MUniverse: A Simulation and Benchmarking Suite for Motor Unit Decomposition

Pranav Mamidanna[1,2,*], **Thomas Klotz**[3,*], **Dimitrios Halatsis**[2,*],
Agnese Grison[2], **Irene Mendez-Guerra**[1,2], **Shihan Ma**[2], **Arnault H. Caillet**[2],
Simon Avrillon[4], **Robin Rohlén**[2,5], **Dario Farina**[2].
[1] I-X Center for AI in Science, Imperial College London, UK
[2] Department of Bioengineering, Imperial College London, UK
[3] Institute for Modelling and Simulation of Biomechanical Systems, University of Stuttgart, Germany
[4] Université Côte d'Azur, LAMHESS, Nice, France
[5] Department of Diagnostics and Intervention, Umeå University, Sweden
[*] These author have equally contributed to the work

Abstract

Neural source separation enables the extraction of individual spike trains from complex electrophysiological recordings. When applied to electromyographic (EMG) signals, it provides a unique window into the motor output of the nervous system by isolating the spiking activity of motor units (MUs). MU decomposition from EMG signals is currently the only scalable neural interfacing approach available in behaving humans and has become foundational in motor neuroscience and neuroprosthetics. However, unlike related domains such as spike sorting or electroencephalography (EEG) analysis, decomposition of EMG signals lacks open benchmarks that reflect the diversity of muscles, movement contexts, and noise sources encountered in practice. To address this gap, we introduce MUniverse, a modular simulation and benchmarking suite for decomposing EMG signals into individual MU spiking activity. MUniverse provides: (1) a simulation stack with a user-friendly interface to a state-of-the-art EMG generator; (2) a curated library of datasets across synthetic, hybrid synthetic-real data with ground truth spikes, and experimental EMG; (3) a set of internal and external decomposition pipelines; and (4) a unified benchmark with well-defined tasks, standard evaluation metrics, and baseline results from established decomposition pipelines. MUniverse is designed for extensibility, reproducibility, and community use, and all datasets are distributed with standardised metadata (Croissant, BIDS). By standardising evaluation and enabling dataset simulation at scale, MUniverse aims to catalyze progress on this long-standing neural signal processing problem.

1 Introduction

Modern neuroscience increasingly relies on source separation techniques that recover latent neural signals from mixtures, aligning with core machine learning challenges of inverse problems and representation learning. These neural (blind) source separation problems are ubiquitous across neuroscience, from spike sorting in extracellular electrophysiological recordings to calcium imaging demixing, and source localization in electroencephalography (EEG) and magnetoencephalography recordings. Each instance requires sophisticated algorithms to invert ill-posed signal models and uncover the underlying neural activity patterns. Electromyographic (EMG) recordings present a particularly compelling case for machine learning researchers: each surface electrode captures a convolutive mixture of motor unit (MU) action potentials propagating through biological tissues, creating a natural testbed for advanced blind source separation (BSS) techniques.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

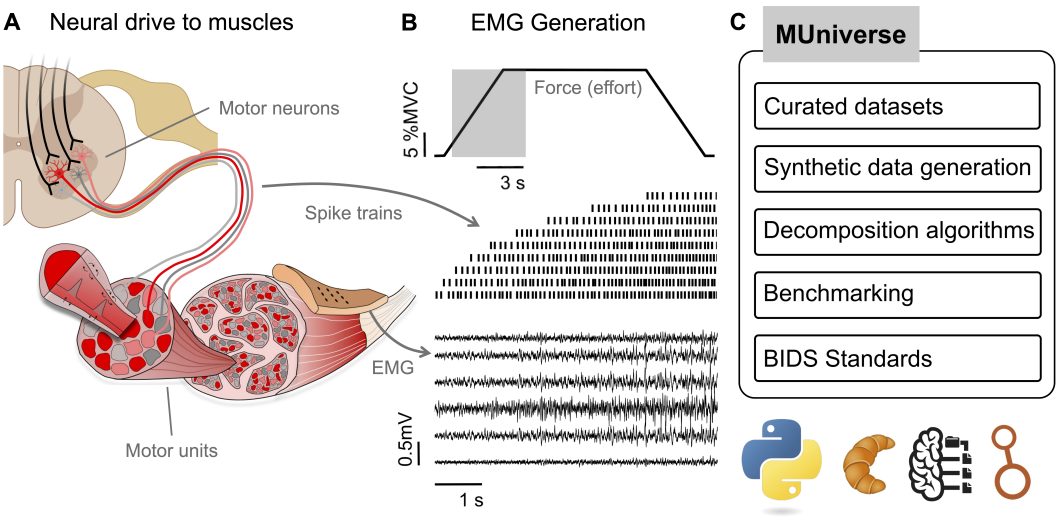

Figure 1: **MUniverse overview and motivation.** (**A**) Decomposition of EMG signals obtained at the surface of the skin into individual motor unit spike trains provides a unique window into the neural control of movement. (**B**) The generation pathway from spike trains to EMG for an isometric contraction with a ramp and hold pattern of effort exertion. Neural commands from the spinal cord (shown as spikes here) drive the motor units to produce a particular pattern of force, and the resulting biopotentials are captured using HD-EMG. (**C**) The scope of MUniverse includes curated datasets, a simulation stack, algorithms, as well as tools to standardize EMG datasets.

The MU is the smallest voluntarily contractible unit comprising a motor neuron and an axon that innervates tens to hundreds of muscle fibres (Figure 1A). EMG-based decomposition into MU activity has played an important role in, e.g., improving our understanding of the neural control of movement [23], developing neural interfaces with applications such as prosthetic control [15] and human augmentation [12]. Decomposition – extracting individual MU activities from the mixed EMG signals – has therefore emerged as a foundational technique with broad scientific and clinical impact.

Over the past two decades, numerous BSS methods have been proposed to decompose high-density EMG (HD-EMG) signals into individual MU spike trains [22, 8, 37, 9, 19]. The gold standard for the validation of BSS methods is the two-source test, i.e., concurrent recording of HD-EMG and intramuscular needle or fine wire electrodes, comparing the rate of agreement between the two methods [14]. However, intramuscular fine wire electrodes provide a limited number of detectable MUs due to their spatial selectivity. Further, there exist very few open datasets with these simultaneous recordings. Further insights on the performance of decomposition algorithms can be obtained from simulated data with known ground-truth, though such data might oversimplify some aspects of the generation and recording of experimental EMG signals. Despite a growing ecosystem of algorithms, EMG decomposition still lacks the rigorous benchmarking culture that catalysed breakthroughs in spike sorting [34] [5]. Methods are evaluated on private data, with heterogeneous pre-processing pipelines and different metrics, making it difficult to gauge real progress or identify failure modes.

The field of neural interfacing and motor neuroscience, therefore, needs a standardised, open, and extensible benchmark that spans simulated, hybrid, and experimental recordings.

Our contributions To address this issue, we introduce MUniverse (Figure 1), an open-source simulation and benchmarking suite for decomposing EMG recordings into MU spike trains. MUniverse includes:

- A containerized simulation stack that integrates musculoskeletal modeling, motor neuron pool dynamics, and generative models of MU action potentials (MUAPs) for end-to-end generation of EMG signals.
- A library of curated, diverse datasets: fully simulated, hybrid (experimentally recorded MUAPs convolved with synthetic spike trains), and experimental EMG.

- A uniform API to a suite of decomposition algorithms, both implemented natively, as well as containerized wrappers to existing algorithms.
- And a standard evaluation framework for decomposition outputs, that generates a report card using standard metrics of source quality and signal reconstruction.

By unifying data generation, evaluation, and reporting under FAIR principles, MUniverse enables the community to measure progress rigorously and accelerate the next decade of EMG-based neural interfacing. MUniverse is currently hosted on GitHub[1].

2 Background and related work

2.1 EMG signal model

HD-EMG signals can be modelled as a linear convolutive mixture [13]:

$$\mathbf{x}(t) \;=\; \sum_{l=0}^{L-1} \mathbf{H}(l,t)\,\mathbf{s}(t-l) \;+\; \boldsymbol{\varepsilon}(t)\;, \tag{1}$$

where t is a discrete time sample, $\mathbf{x}(t) \in \mathbb{R}^M$ is the EMG signal (with M being the number of channels), $\mathbf{H}(l,t) \in \mathbb{R}^{M\times N}$ contains the finite MU impulse responses (i.e., the MUAPs) of length L samples, $\boldsymbol{\varepsilon}(t)$ is additive noise, and $\mathbf{s}(t) \in \mathbb{R}^N$ are the N MU spike trains comprising zeros and ones. The spike train for MU j is defined as $s_j(t) = \sum_{r\in\mathcal{S}_j} \delta(t - t_j^r)$ where $\mathcal{S}_j = \{t_j^1, ..., t_j^{T_j}\}$ is the set of discharge times and $\delta(\cdot)$ the Dirac delta function. Often $\mathbf{H}(l,t)$ is assumed to be stationary (e.g., by considering non-fatiguing and isometric contractions), i.e., not being dependent on time.

2.2 Decomposition methods

The goal of EMG decomposition is to estimate the motor neuron spike trains $\mathbf{s}(t)$ only given the observed EMG signals $\mathbf{x}(t)$ (see Equation (1)). To solve that problem, existing decomposition methods can be classified into three groups: (1) template-matching, (2) convolutive BSS, and (3) deep learning-based methods.

Template matching summarizes a group of (semi-)supervised data analysis methods, which are well-established in many fields of electrophysiological spike sorting problems [e.g., 34]. In short, these methods find, match, and update MUAP templates (waveforms) [10], and resolve complex superpositions [36]. Well-known algorithms in this category are the PD III [10] and PD-IPUS + PD-IGAT [36], where the latter is an extension of the former. Yet, compared to invasive recordings, surface EMG signals have a smaller bandwidth due to low-pass filtering of the volume conductor, making the MUAPs more similar and the inverse problem more challenging. Hence, template matching methods often face limitations in the presence of multiple overlapping sources (e.g., high-intensity contractions or doublet discharges [38]) or signal non-stationarities.

Convolutive BSS (CBSS) is currently the most common method for MU identification from surface EMG signals. In CBSS, MU spike trains are estimated by solving an optimization problem that considers the statistical properties of the MU spike trains [13]. The theory is closely related to independent component analysis [25], which has three main assumptions: linear superposition of the observations, the prior of the MU spike trains and the joint prior of the MU spike trains being factorial (i.e., statistical independence). Existing algorithms vary with respect to the selected objective functions (e.g., higher-order statistical moments such as skewness or kurtosis), optimization methods (e.g., gradient descent or quasi-Newton methods), and (partially algorithm-specific) hyperparameters (often in the range of 10-50). Popular CBSS algorithms include the gradient convolution kernel compensation (gCKC) [21], as well as variants of fast Independent Component Analysis (ICA) with [8] or without peeling off (removing) estimated MU spike trains [37] and reiterating the spike train estimation procedure.

Deep-learning outlook. Recent work [35] shows that a shallow autoencoder with an orthogonally constrained encoder and a sparsity-promoting latent objective can recover spike-like sources from EMG without labels, offering a principled deep learning alternative to traditional handcrafted ICA contrasts. Other approaches are still in an exploratory phase. Recent *non-linear ICA* frameworks offer a principled route to overcome the accuracy–latency limits of classical pipelines and, crucially,

[1] `https://github.com/dfarinagroup/muniverse`

to deal with non-stationary signals. In particular, Time-Contrastive Learning (TCL) trains a classifier to discriminate short, non-stationary segments of a time series; the resulting latent space provably recovers the independent sources up to trivial indeterminacies [24]. Follow-up work with auxiliary variables [26] and identifiable Variational Auto-Encoders [28] generalizes this idea, showing that deep networks can achieve identifiability without ground-truth sources – an essential property for EMG decomposition, where verified MU labels typically do not exist, and if known, might adapt over time. Although these methods have been explored in modalities such as EEG, they have yet to be brought to EMG. Integrating such self-supervised objectives with architectures already validated for *supervised* HD-EMG decomposition [9, 31, 43] could yield low-latency, device-independent systems that learn *on-the-fly* from raw recordings – eliminating the whitening and delay-embedding steps that hamper current approaches and opening the path to robust MU decoding during natural movements.

2.3 Public EMG datasets, algorithms and benchmarking efforts

Over the past few years, a handful of open-source toolboxes [3, 39, 42, 27] have begun to lower the barrier to EMG decomposition research. However, most studies have been based on in-house codes or proprietary packages. Experimental data releases [6, 2, 27] have become more frequent in recent years, yet standards for reporting both data and metadata are still missing, hindering the integration into fully automated pipelines. Other sEMG collections [1, 11, 40, 41] address classification or myoelectric control tasks but lack ground truth in terms of spike trains.

In contrast, fields such as spike sorting (through efforts like SpikeForest [34]) and EEG source localization (through BCI Competitions) have access to large, standardized benchmarks that drive rapid progress. However, EMG decomposition remains fragmented across private data, divergent pipelines, and inconsistent reporting, motivating the need for a unified, extensible benchmark.

3 MUniverse overview

3.1 Design goals

We designed MUniverse to be modular, reproducible, and accessible to both neuroscientists and machine learning researchers. For this purpose, there are three goals for MUniverse:

- **FAIR and BIDS compliance:** to facilitate users in rapidly finding and accessing the datasets, we provide Croissant files for each curated dataset. Each dataset follows the proposed standard for EMG data (in community review [2]). BIDS (Brain Imaging Data Structure [17]) is a community standard for organizing neuroimaging data and metadata in both human and machine-readable formats and comes with existing pipelines to process and ingest data and metadata stored in the format.
- **Rich metadata and provenance:** to ensure that the various datasets we provide and operate on remain end-to-end reproducible, we provide automated logging to carry full provenance on raw signals, simulation/algorithm parameters, and decomposition outputs.
- **Modularity and containerization:** to isolate dependencies and ensure cross-platform reproducibility, we provide a clean and modular API for data generation, decomposition, as well as evaluation, while each of the external packages (NeuroMotion [33] and Swarm-Contrastive Decomposition (SCD) [19]) is containerized.

3.2 High-level architecture

MUniverse is organized into four high-level modules: (1) the simulation stack, (2) decomposition pipelines, (3) evaluation and benchmarking tools, and (4) utilities to handle all data formatting and reporting, enabling users to produce and ingest data and metadata seamlessly (Figure 2).

- **Simulation stack:** MUniverse features a user-friendly API to a state-of-the-art EMG simulator, NeuroMotion [33]. NeuroMotion combines OpenSim-based muscle kinematics, motor neuron-pool models to generate spike trains [16], and a MUAP waveforms generator [32], to synthesize high-fidelity data that closely matches the biophysics of EMG generation. Through MUniverse, users can supply a config file specifying the anatomical, movement, and recording parameters that govern EMG. The simulator module parses this config to invoke NeuroMotion under the hood and emit raw EMG together with ground-truth spike trains, kinematic variables of the degree of freedom under consideration, torque/activation profile that generated the spike trains, and a complete JSON provenance record.

[2] `https://bids-specification--1998.org.readthedocs.build/en/1998/`

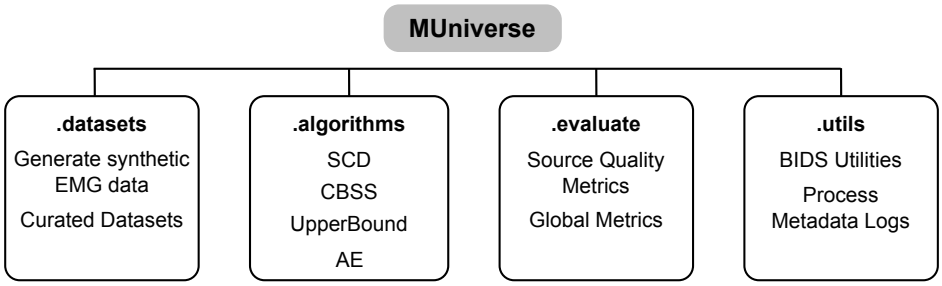

Figure 2: **MUniverse simulation and benchmarking suite**: MUniverse is organized into four high-level modules, providing a user-friendly API to (1) load and simulate data, (2) decompose HD-EMG recordings, (3) evaluate performance and (4) organise data into a standard format.

- **Decomposition routines:** the framework implements a uniform API to four decomposition methods - (1) a well-established CBSS algorithm based on FastICA [37], (2) a containerized version of the SCD algorithm [19], (3) a linear upper-bound method given ground truth MUAPs [29, 30], and (4) a deep autoencoder-based decomposition algorithm [35].
- **Evaluation and benchmarking:** we provide convenience functions to evaluate decomposition algorithms and generate report cards of algorithm performance. We implemented several common metrics to evaluate both the quality of the estimated MU spike trains (individually) and the overall reconstruction accuracy.
- **Data standardisation and logging:** We offer tools to organise experimental HD-EMG datasets into BIDS structure for interoperability with existing tools like MNE [18] that operate on this standardised format.

3.3 Usage and Access

The MUniverse codebase will be made publicly available on GitHub, providing researchers with complete access to all components. Accompanying datasets are hosted on Harvard Dataverse [3] with comprehensive documentation and made available through Croissant files, enabling efficient integration with existing workflows and analytical pipelines. This open-access approach encourages community-driven improvements to the framework.

4 Data generation and curation

4.1 Datasets overview

Our dataset collection covers muscles with different sizes and architectural properties (e.g., tibialis anterior and flexor digitorum muscles), isometric and dynamic movements, and low- to high-density electrode arrays (Table 1), and thereby rigorously challenges decomposition algorithms under realistic variability. We curated three complementary types of HD-EMG datasets – Synthetic, Hybrid, and Experimental – to span a wide spectrum of muscles, contraction types, recording configurations, and noise conditions.

1. **Synthetic:** Generated end-to-end by the NeuroMotion pipeline [33]. Here, a user-defined 'movement profile' (e.g., a trapezoidal isometric contraction of the extensor carpi muscle at 50% MVC) drives a motor neuron-pool model to produce spike trains [16], which are then convolved with 'subject'-specific MUAPs. We generated two synthetic datasets – NeuroMotion Train set and NeuroMotion Test set, of 10,000 and 985 recordings, respectively.
2. **Hybrid:** Uses the same pipeline as in the synthetic datasets, but replaces synthetic MUAPs with experimentally recorded MUAPs (see "Experimental MUAP library" below), preserving realistic spike trains while testing algorithm robustness to real filter shapes. Here, we generated a dataset of 100 recordings of isometric contractions.
3. **Experimental:** Open access HD-EMG recordings of the tibialis anterior muscle during isometric ankle dorsiflexion contractions (Caillet et al. (2023) [7]; Avrillon et al. (2024) [4]; Grison et al. (2025) [19]). For Grison et al. (2025), each file is accompanied by MU spike train

[3] `https://dataverse.harvard.edu/dataverse/muniverse`

estimates extracted in vivo from synchronised high-density *intramuscular* EMG, providing an independent, high-precision reference for benchmarking.

Table 1: MUniverse provides a curated library of datasets, that span a wide spectrum of muscles, contraction types, recording configurations and noise conditions.

Dataset Name	N. Recordings	N. Muscles	Effort (%MVC)	Movement Profiles
NeuroMotion Train	10,000	7	[10, 100]	6
NeuroMotion Test	985	7	[10, 80]	6
Hybrid Tibialis	100	1	[10:5:100]	2
Caillet et al. (2023)	11	1	{30, 50}	1
Avrillon et al. (2024)	124	2	[10:10:70]	1
Grison et al. (2025)	10	1	[10:10:70]	1

4.2 Simulated datasets: synthetic and hybrid

Our synthetic and hybrid datasets were generated through a principled experimental design framework. We employed Latin Hypercube Sampling (LHS) to efficiently explore the high-dimensional parameter space of muscle types, movement profiles, effort levels, and recording configurations. This approach ensures comprehensive coverage of possible EMG signals while minimizing dataset size and computational overhead.

Dataset properties and variants For the synthetic dataset generation pipeline, we first defined the parameter space spanning all relevant factors listed below. We then used LHS to create 10,000 parameter combinations for the train set and 985 for the test set, ensuring balanced representation across the parameter space. Each combination invoked NeuroMotion to generate the recording, while our package enabled reproducible, scalable, and distributed data generation.

- **Target Muscles:** 7 forearm muscles that control the flexion-extension and radial-ulnar deviation of the wrist; tibialis anterior muscle for the hybrid dataset.
- **Movement types:** isometric contractions, dynamic flexion-extension and radial-ulnar deviation movements.
- **Movement profiles:** trapezoid (ramp and hold), triangular, sinusoid, and ballistic contractions. The degree of freedom (DoF) movement angle was varied for dynamic contractions, while effort in terms of maximal voluntary contraction (MVC) was varied for isometric contractions.
- **Effort ranges:** 5 – 80 % MVC.
- **SNR levels:** 10 – 30 dB additive Gaussian noise.
- **Electrode configurations:** grids (10x5, 10x10) and bracelet (10x32).

For the hybrid dataset, we followed a similar approach but with the additional step of collating experimentally recorded MUAPs.

Experimental MUAP library: An experimental MUAP library and the corresponding recruitment thresholds in % MVC were extracted from experimental 256-channel recordings from eight subjects from the dataset Avrillon et al. (2024) [2, 4]. A total of 1031 MUAPs were extracted using spike-triggered averaging with a ± 25 ms window of each of the curated spike trains. To ensure a compact support of the extracted MUAPs due to non-zero edge values, we applied a Tukey window with a cosine fraction of 0.1 [30]. Then, a specified number of the 1031 MUAPs and their recruitment thresholds were randomly sampled for convolving with the generated spike trains for EMG signal generation.

From here, we proceeded to produce 100 recordings by varying the movement profile, effort levels, and noise conditions. Different subsamples of the 1031 MUAPs were obtained to simulate subject-specific parameters.

4.3 Experimental HD-EMG datasets

We re-purposed three state-of-the-art datasets (Caillet et al. (2023) [7]; Avrillon et al. (2024) [4]; Grison et al. (2025) [19]) consisting of HD-EMG signals recorded on the Tibialis Anterior muscle or Vastus Lateralis muscle. All datasets are acquired from published sources, experiments were conducted with informed consent, approved by an institutional review board, and complied with

the Declaration of Helsinki. In Caillet et al. (2023) and Avrillon et al. (2024), four 64-electrode grids were used. In these datasets, trapezoidal contractions of randomized order and in the range of 5-80% MVC were performed. In Grison et al. (2025), two 64-electrode surface grids, as well as intramuscular EMG was recorded from three multi-channel electrode arrays. Intramuscular and surface EMG signals were concurrently sampled at 10,240 Hz.

4.4 Dataset formatting and metadata

All datasets are stored in the open and standardized BIDS format [17] and are rich in metadata to ensure reusability as well as interoperability. The datasets include global metadata (e.g., participant population and protocol descriptions) as well as recording-specific metadata (e.g., hardware specifications, channel annotations, and global electrode coordinate systems). The MUniverse package includes routines for easily generating and reading BIDS datasets, facilitating the enrichment of the existing database as well as easy integration into fully automated decomposition pipelines.

5 Algorithms and benchmarking

5.1 Decomposition algorithms

The convolutive mixture in Equation 1 can be reformulated into a linear instantaneous mixture by introducing an extended vector of MU spike trains and observations, each including the original MU spike trains or observations and their delayed versions [13]. After applying a whitening transformation, spike trains can be estimated recursively by applying a projection vector to the extended and whitened signals. In MUniverse, we implemented and evaluated two of the most successful approaches to EMG decomposition (**CBSS** [37] and **SCD** [19]) along with a deep autoencoder-based decomposition (**AE** [35]) and a linear upper-bound algorithm requiring prior knowledge of the MUAPs (**UB** [30]). We describe an overview of these algorithms, where an extended description is presented in Appendix A.

FastICA-based decomposition (CBSS): The CBSS algorithm uses 25 hyperparameters, capturing signal pre-processing, signal extension, whitening, separation vector optimization, spike detection, as well as automated quality classification of the estimated spike train [37]. The optimization problem (Equation (A.3)) is iteratively optimized by a fixed point algorithm (i.e., a quasi-Newton solver with quadratic convergence) with a sparse-based contrast function. Different strategies on optimization algorithm initialization can be selected as well as strategies in preventing the algorithm from repeatedly converging to the same source (subspace projection methods as vs. peeling off MUs from the EMG signal given the estimated spike trains).

Swarm-Contrastive Decomposition (SCD): SCD preserves the CBSS backbone but replaces the fixed contrast function with a higher-order cumulant selected on-the-fly via particle-swarm search. It uses a peel-off loop to remove each accepted spike train before the next pass. In addition to the 25 hyperparameters, two extra hyperparameters (swarm size and silhouette threshold) are introduced.

Deep-learning baseline (AE): The deep learning approach [35] is fully unsupervised and casts ICA as an autoencoder (AE) problem, where the latent dimension corresponds to the number of MU spike trains. The encoder is an orthogonal rotation that maps extended and whitened observations to latent activations, encouraging separation of different MU spike trains. The decoder is a linear layer followed by tanh-shrink, mapping latents back to the whitened, extended space. Training minimizes a reconstruction term plus a sparsity penalty on the latents.

Upper-bound (UB): In a simulated dataset, given the ground truth MUAP waveforms, the upper-bound (UB) accuracy of a linear CBSS algorithm can be calculated by directly computing the projection vector. Due to the extension, there are multiple delayed copies of the same spike train. Therefore, we select the column of the extended and whitened MUAP with the largest L_2-norm [30].

5.2 Benchmarking workflow

For each recording and pipeline, we executed a standardized workflow: the raw EMG and its JSON sidecar are read into memory; the selected engine produces spike train estimates and predicted sources alongside run metadata (e.g., runtime and the selected hyperparameters). For the synthetic and hybrid datasets, this was followed by a spike train-matching procedure that aligns outputs to reference spike times. Matching was based on cross-correlation within a ±100 ms window, assigning estimated spike trains to ground-truth or intramuscular reference units by solving a linear sum assignment problem with the false positive rate as the cost function and only accepting matches when two spike trains

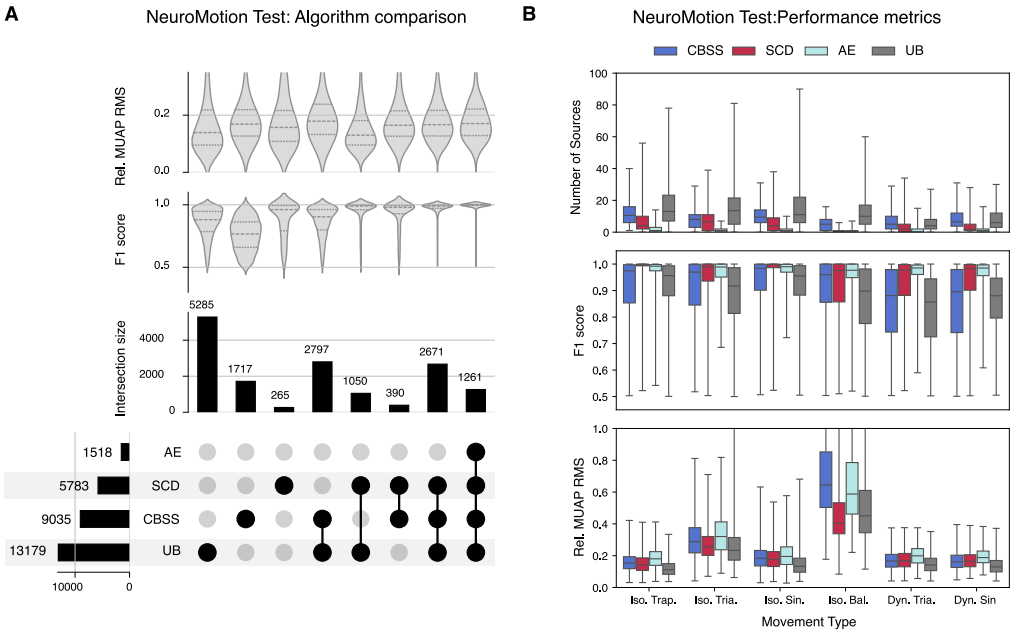

Figure 3: Decomposition performance given the NeuroMotion test dataset, considering all units with an F1 score above 0.5. (A) UpSet plot showing the total number of detected units per algorithm (horizontal bars) and the intersection sizes of units common between different sets of algorithms (vertical bars; only considering sets with a minimum number of 150 units). (B) Boxplots showing the unit yield (top), F1 score (middle) and the relative MUAP amplitude (bottom) depending on the movement type (x-axis) and algorithm (indicated by different colors).

share at least 30 % of common spikes. Finally, each pipeline's outputs were evaluated with the following metrics to produce the final results.

5.3 Evaluation metrics

Each pipeline is evaluated using two sets of metrics. The first set applies across all dataset types and includes (a) source quality metrics such as the silhouette-based score, (b) the number of identified spike trains above the quality thresholds in (a), and (c) the fraction of variance explained (FVE) by the identified and accepted spike trains (see Appendix C). The FVE is computed by recursively peeling off the contribution of each identified spike train. This residual signal lets us directly estimate the variance explained by the identified MU spike trains. The second set of metrics applies to datasets with known ground-truth spike times or expert-annotated decompositions from intramuscular recordings. For these, we computed the number of true positives, false positives, and false negatives (post cross-correlation-based spike train matching, see Appendix C). We report the number of matched units (absolute and relative), along with their F1-scores.

6 Results

Here, we showcase the performance of three different algorithms (plus the linear upper bound estimate) on 5 datasets, leading to the first comprehensive benchmarking effort in EMG decomposition.

6.1 Global summary and source quality metrics

Table 2 summarizes performance on the Avrillon et al. (2024) and Neuromotion Test datasets, chosen as a representative sample of the experimental and synthetic datasets types, respectively (see Table A.1 for the remaining datasets). We report the 10th, 50th (median), and 90th percentiles for each metric across pipelines, and provide brief comments on the most noteworthy trends below.

- **Unit yield:** For experimental data without ground-truth, the unit yield was defined as the number of units with a silhouette score above 0.9 and at least 50 detected spikes. For simulated data, the unit yield was the number of sources with an F1-score larger than 0.5. In all experimental

datasets, SCD shows the best top-level performance (indicated by the 90th percentile yield numbers) and shows the best median performance in 2 out of 3 experimental datasets. On the other hand, CBSS outperforms SCD on the simulated datasets, yet SCD tends to identify units that are of higher quality (median F-1 score for SCD on the NeuroMotion test dataset is 0.99 compared to 0.92 for CBSS). The AE algorithm, however, performs significantly worse than CBSS and SCD.

- **Explained variance:** Similar to the unit yield performance, SCD slightly outperforms CBSS in median FVE on the experimental datasets, while CBSS performs better on the simulated datasets. At the 90th percentile, SCD leads CBSS across four out of the five datasets.

6.2 Evaluation with known ground-truth spike trains

Table 3 shows decomposition performance when a ground-truth is available, i.e., an expert reference decomposition based on invasive EMG (Grision et al. (2025)) or the simulated ground truth (NeuroMotion test, Hybrid Tibialis). We present the total number of predicted units, the fraction of sources matched with respect to the ground-truth, and the percentile of the spike train accuracy quantified by means of F1 score.

- **Summary:** All testes algorithms reliably reconstruct motor unit spike trains, indicated by a median F1 score above F1 score. While SCD shows the highest unit yield in the experimental data, CBSS performed best on the simulated datasets. However, SCD is superior in terms of spike train accuracy. The AE pipeline can reliably decompose MUs; however, it currently can't compete with the established SCD and CBSS algorithms. The relatively low fraction of matched sources for UB (75.8 %) is due to the fact that reconstructing the activity of a full MU pool is often not feasible even if the forward model is known.
- **Insights from simulations:** Figure 3 exemplarily shows for the NeuroMotion test dataset how simulations enable detailed insights into the absolute and relative performance of different decomposition algorithms. We note that there is still a considerable gap between existing decomposition methods and the theoretical upper bound (Figure 3A), however, these units are the most challenging which is represented in the F1 score distribution. Generally, units that are detected by all algorithms show the highest accuracy. Though CBSS shows the highest unit yield, SCD is best in detecting low-amplitude units. Considering different tasks (Figure 3B), it can be observed that for all algorithms the unit yield and decomposition accuracy decrease for dynamic or ballistic contractions. Moreover, it is observed that though CBSS shows the best median unit yield, SCD shows the highest best-case unit yield.

For further details regarding the selected hyperparameters and an in-depth analysis of the Hybrid Tibialis dataset, see Appendix B.

Table 2: High-level performance metrics across datasets and algorithms. (10 | 50 | 90 percentile)

Dataset	Algorithm	Unit yield	FVE	Runtime (s)
Avrillon et al. (2024)	CBSS	0.0 | 14.5 | 52.7	0.07 | 0.34 | 0.61	85 | 195 | 288
	SCD	0.0 | 13.0 | 69.5	0.00 | 0.34 | 0.67	86 | 348 | 805
	AE	2.0 | 10.0 | 37.0	0.00 | 0.10 | 0.23	61 | 88 | 126
NeuroMotion Test	CBSS	1.0 | 8.0 | 18.0	0.07 | 0.27 | 0.46	126 | 381 | 1046
	SCD	0.0 | 3.0 | 16.0	0.00 | 0.12 | 0.47	0 | 147 | 399
	AE	0.0 | 1.0 | 4.0	0.00 | 0.04 | 0.18	0 | 13 | 296
	UB	2.0 | 9.0 | 32.0	0.05 | 0.22 | 0.45	4 | 11 | 90

7 Discussion, limitations and future work

Using the proposed MUniverse framework, we perform a standardised comparison of 3 decomposition pipelines across five synthetic and experimental EMG datasets (1230 recordings), marking the largest such effort ever. Despite the different optimisation strategies and algorithms, all algorithms showed remarkably consistent median performance across experimental and simulated datasets, encouraging neuroscientists and neurophysiologists using these methods to draw robust inferences. No single decomposition pipeline can be considered optimal for arbitrary recordings. Yet, the presented results provide guidance for users in selecting an appropriate decomposition algorithm for specific

Table 3: Decomposition performance with respect to expert reference decomposition (intramuscular recording in the Grison et al. (2025) dataset) or simulated ground truth. (10 | 50 | 90 percentile)

Dataset	Algorithm	N. Sources	Matched Sources (%)	F1-score
Grison et al. (2025)	CBSS	272	12.1	0.74 | 0.92 | 0.98
	SCD	325	11.1	0.78 | 0.92 | 0.99
	AE	159	11.9	0.65 | 0.92 | 0.99
NeuroMotion Test	CBSS	11198	86.0	0.60 | 0.96 | 1.00
	SCD	6648	90.1	0.79 | 1.00 | 1.00
	AE	1717	95.5	0.81 | 0.99 | 1.0
	UB	18040	75.8	0.69 | 0.93 | 1.0
Hybrid Tibialis	CBSS	1420	98.4	0.69 | 0.92 | 0.99
	SCD	843	99.5	0.81 | 0.99 | 1.00
	AE	228	97.4	0.77| 0.97 | 0.99
	UB	4217	97.7	0.79 | 0.91 | 0.98

applications. For example, CBSS is most robust with respect to signal non-stationarities, and SCD is most reliable in fully automated settings. Notably, the inclusion of a theoretical upper bound demonstrates that even for isometric tasks, existing decomposition methods need to be further improved to lower the gap with respect to the theoretical optimal performance. Further, although the deep-learning baseline can currently not compete with established decomposition pipelines, for the first time, we show that an unsupervised neural network architecture can decompose motor unit activity from HDsEMG.

While MUniverse addresses a critical gap in standardized EMG decomposition benchmarks, several limitations remain that we plan to tackle in future releases. In this paper, we have avoided a hyperparameter search for each recording (which is typically done in practice) and instead on a heuristic hyperparameter search over a subset of the most important hyperparameters to manage compute resources. The estimated spike trains have not been post-processed as is typically done in experimental studies. Further, despite the versatility, our current NeuroMotion integration generates EMG only for a single muscle at the moment. In future versions, we will make the API more flexible to be able to directly take kinematic variables as inputs, instead of defining them in a confined manner. Finally, some widely used packages (e.g., MUEdit, Demuse) could not be included in the benchmark. We invite their maintainers to contribute containerized wrappers.

8 Conclusion

We have introduced MUniverse, an open, extensible simulation and benchmarking suite for EMG decomposition that unifies data generation, algorithm interfaces, and evaluation under FAIR principles. By providing both massive synthetic corpora and curated experimental datasets alongside containerized decomposition pipelines and standardized metrics, MUniverse lays the foundation for reproducible comparison and rapid method development. We invite the community to adopt MUniverse for fair benchmarking, to contribute new datasets and algorithms, and to collaboratively drive forward the accuracy and reliability of EMG-based neural interfacing.

9 Broader Impact

MUniverse democratizes access to high-fidelity EMG data and transparent evaluation, accelerating neural-engineering research from basic science to assistive technologies. By lowering barriers to method development, it has the potential to improve clinical diagnostics, prosthetic control, and our understanding of motor disorders—but also underscores the responsibility to validate these methods across diverse populations and tasks before clinical deployment.

References

[1] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. M. Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Müller. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific data*, 1(1):1–13, 2014.

[2] S. Avrillon. Data for the research article 'The decoding of extensive samples of motor units in human muscles reveals the rate coding of entire motoneuron pools', 11 2023.

[3] S. Avrillon, F. Hug, S. N. Baker, C. Gibbs, and D. Farina. Tutorial on MUedit: An open-source software for identifying and analysing the discharge timing of motor units from electromyographic signals. *Journal of Electromyography and Kinesiology*, 77:102886, 2024.

[4] S. Avrillon, F. Hug, R. Enoka, A. H. Caillet, and D. Farina. The decoding of extensive samples of motor units in human muscles reveals the rate coding of entire motoneuron pools. *eLife*, 13, 2024.

[5] A. P. Buccino, C. L. Hurwitz, S. Garcia, J. Magland, J. H. Siegle, R. Hurwitz, and M. H. Hennig. Spikeinterface, a unified framework for spike sorting. *Elife*, 9:e61834, 2020.

[6] A. Caillet. Data for the research article 'Larger and Denser: An Optimal Design for Surface Grids of EMG Electrodes to Identify Greater and More Representative Samples of Motor Units', 8 2023.

[7] A. H. Caillet, S. Avrillon, A. Kundu, T. Yu, A. T. Phillips, L. Modenese, and D. Farina. Larger and denser: an optimal design for surface grids of EMG electrodes to identify greater and more representative samples of motor units. *eNeuro*, 10(9), 2023.

[8] M. Chen and P. Zhou. A novel framework based on FastICA for high density surface EMG decomposition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(1):117–127, 2015.

[9] A. K. Clarke, S. F. Atashzar, A. Del Vecchio, D. Barsakcioglu, S. Muceli, P. Bentley, F. Urh, A. Holobar, and D. Farina. Deep learning for robust decomposition of high-density surface EMG signals. *IEEE Transactions on Biomedical Engineering*, 68(2):526–534, 2020.

[10] C. J. De Luca, A. Adam, R. Wotiz, L. D. Gilmore, and S. H. Nawab. Decomposition of surface EMG signals. *Journal of Neurophysiology*, 96(3):1646–1657, 2006.

[11] E. Eddy, E. Campbell, C. Morrell, H. Williams, S. Bateman, and E. Scheme. Raising the standard: an open source benchmarking platform and data repository to accelerate myoelectric control research. *Machine Learning: Health*, 1(1):010601, 2025.

[12] J. Eden, M. Bräcklein, J. Ibáñez, D. Y. Barsakcioglu, G. Di Pino, D. Farina, E. Burdet, and C. Mehring. Principles of human movement augmentation and the challenges in making it a reality. *Nature Communications*, 13(1):1345, Mar. 2022. Number: 1 Publisher: Nature Publishing Group.

[13] D. Farina and A. Holobar. Characterization of human motor units from surface EMG decomposition. *Proceedings of the IEEE*, 104(2):353–373, 2016.

[14] D. Farina, R. Merletti, and R. M. Enoka. The extraction of neural strategies from the surface EMG: an update. *Journal of Applied Physiology*, 117(11):1215–1230, 2014.

[15] D. Farina, I. Vujaklija, R. Brånemark, A. M. J. Bull, H. Dietl, B. Graimann, L. J. Hargrove, K.-P. Hoffmann, H. H. Huang, T. Ingvarsson, H. B. Janusson, K. Kristjánsson, T. Kuiken, S. Micera, T. Stieglitz, A. Sturma, D. Tyler, R. F. f. Weir, and O. C. Aszmann. Toward higher-performance bionic limbs for wider clinical use. *Nature Biomedical Engineering*, 7(4):473–485, Apr. 2023. Publisher: Nature Publishing Group.

[16] A. J. Fuglevand, D. A. Winter, and A. E. Patla. Models of recruitment and rate coding organization in motor-unit pools. *Journal of Neurophysiology*, 70(6):2470–2488, 1993.

[17] K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, D. A. Handwerker, M. Hanke, D. Keator, X. Li, Z. Michael, C. Maumet, B. N. Nichols, T. E. Nichols, J. Pellman, J.-B. Poline, A. Rokem, G. Schaefer, V. Sochat, W. Triplett, J. A. Turner, G. Varoquaux, and R. A. Poldrack. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3(1):160044, June 2016.

[18] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013.

[19] A. Grison, I. Mendez Guerra, A. K. Clarke, S. Muceli, J. Ibáñez, and D. Farina. Unlocking the full potential of high-density surface EMG: novel non-invasive high-yield motor unit decomposition. *The Journal of Physiology*, 603(8):2281–2300, 2025.

[20] A. Holobar, M. A. Minetto, and D. Farina. Accurate identification of motor unit discharge patterns from high-density surface EMG and validation with a novel signal-based performance metric. *Journal of Neural Engineering*, 11(1):016008, 2014.

[21] A. Holobar and D. Zazula. Gradient convolution kernel compensation applied to surface electromyograms. In *International Conference on Independent Component Analysis and Signal Separation*, pages 617–624. Springer, 2007.

[22] A. Holobar and D. Zazula. Multichannel blind source separation using convolution kernel compensation. *IEEE Transactions on Signal Processing*, 55(9):4487–4496, 2007.

[23] F. Hug, S. Avrillon, J. Ibáñez, D. Farina, and J. Physiol. Common synaptic input, synergies and size principle: Control of spinal motor neurons for movement generation. *The Journal of Physiology*, 601(1):11–20, 1 2023.

[24] A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

[25] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[26] A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pages 859–868. PMLR, 2019.

[27] X. Jiang, X. Liu, J. Fan, X. Ye, C. Dai, E. A. Clancy, M. Akay, and W. Chen. Open access dataset, toolbox and benchmark processing results of high-density surface electromyogram recordings. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1035–1046, 2021.

[28] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.

[29] T. Klotz, L. Lehmann, F. Negro, and O. Röhrle. High-density magnetomyography is superior to high-density surface electromyography for motor unit decomposition: a simulation study. *Journal of Neural Engineering*, 20(4):046022, 2023.

[30] T. Klotz and R. Rohlén. Revisiting convolutive blind source separation for identifying spiking motor neuron activity: From theory to practice. *arXiv preprint arXiv:2502.04065*, 2025.

[31] J. Ma, L. Wang, R. Wu, N. Zhang, J. Wei, J. Li, Q. Li, L. Tan, G. Li, N. Jiang, et al. A multi-label deep residual shrinkage network for high-density surface electromyography decomposition in real-time. *Journal of NeuroEngineering and Rehabilitation*, 22(1):1–19, 2025.

[32] S. Ma, A. K. Clarke, K. Maksymenko, S. Deslauriers-Gauthier, X. Sheng, X. Zhu, and D. Farina. Conditional generative models for simulation of EMG during naturalistic movements. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[33] S. Ma, I. Mendez Guerra, A. H. Caillet, J. Zhao, A. K. Clarke, K. Maksymenko, S. Deslauriers-Gauthier, X. Sheng, X. Zhu, and D. Farina. NeuroMotion: Open-source platform with neuromechanical and deep network modules to generate surface EMG signals during voluntary movement. *PLOS Computational Biology*, 20(7):e1012257, 2024.

[34] J. Magland, J. J. Jun, E. Lovero, A. J. Morley, C. L. Hurwitz, A. P. Buccino, S. Garcia, and A. H. Barnett. SpikeForest, reproducible web-facing ground-truth validation of automated neural spike sorters. *eLife*, 9:e55167, 2020.

[35] K. M. Mayer, A. Del Vecchio, B. M. Eskofier, and D. Farina. Unsupervised neural decoding of signals recorded by thin-film electrode arrays implanted in muscles using autoencoding with a physiologically derived optimisation criterion. *Biomedical Signal Processing and Control*, 86:105178, 2023.

[36] S. H. Nawab, S.-S. Chang, and C. J. De Luca. High-yield decomposition of surface EMG signals. *Clinical Neurophysiology*, 121(10):1602–1615, 2010.

[37] F. Negro, S. Muceli, A. M. Castronovo, A. Holobar, and D. Farina. Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation. *Journal of Neural Engineering*, 13(2):026027, 2016.

[38] M. Piotrkiewicz, O. Sebik, E. Binboğa, D. Młoźniak, B. Kuraszkiewicz, and K. S. Türker. Double discharges in human soleus muscle. *Frontiers in Human Neuroscience*, 7:843, 2013.

[39] J. Rossato, F. Hug, K. Tucker, C. Gibbs, L. Lacourpaille, D. Farina, and S. Avrillon. I-Spin live, an open-source software based on blind-source separation for real-time decoding of motor unit activity in humans. *eLife*, 12:RP88670, 2024.

[40] S. Salter, R. Warren, C. Schlager, A. Spurr, S. Han, R. Bhasin, Y. Cai, P. Walkington, A. Bolarinwa, R. J. Wang, et al. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. *Advances in Neural Information Processing Systems*, 37:55703–55728, 2024.

[41] V. Sivakumar, J. Seely, A. Du, S. Bittner, A. Berenzweig, A. Bolarinwa, A. Gramfort, and M. Mandel. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography. *Advances in Neural Information Processing Systems*, 37:91373–91389, 2024.

[42] G. Valli, P. Ritsche, A. Casolo, F. Negro, and G. De Vito. Tutorial: Analysis of central and peripheral motor unit properties from decomposed high-density surface emg signals with openhdemg. *Journal of Electromyography and Kinesiology*, 74:102850, 2024.

[43] Y. Wen, S. J. Kim, S. Avrillon, J. T. Levine, F. Hug, and J. L. Pons. A deep cnn framework for neural drive estimation from hd-emg across contraction intensities and joint angles. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2950–2959, 2022.

Appendix

A Decomposition algorithms

Many successful BSS methods, such as ICA, consider the inversion of an (often ill-posed) linear system. Such methods can be applied to the decomposition of EMG signals as a convolutive mixture with finite impulse response filters (see Equation 1) can always be written in terms of a linear instantaneous mixture. This reformulation requires introducing an extended vector of MU spike trains and observations, each including the original MU spike trains or observations and their R delayed versions [13]:

$$\tilde{\mathbf{x}}(t) \;=\; \tilde{\mathbf{H}}\tilde{\mathbf{s}}(t) \;+\; \tilde{\boldsymbol{\varepsilon}}(t)\;. \tag{A.1}$$

The signal extension is followed by a whitening transformation, i.e., $\tilde{\mathbf{z}}(t) = \mathbf{V}\tilde{\mathbf{x}}(t)$ where the (non-unique) whitening matrix $\mathbf{V}$ is constructed such that the covariance matrix of extended and whitened observations $\tilde{\mathbf{z}}(t)$ is the identity matrix. This transformation approximately orthogonalizes the mixing matrix and hence, spike trains can be estimated recursively by applying a projection vector $\mathbf{w}_k^T$ to the extended and whitened multichannel EMG signals $\tilde{\mathbf{z}}(t)$:

$$\widehat{s}_k(t) \;=\; \mathbf{w}_k^T\tilde{\mathbf{z}}(t)\;. \tag{A.2}$$

Therein $\widehat{s}_k(t)$ is a non-unique, arbitrarily scaled, and potentially delayed estimate of the kth MU spike train. Different variants of convolutive BSS vary regarding the methods used to estimate the columns $\mathbf{w}_k$ of the inverse mixing matrix.

FastICA-based decomposition (CBSS): MUniverse contains an implementation of a state-of-the-art convolutive BSS (CBSS) algorithm based on FastICA; for details, see [37]. The algorithm uses 25 hyperparameters, capturing signal pre-processing, signal extension, whitening, separation vector optimization, spike detection, as well as automated quality classification of the estimated spike train. In contrast to existing implementations, the objective function is

$$\mathcal{L}^{\mathrm{cbss}}(\mathbf{w}_k) \;=\; \sum_t \mathbb{E}\Big(G(\mathbf{w}_k^T\tilde{\mathbf{z}}(t))\Big)\;, \qquad G(s) \;=\; s(s^2+\epsilon)^{\frac{a-1}{2}}\;, \tag{A.3}$$

where $\epsilon = 0.001$. Adjusting the parameter $a > 0$ allows for fine-tuning the degree of non-linearity based on a simple continuous and infinitely differentiable function. Notably, selecting $a = 3$ is (approximately) equivalent to using skewness as the objective function. The optimization problem posed by Equation (A.3) is iteratively optimized by a fixed point algorithm (i.e., a quasi-Newton solver with quadratic convergence), and different initialization strategies can be selected (random weights or activity index [21]). Moreover, the user can select between different strategies in preventing the algorithm from repeatedly converging to the same source (i.e., different subspace projection methods as well as peeling off MUs from the EMG signal given the estimated spike trains).

Swarm-Contrastive Decomposition (SCD): MUniverse integrates Swarm-Contrastive Decomposition, which preserves the ICA backbone but replaces the fixed contrast with a *source-specific* higher-order cumulant selected on-the-fly via particle-swarm search. For each projection vector $\mathbf{w}_k$ it maximises

$$\mathcal{L}^{\mathrm{scd}}(\mathbf{w}_k) \;=\; \sum_t \mathbb{E}\Big(G(\mathbf{w}_k^T\tilde{\mathbf{z}}(t))\Big)\;, \qquad G(s) \;=\; \mathrm{sign}(s)\,|s|^a\;, \tag{A.4}$$

where the exponent a is optimized, using particle swarm optimization, for every candidate spike train estimate, and a peel-off loop removes each accepted spike train before the next pass. Further, $\mathrm{sgn}(s)$ denotes the signum function. Two extra hyperparameters (swarm size and silhouette threshold) are introduced, allowing SCD to slot into existing frameworks.

Deep-learning baseline (AE): As a deep-learning baseline, Muniverse includes an implementation of the autoencoder approach [35]. This method casts ICA as an autoencoder problem, where the latent dimension corresponds to the number of sources. The encoder is an orthogonal rotation $V \in \mathrm{SO}(mR)$ that maps extended-whitened observations to latent activations, encouraging separation of different sources. The decoder is a linear layer followed by tanhshrink, mapping latents back to the observation space. Training minimizes a reconstruction term plus a sparsity penalty on the latents:

$$\mathcal{L} = \|\tilde{\mathbf{x}}_w - \hat{\mathbf{x}}\|_2^2 \;+\; \lambda\,\log_{10}\!\left(\frac{q}{p}\,\frac{\|\mathbf{s}\|_p}{\|\mathbf{s}\|_q}\right), \qquad 0 < p < q. \tag{A.5}$$

The method is fully unsupervised, aligns with the classical "rotation-after-whitening" view from ICA, and applies to both iEMG and HD-sEMG. After training, a simple peak finding algorithm is used on the latents to produce the estimated sources.

Upper-bound (UB): For robust learning, the projection vector $\mathbf{w}_k^T$ has been shown to converge to a scaled version of the extended and whitened MUAP [30]. Therefore, in a simulated dataset, the upper-bound accuracy of a CBSS algorithm can be calculated by directly computing the projection vector given the ground truth MUAP waveforms. In detail, the spike train of MU k with delay l is

$$\widehat{\tilde{\mathbf{s}}}_{k,l}(t) \;=\; \frac{\tilde{\mathbf{h}}_{k,l}^T}{||\tilde{\mathbf{h}}_{k,l}||}\,\tilde{\mathbf{z}}(t)\;, \tag{A.6}$$

where $\tilde{\mathbf{h}}_{k,l}$ is a single column of the extended and whitened mixing matrix associated with MU k and $||*||$ denotes the L_2-norm. Notably, due to the extension, there are multiple delayed copies of the same spike train. Hence, as for the upper-bound estimate, we maximized the expected spike train amplitude by selecting the column of the extended and whitened MUAP with the largest L_2-norm [30].

Spike train estimation Finally, binary spike trains need to be derived from the estimated MU spike trains $\widehat{s}_k(t)$. Therefore, an asymmetric power function, i.e., $G(s) \;=\; \mathrm{sgn}(s)\cdot|s|^a$ is used to enhance the contrast between the MU spikes and background peaks, where the exponent $a \in \mathbb{R}$ is an adjustable parameter (here $a = 2$). The tallest peaks with a minimal distance of 10 ms were extracted, and K-means clustering with $K = 2$ was used to separate the peak heights into two clusters, where the putative MU spikes are those in the cluster with the largest centroid. Based on these clusters, a silhouette-based score was computed as a quality metric for the estimated spike trains [37].

B Extended results

Here we present an extended analysis of the performance of the algorithms on the Hybrid tibialis dataset. We analyze both global and source-level properties of the decomposition results, and compare their relative performances on recordings of varying difficulty. Further, we present results of our hyperparameter optimisation experiments to quantify how a subset of the most important hyperparameters affect the decomposition performance.

B.1 Insights from simulations

The Hybrid Tibialis dataset strongly mimics experimental conditions that have been extensively used to validate and optimize existing decomposition pipelines. Thereby, all tested algorithms show robust spike train reconstructions, indicated by a median F-1 score larger than 0.9 (Figure A.1B). Nevertheless, Figure A.1A shows that existing decomposition pipelines have not reached the theoretical optimal performance, indicated by 2580 units uniquely detected by UB. While the CBSS algorithm is currently the closest competitor, the SCD algorithm is more refined with respect to identifying low amplitude units. Further, due to the fact that the SCD prediction show the highest F-1 score, it highlights the methods benefits in fully automated settings (i.e., without human-in-the-loop post-processing). Further, the presented data suggests that in triangular contractions less low amplitude units can be detected than in trapezoidal contractions (Figure A.1B, bottom).

B.2 Influence of hyper-parameters

The performance of decomposition algorithms critically depends on the selected hyper-parameters. Achieving an optimal decomposition performance often requires a per-recording hyperparameter optimization. Although a hyper-parameter optimisation is beyond the scope of the proposed manuscript, in this section we exemplarily explore the influence of two hyper-parameters on the performance of the CBSS decomposition algorithm. That is, the extension factor R and the strategy for avoiding finding multiple copies of the same source (subspace projection using Gram-Schmidt orthogonalization vs estimating single source signal contributions and subsequent peel-off). For this purpose, we quasi-randomly selected a subset of 50 recordings from the NeuroMotion test dataset and decomposed it with 6 sets of hyperparameters (i.e., $R = 5, 10, 20$ and using either peel-off or subspace projection). It can be observed that the worst set of hyper-parameters (i.e., $R = 5$ and using subspace projection) detects 72.8 % of the MUs detected by the best performing set of hyperparameters ($R = 10$ and using peel-off), see Figure A.2. Further, for both the peel-off and subspace projection-based approaches the highest MU yield is obtained for an extension factor of $R = 10$. Interestingly, each set of hyper-parameters detects unique MUs (i.e., sources that are not detected by other hyperparameter sets).

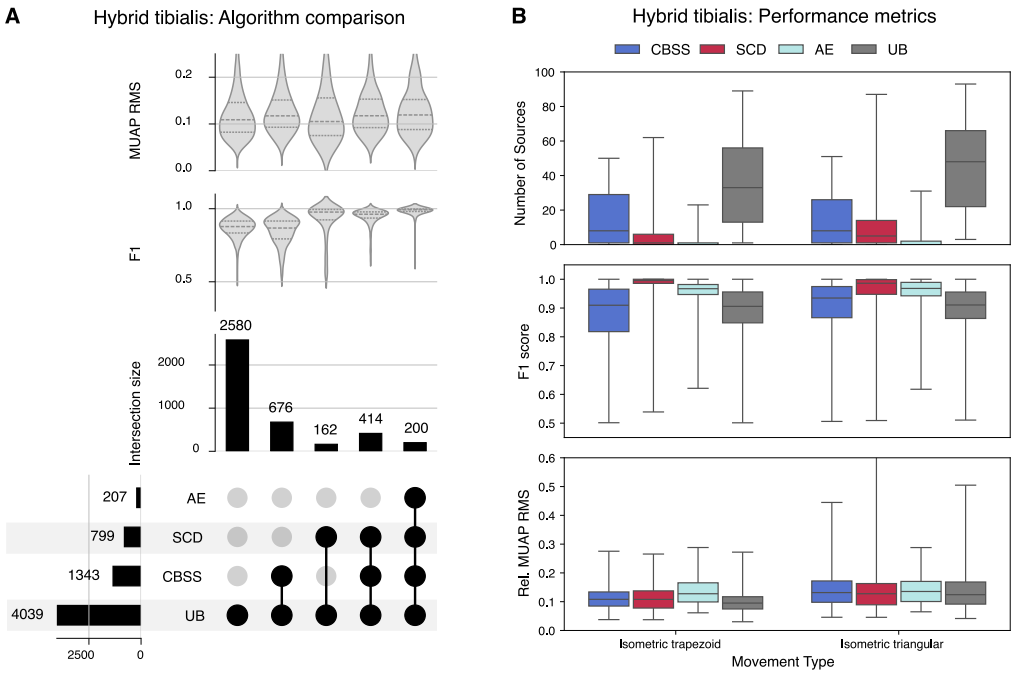

Figure A.1: Decomposition performance on the Hybrid tibialis dataset considering all units with a F1 score above 0.5. (A) UpSet plots showing the total number of detected units per algorithm (horizontal bars) and the intersection sizes of units common between different sets of algorithms (vertical bars; only considering sets with a minimum of 50 units). (B) Boxplots showing the unit yield (top), F-1 score (middle) and the relative MUAP amplitude (bottom) depending on the movement type (x-axis) and algorithm (indicated by different colors).

Table A.1: High-level performance metrics across datasets and algorithms. (10 | 50 | 90 percentile). Bold font indicates best performing algorithm in terms of median yield; UB algorithm highlighted in gray to indicate that it is a linear upperbound estimate.

Dataset	Algorithm	Unit yield	FVE	Runtime (s)
Caillet et al. (2023)	CBSS	4.0 | 32.0 | 48.0	0.33 | 0.51 | 0.59	194 | 254 | 308
	SCD	**20.0 | 38.0 | 60.0**	**0.39 | 0.57 | 0.61**	314 | 504 | 889
	AE	9.0 | 15.0 | 24.0	0.17 | 0.21 | 0.24	77 | 99 | 120
Grison et al. (2025)	CBSS	24.9 | 25.5 | 30.0	0.21 | 0.27 | 0.37	574 | 888 | 1514
	SCD	**27.0 | 32.5 | 34.5**	**0.22 | 0.32 | 0.40**	545 | 802 | 1206
	AE	8.9 | 10.0 | 11.1	0.13 | 0.18 | 0.20	87 | 162 | 186
Hybrid Tibialis	**CBSS**	**0.0 | 8.0 | 35.2**	**0.00 | 0.15 | 0.48**	0 | 623 | 952
	SCD	0.0 | 3.0 | 24.0	0.00 | 0.06 | 0.43	0 | 316 | 718
	AE	0.0 | 0.0 | 6.0	0.00 | 0.00 | 0.15	0 | 0 | 201
	UB	8.0 | 41.5 | 80.3	0.11 | 0.30 | 0.45	50 | 64 | 79

However, these unique sources contain more errors (i.e., false-positive and false-negative spikes), which is indicated by the fact that the corresponding F1-score distributions are shifted towards lower values.

C Performance metrics

The MUniverse evaluation module provides a large set of performance metrics that facilitate a standardized and holistic comparison of different decomposition algorithms. While labeled ground truth spikes always exist for simulated data, this is only partially possible for experimental recordings,

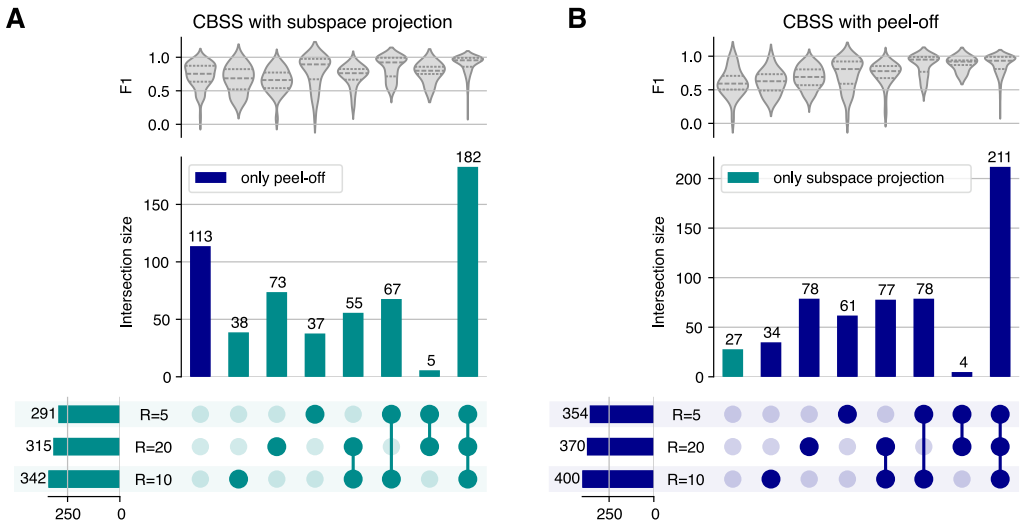

Figure A.2: UpSet plots comparing the performance of different hyper-parameter sets using either subspace-projection (A) or peel-off (B) for avoiding repeated convergence to the same sources as well as different extension factors R. The horizontal bars indicate the overall number of detected MUs per hyper-parameter set. The vertical bars indicate the intersection size of different hyper-parameter sets.

e.g., simultaneously recording invasive signals together with expert knowledge. Thus, next to spike matching-based metrics (Appendix C.1), MUniverse provides several purely signal-based source quality estimates (Appendix C.2).

C.1 Spike-based quality metrics

First, the predicted and ground-truth spike trains are matched by computing for each pair of motor neuron discharges the fraction of common spikes. To do so, each pair of spike trains is aligned in the time domain by computing the lag that maximizes the cross-correlation function. All pairs of spikes with a maximal delay of ± 1 ms are considered common spikes. We label the predicted spike trains based on the highest fraction of common spikes, whereby it is required that two spike trains corresponding to the same MU have a minimum of 30 % common spikes. Next, we compute the total number of matched spike trains as well as the fraction of matched spike trains, i.e., the total number of matched spike trains divided by the total number of ground truth spikes. Further insights into the decomposition performance are obtained by computing the rate-of-agreement (RoA), precision, recall, and the F1-score:

$$\text{RoA} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{A.7a}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{A.7b}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{A.7c}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{A.7d}$$

Therein, TP deontes the number of true positive spikes, i.e., spikes that appear in both spike trains with a maximum delay of ± 1 ms, FP are is the number of false positive spikes that are only part of the predicted spike train, and FN is the number of false positive spikes that are only part of the ground truth spike train.

C.2 Signal-based quality metrics

For quantifying the uncertainty of the predicted sources, we consider a set of quality metrics only relying on measurable quantities (i.e, the EMG signals) together with the predicted sources. First, we compute for each set of predicted spike trains the variance of the EMG signal that can be explained

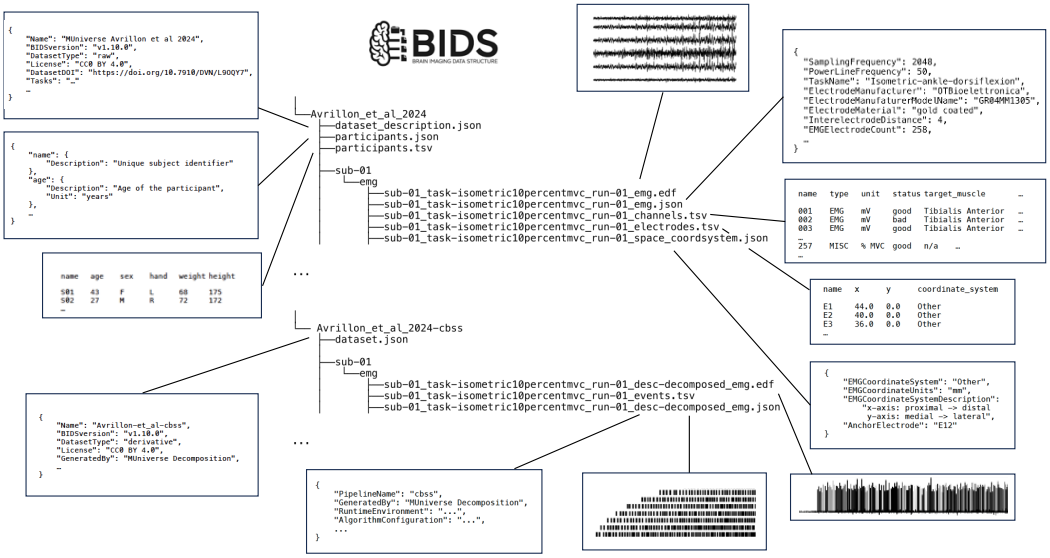

Figure A.3: **Overview of the BIDS specification.** Exemplary illustration of the BIDS folder structure and files used for storing data and metadata in a simultaneously human and machine-readable format. The upper part showcases the BIDS-EMG specification, and the lower part illustrates how decomposition outputs are formatted as BIDS-derivatives.

by the linear convolutive mixture model described in Equation (1). This is achieved by estimating for each predicted MU source the impulse response using spike-triggered averaging (in a window of ± 50 ms around the spike times), and peeling of the MU contribution obtained by convolving the estimated impulse response with the predicted spike trains. The fraction of explained variance (FVE) is given by

$$\text{FVE} \;=\; 1 \;-\; \frac{\text{Var}\left(\mathbf{x}^{\text{res}}(t)\right)}{\text{Var}\left(\mathbf{x}(t)\right)}\,, \tag{A.8}$$

where $\mathbf{x}^{\text{res}}(t)$ is the residual signal obtained after peeling off every predicted source. Purely source-based source quality metrics have been proven useful; however, all metrics face limitations. Thus, MUniverse allows users to simultaneously analyse multiple metrics, and we compute for each source the silhouette score [37], the pulse-to-noise ratio [20], the relative peak separation [30], skewness, kurtosis, the (z-scored) mean peak amplitude, and the mean discharge rate, as well as the coefficient of variation of the interspike intervals.

D MUniverse Datasets

D.1 FAIR principles

MUniverse is built on the premise that the EMG research community can hugely benefit from shared, rigorously standardised, and fully reproducible resources. By bringing various types of HD-EMG signals into a single BIDS-compatible library tied to a transparent evaluation suite, it invites community-driven extension.

FAIR compliance:

- **Findable (F):** Each dataset is issued a Harvard Dataverse DOI and a persistent URL; the collection is further described by a machine-readable Croissant JSON-LD file that indexes every file within the dataset with a version tag, and SHA-256 checksum.
- **Accessible (A):** Harvard Dataverse's open-source platform serves the datasers over HTTPS and via several python/ HTTPS endpoints, so both humans and automated agents can fetch the data without login, paywall, or proprietary tooling.
- **Interoperable (I):** All files follow the BIDS (EMG extension) schema (also see Section D.2) and are encapsulated in a Croissant JSON-LD, so open-source tools like the BIDS-validator or MNE-Python can parse them out-of-the-box.

- **Reusable (R):** Signals are licensed under CC-BY-4.0, and the entire simulation / preprocessing pipeline will be made openly accessible via an MIT license. Due to the rich provenance (simulation config JSONs, container digest, environment variables, etc.) that travels with every recording, MUniverse promotes reusability of both the datasets and code.

D.2 BIDS specification

BIDS is a standardized format for organizing and sharing anatomical and physiological data, to facilitate reproducibility, interoperability, data sharing and the use of automated data processing pipelines. Every BIDS compatible dataset uses a standardized folder structure and naming conventions (see Figure A.3) that are unambiguously related to essential recording metadata such as subject identifiers, the applied protocols and the signal modality. Besides storing the actual data in standardized formats (e.g., EDF), BIDS datasets are rich in metadata stored in JSON or TSV files that are both human and machine readable. While BIDS was originally developed for brain imaging data, it has been proven flexible to capture diverse anatomical or physiological data as well as processed data that is referred to as derivatives. An extension proposal for EMG is currently under review[4] and all MUniverse datasets make use of this preliminary standard. Further, all decomposition results are formatted according to the specifications of BIDS derivatives. The MUniverse utility module contains all essential functionalities to easily load existing BIDS datasets and export new datasets in the BIDS format.

E Author contributions and funding

	PM*	TK*	DH*	AG	IMG	SM	AC	SA	RR	DF
Conceptualization	✓	✓	✓						✓	
Methodology	✓	✓	✓						✓	
Software	✓	✓	✓	✓	✓					
Validation	✓	✓	✓		✓					
Formal Analysis	✓	✓	✓						✓	
Investigation	✓	✓	✓							
Resources – Simulator code						✓	✓			
Resources – Algorithm code		✓	✓	✓						
Resources – Source data				✓			✓	✓		
Data Curation	✓	✓		✓			✓	✓	✓	
Writing – Original draft	✓	✓	✓						✓	
Writing – Review & editing	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Visualization	✓	✓								
Supervision										✓
Project administration	✓									
Funding acquisition	✓	✓			✓				✓	✓

* These authors contributed equally to this work.

P.M. and I.M.G. were supported by the Eric and Wendy Schmidt Postdoctoral Fellowship in AI for Science at the I-X Center for AI in Science, Imperial College London. **T.K.** was supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) through the priority program SPP 2311 (Grant ID: 548605919) and the European Research Council (ERC) through ERC-AdG 'qMOTION' (Grant ID: 101055186). **D. H.** was supported by the Imperial-META Wearable Neural Interfaces Research Centre and the Onassis Foundation under Scholarship ID: F ZT 012-1/2023-2024. **R.R.** was supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe Guarantee (Grant ID: EP/Z002184/1) and the Swedish Brain Foundation (Grant ID: PS2022-0021).

[4] `https://bids-specification--1998.org.readthedocs.build/en/1998/`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract in all aspects of the paper - datasets, algorithms, and benchmarks, are well justified by our results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have explicitly addressed limitations in a dedicated section of the paper, clearly outlining relevant issues.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We fully disclose all the information needed to reproduce the experimental results. Our benchmarking workflow has a dedicated logger to maximize reproducibility, and leaves a data provenance trace at every step.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be published under an MIT license on github, as it is currently hosted anonymously at `https://anonymous.4open.science/r/muniverse-5F37` and sufficiently well documented to reproduce the main results of the benchmark. All hyper parameter values and ranges are clearly specified in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We discuss about the hyperparameter choices of all algorithms tested, as well as keep a rigorous log throughout our benchmarking workflow of the algorithm configurations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide an exploratory analysis of the performance of all algorithms on the datasets, and report 10, 50, and 90 percentile scores of all evaluated metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We both provide an overview of run-times and hardware, as well as make available detailed process metadata containing several resource related details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Ethics statement given for all experiments.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We address the broader societal impacts of method development in neural source separation, specifically EMG decomposition, for the fields of neural engineering and motor neuroscience.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The work does not involve significant risks requiring safeguards for responsible data or model release.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets utilized originate from researchers included among the paper's authors. Each dataset has been made available via a CC-BY 4.0 license and is published on the Harvard Dataverse site along with a curated croissant file.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All code and data is well documented and executable upon publishing.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Participants involved in the original experiments (whose datasets we repurpose) were instructed to perform isometric contractions of the right tibialis anterior muscle.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: All original experiments (whose datasets we repurpose) were conducted with informed consent, approved by an IRB, and complied with the Declaration of Helsinki.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were only used for writing, editing and formatting.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.