Supplementary Methods

PBT Search Space

During PBT, we specified search ranges (min/max) for regularization HPs. Weights were used to control maximum and minimum perturbation magnitudes for different HPs (e.g. a weight of 0.3 results in perturbation factors between 0.7 and 1.3). Keep ratio, dropout probability, and learning rate were limited to their specified ranges, while the KL and L2 penalties could be perturbed outside of the initial ranges.

Supplementary Table 1 | PBT-optimized hyperparameters and search ranges.

hyperparameter	min	max	perturbation weight	initial value
learning rate	1e-4	5e-3	0.3	2e-3
keep ratio	0.3	0.99	0.3	0.5
L2 on generator	1e-5	1e-1	0.8	log-random
L2 on controller	1e-5	1e-1	0.8	log-random
KL weight on controller outputs	1e-6	1e-4	0.8	log-random
KL weight on initial conditions	1e-6	1e-4	0.8	log-random
dropout probability	0.3	1	0.3	random

Generalization Analyses

A consideration when applying powerful deep learning models like AutoLFADS is the extent to which the method can perform when trained on multiple behavioral conditions. While the rat locomotion models were trained on EMG data recorded during gait at 3 distinct speeds (details in Methods), we wanted to further characterize the ability of a single model to be trained on multiple behavioral conditions (**Supplementary Fig. 2**). To assess this, we extended the data used for analysis of rat locomotion to two additional behavioral conditions where the treadmill was set at a fixed incline (+25 degrees/-25 degrees from horizontal) while the rat walked at constant speed (20 m/min). The differing biomechanics of locomotion during the incline conditions extended the variety of muscle activations beyond speeds.

Training AutoLFADS on multiple behavioral conditions. We trained four locomotion AutoLFADS models with different training datasets: (1) “Single-speed”, containing EMG data from gait at a single speed (20 m/min) with no incline, (2) “Multi-speed”, containing EMG data from gait at 3 different speeds (10, 15, 20 m/min) with no incline, (3) “Multi-incline”, containing EMG data from gait at a single speed (20 m/min) with 3 different inclines (+25, 0, -25 degrees), and (4) “Complete”, containing EMG data from all five behavioral conditions (3 speeds with no incline, 2 additional incline conditions at 20 m/min). Each model was trained using the same PBT strategy used in the paper (see *AutoLFADS Training* in Methods).

After training, AutoLFADS output was generated from each of the four models for the speed: 20 m/min, and incline: 0 degrees behavioral condition. Joint angular decoding was performed as described in the paper (see *Predicting joint angular acceleration* in Methods). Performance was evaluated in comparison to optimal Bayesian filtering (i.e., hyperparameters that yielded highest joint angular acceleration decoding performance) and low pass filtering at various cutoff frequencies. We found that performance does not decrease as a result of including data across conditions (**Supplementary Fig. 2 a, b**).

Evaluating AutoLFADS generalization performance. As noted above, when using AutoLFADS to perform offline analysis of data from distinct behaviors, the optimal approach is likely to train the model on data from all behaviors of interest. However, AutoLFADS may also be useful in online applications such as real-time control of a muscle-driven prosthetic or exoskeleton. In online applications, one may encounter data that is outside the behaviors spanned by the training set, and thus it may be important for trained AutoLFADS models to generalize to EMG activity unseen during training. We wanted to test whether model performance would fail at estimating muscle activations for new data, or whether it would be able to generalize reasonably well to behaviors outside of the training set. We evaluated the ability of the original Multi-speed model (trained on multiple speeds) presented in the paper to generalize to the previously-unseen incline conditions, and evaluated performance in decoding kinematic parameters (**Supplementary Fig. 2 c, d**). We compared results to the Complete model (trained on all behavioral conditions), to assess differences in decoding performance when the incline condition EMG data was included in the training set.

We found that the Multi-speed AutoLFADS model (no incline) was able to generalize to the incline conditions quite well, typically matching or significantly improving decoding performance over compared methods. Critically, for either dataset, the Multi-speed AutoLFADS models did not generate estimates of muscle activation that were significantly worse than standard methods (Bayesian filtering and smoothing), suggesting that the AutoLFADS models were not critically overfitting to the training data.

Data Limited Analyses

We wanted to better understand how data quantity affected the ability to effectively train AutoLFADS models to achieve high performance. To test this, we trained AutoLFADS models using datasets of varying sizes (**Supplementary Fig. 3**).

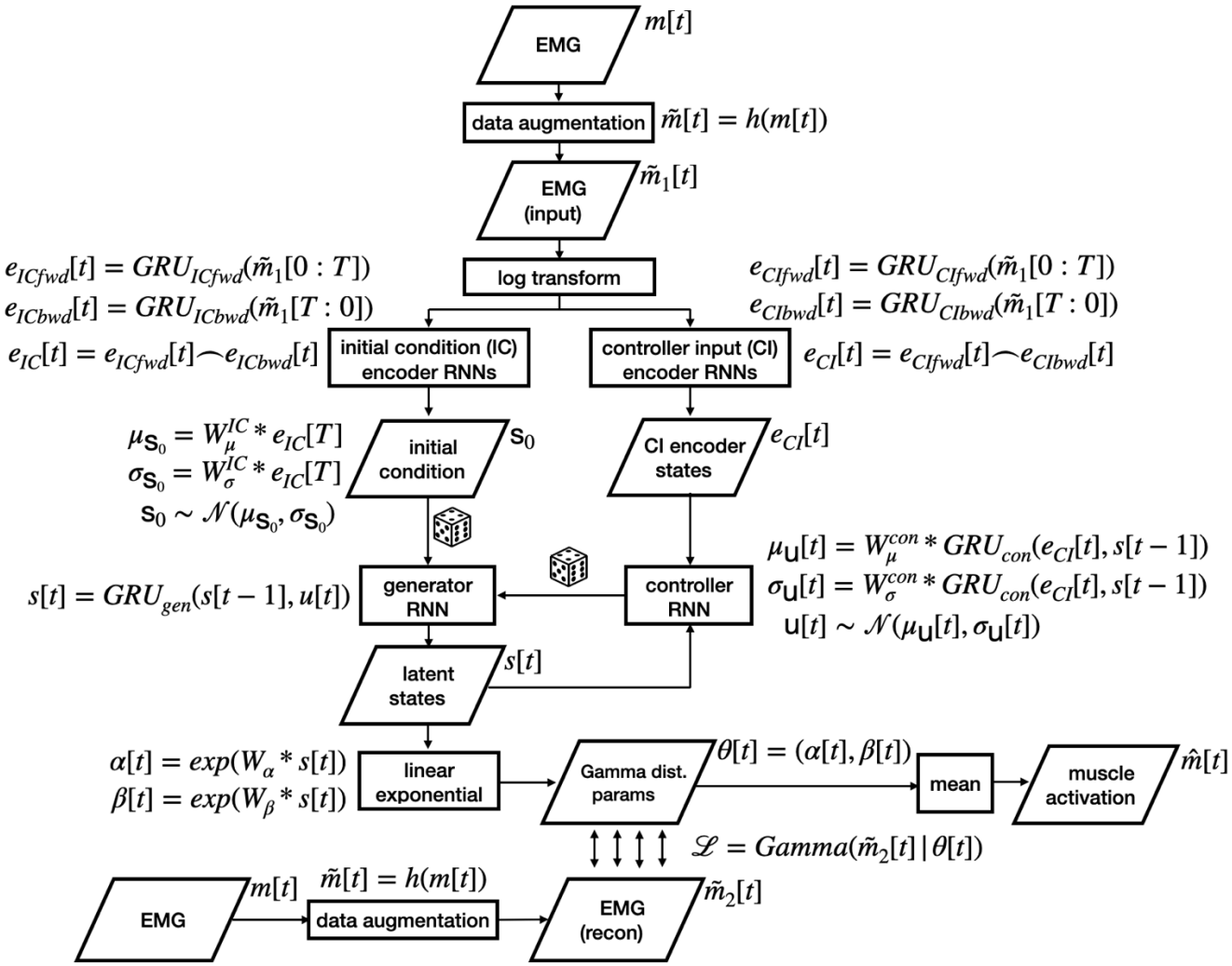
Creating locomotion datasets of varying sizes. We started with a "Multi-speed" dataset containing EMG activity from the three speed conditions (no incline). Since we modeled data using AutoLFADS without regard to behavioral structure, we wanted to ensure that datasets of varying sizes still included a balanced proportion of good cycles from each behavioral condition. To control for this, we limited data used for analysis to cycles that were either labeled as high-quality by the experimenter (i.e., no artifacts, stereotyped behavior) or passed visual screening during the data curation process. For each cycle, EMG data were aligned to timing of foot off (300ms prior, 450ms after) and further chopped to 200ms windows with 50ms overlap for AutoLFADS modeling (see *Data Preprocessing* in the Methods). Datasets of varying sizes were created by randomly selecting 80, 40, 20, 10 and 5% of the cycles from the full dataset.

AutoLFADS training. To ensure PBT training of AutoLFADS models was consistent regardless of dataset size, we matched the batch size across runs (15 samples/batch) and ensured that the number of training steps during KL/L2 ramping and the number of steps per generation of PBT training matched across models (300 training steps for ramping, 180 training steps per generation). This ensured that all models were subject to the same PBT hyperparameter optimization schedule regardless of dataset size. Additionally, we tested the effect of data augmentation on model performance by training a separate model for each dataset with and without temporal shift data augmentation (2 models/dataset size, 6 dataset sizes, 12 models total) Note, although the PBT strategy described above differs from approach presented in the paper, models trained on the full dataset using this PBT strategy matched the performance demonstrated in the paper.

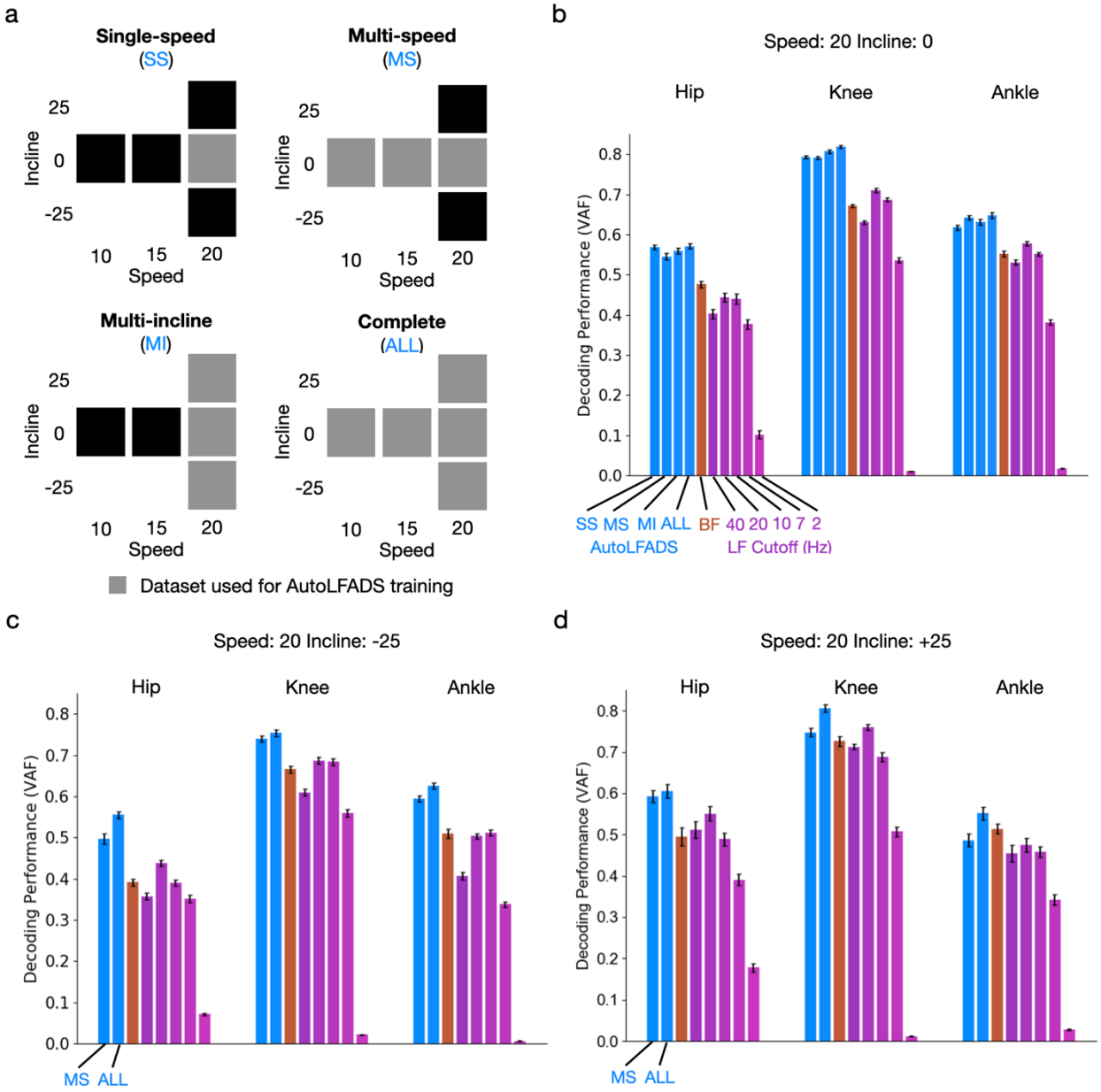
Evaluating AutoLFADS model output. To evaluate AutoLFADS models trained on datasets of differing sizes, we compared the estimates of muscle activation on the full dataset from each trained model. Joint angular decoding was performed as described in the paper (see *Predicting joint angular acceleration* in Methods). To compare decoding performance across models, we normalized decoding performance relative to the highest performing model (i.e., model trained with temporal shift on the full dataset, presented in paper). Additional comparisons were included to understand relative performance of optimal low pass filtering and optimal Bayesian filtering.

We found that training AutoLFADS models on datasets as small as 5% of the full dataset can still maintain ~95% of the decoding performance achieved by training on the full dataset. We also found that the data augmentation strategy we developed for EMG (i.e., Temporal Shift) improves the reliability of the AutoLFADS training regardless of dataset size. When trained on dataset of varying sizes, AutoLFADS models trained *without* Temporal shift do not reliably improve decoding performance over standard methods, i.e., low pass filtering (LF) or Bayesian filtering (BF). In contrast, AutoLFADS models that are trained *with* Temporal Shift, which steadily improves model performance as we incorporate more data into the training process.

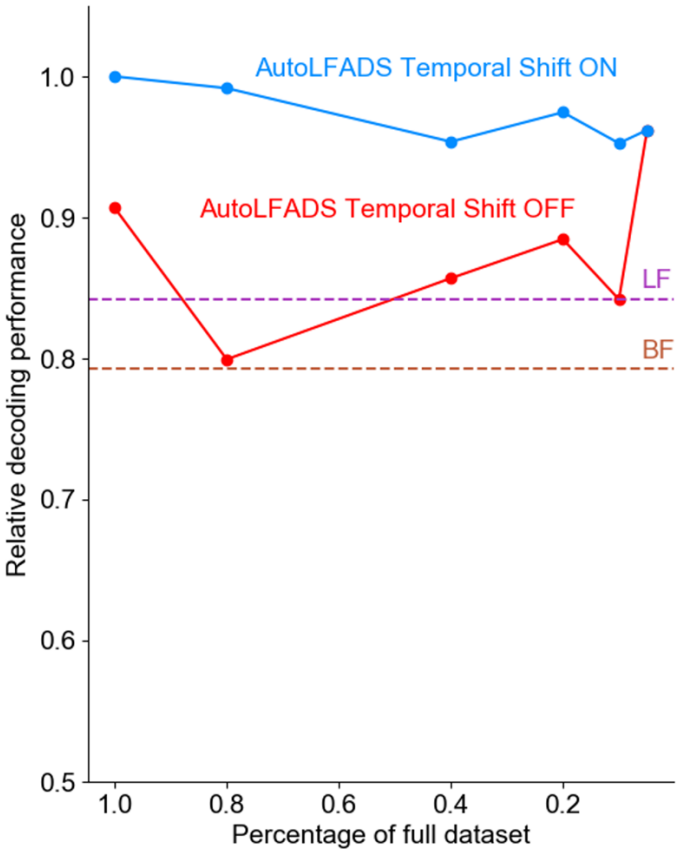
Supplementary Figures



Supplementary Fig. 1 | Flowchart of AutoLFADS modeling for EMG. Data augmentation applied twice to randomize augmentation used for data input to the model and data used to evaluate reconstruction cost (details in *AutoLFADS training*). There are six recurrent neural networks (RNN) implemented using Gated Recurrent Unit (GRU) cells that are trained in AutoLFADS: bi-directional (forward and backward) initial condition (IC) encoder GRU cells, bi-directional controller input (CI) encoder GRU cells, controller GRU cell, and the generator GRU cell. The initial condition ($\mathbf{s_0}$) and the input at each timestep ($\mathbf{u}$[t]) is estimated as a Gaussian distribution. Die graphic denotes sampling from a distribution. "*Gamma"* indicates Gamma log-likelihood.



Supplementary Fig. 2 | AutoLFADS applied across multiple behavioral conditions. (a) Schematic shows corresponding behavioral conditions included in the datasets used to train four different AutoLFADS models (Single-speed, Multi-speed, Multi-incline, Complete) on rat locomotion EMG. (b) Joint angular acceleration decoding performance for the four different AutoLFADS models, quantified using variance accounted for (VAF). Model performance shown in comparison to the optimal Bayesian filtering approach, and multiple cutoff frequencies of low-pass filters. Each bar represents the mean +/- SEM across cross-validated prediction VAF values for a given joint. (c) Joint angular decoding performance for Multi-speed and Complete AutoLFADS models on the Speed: 20, Incline: +25 condition. (d) Same as (c), but for Speed: 20, Incline: -25 condition. Compared approaches in (c) and (d) are same as in (b).



Supplementary Fig. 3 | AutoLFADS applied to datasets of varying size. Joint angular acceleration decoding performance (VAF, mean of hip, knee, and ankle) from AutoLFADS model output as a function of training dataset size. Decoding performance quantified relative to the performance of the AutoLFADS model trained on the full dataset with temporal shift (i.e., strategy used in paper). AutoLFADS performance compared for models trained with temporal shift (blue) or without applying data augmentation (red). Performance also compared to optimal low pass filtering (purple) and optimal Bayesian filtering (orange).