OpenMuscle: An Open Forearm Force-Myography Dataset with Continuous Finger Labels, and Baselines Across Sessions and Wearers

Author: Tory Moghadam (OpenMuscle), ORCID 0009-0001-8126-247X

Report: OpenMuscle technical report v0.3, 2026-11-09 (supersedes v0.2 of 2026-11-03). DOI pending.

Artifacts: open hardware (FlexGrid V4 and V3), open-source capture and analysis pipeline (OpenMuscle-Software, `analysis/paper/`), dataset release 0.1 on Hugging Face, DOI 10.57967/hf/11416 (Section 8)

Abstract

OpenMuscle is an open forearm force-myography (FMG) system: bands of 60 Velostat pressure cells, a capture pipeline, and two sources of continuous finger labels, Meta Quest 3S hand tracking and the LASK5, a handheld labeler with four spring-loaded pistons. The recordings are public (DOI 10.57967/hf/11416). On the main subset, two FlexGrid V4 bands with hand-tracking labels (two wearers, four sessions, 22,291 frames), we report frame-wise random-forest baselines next to a constant baseline that predicts the training-set mean pose. Within a session, finger-flexion mean absolute error is 0.043 of a full curl under a random split and 0.110 under a temporal split (constant 0.218); random splits of adjacent frames overstate within-session skill by 2.4 to 2.6 times at this sampling rate. Without calibration the model does not transfer across a re-donning five days apart (0.327 against a constant 0.205) or across wearers (0.272 against 0.216): it is worse than predicting the average pose. A same-day re-donning kept some skill (0.160 against 0.202). Part of the change between sessions is a per-cell offset and scale shift. Standardizing each cell with statistics from unlabeled data of the test session removes the worse-than-constant failure: with 30 s of calibration data, same-wearer error is 0.169 against 0.223 for the constant on the same frames, while across wearers the error stays near the constant level. A one-band LASK5 subset shows the same pattern for piston position. With two wearers these are observations, not general findings. Every result in the main text reproduces from the public release in a few commands.

1. Introduction

Hand pose is the highest-bandwidth human output channel, yet capturing it outside the lab still requires cameras with line-of-sight or instrumented gloves. Forearm-mounted sensing, whether surface electromyography (sEMG) or, as here, force myography (FMG: the pressure distribution over the forearm surface produced by muscle and tendon displacement) [1], is an unobtrusive alternative for prosthetics, teleoperation and input devices. Two problems slow it down. Ground-truth labels for continuous finger motion are expensive to collect, and models trained in one session often fail in the next, a problem the sEMG literature has documented for years as electrode shift and session and user variability [2, 3].

OpenMuscle addresses the first problem with two cheap label sources: the hand tracking that ships in consumer VR headsets, where a Quest 3S streams 25-joint hand poses into the same clock domain as the band data, and the LASK5, a handheld labeler with four finger pistons. For the second, this report does not propose a method. It measures how a deliberately simple model behaves under evaluation splits of increasing difficulty, each against a constant baseline, and it releases the data so that others can study the shift.

Contributions:

1. An open FMG system with two labelers: band hardware, firmware, and a capture pipeline that stamps wearer, device and session into every recording.

2. A public dataset with continuous finger labels (release 0.1, DOI 10.57967/hf/11416): two FlexGrid V4 bands with hand-tracking labels, and legacy FlexGrid V3 recordings with piston or hand-tracking labels.
3. A generalization ladder (random split, temporal split, cross-session, cross-wearer) in which every rung is reported next to a constant baseline.
4. Evidence that the static per-cell pressure profile shifts between sessions, and a measurement of what label-free per-session standardization does and does not recover.
5. Reproduction of every number from the public release in a few commands (Section 9).

This is version 0.3 of the report. It adds the constant baseline, standardization, blocked cross-validation and the robustness checks, rewrites the claims around them, and withdraws one claim of v0.2 (see the Version note).

2. Related Work

FMG hardware and open bands. Xiao and Menon review two decades of FMG hardware and signal processing and the challenges that remain before everyday use [1]. Radmand, Scheme and Englehart used a high-density array of forearm pressure sensors to classify eight hand and wrist motions in a fixed arm position, and report that changes in arm position introduce error [4]. Ravindra and Castellini compared sEMG, ultrasound and pressure sensing for finger-force detection and found pressure sensing comparable in accuracy to sEMG, more stable over time, and cheaper [5]. Open-source FMG bands exist: Chapman, Dwivedi and Liarokapis released a six-sensor armband that decodes four gestures plus rest [6], and Sagastume, Young and Schofield validated a low-cost open band on 15 participants and 10 gestures [7]. The public FMG datasets we found carry gesture or force labels: Andersen et al. released 18-gesture data from 20 people [8], and Zakia and Menon released FMG recorded during physical human-robot interaction with applied hand force as the target [9]. The OpenMuscle release adds continuous finger labels. Its sensing material, Velostat, shows hysteresis and nonlinearity under static and cyclic load [10].

Continuous hand pose from wearable signals. NinaPro relates sEMG to hand kinematics and forces in a public database [11]. emg2pose pairs 16-channel wrist sEMG with motion-capture hand pose from 193 users and defines held-out user, sensor-placement and stage splits; its authors state that sEMG depends strongly on anatomy and sensor placement [12]. A large sEMG wristband study reports decoding models that generalize across people when trained on data from thousands of participants [13]. With pressure sensing, Shull et al. estimated continuous per-finger flexion angles from ten barometric sensors at the wrist [14], Mizrahi and Sintov mapped forearm FMG to hand poses with a temporal convolutional network and found that fine-tuning on a small amount of data improves transfer to a new user [15], and Kolev et al. recover full hand pose and contact force from a wrist-worn capacitive pressure array [16]. Obukhov et al. fuse EMG, IMU and VR-tracker data for exoskeleton control [17]; their related-work section describes an earlier OpenMuscle prototype, which had no citable reference at the time. We do not compare error numbers with these works: placements, targets and splits differ, and Section 5 shows that the split alone changes the reported error by a factor of more than two.

Session, donning and user shift. In sEMG pattern recognition, shifting the electrodes during donning and doffing degrades classification, more for shifts perpendicular to the muscle fibres [2], and electrode configuration and features can reduce that sensitivity [18]. Kyranou, Vijayakumar and Erden review the causes of degradation outside the lab (non-stationarity, electrode displacement, arm posture, fatigue, differences between users) and the countermeasures for each [3]. Zhai et al. recalibrate an sEMG classifier from session to session without user retraining [19]. We found no FMG-specific study of re-donning shift. This report quantifies the shift for one open FMG system and releases the data to study it.

Evaluation leakage and label accuracy. Kapoor and Narayanan document data leakage across 17 fields and give a taxonomy of its types [20]; random splits of temporally correlated frames are one instance. On labels, Abdlkarim et al. measured the Meta Quest 2 hand tracking at about 1.1 cm average fingertip error, about 9.6 degrees average finger joint-angle error, and about 45 ms delay [21]. The Quest 3S used here is a different headset, so that figure

sets an expectation, not a measured bound. Han et al. describe the four-camera monochrome hand-tracking system that became the default hand tracking on the Oculus Quest [22].

3. System

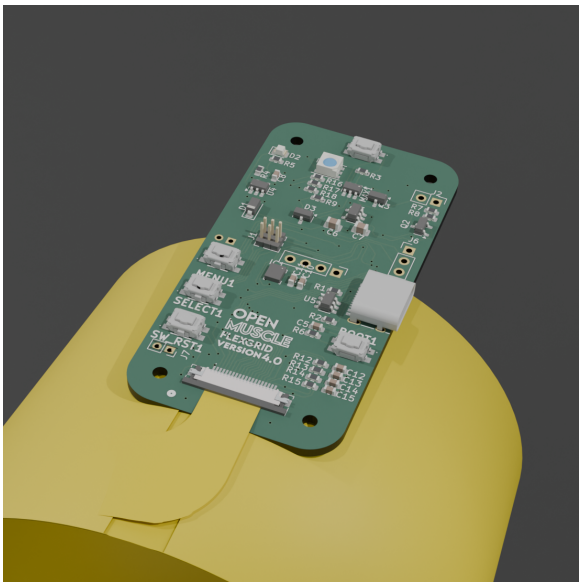


Figure 1: 3D render of the FlexGrid V4 circuit board on a forearm-shaped sensor cuff. This is a computer render, not a photograph; the sensor sheet colour and the strap are not shown as built.

Band. FlexGrid V4: a 15x4 grid of Velostat pressure cells (60 channels) on a flexible PCB, scanned by an ESP32-S3 (12-bit ADC) through an analog multiplexer on a rigid board that also carries a 6-axis IMU (Figure 1). Each band streams at 17-19 Hz over Wi-Fi UDP to a PC hub, and two bands (one per forearm) stream simultaneously. V4 was brought up in June 2026; two V4 bands shipped as the first OpenMuscle starter kit in October 2026. The legacy subset used one band of the previous revision, FlexGrid V3 (same 60-cell grid and ESP32-S3, no IMU columns in these captures), which streamed at 14-25 Hz (about 24 Hz in the LASK5 captures). Velostat is a polyethylene-carbon composite whose response shows hysteresis and nonlinearity [10]; neither is characterized for these bands. Schematics, layout, firmware and the communication protocol are open (Section 8).

VR labels. A WebXR application on a Meta Quest 3S streams both hands' 25-joint poses (position and orientation, 7 floats per joint, 175 values per hand) to the hub, where a temporal matcher pairs each sensor frame with the nearest label frame within a 175 ms window, timestamped on the single hub clock. Each band is routed to the hand on its own arm (left band to left hand). Quest hand tracking is a camera-based estimator [22] with its own error [21].

LASK5 labels. The LASK5 is a handheld labeler with four spring-loaded pistons, one per long finger, and a joystick. It reports each piston's position in [0,1] and is matched to sensor frames within a 100 ms window. The fingers press against the springs, so every labeled movement is a loaded contraction; under VR labels the fingers move in free air.

Canonical targets. Raw joint positions are world-frame and therefore unusable for cross-session comparison (a model would learn where the wearer stood). All VR-labeled experiments regress 15 wrist-relative angular degrees of freedom per hand, derived deterministically from the raw joints: per-finger flexion normalized to [0,1] (5 targets; 0 = extended, 1 = curled) and per-joint flexion angles in degrees for MCP and PIP of each finger and MCP and IP of the thumb (10 targets). For this report they are re-derived uniformly from the raw labels with `openmuscle.hand_angles`. LASK5 experiments regress the four piston positions directly.

4. Dataset

The report uses three subsets: FlexGrid V4 with VR labels (main), and legacy FlexGrid V3 recordings with LASK5 labels and with VR labels. No V4 recording with LASK5 labels exists: every LASK5-labeled capture names the V3 band, and all of them (May 2026) predate V4 bring-up.

The author attests that every single-band recording in the release's LASK5, V3 Quest and 2025 legacy sets is wearer W1's right forearm, including those whose metadata records no wearer, and that every two-band recording is both forearms of one wearer.

4.1 FlexGrid V4 + Quest (main subset)

Two wearers (W1, W2), four sessions, eight two-band captures with both hands labeled, 602.7 s of recording summed over the captures (about 10 minutes per hand). Rows without a full 25-joint label are dropped; the resulting set is:

Wearer	Session	Date (2026), start	Hands	Rows (L/R)	Captures
W1	W1-s1	Jun 28, 11:00	both	3,010 / 3,013	3
W2	W2-s1	Jul 3, 15:49	both	934 / 929	1
W2	W2-s2	Jul 3, 16:20	both	4,361 / 4,350	2
W1	W1-s2	Jul 3, 16:34	both	2,840 / 2,854	2
Total				**11,145 / 11,146**	8 (22,291 frames)

Start times are local clock times from the capture metadata. W2-s1 and W2-s2 are separate sessions about 30 minutes apart; W1-s2 followed W2-s2 on the same two bands, five days after W1-s1. Two captures from these sessions are excluded and not released (named in `analyze.py`): a 9.7 s left-band-only fragment of W2-s1, and a W2-s2 capture with a 29.5 s hand-tracking dropout.

One sensor column (C3, four cells) on the left band was dead for W1-s1: three of its cells read exactly zero in every analysed row of that session and the fourth (R3C3) in 92% of them; in W1-s2 the same cells read zero in at most 1% of rows. Section 5.3 tests whether this column drives the W1 result.

4.2 FlexGrid V3 + LASK5 (legacy, May 2026)

Twenty captures in seven sessions (L1 to L7; a session is a run of captures with less than an hour between starts), 35,345 frames, 24.4 minutes, all on one V3 band on W1's right forearm. One empty capture is excluded.

Session	Date (2026), start	Captures	Rows	Seconds
L1	May 16, 08:38	1	3,952	167
L2	May 16, 22:01	4	9,169	384
L3	May 17, 07:52	4	8,102	333
L4	May 17, 21:57	1	1,190	50
L5	May 24, 14:59	3	3,948	162
L6	May 24, 18:59	6	7,773	317
L7	May 31, 09:57	1	1,211	50

4.3 FlexGrid V3 + Quest (legacy, May to June 2026)

Ten right-hand captures of W1 with VR labels on the same V3 band (May 25 to June 4), 11,619 frames, each at least 500 rows (about 20 s); one shorter fragment is left out. They are used only for the labeler comparison in Section 5.7.

4.4 Participants, movement and ethics

W1 is the author and W2 an adult member of his family. Both took part voluntarily, and the recordings are released in pseudonymised form with their agreement. No institutional ethics review was sought, because this is independent, self-funded work carried out outside any institution. Both wearers have intact hands; no data from people with limb difference is included.

No movement routine was recorded for these sessions. The capture metadata hold no movement list, prompt sequence or cue schedule, and the project's written movement guide (a 30 s take with an initial rest, followed by open-close, single-finger, grip or free movement) was published in October 2026, after every recording analysed here. In the two-band sessions the two hands moved largely in mirror: the correlation between left and right flexion, on time-aligned frames, is 0.78 to 0.85 per finger (mean over fingers) and 0.86 to 0.94 for mean flexion across the four sessions. Band placement (distance from the elbow, rotation, strap tension) was not measured or recorded in any session.

4.5 Mapping to the public release

In release 0.1 (Hugging Face `openmuscle/openmuscle-fmg`, DOI 10.57967/hf/11416, pinned revision `8e4fa1f`), W1 is `P01` and W2 is `P02`. The manifest column `paper_session` names the report session (W1-s1, W1-s2, W2-s1, W2-s2) of each analysed capture. The release holds 14 V4 Quest captures, of which the 8 with a `paper_session` are analysed. The analysed set contains only two-band captures; the five single-band V4 captures (four of P01's right forearm, June 26 to 28, and one of P02's left forearm, July 3) are released but not analysed. A ninth two-band capture, `capture_1783111690` (P02, 27.5 s, July 3), is published but not analysed, and no reason for leaving it out was recorded. `REPRODUCE.md` lists every capture used per session.

5. Experiments

5.1 Models, protocols and baselines

V4 + Quest. One random forest (100 trees, seed 42) per hand per condition maps the 60 raw cell values of a single frame to the 15 canonical targets. There is no temporal context, so that differences between evaluation protocols cannot be attributed to model capacity. E1: random 80/20 split within a session. E2: temporal 80/20 split within a session (train on the first 80% of the timeline). E3: train on one session, test on another session of the same wearer. E4: train on all sessions of one wearer, test on all of the other's.

Constant baseline. Every run is also scored for a constant predictor that outputs the training-set mean target vector for every test frame. A model that scores worse than this constant has learned nothing that transfers to the test data.

Label-free per-session standardization. Each cell is z-scored by its own session's per-cell mean and standard deviation (SD floor 1 ADC count), for the training and the test sessions separately; in E4 the training sessions are standardized one by one and then pooled. No labels from the test session are used. In the *transductive* variant the test session's statistics come from the whole test session, which a live system would not have, so it is an upper bound for this method. In the *causal* variant they come from only the first T seconds of the test session (T = 10, 30 or 60 s), and the model, the raw model and the constant baseline are all scored on the frames after T. A test session with less than 15 s left after T is skipped. Standardization was first seen to help in the May 2026 LASK5 benchmark (Appendix A) and in one exploratory run on these data. It was then fixed as described, with nothing tuned, and run once; no alternative normalizations were compared.

Blocked cross-validation. As a second check on split leakage, each session and hand is cut into five contiguous time blocks; each block is tested by a model trained on the other blocks minus the frames within 2 s of it, and the result is compared with shuffled 5-fold cross-validation on the same frames.

V3 + LASK5. The same forest and a ridge regression (alpha 1.0) map the 60 raw cells to the four piston positions. E1 and E2 are random and temporal 80/20 splits within each capture; E3a leaves one capture out of its session and

trains on the session's other captures; E3b leaves one whole session out and trains on the other six. With one wearer there is no E4. Standardization is applied per capture, transductively.

The V4 numbers come from `results.json` and the V3 numbers from `results_lask5.json`; `paper_numbers.py` prints every summary value quoted here from those two files.

5.2 V4 + Quest: the generalization ladder

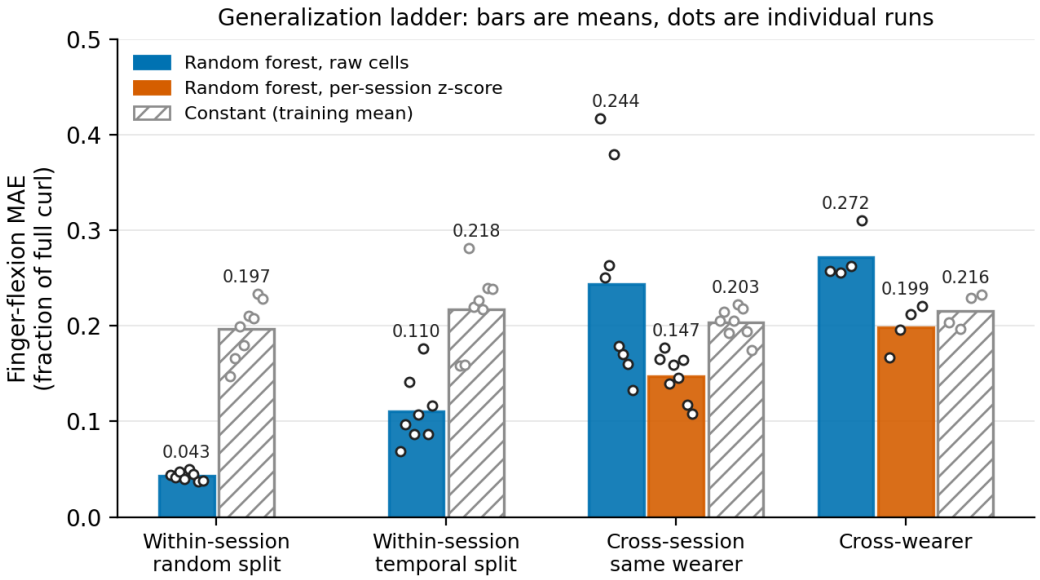


Figure 2: generalization ladder. Bars are means over runs, dots are individual runs; raw-cell forest, forest on per-session standardized cells (transductive, E3 and E4 only), and the constant baseline.

Mean finger-flexion MAE (fraction of a full curl) over runs, with the mean per-joint angular MAE in parentheses:

Protocol	Runs	Model	Constant	Model range	Model flexion R^2
E1 within-session, random	8	**0.043** (4.9 deg)	0.197 (20.9 deg)	0.037 to 0.050	0.84 to 0.91
E2 within-session, temporal	8	**0.110** (12.7 deg)	0.218 (23.0 deg)	0.069 to 0.176	-2.37 to 0.77; negative in 2
E3 cross-session, same wearer	8	**0.244** (25.7 deg)	0.203 (21.9 deg)	0.133 to 0.417	-2.34 to 0.35; negative in 4
E3, W1 (five days apart)	4	0.327	0.205	0.251 to 0.417	
E3, W2 (30 minutes apart)	4	0.160	0.202	0.133 to 0.179	
E4 cross-wearer	4	**0.272** (28.5 deg)	0.216 (23.0 deg)	0.256 to 0.310	-0.96 to -0.35; all negative

(1) Random splits overstate within-session skill. E1 and E2 use the same data and model; only the split changes, and the error rises 2.6 times (0.043 to 0.110). Blocked 5-fold cross-validation with a 2 s guard gap gives 0.104 against 0.043 for shuffled 5-fold on the same frames, a ratio of 2.39. At 17-19 Hz adjacent frames are near-duplicates, so a random split places near-copies of most test frames in the training set, a form of leakage [20]. The factor is specific to this sampling rate and these movement speeds; it is not a constant to apply to other systems. The negative temporal-split R^2 values are both hands of W2-s1, the 49 s session (Section 5.5).

(2) Without calibration, the frame-wise model does not transfer across a re-donning days apart or across wearers. Five days and a re-donning apart, W1's error (0.327) is worse than the constant's (0.205) in every one of the four runs; across wearers the model (0.272) is also worse than the constant (0.216). W2's second session, 30 minutes and a re-donning after the first, kept some skill (0.160 against 0.202), though far less than within a session. This is the same class of failure the sEMG literature reports under electrode shift and between users [2, 3]; Sections

5.3 and 5.4 look at part of its cause.

(3) Within-session error against label accuracy. The temporal-split error of 12.7 deg per joint (7.8 to 18.7 deg over runs) should be read against the joint-angle error of the labels themselves: about 9.6 deg was reported for Quest 2 hand tracking [21], a different headset from the Quest 3S used here. The 4.9 deg random-split figure is below the label error reported for that headset; it reflects near-duplicate neighbouring frames more than finger tracking.

Label lag. Shifting the labels by -4 to +4 rows (about 54 ms per row) on the E2 split gives its lowest error at +1 row, 0.107 against 0.110 unshifted, a negligible difference.

5.3 What changes between sessions

The constant baseline does not depend on the sensor, so the failure in (2) lies in the mapping from cells to pose. One visible part of it is the static pressure profile of each band. For each band and session we take the per-cell mean and SD of the 60 raw cells over the session's frames and compare sessions:

Sessions compared	Context	Band	r(cell mean)	r(cell SD)	Median shift of cell mean (counts)	Median within-session SD (counts)
W2-s1, W2-s2	same wearer, about 30 min	right	0.94	0.72	23.4	21.9 / 21.6
W2-s1, W2-s2	same wearer, about 30 min	left	0.97	0.82	24.3	21.0 / 24.1
W1-s1, W1-s2	same wearer, 5 days	right	0.63	0.39	57.3	24.7 / 30.5
W1-s1, W1-s2	same wearer, 5 days	left	0.54	0.05	58.3	23.6 / 21.0
W2-s2, W1-s2	different wearers, same day	right	0.20	0.39	91.1	21.6 / 30.5
W2-s2, W1-s2	different wearers, same day	left	0.77	0.52	32.9	24.1 / 21.0

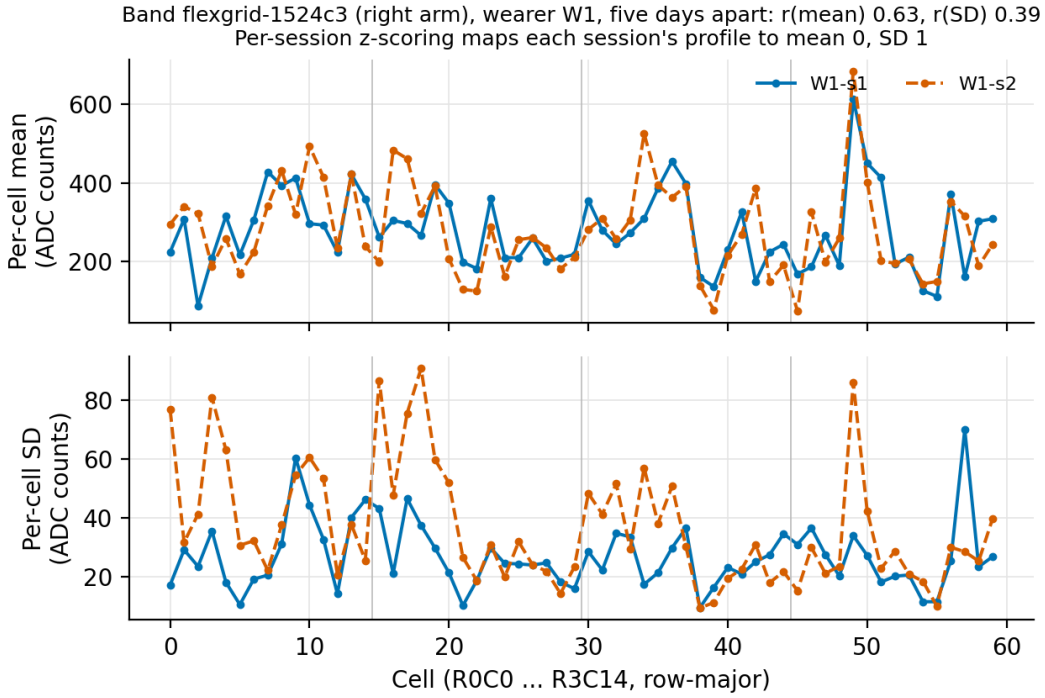


Figure 3: per-cell mean and SD of the right band (flexgrid-1524c3) in W1's two sessions, five days apart.

Thirty minutes apart, W2's profiles correlate at 0.94 to 0.97 and the median cell mean moves by about 24 counts, close to the within-session SD of a cell (21 to 24 counts). Five days apart, W1's profiles correlate at only 0.54 to 0.63 and the median cell mean moves by about 58 counts, about twice the within-session SD (21 to 31 counts); the per-cell SD pattern also changes (Figure 3). A model that maps absolute cell values to pose meets inputs it has not seen. This five-day contrast confounds several causes that these data do not separate: band placement, elapsed time, strap tension, sensor ageing and Velostat hysteresis [10]. Across wearers on the same day and the same bands the correlation is 0.20 for the right band and 0.77 for the left.

Dead column. Dropping the four cells of the dead left-band column (Section 4.1) changes W1's left-hand E3 error from 0.340 to 0.320 for the raw model and from 0.152 to 0.161 with standardization (means of the two directions). The column does not drive the W1 result.

Band-to-arm check. Each band is routed to the hand on its own arm. To check the routing, we paired each band with each hand on time-aligned frames (E2 split). Own-arm pairing scores better than opposite-arm pairing in 3 of the 4 sessions (0.077 against 0.101, 0.111 against 0.146, 0.100 against 0.110, and 0.139 against 0.138 in W2-s2). Because the hands moved largely in mirror (Section 4.4), opposite-arm pairing is expected to work fairly well too, so the check is suggestive only.

5.4 Label-free standardization

Mean flexion MAE with standardized cells, with the raw-cell model and the constant baseline scored on the same frames. The transductive rows use the whole test session; the causal rows use the first T seconds of the test session for the statistics and score the frames after T.

Protocol	Calibration	Runs	Standardized	Raw model	Constant
E3, same wearer	whole session	8	**0.147**	0.244	0.203
E3, same wearer	first 10 s	8	0.179	0.246	0.210
E3, same wearer	first 30 s	8	**0.169**	0.250	0.223
E3, same wearer	first 60 s	6	0.172	0.274	0.220
E3, W1 (five days)	whole session	4	0.160	0.327	0.205
E3, W1 (five days)	first 30 s	4	0.190	0.327	0.209
E3, W2 (30 minutes)	whole session	4	0.134	0.160	0.202
E3, W2 (30 minutes)	first 30 s	4	0.149	0.173	0.236
E4, cross-wearer	whole session	4	**0.199**	0.272	0.216
E4, cross-wearer	first 10 s	4	0.217	0.275	0.218
E4, cross-wearer	first 30 s	4	0.196	0.279	0.221
E4, cross-wearer	first 60 s	4	0.187	0.284	0.218

At 60 s the 49 s session W2-s1 has too little data left and is skipped: E3 drops to 6 runs, and the two E4 runs that test on W2 are scored on W2-s2 alone.

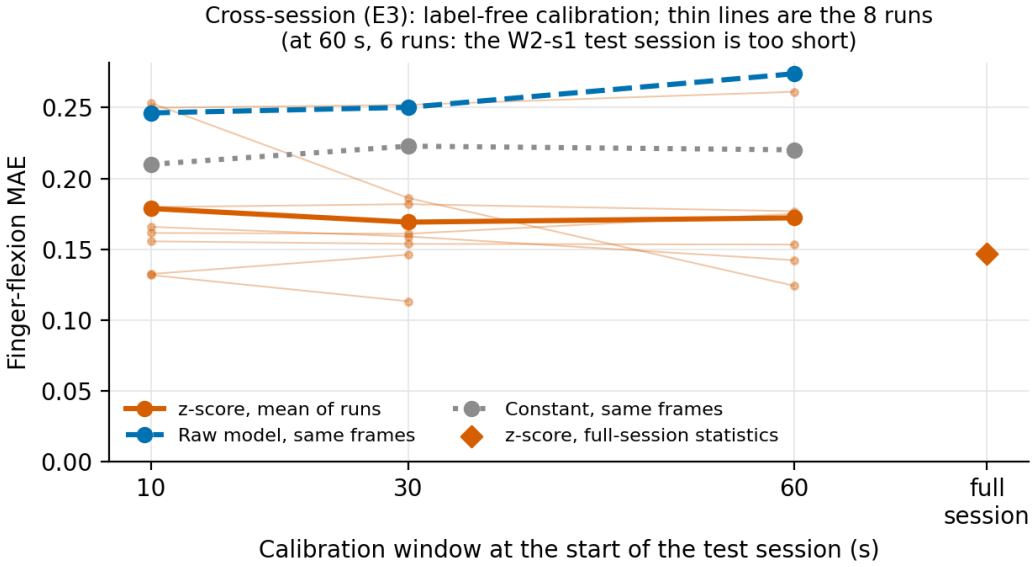


Figure 4: E3 flexion MAE against the length of the calibration window at the start of the test session, with the raw model and the constant baseline on the same frames.

Same wearer. Standardization removes the worse-than-constant failure and recovers part of the within-session skill. With the whole test session (transductive, an upper bound) the same-wearer error is 0.147, against 0.203 for the constant and 0.110 within a session; all eight runs fall between 0.108 and 0.177, below their constants. With 30 s of unlabeled calibration data the error is 0.169 against 0.223 for the constant on the same frames. For W1, five days apart, the error drops from 0.327 raw to 0.160 transductive and 0.190 with 30 s of calibration, against a constant of about 0.21. For W2, 30 minutes apart, the gain is smaller because the raw model already transferred (0.160 to 0.134). With 30 s of calibration the error sits about midway between the constant and the within-session error. Standardization removes an offset and a scale per cell; it cannot undo a change in which cells respond to which finger.

Across wearers. Standardization lifts the cross-wearer model from 0.272 to 0.199 transductive and 0.187 to 0.217 causal, against constants of 0.216 to 0.221: roughly the constant level, and no better than it by a margin that matters at this sample size. The model on standardized cells explains none of the cross-wearer variance (mean flexion R^2 -0.04, transductive).

Same wearer against cross-wearer. After standardization, same-wearer transfer (0.147) is better than cross-wearer transfer (0.199), the reverse of the raw ordering for W1. With two wearers and one contrast each, this ordering is an observation, not a finding.

5.5 Per-finger breakdown, session length, and an example trace

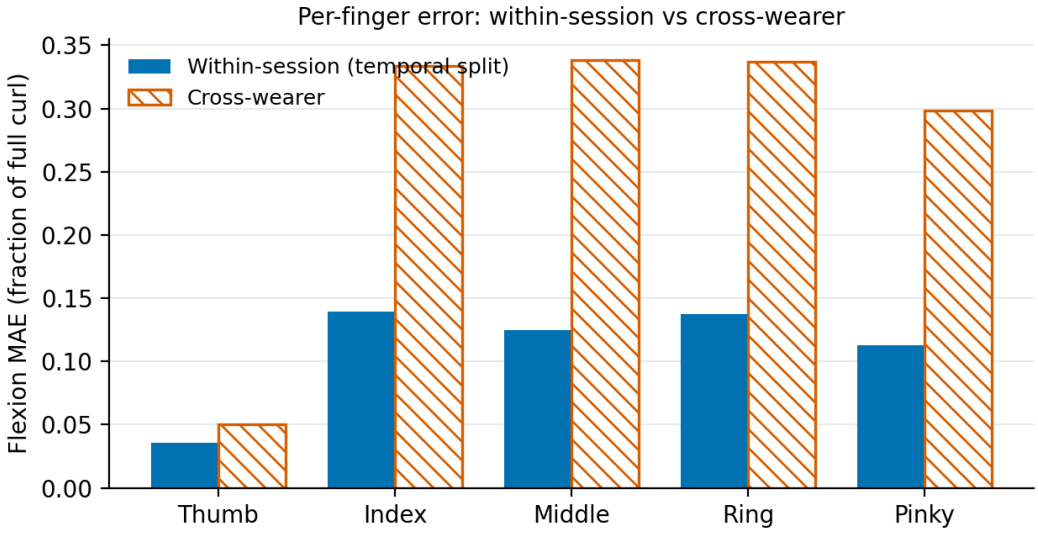


Figure 5: per-finger flexion MAE, within-session temporal split against cross-wearer.

Within a session (temporal split), the four long fingers land between 0.113 and 0.140 flexion MAE, and under cross-wearer transfer between 0.298 and 0.338; no long finger transfers markedly better than the others. The thumb scores 0.036 and 0.050, but this reflects a target-variance effect rather than better thumb decoding: flexion is the mean of the normalized MCP and IP bone angles, and the thumb's target moves much less than the long fingers' (standard deviation 0.081 over the analysis set, against 0.270 to 0.304), so even a weak predictor scores a low MAE. Per-finger errors should be read against that spread, which `paper_numbers.py` prints.

The shortest session (W2-s1, 49 s per hand) gives the worst temporal-split scores (flexion R^2 -0.19 left, -2.37 right) although its random-split scores (0.87 and 0.88) match the other sessions. With less than a minute of data, the last 20% of the timeline contains poses the first 80% never visited.

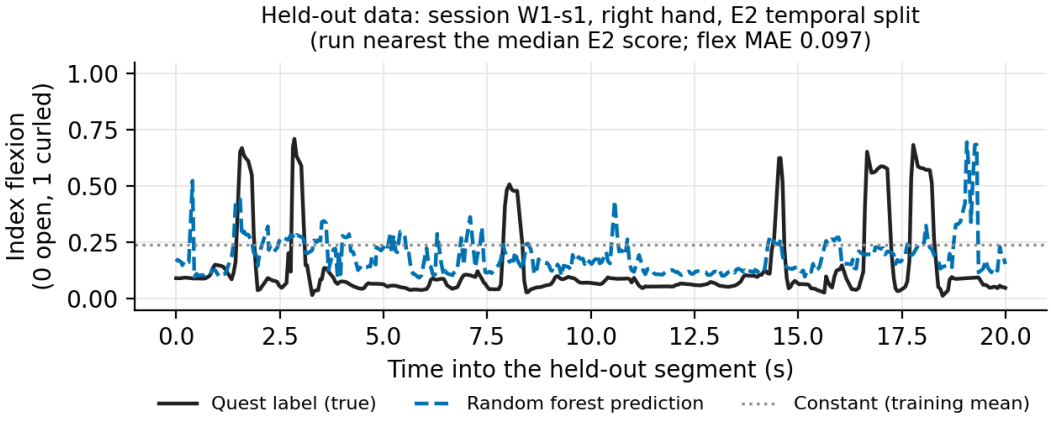


Figure 6: held-out index-finger flexion for the E2 run nearest the median E2 score (W1-s1, right hand), with the forest's prediction and the constant baseline.

Figure 6 shows 20 s of held-out data from the E2 run nearest the median (W1-s1 right hand, flexion MAE 0.097 against a median of 0.102). The prediction follows the resting level and some of the curls, but it reaches only part of

the height of most peaks and adds spikes where the finger did not move.

5.6 V3 + LASK5: the ladder on piston labels

Piston-position R^2 (mean and median over runs) and MAE (piston travel, 0 to 1), with the constant baseline. Medians are given because in E2 one outlier capture dominates the mean (see below).

Protocol (runs)	Forest R^2 mean / median	Forest MAE	Standardized forest: R^2 mean / median, MAE	Constant: R^2 mean / median, MAE	Ridge R^2 (median)
E1 random, within capture (20)	0.93 / 0.93	0.039		0.00 / 0.00, 0.233	0.75
E2 temporal, within capture (20)	-1.79 / 0.55; negative in 3	0.126		-1.54 / -0.04, 0.240	0.29
E3a cross-capture, same session (17)	0.38 / 0.45; negative in 1	0.151	0.53 / 0.56, 0.123	-0.02 / -0.01, 0.240	0.21
E3b cross-session (7)	-0.25 / -0.15; all negative	0.268	0.03 / 0.08, 0.212	-0.01 / -0.01, 0.231	-4.43

The ladder has the same shape as on the V4 data. Under the temporal split the forest's MAE rises 3.3 times over the random split (0.039 to 0.126), and with a whole session held out it is 6.9 times higher (0.268). Within a capture and across captures of one session the forest does better than the constant (R^2 about 0); with a whole session held out it does worse in all seven runs, as does ridge. Capture-level standardization raises the cross-capture forest from a mean R^2 of 0.38 to 0.53, and lifts the cross-session forest from -0.25 to 0.03 (median 0.08, range -0.23 to 0.21, negative in 2 of 7), that is, to about the constant level and no further.

The E2 mean is not a useful summary: in one capture (`capture_1779653145`) piston 4 barely moves in the held-out final 20%, which drives that piston's R^2 to -257 for ridge and makes the constant baseline's mean R^2 -1.54. The medians and the MAE are robust to it.

5.7 Loaded versus free-air finger motion

The author's experience is that the LASK5-labeled recordings are the strongest the project has made, and the proposed mechanism is mechanical: with VR labels the fingers curl in free air, whereas the LASK5 pistons are spring-loaded, so the forearm muscles contract more and the pressure signal should be stronger. The two labelers' R^2 values cannot test this, because their targets differ, so we compare the sensor signal of the one V3 band that recorded under both labelers on the right forearm. Per capture we take the median over the 60 cells of the per-cell SD, the median per-cell dynamic range (99th minus 1st percentile), and the fraction of cells whose SD reaches 5 ADC counts (p values: two-sided Mann-Whitney U over captures).

V3 band, LASK5 against Quest	Per-cell SD (counts)	Dynamic range (counts)	Cells moving
All LASK5 (20) vs Quest (10)	26.8 vs 15.6 (1.72x, p = 0.003)	121 vs 67 (1.80x, p = 0.005)	1.00 vs 1.00
Late-May LASK5, L5 to L7 (10) vs Quest (10)	17.5 vs 15.6 (1.12x, p = 0.16)	74 vs 67 (1.10x, p = 0.31)	1.00 vs 1.00
Same day (1 vs 1)	16.1 vs 16.8	71 vs 82	1.00 vs 1.00

Over all captures the LASK5 recordings vary about 1.7 times more per cell, but the difference is carried by the four mid-May sessions (L1 to L4), whose per-cell SD (31 to 47 counts per capture) exceeds that of every later capture under either labeler (16 to 22 for L5 to L7, 11 to 27 for Quest). Against the LASK5 sessions nearest in date to the Quest recordings the ratio is 1.1 and not significant, and in the one same-day pair the Quest capture varies slightly more. The mean pressure level does not differ (268 against 279 counts overall). These data neither confirm nor rule out the hypothesis: any labeler effect is smaller than the session-to-session variation of a single band, and the movements performed with each labeler were not recorded. A direct test would alternate LASK5 and free-air blocks within one session on a band left in place.

6. Limitations

Wearers. There are two wearers: the author, an experienced user of the system, and an adult member of his family. Cross-wearer results rest on one pair of wearers in each direction, and both have intact hands, although prosthetic control is part of the motivation. Nothing here says how the system behaves for people with limb difference.

Movement and design. No movement routine was recorded, the two hands moved largely in mirror, and sessions spanned from 49 s to 777 s on the hub clock. The two same-wearer contrasts differ in wearer and in elapsed time (30 minutes against five days), so the five-day contrast confounds placement, elapsed time, strap tension and sensor state, and band placement was not measured in any session. Everything was recorded in one lab.

Labels. VR labels come from consumer headset hand tracking, itself an estimator whose error on this headset was not measured; about 9.6 deg was reported for the Quest 2 [21]. Label error sets a floor under every number here. LASK5 labels measure piston travel, not joint angles, so the two labelers' errors are not on a common scale.

Sensor. The bands sample at 17-19 Hz, which limits the movement speeds that can be followed and sets the size of the split-leakage factor. Velostat hysteresis and nonlinearity [10], and any drift of these bands over days, were not characterized.

Statistics. Runs share training data and sessions (each E3 pair is run in both directions, for both hands), so they are not independent, and we give no confidence intervals. Ranges over runs are reported instead.

Standardization. The transductive variant uses statistics from the whole test session and is an upper bound; only the causal variant could run live, and it was tested with T up to 60 s on test sessions as short as 49 s.

Models. The baselines are frame-wise. Temporal models and supervised per-session calibration may change every rung of the ladder; flexion uses nominal joint ranges rather than per-wearer calibration, and the band IMU is recorded but not used.

7. Conclusion

On an open 60-cell forearm FMG system, a frame-wise model tracks finger flexion within a session at 0.110 of a full curl on held-out time (12.7 deg per joint), while a random split of adjacent frames reports 2.4 to 2.6 times less error. Without calibration the same model is worse than predicting the average pose after a re-donning five days later and across wearers. Per-cell offset and scale change between sessions, and label-free standardization removes the worse-than-constant failure for the same wearer and recovers part of the within-session skill, while across wearers it reaches about the constant level. The data and the code that produce these numbers are public, so that the session and wearer shift can be studied with better models and more people.

8. Data and Code Availability

- **Dataset:** https://huggingface.co/datasets/openmuscle/openmuscle-fmg, release 0.1, DOI 10.57967/hf/11416 (https://doi.org/10.57967/hf/11416), licensed CC BY 4.0. The analysis is pinned to revision `8e4fa1fdaec38188476ad5035b8e9f8c853cb1d5` (`data_source.py`, `HF_REVISION`). Wearers are identified only as P01 (W1) and P02 (W2).
- **Earlier captures** (FlexGrid V3 and earlier prototypes): https://github.com/Open-Muscle/OpenMuscle-Dataset, CC BY 4.0.
- **Software:** https://github.com/Open-Muscle/OpenMuscle-Software, MIT license. Capture studio, training, inference, and the analysis for this report (`analysis/paper/`, with `REPRODUCE.md`; `analysis/anniversary/`; `analysis/controller/`).
- **Hardware:** https://github.com/Open-Muscle/OpenMuscle-FlexGrid, CERN-OHL-S-2.0 (KiCad sources, Gerbers, BOMs and design PDFs for every FlexGrid revision). Band firmware:

https://github.com/Open-Muscle/FlexGridV4-Firmware, MIT license.

- **How to cite:** Moghadam, T. (2026). OpenMuscle: An Open Forearm Force-Myography Dataset with Continuous Finger Labels, and Baselines Across Sessions and Wearers. OpenMuscle technical report v0.3. https://openmuscle.org/research. DOI pending. The repository's `CITATION.cff` carries the same reference.

9. Reproducibility

Every table and figure in Sections 4 and 5 is produced from the public release by the scripts in `analysis/paper/`; `REPRODUCE.md` gives the full procedure. With Python 3.12 and git (Windows form; `REPRODUCE.md` also gives the Linux and macOS form):

```
git clone https://github.com/Open-Muscle/OpenMuscle-Software.git
cd OpenMuscle-Software
py -3.12 -m venv .venv-paper
.venv-paper\Scripts\python -m pip install -r analysis/paper/requirements.txt
.venv-paper\Scripts\python analysis/paper/analyze.py --source hf --out results_check.json
.venv-paper\Scripts\python analysis/paper/analyze_lask5.py --source hf --out results_lask5_check.json
.venv-paper\Scripts\python analysis/paper/paper_numbers.py
.venv-paper\Scripts\python analysis/paper/figures.py
```

`requirements.txt` pins the package versions used (NumPy 2.2.6, pandas 3.0.1, scikit-learn 1.8.0 among them). `render_pdf.py`, which builds this PDF from `paper.md`, additionally needs reportlab. The forests use a fixed seed, so the same versions give the same numbers. Figures 2 to 6 are the files `fig1` to `fig5` written by `figures.py` (Figure 2 is `fig1_generalization_ladder.png`, Figure 3 `fig5_baseline_shift.png`, Figure 4 `fig4_calibration.png`, Figure 5 `fig2_per_finger.png`, Figure 6 `fig3_trace.png`).

In a clean-room run (fresh clone, fresh virtual environment, empty Hugging Face cache, Windows 11) the data came down anonymously, about 50 MB, and the whole procedure took about 12 minutes. The recomputed `results.json` and `results_lask5.json` matched the committed files to within 1e-13, and `figures.py` rewrote all five figures byte-identically.

Appendix A comes from `analysis/anniversary/bench.py` and `walkforward.py` (each takes `--source hf`), and Appendix B from `analysis/controller/split_analysis.py` on a capture that is not in the release.

Ethics statement

The two wearers are the author and an adult member of his family, who took part voluntarily and agreed to the pseudonymised release of their recordings. No institutional ethics review was sought, because the work is independent and self-funded and was carried out outside any institution. Section 4.4 gives the details.

AI assistance

Analysis code and drafting were done with Claude-based coding agents under the author's direction; the author is responsible for the content.

Version note

v0.3 (2026-11-09): new title (v0.2 was "OpenMuscle: Forearm Force-Myography with VR Hand-Tracking Ground Truth: Dataset and Honest Baselines"). Added the constant baseline on every run, label-free per-session standardization (transductive and causal), blocked temporal cross-validation, the per-cell profile comparison between sessions, the dead-column, label-lag and band-to-arm checks, the LASK5 constant baseline and capture-level standardization, Figures 1, 3, 4 and 6 (Figures 2 and 5 correspond to the v0.2 Figures 1 and 2; Figure 2 now also shows the constant baseline and the standardized model), the participants, movement and ethics section, the mapping to the public release, a related-work section built on verified references, and reproduction from the public release. Rewrote the abstract, introduction and conclusion around these results. **Withdrawn:** v0.2 claimed that "donning variance rivals wearer identity" because W1's five-day cross-session error (0.33) was as high as or higher than the cross-wearer error (0.27). Both raw numbers are worse than the constant baseline (0.205 and 0.216), so neither ordering of two failed models carries information about donning or identity; after standardization the ordering reverses (Section 5.4). Also corrected: the dead-column description (three cells read zero in every analysed W1-s1 row, the fourth in 92%), and removed the statement that no cell reached ADC full scale, which no released script computes. The thumb target spread is now computed by `analyze.py`. The first four rows of Appendix A were re-run from the public release.

v0.2 (2026-11-03): every number re-checked against `results.json` and the raw captures. Corrected the W2-s1 recording time (afternoon, not morning), the band frame rate (17-19 Hz, not 18-20 Hz), the data-quality statement (one dead sensor column in W1-s1, not zero flat cells), the E2 and E3 R^2 ranges, the E3 summary (mean 0.24, per-wearer 0.16 and 0.33, matching Fig 1), and the temporal-split angular error (7.8-18.7 deg, not 9-13 deg). Removed a cross-wearer error ratio from an earlier world-frame analysis that the released scripts do not reproduce. Added the LASK5 labeler, the V3 subsets, the V3 + LASK5 ladder (`analyze_lask5.py`), the labeler signal comparison, related work, the data, code, and reproducibility sections, the author line, and the AI-assistance statement.

v0.1 (2026-07-03): first draft.

Appendix A. Feature and model benchmark (LASK5 session L2, May 2026)

Session L2 (four recordings, 9,169 frames) was benchmarked in May 2026 with seven model configurations and three feature sets: raw cells (60), raw plus first differences (120), and raw plus differences plus a 5-frame rolling mean (180). Results are in `analysis/anniversary/results/bench.json` and `walkforward.json`.

Comparison	Result (R^2)
Last 20% of each recording held out, best (ridge, 180 features)	0.70 (gradient boosting and extra trees also 0.70)
Same split, raw cells: ridge / 100-tree forest	0.53 / 0.59
Shuffled split: best (extra trees, 180 features) / raw-cell forest	0.97 / 0.92
Small MLP, either split	negative in five of six runs
Train on one recording, test on the next	negative for 9 of 21 model and feature pairs (ridge, raw cells: -0.80)
Train on two or three recordings, test on the next	all 21 positive (ridge, raw cells: 0.50 and 0.50)
Leave one recording out, best mean	0.64 (gradient boosting, per-recording z-scored cells plus differences); raw cells 0.43 to 0.51
Cross-recording matrix (extra trees), off-diagonal range	raw cells -0.19 to 0.63; per-recording z-score 0.39 to 0.68

The shuffled split shows the same adjacent-frame leakage as Section 5.2. The per-recording z-score result is where standardization was first seen to help (Section 5.1). The piston labels sit at the fully closed rail in 64% to 79% of frames; R^2 on the in-motion frames alone is only slightly lower (0.62 against 0.64 for the best leave-one-out

model).

For v0.3, `bench.py --source hf` was re-run from the public release (the four L2 captures at the pinned revision). It reproduced every score in the committed `bench.json` exactly; only the timing fields differed. The run took about 25 minutes. `walkforward.py` was not re-run to completion for v0.3 (it runs for well over 20 minutes), so the last four rows of the table and the label figures below the table come from the committed `walkforward.json`, computed in May 2026 from local copies of the same four captures. The script takes `--source hf`, and the four captures load identically from the release.

Appendix B. Gamepad decode (July 2026)

In one two-band capture (54,869 rows over both bands), the labels were the state of a game controller held during play, and each band's model (a 100-tree random forest) predicted only its own hand's controls. Analog-channel R^2 was 0.25 (left) and 0.27 (right) under a random split and -0.27 and -0.08 under a temporal split (`analysis/controller/results.json`). Play was dominated by the thumbsticks, which are thumb-driven and give a weak forearm signal, while the triggers and most buttons were rarely used, leaving little to learn for the controls with the strongest expected forearm signal. We report it as a negative result and as one more instance of the gap between random and temporal splits. This capture is not part of release 0.1, so this appendix cannot be reproduced from the public data.

Appendix C. Lessons encoded in the pipeline

Two silent failure modes each cost a capture session and are now guarded in the released code:

1. **Silent model routing.** A server restart restored the *previous wearer's* models, and training runs that produced incompatible (bilateral, 120-feature) models were refused by a shape guard without any message. The wearer's entire session ran inference with someone else's model. The pipeline now stamps wearer, session and source captures into model metadata, shows the active model per hand in the UI, warns on a wearer mismatch at session start, and reports activation failures.
2. **Flat-cell captures.** A capture can look structurally complete while 50 of 60 cells never move (band not engaged). A live per-cell activity check now flags this during recording, and the affected historical file ships as a regression fixture.

References

[1] Zhen Gang Xiao, Carlo Menon. A Review of Force Myography Research and Development. Sensors 19(20):4557, 2019. https://doi.org/10.3390/s19204557

[2] A. J. Young, L. J. Hargrove, T. A. Kuiken. The Effects of Electrode Size and Orientation on the Sensitivity of Myoelectric Pattern Recognition Systems to Electrode Shift. IEEE Transactions on Biomedical Engineering 58(9):2537-2544, 2011. https://doi.org/10.1109/TBME.2011.2159216

[3] Iris Kyranou, Sethu Vijayakumar, Mustafa Suphi Erden. Causes of Performance Degradation in Non-invasive Electromyographic Pattern Recognition in Upper Limb Prostheses. Frontiers in Neurorobotics 12:58, 2018. https://doi.org/10.3389/fnbot.2018.00058

[4] Ashkan Radmand, Erik Scheme, Kevin Englehart. High-density force myography: A possible alternative for upper-limb prosthetic control. Journal of Rehabilitation Research and Development 53(4):443-456, 2016. https://doi.org/10.1682/jrrd.2015.03.0041

[5] Vikram Ravindra, Claudio Castellini. A Comparative Analysis of Three Non-Invasive Human-Machine Interfaces for the Disabled. Frontiers in Neurorobotics 8, 2014. https://doi.org/10.3389/fnbot.2014.00024

[6] Jayden Chapman, Anany Dwivedi, Minas Liarokapis. A Wearable, Open-Source, Lightweight Forcemyography Armband: On Intuitive, Robust Muscle-Machine Interfaces. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 4138-4143, 2021. https://doi.org/10.1109/IROS51168.2021.9636345

[7] Giancarlo K. Sagastume, Peyton R. Young, Jonathon S. Schofield. Development and Validation of a Low-Cost Open-Source Force Myography Band for Hand Gesture Recognition. 2025 International Conference On Rehabilitation Robotics (ICORR), pages 919-924, 2025. https://doi.org/10.1109/ICORR66766.2025.11062965

[8] Thomas Buhl Andersen, Rógvi Eliasen, Mikkel Jarlund, Bin Yang. Force myography benchmark data for hand gesture recognition and transfer learning. arXiv preprint arXiv:2007.14918, 2020.

[9] Umme Zakia, Carlo Menon. Dataset on Force Myography for Human-Robot Interactions. Data 7(11):154, 2022. https://doi.org/10.3390/data7110154

[10] Andrius Dzedzickis, Ernestas Sutinys, Vytautas Bucinskas, Urte Samukaite-Bubniene, Baltramiejus Jakstys, Arunas Ramanavicius, et al. Polyethylene-Carbon Composite (Velostat) Based Tactile Sensor. Polymers 12(12):2905, 2020. https://doi.org/10.3390/polym12122905

[11] Manfredo Atzori, Arjan Gijsberts, Claudio Castellini, Barbara Caputo, Anne-Gabrielle Mittaz Hager, Simone Elsig, et al. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. Scientific Data 1:140053, 2014. https://doi.org/10.1038/sdata.2014.53

[12] Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, et al. emg2pose: A Large and Diverse Benchmark for Surface Electromyographic Hand Pose Estimation. Advances in Neural Information Processing Systems 37 (NeurIPS 2024, Datasets and Benchmarks Track), pages 55703-55728, 2024. https://doi.org/10.52202/079017-1770 (arXiv:2412.02725)

[13] Patrick Kaifosh, Thomas R. Reardon, CTRL-labs at Reality Labs, Brian D. Allen, Chris Anderson, Sacha Arnoud, et al. A generic non-invasive neuromotor interface for human-computer interaction. Nature 645(8081):702-711, 2025. https://doi.org/10.1038/s41586-025-09255-w

[14] Peter B. Shull, Shuo Jiang, Yuhui Zhu, Xiangyang Zhu. Hand Gesture Recognition and Finger Angle Estimation via Wrist-Worn Modified Barometric Pressure Sensing. IEEE Transactions on Neural Systems and Rehabilitation Engineering 27(4):724-732, 2019. https://doi.org/10.1109/TNSRE.2019.2905658

[15] Alon Mizrahi, Avishai Sintov. TeleFMG: A Wearable Force-Myography Device for Natural Teleoperation of Multi-Finger Robotic Hands. IEEE Robotics and Automation Letters 9(3):2933-2940, 2024. https://doi.org/10.1109/LRA.2024.3362679

[16] Svetoslav Kolev, Lingni Ma, Michael Goesele, Renzo De Nardi, Jakob Engel, Richard Newcombe. Beyond Gestures: Estimating Full Hand Pose and Contact Forces from Wrist-Worn Pressure Sensor Array. arXiv preprint arXiv:2609.16518, 2026.

[17] Artem Obukhov, Mikhail Krasnyansky, Yaroslav Merkuryev, Maxim Rybachok. Development of a System for Recognising and Classifying Motor Activity to Control an Upper-Limb Exoskeleton. Applied System Innovation 8(4):114, 2025. https://doi.org/10.3390/asi8040114

[18] A. J. Young, L. J. Hargrove, T. A. Kuiken. Improving Myoelectric Pattern Recognition Robustness to Electrode Shift by Changing Interelectrode Distance and Electrode Configuration. IEEE Transactions on Biomedical Engineering 59(3):645-652, 2012. https://doi.org/10.1109/TBME.2011.2177662

[19] Xiaolong Zhai, Beth Jelfs, Rosa H. M. Chan, Chung Tin. Self-Recalibrating Surface EMG Pattern Recognition for Neuroprosthesis Control Based on Convolutional Neural Network. Frontiers in Neuroscience 11:379, 2017. https://doi.org/10.3389/fnins.2017.00379

[20] Sayash Kapoor, Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4(9):100804, 2023. https://doi.org/10.1016/j.patter.2023.100804

[21] Diar Abdlkarim, Massimiliano Di Luca, Poppy Aves, Mohamed Maaroufi, Sang-Hoon Yeo, R. Chris Miall, et al. A methodological framework to assess the accuracy of virtual reality hand-tracking systems: A case study with the Meta Quest 2. Behavior Research Methods 56(2):1052-1063, published online 2023. https://doi.org/10.3758/s13428-022-02051-8

[22] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D. Twigg, Peizhao Zhang, Jeff Petkau, et al. MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality. ACM Transactions on Graphics 39(4), 2020. https://doi.org/10.1145/3386569.3392452