

Biosignal foundation models for wearables

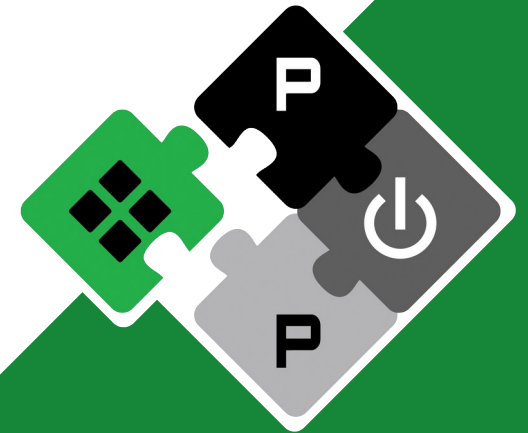
From EEG and EMG to multimodal models at the edge

Thorir Mar Ingolfsson

ARNI working group

Frontier Models for Neuroscience and Behavior

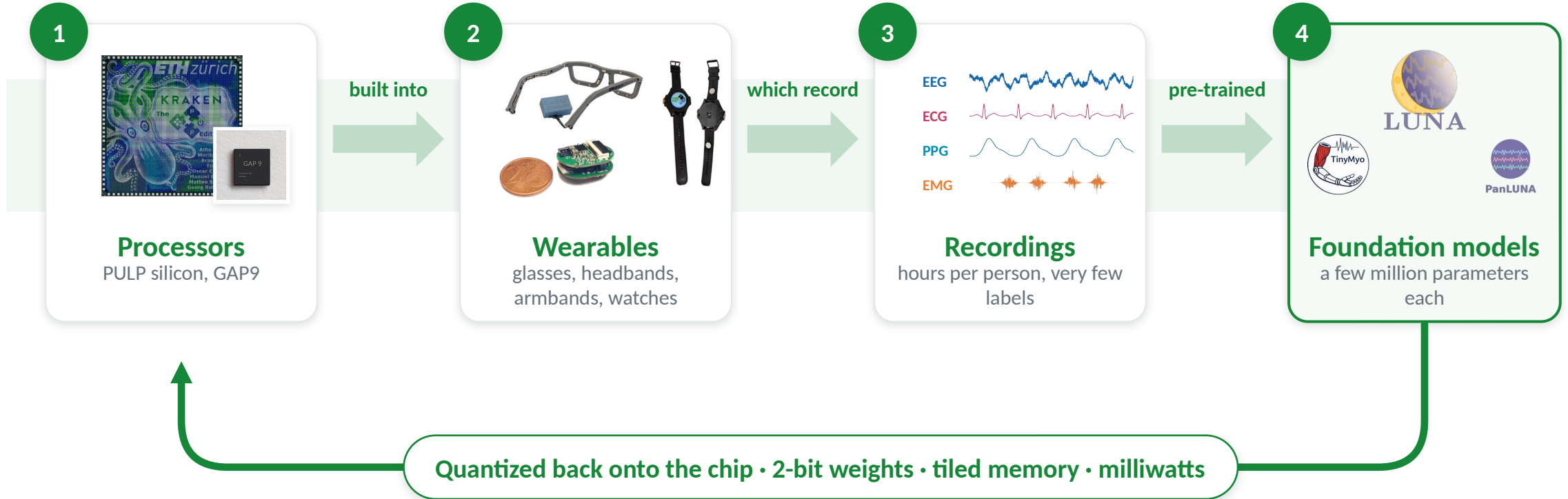
19 October 2026 • Integrated Systems Laboratory, ETH Zürich



Where we come from



Chips become wearables, wearables produce data, and the models go back onto the chip

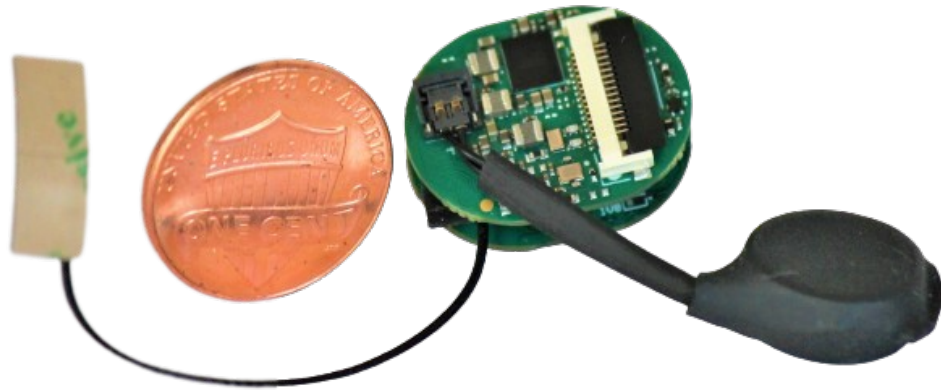


Pre-train on what the devices record, and keep the model the size of the chip.



What the wearable asks of the model

Three constraints that shaped every model in this talk



BioGAP wearable platform

1 • Electrode layouts

Useful features across recording setups

2 • Long recordings

Compute must grow at most linearly with recording length

3 • Local inference

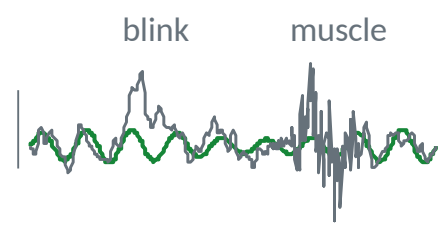
Latency and energy measured on the device



Why EEG is hard

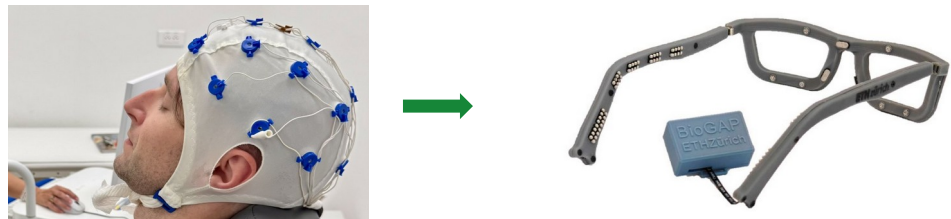
The same physiology, very different measurements

Weak signals, strong interference



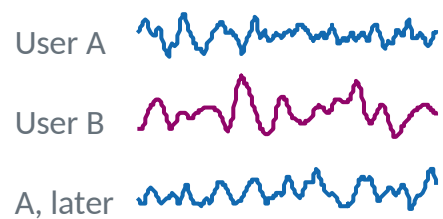
Blinks, muscle and motion can bury brain rhythms

A changing sensor interface



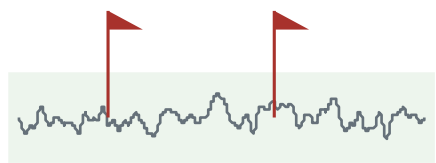
Fewer sensors, different positions and contact

People and sessions change



Signals vary across users and sessions

Few reliable labels



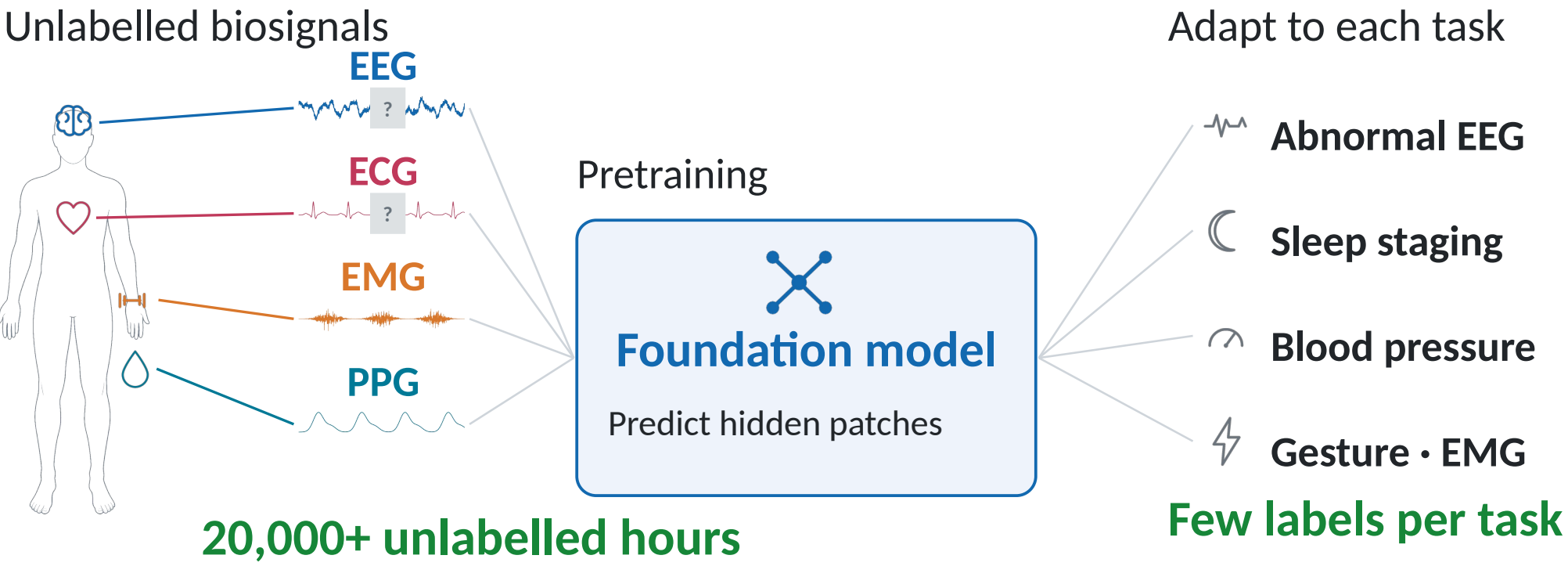
Expert labels are costly and specific to each task

Learn useful structure that survives these changes



Learning from unlabelled biosignals

Learn from unlabelled signals, then adapt with few labels



Pretrain once, then adapt each task with few labels



The model family

Each model adds a capability to the same research path

Flexible inputs



LUNA
Channel unification



PanLUNA
EEG + ECG + PPG

constraint 1: electrode layouts

Efficient encoders



FEMBA
State-space model



LuMamba
Unified SSM

constraint 2: long recordings

Beyond EEG



TinyMyo
EMG Transformer



GAP9
Wearable processor
constraint 3: local inference

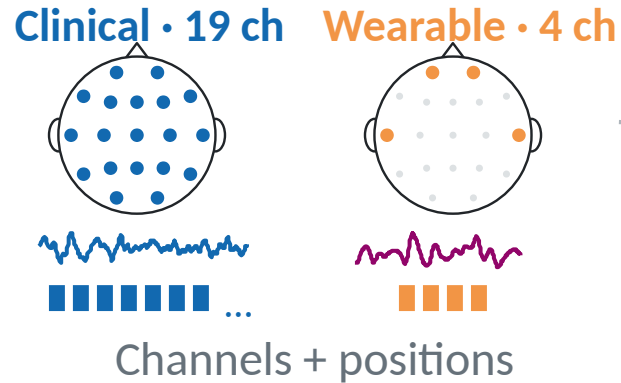
Each constraint is answered by at least one model; GAP9 is where all of them have to run



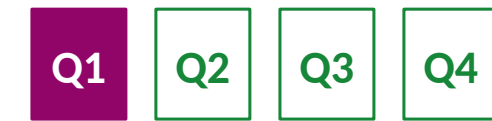
Different layouts, a shared representation

Constraint 1, electrode layouts: one query set reads varying montages

Different inputs

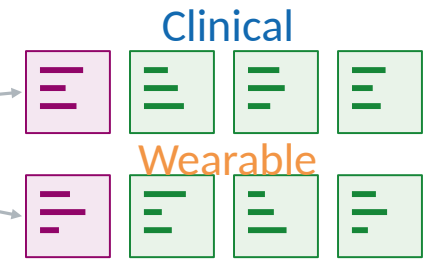


Same learned queries



Read available channels

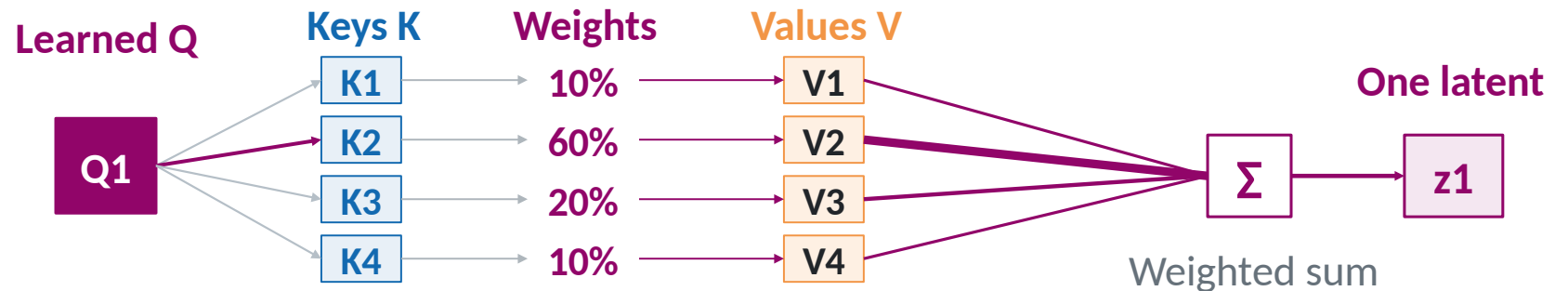
Same latent shape



4 tokens per time patch

One query, up close

Match Q with K.
Weight and sum V.

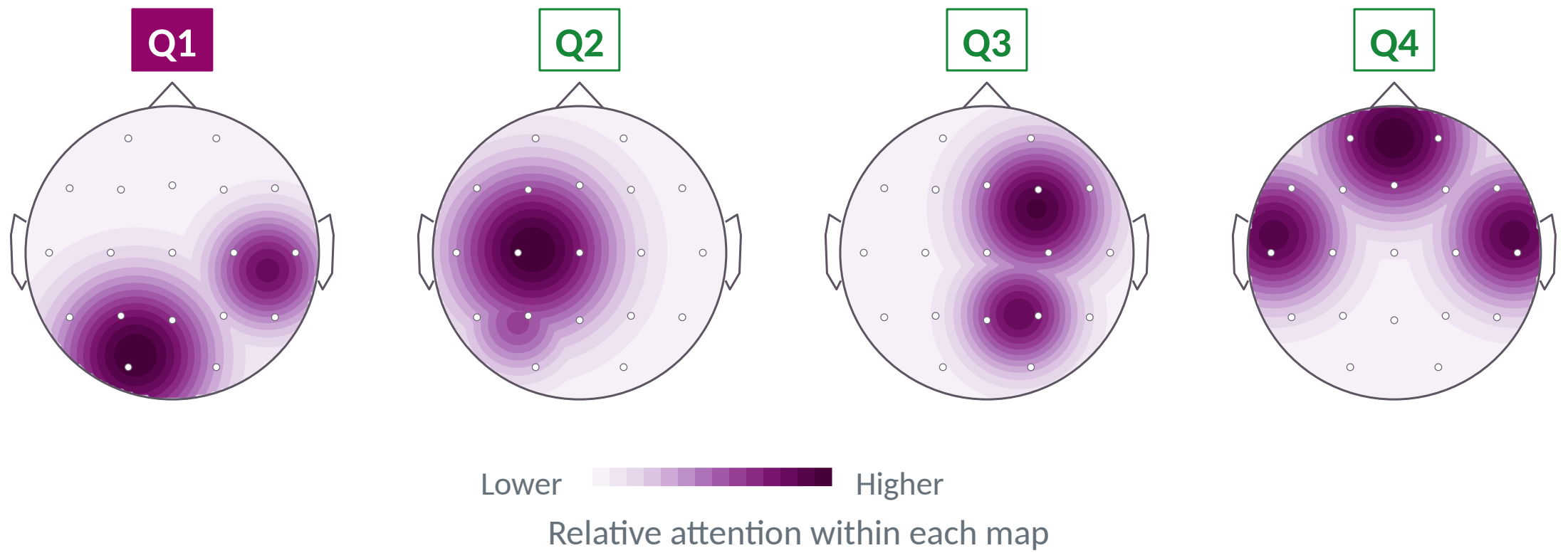


One learned query combines the channels into one latent token

Learned queries focus on different scalp regions

Each query weights the available channels differently

Illustrative attention patterns



Shared representation, different spatial preferences

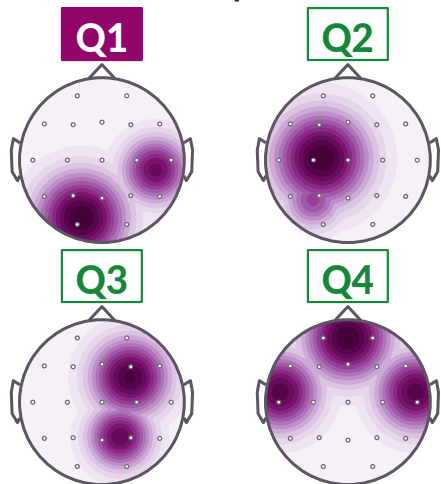
LuMamba: flexible inputs, efficient sequences



Constraint 2, long recordings: FEMBA carries time at linear cost

Spatial readouts

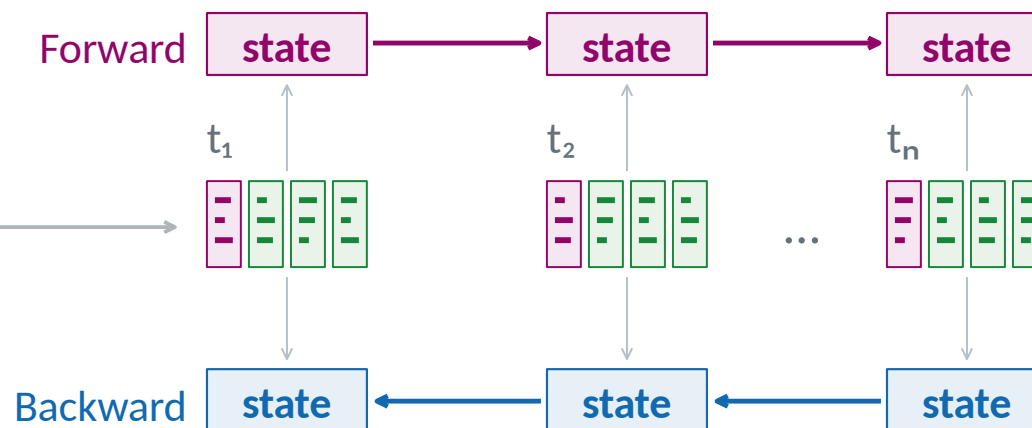
LUNA queries



Reuse queries
at each time patch

Efficient temporal encoding

FEMBA · bidirectional state-space model (SSM)



Carry state across time

Why FEMBA?

Linear compute growth
with sequence length

$O(N)$

Full attention: $O(N^2)$

26.5×
fewer FLOPs

LuMamba vs LUNA · FLOPs at the longest
window LUNA fits on one A100, batch 1

LuMamba combines LUNA and FEMBA · 4.6M parameters

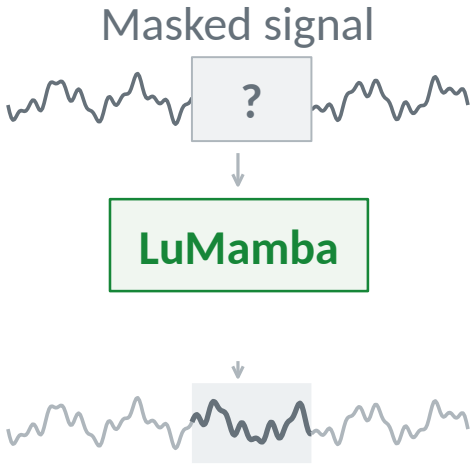


Training LuMamba for transfer

Same encoder, different objectives: waveform detail and wider context

1 Reconstruction

Recover missing signal



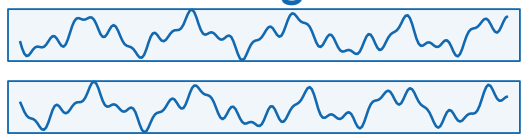
Learn waveform detail

2 LeJEPa

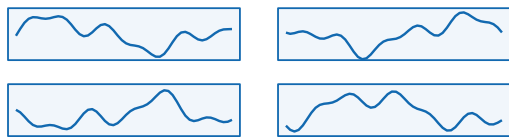
Short crops predict the mean embedding of long windows

Same EEG recording

2 GLOBAL • long windows



4 LOCAL • short crops



Shared weights

Global embeddings



Global mean



Align to mean



Local embeddings

SIGReg keeps the embeddings spread out, so all crops do not map to one point

3 Combined

Waveform detail + wider context

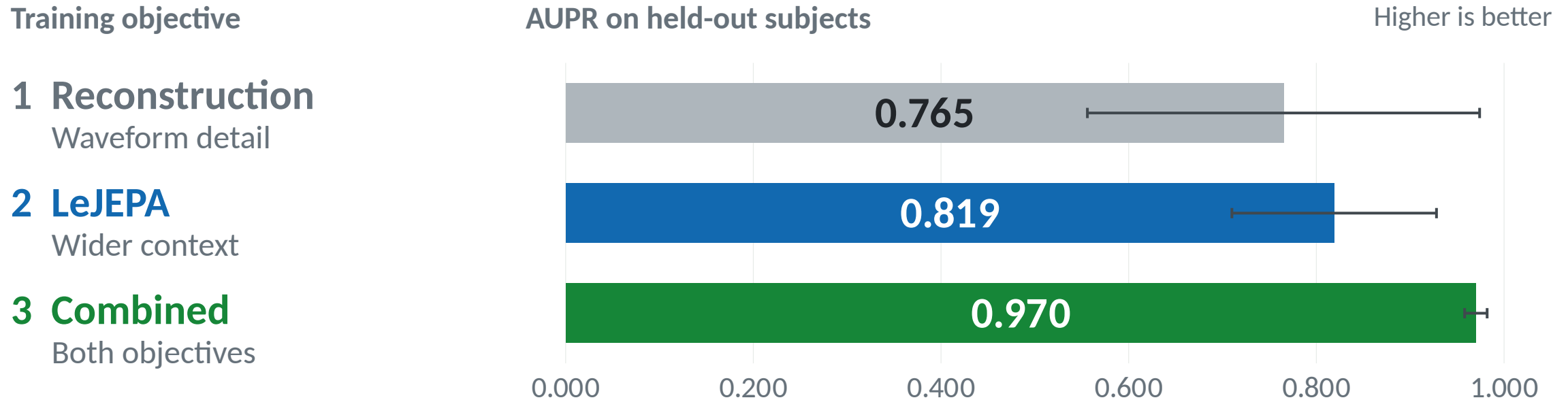
Does this improve transfer to an unseen montage?

Combined objective transfers best to a new montage



APAVA: Alzheimer's EEG on a 16-channel montage unseen in pre-training

EEG pretraining → APAVA fine-tuning → Held-out subjects



Mean $\pm$ SD over 3 seeds, 4 held-out test subjects · reconstruction alone still leads on the Temple artifact and slowing tasks (TUAR, TUSL)

+0.205 AUPR

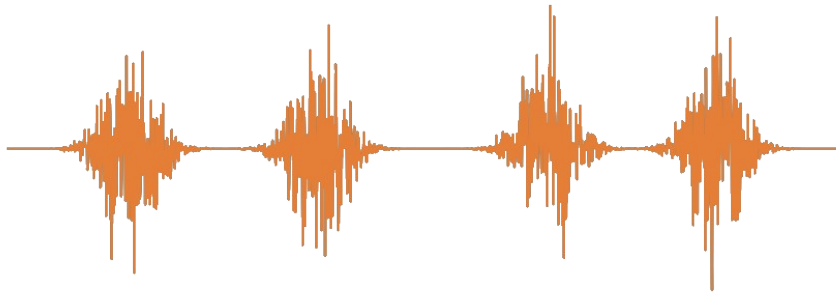
combined vs. reconstruction



Surface EMG: sensing neuromuscular activity

Forearm, wrist or face: muscle activation at the skin, with shifts the model has to absorb

One surface electrode, four bursts of muscle activity



Summed motor-unit potentials, 20 to 500 Hz, millivolt scale

What gets decoded from it

Hand gestures

Wrist and finger angles

Silent speech

Why one model is hard

Across people and days

Anatomy, placement and skin impedance reshape the waveform, and it drifts again day to day

Across hardware

8 to 16 channels, 200 Hz to 2 kHz, armband, wristband or face electrodes, dry or wet

Across tasks

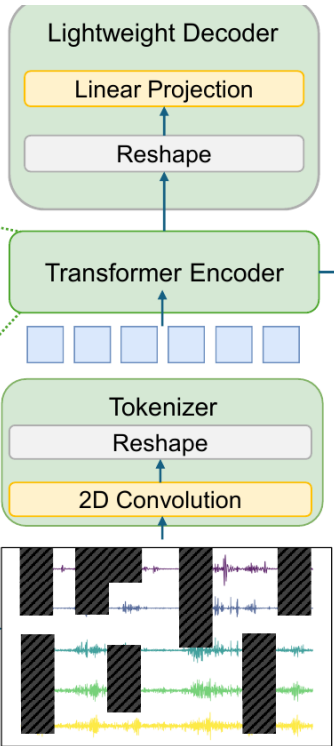
Gestures, continuous kinematics and speech ask different questions of the same signal

Same recipe as for EEG: pre-train on unlabelled sEMG, then adapt each task with a small head



TinyMyo: a 3.6M-parameter EMG encoder

Masked reconstruction on public sEMG, with channel-independent patches



■ Masked Patches

TinyMyo pre-training pipeline

1 Tokenize each channel on its own

One patch per channel and time step, one shared linear projection
Fewer channels simply means fewer tokens: no padding at fine-tuning

2 Mask half, reconstruct the waveform

50% of patches replaced by a learned token; a 3.9k-parameter linear decoder rebuilds them
A deliberately tiny decoder keeps the representation in the encoder

3 Pre-train on heterogeneous public data

NinaPro DB6 and DB7 (lab-grade forearm sensors, 12 to 14 channels) and emg2pose (193 people, 16-channel wristband), all at 2 kHz

Width 192, 3 heads • TinyMyo: 8 blocks, 3.6M parameters
TinyissimoMyo: 4 blocks, 1.9M parameters, built for latency



One encoder, four EMG tasks

Pre-trained once, adapted with a small task head • mean of five seeds

Hand gestures

88.0%

higher is better

Accuracy, NinaPro DB5: 52 gestures, 10 subjects

96.6% on EPN-612, 96.5% on UCI-EMG
WaveFormer (3.1M): 87.5%, 95.2%, 93.1%

Wristband gestures

0.146

lower is better

Classification error rate (CLER), Meta neuromotor-interface data, 100 subjects

Hand kinematics

8.8°

lower is better

Mean absolute error, NinaPro DB8, 5 degrees of freedom, one model for all subjects

Subject-specific TCN: 6.9°, one model per subject

Silent speech

33.5%

lower is better

Word error rate, speech production, Gaddy dataset

Recognition: 34.0%
54M task-specific Transformer: 28.8%

Pre-training helped most on the smallest set (DB5, +4.9 points)

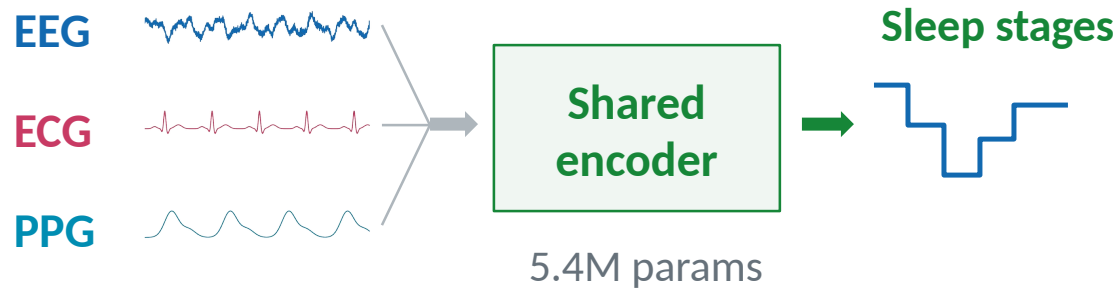


PanLUNA: one encoder for EEG, ECG and PPG

Learned queries unify channels and modalities in a single 5.4M-parameter encoder

Sensor-type tags, shared queries

Each modality is tokenized, tagged and read by the same queries



Shared features across sensor types

What the shared encoder buys

10 of 13 tasks: state of the art or competitive, single- and multi-modal

Missing modalities: handled by construction, since the queries read whatever sensors are present

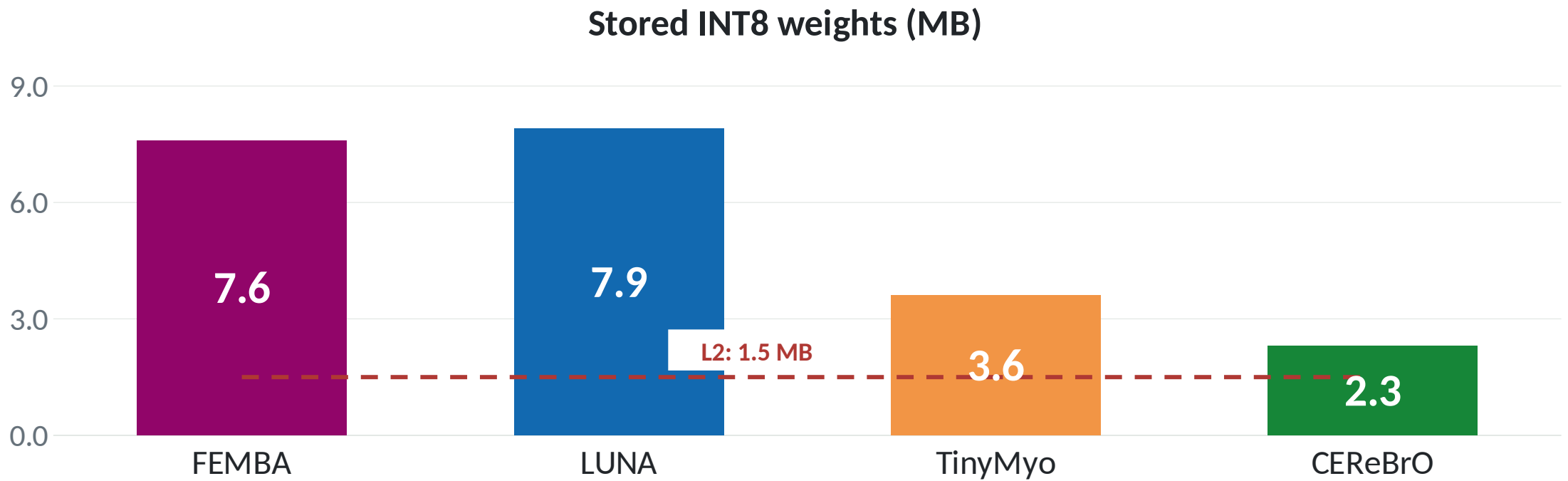
Specialists up to 57× larger: matched by one 5.4M model

On GAP9: 326 ms and 18.8 mJ per 10-second ECG window



Even compact models exceed local memory

Constraint 3, local inference: INT8 weights exceed the 1.5 MB on-chip memory (L2)



Reduce storage, then schedule what stays on chip (hide memory behind compute)

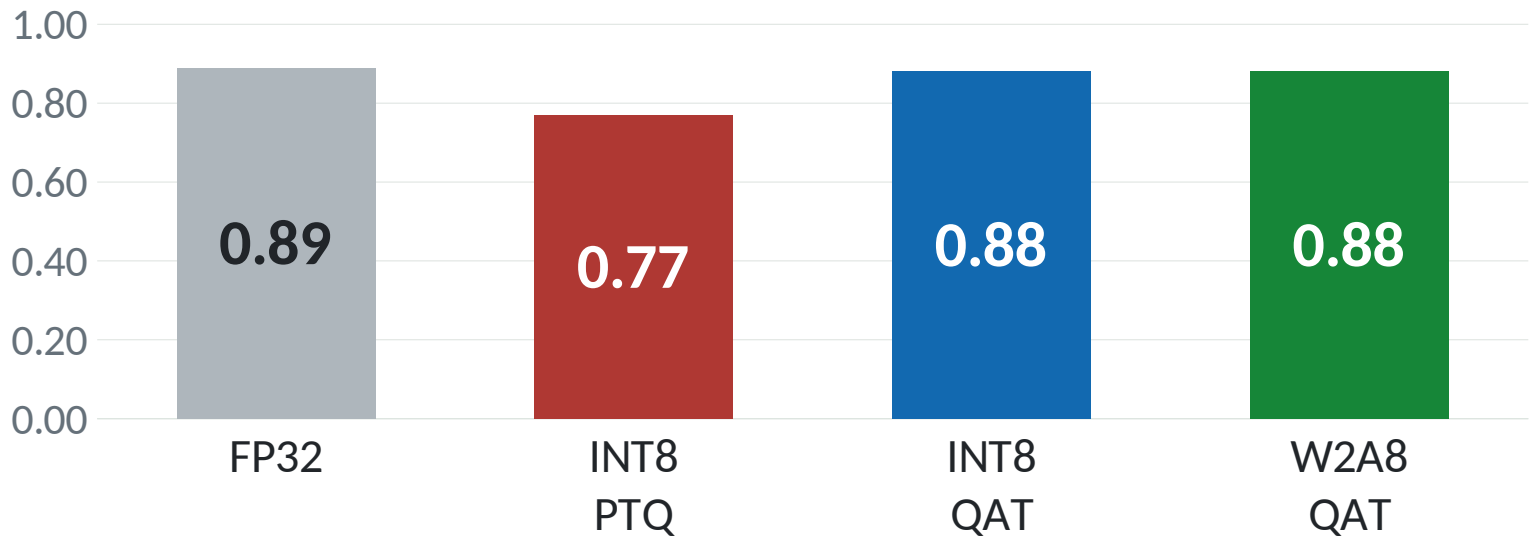


Training recovers performance at low precision

FEMBA on TUAB (Temple abnormal-EEG benchmark): quantization-aware training recovers AUROC

Original Direct conversion Train at low precision

TUAB AUROC • higher is better



≈ 2 MB

stored weights

2-bit weights
8-bit activations

Still above 1.5 MB: weights stream in tiles from external RAM while the previous tile computes

Training is part of the hardware solution



Foundation-model encoders run on GAP9

Measured latency at 370 MHz • TinyMyo draws 57 mW on average

Model • input	Latency	Energy
FEMBA • EEG	1,324 ms	39.7 mJ
TinyMyo • 5-s EMG	785 ms	44.9 mJ
TinyissimoMyo • 5-s EMG	496 ms	≈ 28 mJ*
PanLUNA • 10-s ECG	326 ms	18.8 mJ
LUNA • EEG	270 ms	14.8 mJ
CEReBrO • EEG	146 ms	8.0 mJ
TinyissimoMyo • 1-s EMG	89 ms	≈ 5 mJ*

* TinyissimoMyo energy estimated from its latency at the measured 57 mW envelope

FEMBA has the fewest FLOPs and the longest latency: the state-space recurrence serializes on a parallel core, which is why model and hardware mapping are designed together

Two sizes, one recipe: 3.6M for accuracy, 1.9M for sub-100 ms latency



What we have built

Open biosignal models, developed for local inference

Electrode layouts

LUNA and PanLUNA: varying layouts of EEG, ECG, PPG.
TinyMyo: per-channel EMG



Long recordings

LuMamba: linear cost, 26.5× fewer FLOPs, combined objective best on APAVA



Local inference

INT8 encoders run in 89 ms to 1.3 s on GAP9; 2-bit weights keep accuracy



40,000+

model downloads on Hugging Face
Open code and model weights

Biosignal learning and hardware execution, developed together