (0, green example), using a 1000 time step input vector and delivering a class score as output. **C** Result of Model I classifier. For each SPL, the binary decision (0/1) and a class score is generated. **D** A second classifier (Model II) uses the class score outputs from Model I as input vector and predicts the hearing threshold (HT, red). Both models are trained on the actual hearing threshold label

network approach (see "Supervised artificial neural network (NN)" section) than the measurements receiving a "don't agree" or "can't decide" validation result. Using the app, a large number of ABR measurements could be re-evaluated in short time in a blinded fashion, since no information about the mouse, the stimulus, or the original reader is provided whatsoever.

The hearing thresholds of the ING data set were not re-validated, but used as provided.

Supervised artificial neural network (NN)

For modelling the human threshold finding process, we used a two-stage artificial neural network approach, which is illustrated in Fig. 2. Briefly, artificial neural networks consist of interconnected layers of weighted algorithmic nodes that mimic neurons. The weights of each node are adjusted during extensive training using a dataset of known outputs called ground truth labels. In this way, a trained artificial neural network represents an implicit algorithm that can be used for tasks that defy rule-based approaches, such as complex classification or regression tasks.

A first convolutional neural network (Model I) is trained as classifier to predict if an ABR response is present or not present in a single stimulus curve (one frequency, one sound pressure level). The required labels for Model I are derived from the original hearing thresholds under the assumption that all sub-threshold SPL curves represent non-hearing, while threshold and supra-threshold SPL curves represent hearing. A second convolutional neural network (Model II) is then trained as classifier for every stimulus to predict the hearing threshold using the respective class score outputs of Model I as input and the original hearing thresholds as labels. A five-fold grouped cross-validation approach with the mice as groups was followed. First, mice were randomly split 4:1 into training and test mice. This training data was then randomly split 4:1 into training and validation mice in each fold. The architecture of both models is provided in Additional file 1: S1- Neural network model architectures. For reference, this method will be addressed as "NN" in this work.

Self-supervised Sound Level Regression (SLR)

A scheme of the new threshold detection method called "Sound Level Regression" is shown in Fig. 3 . For reference, this method will be addressed as "SLR" in this work. In short, it consists of two steps, which are performed on each stimulus frequency and click separately:

A *Sound level estimation from single curves* In this step, the sound level of the stimulus is estimated from the time series data of its evoked signal curve using a supervised regression method. More precisely, as the

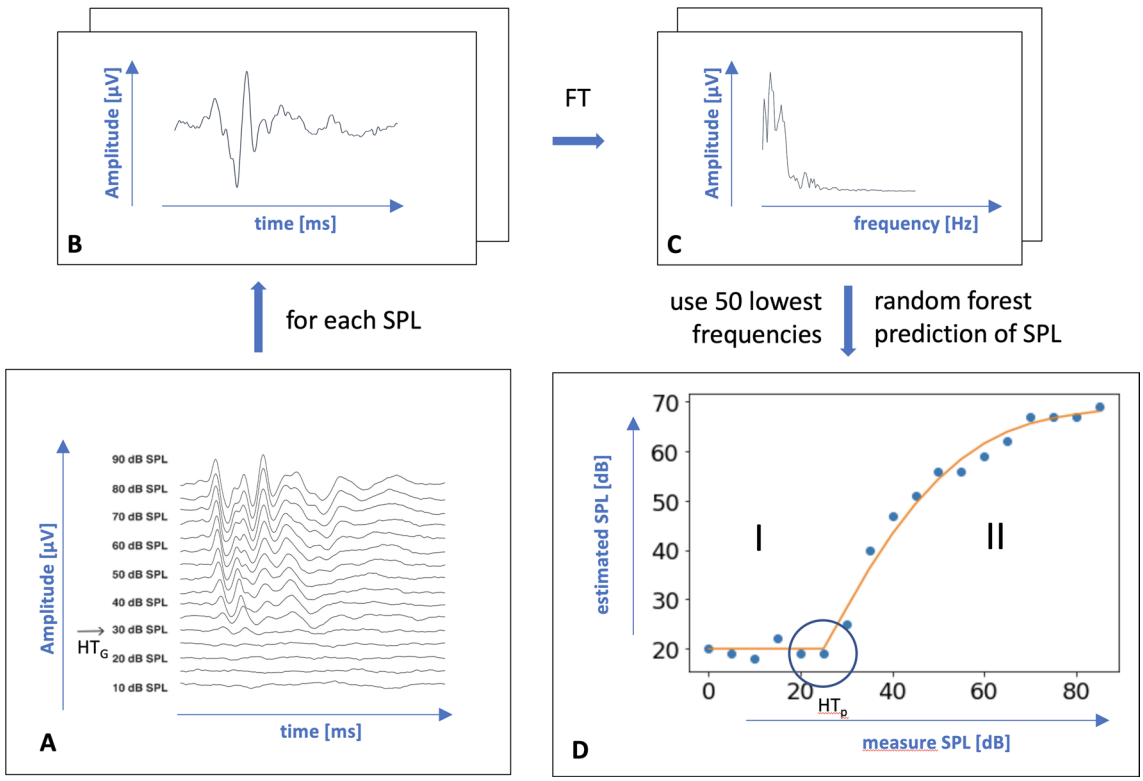


Fig. 3 Scheme illustrating self-supervised prediction of hearing thresholds from evoked ABR signals. **A** Stacked ABR time series for different given sound pressure levels (SPL) at a given stimulation frequency, e.g. 18 kHz. The arrow indicates the manual hearing threshold HT_G, which is only used for validation in this approach. **B** Amplitude signal in the time domain for a single SPL. **C** Amplitude in the frequency domain after Fourier transformation (FT). The lowest 50 frequencies are extracted from the power spectrum as features. **D** estimated vs. measure SPL. Blue dots: sound pressure levels estimated by random forest regression plotted against the actual measure sound levels. Orange line: a piece-wise function composed of a constant (I) and a polynomial (II) part. The polynomial part of the function is fitted using an elastic net. The circle indicates where the constant and polynomial function meet, which determines the predicted hearing threshold HT_P

sound level is given in the data itself, it is called a self-supervised method. The core idea is that such a prediction can only work if the sound level of the stimulus that leads to the evoked signal curve is above the hearing threshold. As otherwise, per definition, no information about the sound level should be contained in the resulting time series.

B *Hearing threshold estimation from sound level estimates* In this step, the hearing threshold for a given frequency can be determined from all respective single hearing curves by comparison of the sound levels predicted in the previous step against the known sound levels according to the following logic: it can be expected that for sub-threshold conditions, the predicted sound levels fluctuate around a constant value, while for supra-threshold conditions, the predicted sound levels follow a monotonically increasing function of the actual sound level (see Fig. 3D). By fitting a piece-wise function that is constant up to a certain value and then monotonically increasing, the break point can be used as an estimate of the hearing threshold.

In the following, the two steps are described in more detail.

Step A: Estimate sound levels for hearing curves using machine learning

In order to estimate the sound levels, the time series are first transformed into a feature space. The Fourier transform (FT) is used to transform each hearing curve from the time domain to the frequency domain. The resulting power spectrum consists of a set of discrete frequency signals that can be considered as features. To reduce the dimension of this feature space, only the lowest 50 frequencies of the power spectrum are used in the following steps. Then, a random forest regression model is trained to estimate the sound level for each hearing curve from the corresponding feature vector.

To avoid overfitting, training and prediction are embedded in a fivefold, mice are divided into a training and a test group. The random forest is trained only on time series from the training group and makes the prediction for the test group. This way, training is strictly separated from the test data and prediction is still possible for each time series.

Step B: Determine hearing thresholds from sound level estimates

Next, the predicted sound levels are used to determine the hearing threshold. As described above, a piece-wise function is fitted, consisting of a constant part and a monotonically increasing part, which is modeled as a polynomial function. In principle, the breakpoint of the fitted function could be used directly to determine the hearing threshold. However, we have lfound that this is not very robust, since it is possible that the polynomial starts as a very flat function that is still quite similar to a constant function. Therefore, the hearing threshold is determined as the sound level at which the polynomial part of the fitted function deviates from the constant part by more than 4 dB. In the remainder of this section, details about the fitting process are described.

Determine upper and lower bounds for threshold First, the search space for the hearing threshold is narrowed by calculating a rough estimate of its upper and lower bounds.

The upper bound of the threshold is determined by the largest sound level for which all estimated values above that limit show a significant positive correlation to the actual sound level used. This is calculated by testing the hypothesis for each sound level in question to see if the sound levels greater than that level have a positive correlation with the corresponding predicted sound levels. The largest value for which the p-value is greater than 5 percent after a Bonferroni correction is used.

As the lower limit for the threshold is determined by the first increase of a function learned by isotonic regression, which empirically was found to be a conservative lower limit for the hearing threshold.

Fitting a piece-wise function What remains is a range between these upper and lower thresholds as candidates for the threshold. Since measurements are taken in steps of 5 dB, possible candidates for the threshold are also limited to a grid of 5 dB.

For each possible threshold value, a piece-wise function with the breakpoint at the possible threshold position is fitted. The function consists of a constant function on the left side of the breakpoint and a polynomial of the fourth degree for sound levels larger than the breakpoint. An elastic net with l1 ratios of 0.5 and 0.99 and 5-fold cross-validation with automatic determination of the regularisation parameter is used for fitting. Of the various functions used for fitting, one for each possible breakpoint, the one that has the least cross-validation error is selected.

With this procedure it can happen that the true threshold value is e.g. 25.1 dB and therefore a threshold value of 25 is estimated. However, the threshold should be the lowest recorded sound level at which the mouse exhibits stimulus-induced ABR activity, which in this case would be 30 dB.

Therefore, also the piece-wise function for sound levels that are 0.5 dB lower and higher than the selected breakpoint is fitted. If either of these show a cross-validation error that is lower than the current optimum breakpoint, the new value is considered the new optimum and therefore the final predicted hearing threshold.

Evaluation curves

To avoid the use of human-derived ground truth labels in the quality assessment of two hearing threshold finding methods, evaluation curves were developed as a visual quality assessment method. This section describes the theoretical concept behind it.

Assuming that the true hearing thresholds are known, the sample average of all super-threshold curves and the sample average of all sub-threshold curves can be calculated. When taking the sample average of all super-threshold curves, a temporal pattern should emerge, since the mice react to the signal tone in a temporally coherent manner. In contrast, averaging the sub-threshold ABR curves should result in a constant signal as the ABR curves are/have to be temporally incoherent due to the absence of a perceived signal and therefore any temporal pattern is averaged out when taking the mean.

From this argumentation, measures to assess the quality of any threshold finding method can be derived. To this end, all ABR curves from all mice that correspond to a specific stimulus (e.g. click) are given an index $i \in [0, N]$, with N being the total number of measured ABR curves for all mice, but restricted to this stimulus.

Now

$$\mathrm{l}(i) := \frac{\mathrm{soundlevel}(i)}{\mathrm{threshold}(i)}$$

is defined as the threshold normalized sound level. The ABR curves are sorted by $\mathrm{l}(i)$, so that $\mathrm{l}(i) < \mathrm{l}(i+1)$. Let $x_i(t)$ with $t \in [0, T]$ be the time series of the ABR curve with index i.

The cumulative average for the first n curves with the lowest threshold normalized sound levels can be computed as

Table 1 Basic dataset properties.

	GMC dataset			ING data set		
	Mutants	**Controls**	**Total**	**Mutants**	**Controls**	**Total**
(A) Number of mice						
Males	1331	849	2180	–	–	–
Females	1323	858	2181	–	–	–
Total	2654	1707	**4361**	6130	1900	**8030**
(B) Gene cohort size median [5%;95%]						
Gene cohort size	8 [3;11]			4 [4;10]		
(C) Number of genes						
Distinct genes	352			1152		
Common genes	12					

(A) Number of mice: shown are numbers of distinct, individual mice. In the ING data set, no distinction between male and female numbers was possible, so only total numbers are given. (B) Gene cohort size median: the median size of cohorts of animals with the same affected distinct knockout gene are given along with the 5% and 95% quantiles. (C) Number of genes: the number of distinct knockout genes per data set is given. Common genes provides the number of genes occurring in both datasets: *Bach2, Cdkal1, Dbn1, Dnase1l2, Entpd1, Gsk3a, Hdac1, Klk5, Nxn, Rnf10, Slc20a2, Ubash3a*

$$\bar{X}_n(t) := \frac{1}{n}\sum_{i=0}^{n} x_i(t),$$

where $x_i(t)$ is the time series of the ABR curve with index i defined in the measurement interval $t \in [0, T]$. Now let i_{crit} be the largest index for curves which are still below the threshold value, i.e. for which $\mathrm{l}(i > i_{\text{crit}}) \geq 1$ and $\mathrm{l}(i < i_{\text{crit}}) < 1$. Then for $n \leq i_{\text{crit}}$, $\bar{X}_n(t)$ should be an approximately[1] constant signal with a vanishing temporal variance

$$S^2(n) := \frac{1}{T}\int_0^T \bar{X}_n(t)^2 dt - \left(\frac{1}{T}\int_0^T \bar{X}_n(t) dt\right)^2 \approx 0.$$

If ground truth threshold is used for this sorting, the averaged curve should not deviate significantly from a constant signal until all sub-threshold curves have been added to the cumulative mean $\bar{X}_n(t)$. However, if sub-optimal threshold values are used, the averaged signal should start to deviate from a constant signal earlier, because true sub-threshold curves are mixed with super-threshold curves.

As an example, there might be a total of i_{crit} =3000 real sub-threshold curves in the set of all curves. Then $S^2(n) \approx 0$ for $n \leq 3000$, given the true thresholds are used for sorting. However, if erroneous thresholds are used for sorting, then $S^2(3000)$ can only be zero if the error of the thresholds is a systematic and constant shift of the thresholds. However, if the error is due to an inconsistency in the threshold labeling, then $S^2(3000) > 0$, since lower and upper threshold curves are mixed.

[1] The 'approximately' is due to the finite sample size.

Based on this, evaluation curves can be constructed that compare the quality of threshold value procedures: the (normalized) time variance of the averaged signal

$$\frac{S^2(n)}{S^2(N)}$$

is plotted versus $\frac{n}{N}$, the total percentage of ABR curves included in the cumulative average.

For the ground truth threshold, this curve should be approximately zero until $\frac{n}{N}$ is equal to the number of sub-threshold curves divided by the total number of ABR curves (= sub + super threshold). After that it should increase. For suboptimal thresholds, the curve should start to deviate from zero already at smaller levels of $\frac{n}{N}$. The more error-prone the threshold values are, the faster the corresponding evaluation curve deviates from zero. However, rating curves are not meant to be interpreted quantitatively, but rather are a qualitative tool to compare which method is better. As a visual tool, it is limited by resolution when two curves are close to each other. In this case, the difference could be negligible anyway.

Results and discussion

Pre-processing and characterisation of working data sets

Following pre-processing and validation of raw data, two independent working data sets were produced as described in "Data generation". In short, the GMC data set is based on *in-house* data collected at the German Mouse Clinic, whereas the ING data set is based on a large published ABR data ressource. Table 1 summarises basic properties of the two data sets.

They comprise data of a combined total of 12,391 mice, of which 8784 (2654 + 6130) are mutants and 3607 (1707 + 1900) are controls. In the GMC data set,

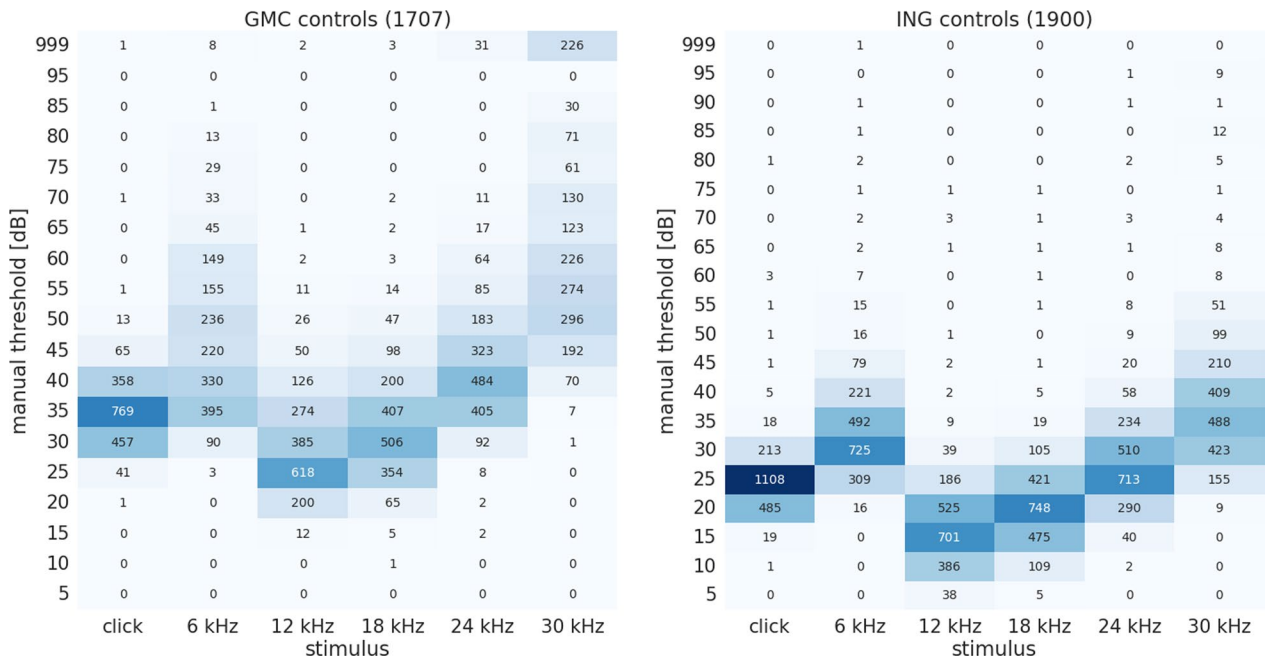


Fig. 4 Available threshold labels for wildtype (control) animals in the GMC (left) and the ING (right) data set. Numbers in parentheses denote the size of the respective data set. Columns indicate the stimulus (click, 6, 12, 18, 24, 30 kHz), rows indicate the hearing threshold assigned by human readers. Non-hearing was arbitrarily assigned 999. Numbers in cells correspond to the distinct number of animals exhibiting the respective hearing threshold at the given stimulus. To facilitate visual comparison of the data sets, numbers in cells are colour-coded in shades of blue

male and female mice are represented equally, both in the mutant and the control groups. For the ING data set, no information about sex is given. The number of knockout genes represented in the GMC and ING data set is 352 and 1152, respectively. Twelve genes (*Bach2, Cdkal1, Dbn1, Dnase1l2, Entpd1, Gsk3a, Hdac1, Klk5, Nxn, Rnf10, Slc20a2, Ubash3a*) are common to both sets, resulting in a combined total of 1492 (352 + 1152 − 12) knockout genes. The median size of mutant cohorts in the GMC data set is 8, compared to 4 in the ING data set.

To investigate the distribution of human-assigned hearing thresholds in the data sets, the according numbers of control (wildtype) mice have been compiled from raw data and visualised in Fig. 4 . While the pattern of the hearing threshold labels reflects the typical U-shaped appearance of a hearing curve, it is obvious that there is a 10–15 dB shift towards lower thresholds in the ING data set compared to the GMC data set. Also, threshold variance is smaller for the ING data set. Notably, there is a considerable number of "non-hearing" (999) labels in the GMC data for 24 kHz and 30 kHz, whereas this is not the case for the ING data. Naturally, the distribution of hearing thresholds is not uniform, i.e. most mice exhibit a hearing threshold only in a small frequent-specific range. Evidently, for any supervised approach, this means that for non-normal thresholds, there are almost no training cases.

Overall, considerable numbers of same-standard, quality-controlled ABR raw data, including metadata and human-assigned threshold labels, have been compiled into two working data sets for further use.

Table 2 Experiment overview

Tested on	NN trained on		SLR calibrated on	
	GMC	ING	GMC	ING
GMC	*Experiment 1* *"NN GMC-GMC"*	Experiment 3 "NN ING-GMC"	*Experiment 5* *"SLR GMC-GMC"*	Experiment 7 "SLR ING-GMC"
ING	Experiment 2 "NN GMC-ING"	*Experiment 4* *"NN ING-ING"*	Experiment 6 "SLR GMC-ING"	*Experiment 8* *"SLR ING-ING"*

Two different ABR threshold finding methods were tested on two different data sets (GMC and ING). The first two columns contain experiments with the two-stage neural network (NN), the last two columns contain experiments with the sound level regression method (SLR). Sub-columns specify the data set that was used for training (NN) or calibration (SLR), respectively. The two rows indicate the data set that was used for testing of the trained NN or the calibrated SLR model. Cells provide the experiment number and the name of the experiment as used in the text. Experiments that use data from the same data set for training/calibration and testing are highlighted in grey

Table 3 NN accuracies for two data sets, stimuli and match levels

Stimulus	NN GMC-GMC			NN ING-ING		
	Accuracy [%]			Accuracy [%]		
	Exact	±5 dB	±10 dB	Exact	±5 dB	±10 dB
Click	19.5	90.3	98.5	12.6	83.9	99.3
6 kHz	28.8	70.3	87.9	22.0	77.1	94.9
12 kHz	32.8	80.1	93.9	22.2	79.5	95.9
18 kHz	28.3	78.4	93.4	15.1	75.3	96.0
24 kHz	25.0	73.9	88.0	14.1	76.9	96.6
30 kHz	21.6	60.5	77.4	16.5	73.6	95.4
Overall	26.0	75.6	89.8	17.1	77.7	96.3

Major columns "NN GMC-GMC" and "NN ING-ING" correspond to two experiments introduced in Table 2. The "exact" columns contain accuracy values when requiring exact match of human-assigned threshold label and NN prediction. The "±5 dB" and "±10 dB" columns contain accuracy values when allowing 5 dB and 10 dB tolerance, respectively. Numbers in cells denote the accuracy of the model prediction at the stimulus and match level

Evaluation and comparison of two new threshold finding methods

In order to comprehensively examine and compare the performance of the two threshold finding methods introduced in this work, a scheme of eight experiments was conceived as shown in Table 2. First, both methods were evaluated in a way that the neural network based method and the Sound Level Regression were tested on subsets of mice from the same data set used for training and calibration, respectively. In a next step, the robustness of both methods was evaluated, to find out to what extent a method trained/calibrated on the GMC data set can be applied on the ING data set and vice versa.

For all experiments, data set specific labels assigned by human readers were used to calculate accuracy as a quality measure. To take into account that hearing thresholds (a) were assigned with a granularity of only 5 dB and (b) human threshold finding is prone to variability, accuracies were calculated using three match levels—"exact": requiring an exact match of label and predicted/estimated threshold, "±5 dB" and "±10 dB": allowing 5 dB and 10 dB mismatch between label and predicted/estimated threshold to still be considered accurate.

The neural network model (NN) can objectively predict hearing thresholds from averaged ABR raw data

With each of both data sets, the NN models were trained and tested with subsets of mutant and control mice from the same data set. This corresponds to experiment 1: "NN GMC-GMC" and experiment 4: "NN ING-ING" as introduced in Table 2.

Five-fold cross-validation showed that the method is robust and predictions can be generalized to the whole data set (not shown). Accuracies calculated for three match levels (see Table 3) show that requiring exact match is not fit for practical use. However, allowing 5 dB and 10 dB mismatch achieves reasonable overall accuracies. This is not surprising, as labels are assigned by human readers and human threshold variance is well-established in literature [9–11] and confirmed by own evaluation experiments with GMC data (see "Data validation" section , data not shown). In general, accuracies are highest for the click stimulus. For both mismatch levels, ING accuracies are higher. This may be due to the observed lower label variability in the ING data set, which hints on more consistent label reading.

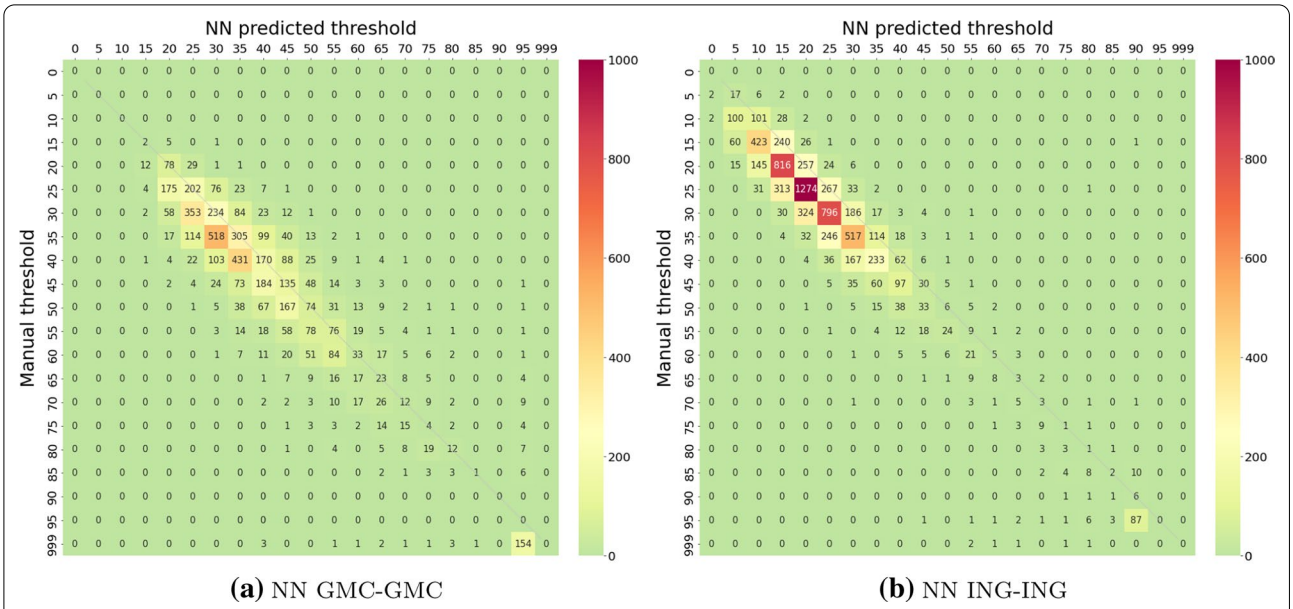
An overall comparison of manual vs. NN predicted thresholds is given in Fig. 5 for both experiments. Interestingly, both experiments reveal a 5 dB shift towards lower predicted thresholds. However, since manual thresholds are used as labels, but do not necessarily provide a ground truth for the hearing threshold, the question remains whether this difference is due to an inaccuracy in manual thresholds or algorithmic prediction.

Sound Level Regression (SLR) can objectively predict hearing thresholds from averaged ABR raw data

With both data sets, the SLR models were calibrated and tested with subsets of mutant and control mice from the same data set. This corresponds to experiment 5: "SLR GMC-GMC" and experiment 8: "SLR ING-ING" as introduced in Table 2.

Quite similar as with the NN approach, SLR accuracies calculated for three match levels (see Table 4) show that exact match accuracies are far below practical applicability. Again, allowing 5 dB and 10 dB mismatch achieves reasonable accuracies, however lower than with the NN approach. SLR accuracies were consistently highest for the click stimulus.

An overall comparison of manual vs. SLR estimated thresholds is given in Fig. 6 for both experiments. In



(a) NN GMC-GMC **(b)** NN ING-ING

Fig. 5 Confusion matrix of manual vs. NN predicted hearing thresholds. In both experiments, named after the scheme introduced in Table 2 , the two-stage neural network was trained and tested on subsets of the same data set (left: GMC, right: ING). The threshold manually assigned by human readers is given on the y-axis, the NN predicted thresholds are given on the x-axis. Numbers in cells are numbers of cases across all stimuli. To facilitate interpretation, numbers are coded by colour intensity according to the colour bar. The thin bisecting line indicates an ideal prediction

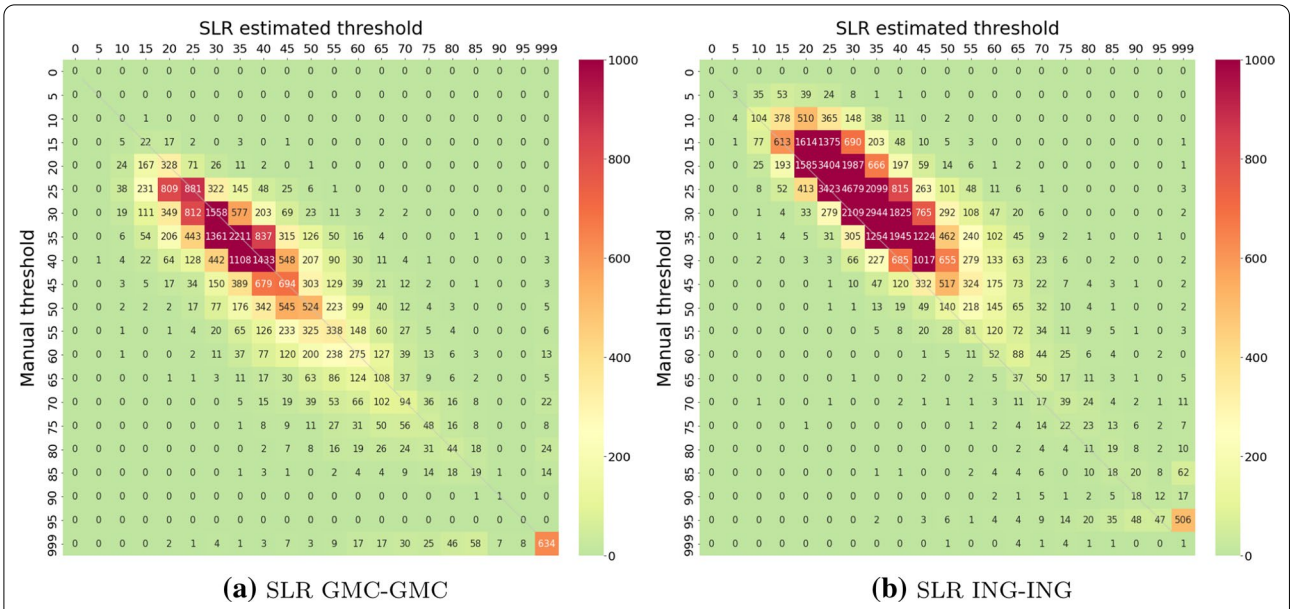
Table 4 SLR accuracies for two data sets, stimuli and match levels

Stimulus	SLR GMC-GMC			SLR ING-ING		
	Accuracy [%]			Accuracy [%]		
	Exact	±5 dB	±10 dB	Exact	±5 dB	±10 dB
Click	59.5	95.4	98.7	44.0	91.2	98.3
6 kHz	24.4	58.7	79.8	14.3	48.9	74.5
12 kHz	27.7	66.4	85.9	17.5	54.1	79.6
18 kHz	30.4	69.9	89.7	18.7	58.2	83.6
24 kHz	33.5	73.3	89.3	18.7	58.5	84.2
30 kHz	35.7	69.0	84.7	19.2	58.4	83.2
Overall	35.2	72.1	88.0	22.0	61.5	83.9

Major columns "SLR GMC-GMC" and "SLR ING-ING" correspond to two experiments introduced in table 2. The "exact" columns contain accuracy values when requiring exact match of human-assigned threshold label and SLR estimation. The "±5 dB" and "±10 dB" columns contain accuracy values when allowing 5 dB and 10 dB tolerance, respectively. Numbers in cells denote the accuracy of the model estimation at the stimulus and match level

contrast to the NN approach, the ING experiment reveals a 5–10 dB shift towards higher estimated thresholds, while the estimation fits quite well in the GMC experiment. Also in contrast to the NN approach, for both mismatch levels, accuracies are higher for the GMC data set. As the SLR method is independent from human labels, this may hint towards systematic differences in human curve reader training or criteria between the data sets, which is also supported by the visible shift in the manual thresholds (Fig. 4).

As an overall evaluation of both methods, NN as well as SLR both work well and deliver good results compared to human labels, provided 5 dB or even 10 dB mismatch are allowed. Depending on the level of reader training and quality control, this may be acceptable to many laboratories. More even so, since both methods have the advantage of delivering reproducible results and are applicable to large ABR data collections while avoiding reader bias.



(a) SLR GMC-GMC **(b)** SLR ING-ING

Fig. 6 Confusion matrix of manual vs. SLR estimated hearing thresholds. In both experiments, named after the scheme introduced in Table , the SLR model was tested on the complete data set after being calibrated on a subset of it (left: GMC, right: ING). The threshold manually assigned by human readers is given on the y-axis, the SLR estimated thresholds are given on the x-axis. Numbers in cells are numbers of cases across all stimuli. To facilitate interpretation, numbers are coded by colour intensity according to the colour bar. The thin bisecting line indicates an ideal estimation. Note: compared to Fig. , numbers are much higher, since SLR methodically allows testing with the complete data set

NN shows higher accuracy than SLR, however SLR is more robust

To investigate robustness of both methods, cross-over experiments according to the scheme laid out in Table 2 were performed. In this experiment series, both methods were systematically trained/calibrated on one data set and applied to the other one. For NN, this corresponds to experiment 2: "NN GMC-ING" and experiment 3: "NN ING-GMC". For SLR, the respective experiments are 6: "SLR GMC-ING" and 7: "SLR ING-GMC". Resulting accuracies are shown in a large overview Table 5 for all three match levels.

For all experiments, "exact match" accuracies are only shown for the sake of completeness and are not further discussed, since they are consistently far below any usability. However, for both ±5 dB and ±10 dB match level, a similar pattern emerges: for both data sets, NN almost always shows highest accuracies when trained on the later test data set (experiments 1 and 4). This is not the case for cross-over experiments 2 and 3, where accuracies collapse. In contrast, SLR exhibits almost consistent accuracies throughout all experiments, thus seems robust and invariant against transfer between calibration and test data sets. This is not surprising, since SLR methodically is not dependent on human labels at all and marginal differences between data sets may only be explained by differences in experimental settings or primary data capture which influence raw data properties.

Overall, under the condition of having a large amount of high quality human labels delivering a consistent hearing threshold ground truth for training, NN shows to be a good replacement or supplementary method for human threshold reading. SLR could be the method of choice if no or not sufficient consistent human labels are available, since it works purely data-driven and ready calibrated SLR methods can be transferred between data sets.

Evaluation curves allow ground-truth free comparison of threshold finding methods

Standard accuracy measurement requires gold standard labels delivering a ground truth. While large, specialised groups may be able to maintain a high level of human reader training and quality control consistently over many years, ABR threshold data generated in smaller groups may show more reader bias and higher variability. Therefore, it seems sensible to measure the quality of any hearing threshold determining method without requiring a gold standard.

Evaluation curves developed in this work are such a method, which has been used to compare human, NN and SLR threshold finding. A forth method always returns a constant threshold, arbitrarily set to 50 dB[2], and is used as a control, since it can assumed to be the worst method. In

[2] Note that the curve looks the same for any constant threshold.

Table 5 Accuracy overview

		Neural Network (NN)		Sound level regression (SLR)	
Test		**Trained on**		**Calibrated on**	
Data	**Stimulus**	**GMC**	**ING**	**GMC**	**ING**
***(A)* Exact match**		**Experiment 1**	**Experiment 3**	**Experiment 5**	**Experiment 7**
GMC data	Overall	26.0 %	9.7 %	35.2 %	**36.0 %**
	Click	19.5 %	16.1 %	**59.5 %**	58.5 %
	6 kHz	**28.8 %**	10.7 %	24.4 %	24.1 %
	12 kHz	**32.8 %**	5.5 %	27.7 %	29.7 %
	18 kHz	28.3 %	7.0 %	30.4 %	**32.7 %**
	24 kHz	25.0 %	9.8 %	**33.5 %**	**33.5 %**
	30 kHz	21.6 %	9.1 %	35.7 %	**37.1 %**
		Experiment 2	**Experiment 4**	**Experiment 6**	**Experiment 8**
ING data	Overall	17.6 %	17.1 %	25.0 %	**22.0 %**
	Click	**65.2 %**	12.6 %	37.3 %	44.0 %
	6 kHz	1.4 %	22.0 %	**16.4 %**	14.3 %
	12 kHz	1.4 %	22.2 %	**25.8 %**	17.5 %
	18 kHz	6.3 %	15.1 %	**24.0 %**	18.7 %
	24 kHz	12.1 %	14.1 %	**23.8 %**	18.7 %
	30 kHz	20.2 %	16.5 %	**22.9 %**	19.2 %
(B)*±*5dB match		**Experiment 1**	**Experiment 3**	**Experiment 5**	**Experiment 7**
GMC data	Overall	**75.6 %**	35.7 %	72.1 %	73 %
	Click	90.3 %	58.3 %	**95.4 %**	95.2 %
	6 kHz	**70.3 %**	36.0 %	58.7 %	57.6 %
	12 kHz	**80.1 %**	28.1 %	66.4 %	69.9 %
	18 kHz	**78.4 %**	28.1 %	69.9 %	72.7 %
	24 kHz	**73.9 %**	32.5 %	73.3 %	72.1 %
	30 kHz	60.5 %	31.0 %	69 %	**70.7 %**
		Experiment 2	**Experiment 4**	**Experiment 6**	**Experiment 8**
ING data	Overall	40.1 %	**77.7 %**	66.3 %	61.5 %
	Click	**98.3 %**	83.9 %	89.9 %	91.2 %
	6 kHz	3.0 %	**77.1 %**	51.7 %	48.9 %
	12 kHz	3.6 %	**79.5 %**	63.4 %	54.1 %
	18 kHz	34.2 %	**75.3 %**	62.8 %	58.2 %
	24 kHz	47.0 %	**76.9 %**	65.2 %	58.5 %
	30 kHz	55.4 %	**73.6 %**	65.1 %	58.4 %
(C)* ±*10 dB match		**Experiment 1**	**Experiment 3**	**Experiment 5**	**Experiment 7**
GMC data	Overall	**89.8 %**	62.3 %	88.0 %	88.2 %
	Click	98.5 %	85.4 %	98.7 %	**98.9 %**
	6 kHz	**87.9 %**	59.0 %	79.8 %	78.5 %
	12 kHz	**93.9 %**	62.8 %	85.9 %	87.5 %
	18 kHz	**93.4 %**	57.4 %	89.7 %	90.1 %
	24 kHz	88.0 %	55.4 %	**89.3 %**	88.8 %
	30 kHz	77.4 %	53.4 %	84.7 %	**85.5 %**

Table 5 (continued)

		Neural Network (NN)		Sound level regression (SLR)	
Test		Trained on		Calibrated on	
Data	Stimulus	GMC	ING	GMC	ING
		Experiment 2	Experiment 4	Experiment 6	Experiment 8
ING data	Overall	58.9 %	**96.3 %**	85.9 %	83.9 %
	Click	**99.7 %**	99.3 %	98.1 %	98.3 %
	6 kHz	8.6 %	**94.9 %**	75.0 %	74.5 %
	12 kHz	14.6 %	**95.9 %**	83.8 %	79.6 %
	18 kHz	71.0 %	**96.0 %**	85.3 %	83.6 %
	24 kHz	79.5 %	**96.6 %**	87.3 %	84.2 %
	30 kHz	80.5 %	**95.4 %**	86.0 %	83.2 %

For eight experiments, as introduced in Table , the table shows overall and stimulus-specific prediction accuracies. In short, columns determine the applied method (NN or SLR) and the training/calibration data set. The header columns denote the data set that was used for testing and the stimulus, respectively. Cells contain the accuracy values. To facilitate interpretation, the best accuracy in each row is marked in bold. Three blocks correspond to the required match level for accuracy calculation: (A) exact match, (B) ± 5 dB, and (C) ± 10 dB tolerance

short, a method is better than another method, the longer its curve stays closer to zero.

Figure 7 shows evaluation curves for data from experiments 1 and 5 (GMC) and from experiments 4 and 8 (ING). Evaluation curves of cross-over experiments 2, 6, 3 and 7 are shown in Additional file 1: Table S1. All methods begin to deviate from zero quite early, so none of them seems to be perfect. However, curves show that for GMC data, both NN and SLR seem to work better than manual threshold finding, with NN overall being slightly better than SLR. In contrast, for ING data, the three methods (human, NN, SLR) differ only marginally, with SLR overall being best. In the context of evaluation curves, "better" means that a method delivers less sub-optimal thresholds than another method.

Using evaluation curves as an unbiased tool, it can be concluded that human threshold finding cannot automatically assumed to be the best method. Data sets may exhibit different levels of variability and human bias. In this regard, the ING data set is more consistent than the GMC data set, which only underpins the need for unbiased threshold finding methods.

Results from evaluation curves in part contradict the assumptions behind the accuracy based evaluation which treat the human labeled thresholds as ground truth. Obviously, this is not always the case and seems to depend on the level of variability and human bias represented in a data set. When abandoning the premise that human threshold reading always delivers the ground truth, both methods introduced in this work perform very well.

Both NN and SLR methods perform well in an end-to-end phenotyping pipeline

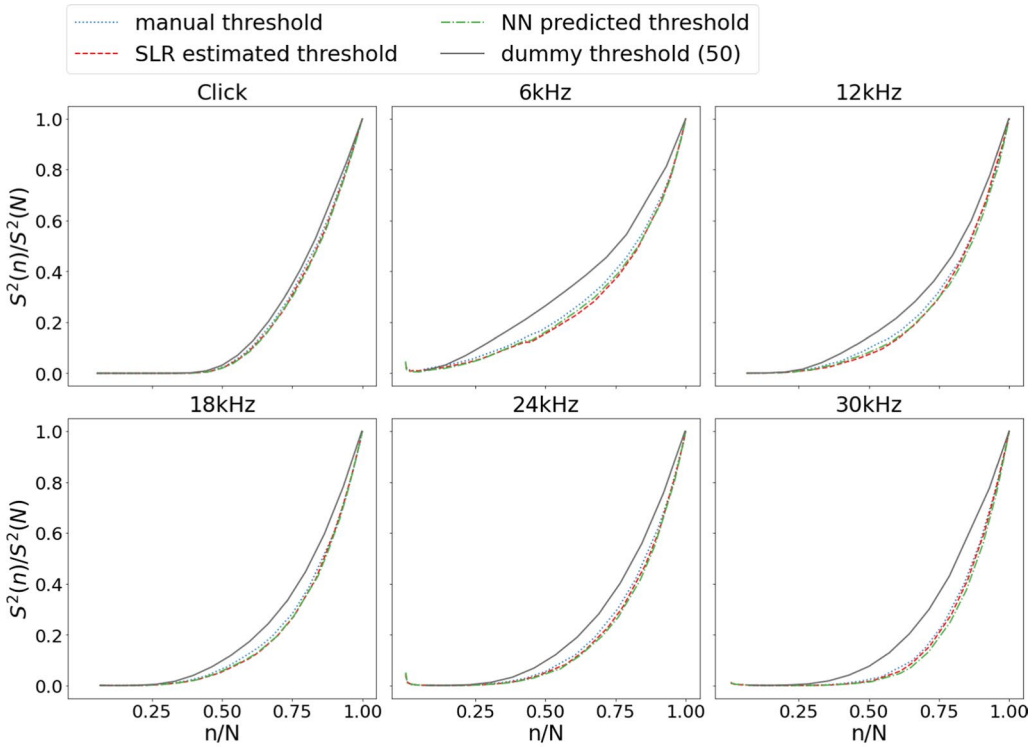
While it is interesting to know that NN and SLR work well for unrelated single stimuli on single mice, using them for routine hearing assessment in high throughput mouse phenotyping is a different matter. In such a scenario, found thresholds are usually aggregated on two levels: first, thresholds for all stimuli of one individual are aggregated to a hearing curve, then, hearing curves are aggregated to display mutant vs. control threshold medians or means.

To find out whether NN and SLR are able to identify mouse lines with biologically relevant changes in such a scenario, the following approach was applied: complete raw data from both data sets was subjected to NN and SLR threshold finding. However, for downstream gene-based analysis steps, data from some mice had to be excluded. In the GMC data set, 45 mutants without clearly assigned reference controls and in the ING data set, 48 mice without valid gene label were affected.

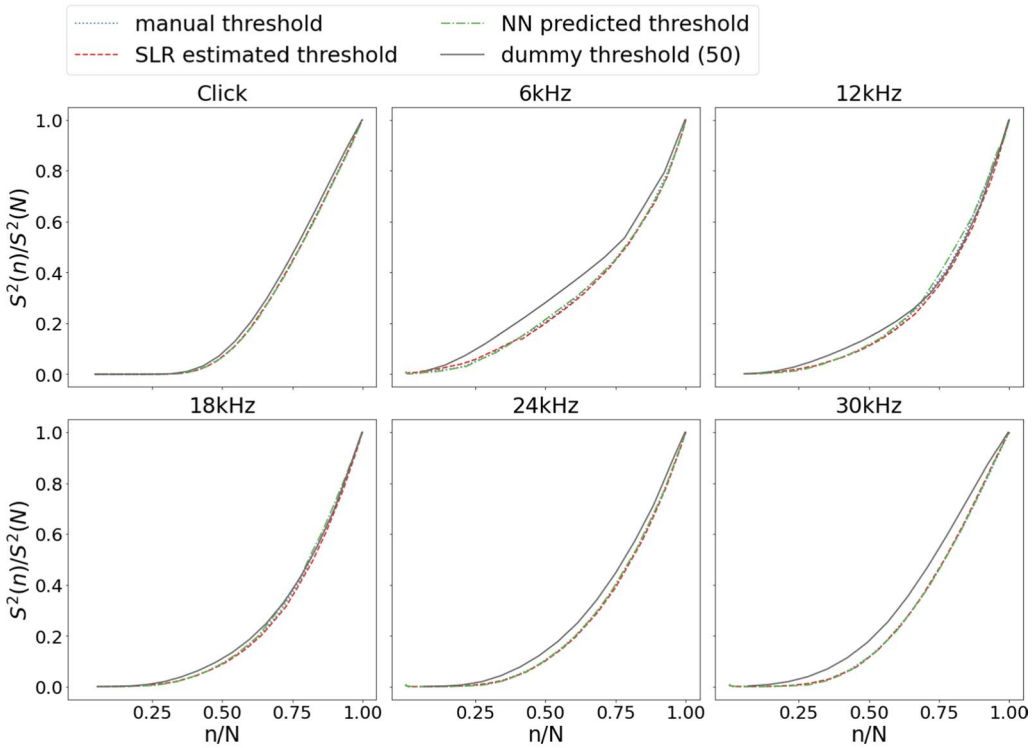
Visual identification of candidate genes Using resulting thresholds, a series of high-level visualisation have been generated that can be used for visual identification of candidate genes. Figure 8 shows an example of an audiogram, which has been generated for every single

(See figure on next page.)

Fig. 7 Objective comparison of threshold finding methods using evaluation curves. Four methods are compared: manual thresholds (blue, dotted lines), SLR estimations (red, dashed lines), NN predictions (green, dash-dotted lines), and an "always 50 dB" control method (grey, solid lines). Separate plots show evaluation curves for each stimulus (click, 6, 12 18, 24, 30 kHz). Plots show the normalized time variance of the averaged signal $S^2(n)/S^2(N)$ (y-axis) vs. the total percentage of ABR curves included in the cumulative average n/N (x-axis). a shows NN predictions and SLR estimations from experiments 1 and 5, b shows NN predictions and SLR estimations from experiments 4 and 8, as introduced in Table . Two methods can be compared in a way that the curve of the better method stays longer close to zero



(a) NN/SLR GMC-GMC (experiments 1 and 5)



(b) NN/SLR ING-ING (experiments 4 and 8)

Fig. 7 (See legend on previous page.)

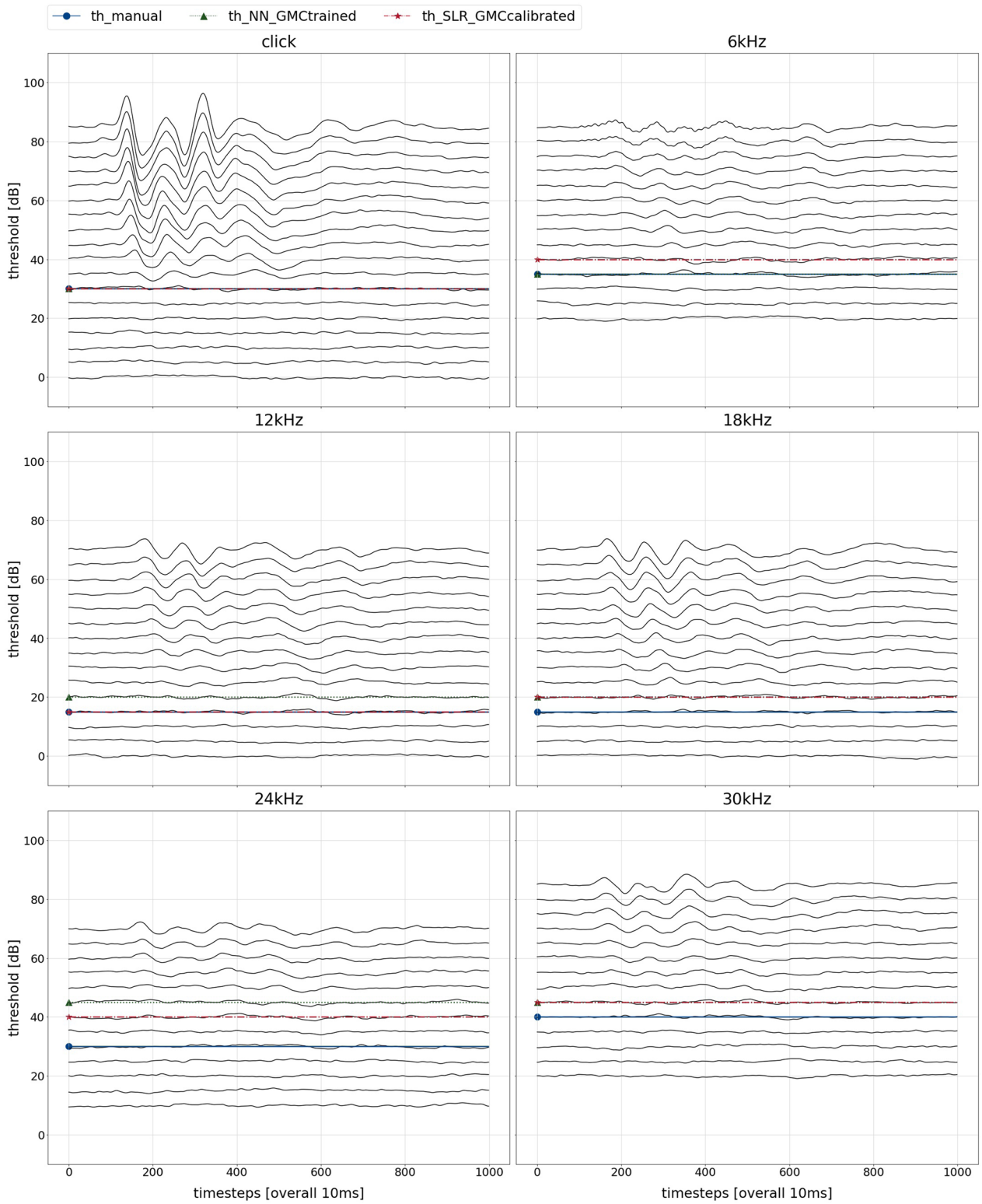


Fig. 8 Audiograms and hearing thresholds of a single GMC mouse. For all six stimuli, the stacked averaged response signals of an individual mouse are shown. The x-axis covers a time span of 10 ms in 1000 time steps. The y-axis shows stacked response signal strengths, with each curve corresponding to a sound pressure level. Ticks in 20 dB steps indicate where each SPL curve begins. Overlaid horizontal lines indicate hearing thresholds assigned by three methods: manual, by GMC reader (blue, solid line and circle); NN, GMC-trained (green, dotted line and triangle); SLR, GMC-calibrated (red, dashed line and star)

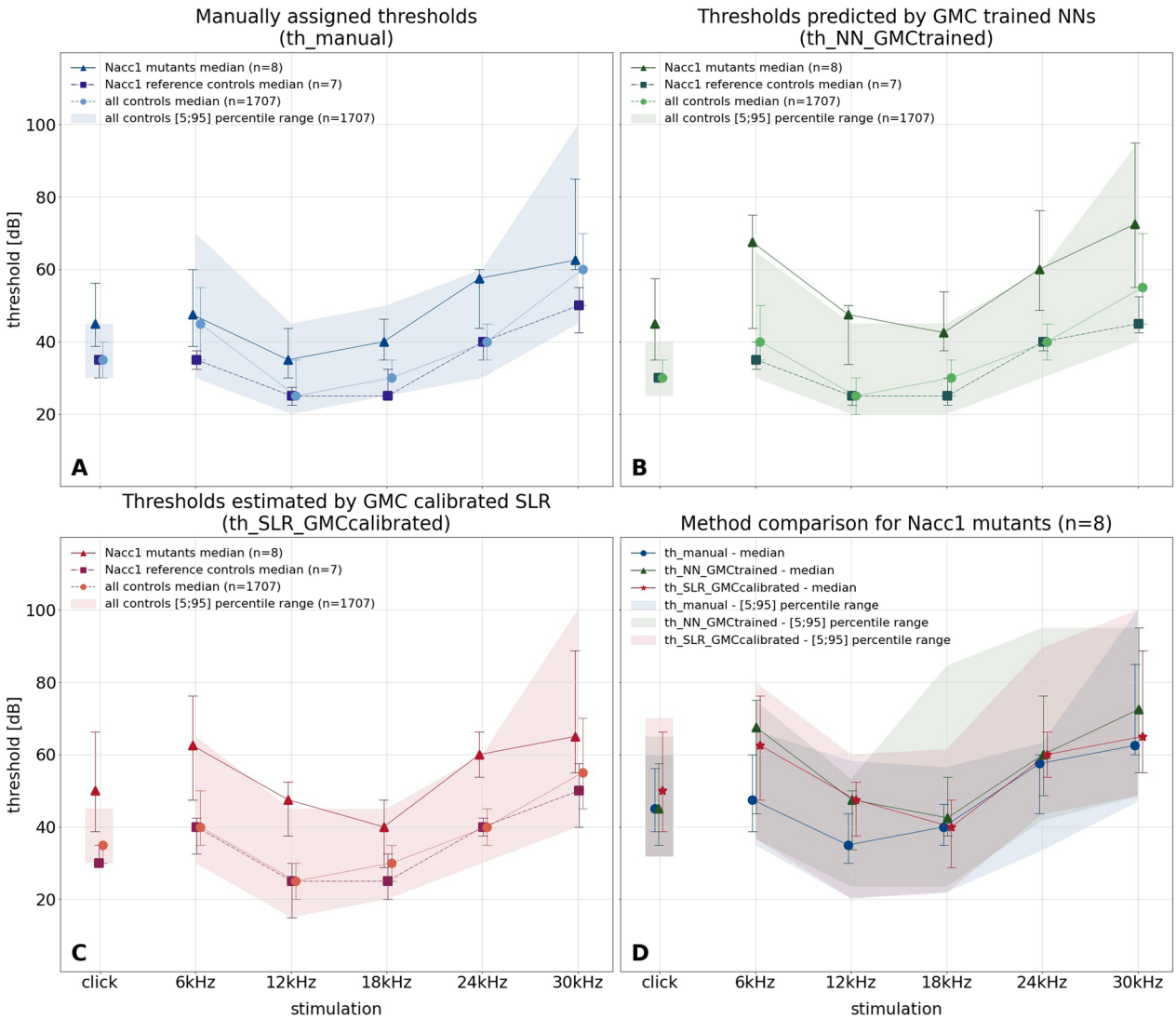


Fig. 9 Group-based comparison of three threshold finding methods. On the example of the *Nacc1*$^{em1(IMPC)Hmgu}$ mouse line, the figure shows an end-to-end comparison of the three threshold finding methods for the use case of identifying candidate genes with hearing function involvement. For each method (manual: top left, NN: top right, SLR: bottom left), stimulus-specific hearing thresholds are compiled to median hearing curves of mutants (triangles and solid lines), same-day reference controls (squares and dashed lines) and "all controls" (circles and dotted lines) and shown as symbols, which are connected for the non-click stimuli. The ribbon shows the [5;95] percentile range of all controls. Numbers of mice in each group are shown in the legend. On the y-axis, hearing threshold is given in 20 dB ticks. The x-axis shows the stimulus. Whiskers indicate inter-quartile ranges. Each method-specific plot allows comparison of mutant vs. control hearing curves with thresholds based on that method. The bottom-right figure shows mutant-only median hearing curves and [5;95] percentile ranges of the three methods (manual: blue, solid line and circle; NN: green, dotted line and triangle; SLR: red, dashed line and star) to allow method comparison

mouse in the data sets. For all six stimuli, ABR responses as well as respective manual, NN, and SLR thresholds are plotted.

Next, for all GMC lines, hearing curves were generated that show mutant vs. control group medians, with a background indicating the [5;95] percentile range of all control animals. This is done in separate subplots for manual, NN, and SLR thresholds, to allow comparison of hearing curve differences of mutants and controls between methods. A forth subplot only shows overlaid mutant median hearing curves for all three methods. Figure 9 shows on the example of the *Nacc1*$^{em1(IMPC)Hmgu}$ mouse line, that all methods are able to detect the shift of the hearing curve in mutants. This use case shows a clear advantage of the algorithmic methods: there may be a systematic shift with regards to the manual method. However,

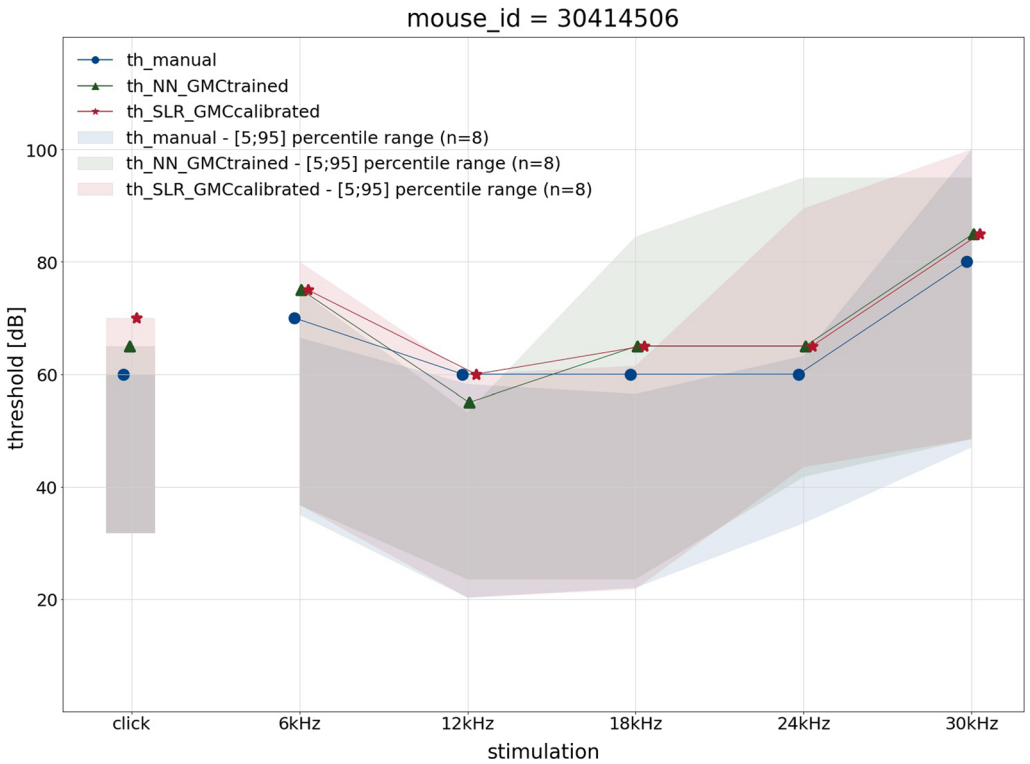


Fig. 10 Mouse-based comparison of three threshold finding methods. On the example of an individual *Nacc1*$^{em1(IMPC)Hmgu}$ mutant mouse, the figure shows an end-to-end comparison of the three threshold finding methods. Stimulus-specific hearing thresholds are displayed as hearing curves gained by manual (blue, solid line and circle), NN (green, dotted line and triangle) and SLR (red, dashed line and star) threshold finding. The ribbons show the respective [5;95] percentile range of all mutants of the same gene in the respective colour. On the y-axis, hearing threshold is given in 20 dB ticks. The x-axis shows the stimulus

it applies to both mutants and controls, conserving any differences between both. Both methods can also be considered blinded, as they are not aware to which group an ABR response signal belongs.

Finally, for each mouse, another plot shows an overlay of hearing curves for the three methods in comparison. Figure 10 shows an example of a *Nacc1*$^{em1(IMPC)Hmgu}$ mouse where all three methods agree quite well.

All plots are made publicly available and can be used to validate and compare the methods on the original data.

Fully automated identification of candidate genes Visual comparison of hearing curves is indispensable for evaluation purposes, however not feasible for screening, since it is laborious and, similar to curve reading, it may be prone to bias. Therefore, a programmatic approach has been implemented that uses two measures as criteria to detect mutant mouse lines that exhibit potential biologically meaningful changes in hearing. First, effect size, which descriptively spoken measures the degree of overlap between mutant and control group distribution of a stimulus-specific threshold. As no normal distribution can be assumed, Cliff's Delta was used, which ranges between -1 and 1. Second, significance, using p-values resulting from a Wilcoxon rank sum test, defined as the probability of getting a test statistics as large or larger assuming mutant and control distribution are the same. A well-established way of displaying these two measures is the so-called volcano plot. Figure 11 shows such volcano plots for click and 30 kHz thresholds of GMC lines for all three methods. Here, interesting lines—i.e. lines that exhibit a biologically meaningful hearing phenotype—are supposed to be those that show high significance and a large effect size at the same time. Using $p < 0.05$ and $|d| > 0.474$ for large effects [38], candidate mouse lines can be found in the upper left (lower threshold) and upper right (higher threshold) area of the plots and of course can be directly filtered to result lists.

Additional file 1: Tables S1 and S2 each show a method comparison of top candidate lines/genes

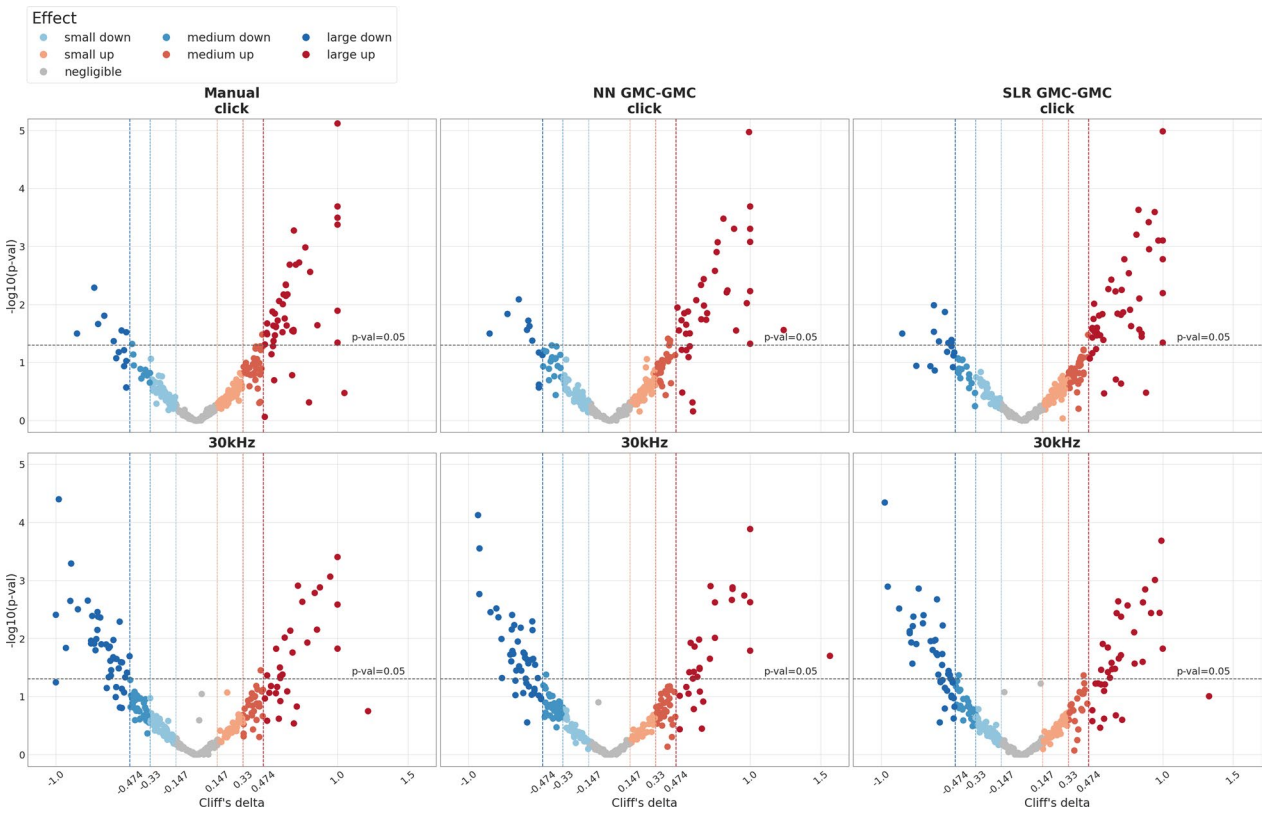


Fig. 11 Biologically relevant changes in hearing thresholds—GMC lines, click and 30 kHz. Volcano plots show significance vs. effect size for all GMC lines. For each mouse line, represented by a dot, hearing thresholds were used to calculate significance (Wilcoxon rank sum test, y-axis) and non-parametric effect size (Cliff's delta [39], x-axis) of mutant vs. control animals. Vertical lines indicate margins for small (0.147), medium (0.33) and large (0.474) effects as suggested in [38]. An effect size of 1 indicates that the difference between means of mutants and controls is equal to one standard deviation. The horizontal line indicates the 0.05 significance threshold level. Accordingly, dots in the upper left and upper right areas denote GMC lines with significant as well as relevant changes and thus are considered worthwhile candidates (see Additional file 1: Tables S1 and S2). Dot colours in addition represent effect size as shown in the legend. Plot rows represent click (upper) and 30 kHz (lower) stimulus data. Columns compare the three hearing threshold finding methods compared in this work (left: manual, middle: NN, right: SLR)

for modified threshold at click and 30 kHz stimulus, respectively. Not surprisingly, lists are largely similar, although not completely. For example, all three methods identified *Gpsm2* as well as *Rest*, two well-known hearing loss genes [40], while other hits differ at least at single frequencies. To further improve facilitated identification of candidate genes, a new plot displays calculated effect sizes for all stimuli and all three methods. Figure 12shows on two examples of this highly integrated plot, that it allows to rapidly evaluate ABR results in two ways: (a) assess effect sizes for the different stimuli and thus judge the nature of hearing impairment, (b) compare effect sizes derived from different methods. As can be seen for *Hunk*, all three methods end up in almost identical conventional hearing curves and, accordingly, effect size plots. For another gene, *Ngdn*, both plot types show that the automated methods differ from manual threshold finding in delivering consistently smaller effects.

An end-to-end analysis pipeline using SLR based thresholds reveals 76 candidate genes with impact on hearing sensitivity In a re-analysis of the GMC raw data set, hearing thresholds derived from both automated methods (SLR and NN) were used for identification of candidate lines as described above. For click stimulation, the visual and/or the fully automated method identified six genes (*Vps13c, Rabgap1, Ttll12, Hdac1, Adprm, and Kansl1l*, Fig. 13) with strong effects that had not been detected so far using manual thresholds only. For 30 kHz stimulation,

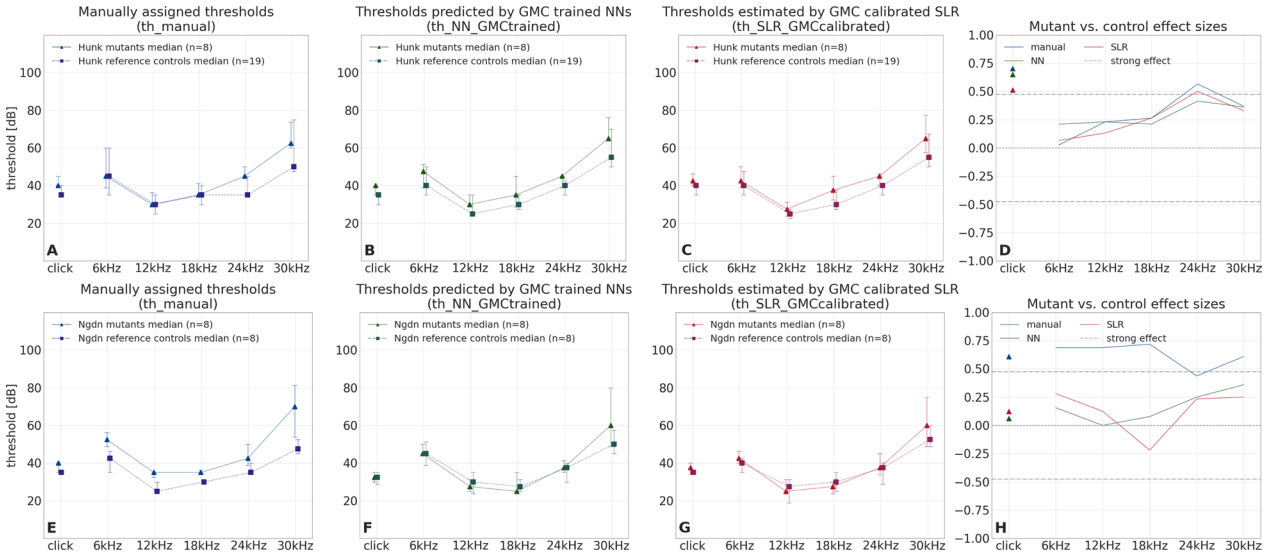


Fig. 12 Comparison of mutant vs. control effect sizes between three hearing threshold finding methods. For two genes (Top: *Hunk*, Bottom: *Ngdn*), hearing thresholds determined by three different methods (manual, NN, SLR) are compared using conventional plots (panels **A**–**C**, **E**–**G**) and a new effect size plot (panels **D**, **H**). In the conventional plots, stimulus-specific hearing thresholds are compiled to median hearing curves of mutants (triangles and solid lines) and controls (squares and dashed lines), which are connected for the non-click stimuli. Whiskers indicate inter-quartile ranges. Numbers of mice in each group are shown in the legend. On the y-axis, hearing threshold is given in 20 dB ticks. The x-axis shows the stimulus. Effect size plots (**D**, **H**) show non-parametric effect size (Cliff's delta [39], y-axis) of mutant vs. control animals for all six stimuli (x-axis). Colours indicate the threshold finding method (manual: blue, NN: red, SLR: orange). The grey dashdotted horizontal lines show the thresholds for strong effects at ±0.474. The central dotted horizontal line indicates the zero effect size level. Lines connect effect sizes of the same method. Top: For *Hunk*, effect sizes differ only marginally for all stimuli, consistent with mutant and control curves running close together in the conventional plots **A**–**C**. Bottom: For *Ngdn*, smaller effects are found with NN and SLR consistently throughout all stimuli, whereas manual thresholds show large effects. Again, this is consistent with the conventional plots (**E**–**G**), where mutant and control curves are clearly apart for manual thresholds (**E**), but close together for NN (**F**) and SLR (**G**). Interpretation aid: the effect size relates the difference between mutant and control medians to data variance (indicated by the interquartile range whiskers)

two new candidate genes (*Alkbh6 and Mgat1*, Fig. 14) were identified. For a set of three other manually identified candidate genes (*Ngdn, Gpatch2l, Gdi2*, Fig. 15), NN and SLR derived thresholds did not lead to strong effects at click or 30 kHz.

Of course, evaluation of hearing deficits is not relying on differences at single frequencies. For identifying genes with impact on hearing sensitivity, the evaluation of single thresholds is the basis for analysis. Additional steps will include the definition of relevant effect sizes and patterns of alteration. To further explore these potential hearing genes, databases for human variants, expression patterns, pathways etc. will have to strengthen the evidence for candidate genes. In addition, confirmation of results with calculated sample sizes and/or separation of sexes is needed in some cases.

Altogether, 76 potential hearing genes have been detected by automated analysis starting from raw data using SLR (see Additional file 1: Tables S1 and S2s, unique entries from combined SLR columns). For four of them (*Hoxa2, Aspa, Gpsm2*, and *Rest*), human orthologue genes have published annotations for human hearing loss according to OMIM® [41]. Inner ear gene expression was evaluated by literature [42, 43] and eleven of the genes were reported to be expressed in hair cells or surrounding cells. For 35 of the genes, no mouse model was yet listed at the Mouse Genome Database (MGD) [44], while for 37 of them with a mouse model available no information about hearing sensitivity was provided. Solely for four of the mouse models, either altered hearing or middle ear morphology was reported (*Rest, Gpsm2, Aspa*, and *Hoxa2*). Some of the genes are already associated with human disease, underlining the pleiotrophy of gene functions and phenotypes. For example, *Btbd 9* is associated with restleg legs syndrome (RLS, OMIM 611185), but is also expressed in outer hair cells [43[, thus

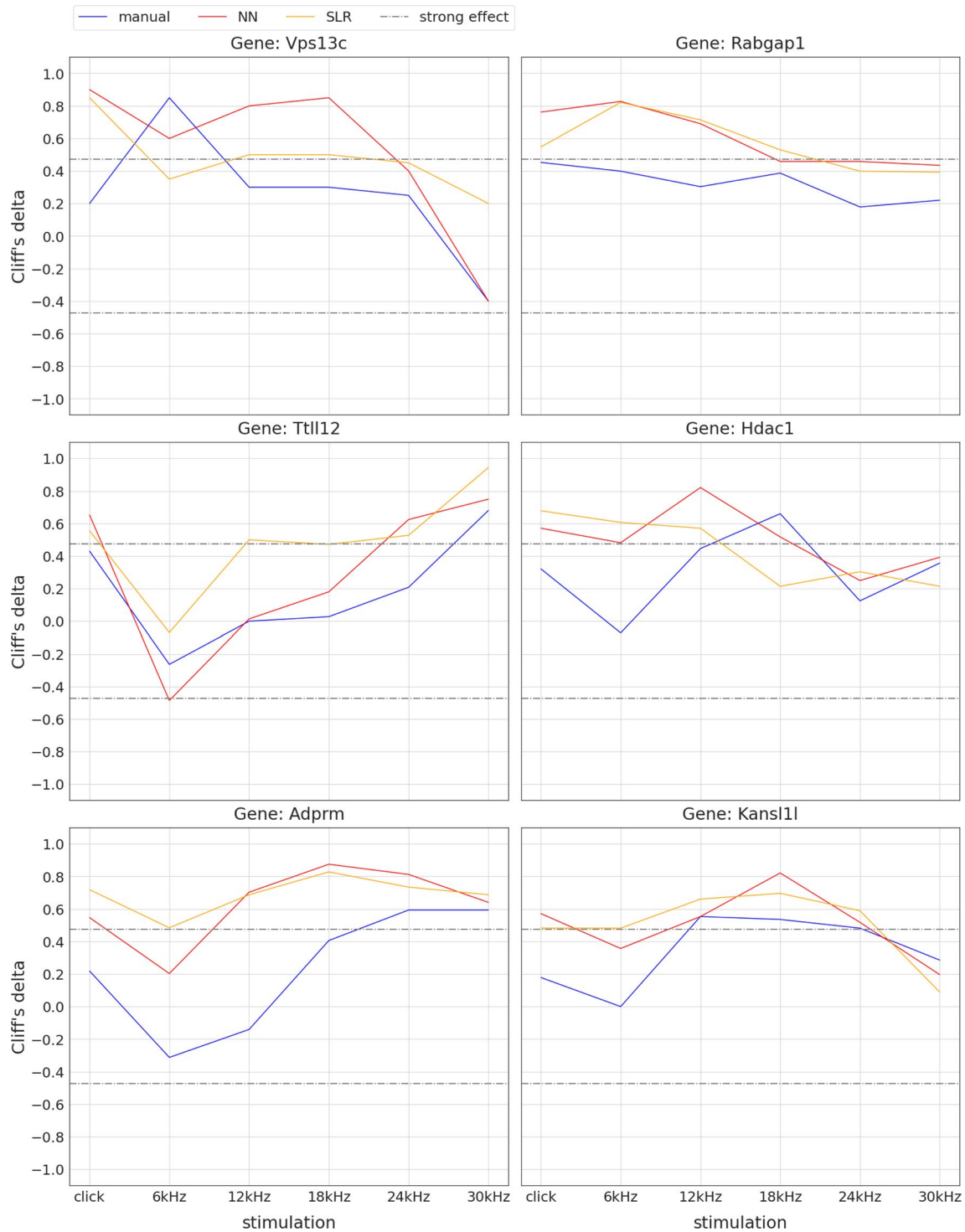


Fig. 13 Visual identification of six new candidate genes with hearing impact at click stimulation. Plots show non-parametric effect size (Cliff's delta [39], y-axis) of mutant vs. control animals for all six stimuli (x-axis). Colours indicate the threshold finding method (manual: blue, NN: red, SLR: orange). The grey dashdotted horizontal lines show the thresholds for strong effects at ±0.474. For convenience, lines connect effect sizes of the same method (Note: this is not a hearing curve, click has to be interpreted separately). For each of the six genes shown, click effect sizes of NN and SLR derived thresholds are above the dashdotted line, indicating strong effects, whereas manual threshold effects are below

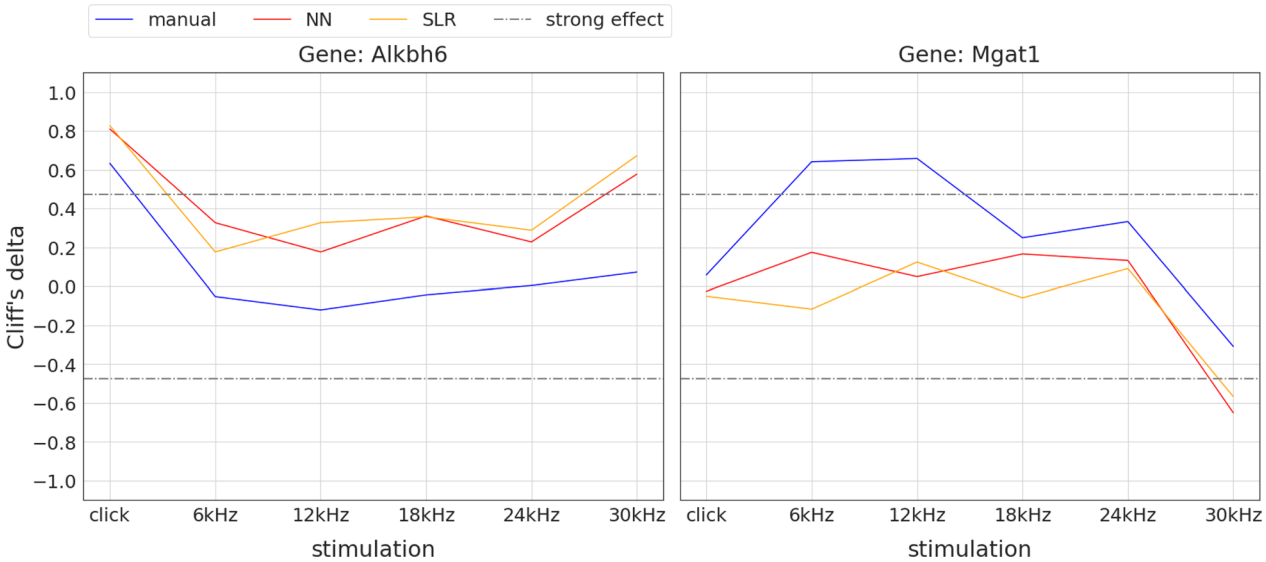


Fig. 14 Visual identification of two new candidate genes with hearing impact at 30 kHz stimulation. Plots show non-parametric effect size (Cliff's delta [39], y-axis) of mutant vs. control animals for all six stimuli (x-axis). Colours indicate the threshold finding method (manual: blue, NN: red, SLR: orange). The grey dashdotted horizontal lines show the thresholds for strong effects at ±0.474. For convenience, lines connect effect sizes of the same method (Note: this is not a hearing curve, click has to be interpreted separately). For the two genes shown, 30 kHz effect sizes of NN and SLR derived thresholds are above the +0.474 (left: *Alkbh6*) and below the −0.474 (right: *Mgat1*) dashdotted line, indicating strong effects, whereas manual threshold effects don't

providing a possible link to the detected hearing alteration. Further analysis will be needed for the possible candidate genes to uncover the nature of gene-phenotype association.

Conclusions

Using two independent and large data sets, this work shows that two new methods are robust and able to objectively detect hearing thresholds from averaged ABR raw data. While the supervised NN method, using two neural networks, achieves higher accuracies for manual ground truth, it requires training with large numbers of human-assigned labels and cannot be transferred between data sets. Thus, it may be preferred by large laboratories with high level manual thresholding standards. The self-supervised Sound Level Regression—SLR—method does not depend on labels and thus can be directly applied to any ABR data set.

Both methods have the advantage of delivering highly consistent results. As they can be employed in fully-integrated end-to-end pipelines, they are predestined for use in routine measurements, quality control, and automated retrospective re-analysis of large ABR data collections. Since SLR is invariant to the data set, it offers itself as a method for meta analysis of ABR data from different institutions.

In a mutant screening environment, both NN and SLR can be integrated into a fully automated end-to-end pipeline, starting from raw averaged ABR data and finally producing candidate lists and plots. By adding SLR-based threshold calling to the ABR curve reading tool of the German Mouse Clinic, a time saving of approx. 3 person hours/40 mice is sought. In the current phase, SLR calls still need human validation. However, a much greater advantage of automatic thresholding is that it allows for consistent re-analysis of large raw ABR data sets collected over large periods of time. What is hardly feasible with a manual approach is a matter of hours with SLR.

The decision to trust NN- and SLR-derived thresholds over manual derived thresholds is subjective. However, this work—using two independent data sets—supplies a solid foundation of data, results and comparative plots to allow external validation by experts, using visual curve reading. In addition, the provided methods allow comparative analysis of all methods using own data.

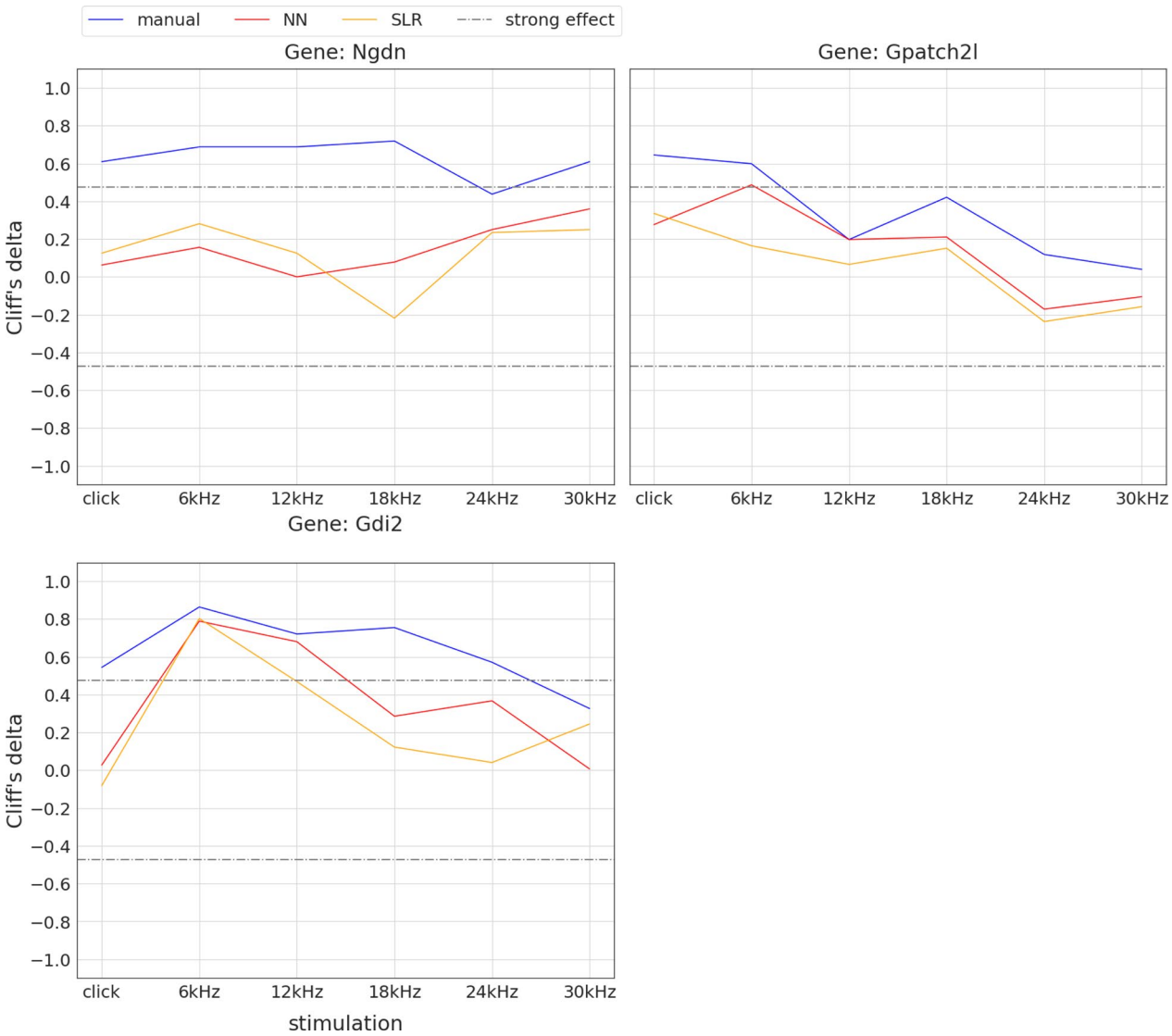


Fig. 15 Loss of hearing sensitivity gene candidate status using automated threshold finding. Plots show non-parametric effect size (Cliff's delta [39], y-axis) of mutant vs. control animals for all six stimuli (x-axis). Colours indicate the threshold finding method (manual: blue, NN: red, SLR: orange). The grey dashdotted horizontal lines show the thresholds for strong effects at ±0.474. For convenience, lines connect effect sizes of the same method (Note: this is not a hearing curve, click has to be interpreted separately). For the three genes shown, click and/or 30 kHz manual threshold effect sizes are above the +0.474 dashdotted line, indicating strong effects, whereas SLR and NN threshold effects don't

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12868-022-00758-0.

Additional file 1: **S1.** Neural network model architectures. **S2.** Evaluation curves. **S3.** Volcano plots of GMC mutant lines. **S4.** Comparison of top click candidate genes with hearing threshold changes. **S5.** Comparison of top 30 kHz candidate genes with hearing threshold changes

Additional file 2: **S6.** Information on 76 SLR-based candidate genes

Acknowledgements
We thank D. Feeser, A. Badmann, R. Fischer, E. Köfferlein, and F. Schleicher for ABR measurements and identification of hearing thresholds. We also thank R. Steinkamp for data capture as well as V. Gailus-Durner and H. Fuchs for critically reading the manuscript.

Author contributions
DT: conceptualisation, methodology, software, formal analysis, writing—original draft, visualisation ES: software, validation, data curation, visualisation GM: conceptualisation, methodology, software, formal analysis, data curation AH: software, visualisation MHdA: resources, supervision, funding acquisition,

writing—review & Editing LB: validation, investigation, data curation, writing—review & editing CLM: conceptualisation, writing —review & editing, supervision HM: conceptualisation, writing—original draft, writing—review & editing, supervision, Project administration. All authors read and approved the final manuscript.

Authors' information
Dominik Thalmeier and Gregor Miller have equal distribution. Correspondence to Christian L. Müller and Martin Hrabě de Angelis.

Funding
Open Access funding enabled and organized by Projekt DEAL. D. Thalmeier and C.L. Müller were funded by Helmholtz Association's Initiative and Networking Fund through Helmholtz AI.

Availability of data and materials
The datasets generated and/or analysed during the current study are available in the zenodo.org repository, (http://dx.doi.org/doi:10.5281/zenodo.5779876). Code of this work is available at the GitHub repository, https://github.com/ExperimentalGenetics/ABR_thresholder.

Declarations

Ethics approval and consent to participate
Original mouse husbandry and animal experiments were carried out in accordance with European Directive 2010/63/EU and following the approval by the responsible local authority of the ***Regierung von Oberbayern***, Germany.

Consent for publication
Not applicable.

Competing interests
The authors of this manuscript declare no potential financial, personal or other conflicts with other people or organisations that could inappropriately influence their work.

Author details
[1]Institute of Computational Biology, Helmholtz Zentrum München, München, Germany. [2]Institute of Experimental Genetics, Helmholtz Zentrum München, München, Germany. [3]Helmholtz AI, Helmholtz Zentrum München, München, Germany. [4]Department of Statistics, LMU München, München, Germany. [5]Center for Computational Mathematics, Flatiron Institute, New York, USA. [6]German Center for Diabetes Research (DZD), Neuherberg, Germany. [7]Chair of Experimental Genetics, School of Life Science Weihenstephan, Technische Universität München, Freising, Germany.

Received: 7 April 2022 Accepted: 18 November 2022
Published online: 27 December 2022

References

1. James S, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392:1789–858. https://doi.org/10.1016/S0140-6736(18)32279-7.
2. Cunningham L, Tucci D. Hearing loss in adults. New Engl J Med. 2017;377:2465–73. https://doi.org/10.1056/NEJMra1616601.
3. Ingham NJ, Pearson SA, Vancollie VE, et al. Mouse screen reveals multiple new genes underlying mouse and human hearing loss. PLoS Biol. 2019;17.4: e3000194. https://doi.org/10.1371/journal.pbio.3000194.
4. Bowl M, et al. A large scale hearing loss screen reveals an extensive unexplored genetic landscape for auditory dysfunction. Nat Commun. 2017. https://doi.org/10.1038/s41467-017-00595-4.
5. Dickinson M, et al. High-throughput discovery of novel developmental phenotypes. Nature. 2016. https://doi.org/10.1038/nature19356.
6. Meehan TF, Conte N, West DB, et al. Disease model discovery from 3,328 gene knockouts by the international mouse phenotyping consortium. Nat Gen. 2017;49:1231–8. https://doi.org/10.1038/ng.3901.
7. Ingham NJ, Pearson S, Steel KP. Using the Auditory Brainstem Response (ABR) to determine sensitivity of hearing in mutant mice. Curr Protocols Mouse Biol. 2011;1(2):279–87. https://doi.org/10.1002/9780470942390.mo110059.
8. Gans D, Zotto DD, Gans KD. Bias in scoring auditory brainstem responses. Br J Audiol. 1992;26(6):363–8. https://doi.org/10.3109/03005369209076660.
9. Arnold SA. Objective versus visual detection of the auditory brain stem response. Ear Hear. 1985;6(3):144–50. https://doi.org/10.1097/00003446-198505000-00004.
10. Vidler M, Parkert D. Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test. Int J Audiol. 2004;43:417–29. https://doi.org/10.1080/14992020400050053.
11. Zaitoun M, Cumming S, Purcell A. Inter and intra-reader agreement among audiologists in reading auditory brainstem response waves. Can J Speech-Language Pathol Audiol. 2014;38:440–9.
12. Pandiyan PM, et al. A machine learning approach for distinguishing hearing perception level using auditory evoked potentials. IEEE Conference on Biomed Eng and Sci (IECBES). 2014. https://doi.org/10.1109/IECBES.2014.7047661.
13. Carter L, et al. The detection of infant cortical auditory evoked potentials (CAEPs) Using Statistical and Visual Detection Techniques. J Am Acad Audiol. 2010;21:347–56. https://doi.org/10.3766/jaaa.21.5.6.
14. Alpsan D, et al. Determining hearing threshold from Brain Stem Evoked Potentials: optimising a neural network to improve classification performance. Eng Med Biol Mag. 1994;13:465–71. https://doi.org/10.1109/51.310986.
15. Vannier E, Adam O, Motsch JF. Objective detection of brainstem auditory evoked potentials with a priori information from higher presentation levels. Artif Intell Med. 2002;25:283–301. https://doi.org/10.1016/S0933-3657(02)00029-5.
16. Lundt A, et al. Data acquisition and analysis In brainstem evoked response audiometry in mice. J Vis Exp JoVE. 2019. https://doi.org/10.3791/59200.
17. Dobie RA, Wilson MJ. Analysis of auditory evoked potentials by magnitude-squared coherence. Ear Hear. 1989;10(1):2–13. https://doi.org/10.1097/00003446-198902000-00002.
18. Acır N, Özdamar Ö, Güzeliş C. Automatic classification of auditory brainstem responses using SVM-based feature selection algorithm for threshold detection. Eng Appl Art Intelligence. 2006;19.2:209–18. https://doi.org/10.1016/j.engappai.2005.08.004.
19. Berninger E, Olofsson Å, Leijon A. Analysis of click-evoked auditory brainstem responses using time domain cross-correlations between interleaved responses. Ear Hear. 2014;35(3):318–29. https://doi.org/10.1097/01.aud.0000441035.40169.f2.
20. Bogaerts S, et al. Automated threshold detection for auditory brainstem responses: comparison with visual estimation in a stem cell transplantation study. BMC Neurosci. 2009;10:104. https://doi.org/10.1186/1471-2202-10-104.
21. Cebulla M, Stürzebecher E, Wernecke K-D. Objective detection of auditory brainstem potentials: comparison of statistical tests in the time and frequency domains. Scand Audiol. 2000;29(1):44–51. https://doi.org/10.1080/010503900424598.
22. Chesnaye M, et al. Objective measures for detecting the auditory brainstem response: comparisons of specificity, sensitivity and detection time. Int J Audiol. 2018;57:1–11. https://doi.org/10.1080/14992027.2018.1447697.
23. Cone-Wesson BK, Hill KG, Liu G-B. Auditory brainstem response in tammar wallaby (*Macropus eugenii*). Hearing Res. 1997;105(1–2):119–29. https://doi.org/10.1016/s0378-5955(96)00199-2.
24. Dobrowolski A, et al. Classification of auditory brainstem response using wavelet decomposition and SVM network. Biocybernet Biomed Eng. 2016;36.2:427–36. https://doi.org/10.1016/j.bbe.2016.01.003.
25. Mccullagh P, et al. A comparison of supervised classification methods for auditory brainstem response determination. Stud Health Technol Inform. 2007;129:1289–93. https://doi.org/10.3233/978-1-58603-774-1-1289.
26. Özdamar O, et al. Computer methods for on-line hearing testing with auditory brain stem responses. Ear Hear. 1990;11(6):417–29. https://doi.org/10.1097/00003446-199012000-00003.

27. Özdamar O, et al. Automated electrophysiologic hearing testing using a thresholdseeking algorithm. J Am Acad Audiol. 1994;5(2):77–88.
28. Achim S, Richard G, Patrick K, et al. Objective estimation of sensory thresholds based on neurophysiological parameters. Front Neurosci. 2019;13:481. https://doi.org/10.3389/fnins.2019.00481.
29. Suthakar K, Liberman M. A simple algorithm for objective threshold determination of auditory brainstem responses. Hear Res. 2019;381:107782. https://doi.org/10.1016/j.heares.2019.107782.
30. Wang H, et al. Automated threshold determination of auditory evoked brainstem responses by cross-correlation analysis with varying sweep number. medRxiv. 2020. https://doi.org/10.1101/19003301.
31. Lv J, Simpson D, Bell S. Objective detection of evoked potentials using a bootstrap technique. Med Eng Phys. 2007;29:191–8. https://doi.org/10.1016/j.medengphy.2006.03.001.
32. Davey R, et al. Auditory brainstem response classification: a hybrid model using time and frequency features. Artificial Intelligence Med. 2007;40:1–14. https://doi.org/10.1016/j.artmed.2006.07.001.
33. McKearney RM, MacKinnon RC. Objective auditory brainstem response classification using machine learning. Int J Audiol. 2019;58(4):224–30. https://doi.org/10.1080/14992027.2018.1551633.
34. Cheng C, Li Z, Xiaoxin P, et al. Automatic recognition of auditory brainstem response characteristic waveform based on bidirectional long short-term memory. Front Med. 2021;7:1027. https://doi.org/10.3389/fmed.2020.613708.
35. Gailus-Durner V, et al. Introducing the German Mouse Clinic: open access platform for standardized phenotyping. Nat Method. 2005;2:403–4. https://doi.org/10.1038/nmeth0605-403.
36. Helmut F, et al. The German Mouse Clinic: a platform for systemic phenotype analysis of mouse models. Curr Pharm Biotechnol. 2009. https://doi.org/10.2174/138920109787315051.
37. Ingham NJ, Pearson SA, Vancollie VE, et al. Data from: mouse screen reveals multiple new genes underlying mouse and human hearing loss. 2019. https://doi.org/10.5061/DRYAD.CV803RV.
38. Romano Jeanine, Kromrey Jeffrey (2006) Appropriate Statistics for Ordinal Level Data: Should We Really Be Using t-test and Cohens d for Evaluating Group Differences on the NSSE and other Surveys? In: Annual meeting of the Southern Association for Institutional Research. url:https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.595.6157 &rep=rep1 &type=pdf.
39. Cli N. Dominance statistics: ordinal analyses to answer ordinal questions. Psychol Bull. 1993;114:494–509. https://doi.org/10.1037/0033-2909.114.3.494.
40. Camp G Van, Smith RJH (2021). Hereditary Hearing Loss Homepage. https://hereditaryhearingloss.org. Accessed 29 Nov 2021.
41. Amberger JS, Bocchini CA, Schiettecatte F, et al. OMIM.org: online Mendelian Inheritance in Man (OMIMR), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2014;43(D1):D789–98. https://doi.org/10.1093/nar/gku1205.
42. Scheffer DI, et al. Gene expression by mouse inner ear hair cells during development. J Neurosci. 2015;35(16):6366–80. https://doi.org/10.1523/jneurosci.5126-14.2015.
43. Ranum PT, Goodwin AT, Yoshimura H, et al. Insights into the biology of hearing and deafness revealed by single-cell RNA sequencing. Cell Rep. 2019;26(11):3160-3171.e3. https://doi.org/10.1016/j.celrep.2019.02.053.
44. Bult CJ, et al. Mouse genome database (MGD) 2019. Nucleic Acids Res. 2018;47(D1):D801–6. https://doi.org/10.1093/nar/gky1056.

Publisher's Note