OpenMuscle: Forearm Force-Myography with VR Hand-Tracking Ground Truth: Dataset and Honest Baselines

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

Report: OpenMuscle technical report v0.2, 2026-11-03 (supersedes the v0.1 draft of 2026-07-03). DOI pending.

Artifacts: open hardware (FlexGrid V4 and V3), open-source capture and analysis pipeline (OpenMuscle-Software, `analysis/paper/`), dataset (Hugging Face; see Section 10)

Abstract

Decoding finger motion from the forearm without cameras or gloves would make prosthetics, teleoperation, and low-cost VR input more practical. We present OpenMuscle, an end-to-end open system built around low-cost 60-cell forearm force-myography (FMG) bands, with two ground-truth labelers: consumer VR hand tracking (Meta Quest 3S) and the LASK5, a handheld labeler whose four spring-loaded pistons the fingers press. The main dataset pairs the current FlexGrid V4 bands with VR labels (two wearers, about 22,000 frames). A deliberately simple per-frame random-forest baseline predicting 15 wrist-relative finger degrees of freedom per hand gives a *generalization ladder* that we argue any wearable-sensing evaluation should report: a finger-flexion mean absolute error (MAE) of 0.043 (fraction of full curl; 4.9 deg per joint) under a random train/test split, 0.110 under a temporal split, 0.24 when the band is re-donned for a new session of the same wearer (0.16 and 0.33 for the two wearers), and 0.27 across wearers. Random splits therefore overstate within-session skill by about 2.6x, and donning variance rivals wearer identity: a model tested five days later on the *same* wearer degraded at least as much as one tested on a *different* wearer. A legacy subset of 20 LASK5-labeled recordings on the previous band revision (FlexGrid V3, one band, seven sessions) shows the same ladder for piston position: R^2 0.93 under a random split, median 0.55 under a temporal split, and negative in all seven cross-session tests. We hypothesised that pressing against springs loads the forearm muscles and gives a stronger pressure signal than free-air finger motion; on the same V3 band the LASK5 recordings varied 1.7x more per cell than the Quest recordings overall, but only 1.1x (not significant) when the comparison is limited to the LASK5 sessions nearest in date, so these data do not yet separate the labeler from the recording date. We release the hardware, datasets, capture pipeline, and all analysis code.

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 EMG or, as here, force myography (FMG: the pressure distribution over the forearm surface produced by muscle and tendon displacement), promises an unobtrusive alternative, but the field has two chronic problems: ground-truth labels are expensive, and reported accuracy often fails to survive a change of session or user.

OpenMuscle attacks both. For labels, we use two cheap sources: the hand-tracking stack that ships in consumer VR headsets, where a wearer simply moves their hands while a Quest 3S streams 25-joint poses (about 25 Hz) into the same clock domain as the band data, and the LASK5, a handheld labeler with four finger pistons. Both produce labeled data at zero marginal annotation cost. For honesty, we report every result under splits of increasing difficulty and release the pipeline that enforces provenance (per-session wearer, device and firmware stamping) and live capture-quality verdicts, so future datasets are clean by construction.

Contributions: (1) an open FMG system with two labelers (hardware, firmware, capture studio); (2) a two-wearer, four-session FlexGrid V4 dataset with wrist-relative canonical finger-angle labels, plus a legacy one-band FlexGrid V3 subset labeled by piston position; (3) baselines under a generalization ladder on both labelers, quantifying split leakage (about 2.6x in MAE on V4, 3.3x on V3), donning variance, and wearer specificity; (4) a first test, on one band, of whether loaded finger presses give a stronger signal than free-air motion; (5) tooling lessons (silent

model-routing failures, flat-cell duds) encoded as guards in the released pipeline.

This is version 0.2 of the report. It re-checks every number against the released results files and the raw captures, corrects several statements of the July 2026 draft (see the Version note), and adds the LASK5 subset, related work, and data, code, and reproducibility statements so that the project can be cited.

2. Related Work

Open FMG hardware. Chapman, Dwivedi, and Liarokapis presented a wearable, open-source, lightweight force-myography armband [1], and Sagastume, Young, and Schofield developed and validated a low-cost open-source FMG band for hand gesture recognition [2]. OpenMuscle shares their premise that FMG hardware can be cheap and open. It differs in regressing continuous finger angles and piston positions instead of recognizing gestures, in running two bands at once (one per forearm), and in labeling with consumer VR hand tracking or a piston labeler.

Continuous hand pose from pressure. Kolev et al. estimate full hand pose and contact forces from a wrist-worn pressure sensor array [3], the closest neighbour in goal. The OpenMuscle band sits on the forearm rather than the wrist, and this report is concerned less with the best achievable error than with how much the evaluation protocol changes the error that gets reported.

Exoskeleton control. Obukhov, Krasnyansky, Merkuryev, and Rybachok built a system for recognising and classifying motor activity to control an upper-limb exoskeleton [4]. Their related-work section describes the OpenMuscle project through an earlier 12-sensor Hall-effect bracelet; the project had no citable reference at the time, which this report is meant to provide.

We do not compare error numbers with these works: placements, targets, and evaluation splits differ, and Section 5 shows that the split alone can change a reported error by a factor of 2.6 or more. A fuller survey (sEMG cross-session and cross-user adaptation, accuracy of headset hand tracking, leakage in time-series evaluation) is left for a later version.

3. System

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. 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 and is the current validated revision; 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 (24-25 Hz in the LASK5 captures). Schematics, layout, firmware, and the communication protocol are open (Section 10).

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).

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 memorize where the wearer stood). All VR-labeled experiments here regress 15 wrist-relative angular DOFs 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). The released pipeline writes these canonical labels at capture time; for this report they are re-derived uniformly from the raw labels across all sessions, including sessions predating that feature. LASK5 experiments regress the four piston positions directly.

4. Dataset

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

4.1 FlexGrid V4 + Quest (main subset)

Two wearers (W1, W2), four sessions, eight two-hand captures, all recorded with both bands and two-hand labeling, about 10 minutes of recording per hand (603 s summed over captures). Rows without a full 25-joint label are dropped by the analysis; 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.

Excluded (documented in the data audit and in `analyze.py`): a 9.7 s single-band fragment from the W2-s1 session and a W2-s2 capture with a 29.5 s hand-tracking dropout (23% of the take). Among the retained captures, no cell reached ADC full scale (the largest raw value in the set is 850 of 4095), and labels are fully populated and varying. One sensor column (C3, four cells) on the left band read zero for most of W1-s1: two of its cells in all three captures, the other two in two of the three. No other band-session has a constant cell.

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 the right forearm where the arm is recorded. The wearer is recorded as W1 in 9 of the 20 captures and not recorded in the others. All twenty were rated publishable in the July 2026 data audit; 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 with VR labels on the same V3 band (May 25 to June 4), 11,619 frames, publishable in the audit and at least 500 rows (about 20 s) each; shorter fragments are left out. No wearer is recorded in their metadata. They are used only for the labeler comparison in Section 5.6.

5. Experiments

5.1 Models and protocols

V4 + Quest. One random forest (100 trees) per hand per condition, mapping the 60 raw cell values of a single frame to the 15 canonical targets. No temporal context, no calibration, no normalization: this is a floor, chosen 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 of one wearer, test on all of the other. These numbers come from `results.json`.

V3 + LASK5. The same random forest and a ridge regression (alpha 1.0), each mapping the 60 raw cells of a frame to the four piston positions. E1 and E2 are the 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. These numbers come from `results_lask5.json`, and `paper_numbers.py` prints every summary value from both files.

5.2 V4 + Quest: the generalization ladder

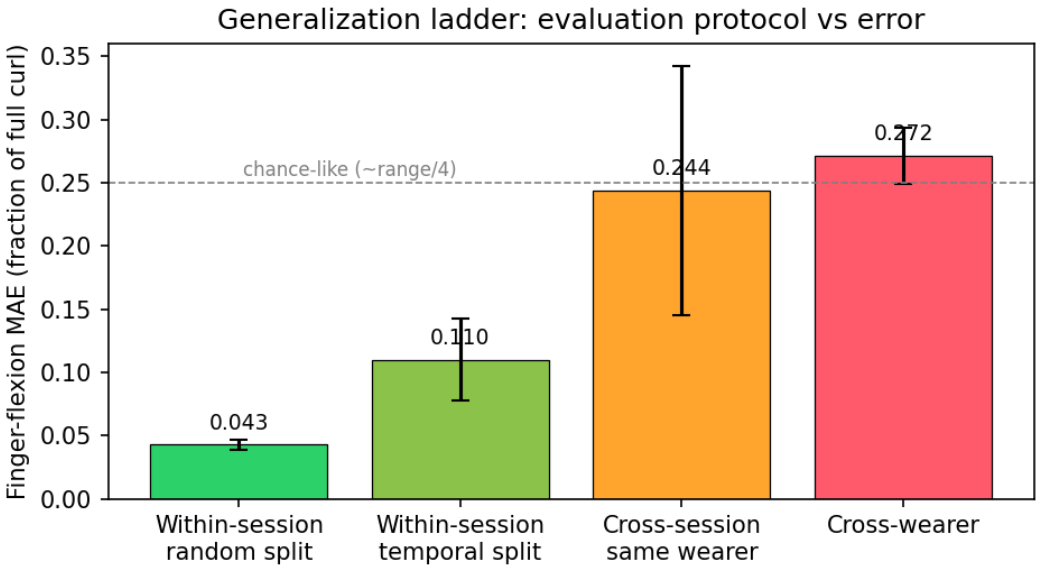


Fig 1: generalization ladder (bars: mean flexion MAE; error bars: one standard deviation over sessions and hands)

Mean finger-flexion MAE (fraction of full curl; mean over sessions and hands, 8 runs per protocol for E1 to E3 and 4 for E4), with the mean per-joint angular MAE in parentheses:

Protocol	Flexion MAE	Angular MAE (range)	Flexion R^2 (range)
E1 within-session, random split	**0.043** (4.9 deg)	4.1-6.0 deg	0.84 to 0.91
E2 within-session, temporal split	**0.110** (12.7 deg)	7.8-18.7 deg	0.21 to 0.77 in six runs; negative for both hands of W2-s1

Protocol	Flexion MAE	Angular MAE (range)	Flexion R^2 (range)
E3 cross-session, same wearer	**0.24** (25.7 deg); W2 0.16, W1 0.33	16.0-42.0 deg	-2.34 to 0.35; negative in 4 of 8
E4 cross-wearer	**0.27** (28.5 deg)	27.1-31.9 deg	-0.96 to -0.35 (all negative)

Three observations:

(1) Random splits flatter the model about 2.6x. E1 and E2 use identical data and model; only the split changes. At 17-19 Hz, adjacent frames are near-duplicates, so a random split places near-copies of most test frames in the training set. Reported wearable-sensing accuracy that relies on random splits should be read with this factor in mind.

(2) Donning variance rivals wearer identity. A second session by the same wearer on the same day (W2-s1 and W2-s2, about 30 minutes apart, band re-donned) gives 0.13-0.18 flexion MAE (mean 0.16); five days and a re-donning apart (W1-s1 and W1-s2), it is 0.25-0.42 (mean 0.33), *as bad as or worse than* transferring across wearers (0.26-0.31, mean 0.27). The dominant failure mode of this sensing modality is not "your muscles differ from mine" but "the band never sits in the same place twice." Measured against the within-session temporal baseline (0.110), cross-wearer transfer costs about 2.5x in the frame-invariant angle space used here.

(3) Within-session, FMG is usable today. A mean per-joint error of about 5 deg under the random split (4.1-6.0 deg) and about 13 deg under the temporal split (7.8-18.7 deg) supports live per-hand "ghost hand" rendering, which the released system demonstrates end-to-end (train-to-prediction in under a minute on consumer hardware).

5.3 V4 + Quest: per-finger breakdown and session length

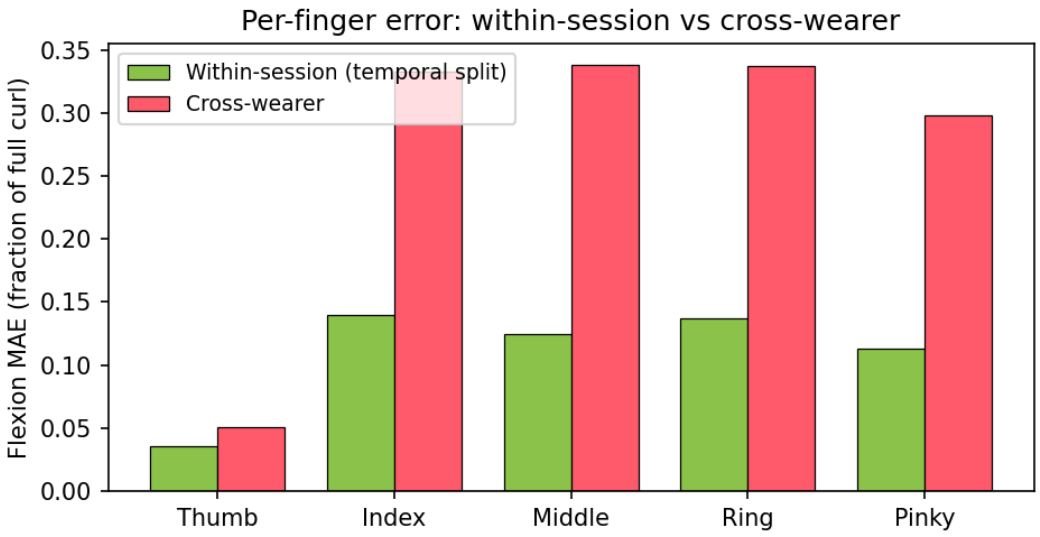


Fig 2: per-finger flexion MAE, within-session temporal split against cross-wearer

Within-session (temporal), the four long fingers land between 0.11 and 0.14 flexion MAE and rise uniformly to 0.30-0.34 under cross-wearer transfer; no long finger transfers meaningfully better than the others. The thumb *appears* dramatically easier (0.036 within, 0.050 cross-wearer), but this is largely a target-variance artifact: with flexion defined as the mean of the normalized MCP and IP bone angles, the thumb's values occupy a much narrower band than the long fingers' (standard deviation 0.08 against 0.27-0.30 over the analysis set), so even weak predictors score a low MAE on it. Per-finger errors should be read against per-finger target spread; the targets can be re-derived from the raw labels with `openmuscle.hand_angles`.

The shortest session (W2-s1, 49 s per hand) shows the worst temporal-split scores (flexion R^2 -0.19 left, -2.37 right) despite random-split scores in line with the other sessions (0.87 and 0.88). With less than a minute of data, the last 20% of the timeline contains poses the first 80% never visited. In this small set, session length appears to matter more than take count.

5.4 V3 + LASK5: the ladder on piston labels

Piston-position R^2 (median over runs, with the range) and mean MAE (piston travel, 0 to 1):

Protocol (runs)	Forest R^2	Forest MAE	Ridge R^2	Ridge MAE
E1 within-capture, random (20)	**0.93** (0.85 to 0.96)	0.039	0.75 (0.44 to 0.88)	0.109
E2 within-capture, temporal (20)	**0.55** (-45.19 to 0.83); negative in 3	0.126	0.29; negative in 4	0.191
E3a cross-capture, same session (17)	**0.45** (-0.69 to 0.73); negative in 1	0.151	0.21; negative in 5	0.275
E3b cross-session (7)	**-0.15** (-0.80 to -0.04); all negative	0.268	-4.43; all negative	0.508

The ladder has the same shape as on the V4 data. Under the temporal split the forest's MAE rises 3.3x over the random split (0.039 to 0.126; compare 2.6x for V4 flexion), and with a whole session held out it is 6.9x higher (0.268), with negative R^2 for every held-out session and both models. Across captures within a session the forest keeps a median R^2 of 0.45, so the steepest drop comes between sessions, consistent with the donning reading of Section 5.2. The four pistons behave alike under the random split (forest R^2 0.91 to 0.94 per piston). Mean R^2 is not a useful summary for the temporal split: in one capture piston 4 barely moves in the held-out final 20% (standard deviation 0.013, against 0.25 in the training span), which drives that piston's R^2 to -182 for the forest and -257 for ridge; the medians and MAE above are robust to it.

5.5 V3 + LASK5: features and models (May 2026 benchmark)

Session L2 (four recordings, 9,169 frames) was also benchmarked 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). With the last 20% of each recording held out, ridge regression on the 180-feature set reached R^2 0.70, level with histogram gradient boosting and extra trees (both 0.70). On raw cells alone, ridge reached 0.53 and a 100-tree random forest 0.59. A shuffled split of the same data raised the best score to 0.97 (extra trees) and the raw-cell forest from 0.59 to 0.92, the same adjacent-frame leakage as in Section 5.2. A small MLP failed under both splits (negative R^2 in five of its six runs).

Across recordings the picture is mixed. Trained on one recording and tested on the next, 9 of 21 model and feature pairs (ridge, extra trees, gradient boosting) score negative R^2 (ridge on raw cells: -0.80). Trained on two or three recordings and tested on the following one, all 21 are positive (ridge on raw cells: 0.50 and 0.50). Leaving one recording out in turn, the best mean R^2 is 0.64 (gradient boosting on per-recording z-scored cells plus differences), against 0.43-0.51 for raw cells. Per-recording z-scoring also narrows the cross-recording matrix (extra trees, trained on one recording and tested on another) from -0.19 to 0.63 with raw cells to 0.39 to 0.68, consistent with the donning reading: normalizing each recording cancels part of the placement shift. The piston labels sit at the fully-closed rail in 64-79% 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).

5.6 Loaded versus free-air finger motion

In the author's experience the LASK5-labeled recordings are the strongest the project has made, and his proposed mechanism is mechanical: with camera or VR labels the fingers curl in free air against no resistance, so the forearm muscles contract less, whereas the LASK5 pistons are spring-loaded. We hypothesise that loaded contraction makes

the forearm pressure signal more pronounced. The two labelers' R^2 values cannot settle this, because their targets differ, so we compare the sensor signal itself. The V3 band recorded under both labelers on the right forearm, which allows a comparison within one band and removes band-to-band differences in sensitivity (the script also normalises each band by its own median, which leaves these within-band ratios unchanged); no V4 band has LASK5 data. Per capture we take the median over the 60 cells of the per-cell standard deviation, the median per-cell dynamic range (99th minus 1st percentile), and the fraction of cells whose standard deviation 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, 12 minutes apart (1 vs 1)	16.1 vs 16.8	71 vs 82	1.00 vs 1.00

Over all captures the LASK5 recordings vary about 1.7x more per cell, as the hypothesis predicts. The difference is carried by the four mid-May sessions (L1 to L4), whose per-cell standard deviation (31 to 47 counts per capture) exceeds that of every later capture under either labeler (16 to 22 for LASK5 sessions L5 to L7, 11 to 27 for Quest). Against the LASK5 sessions nearest in date to the Quest recordings, the ratio falls to 1.1x and is not significant, and in the one same-day pair the Quest capture varies slightly more. The fraction of moving cells does not discriminate (nearly every cell passes 5 counts in every capture), and the mean pressure level, which a sustained grip on the handheld labeler might raise, does not differ (268 against 279 counts overall; ratio 1.00 in late May). These data therefore neither confirm nor rule out the hypothesis: any labeler effect is smaller than the session-to-session variation of a single band, and the two labelers were also used for different movement routines. A direct test would alternate LASK5 and free-air blocks within one session on a V4 band left in place.

6. Supplementary result: 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. The capture explains much of this: play was dominated by the thumbsticks, which are thumb-driven and give a weak forearm signal, while the triggers and most buttons were barely used (most buttons were pressed in under 1% of test frames), 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.

7. Lessons encoded in the pipeline

Two silent failure modes cost a capture session each and are now guarded in the released code, because dataset papers rarely say how the sausage was made:

1. **Silent model routing.** A server restart auto-restored the *previous wearer's* models; training runs that produced incompatible (bilateral, 120-feature) models were refused by a shape guard *silently*. The wearer's entire session ran inference with someone else's model. The pipeline now stamps wearer, session, and source captures into model metadata, badges the active model per hand in the UI, warns on wearer mismatch at session start, and surfaces activation failures loudly.
2. **Flat-cell duds.** A capture can look structurally perfect while 50 of 60 cells never move (band not engaged). A live per-cell activity verdict now flags this during recording, and the offending historical file ships as a regression fixture.

8. Limitations

Scale. The V4 subset has two wearers, four sessions, and about 22,000 frames (11,145 left, 11,146 right); the V3 LASK5 subset has one band, and W1 is the only wearer its metadata records. All of it was recorded in a single lab. Conclusions about cross-wearer transfer are directional, not definitive, and there is no V4 recording with LASK5 labels yet.

Ground truth. VR labels come from consumer headset hand tracking, which is itself an estimator. Its error is not measured here; label noise from tracking loss is mitigated by the quality verdicts and the exclusion above but not eliminated, and label error sets a floor under every number in this report. LASK5 labels measure piston travel, not joint angles, so the two labelers' errors are not on a common scale.

Donning variance. Band placement was neither measured nor controlled between sessions, so donning variance is inferred from the cross-session gap rather than isolated. The same-day contrast (W2) and the five-day contrast (W1) also differ in wearer and elapsed time.

Labeler comparison. The signal comparison in Section 5.6 is confounded by recording date and by the movement routine used with each labeler; it uses one V3 band only.

Baseline and features. The baselines are deliberately frame-wise; temporal models (TCN or GRU) and per-session calibration are expected to lift every rung of the ladder and are the obvious next experiments. Flexion normalization uses nominal joint ranges rather than per-subject calibration. Gravity-relative forearm orientation from the band IMU is captured but not yet used as a feature. One sensor column was dead for most of W1-s1 on the left band (Section 4.1).

9. Conclusion

OpenMuscle shows that consumer VR hand tracking and a cheap piston labeler turn FMG data collection from an annotation problem into an afternoon activity, and that the resulting honesty, applied to our own system, relocates the field's central challenge from decoding accuracy to *placement and session robustness*. Within a session, a 100-tree forest on 60 pressure cells tracks fifteen finger DOFs to a mean per-joint error of about 5 deg (random split) to 13 deg (temporal split), and four piston positions to a median R^2 of 0.55 under a temporal split; across a re-donning, neither holds up. Whether loaded presses give a stronger signal than free-air motion remains open. We release everything needed to reproduce, refute, or extend these numbers.

10. Data and Code Availability

- **Dataset:** https://huggingface.co/datasets/openmuscle/openmuscle-fmg (DOI 10.57967/hf/11416, https://doi.org/10.57967/hf/11416), licensed CC BY 4.0. Wearers are identified only as W1 and W2. The capture files analysed here are named in `analysis/paper/analyze.py` (V4) and `analysis/paper/analyze_lask5.py` (V3).
- **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/`, `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: Forearm Force-Myography with VR Hand-Tracking Ground Truth: Dataset and Honest Baselines. OpenMuscle technical report v0.2. https://openmuscle.org/research. DOI

pending. Each repository above carries a `CITATION.cff` with the same reference.

11. Reproducibility

The tables, figures, and results in Sections 4 and 5 (except 5.5) are produced by the scripts in `analysis/paper/` from the raw captures; a few descriptive values in the text (recorded seconds, raw value range, dead cells, target spread) come from direct inspection of the same captures. From a clone of OpenMuscle-Software, with the capture CSVs and their `meta.json` files placed in `pc/data/raw/merged/`:

```
python -m pip install numpy pandas scipy scikit-learn matplotlib reportlab pillow
python analysis/paper/analyze.py        # V4 + Quest: results.json (about 30 s)
python analysis/paper/analyze_lask5.py  # V3: results_lask5.json (about 90 s)
python analysis/paper/paper_numbers.py  # prints every summary value quoted here
python analysis/paper/figures.py        # writes fig1 and fig2 from results.json
python analysis/paper/render_pdf.py     # renders paper.md to paper.pdf
```

For v0.2, `analyze.py` was re-run on the raw captures with Python 3.12.10, scikit-learn 1.8.0, pandas 3.0.1, and NumPy 2.2.6. It reproduced the released `results.json` to floating-point precision (largest difference 3.6e-15 over 260 values), and `figures.py` reproduced both figures pixel for pixel. `analyze_lask5.py` gave identical results on two runs. The forests use a fixed random seed (42); other scikit-learn versions may differ in the last digits.

Section 5.5 comes from `analysis/anniversary/bench.py` and `analysis/anniversary/walkforward.py` (run from that folder with the four L2 recordings in its `data/` folder; outputs in `results/`), and Section 6 from `python analysis/controller/split_analysis.py <capture.csv>`.

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

v0.2 (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.

References

[1] Chapman, Dwivedi, Liarokapis. A Wearable, Open-Source, Lightweight Forcemyography Armband. IROS 2021.

[2] Sagastume, Young, Schofield. Development and Validation of a Low-Cost Open-Source Force Myography Band for Hand Gesture Recognition. ICORR 2025. https://doi.org/10.1109/ICORR66766.2025.11062965

[3] Kolev et al. Beyond Gestures: Estimating Full Hand Pose and Contact Forces from Wrist-Worn Pressure Sensor Array. arXiv:2609.16518, 2026.

[4] Obukhov, Krasnyansky, Merkuryev, 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