Electrode Shape Reconstruction via a Dual-Channel 1D Convolutional Neural Network

1st Murad Redzheb
Lab. Neurotechnology, PAML-LN
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Sofia, Bulgaria
0000-0003-1814-0768

2nd Plamen Pekov
Lab. Neurotechnology, PAML-LN
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Sofia, Bulgaria
plamen.pekov@iict.bas.bg

3rd Dimiter Prodanov
Lab. Neurotechnology, PAML-LN
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Sofia, Bulgaria
0000-0001-8694-0535

Chronically implanted neural electrodes deform within soft, viscoelastic brain tissue, yet this mechanical interaction cannot be observed directly. We present a machine-learning framework that reconstructs the shape of a sensor probe having the form-factor of a deep brain stimulation electrode embedded in a hydrogel brain-phantom from a strain-gauge signal. A flexure transducer couples the probe to a Wheatstone bridge, and a dual-channel 1D convolutional neural network maps the normalized bridge signal to 49 deflection targets. Across 33 dynamic sessions, reconstruction reached a mean RMSE of 20.6 µm, which is about 3.4% of the probe width. An ablation showed that input normalization, rather than temporal architecture, was the dominant driver of accuracy. Generalization was the principal limitation: the worst leave-one-day-out case reached 110 µm, and a dynamically trained model extrapolated poorly to static relaxation holds.

Keywords— electrode shape reconstruction; brain phantom; hydrogel compression; Wheatstone bridge; machine learning

I. Introduction

Chronically implanted neural electrodes — such as those used for deep brain stimulation and cortical recording — bend rigidly within soft, viscoelastic brain tissue, where mechanical mismatch and micromotion drive the foreign-body response, gliosis, and eventual signal loss [1]. Characterizing how such an electrode is loaded and deformed by and in turn strains the surrounding tissue is therefore central to designing safer, longer-lived neural interfaces. This process can be studied in soft gel-based brain tissue phantoms, yet it is hard to observe directly in an opaque, time-dependent medium. As a step in that direction, with this work, we build on our existing electromechanical brain-phantom platform [2] that couples a probe embedded in a hydrogel to a strain-gauged flexure and reads out electrically the global probe deformation, and ask whether the shape of the loaded probe can be reconstructed from a global deformation bridge signal.

Accurate shape information is essential to reconstruct the bending forces distribution of a loaded implanted electrode. Full reconstruction can be achieved by optical means in transparent or semi-opaque media, which is the objective of the present contribution. In optically opaque media other means of reconstruction are necessary. At first sight, the deformation of the embedded probe appears to be a classical beam-mechanics problem: if the probe stiffness, boundary conditions, and distributed load were known, the bending profile would agree with Euler–Bernoulli beam theory. Here, however, the measured quantity is not the load distribution or curvature but a proxy signal generated by the coupled hydrogel–probe–flexure system. Machine learning is therefore an empirical inverse calibration strategy, valid within the constrained family of experimentally realised shapes rather than a replacement for beam mechanics. Within this domain, temporal modelling captures the history- and rate-dependent hydrogel–probe load transfer, while leakage-controlled normalisation reduces the influence of sensor offset and session-to-session preload differences.

That inverse problem is challenging in hydrogels, because hydrogel mechanics are strongly time-dependent. Their response under compression reflects not only elastic deformation but also viscoelastic and poroelastic effects, so the measured signal depends on loading rate, and history, and relaxation dynamics rather than on instantaneous displacement alone [3]. This makes time-domain modelling a natural analytical choice. Convolutional sequence models, including temporal convolutional networks with dilated receptive fields, have shown that long-range dependencies can often be captured effectively without recurrent architectures, which is relevant when hysteresis, lag, or relaxation unfold over the duration of the sensing window [4].

The present setup incorporates a mechanically coupled flexure transducer in a T-bridge configuration, instrumented with strain gauges in a Wheatstone bridge to provide a global deformation signal [2]. We hypothesize that the temporal dynamics of this signal contain sufficient information to reconstruct the distributed electrode-deflection profile during hydrogel compression. To test this, we developed a dual-channel convolutional neural network (CNN) that maps normalized bridge signals to 49 electrode-shape targets extracted from optical images acquired orthogonal to the direction of deformation. The framework combines a physically motivated transducer model, a leakage-controlled target definition, a dual-channel normalization strategy, and a systematic evaluation across in-session and generalization tests.

II. Transducer Principle and Learning Problem

A. Wheatstone Bridge Flexure Sensing

A custom flexure transducer was used to convert hydrogel compression into an electrical descriptor of mechanical response. The transducer consists of a 3D-printed steel probe, 600 µm wide, embedded in the hydrogel and mechanically coupled normal to a cross-shaped base (Figure 1a) [2]. The

979-8-3195-3402-6/26/$31.00 ©2026 IEEE

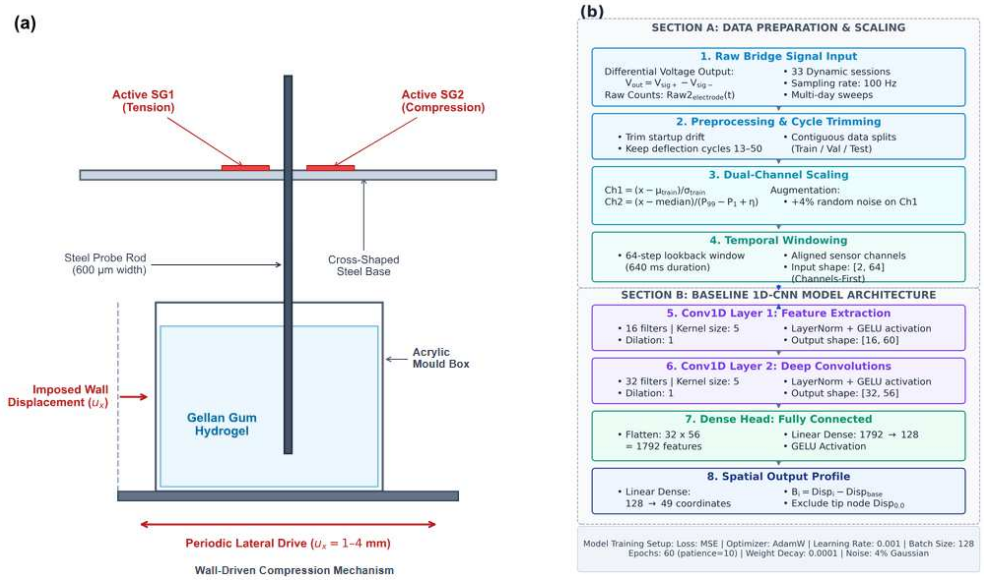


Figure 1. Experimental setup and CNN workflow. (a) Wall-driven compression of gellan gum hydrogel with embedded steel probe and strain gauges. (b) Data preparation and 1D-CNN architecture used to reconstruct the spatial displacement profile from sensor signals.

compressed medium was a 0.4 % (w/v) low-acyl gellan-gum hydrogel (4.4 L) gelled by hydronium (H_3O^+) cations at pH 3.0, with a storage modulus $G' \approx 1$ kPa, comparable to that of a human brain tissue [5], serving as a viscoelastic brain-phantom analogue. The gum was dissolved in distilled water at 90°C; after the complete dissolution, the pH was lowered to 3 with concentrated HCl, and the solution was cast into a plexiglass box. Following casting, the steel probe was inserted into the still-liquid solution so that the hydrogel gelled around the rod, ensuring continuous, gap-free rod–hydrogel coupling and reliable load transfer. During cyclic compression and relaxation, hydrogel displacement drives the probe and bends the base, causing two shoulder-mounted active gauges to experience opposite strain states. The gauges were wired as an active half bridge using two EA-06-015DJA-120 foil strain gauges connected in series, each with nominal resistance 120 Ω and gauge factor 1.820 ± 2%. Two additional precision resistors were incorporated (R3 and R4 – 1kΩ) to complete the Wheatstone bridge, resulting in an equivalent full-bridge configuration (Figure 1a). The bridge was excited at 5.0 V DC. Because the hydrogel is subjected to compression and relaxation, the bridge output is treated as the time-dependent mechanical response transmitted through the hydrogel–probe–base system.

B. Shape Target Definition

Video-based coordinate tracking (DMK 33UX290 camera, Sony IMX290 CMOS sensor, COMPUTAR 4–8 mm f/1.4 lens) was used to obtain the ground-truth probe shape during dynamic compression. Each frame was converted to 8 bit grayscale, smoothed with a 5×5 Gaussian kernel, and the probe region was localized by Otsu thresholding. Probe boundaries were then detected along horizontal scanlines at 2-pixel vertical intervals and refined by sub-pixel interpolation; the centerline, taken as the midpoint between the boundaries, was interpolated across any missing rows and denoised with a second-order Savitzky–Golay filter. The filter window was set adaptively to one-fifth of the rod's pixel length, rounded to the nearest odd integer and constrained to 5–51 pixels (11 pixels for a typical 50-pixel rod).

The smoothed centerline was resampled into 50 geometry-weighted nodes using an exponential distribution that increases sampling density near the clamped base. To reduce the influence of rigid-body motion, fixture compliance, and mounting-arm flexure, the tracked coordinates were corrected relative to the base coordinate, and the distal tip node was excluded from the supervised target because tracking at the image boundary frequently produced noisy, non-physical coordinates. The final supervised output therefore contains 49 active base-corrected bending deflections. Reconstruction was treated as a one-axis bending problem because the imposed compression produced a dominant loading direction and no lateral deflection was evident.

C. Dataset and Experimental Conditions

The recordings span four nominal compression amplitudes (1, 2, 3, and 4 mm; strain 0.5–2.0%) and four frequencies (0.1, 0.2, 0.5, and 1.0 Hz) across three days, giving 48 amplitude–frequency condition cells; after excluding sessions with sensor dropouts or mechanically invalid data, 33 valid dynamic sessions remain (Table 1). Ambient temperature was held at 22 ± 2 °C, limiting thermal contributions to day-to-day signal variation, and the hydrogel was not re-hydrated after gelation. Since every experiment contained 50 loading cycles, low-frequency trials lasted longer and generated more 64-sample windows. Thus, slower sessions therefore had a proportionally larger influence on training.

Table 1. Valid dynamic sessions by amplitude-frequency condition; total = 33 sessions.

Amplitude / Freq	0.1 Hz	0.2 Hz	0.5 Hz	1.0 Hz
1 mm	2	2	2	2
2 mm	1	1	2	1
3 mm	2	3	2	2
4 mm	3	3	3	2

D. Preprocessing and Model Architecture

1) Leakage-Controlled Signal Scaling

The Wheatstone bridge output is transformed into two complementary normalized channels before being provided to the learning model. Channel 1 preserves the global amplitude scale of the experiment by applying z-score normalization

using only the mean and standard deviation computed from the training fold. This retains the absolute magnitude of the bridge response relative to the training distribution, linking the transducer's mechanical response to probe-shape deformation:

$$Ch1_t = (x_t - \mu_{train})/\sigma_{train}.$$

Channel 2 compensates for local baseline shifts in the bridge signal using robust baseline scaling based on the available flat pre-compression segment:

$$Ch2_t = \left(x_t - median\left(X_{pre}\right)\right)/((P_{99} - P_1) + \eta),$$

where X_{pre} represents the available flat pre-compression baseline, $(P_{99} - P_1)$ is the 1st-to-99th percentile range, and η is a stabilizing constant that prevents division by very small values. The second channel improves tolerance to baseline shifts without fitting statistics on validation or test targets. The median provides a robust estimate of the local baseline offset, while the 1st-to-99th percentile range provides a robust estimate of baseline signal spread. This makes the second channel less sensitive to outliers, segmentation artefacts, and day-to-day baseline offset drift than a mean/std normalization computed from the same local segment.

Both transformations are leakage controlled: the global statistics are fitted only on the training fold, and the robust local statistics are computed only from the pre-compression baseline available before the corresponding loading interval, never from validation or test data, so no future information enters the model input.

2) Dual-Channel CNN

The baseline model is a dual-channel 1D CNN that maps the normalized Wheatstone bridge temporal descriptor to the probe-shape target vector (Figure 1b). The input is a 64-sample window with two normalized channels. The convolutional trunk consists of two sequential Conv1D blocks. The first block uses 16 filters and the second uses 32 filters, both with kernel size 5, stride 1, and no (valid) padding. Each convolutional layer is followed by layer normalization [6] and GELU activation [7]. The two valid convolutions reduce the 64-sample window to 56 samples, giving a 32-channel × 56-sample feature map that is flattened into a 1792-dimensional vector and passed through a dense layer with 128 units and GELU activation. A final linear layer maps this representation to the 49 active deflection targets. The training loss represents the mean squared error (MSE) over the 49 targets.

The CNN was trained using AdamW optimization [8] with decoupled weight decay. The initial learning rate was 1×10^{-3}, and weight decay was set to 1×10^{-4}. A ReduceLROnPlateau scheduler reduced the learning rate by a factor of 0.5 after four validation epochs without improvement, with a minimum learning rate of 1×10^{-6}. Training used mini-batches of 128 windows, a 60-epoch budget, and early stopping with patience 10 on validation MSE; no dropout was used in the convolutional trunk. During training, Gaussian noise was injected into the Channel 1 input [9] to improve robustness against sensor baseline offset drift. A sensitivity sweep from 1% to 6% of the raw signal scale gave mean test RMSEs values with no statistically significant differences (Friedman test, $\chi^2 = 9.05$, $p = 0.11$); the minimum, at 4% noise, was selected for all evaluated runs. The reported runs used standard unweighted MSE loss.

The rationale for the two channels is the following: the global z-score channel preserves amplitude relative to the training distribution, helping the network exploit absolute signal magnitude near the training domain, while the local robust channel centres each sequence on its pre-compression baseline and rescales it with statistics less sensitive to outliers than mean/standard-deviation estimates. This was needed due to observed baseline artefacts and inter-day drift.

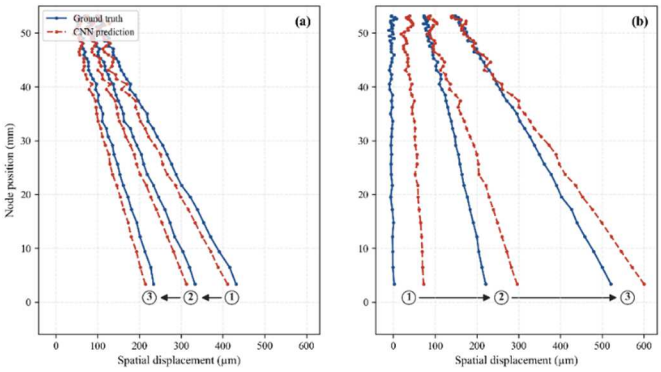


Figure 2. Reconstructed vs. ground-truth electrode shape at three chronological snapshots for (a) the best session (1 mm, 0.1 Hz; RMSE 8.5 μm) and (b) the worst (1 mm, 1.0 Hz; RMSE 37.7 μm).

3) Alternative Temporal Architectures

A narrow standard CNN has a limited receptive field and may not capture hysteresis or viscoelastic lag across the full 64-sample window. A dilated causal residual trunk with kernel size 3 and dilations d = [1, 2, 4, 8, 16] expands the receptive field to 63 samples, nearly spanning the input window while preserving causal temporal structure. Thus, the Temporal Convolutional Network (TCN) is included as an architecture variant for long-memory dynamics. As a recurrent alternative, a single-layer gated recurrent unit (GRU [10], hidden size 64, 27.7k parameters) processes the same 64-sample window step by step and regresses the shape targets from its final hidden state, testing whether explicit recurrent state captures a potential hysteresis better.

III. RESULTS

A. Dynamic Deflection Performance

Using the proposed dual-channel scaling pipeline, the model achieved a mean test Root Mean Square Error (RMSE) of 20.6±1.8 μm — approximately 3.4% of the nominal 600 μm probe width. Error remained essentially flat across the three testing days despite measured baseline drift (pooled RMSE 21.3, 20.3, and 20.7 μm for Days 1–3). The best-performing dynamic condition was the 1 mm amplitude, 0.1 Hz frequency trial from Day 2, which achieved a median test RMSE of 8.8±0.5 μm (Figure 2a). In contrast, the worst-performing dynamic condition was the 1 mm amplitude, 1.0 Hz frequency trial from Day 2, yielding a test RMSE of 43.0±2.0 μm (Figure 2b). Across amplitudes, reconstruction errors generally increased at higher loading frequencies; for example, at 1 mm amplitude, RMSE increased from 8.8±0.5 μm at 0.1 Hz to 43.0±2.0 μm at 1.0 Hz. Notably, training without input normalization (raw voltage) caused catastrophic breakdown: mean test RMSE 346.3 μm, a 16.8× inflation over the dual-channel model (20.6 μm) and 14.1× over the single-channel standardized baseline (24.54 μm).

B. Generalization Across Evaluation Splits

To evaluate reliability under different experimental splits, four regimes were analysed (Figure 3): in-session dynamic

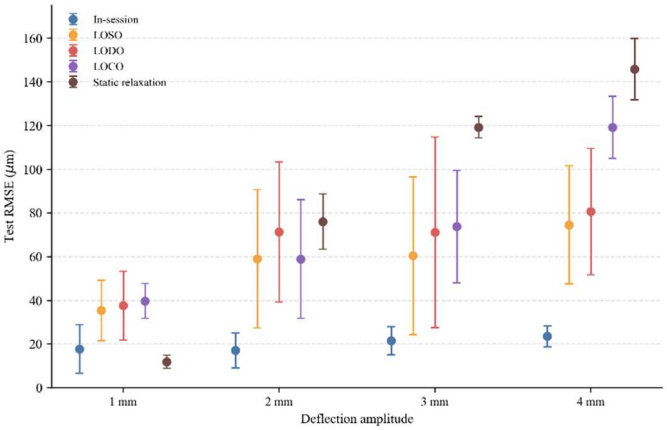


Figure 3.Test RMSE by deflection amplitude across five evaluation regimes. Error bars: ±1 SD.

splits (interpolation), leave-one-session-out (LOSO) cross-validation (held-out session), leave-one-day-out (LODO) cross-validation (cross-day generalization), and leave-one-condition-out (LOCO) cross-validation (extrapolation to an unseen amplitude–frequency condition). In-session interpolation achieved a worst-case test RMSE of 43.0 µm. Under LOSO, holding out one session (day × condition) at a time, the mean test RMSE was 66.0 µm across three seeds (worst fold 166.6 µm) — intermediate between in-session interpolation and the harder cross-day and cross-condition splits below. Under LODO, the average test RMSE was 81.3±24.6 µm (worst fold: 110.2 µm), a ~156% increase in worst-case error over in-session; adding additive-noise (σ = 0.04) and gain (0.85–1.15) augmentation gave no improvement (80.6±25.4 µm). Under LOCO, where an entire amplitude–frequency condition is withheld from training, the mean held-out RMSE was 74.2 ± 32.4 µm across three seeds, with the worst condition (4 mm, 0.1 Hz) reaching 110.9 µm — a 4.0× increase over condition-specific training (18.31 µm). Pilot attempts to close this gap using Feature-wise Linear Modulation (FiLM) [11] and late-concatenation metadata architectures did not succeed, yielding held-out test MSEs of 1,297,848.63 and 94,275.97 respectively on the worst-case 4 mm, 0.5 Hz condition.

C. *Generalization to Static Relaxation Holds*

To evaluate reconstruction under constant strain, the model was tested on four 300 s stress-relaxation holds, one per nominal amplitude (1–4 mm). A model trained strictly on dynamic cyclic sweeps generalized poorly to these flat-input holds (mean RMSE 102.2 ± 1.6 µm; worst 104.2 µm; Figure 3). Adding the holds to training recovered within-hold reconstruction to 6.9 ± 1.4 µm on held-out late-phase windows — effectively an in-session interpolation task. The stricter extrapolation case, Leave-One-Relaxation-amplitude-Out (LORO) cross-validation, withheld an entire hold amplitude from training: the mean unseen-amplitude RMSE was 84.1 µm, rising monotonically from 15.7 µm (1 mm withheld) to 138.9 µm (4 mm withheld).

D. *Architecture Ablation*

As a secondary architecture ablation, a dilated causal TCN was trained on the pooled dataset under identical single-channel scaling and compared against the single-channel 1D CNN. The TCN reduced the mean test RMSE from 24.54 to 23.87 µm (5.2 % lower MSE), a difference within the seed-to-seed variability of the headline model (±1.8 µm) and obtained at a single seed. The benefit was concentrated in conditions where the 64-sample window undersamples the loading cycle. Under the same single-channel scaling and in-session split, a single-layer GRU (hidden size 64, 27.7k parameters) reached a mean test RMSE of 30.4 µm, worse than both convolutional variants, beyond the seed-to-seed variability, observed from the CNN training runs.

IV. DISCUSSION AND CONCLUSION

The two-active-gauge Wheatstone bridge measures bending at the flexure base induced by lateral displacement of the coupled probe. It therefore provides a global temporal descriptor of the hydrogel–probe–base mechanical response, rather than a direct measurement of hydrogel strain or local electrode curvature. This distinction explains both the success and the limits of the tested models. Repeated loading produces a reproducible mapping from temporal base response to distributed probe shape, but the inverse relation remains indirect and potentially non-unique outside the calibrated operating range. The generalization test show that reconstruction accuracy depends on support coverage across loading amplitude, frequency, hold state, and temporal history. Overall, the results support the central hypothesis that temporally modelled bridge signals contain sufficient information to reconstruct electrode-deflection profiles during hydrogel compression, but only within a bounded manifold.

A key finding of the present study is that normalisation mattered more than the change in temporal architecture. Performance collapsed without normalisation, recovered substantially with global z-scoring, and was strongest with the proposed dual-channel representation (data not shown). By contrast, neither the dilated causal TCN, which improved only modestly over the CNN, nor a recurrent GRU variant, which performed worse, changed this picture in the reported head-to-head comparison. This suggests that the dominant learning challenge is not only long-memory modelling, but also handling non-stationary signal statistics, particularly baseline drift and session-to-session scale variation.

The cross-day results reveal a more fundamental limitation. Local baseline subtraction removes additive offset drift, but a zero-load segment contains no information about sensor slope and therefore cannot correct multiplicative gain or sensitivity changes. Performance deteriorated sharply under LODO evaluation and was not recovered by additive-noise or gain augmentation. The cross-day gap is therefore better interpreted as a distribution-coverage limitation than as a simple correctable offset or gain error. Although the same hydrogel was used throughout, progressive stiffening through syneresis and/or drying likely shifted the operating point across testing days, placing held-out days partly outside the trained envelope.

Temporal modelling is not sufficient unless the relevant histories are represented during training. A model trained only on cyclic dynamics failed on 300 s holds, whereas adding static holds reduced the error to nearly the dynamic level. Sustained holds therefore occupy a distinct region of temporal state space not spanned by cyclic windows alone. This is consistent with viscoelastic and poroelastic relaxation, where constant-strain holds reveal redistribution processes that are not equivalent to short oscillatory segments [3]. The strong LORO penalty when an entire hold amplitude was withheld further shows that the network learns a data-supported mapping rather than a general relaxation law.

The preprocessing choices improve the physical relevance of the supervision. Base-corrected coordinates reduce rigid-body and mount-compliance artefacts, while tip-node exclusion and clamp-concentrated node density reduce edge-of-image noise and prioritise the mechanically informative proximal region. Because preprocessing can inflate apparent performance if future-label or test-statistic information leaks into training [12], fitting global statistics on training folds only and local statistics on the pre-compression segment only is an important methodological safeguard. However, the reported error should be interpreted over the reliably observable active profile rather than the entire visible rod.

Future work will focus on calibration transfer, uncertainty under unseen loading programmes, and physical admissibility of predicted shapes. For deployment, unsupervised domain adaptation such as covariance-alignment domain adaptation [13] could align latent-feature covariances between training and daily deployment domains using a short buffer of unlabeled voltages, while online input standardisation could suppress low-frequency baseline drift. The spread across independently trained models [14] could further flag inputs from unseen loading programmes as uncertain. Finally, physics-guided geometric or energy-consistency penalties may encourage physically plausible shapes consistent with clamped-base cantilever bending, while avoiding the optimisation difficulties associated with higher-order beam constraints [15], [16].

In conclusion, we showed that the shape of a slender probe embedded in a hydrogel brain-phantom analogue can be reconstructed from a single proximal Wheatstone bridge signal, without sensing distributed along the device. A flexure transducer couples the probe to a strain-gauged cross-base, and a dual-channel 1D CNN maps the normalized bridge signal to a 49-node base-relative deflection profile. Within the trained loading envelope the model achieved a mean dynamic reconstruction error of 20.6 — about 3.4 % of the 600 µm probe width. An ablation established the leakage-controlled dual-channel normalization, rather than network depth or temporal receptive field, as the principal driver of this accuracy. The main limitations are out-of-distribution: cross-day transfer and extrapolation to unseen amplitudes or sustained holds degrade to $\approx$110-139 µm ($\approx$18-23% of probe width) and are not repaired by noise or gain augmentation, consistent with the hydrogel stiffening across days and shifting the operating point outside the trained envelope. The system is therefore best understood as a calibrated inverse decoder — accurate within its measured mechanical envelope, not a universal mechanical observer.

ACKNOWLEDGMENT

The authors acknowledge the European Union's ERA Chair program for funding the VIBraTE project ("Viscoelastic Interactions of the Brain Tissue with Brain Computer Interfaces", 101086815).

REFERENCES

[1] D. Prodanov and J. Delbeke, "Mechanical and biological interactions of implants with the brain and their impact on implant design," *Front. Neurosci.*, vol. 10, no. FEB, 2016, doi: 10.3389/fnins.2016.00011.

[2] H. I. Hristov, P. Pekov, T. Minev, and D. Prodanov, "Electromechanical Setup for Measuring Viscoelastic Interactions of a Brain Phantom," in *2025 6th International Conference on Communications, Information, Electronic and Energy Systems (CIEES)*, IEEE, Nov. 2025, pp. 1–6. doi: 10.1109/CIEES66347.2025.11300229.

[3] M. L. Oyen, "Mechanical characterisation of hydrogel materials," *Int. Mater. Rev.*, vol. 59, no. 1, pp. 44–59, 2014, doi: 10.1179/1743280413Y.0000000022.

[4] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1003–1012, 2017, doi: 10.1109/CVPR.2017.113.

[5] S. Budday, T. C. Ovaert, G. A. Holzapfel, P. Steinmann, and E. Kuhl, "Fifty Shades of Brain: A Review on the Mechanical Testing and Modeling of Brain Tissue," *Arch. Comput. Methods Eng.*, vol. 27, no. 4, pp. 1187–1230, Sep. 2020, doi: 10.1007/s11831-019-09352-w.

[6] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," 2016, [Online]. Available: http://arxiv.org/abs/1607.06450

[7] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," Jun. 2023, [Online]. Available: http://arxiv.org/abs/1606.08415

[8] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *7th Int. Conf. Learn. Represent. ICLR 2019*, 2019.

[9] C. M. Bishop, "Training with Noise is Equivalent to Tikhonov Regularization," *Neural Comput.*, vol. 7, no. 1, pp. 108–116, 1995, doi: 10.1162/neco.1995.7.1.108.

[10] S. Bhaduri, "Evaluation of different techniques for detection of virulence in Yersinia enterocolitica," *J. Clin. Microbiol.*, vol. 28, no. 4, pp. 828–829, 1990, doi: 10.1128/jcm.28.4.828-829.1990.

[11] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," *32nd AAAI Conf. Artif. Intell. AAAI 2018*, pp. 3942–3951, 2018, doi: 10.1609/aaai.v32i1.11671.

[12] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine- learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[13] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," *Adv. Comput. Vis. Pattern Recognit.*, no. 9783319583464, pp. 153–171, 2017, doi: 10.1007/978-3-319-58347-1_8.

[14] J. Gawlikowski *et al.*, *A survey of uncertainty in deep neural networks*, vol. 56, no. s1. Springer Netherlands, 2023. doi: 10.1007/s10462-023-10562-

9.

[15] T. Kapoor, H. Wang, A. Nunez, and R. Dollevoet, “Physics-Informed Neural Networks for Solving Forward and Inverse Problems in Complex Beam Systems,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 5, pp. 5981–5995, 2024, doi: 10.1109/TNNLS.2023.3310585.

[16] A. S. Krishnapriyan, A. Gholami, S. Zhe, R. M. Kirby, and M. W. Mahoney, “Characterizing possible failure modes in physics-informed neural networks,” *Adv. Neural Inf. Process. Syst.*, vol. 32, no. ML, pp. 26548–26560, 2021.