DISENTANGLING DEVICE AND NEURAL CONTRIBUTIONS TO SPATIAL HEARING DEFICITS IN COCHLEAR IMPLANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial hearing is essential for spatial orientation, auditory scene analysis, and speech perception. Cochlear implant (CI) users localize sounds poorly, and these deficits can arise from both device-related factors that distort spatial cues and from neural degeneration in the peripheral and central auditory pathways. Because these factors are typically intertwined, their individual contributions to localization deficits are difficult to isolate. To disentangle these contributions, we introduce *Hybrid-CINN*, a model that combines a biophysical CI simulation with a convolutional neural network trained to localize natural sounds and rendering spatial signals with behind-the-ear (BTE) head-related transfer functions (HRTFs) to reflect realistic CI listening. Our model, trained under ideal conditions reproduces key characteristics of CI sound localization effects, including lateral response compression and a strong reliance on interaural level-difference (ILD) cues. We then introduce device-related factors and neural degeneration individually, retraining the network under each condition to simulate chronic adaptation. We find that electrode insertion mismatch degrades localization performance, while linking bilateral automatic gain control improves it; and modeled central degeneration has a greater impact on localization than peripheral degeneration. Together, these results establish *Hybrid-CINN* as a testbed for disentangling the factors that limit spatial hearing in CI users and for evaluating CI processing and fitting strategies before they reach patients.

1 INTRODUCTION

Spatial hearing (SH) is a fundamental aspect of auditory perception. It allows listeners to identify the direction of sound sources, orient attention, and separate target signals from competing background noise (Grothe et al., 2010). Normal-hearing (NH) listeners localize sources in the horizontal plane with a root-mean-square error (RMSE) of about 6° (Dorman et al., 2016), combining both binaural cues such as interaural time differences (ITDs), interaural level differences (ILDs), and monoaural cues e.g. spectral cues imposed by the pinnae (Macpherson & Middlebrooks, 2002). For cochlear implant (CIs) users, spatial hearing remains substantially impaired with bilateral CI localization errors of 20–30° RMSE(s) and single-implant users exhibit close to no SH (Grantham et al., 2007; Kerber & Seeber, 2012; Dorman et al., 2016). Candidate causes fall into three groups: device-related, neural, and listener-specific. On the device side, CI coding strategies discard temporal fine structure, leaving the ITD cue, which dominates localization in NH listeners, largely inaccessible (Laback et al., 2015; Carlyon et al., 2025). The number of stimulating electrodes limits the spectral resolution of sounds, CI processing compresses ILDs (Dorman et al., 2014), and automatic gain control (AGC) acting independently in each processor further distorts them (Archer-Boyd & Carlyon, 2019). The two electrode arrays are rarely inserted to matched depths in each cochlea leading to excitation mismatches in frequency (Goupell et al., 2022; Kan et al., 2013). Behind-the-ear (BTE) microphones largely remove the sounds monaural spectral cues. On the neural side, deafness depletes the spiral ganglion to an extent that depends on etiology and duration and can differ between the ears (Nadol et al., 1989; Anderson et al., 2024). Prolonged acoustic deprivation further degrades neural ITD coding and reorganizes auditory cortex (Hancock et al., 2013; Kral & Sharma, 2012). Neural survival cannot currently be directly measured in a living listener (Banerjee et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Hence, direct evidence on how degeneration affects bilateral CI localization is scarce. Because each patient carries a different combination of these factors, behavioral data cannot separate their individual contributions. This leaves open a question central to CI development: which limitations could be addressed by better processing and fitting, and which are set by neural degeneration?

Computational models can be used to vary these factors individually. Mechanistic models simulate the CI processor, the electrode–nerve interface, and brainstem binaural neurons in detail, but read out with a hand-designed decision stage (Kelvasa & Dietz, 2015; Klug et al., 2020; Hu et al., 2023). A fixed read-out confounds the available information in input vs. ability to extract information. Task-optimized networks avoid this by learning the read-out and when trained to localize natural sounds they reproduce a broad set of psycho-acoustic SH-effects in NH listeners (Francl & McDermott, 2022; Saddler & McDermott, 2024). Banerjee et al. (2025) recently applied this approach to simulated CI input and matched speech recognition and localization of the best CI users. Their analysis, however, focused on speech in noise and no task optimized model has yet attributed CI localization error to device versus neural factors.

Here, we introduce *Hybrid-CINN* (Hybrid Cochlear Implant Neural Network, Fig. 1), a model of a bilateral CI listener that couples a biophysical CI front end Hu et al. (2023) to a convolutional network trained to localize natural sounds from simulated auditory-nerve spike trains. Unlike previous studies, sounds are rendered with BTE head-related transfer functions (HRTFs) matching the input captured by CI user's microphones. Our design allows manipulation at two levels: device and peripheral factors (insertion depth, AGC linking, fiber survival) as physical parameters and central capacity by lesioning the learned decoder. We then degrade the pipeline one factor at a time and retrain from scratch under each condition. Each model thus adapts to its own impairment, analogous to chronic adaptation in implanted listeners, and each factor's cost reflects what adaptation cannot compensate. Unless noted, model–human comparisons use rooms and HRTFs held out of training, and we restrict our claims to horizontal-plane localization. Our contributions are:

- **A hybrid model of CI localization** *Hybrid-CINN* couples a biophysical CI front end to a task-optimized decoder, which reproduces key signatures of CI localization: lateral response compression, reliance on ILDs over ITDs, and poor localization of low-frequency sounds.
- **A lesion-and-retrain protocol** Degrading each factor separately and retraining under impairment separates the information the CI signal carries from the decoder that reads it, and measures each impairment's cost after adaptation.
- **Attributing effects on localization** In the model, insertion mismatch, asymmetric or basal fiber loss and central loss impair localization, whereas linking AGCs improve it.

2 Related Work

Task-optimized models of hearing. Networks trained on auditory tasks are a standard tool for asking which behaviors follow from the task and the input statistics alone (Yamins & DiCarlo, 2016; Kell et al., 2018). For localization, convolutional neural networks (CNNs) trained on binaural signals in simulated rooms develop human-like cue weighting, spectral-cue use, and precedence effects (Francl & McDermott, 2022), and the temporal precision they need depends on the task (Saddler & McDermott, 2024). These models, and the likes of end-to-end (Vecchiotti et al., 2019) and Bayesian (Barumerli et al., 2023) localizers, address NH listening only. In CI research, machine learning has served mainly for front-end signal enhancement (Goehring et al., 2017). A notable exception, and the work closest to ours, is Banerjee et al. (2025), who trained deep networks on simulated CI input for speech recognition and localization, using a linear electrode–nerve interface and KEMAR HRTFs. We differ in three ways. We use a stochastic nerve model with BTE HRTFs. We treat localization as the object of study. And we sweep device-fitting and neural-lesion factors.

Mechanistic models of CI hearing. Mechanistic models chain CI processing, the electrode–nerve interface, and brainstem stages into full pipelines (Kelvasa & Dietz, 2015; Klug et al., 2020). Hu et al. (2023) assembled these into a modular, open-source artificial bilateral CI user, whose front end we adopt. They also note that these models rarely yields an accurate simulation of localization patterns. We therefore replace the hand-designed decision stage with a task-optimized decoder.

3 METHODS

Hybrid-CINN has two parts (Fig. 1). A fixed biophysical front end maps a binaural sound $\boldsymbol{S}$ to the auditory-nerve response (ANR) of a bilateral CI user, $\mathbf{R} = f(\boldsymbol{S}; \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ collects the device and peripheral parameters. A learned decoder with weights $\boldsymbol{\omega}$ maps the ANR to a source direction, $g(\mathbf{R}; \boldsymbol{\omega})$. We first train $\boldsymbol{\omega}$ with $\boldsymbol{\xi}$ at its default values, which gives the ideal CI listener (§3.3). We then change one component of $\boldsymbol{\xi}$, or lesion g itself, and retrain $\boldsymbol{\omega}$ under each condition (§3.4). Parameter settings, dataset composition and analysis details are in Appendix A.4.

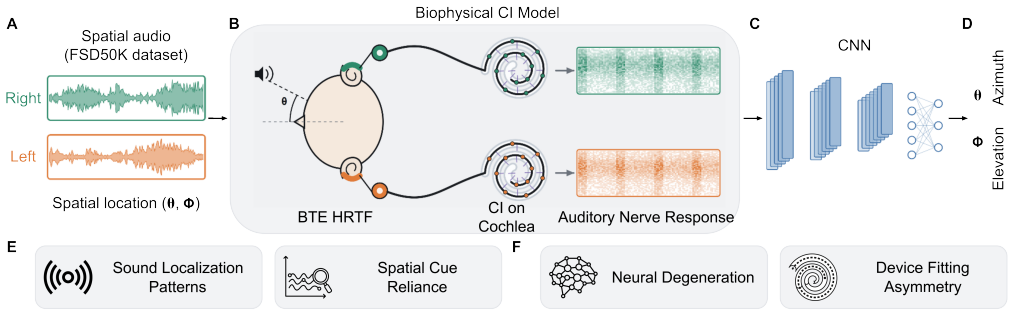


Figure 1: **A biophysically grounded framework for modeling spatial hearing with bilateral cochlear implants. (A)** Natural sounds are rendered binaurally at azimuth θ and elevation ϕ using behind-the-ear (BTE) head-related transfer functions (HRTFs). **(B)** A biophysical model of CI processing converts the binaural signals into an auditory-nerve response (ANR). **(C)** A convolutional neural network (CNN) is trained to localize the source from the ANR. **(D)** The network jointly classifies azimuth θ (72 classes at $5°$ spacing) and elevation ϕ (3 classes: $-25°$, $0°$, $+25°$). **(E)** Model behavior is compared with CI listeners' localization error patterns and spatial cue reliance. **(F)** The effects of (i) neural degeneration (peripheral and central) and (ii) device fitting (insertion-depth mismatch and loudness imbalance) are quantified.

3.1 FRONT END: FROM SOUND TO AUDITORY-NERVE RESPONSE

Stimuli. Natural sounds from FSD50K (Fonseca et al., 2022) (Appendix A.2) were convolved with binaural room impulse responses computed with the image-source method (Saddler & McDermott, 2024) and the BTE HRTFs of five listeners from the ARI database (Majdak et al., 2007). Sources were placed at 216 directions (72 azimuths in $5°$ steps $\times$ 3 elevations, which add variability to the localization task) in twelve rooms ranging from anechoic to fully reverberant. Listeners and sources are static and there is no background noise, matching the laboratory experiments we compare against (room and sound split details in Appendix A.3).

CI and auditory-nerve simulation. The binaural signal then passes through our Python port of the bilateral CI model of Hu et al. (2023), which runs alongside PyTorch. Each of the 12 electrodes per ear produces a monopolar, exponentially decaying potential that stimulates 980 auditory nerve fibers spaced uniformly along 35 mm of cochlea, and each fiber responds to each pulse stochastically, depending on refractoriness and membrane noise (Fredelake & Hohmann, 2012). Spike times are binned at 1 ms and binarized. The resulting ANR, $\mathbf{R} \in \{0, 1\}^{2\times980\times1000}$ (ear $\times$ fiber $\times$ time), is the decoder's sole input. By default, the two ears have independent AGCs and identical frequency maps and insertion depths (Appendix A.4).

3.2 NEURAL NETWORK AND TRAINING

We pose localization as classification over the source directions, $y \in \{1, \ldots, 216\}$, and take the argmax of the logits $g(\mathbf{R}; \boldsymbol{\omega})$ as the response. The decoder follows Banerjee et al. (2025) and Saddler & McDermott (2024) and treats the ANR as a two-channel fiber $\times$ time image. A stride-4 5×5 convolution combines the two ears from the first layer, three depthwise-separable blocks reduce the input to $64 \times 15 \times 15$, and global average pooling and one 128-unit hidden layer map to the

Table 1: Device and neural factors, each trained with 12 seeds per condition (Appendix A.8).

Factor	Locus	Conditions
Insertion mismatch	Device	$\Delta \in \{0, \pm 1.5, \pm 3, \pm 6\}$ mm, right ear
AGC linking	Device	independent vs. linked
Peripheral extent	Neural	14 levels, 0–97.5% of fibers per ear
Peripheral asymmetry	Neural	50/50, 30/70, 10/90 (50% loss across both ears)
Peripheral place	Neural	basal vs. apical half, both ears
Central extent	Neural	13 levels, 0–95% of units per layer

logits (42,836 parameters; Appendix A.6). We explored other models, a CNN with nine times the parameters, structured state-space (S4) models and low-rank recurrent networks did not outperform it (Appendix A.6). Models were trained with Adam on label-smoothed cross-entropy, with warm-up, cosine decay and early stopping, for at most 60 epochs (Appendix A.6); each seed is an independent training run. As a mechanistic reference, a parameter-free excitation–inhibition read-out (Hu et al., 2023) was applied to the same ANRs (Appendix A.6).

3.3 Ideal CI listener: evaluation against human data

Generalization splits. For each (room, HRTF, direction) cell we drew 75 training, 10 validation and 10 test sounds randomly (Appendix A.5). We used five-fold cross-validation over listeners, each fold holding out one HRTF and training on the other four, with six seeds per fold. The two rooms in which the human data were collected are held out of every fold(Appendix Table S2). Test performance is reported on unseen sounds, rooms and listeners.

Metric and statistics. The main metric is folded azimuth RMSE: targets and responses are folded into the frontal hemifield before computing the error, matching the frontal loudspeaker arrays of the human studies. We report medians and interquartile ranges across units (model seeds or listeners), and use two-sided Mann–Whitney U tests for independent groups and Wilcoxon signed-rank tests for paired contrasts (Appendix A.7). Seeds share folds and stimuli, so model-side p-values should be read as descriptive.

Comparisons with CI listeners. We ran three comparisons against published bilateral CI data. *Response function* (Fig. 2A–C): models were scored at the 19 loudspeaker positions (-90° to $+90^\circ$) used for the 48 listeners of Anderson et al. (2022), using FSD50K sounds. *Cue weighting* (Fig. 2D–E): on a shared probe bank of low-band (125–750 Hz) and high-band (0.75–8.5 kHz) noise near the midline ($\pm 20^\circ$), we imposed an ITD offset (up to $\pm 600\,\mu$s) or an ILD offset (up to ± 20 dB) and measured the fraction of the imposed cue carried by the model's response (Appendix A.7). *Spectral dependency* (Fig. 2F): following Dorman et al. (2014), we presented wideband, high-pass and low-pass noise from 11 loudspeakers ($\pm 75^\circ$) and expressed each unit's RMSE relative to its own high-pass RMSE, for comparison with the 11 listeners of Loiselle et al. (2016) (Appendix A.7).

3.4 Device and neural factors

Every factor is chronic: it is present in both training and testing, and a separate model is trained from scratch for each condition (Table 1). Because this requires hundreds of models, these experiments use a 10% subset of the corpus, which was sufficient to show the effect without overfitting (Appendix A.5), so absolute RMSEs are not comparable with §3.3 and we interpret only contrasts within each experiment.

Insertion mismatch. All right-ear electrode positions are shifted by Δ mm ($\Delta > 0$ deeper) while the left ear stays in place, so $|\Delta|$ is an interaural place–frequency mismatch. *AGC linking.* Both arms use the fast-acting, more compressive AGC setting of Potts et al. (2019). In the independent arm each ear applies its own gain; in the linked arm both apply the larger attenuation, $\min(G_L, G_R)$, which preserves the interaural level ratio. *Peripheral loss.* A random subset of fibers is silenced in each ear. We vary its extent (dose d, the fraction of fibers silenced per ear), its split between the ears at a fixed mean of 50%, and its place (the basal or apical half of the fibers in both ears). *Central loss.* A random fraction d of units in every decoder layer is silenced, with inverted-dropout

scaling $1/(1-d)$ so that $d = 0$ is the intact decoder. Dose–response curves are summarized with continuous piecewise-linear fits, with seed-level bootstrap confidence intervals on the breakpoints (Appendix A.8).

4 RESULTS

We first test whether a decoder trained only to localize sounds from simulated CI nerve activity behaves like a bilateral CI listener (§4.1, Fig. 2). We then degrade the device (§4.2, Fig. 3) and the neural pipeline (§4.3, Fig. 4) one factor at a time. Finally, we test whether the costs of combined factors add (§A.1, Fig. S1).

4.1 THE IDEAL CI LISTENER MODEL EXHIBITS CI LOCALIZATION PATTERNS DESPITE A SUPERIOR PERFORMANCE.

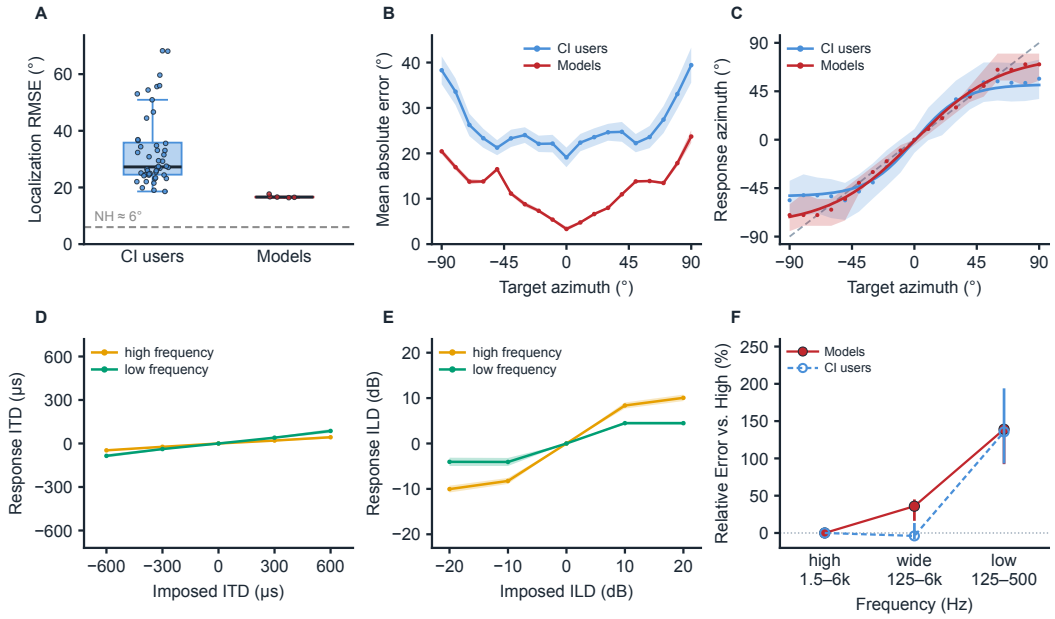


Figure 2: **The ideal CI listener outperforms every CI user, still qualitatively reproduces the signatures of CI localization.** Red, models; blue, bilateral CI users. Unless noted, one model unit is one held-out HRTF fold (the mean of six training seeds; $n = 5$). **(A)** Localization error (RMSE); boxes, median and IQR; dashed line, normal hearing. **(B)** In both CI and models, error increases laterally; shading, ± 1 SEM across listeners and models. **(C)** Both compress lateral responses along a sigmoid, the model less so; shading, interquartile range of responses; dashed diagonal, veridical. **(D, E)** An imposed ILD (E) shifts model response far more than an imposed ITD (D), similar to in CI users; shading, ± 1 SD across folds. **(F)** Both localize low-pass noise worse than high-pass; error relative to each unit's own high-pass error, median and IQR.

The model is trained only on simulated auditory-nerve responses and never sees human responses, so every agreement with CI users below is a prediction from signal rather than a behavioral fit.

Accuracy relative to CI users and normal hearing. Across five HRTF folds (six seeds each), the ideal listener localized with a median RMSE of 16.6° (IQR 16.5–16.7°; Fig. 2A), versus 27.3° (24.5–35.8°) for the 48 bilateral CI users of Anderson et al. (2022). Even the worst fold (17.7°) beat the best listener (18.6°), and the model's spread was far narrower (SD 0.5° vs. 13.0°). This is expected: the folds differ only in HRTF, and the model lacks listener noise sources such as attention, decision noise and motor execution. It thus approximates what the simulated CI signal can support under idealized conditions, not typical performance. Even so, its error is close to three times that of

normal-hearing listeners ($\approx 6°$; Dorman et al. 2016), so a substantial deficit is already present in the CI signal itself.

Lateral error pattern. Error increased laterally in both populations, with the model more accurate at every azimuth (Fig. 2B). Both response functions follow a sigmoid that flattens before $\pm 90°$, placing lateral targets closer to the midline, a medial bias characteristic of CI users (Anderson et al., 2022). Four-parameter logistic fits to each unit (Anderson et al., 2022) show milder compression in the model: a wider response range ($154°$ [152, 167] vs. $119°$ [101, 128]; $p = 4.9 \times 10^{-6}$) and a midline gain closer to veridical (1.30 [1.24, 1.33] vs. 1.67 [1.37, 2.15] deg/deg; $p = 0.022$). The model thus shares the sigmoid shape of the CI response function but lies closer to the identity line.

Reliance on ILD. To measure cue reliance, we imposed an ITD or an ILD on probe noise near the midline and measured how much of the imposed cue the model's responses carried (Fig. 2D, E). In the high band (0.75–8.5 kHz), responses shifted by $57 \pm 2\%$ of an imposed ILD but only $7 \pm 1\%$ of an imposed ITD (transfer slopes, mean $\pm$ SD across folds). In the low band (125–750 Hz), both cues transferred weakly, though ILD still exceeded ITD ($26 \pm 2\%$ vs. $14 \pm 2\%$). Bilateral CI users show a similar ILD dominance (Dorman et al., 2014; Loiselle et al., 2016). In the model, this is consistent with its input: the CI front end discards temporal fine structure, so little usable ITD reaches the decoder (§5).

Poor localization of low-frequency sounds. A listener that reads direction from ILDs should struggle with low-frequency sounds, which carry small ILDs. Following Dorman et al. (2014), we compared wideband, high-pass and low-pass noise, measuring each unit's error relative to its own high-pass error (Fig. 2F). Low-pass noise raised error by 139% [92, 145] in models and by 136% [93, 194] in 11 bilateral CI users (Loiselle et al., 2016). Both increases were significant (models: one-sample t-test on the five fold means, $p = 2.4 \times 10^{-3}$, with an increase in every fold, 61–146%; listeners: Wilcoxon signed-rank, $p = 9.8 \times 10^{-4}$) and did not differ between populations ($p = 0.83$). The profiles diverged only for wideband noise, which raised model error by 36% [16, 45] but left listener error unchanged (-4% [-9, 13]; $p = 0.052$). Adding low-frequency energy to a high-frequency sound thus tends to cost the model accuracy, but not CI users.

The ideal model therefore retains four signatures of CI spatial hearing: greater lateral error, lateral response compression, ILD dominance and poor low-frequency localization. We use this ideal listener framework as the reference for measuring the costs of device and neural factors in the next section.

4.2 Interaural device mismatches degrade localization

The model reads direction mainly from ILDs, so device settings that disturb the comparison between the two ears should cost accuracy. We tested two: mismatched insertion depths and unlinked AGCs.

Interaural electrode-place insertion mismatch. We shifted the right-ear array basally (shallow) or apically (deep) by up to 6 mm (Fig. 3A). The left ear stayed in place and filterbank frequencies were unchanged. We trained 12 models (seeds) at each shift. Error rose monotonically with the size of the mismatch, from $17.78°$ without a shift to $19.33°$ at 6 mm (Fig. 3B; Spearman $\rho = 0.97$ across the seven conditions, $p = 2.3 \times 10^{-4}$). This corresponds to a slope of $0.28°$/mm (95% CI 0.20–0.37; crossed bootstrap over seeds and test sounds). The direction of the shift did not matter (direction $\times$ mismatch, $p = 0.93$). The cost therefore comes from the two ears disagreeing about place, not from a shift of the absolute place–frequency map. This is consistent with the observation that interaural place mismatch impairs binaural hearing in CI users (Goupell et al., 2022; Litovsky et al., 2009).

AGC linking. Compression in the AGC grows with input level. Independent AGCs therefore attenuate the louder ear more than the quieter one, which shrinks the ILD. Linking them, so that both ears receive the smaller of the two gains, preserves the interaural level ratio (Fig. 3C). Linking AGC is a factor that enhances localization, rather than degradation.

We used the fast-acting setting of Potts et al. (2019). With this setting, linked AGCs lowered median error from $16.58°$ to $16.16°$ (Fig. 3D). The seed-paired difference was $0.39°$ (95 % CI 0.06–$0.77°$; crossed bootstrap over seeds and test sounds). The benefit is small in absolute terms but it goes in

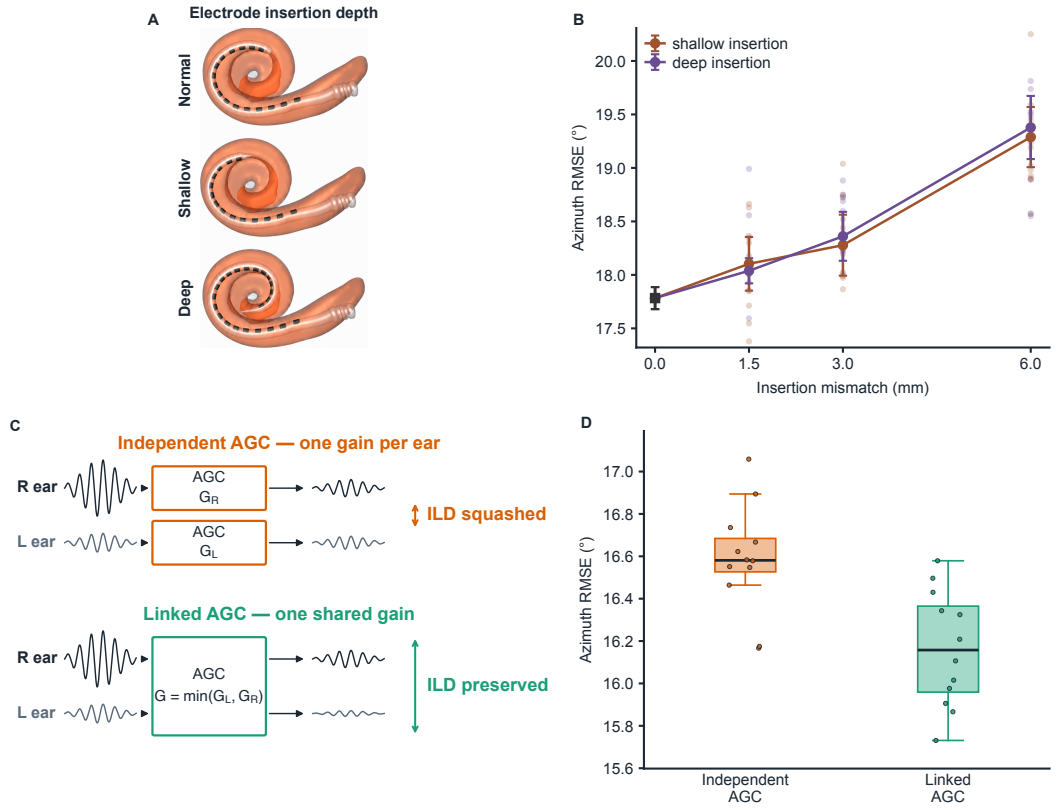


Figure 3: **Interaural device mismatches degrade localization.** **(A)** Shifting one electrode array shallower or deeper than the other creates an interaural place mismatch. **(B)** Error grows with the size of the mismatch, whichever way the array is shifted; black square, no shift. **(C)** Independent AGCs compress the louder ear more and shrink the ILD; linked AGCs apply one gain to both ears and preserve it. **(D)** Linking the AGCs lowers error; dots, individual seeds.

the same direction as the improvement that Potts et al. (2019) reported for bilateral CI users. This AGC setting differs from the default, so absolute errors here are not comparable with Fig. 3B.

4.3 Central loss costs more than peripheral loss, and where peripheral loss falls matters

We next removed neural substrate at two loci: auditory-nerve fibers (peripheral) and decoder units (central). Lesioned models were retrained to adapt to the damage, as in chronic degeneration.

Extent of degeneration. Silencing half of the auditory-nerve fibers roughly halves the spikes reaching the decoder (Fig. 4A), yet localization barely changed. Peripheral error stayed near its intact level up to about two-thirds loss (Fig. 4B, green), then rose gradually and collapsed only when almost the entire nerve was gone. A three-segment fit placed the breakpoints at $\tau_1 = 65\%$ (poorly constrained; bootstrap 95% CI 38.0–85.5%) and $\tau_2 = 94\%$ (89.0–95.5%). Beyond τ_2, error rose by $0.452°$ per percentage point, about ten times faster than between the breakpoints. Central degeneration was more damaging (Fig. 4B, purple): its error exceeded peripheral error at every shared non-zero fraction, and beyond a single breakpoint at 88% (CI 82.7–89.1%) it rose by $1.249°$ per point, about three times as steeply as the peripheral collapse. For localization, the nerve thus carries substantial redundancy, whereas the decoder has less spare capacity. Because the two sweeps remove different kinds of units, they are compared at matched fractions rather than matched information. These lesions were applied under otherwise ideal conditions; Supplementary A.1 tests how they combine with other factors.

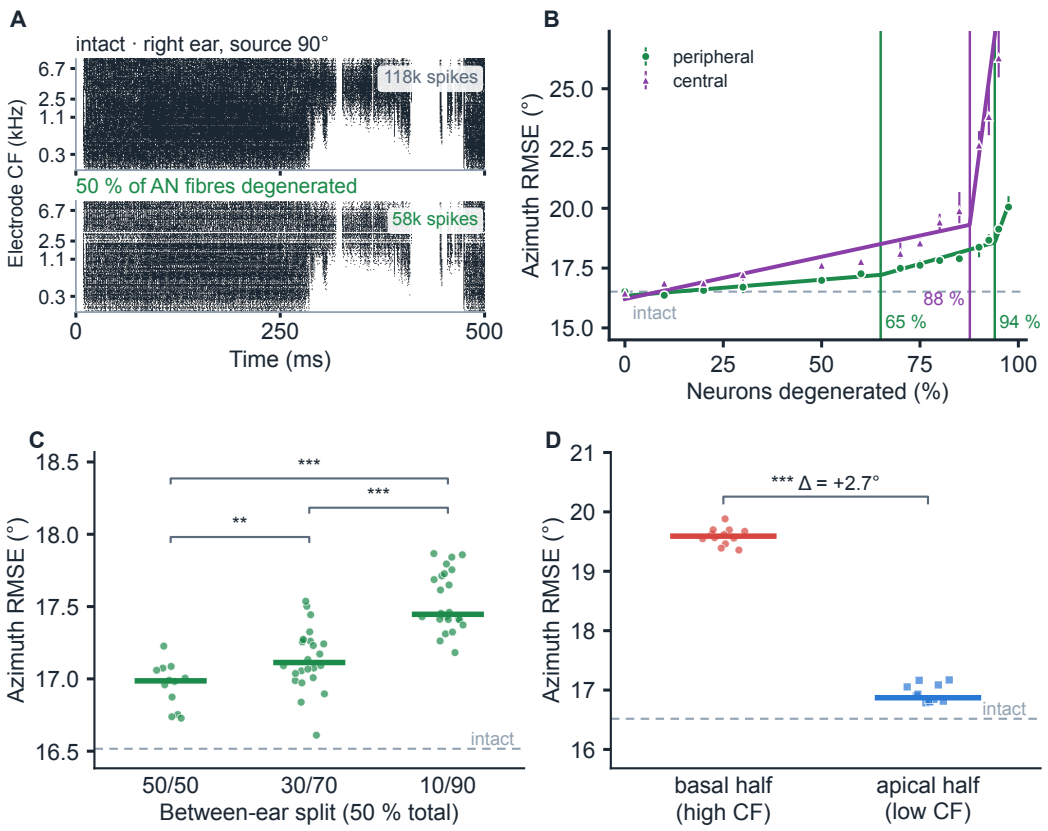


Figure 4: **Central degeneration costs more than peripheral degeneration, and asymmetric or basal peripheral loss is the most costly.** All lesions are present during both training and testing. **(A)** Auditory-nerve spikes, intact (top) and with half of the fibers silenced (bottom). **(B)** Error against the fraction of neurons lost. Peripheral loss (green) is tolerated until severe; central loss (purple) costs more and collapses more steeply. Lines, piecewise-linear fits; vertical lines, breakpoints. **(C)** At the same total loss, the more unequal the two ears, the larger the error. **(D)** At the same number of surviving fibers, losing the basal half costs far more than losing the apical half. Dots, individual seeds; dashed lines, intact model. $^{**}\,p < 0.01$, $^{***}\,p < 0.001$.

Interaural asymmetry of fiber loss. Because localization compares the two ears and fiber loss are usually asymmetric, we asked whether it matters how a given loss is distributed between them. Holding the mean loss at 50% of fibers per ear, we split it evenly (50/50) or unevenly (30/70, 10/90) (Fig. 4C). Error rose with asymmetry, from 17.0° (50/50) to 17.1° (30/70) and 17.5° (10/90), with all pairwise differences significant (Mann–Whitney: 50/50 vs. 30/70 $p = 6.9 \times 10^{-3}$; 30/70 vs. 10/90 $p = 1.2 \times 10^{-6}$; 50/50 vs. 10/90 $p = 1.7 \times 10^{-6}$). The same total loss is thus more costly when it falls unevenly on the two ears.

Basal versus apical fiber loss. Because the cochlea is tonotopic, silencing one end of the nerve removes a frequency region rather than a random sample of fibers. We therefore silenced either the basal (high-CF) or apical (low-CF) half of the fibers in both ears (Fig. 4D), leaving 490 fibers per ear, as in the 50/50 condition. Basal loss raised error to $19.59 \pm 0.14°$ (mean ± SD), whereas apical loss left it at $16.93 \pm 0.15°$ ($\Delta = +2.66°$; Mann–Whitney $p = 3.7 \times 10^{-5}$). Apical loss was thus no worse than random loss of the same number of fibers (median 17.0°, Fig. 4C), while basal loss cost 2.6° more: which fibers are lost matters more than how many. The basal half carries the high-frequency ILDs on which the model relies (Fig. 2D, E), making its loss the neural counterpart of the low-pass penalty in Fig. 2F.

In summary, the device factors cost at most $1.5°$ at the settings tested, peripheral loss is well tolerated unless severe, and central loss is the most damaging.

5 Discussions and Conclusion

The implant signal itself limits localization. Trained only on simulated auditory-nerve activity, without listener-specific factors or cognitive noise, the model localized more accurately than every CI user, yet its error remained close to three times that of normal-hearing listeners (Fig. 2A). None of the alternative decoders we tested did better (Appendix A.6). Within our simulation, much of the deficit is thus already present in the nerve signal, regardless of decoding; how well this transfers to real implants depends on the simulated front end.

CI-like behavior emerges without fitting. Although never fit to human responses, the model reproduced three signatures of bilateral CI localization (§4.1): compression of lateral responses toward the midline, reduced accuracy for low-frequency sounds, and reliance on ILDs over ITDs. None was imposed on the decoder; each emerged from optimizing localization on the simulated CI signal, suggesting that these signatures arise from the implant signal rather than from listener-specific factors.

High-frequency cues and interaural symmetry are key. Across the device and neural experiments, two input properties determined accuracy. The first is high-frequency level information: the model relies on interaural level differences, which are largest at high frequencies, so every manipulation that degraded this information increased error, whether removing high frequencies from the stimulus (Fig. 2F), silencing the basal fibers that encode them (Fig. 4D), or applying independent gain control in each ear, which compresses level differences (Fig. 3D). The second is interaural symmetry: error increased whenever the two ears differed. For insertion depth, only the size of the mismatch mattered, not its direction (Fig. 3B); for nerve survival, uneven loss was costlier than even loss of the same total (Fig. 4C). These results suggest two clinical priorities: fitting the two ears to match each other rather than each to a standard frequency map, and assessing nerve health separately for each ear and cochlear region.

Combined factors. Patients carry several impairments at once. When insertion mismatch and peripheral and central loss were combined, their joint cost was indistinguishable from the sum of the individual costs, except when all three were at their most severe (§A.1). Single-factor costs may therefore approximate combined effects, although moderate interactions cannot be ruled out with six seeds per condition.

Limitations and opportunities. The main limitation is that we do not model the listener factors known to predict bilateral CI outcomes: age at onset and duration of deafness, age at implantation, prior acoustic experience, and the interval between implantations; which are predictors of CI performance (Anderson et al., 2022). The model therefore approximates localization under idealized conditions, not the spread of outcomes. Comparisons with CI users are qualitative, because stimuli, rooms and response methods differ and the model lacks error sources present in every listener (decision noise, lapses, memory, response execution). The front end is representative rather than exhaustive: it uses CIS, 12 electrodes, symmetric fitting and 1-ms spike binning, and models degeneration as silencing. However, each choice is a parameter of the modular pipeline and can be varied. We assess azimuth only, as behavioral data for comparing elevation are lacking, and the effects of the suggested fitting changes on speech perception remain open. Extending the framework to other spatial tasks and patient-specific conditions will open a broader range of clinical and engineering questions.

Conclusion. We developed Hybrid-CINN, an end-to-end, biophysically grounded model of spatial hearing with bilateral cochlear implants. Natural sounds are rendered with behind-the-ear HRTFs and passed through models of CI processing, the electrode–nerve interface and auditory-nerve responses, followed by a learned localization model. Without being fit to human data, the model captures key characteristics of CI localization and showed that high-frequency level cues and interaural symmetry are the main determinants of accuracy. By systematically varying device-related and neural factors, we isolated effects of degradation factors, a question difficult to study in patients. The framework thus provides a controlled test bed for tracing the sources of spatial-hearing deficits in CI users and for evaluating device and fitting strategies before clinical testing.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, design or provide feedback on research methodology and experiments, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools for research ideation, to propose or refine hypotheses, or to help develop theoretical models or conceptual frameworks, and the remaining required-disclosure tasks (generating synthetic datasets, formulating mathematical claims or assisting with proofs, translation, and cleaning or reformatting datasets) are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, suggest experimental parameters, create or modify scientific figures and images, create artifacts, identify relevant literature, summarize and analyze existing literature, draft parts of the paper, edit the paper to improve readability, and format the manuscript and references in LaTeX.

We have reviewed all AI-assisted work. Methods were implemented with Claude Code, an agentic coding framework operating under author supervision, including the conversion of our original MATLAB modules to Python; all code changes were reviewed by the authors, and the Python ports were validated against the MATLAB originals on identical inputs. Claude Code was also used to optimize experiment execution (job batching and scheduling on our compute cluster) and to write plotting scripts, whose outputs we checked against the underlying data. AI feedback on methodology and experimental design, as well as AI-suggested experimental parameters, was treated as advisory, and all design and parameter choices were made by the authors. AI-assisted analysis and interpretation served only as a starting point; the authors independently examined all results, and the conclusions reported here are our own. For the literature review, every AI-identified paper was read by the authors, and each claim attributed to prior work was verified against the original source before inclusion. AI writing tools were used for initial drafts of parts of the manuscript and for paraphrasing, and all such text was revised by the authors. AI image generation was limited to illustrative figures and was not used for any figure reporting results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

An anonymized repository accompanying this submission (`https://anonymous.4open.science/r/hybrid_cinn-8EBC`) contains a minimal, self-contained implementation of the full Hybrid-CINN pipeline: binaural spatialization of natural sounds with simulated room impulse responses and behind-the-ear HRTFs, our Python port of the bilateral CI and auditory-nerve model of (Hu et al., 2023) that generates the auditory-nerve response, the compact CNN decoder that classifies azimuth and elevation, and the folded-azimuth RMSE evaluation. Every manipulation in §3.4 is exposed as a parameter of this code: interaural insertion mismatch and AGC linking at the device level, and peripheral fiber silencing (extent, interaural asymmetry and place) and central unit silencing at the neural level. The model is described in §3. All front-end settings (pre-emphasis, AGC, filterbank, CIS pulse rate, loudness mapping and electrode–nerve interface) are listed in Appendix A.4. The decoder architecture, optimizer, learning-rate schedule and stopping criteria are in Appendix A.6. All data are publicly available. The FSD50K clip selection and sound splits are in Appendix A.2. The ARI BTE HRTFs, the room simulation and the geometry and surface materials of all twelve rooms are in Appendix A.3 and Table S1. The sampling procedure and per-fold dataset composition are in Appendix A.5 and Table S2. Human comparison data are taken from published studies (Anderson et al., 2022). The condition grids, training seeds, lesion-masking procedure, piecewise-linear fits and bootstrap procedures for the factor experiments are given in Table 1 and Appendix A.8. Compute requirements are reported in Appendices A.3 and A.6.

REFERENCES

Acoustics Research Institute, Austrian Academy of Sciences. ARI HRTF database: Behind-the-ear HRTFs. https://www.oeaw.ac.at/isf/outreach/software/hrtf-database. SOFA files (folder "ari (bte)"): https://sofacoustics.org/data/database/. Accessed: 2026-09-16.

Sean R. Anderson, Rachael Jocewicz, Alan Kan, Jun Zhu, ShengLi Tzeng, and Ruth Y. Litovsky. Sound source localization patterns and bilateral cochlear implants: Age at onset of deafness effects. *PLOS ONE*, 17(2):e0263516, February 2022. doi: 10.1371/journal.pone.0263516. URL https://dx.plos.org/10.1371/journal.pone.0263516.

Sean R. Anderson, Emily Burg, Lukas Suveg, and Ruth Y. Litovsky. Review of Binaural Processing With Asymmetrical Hearing Outcomes in Patients With Bilateral Cochlear Implants. *Trends in Hearing*, 28:23312165241229880, January 2024. doi: 10.1177/23312165241229880. URL https://journals.sagepub.com/doi/10.1177/23312165241229880.

Alan W. Archer-Boyd and Robert P. Carlyon. Simulations of the effect of unlinked cochlear-implant automatic gain control and head movement on interaural level differences. *The Journal of the Acoustical Society of America*, 145(3):1389–1400, 2019. doi: 10.1121/1.5093623. URL https://doi.org/10.1121/1.5093623.

Annesya Banerjee, Mark R. Saddler, Julie G. Arenberg, and Josh H. McDermott. A deep learning framework for understanding cochlear implants. bioRxiv preprint, July 2025. URL http://biorxiv.org/lookup/doi/10.1101/2025.07.16.665227.

Roberto Barumerli, Piotr Majdak, Michele Geronazzo, David Meijer, Federico Avanzini, and Robert Baumgartner. A Bayesian model for human directional localization of broadband static sound sources. *Acta Acustica*, 7:12, 2023. doi: 10.1051/aacus/2023006. URL https://acta-acustica.edpsciences.org/10.1051/aacus/2023006.

Robert P. Carlyon, John M. Deeks, Bertrand Delgutte, Yoojin Chung, Maike Vollmer, Frank W. Ohl, Andrej Kral, Jochen Tillein, Ruth Y. Litovsky, Jan Schnupp, Nicole Rosskothen-Kuhl, and Raymond L. Goldsworthy. Limitations on Temporal Processing by Cochlear Implant Users: A Compilation of Viewpoints. *Trends in Hearing*, 29:23312165251317006, March 2025. doi: 10.1177/23312165251317006. URL https://journals.sagepub.com/doi/10.1177/23312165251317006.

Michael F. Dorman, Louise Loiselle, Josh Stohl, William A. Yost, Anthony Spahr, Chris Brown, and Sarah Cook. Interaural Level Differences and Sound Source Localization for Bilateral Cochlear Implant Patients. *Ear and Hearing*, 35(6):633–640, November 2014. doi: 10.1097/AUD.0000000000000057. URL https://journals.lww.com/00003446-201411000-00006.

Michael F. Dorman, Louise H. Loiselle, Sarah J. Cook, William A. Yost, and René H. Gifford. Sound Source Localization by Normal-Hearing Listeners, Hearing-Impaired Listeners and Cochlear Implant Listeners. *Audiology and Neurotology*, 21(3):127–131, April 2016. doi: 10.1159/000444740. URL https://doi.org/10.1159/000444740.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022. doi: 10.1109/TASLP.2021.3133208. URL https://doi.org/10.1109/TASLP.2021.3133208.

Andrew Francl and Josh H. McDermott. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, 6(1):111–133, January 2022. doi: 10.1038/s41562-021-01244-z. URL https://www.nature.com/articles/s41562-021-01244-z.

Stefan Fredelake and Volker Hohmann. Factors affecting predicted speech intelligibility with cochlear implants in an auditory model for electrical stimulation. *Hearing Research*, 287(1-2):76–90, May 2012. doi: 10.1016/j.heares.2012.03.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0378595512000639.

Tobias Goehring, Federico Bolner, Jessica J. M. Monaghan, Bas van Dijk, Andrzej Zarowski, and Stefan Bleeck. Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hearing Research*, 344:183–194, February 2017. doi: 10.1016/j.heares.2016.11.012. URL https://www.sciencedirect.com/science/article/pii/S0378595516304348.

Matthew J. Goupell, Jack H. Noble, Sandeep A. Phatak, Elizabeth Kolberg, Miranda Cleary, Olga A. Stakhovskaya, Kenneth K. Jensen, Michael Hoa, Hung Jeffrey Kim, and Joshua G. W. Bernstein. Computed-Tomography Estimates of Interaural Mismatch in Insertion Depth and Scalar Location in Bilateral Cochlear-Implant Users. *Otology & Neurotology*, 43(6):666–675, July 2022. doi: 10.1097/MAO.0000000000003538. URL https://journals.lww.com/10.1097/MAO.0000000000003538.

D. Wesley Grantham, Daniel H. Ashmead, Todd A. Ricketts, Robert F. Labadie, and David S. Haynes. Horizontal-Plane Localization of Noise and Speech Signals by Postlingually Deafened Adults Fitted With Bilateral Cochlear Implants. *Ear and Hearing*, 28(4):524–541, August 2007. doi: 10.1097/AUD.0b013e31806dc21a. URL https://journals.lww.com/ear-hearing/pages/articleviewer.aspx?year=2007&issue=08000&article=00009&type=Fulltext.

Benedikt Grothe, Michael Pecka, and David McAlpine. Mechanisms of Sound Localization in Mammals. *Physiological Reviews*, 90(3):983–1012, July 2010. doi: 10.1152/physrev.00026.2009. URL https://www.physiology.org/doi/10.1152/physrev.00026.2009.

Kenneth E. Hancock, Yoojin Chung, and Bertrand Delgutte. Congenital and Prolonged Adult-Onset Deafness Cause Distinct Degradations in Neural ITD Coding with Bilateral Cochlear Implants. *Journal of the Association for Research in Otolaryngology*, 14(3):393–411, June 2013. doi: 10.1007/s10162-013-0380-5. URL http://link.springer.com/10.1007/s10162-013-0380-5.

Hongmei Hu, Sebastián A. Ausili, Ben Williges, Jonas Klug, Rebecca C. Felsheim, Deborah Vickers, and Mathias Dietz. A model framework for simulating spatial hearing of bilateral cochlear implant users. *Acta Acustica*, 7:42, 2023. doi: 10.1051/aacus/2023036. URL https://acta-acustica.edpsciences.org/10.1051/aacus/2023036.

Alan Kan, Corey Stoelb, Ruth Y. Litovsky, and Matthew J. Goupell. Effect of mismatched place-of-stimulation on binaural fusion and lateralization in bilateral cochlear-implant users. *The Journal of the Acoustical Society of America*, 134(4):2923–2936, October 2013. doi: 10.1121/1.4820889. URL https://pubs.aip.org/jasa/article/134/4/2923/897601/Effect-of-mismatched-place-of-stimulation-on.

Alexander J.E. Kell, Daniel L.K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, May 2018. doi: 10.1016/j.neuron.2018.03.044. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627318302502.

Daryl Kelvasa and Mathias Dietz. Auditory Model-Based Sound Direction Estimation With Bilateral Cochlear Implants. *Trends in Hearing*, 19:2331216515616378, December 2015. doi: 10.1177/2331216515616378. URL https://journals.sagepub.com/doi/10.1177/2331216515616378.

Stefan Kerber and Bernhard U. Seeber. Sound Localization in Noise by Normal-Hearing Listeners and Cochlear Implant Users. *Ear and Hearing*, 33(4):445–457, July 2012. doi: 10.1097/aud.0b013e318257607b. URL https://journals.lww.com/00003446-201207000-00001.

Jonas Klug, Lisa Schmors, Go Ashida, and Mathias Dietz. Neural rate difference model can account for lateralization of high-frequency stimuli. *The Journal of the Acoustical Society of America*, 148(2):678–691, August 2020. doi: 10.1121/10.0001602. URL https://pubs.aip.org/jasa/article/148/2/678/958155/Neural-rate-difference-model-can-account-for.

Andrej Kral and Anu Sharma. Developmental neuroplasticity after cochlear implantation. *Trends in Neurosciences*, 35(2):111–122, 2012. doi: 10.1016/j.tins.2011.09.004. URL https://doi.org/10.1016/j.tins.2011.09.004.

Bernhard Laback, Katharina Egger, and Piotr Majdak. Perception and coding of interaural time differences with bilateral cochlear implants. *Hearing Research*, 322:138–150, April 2015. doi: 10.1016/j.heares.2014.10.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0378595514001683.

Johan Laneau, Jan Wouters, and Marc Moonen. Improved Music Perception with Explicit Pitch Coding in Cochlear Implants. *Audiology and Neurotology*, 11(1):38–52, 2006. doi: 10.1159/000088853. URL https://karger.com/article/doi/10.1159/000088853.

Ruth Y. Litovsky, Aaron Parkinson, and Jennifer Arcaroli. Spatial Hearing and Speech Intelligibility in Bilateral Cochlear Implant Users. *Ear and Hearing*, 30(4):419–431, August 2009. doi: 10.1097/AUD.0b013e3181a165be. URL https://journals.lww.com/00003446-200908000-00005.

Louise H. Loiselle, Michael F. Dorman, William A. Yost, Sarah J. Cook, and Rene H. Gifford. Using ILD or ITD Cues for Sound Source Localization and Speech Understanding in a Complex Listening Environment by Listeners With Bilateral and With Hearing-Preservation Cochlear Implants. *Journal of Speech, Language, and Hearing Research*, 59(4):810–818, August 2016. doi: 10.1044/2015_JSLHR-H-14-0355. URL http://pubs.asha.org/doi/10.1044/2015_JSLHR-H-14-0355.

Ewan A. Macpherson and John C. Middlebrooks. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5):2219–2236, May 2002. doi: 10.1121/1.1471898. URL https://pubs.aip.org/jasa/article/111/5/2219/551392/Listener-weighting-of-cues-for-lateral-angle-The.

Piotr Majdak, Peter Balazs, and Bernhard Laback. Multiple exponential sweep method for fast measurement of head-related transfer functions. *Journal of the Audio Engineering Society*, 55(7/8):623–637, 2007. URL https://aes2.org/publications/elibrary-page/?id=14190.

Piotr Majdak, Matthew J. Goupell, and Bernhard Laback. Two-Dimensional Localization of Virtual Sound Sources in Cochlear-Implant Listeners. *Ear and Hearing*, 32(2):198–208, March 2011. doi: 10.1097/AUD.0b013e3181f4dfe9. URL https://journals.lww.com/00003446-201103000-00005.

Joseph B. Nadol, Yi-Shyang Young, and Robert J. Glynn. Survival of Spiral Ganglion Cells in Profound Sensorineural Hearing Loss: Implications for Cochlear Implantation. *Annals of Otology, Rhinology & Laryngology*, 98(6):411–416, June 1989. doi: 10.1177/000348948909800602. URL https://journals.sagepub.com/doi/10.1177/000348948909800602.

Wendy B. Potts, Lakshmish Ramanna, Trevor Perry, and Christopher J. Long. Improving Localization and Speech Reception in Noise for Bilateral Cochlear Implant Recipients. *Trends in Hearing*, 23:2331216519831492, January 2019. doi: 10.1177/2331216519831492. URL https://journals.sagepub.com/doi/10.1177/2331216519831492.

Mark R. Saddler and Josh H. McDermott. Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing. *Nature Communications*, 15(1):10590, December 2024. doi: 10.1038/s41467-024-54700-5. URL https://www.nature.com/articles/s41467-024-54700-5.

Chiara Valzolgher, Julie Gatel, Sabrina Bouzaid, Solene Grenouillet, Michela Todeschini, Gregoire Verdelet, Romeo Salemme, Valerie Gaveau, Eric Truy, Alessandro Farnè, and Francesco Pavani. Reaching to Sounds Improves Spatial Hearing in Bilateral Cochlear Implant Users. *Ear and Hearing*, 44(1):189–198, January 2023. doi: 10.1097/aud.0000000000001267. URL https://journals.lww.com/10.1097/AUD.0000000000001267.

Paolo Vecchiotti, Ning Ma, Stefano Squartini, and Guy J. Brown. End-to-end Binaural Sound Localisation from the Raw Waveform. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 451–455, 2019. doi: 10.1109/ICASSP.2019.8683732. URL https://doi.org/10.1109/ICASSP.2019.8683732.

Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, March 2016. doi: 10.1038/nn.4244. URL https://www.nature.com/articles/nn.4244.

A Appendix

A.1 Combined factors

Patients rarely carry a single impairment, so we trained models with two or three factors at once (App. 7.1, Fig. S1) and asked whether their costs add. For each combination, we compared the increase in error over its own intact control with the sum of the increases each factor produced alone (ratio 1 = additive, > 1 = compounding; seed-bootstrap 95% CIs). At moderate doses (3 mm insertion shift, peripheral and central degeneration up to 70%), costs were close to additive: across the ten combinations the ratio ranged from 0.90 to 1.15 and every CI included 1. With all three factors at 3 mm/50%/50%, error rose by 2.19° against a summed 2.44° (ratio 0.90 [0.71, 1.16]). At high doses, costs compounded only where peripheral and central degeneration were both at 90%: combining them raised error 1.12 [0.91, 1.35] times the sum of their separate costs, and 1.21 [0.96, 1.50] times with the insertion shift added (1.68° above the sum, CI -0.31 to 3.82°). The other eight high-dose combinations stayed between 0.86 and 1.03. Insertion mismatch combined additively with either kind of degeneration (ratios 1.04 and 1.01 at 50%), so the only interaction appears when both neural factors are at their highest tested dose, where central degeneration is already past its breakpoint (Fig. 4B). Independent error sources add in squares, so a ratio of 1 on RMSE already implies slightly more than independent combination. On squared RMSE the ratios were 0.92–1.40, and only the three-factor 90% corner excluded 1 (1.40 [1.06, 1.81]). With six seeds per condition, the data therefore cannot distinguish additive costs from independent error sources below that corner.

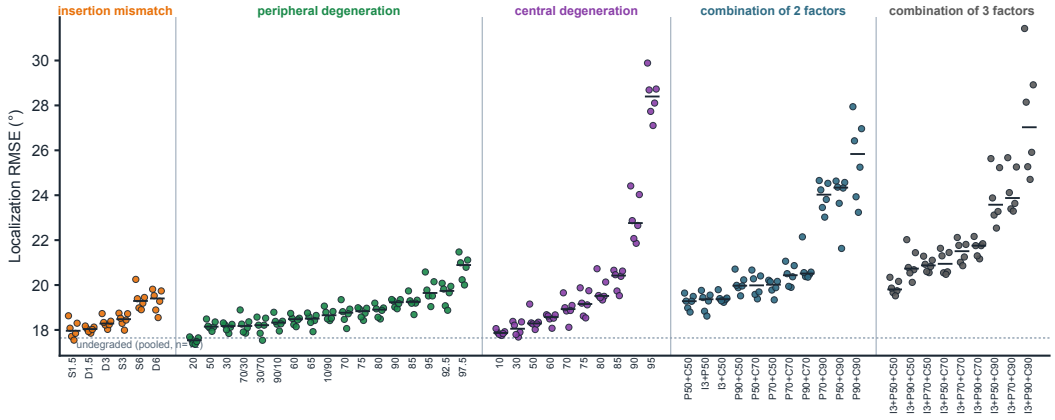


Figure S1: Localization error for all factor conditions, alone and combined. Dots, individual seeds; bars, medians. All models were trained on the reduced corpus (§3.4) and tested in the SRL cell of the b_ci8 fold, so errors are comparable within this figure but not with Fig. 2 or with CI users. Dotted lines, the intact controls of the device (upper) and neural (lower) training subsets. S/D, shallow/deep insertion shift of the right array (mm); P, peripheral fiber loss per ear (%; left/right where asymmetric); C, central unit dropout (%); I3, 3 mm shallow insertion shift.

A.2 SOUND CORPUS

Source sounds were taken from FSD50K (Fonseca et al., 2022). From 5,763 single-label clips in 143 AudioSet classes, we cropped each candidate to its highest-energy 1 s window and kept 20 clips per class, giving 2,860 clips at 44.1 kHz. Clips were split within each class 8:1:1 into 2,288 training, 286 validation and 286 test sounds, so that no sound identity appears in more than one split.

A.3 SPATIALIZATION

Rendering. Each monaural clip was convolved with a binaural room impulse response (BRIR) computed with the image-source method, which models room geometry and frequency-dependent wall absorption, using a Python port of the room simulator of Saddler & McDermott (2024). Signals were resampled to 48 kHz, given a 20 ms cosine onset/offset ramp, and set to 60 dB SPL.

Listeners (HRTFs). We used five behind-the-ear (BTE) HRTF sets from the ARI database (Acoustics Research Institute, Austrian Academy of Sciences) (`b_ci3`, `b_ci8`, `b_ci24`, `b_ci25` and `b_ci52`), measured as described by Majdak et al. (2007; 2011). Because they are recorded at the processor microphones rather than at the ear canal, they include individual head and pinna asymmetries as a CI processor receives them.

Rooms. Twelve shoebox rooms span the acoustic conditions of the bilateral-CI localization literature, from anechoic to a reverberant classroom (Table S1). Nine are modeled on specific published test environments; the other three (a small clinical booth, a moderately reverberant lab and a living room) are representative rather than replicas. Each of the six surfaces was assigned a material with published octave-band absorption coefficients, so walls, floor and ceiling can differ. A static listener stood at the horizontal centre of the room, at ear height, facing $0°$. Sources were placed at a room-specific distance d (0.85–2.0 m), set to the loudspeaker distance of the corresponding study. The two rooms in which the human behavioral data were collected (bold in Table S1) were held out of all training.

Table S1: The twelve simulated rooms. d: source distance. Surfaces are listed as walls / floor / ceiling. Rooms in bold are the rooms in which the human behavioral data were collected (Anderson et al., 2022) and (Valzolgher et al., 2023); both are held out of all training.

Room (modeled on)	$L \times W \times H$ (m)	d (m)	Surfaces
Large anechoic	$10 \times 10 \times 3$	1.40	anechoic (α=1)
IAC booth	$2.90 \times 2.74 \times 2.44$	1.20	foam, α=0.8 all surfaces
Foam-lined room	$3.05 \times 4.57 \times 3.05$	1.50	foam, α=0.8 all surfaces
Small clinical booth	$2.54 \times 2.64 \times 2.10$	1.00	α=0.6 / carpet / plaster
Clinic room	$3.50 \times 4.00 \times 2.70$	1.00	plaster / parquet / plaster
Semi-anechoic room	$5.00 \times 5.00 \times 3.00$	1.20	anechoic / parquet / anechoic
Anechoic room	$3.00 \times 3.00 \times 3.00$	0.85	anechoic (α=1)
Moderately reverberant lab	$5.00 \times 6.00 \times 2.80$	1.50	α=0.14 / parquet / plaster
Living room	$4.50 \times 5.50 \times 2.50$	1.50	plywood, drapery, glass, plaster / carpet on foam / plaster
Classroom	$8.00 \times 10.00 \times 3.00$	2.00	painted concrete / linoleum / plaster
Hospital room	$4.00 \times 4.50 \times 2.70$	0.90	plaster / linoleum / plaster
Loudspeaker array room	$5.50 \times 5.50 \times 3.00$	1.50	α=0.8 / carpet / α=0.8

Directions. Sources were placed on a sphere of room-specific radius (0.85–2.0 m) at 72 azimuths ($-180°$ to $+175°$ in $5°$ steps) and 3 elevations ($-25°$, $0°$, $+25°$), giving 216 directions. Elevation adds variability to the azimuth task and is not evaluated.

Compute. The full set comprises 12,960 BRIRs (12 rooms $\times$ 5 HRTFs $\times$ 216 directions). It took about 930 CPU-hours, generated in SLURM arrays of 4-core jobs (at most 144 cores at once; about 16 h wall-clock). Reverberant rooms dominate the cost: about 100 task-hours for the slowest room against 6 for the anechoic room.

Table S2: Per-fold dataset composition. One listener HRTF is held out per fold, and rooms 2 and 11 are held out of all folds. Training, validation and the S test set use the 10 training rooms and 4 training HRTFs, with disjoint source sounds. The SRL test set uses only the held-out rooms and the fold's held-out HRTF.

Split	Rooms	HRTFs	Samples
Train	10	4	648,000
Validation	10	4	86,400
Test, S (unseen sound)	10	4	86,400
Test, SRL (unseen sound, room, HRTF)	2	1	4,320

Limitations. Listeners and sources are static and there is no background noise, matching the laboratory experiments we compare against. BRIRs are truncated at 500 ms, which shortens the reverberant tail in the three most reverberant rooms, and the image-source method omits diffraction and scattering.

A.4 Biophysical CI and auditory-nerve simulation

Implementation. The simulation is our Python port of the bilateral CI model of Hu et al. (2023), whose auditory-nerve stage follows the stochastic model of pulsatile stimulation of (Fredelake & Hohmann, 2012). The port runs natively alongside PyTorch, which made the dataset sizes used here feasible, and is released with this paper.

Sound processor. Signals were resampled to an internal rate of 100 kHz and pre-emphasized with a first-order high-pass filter (corner frequency 1.2 kHz). A single-band AGC followed (knee point 50 dB SPL, ratio 2:1, attack 50 ms, release 200 ms). A 12-channel fourth-order gammatone filterbank with ERB-spaced centre frequencies from 298 to 6,679 Hz extracted channel envelopes, which the CIS strategy sampled into non-overlapping pulse trains at 900 pulses/s per electrode. The Nucleus loudness growth function mapped a 40 dB acoustic input range onto the clinical range between threshold and comfort level (120–170 clinical units, 153–377 µA) (Laneau et al., 2006). The output is the electrodogram: the current delivered by each electrode in each ear over time.

Electrode–nerve interface and auditory-nerve response. The nerve is represented by 980 fibers spaced uniformly along 35 mm of cochlea, from the base (0 mm, high characteristic frequency) to the apex. The 12 electrode positions $\boldsymbol{x}_{\mathrm{el}}$ are equally spaced from 5 to 26 mm, and each electrode produces a monopolar, exponentially decaying potential field with a broad spread constant ($\lambda = 9$ mm). For every pulse, all 980 fibers are evaluated in parallel. The effective current at each fiber drives a strength–duration (chronaxie) law; the resulting depolarization is compared with a threshold that is raised by the fiber's refractory state and perturbed by membrane noise; and fibers that fire do so after a latency whose mean and jitter depend on the firing probability. Spike times were binned at 1 ms over the 1 s stimulus and binarized. The result is the auditory-nerve response (ANR), $\mathbf{R} \in \{0,1\}^{2\times 980\times 1000}$ (ear × fiber × time; channel 0 = left ear), stored bit-packed in HDF5.

A.5 Datasets and generalization splits

Sampling. Rendering every sound in every (room, HRTF) cell and direction would require $2{,}860 \times 60 \times 216 \approx 37$ M ANRs. Instead, for each cell and direction we drew $k = 75$ distinct training sounds ($k = 10$ for validation and for test), chosen to give about 15 training samples per model parameter.

Evaluation folds. Each fold holds out one listener HRTF and trains on the other four; rooms 2 (`jones2014_IAC_booth`) and 11 (`valzolgher2023_lyon_hospital`) are held out of every fold. Each fold therefore trains on 10 rooms × 4 HRTFs × 216 directions × 75 sounds = 648,000 samples (15.1 per parameter; Table S2).

Factor datasets (FSD50K-p10). Because the factor experiments require hundreds of models, they use a reduced corpus built on a single fold. Listener `b_ci8` is held out as a representative listener: its ILD at 0° azimuth is the median of the five. Training and validation use the remaining 40 cells

(10 rooms $\times$ 4 HRTFs; Table S2). The training subset is built in one of two ways. For the device factors, whose ANRs are regenerated for each condition, it keeps 8 of the 75 training sounds and 1 of the 10 validation sounds per cell and direction (40 cells $\times$ 216 directions $\times$ 8 = 69,120 training and 8,640 validation samples). For the neural factors, whose lesions are applied to the stored ANRs at load time, it is a uniform random 10 % of each split (64,800 training and 8,640 validation samples). Within each experiment, every condition uses the same rows. All factor models are tested in the SRL cell of the `b_ci8` fold: 2 held-out rooms $\times$ 1 held-out HRTF $\times$ 216 directions $\times$ 10 sounds = 4,320 samples. This is the same test set as the ideal listener's `b_ci8` fold, so the intact baselines of Fig. 3 and Fig. 4 differ only through their training subsets. This single-fold design shows the direction and relative size of each factor's effect; a full five-fold evaluation on the complete corpus will be added during the discussion period.

A.6 Decoder and training

Architecture. The decoder of Hybrid-CINN is a compact depthwise-separable CNN (42,836 trainable parameters) that follows the design of Banerjee et al. (2025); Francl & McDermott (2022); Saddler & McDermott (2024) and treats the ANR as a two-channel fiber $\times$ time image. A 5×5, stride-4 convolution maps the two ears onto 12 feature maps, so binaural comparison is possible from the first layer. Three separable blocks (24, 48 and 64 channels; 3×3 depthwise and 1×1 pointwise convolution, batch normalization, ReLU, striding and max pooling) reduce the 980×1000 input to $64 \times 15 \times 15$. Global average pooling gives a 64-dimensional vector, and one hidden layer (128 units, batch normalization, 40 % dropout) maps it to the 216 logits. Only 6,396 parameters are in the convolutional stack; the final $128 \rightarrow 216$ projection alone has 27,864.

Training. All models were trained with Adam (learning rate 10^{-3}, weight decay 10^{-4}) on cross-entropy with label smoothing 0.1, using a two-epoch linear warm-up followed by cosine decay, gradient-norm clipping at 1.0 and mixed precision, for at most 60 epochs. Training stopped early when validation loss had not improved for 15 epochs, but never before epoch 40. The checkpoint with the highest validation accuracy was evaluated. The batch size was 1,400 for the evaluation folds and 400 for FSD50K-p10. Each fold model trained on one NVIDIA A100 (40 GB) GPU in 6.0 ± 0.9 h.

Alternative architectures. Before settling on this decoder, we evaluated three alternatives. A conventional four-layer CNN (≈380k parameters) matched the compact model to within seed noise with nine times the parameters. A structured state-space model (S4; 67k–4M parameters), which keeps 1 ms resolution over the full second and encodes each ear separately, reached no better accuracy at roughly ten times the cost per epoch. Low-rank recurrent networks ($\boldsymbol{W} = \boldsymbol{M}\boldsymbol{N}^\top$), both vanilla and a biologically constrained leaky-integrator variant with learnable 5–50 ms time constants and recurrent noise, trained well above chance but below convolutional accuracy.

Excitation–inhibition reference decoder. To measure how much spatial information a fixed, physiologically motivated decoder can extract from the same input, we also evaluated a parameter-free excitation–inhibition model (Kelvasa & Dietz, 2015; Hu et al., 2023). Each ear's 980 fibers are grouped into 35 bundles. Each bundle drives an LSO-like unit with ipsilateral excitation (20 fibers, 1.1 ms window) and contralateral inhibition (8 fibers, 3.1 ms window); the unit's thresholded, refractory output defines a firing rate. Azimuth is read out from the normalized interaural difference in population rate, mapped linearly onto $\pm 90°$. Because this read-out is monotonic, it cannot resolve front from back or judge elevation, so only its folded azimuth error is interpretable.

A.7 Evaluation against CI listeners

Comparison with CI listeners (Fig. 2A, 2B). The 48 bilateral CI users of Anderson et al. (2022) localized sounds from 19 loudspeakers ($-90°$ to $+90°$ in $10°$ steps, 15 repetitions each). Model responses were read at the same 19 targets in the SRL cell (unseen sounds, rooms and listener), using the FSD50K test sounds rather than the listeners' stimuli. Fig. 2A compares per-unit RMSE (Mann–Whitney U, with the common-language effect size). Fig. 2B shows mean absolute error at each target (mean $\pm$ SEM across folds).

Sigmoid fitting of the response function (Fig. 2C). For each unit, we fitted the four-parameter logistic of Anderson et al. (2022) (their Eq. 1) to trial-level target–response pairs (θ, r):

$$r(\theta) = a + \frac{b - a}{1 + \exp\big(-(\mu - \theta)/\sigma\big)}, \tag{1}$$

by non-linear least squares (initial values $a = 90°$, $b = -90°$, $\mu = 0°$, $\sigma = -50°$). We report three quantities: the *response range* $b - a$ (the arc of responses used), the *inflection point* μ, and the *midline gain* $|b - a|/(4|\sigma|)$ in deg/deg (the slope at μ). Goodness of fit is R^2, and parameters are compared as medians across units. For display, a single pooled fit per population is also drawn.

ITD and ILD cue weighting: probe stimuli (Fig. 2D, 2E). Each model was probed with the HRTF held out of its own training, so that, as for the rest of Fig. 2, sound, room and listener were all unseen (SRL cell). We generated five independent noise tokens in each of two bands, high (750 Hz–8.5 kHz) and low (125–750 Hz), and spatialized them in room 2, which is held out of every fold, with each of the five HRTFs at 9 source azimuths ($-20°$ to $+20°$ in $5°$ steps, elevation $0°$). Before CI processing, one cue was offset on top of the natural spatial cues: ITD by 0, ± 300 or ± 600 µs, by delaying one ear; or ILD by 0, ± 10 or ± 20 dB, applied as $\pm$half the offset to each ear. Each condition was simulated with 4 independent spike-noise draws, each with its own seed: 2 bands $\times$ 2 cues $\times$ 9 sources $\times$ 5 offsets $\times$ 5 tokens $\times$ 4 draws = 3,600 ANRs per HRTF, 18,000 in total, with the same conditions and seeds for every HRTF. Every model also scored the ear-swapped ANRs to check left/right symmetry.

Cue read-out. For each HRTF and azimuth, we measured the natural cue in the unmodified spatialized tokens: ITD as the lag of the interaural cross-correlation peak (within ± 1 ms), and ILD as the interaural RMS level ratio, both averaged over tokens. Reported azimuths were folded into the frontal hemisphere; elevation was not used. The *response cue* is the natural cue at the reported azimuth minus that at the source azimuth, both taken from the model's held-out HRTF. Responses were anchored to the mean response at zero offset for each source and token, then averaged over sources, tokens and draws. This gives one transfer curve per model, from imposed cue to returned cue, whose linear slope is the fraction of the imposed cue that the model's response returns. ILD and ITD slopes, and high- and low-band slopes, were compared with paired t-tests; ratios are reported as the mean of per-model ratios.

Spectral dependency (Fig. 2F). Following the paradigm of Dorman et al. (2014), 11 loudspeakers from $-75°$ to $+75°$ in $15°$ steps at elevation $0°$ played wideband (125 Hz–6 kHz), high-pass (1.5–6 kHz) or low-pass (125–500 Hz) noise. The stimuli were rendered in room 2, which is held out of every fold, and scored with each fold's held-out HRTF (SRL cell). For each unit and band, RMSE is expressed as a percentage change from that unit's own high-pass RMSE, $100\,(\mathrm{RMSE}_{\mathrm{band}}/\mathrm{RMSE}_{\mathrm{HP}} - 1)$. The models were compared with the 11 individual bilateral CI users of Loiselle et al. (2016) (their Table 3).

A.8 Device and neural factors

Interaural insertion mismatch (Fig. 3A, 3B). Right-ear electrode positions were shifted, $\boldsymbol{x}_{\mathrm{el}} \rightarrow \boldsymbol{x}_{\mathrm{el}} + \Delta$ ($\Delta > 0$ deeper, i.e. apical), in the voltage-spread model, with filterbank centre frequencies unchanged. The left ear stayed at 0 mm, so $|\Delta|$ is the interaural place–frequency mismatch. Each of the seven shifts was trained with 12 seeds. The dose–response was tested with Spearman's ρ between $|\Delta|$ and RMSE and with a Kruskal–Wallis test across $|\Delta|$ levels. Each level was compared with the unshifted control (both directions pooled), and shallow with deep insertion at matched $|\Delta|$, using Mann–Whitney U tests.

Linked versus independent AGC (Fig. 3C, 3D). Both arms used the fast-acting AGC setting of Potts et al. (2019), Experiment 2: knee point 65 dB SPL, ratio 10:1, attack 5 ms and release 200 ms, with stimuli presented at 75 dB SPL. This setting differs from the default AGC of Appendix A.4. The AGC computes a smoothed gain that is 0 dB below the knee and negative above it, so it can only attenuate. In the *independent* arm, each ear applies its own gain, G_L or G_R. In the *linked* arm, both ears apply the same gain at every sample, $G = \min(G_L, G_R)$, i.e. the larger attenuation, so the interaural level ratio is preserved. Linking was the only difference between the arms. Before

training, we confirmed on the front end that this setting gave the largest ILD difference between the two arms ($+6.4$ dB on average at lateral azimuths). Fig. 3C shows input and output levels measured through the front end for a source at $+60°$.

Peripheral degeneration: dose–response (Fig. 4A, 44B). A random subset of the 980 fibers in each ear was set to zero across the whole ANR, in both training and testing (Fig. 4A). Silenced fibers were drawn independently for each ear from a mask seed equal to the training seed, so the spread across seeds includes variation between lesion patterns. Symmetric bilateral loss was tested at 14 doses d (0, 10, 20, 30, 50, 60, 70, 75, 80, 85, 90, 92.5, 95 and 97.5 % of fibers per ear), with 12 seeds each.

RMSE was fitted as a continuous piecewise-linear (hinge) function of dose d with 1, 2 or 3 segments:

$$\mathrm{RMSE}(d) = a + b_1 d + \textstyle\sum_j c_j\,(d - \tau_j)^+, \tag{2}$$

where $x^+ = \max(0, x)$. For each set of breakpoints τ_j, the coefficients were solved by least squares, and the breakpoints were found by exhaustive grid search. The fit used per-seed points, and the number of segments was selected by BIC. 95 % CIs for breakpoints and slopes came from a seed-level bootstrap: seeds were resampled within each dose and the selected model was refitted.

Central degeneration (Fig. 4B). A random fraction d of units was silenced in every layer: the output channels of the four convolutional blocks (12, 24, 48 and 64 channels) and the 128-unit hidden layer. Each layer's mask was built once from the mask seed and a layer index, and applied identically in training and testing, with inverted-dropout scaling $1/(1-d)$ so that $d = 0$ leaves the model unchanged. We tested 13 levels ($d = 0$, 10, 20, 30, 50, 60, 70, 75, 80, 85, 90, 92.5 and 95 %) with 12 seeds each; at 97.5 % the 12-channel first layer loses every unit. RMSE against d was fitted with a continuous two-segment piecewise-linear model by non-linear least squares, restarted from a grid of initial breakpoints. The breakpoint's 95 % CI came from 2,000 bootstrap resamples of seeds.

Interaural asymmetry of degeneration (Fig. 4C). Total fiber loss was fixed at 50 % and split between the ears as 50/50, 30/70 or 10/90. Mirror-image splits (70/30, 90/10) were trained as separate conditions and pooled with their counterparts. Each condition had 12 seeds, so the 30/70 and 10/90 groups have 24 runs each. Groups were compared with Mann–Whitney U tests.

Place of degeneration: basal versus apical (Fig. 4D). Fiber index maps monotonically onto cochlear place and characteristic frequency. We silenced either the whole *basal* half (fibers 0–489; electrode CFs $\approx$ 1.1–6.7 kHz) or the whole *apical* half (fibers 490–979; CFs $\approx$ 0.3–1.1 kHz) in both ears. Both remove 490 fibers per ear, so lesion size is matched and only location differs. Partial lesions of each half (50 % and 90 %, 12 seeds) are reported in the supplementary results. The two groups were compared with Welch's t-test and Mann–Whitney U, with Hedges' g and the mean difference (95 % CI) as effect sizes.

Additivity of combined factors. For a combination c with intact control b_c, and each of its single factors f at the same dose with its own intact control b_f,

$$\Delta_c = \overline{\mathrm{RMSE}}_c - \overline{\mathrm{RMSE}}_{b_c}, \qquad \Delta_f = \overline{\mathrm{RMSE}}_f - \overline{\mathrm{RMSE}}_{b_f}, \qquad R_c = \Delta_c \Big/ \textstyle\sum_{f \in c} \Delta_f, \tag{3}$$

where the bar is the mean over six training seeds. Seed means are used rather than medians because a six-seed median takes too few distinct bootstrap values to give a usable interval. The insertion factor and every combination containing it are trained on the device subset and scored against its intact control. Peripheral and central degeneration and their pair are trained on the neural subset and scored against its control (App. A.5). The 95% CIs are percentile intervals over 10,000 bootstrap resamples, drawing seeds with replacement independently within every condition, the intact controls included. A control shared by a combination and its factors is the same draw in both. Every dose used in a combination was also trained alone with the same fixed lesion mask, so no single-factor cost is interpolated.