# Neural Data Transformer 3: A Generalist Motor Decoder

Joel Ye[1,2,4], Fabio Rizzoglio[7], Adam Smoulder[1], Hongwei Mao[3], Xuan Ma[7], Patrick Marino[3], Raeed Chowdhury[3], Dalton Moore[4,6], Gary Blumenthal[2,3], William Hockeimer[2,3,4], Nicolas G. Kunigk[2,3,4], J. Patrick Mayo[3], Adam Rouse[5], Aaron Batista[3], Steven Chase[1], Charles Greenspon[4,6], Lee Miller[4,7], Nicholas Hatsopoulos[4,6], Andrew Schwartz[3], Jennifer L. Collinger[2,3,4], Leila Wehbe[1], Robert Gaunt[2,3,4]
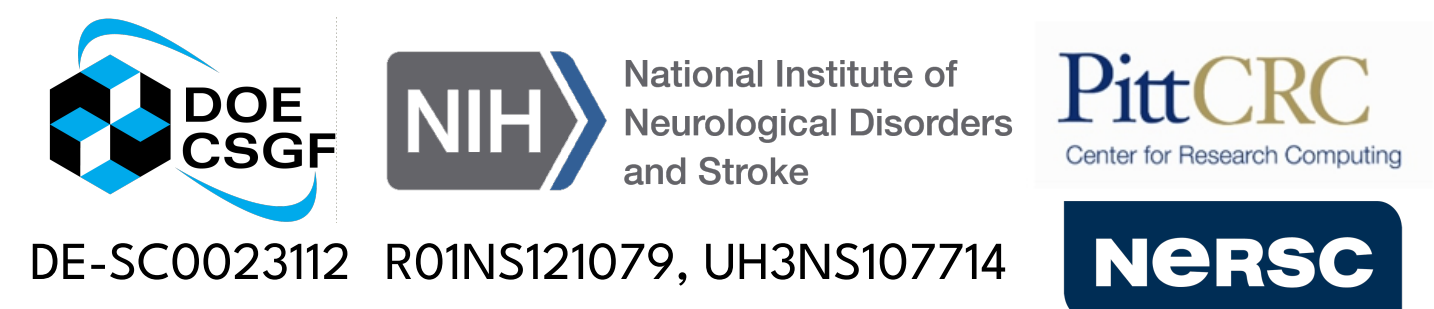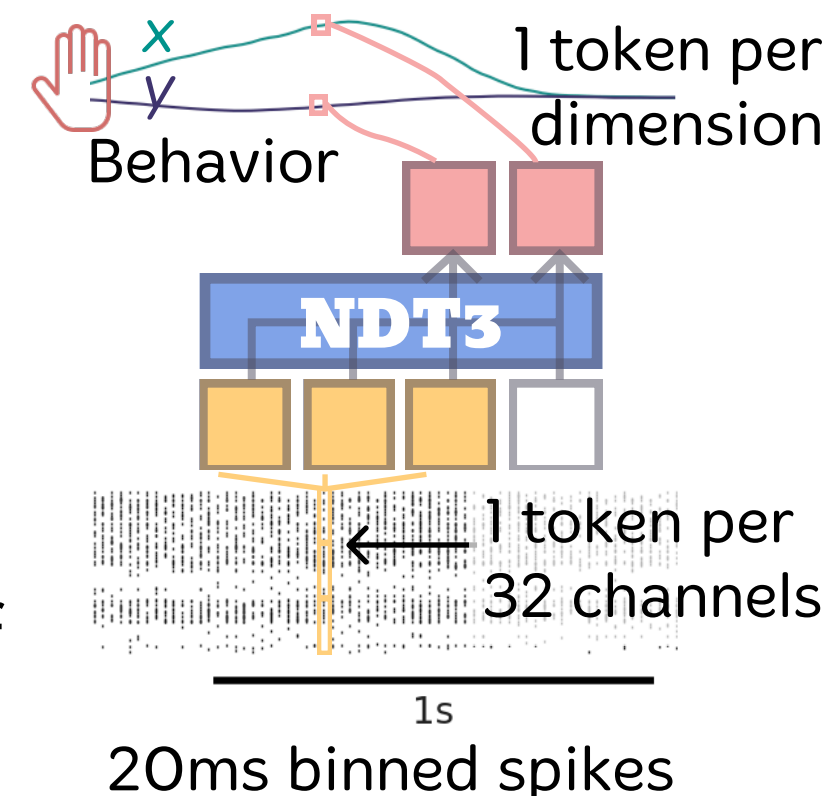
Methods

Training on heterogeneous data

Datasets vary in number of recorded neurons and covariates.
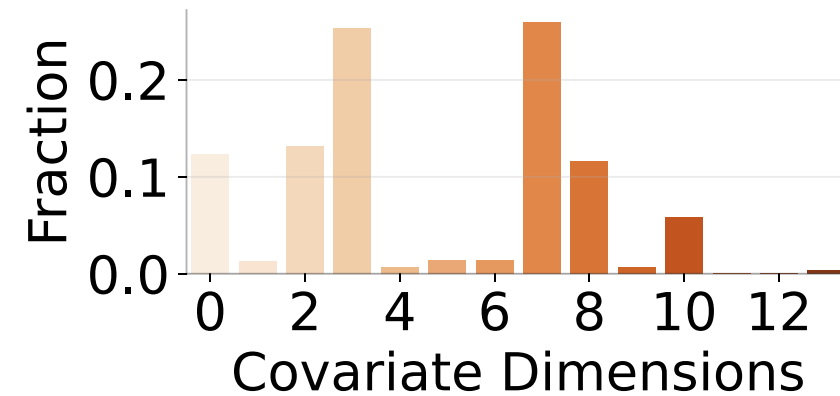
Varied **dimensions** in **data** can be converted to a varied number of **tokens** in a **Transformer**.

Behavior — 1 token per dimension

NDT3 — 1 token per 32 channels

Transformer training is highly parallel. Training 350M parameters on 2000 hours of data uses 20K GPU-hours.

20ms binned spikes

Evaluating decoding on 8 datasets
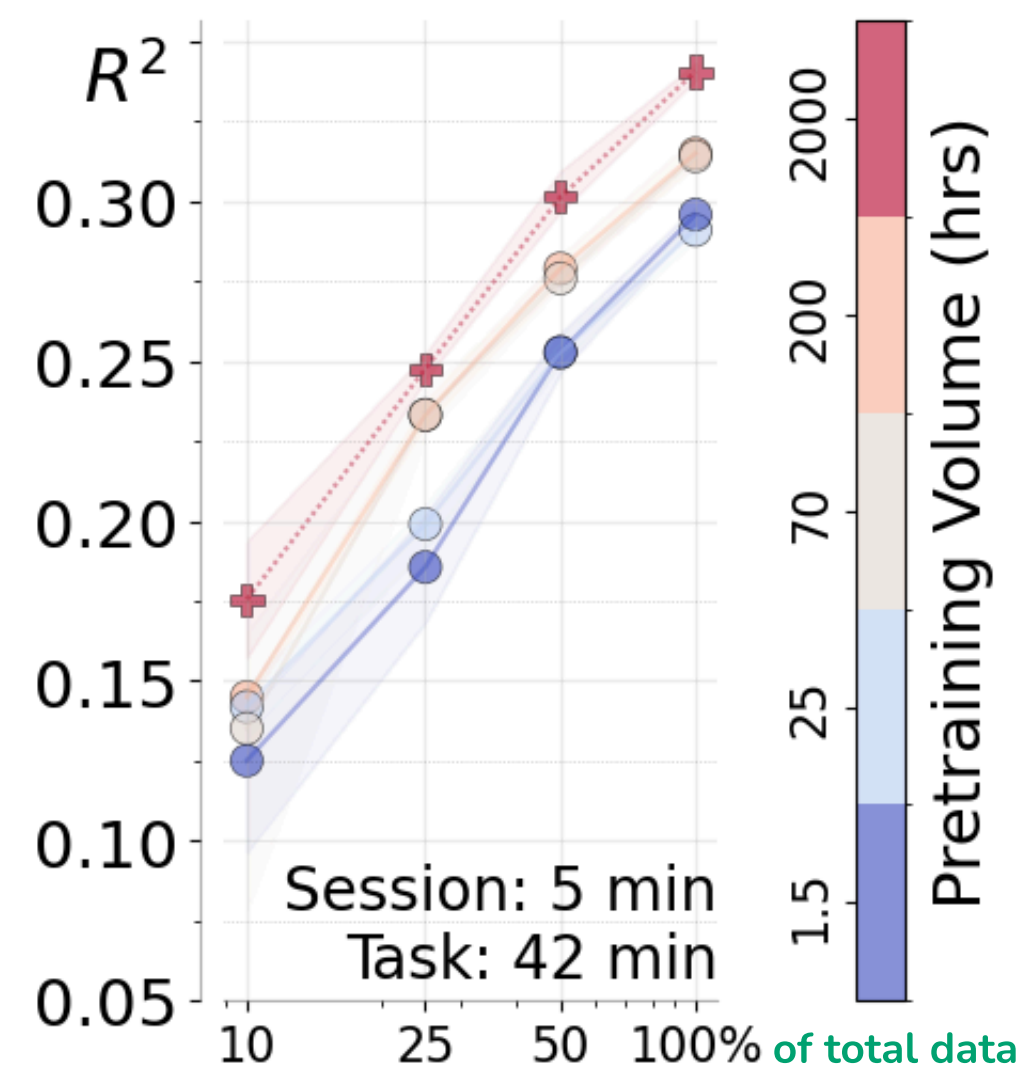
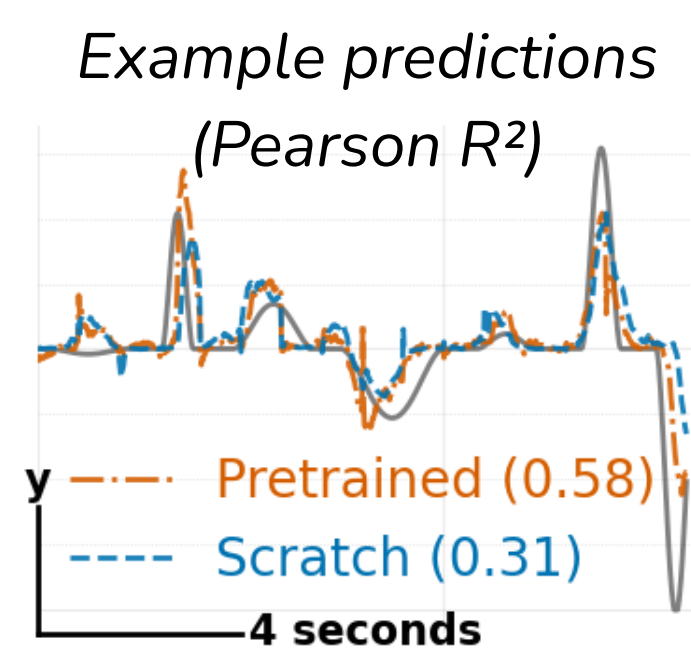FALCON benchmark (Karpowicz 24)
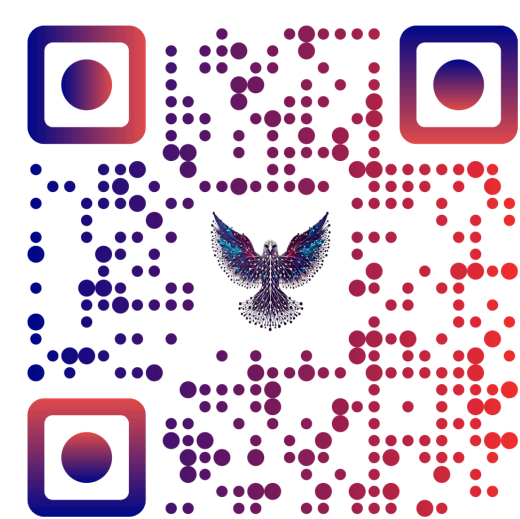
Monkey Movement
- 1D Critical Stability Task
- 2D Finger Pose ←
- 2D Self-paced reach
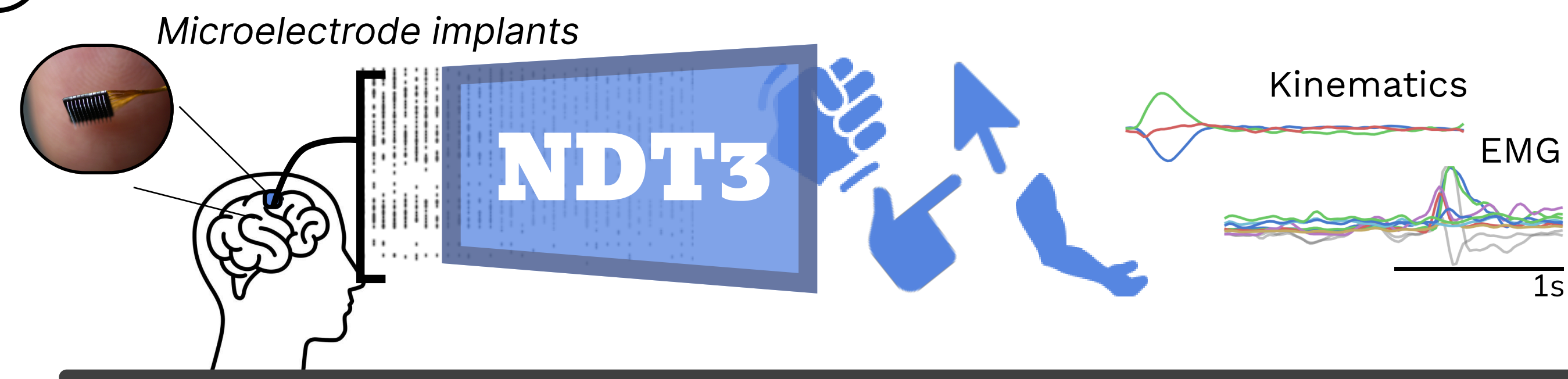- 16D Reach / Grasp EMG ←

Human iBCI (Open-Loop)
- 1D Grasp Force
- 2D Cursor + Click
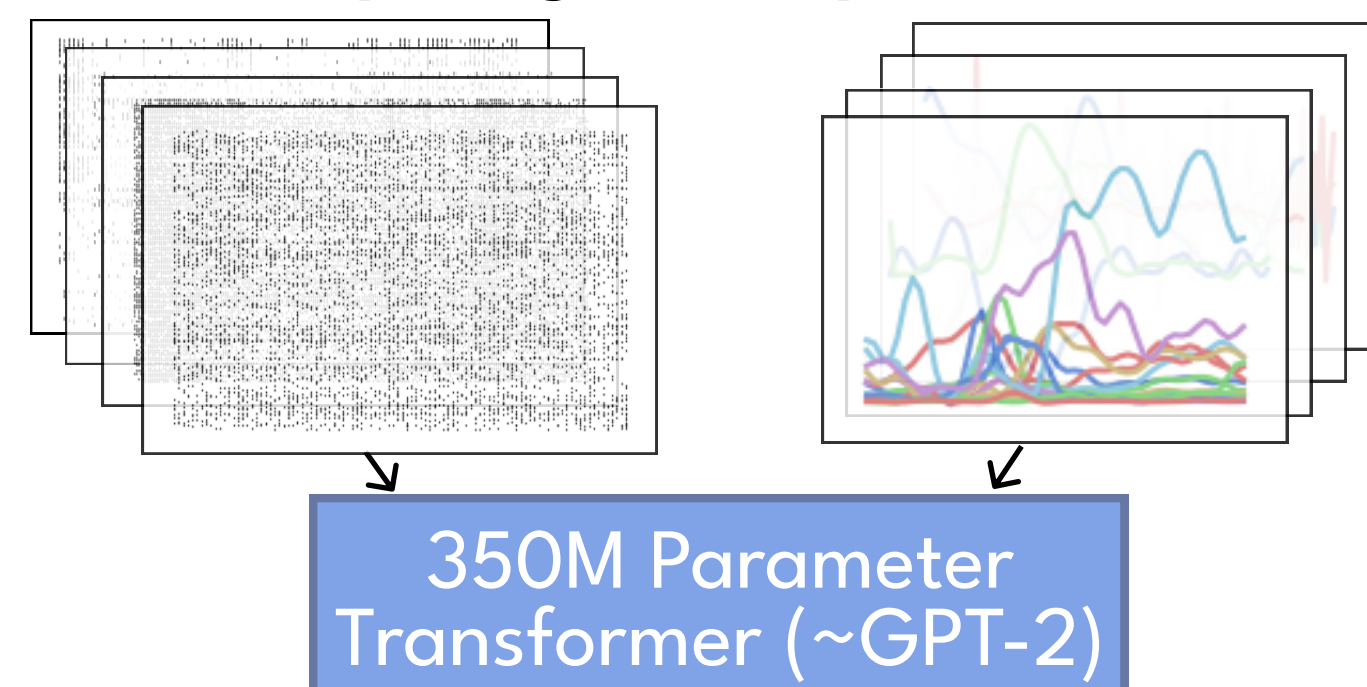- 4D Bimanual Cursor
- 7D Reach / Grasp Kinematics ←

Model must fine-tune to each dataset. We evaluate at **multiple task data scales**.
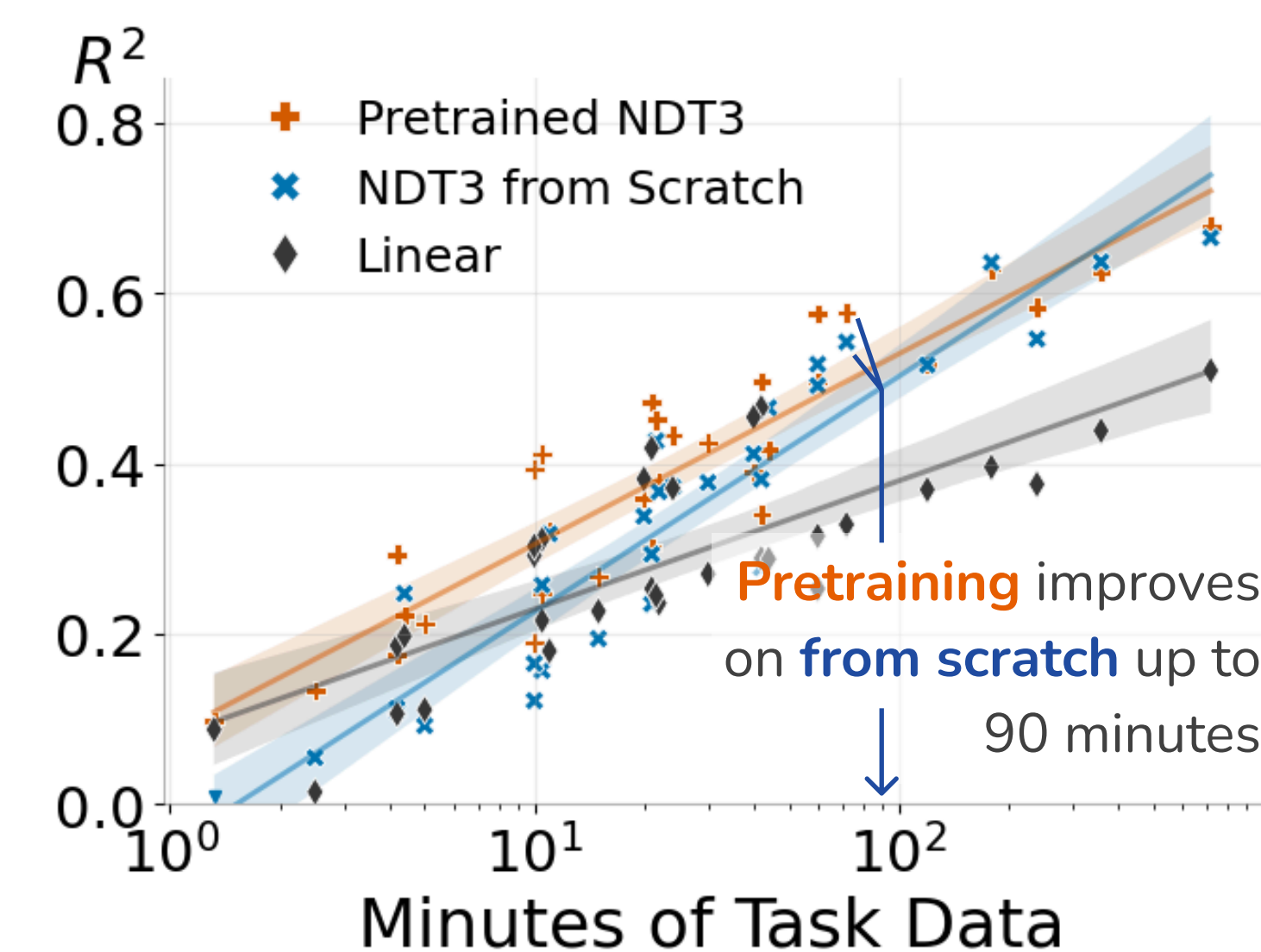
R^2 — Pretraining Volume (hrs)

Session: 5 min
Task: 42 min

Example predictions (Pearson R^2)

Pretrained (0.58)
Scratch (0.31)
4 seconds

10 25 50 100% of total data

① Can one model provide benefits across individuals and effectors?

Microelectrode implants

NDT3

Kinematics

EMG

1s

② To do this, we pretrain a deep neural network on

30+ monkeys
6 humans
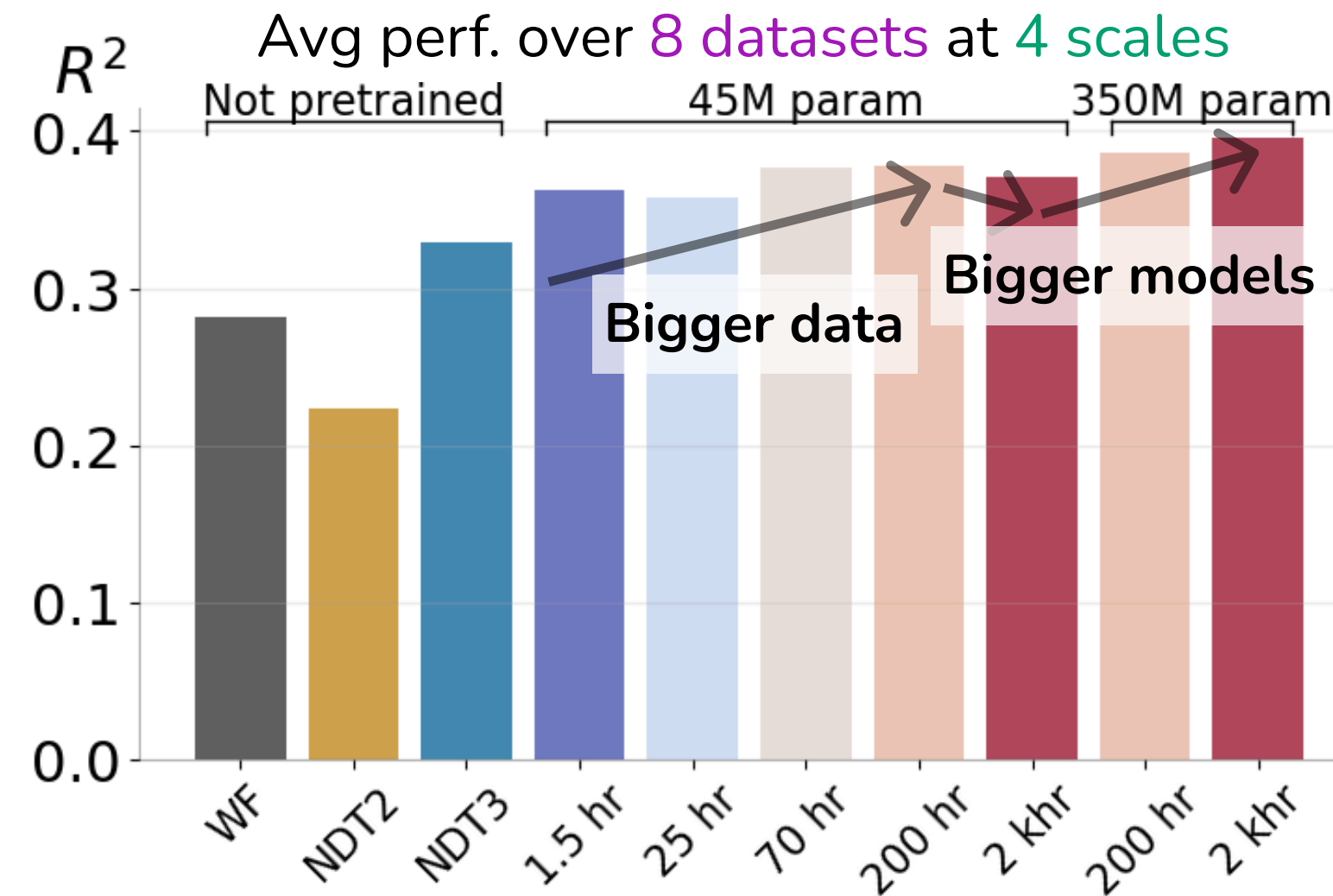**2000 hours**
of neural spiking activity and behavior.

Hours 1 10 100

350M Parameter Transformer (~GPT-2)

③ This improves data efficiency in decoding new behaviors in new individuals.

R^2

- Pretrained NDT3
- NDT3 from Scratch
- Linear

Pretraining improves on **from scratch** up to 90 minutes

Minutes of Task Data 10^0 10^1 10^2

④ Take-Home Message
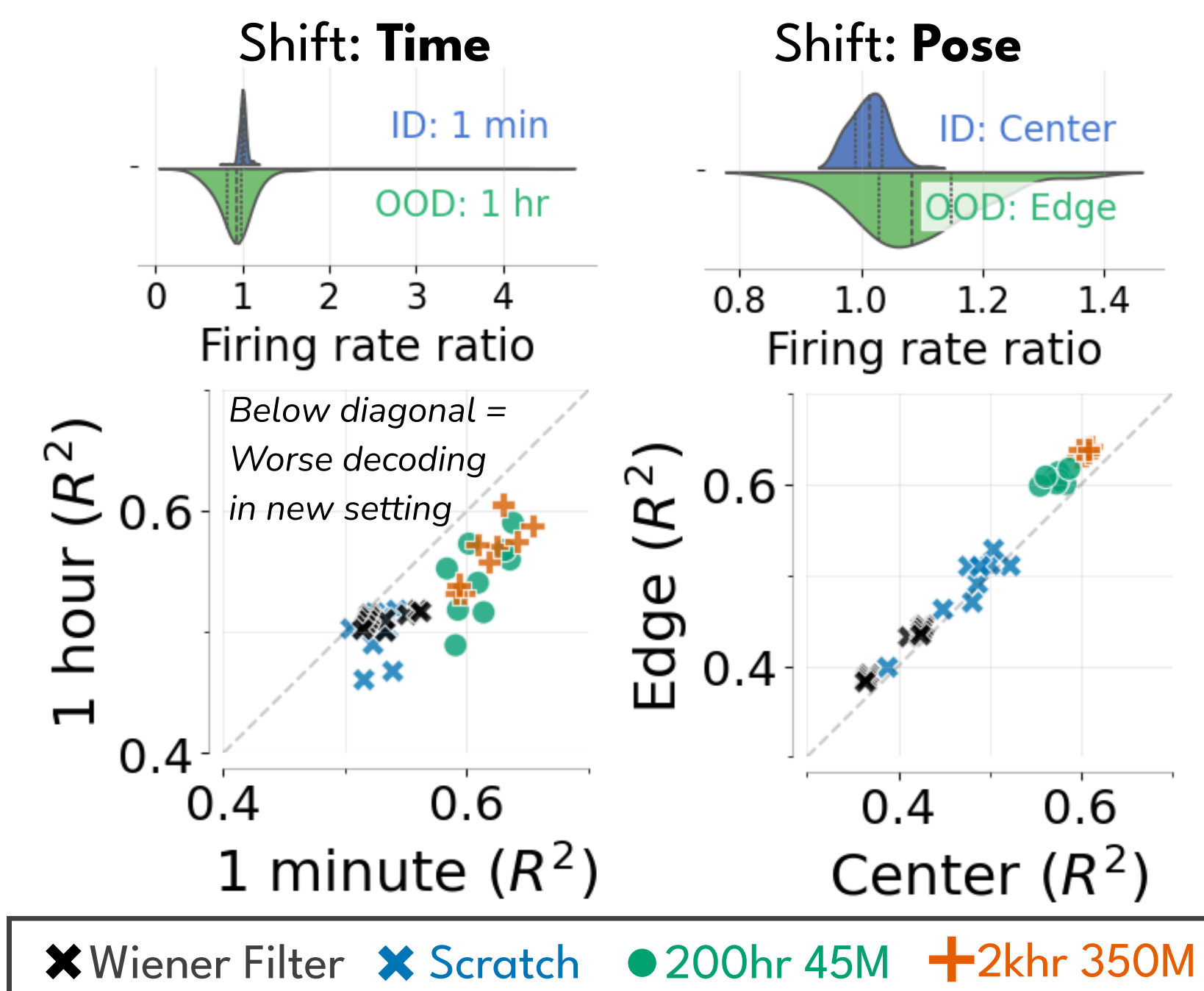
NDT3 benefits low-data decoding on many tasks.

Scaling more may further improve some tasks, but will require larger models.
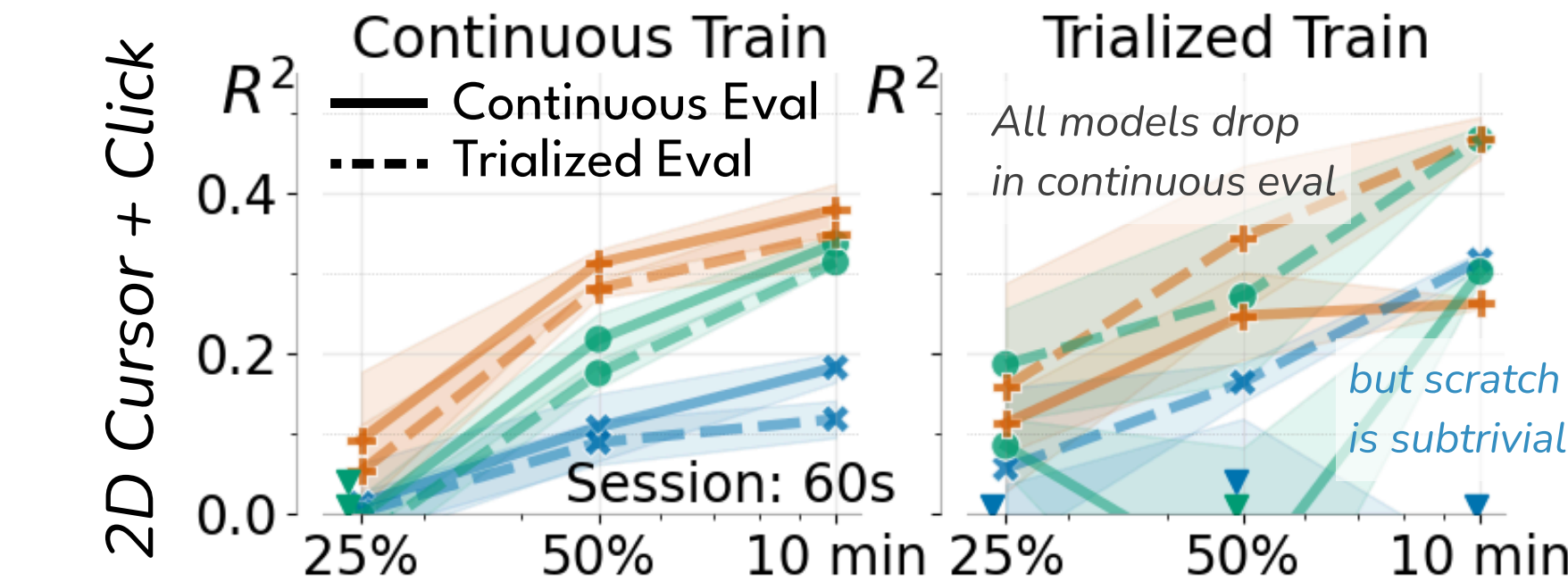
R^2 — Avg perf. over 8 datasets at 4 scales
Not pretrained | 45M param | 350M param

Bigger data → Bigger models

WF NDT2 NDT3 1.5 hr 25 hr 70 hr 200 hr 2 khr 200 hr 2 khr

Model Benefits

Pretraining gains persist through changes in neural firing patterns in reaching tasks.

Shift: Time
ID: 1 min
OOD: 1 hr
Firing rate ratio 0 1 2 3 4

Shift: Pose
ID: Center
OOD: Edge
Firing rate ratio 0.8 1.0 1.2 1.4

1 hour (R^2) vs 1 minute (R^2)
Below diagonal = Worse decoding in new setting

Edge (R^2) vs Center (R^2)

✖ Wiener Filter   ✖ Scratch   ● 200hr 45M   ✚ 2khr 350M

Pretrained NDT3 resists overfit to trial structure

Continuous Train / **Trialized Train**

2D Cursor + Click (R^2)
— Continuous Eval
-- Trialized Eval

All models drop in continuous eval

but scratch is subtrivial

25% 50% 10 min   Session: 60s

Pretraining benefits decoding outside motor cortex

R^2 **Somatosensory**
Self-paced reach
Session: 4 min
25% 50% 28 min

R^2 **Oculomotor**
4-Target Pursuit
Session: 40 min
25% 50% 10 hr