# KinEMbed: Decoding Kinematics from Electromyography via Cross-Modal Contrastive Learning

**Anonymous Authors**[1]

## Abstract

Decoding hand kinematics from surface electromyography (EMG) is a core challenge in wearable biosignal processing with clinical relevance for prosthetic control and motor rehabilitation. Most representation learning approaches for EMG focus on discrete gesture classification, and few focus on continuous regression. We present **KinEMbed**, a cross-modal contrastive learning framework for hand kinematics regression that jointly trains dual encoders – one for windowed EMG features and one for kinematic (joint angle) targets. The resulting embeddings inherit the geometric structure of the kinematic space without requiring kinematic signals at inference time. Evaluating on the NinaPro DB8 dataset that includes both able-bodied subjects and subjects with limb difference ($N$=11), KinEMbed outperforms PCA, PLS, autoencoder and contrastive (CEBRA) baselines on held-out sessions, with largest gains on the most challenging thumb degrees of articulation. We position this work as a first step toward contrastive representation learning for regression of hand kinematics from structured wearable biosignals.

## 1. Introduction

Decoding continuous hand joint kinematics from surface electromyography (EMG) is a longstanding challenge in rehabilitation engineering with direct implications for prosthetic control and wearable human–machine interfaces. Despite decades of progress in myoelectric prosthetics, upper-limb prosthesis abandonment rates have remained at approximately 44% over the past two decades, with inadequate device control cited as a primary driver of rejection (Biddiss & Chau, 2007; Salminger et al., 2022). A key bottleneck is the gap between the discrete, sequential control strategies deployed in commercial devices (Jiang et al., 2012) and the continuous, simultaneous, proportional control required for natural hand use.

The dominant paradigm remains discrete gesture classification (Farina et al., 2014; Phinyomark & Scheme, 2018; Raghu et al., 2025; Yang et al., 2025), in which EMG windows are assigned to one of a finite set of predefined postures – an approach that cannot represent the continuous, graded nature of hand movement, limiting user utility.

Contrastive representation learning is a powerful framework for learning transferable, structured embeddings within and across modalities (Chen et al., 2020; Radford et al., 2021; van den Oord et al., 2018). CEBRA (Schneider et al., 2023) uses auxiliary variables to guide contrastive sampling for single-encoder neural embeddings. CPEP (Cui et al., 2025) and EMBridge (Cui et al., 2026) use cross-modal contrastive pre-training discrete for gesture classification from EMG. No prior work has applied cross-modal contrastive alignment to *continuous* kinematic regression from EMG.

We present **KinEMbed**, the first cross-modal contrastive framework for continuous EMG-to-kinematics regression. KinEMbed jointly trains dual encoders – one for windowed EMG features and one for kinematic (joint angle) targets – to align both modalities on a shared unit hypersphere via the NT-Xent loss (Sohn, 2016). The resulting EMG embedding inherits the geometric structure of the kinematic space as an inductive bias for downstream regression, *without* requiring kinematic signals at inference time. We evaluate on the NinaPro DB8 dataset (Krasoulis et al., 2019b) ($N$=11 subjects, including 2 subjects with limb difference (LD) under a strict cross-session protocol in which the test session is independently recorded and never used during model selection. Our main contributions are:

**1.** We introduce KinEMbed, the first cross-modal contrastive learning framework for continuous hand kinematics regression from EMG, embedding EMG and joint angle vectors as co-equal modalities in a shared embedding space.

**2.** We benchmark KinEMbed against five representation-learning baselines (PCA, PLS, AE-Recon, AE-Super, CEBRA) spanning the supervised/unsupervised and linear/nonlinear axes, plus ARIMA as a temporal baseline.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

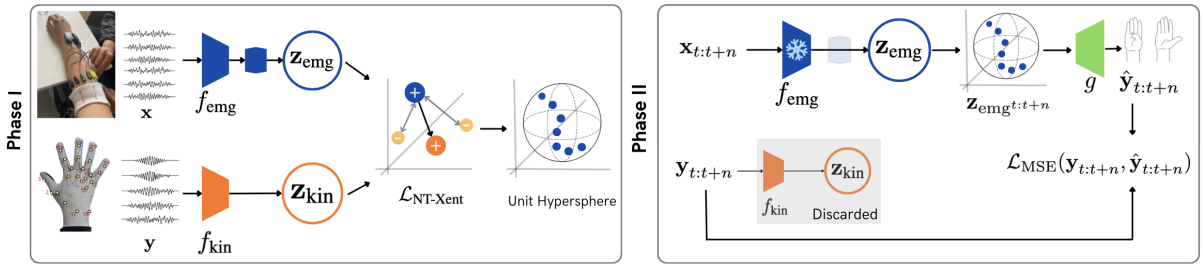



*Figure 1.* KinEMbed architecture and training phases. In Phase I, two MLP encoders ($f_{\text{emg}}$, $f_{\text{kin}}$) project EMG features $\mathbf{x}$ and kinematic angles $\mathbf{y}$ onto a shared 16-dim unit hypersphere trained via a contrastive objective with L2 normalisation. In Phase II, the frozen EMG embedding is then passed to a TCN decoder ($g$) trained via an MSE objective for continuous DoA regression. Full details in Section 3.

## 2. Related Work

**Continuous joint angle regression from EMG.** While discrete gesture classification has dominated the EMG decoding literature, a substantial body of work has pursued continuous joint angle regression for proportional prosthetic control. Classical approaches apply linear regression, kernel ridge regression, Gaussian process regression, or support vector regression to hand-crafted time-domain and frequency-domain feature windows (Hahne et al., 2014; Muceli & Farina, 2011; Xiloyannis et al., 2017). Krasoulis et al. (2019a) evaluated regression-based decoding of finger kinematics on the NinaPro DB8 dataset – the same dataset used in this work – and demonstrated accurate per-subject prediction of five degrees of actuation using EMG features and a regularized Wiener filter. More recent deep learning approaches apply CNNs, LSTMs, and temporal convolutional networks (TCNs) directly to raw or lightly processed EMG signals (Zanghieri et al., 2019).

**Contrastive and self-supervised learning for biosignals.** SimCLR (Chen et al., 2020) and InfoNCE (van den Oord et al., 2018) established that contrastive objectives learn representations that transfer well across tasks, with the SimCLR projection head - discarded at inference - decoupling the contrastive geometry from the downstream task. VICReg (Bardes et al., 2022) extends this to non-contrastive variance-invariance-covariance objectives. CEBRA (Schneider et al., 2023) uses contrastive learning to align embeddings of neural data with an auxiliary modality (e.g., behavioural data). The latter is used as a sampling oracle to define positive *time offsets* for the contrastive objective.

Closely related work, CPEP (Cui et al., 2025) and EMBridge (Cui et al., 2026), use cross-modal contrastive pre-training for *discrete gesture classification*, aligning EMG with pose representations, including from large-scale datasets such as emg2pose (Salter et al., 2024) and kinematics data. In contrast, KinEMbed targets *continuous joint angle regression*, operating directly on fine-grained kinematics in low-data per-subject clinical settings (including limb difference), rather than predicting gesture labels.

## 3. Method

**Problem Formulation.** Given a window of $C{=}16$ surface EMG channels, we extract a feature vector $\mathbf{x} \in \mathbb{R}^{256}$ (detailed below) and wish to predict five continuous degrees of articulation (DoA): thumb rotation, thumb flexion, index flexion, middle flexion, and ring/little flexion, collectively $\mathbf{y} \in \mathbb{R}^5$. Degrees of articulation (DoA) are extracted from a hand motion capture glove as described in Appendix A.1.

**EMG Feature Extraction** EMG signals are windowed at 128 ms (256 samples at 2000 Hz) with a 52 ms stride. Per channel, we compute time-domain features (RMS, Waveform Length, Log Variance, Zero Crossings, Slope Sign Changes, MAV, Willison Amplitude, Hjorth Mobility and Complexity), spectral features (Mean Power Frequency, Median Frequency), and STFT magnitude across five bands. Concatenating over 16 channels yields a 256-dimensional feature vector. This feature group was selected over three alternative variants by measuring baseline regression performance under cross-validated session splits. All features are standardised using per-training-set $z$-score normalisation.

### 3.1. KinEMbed Architecture

**Dual encoders.** We train two separate MLPs jointly (Figure 1). The *EMG encoder* $f_{\text{emg}}$ has hidden dimensions $[256, 128]$ with batch normalisation, ReLU activations, and dropout ($p{=}0.1$). The *kinematic encoder* $f_{\text{kin}}$ has hidden dimensions $[64, 32]$ with the same activations. Both encoders project to an $E$-dimensional space ($E{=}16$) followed by L2 normalisation onto a unit hypersphere:

$$\mathbf{z}_{\text{emg}} = \ell_2(f_{\text{emg}}(\mathbf{x})), \quad \mathbf{z}_{\text{emg}} \in \mathbb{R}^{16}, \tag{1}$$

$$\mathbf{z}_{\text{kin}} = \ell_2(f_{\text{kin}}(\mathbf{y})), \quad \mathbf{z}_{\text{kin}} \in \mathbb{R}^{16}. \tag{2}$$

**Projection head.** Following SimCLR (Chen et al., 2020), a two-layer nonlinear projection head is appended to $f_{\text{emg}}$ during contrastive training. At inference, only the encoder trunk is used; the projection head is discarded. This decouples the geometry optimised by the contrastive loss from the geometry seen by the downstream decoder.

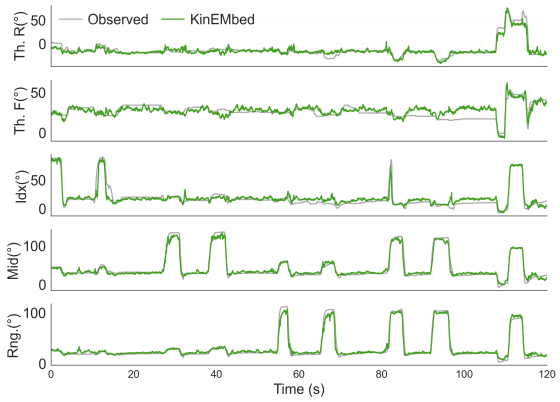



*Figure 2.* Example 120-second continuous prediction trace for subject 10 (able-bodied). All five DoA are decoded simultaneously from EMG using the frozen $f_{emg}$ encoder.

*Table 1.* Overall mean $R^2$ (across 5 DoA) on held-out session `d3`. Values: mean ± std across subjects ($R^2$). Best per column in **bold**.

| Method | All ($n$=11) | | AB ($n$=9) | | limb difference ($n$=2) | |
|---|---|---|---|---|---|---|
| ARIMA | -0.072 | 0.056 | -0.075 | 0.056 | -0.062 | 0.074 |
| PCA | 0.717 | 0.113 | 0.748 | 0.077 | 0.560 | 0.170 |
| PLS | 0.722 | 0.118 | 0.753 | 0.079 | 0.566 | 0.201 |
| AE-Recon | 0.704 | 0.111 | 0.733 | 0.081 | 0.558 | 0.159 |
| AE-Super | 0.703 | 0.132 | 0.743 | 0.076 | 0.523 | 0.221 |
| CEBRA | 0.713 | 0.115 | 0.745 | 0.078 | **0.570** | 0.181 |
| **KinEMbed** | **0.732** | 0.129 | **0.769** | 0.090 | 0.568 | 0.191 |

**Contrastive objective.** Given a batch of $N$ synchronised (EMG, kinematic) pairs, we use symmetric NT-Xent loss:

$$\mathcal{L} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{e^{\mathbf{z}_i^{\text{emg}}\cdot\mathbf{z}_i^{\text{kin}}/\tau}}{\sum_j e^{\mathbf{z}_i^{\text{emg}}\cdot\mathbf{z}_j^{\text{kin}}/\tau}} + \log\frac{e^{\mathbf{z}_i^{\text{kin}}\cdot\mathbf{z}_i^{\text{emg}}/\tau}}{\sum_j e^{\mathbf{z}_i^{\text{kin}}\cdot\mathbf{z}_j^{\text{emg}}/\tau}}\right], \tag{3}$$

where $\tau$=0.2 is the temperature and positive pairs are temporally aligned (EMG window $i$, DoA measurement $i$). We also evaluated a *Soft-NT-Xent* variant that replaces the one-hot contrastive target with a Gaussian-kernel soft target over DoA distances (addressing the false-negative problem for regression), and VICReg (Bardes et al., 2022). NT-Xent was selected via grid search (Section 4).

**Downstream TCN decoder.** After contrastive pre-training, the EMG encoder is frozen and a Temporal Convolutional Network (TCN) (Bai et al., 2018) decoder is trained to map sequences of EMG embeddings to 5-dimensional DoA predictions. The TCN ($g$) uses dilated causal convolutions (64 channels, kernel size 3, dilations $[1, 2, 4]$) with weight normalisation and residual connections. The decoder is trained for 100 epochs with Adam (Kingma & Ba, 2015) ($\eta$=$10^{-3}$, MSE loss). All baselines share this decoder, ensuring that embedding quality is the only variable under comparison. Full training details in Appendix B, code to be released soon.

*Table 2.* Per-DoA $R^2$ on held-out session d3, able-bodied subjects ($n$=9). Best per column in **bold**. DoA names abbreviated: Th.R = thumb rotation, Th.F = thumb flexion, Idx = index, Mid = middle, Rng = ring/little.

| Method | Th.R | Th.F | Idx | Mid | Rng | Mean |
|---|---|---|---|---|---|---|
| ARIMA | −0.045 | −0.025 | −0.009 | −0.008 | −0.286 | −0.075 |
| PCA | 0.660 | 0.611 | 0.790 | 0.828 | 0.853 | 0.748 |
| PLS | 0.680 | 0.596 | 0.787 | 0.834 | **0.867** | 0.753 |
| AE-Recon | 0.655 | 0.591 | 0.772 | 0.815 | 0.834 | 0.733 |
| AE-Super | 0.647 | 0.570 | 0.805 | 0.834 | 0.857 | 0.743 |
| CEBRA | 0.672 | 0.621 | 0.787 | 0.826 | 0.818 | 0.745 |
| **KinEMbed** | **0.694** | **0.644** | **0.812** | **0.845** | 0.850 | **0.769** |

## 4. Experiments

**Dataset and Evaluation Protocol.** We evaluate on **NinaPro DB8** (Krasoulis et al., 2019b), which comprises synchronised EMG and instrumented glove recordings from 12 subjects performing continuous hand grasps and a wide range of finger movements. Subject 4 was excluded due to data fidelity issues yielding $N$=11 participants: 9 able-bodied (AB) and 2 LD subjects. Each subject completed three sessions (`d1`, `d2`, `d3`). Sessions `d1` and `d2` are used for training; `d3` is the strictly held-out test set, never touched during model selection. We use 2-fold session CV during hyperparameter tuning. We report the mean coefficient of determination ($R^2$) across the 5 DoA channels on the held-out session `d3`. We report mean ± std across three seeds.

**Baseline selection.** We initially evaluated seven candidate DR methods in preliminary experiments: PCA, UMAP, PLS, a reconstruction autoencoder (AE-Recon), a supervised autoencoder (AE-Super), a contrastive autoencoder, a conditional VAE (CVAE) and CEBRA. The five baselines retained for the final comparison – PCA, PLS, AE-Recon, AE-Super and CEBRA – were selected by ranking all candidates on mean $R^2$ under 2-fold session CV and keeping the top performers. The set spans the key methodological axes: unsupervised linear (PCA), supervised linear (PLS), unsupervised nonlinear (AE-Recon), and supervised nonlinear (AE-Super). *As far as we are aware, this is also the first instance of CEBRA applied to EMG decoding.* We also included a non–representation-learning baseline, Autoregressive Integrated Moving Average (ARIMA), to benchmark against a classical statistical time-series model.

**Baseline configurations.** Because the aim of this work is to *introduce* KinEMbed rather than to produce an exhaustive benchmark, baseline architectures follow standard designs with minimal tuning. AE-Recon uses a symmetric MLP encoder-decoder ($144 \to 128 \to 64 \to d$) trained with MSE reconstruction loss (600 epochs, Adam, $\eta$=$10^{-4}$). AE-Super shares the same encoder but adds an auxiliary DoA regression head from the bottleneck; the reconstruction/regression loss weighting $\alpha$ is selected from $\{0.1, 0.5, 1.0\}$ via 2-fold session CV. PCA and PLS are fit

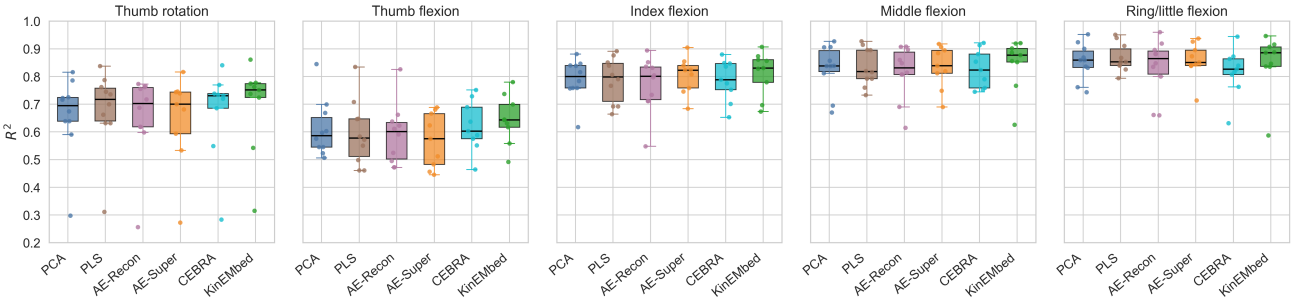



*Figure 3.* Per-DoA $R^2$ for able-bodied subjects. KinEMbed shows the most pronounced gains on thumb rotation and thumb flexion, the most challenging and variable degrees of articulation. ARIMA is omitted due to poor performance (negative $R^2$).

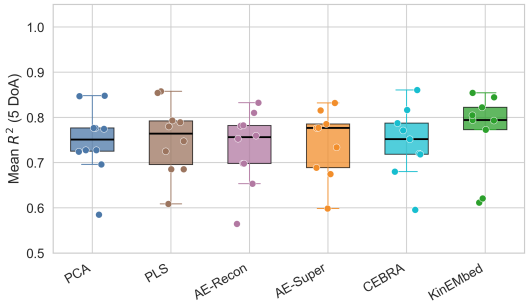



*Figure 4.* Distribution of mean $R^2$ across able-bodied subjects ($n$=9) on held-out session d3. KinEMbed achieves the highest median and mean $R^2$.

on the training split using standard `sklearn` implementations. CEBRA was tuned as described in Appendix B.4. To ensure a fair comparison, each baseline's embedding dimension is independently optimised via grid search over $d \in \{2, 3, 4, 8, 16, 32\}$ using 2-fold session CV. The canonical dimension is the mode of per-subject best values (PCA: 32, PLS: 32, AE-Recon: 32, AE-Super: 16, CEBRA: 8).

### 4.1. Results

Table 1 reports overall mean $R^2$ across the 5 DoA for all subjects. KinEMbed achieves the highest mean $R^2$ overall, with a mean of **0.732** (all subjects), **0.769** (AB), and **0.568** (LD). Figure 4 illustrates the $R^2$ distribution across AB subjects. ARIMA, which exploits only the temporal autocorrelation of the DoA signal without any EMG input, achieves *negative* $R^2$ across all groups, confirming that hand joint states cannot be inferred from their own temporal dynamics and that EMG carries essential non-redundant kinematic information. On the limited LD cohort, CEBRA achieves marginally higher mean $R^2$ (0.570 vs. 0.568), but differences are small and the cohort is limited to two subjects with high inter-subject variance.

Table 2 breaks results down by DoA for AB subjects. KinEMbed achieves the best mean $R^2$ and leads on four of five DoA. The largest gains are on *thumb flexion* (+4.8 pp over next best) and *thumb rotation* (+1.4 pp), which are the most variable and clinically relevant degrees. The ring/little finger DoA is already well-predicted by all methods and differences are negligible, as shown in Figure 3.

## 5. Discussion and Conclusion

KinEMbed introduces cross-modal contrastive learning as a new paradigm for continuous hand joint regression from EMG. By treating EMG and kinematics as co-equal modalities and aligning their latent representations on a shared unit sphere, the EMG encoder inherits the geometric structure of the kinematic space – structure that reconstructive and linear baselines do not exploit. Consistent empirical improvement across subjects and DoA suggests that this geometric alignment provides a meaningful inductive bias for the regression task. The gains are most pronounced for thumb movements, which are mechanically decoupled from the other fingers and produce more variable EMG patterns, precisely where discriminative contrastive pressure provides the most benefit over reconstructive objectives.

**Limitations.** This work introduces our method and presents initial evaluation of our method against relevant baselines. Further evaluation would be desirable, including additional tuning for KinEMbed and baselines. Improvements are modest in absolute terms and the LD sample is small. Additional datasets could be used for evaluation, to test cross-user generalization, and to draw firm conclusions on the relevance of these methods for prosthetic control.

**Future directions.** Several natural extensions follow from this work. *Temporal contrastive objectives* (e.g., CPC-style) could exploit the sequential structure of EMG more explicitly. *Cross-subject and cross-session adaptation* via contrastive pre-training on pooled data, followed by subject-specific fine-tuning, is a direct path to clinical deployment. The dual-encoder structure is also well-suited to *foundation model pre-training*: a kinematic encoder pre-trained on large motion-capture datasets could provide a rich target embedding space without requiring synchronised EMG.

We hope KinEMbed serves as a useful baseline for EMG regression tasks and conceptual stepping stone for representation learning on structured wearable biosignals.

## References


Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, 2019.

Atzori, M., Gijsberts, A., Heynen, S., Hager, A.-G. M., Deriaz, O., van der Smagt, P., Castellini, C., Caputo, B., and Müller, H. Building the ninapro database: A resource for the biorobotics community. In *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pp. 1258–1265, 2012. doi: 10.1109/BioRob.2012.6290287.

Bai, S., Kolter, J. Z., and Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Bardes, A., Ponce, J., and LeCun, Y. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Biddiss, E. A. and Chau, T. T. Upper limb prosthesis use and abandonment: a survey of the last 25 years. *Prosthetics and orthotics international*, 31(3):236–257, 2007.

Cabibihan, J.-J., Alkhatib, F., Mudassir, M., Lambert, L. A., Al-Kwifi, O. S., Diab, K., and Mahdi, E. Suitability of the openly accessible 3d printed prosthetic hands for war-wounded children. *Frontiers in Robotics and AI*, 7: 594196, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML 2020, pp. 1597–1607, 2020.

Cui, W., Sandino, C., Pouransari, H., Liu, R., Minxha, J., Zippi, E., Verma, A., Sedlackova, A., Azemi, E., and Mahasseni, B. Cpep: Contrastive pose-emg pre-training enhances gesture generalization on emg signals. *arXiv preprint arXiv:2509.04699*, 2025.

Cui, W., Sandino, C. M., Pouransari, H., Liu, R., Minxha, J., Zippi, E. L., Azemi, E., and Mahasseni, B. Embridge: Enhancing gesture generalization from emg signals through cross-modal representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Farina, D., Jiang, N., Rehbaum, H., Holobar, A., Graimann, B., Dietl, H., and Aszmann, O. C. The extraction of neural information from the surface EMG for the control of upper-limb prostheses. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(4):797–809, 2014.

Hahne, J. M., Biessmann, F., Jiang, N., Rehbaum, H., Farina, D., Meinecke, F. C., Müller, K.-R., and Parra, L. C. Linear and nonlinear regression techniques for simultaneous and proportional myoelectric control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(2): 269–279, 2014.

Hannan, E. J. and Quinn, B. G. The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–195, 1979.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Hyndman, R. J. and Khandakar, Y. Automatic time series forecasting: the forecast package for r. *Journal of statistical software*, 27:1–22, 2008.

Jiang, N., Dosen, S., Muller, K.-R., and Farina, D. Myoelectric control of artificial limbs—is there a need to change focus?[in the spotlight]. *IEEE Signal Processing Magazine*, 29(5):152–150, 2012.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Krasoulis, A., Kyranou, I., Erwin Veltink, M., Nazarpour, K., and Vijayakumar, S. Improved prosthetic hand control with concurrent use of myoelectric and inertial measurements. In *Journal of NeuroEngineering and Rehabilitation*, volume 14, pp. 71, 2017.

Krasoulis, A., Vijayakumar, S., and Nazarpour, K. Effect of user practice on prosthetic finger control with an intuitive myoelectric decoder. *Frontiers in neuroscience*, 13:891, 2019a.

Krasoulis, A., Vijayakumar, S., and Nazarpour, K. Effect of user practice on prosthetic finger control with an intuitive myoelectric decoder. *Frontiers in Neuroscience*, 13:891, 2019b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Muceli, S. and Farina, D. Simultaneous and proportional estimation of hand kinematics from emg during mirrored movements at multiple degrees-of-freedom. *IEEE transactions on neural systems and rehabilitation engineering*, 20(3):371–378, 2011.

Phinyomark, A. and Scheme, E. A feature extraction issue for myoelectric control based on wearable EMG sensors. *Sensors*, 18(5):1615, 2018.

Prensilia. IH2 Azzurra Robotic Hand. https://www.prensilia.com/ih2-azzurra-hand/, 2023. Accessed: 2023-03-23.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Raghu, S. T. P., MacIsaac, D. T., and Scheme, E. J. Self-supervised learning via vicreg enables training of emg pattern recognition using continuous data with unclear labels. *Computers in Biology and Medicine*, 185:109479, 2025.

Salminger, S., Stino, H., Pichler, L. H., Gstoettner, C., Sturma, A., Mayer, J. A., Szivak, M., and Aszmann, O. C. Current rates of prosthetic usage in upper-limb amputees–have innovations had an impact on device acceptance? *Disability and rehabilitation*, 44(14):3708–3713, 2022.

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

Schneider, S., Lee, J. H., Bhatt, S., Bhatt, D., and Ecker, A. S. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617:360–368, 2023.

Smith, T. G. et al. pmdarima: Arima estimators for Python, 2017–. URL http://www.alkaline-ml.com/pmdarima. [Online; accessed April 5th 2026].

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Steffen, J., Maycock, J., and Ritter, H. Robust dataglove mapping for recording human hand postures. In *International Conference on Intelligent Robotics and Applications*, pp. 34–45. Springer, 2011.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Wang, Y. and Neff, M. Data-driven glove calibration for hand motion capture. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 15–24, 2013.

Xiloyannis, M., Gavriel, C., Thomik, A. A., and Faisal, A. A. Gaussian process autoregression for simultaneous proportional multi-modal prosthetic control with natural hand kinematics. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1785–1801, 2017.

Yang, J., Cha, D., Lee, D.-G., and Ahn, S. Stcnet: Spatio-temporal cross network with subject-aware contrastive learning for hand gesture recognition in surface emg. *Computers in Biology and Medicine*, 185:109525, 2025.

Zanghieri, M., Benatti, S., Burrello, A., Kartsch, V., Conti, F., and Benini, L. Robust real-time embedded emg recognition framework using temporal convolutional networks on a multicore iot processor. *IEEE transactions on biomedical circuits and systems*, 14(2):244–256, 2019.

# A. Dataset

We evaluate KinEMbed and baselines on the Ninapro DB8 dataset. The EMG data in DB8 was recorded using 16 active double-differential wireless sensors from a Delsys Trigno IM Wireless EMG system (Atzori et al., 2012) as shown in Figure 5. Muscle activity was recorded from the participants' right forearm (i.e., the remnant limb for subjects with limb difference). Motion capture data was recorded with a Cyberglove II - a motion capture glove that contains 18 joint-angle measurement sensors, distributed as shown in Figure 6b.

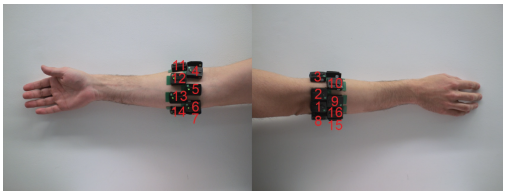
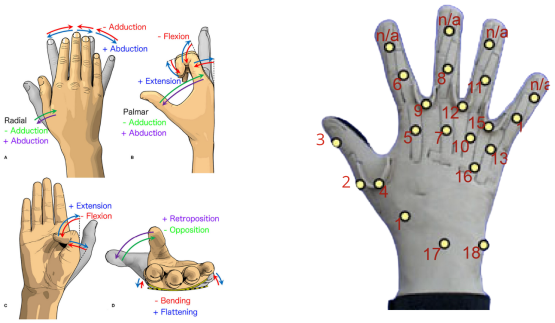


*Figure 5.* The positioning of the 16 electrodes in the NinaPro dataset. Image reproduced from (Krasoulis et al., 2019b).




(a) (b)

*Figure 6.* A visualisation of the movements of the hand and wrist (a). Image reproduced from (Cabibihan et al., 2021). The positioning of the joint state sensors (b). Image reproduced from (Krasoulis et al., 2019b).

## A.1. Cyberglove Data

Although the Cyberglove II has 18 joint measurement sensors, the objective of this research is the control of an external device such as a dexterous robotic hand, or a simulated hand. Hence, a mapping strategy is employed to translate the 18 channels of joint angle data to DoA of a robotic hand. The mapping of Cyberglove joint states to hand position reconstruction has been explored for visualisation (Wang & Neff, 2013) and robot manipulation (Steffen et al., 2011) and has further been refined for the actuation of a robotic hand for prostheses using a linear mapping from the 18 joint states to the 5 DoA of the IH2 Azzurra robotic hand (Krasoulis et al., 2019a; Prensilia, 2023). In this transformation, the 18 calibrated measurements of the dataglove ($\mathbf{x} \in \mathbb{R}^{18}$) are mapped to the DoA of the robotic hand ($\mathbf{y} \in \mathbb{R}^{5}$) via the transformation matrix 6. The DoA correspond to the movement of the five fingers: $y_1$ , thumb rotation; $y_2$, thumb flexion; $y_3$, index flexion; $y_4$, middle flexion; $y_5$, ring/little finger flexion. The ring and little finger are controlled together due to a mechanical coupling in the Azzurra robotic hand (Krasoulis et al., 2019a). Using this model allows for a more explainable decoding performance and renders the model more directly transferable for the control of an external device.

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{4}$$

$$A^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \tag{5}$$

$$A^T = \begin{bmatrix} 0.639 & 0 & 0 & 0 & 0 \\ 0.383 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -0.639 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1667 \\ 0 & 0 & 0 & 0 & 0.3333 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1667 \\ 0 & 0 & 0 & 0 & 0.3333 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -0.19 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{6}$$

# B. Training and Hyperparameter Optimisation Details

## B.1. Decoder

An MLP, GRU, and TCN were evalauted as potential decoders. TCN was chosen as the best performing on average across all methods, based on the $R^2$ across a subsection of tested subjects via 2-fold CV.

## B.2. KinEMbed

Contrastive pre-training uses AdamW (Loshchilov & Hutter, 2019) with weight decay $10^{-4}$, learning rate $3\times10^{-3}$, cosine annealing ($T_{\text{max}}$=200), batch size 256, and gradient clipping (norm 1.0). Training runs for up to 200 epochs with early stopping (patience 20) on a fixed 10% validation

split.

Hyperparameters were selected via a principled 4-stage sequential grid search, designed to explore each axis of variation independently while fixing decisions from prior stages. This staged strategy reduces the effective search space from a prohibitive full factorial (tens of thousands of combinations) to targeted combinations evaluated on 5 representative subjects (1, 3, 6, 8, 11) via 2-fold session CV, using mean CV $R^2$ as the selection criterion.

**Stage 1**: loss $\in$ {NT-Xent, Soft-NT-Xent, VICReg}, temperature $\tau \in \{0.05, 0.07, 0.1, 0.2, 0.5\}$, embedding dim $d \in \{4, 8, 16, 32, 64\}$. **Winner**: NT-Xent, $\tau$=0.2, $d$=16.

**Stage 2**: architecture variant $\in$ {standard, projection head, residual, asymmetric, momentum EMA}. **Winner**: projection head.

**Stage 3**: EMG hidden dims $\in \{[128, 64], [256, 128], [512, 256, 128], [512, 256, 128, 64]\}$, kinematic hidden dims $\in \{[32], [64, 32], [128, 64]\}$. **Winner**: EMG $[256, 128]$, kin $[64, 32]$.

**Stage 4**: learning rate $\in \{10^{-4}, 5{\times}10^{-4}, 10^{-3}, 3{\times}10^{-3}\}$, batch size $\in \{128, 256, 512\}$, dropout $\in \{0.0, 0.1, 0.2\}$. **Winner**: $\eta$=$3{\times}10^{-3}$, batch 256, dropout 0.1.

### B.3. ARIMA

As a time-series baseline, we evaluated an autoregressive integrated moving average (ARIMA) model applied independently to each degree of action (DoA). Unlike the EMG-feature-based methods, ARIMA operates solely on the DoA time series, exploiting temporal autocorrelation in the kinematic signal rather than instantaneous EMG observations.

Model orders $(p, d, q)$ were selected per-DoA and per-subject using the `auto_arima` procedure from the `pmdarima` library (Smith et al., 2017–), with the Bayesian Information Criterion (BIC) as the selection objective. BIC was preferred over AIC as its stronger complexity penalty yields parsimonious models that generalise well from limited data. A stepwise search (Hyndman & Khandakar, 2008) was used in place of an exhaustive grid to reduce computation. Seasonal components were not modelled, as the DoA signal has no fixed periodic structure.

Order selection was performed on the final 20% of the second training session (dataset 2), a single contiguous recording that is temporally adjacent to the held-out test session (dataset 3). This segment provides a representative characterisation of the signal's local autocorrelation structure; BIC-based order selection is known to stabilise with a few hundred observations (Hannan & Quinn, 1979).

Evaluation followed the same protocol as all other methods: models were applied to dataset 3.

### B.4. CEBRA

CEBRA hyperparameters were tuned via Optuna Tree-structured Parzen Estimator (TPE) search (Akiba et al., 2019), run independently for each subject using 2-fold session cross-validation with mean $R^2$ as the objective. *Up to 50 trials per subject* were evaluated over a categorical search space spanning training iterations $\in \{1000, 3000, 5000, 10000\}$, batch size $\in \{256, 512, 1024\}$, learning rate $\in \{10^{-4}, 5{\times}10^{-4}, 10^{-3}\}$, temperature $\in \{0.5, 1.0, 2.0\}$, and time offset $\in \{5, 10, 20\}$ windows. Canonical hyperparameters were set to the mode across subjects, yielding: batch size 512, learning rate $10^{-3}$, 1000 iterations, temperature 2.0, time offset 10, and embedding dimension 8 (selected in a separate dimensionality study); these values were fixed for all final evaluations.

## C. Relationship to CEBRA

CEBRA (Schneider et al., 2023) is a contrastive representation learning framework designed to produce structured latent embeddings of neural or physiological recordings by leveraging a continuous auxiliary behavioural variable. It has shown strong performance for neural data, and is thus a relevant baseline for our work. Although KinEMbed shares the broad goal of learning a auxiliary-modality-structured embedding of a primary physiological signal, the two methods differ in architecture, supervisory mechanism, and the geometric properties of the resulting embedding space.

**Architecture.** CEBRA employs a *single-encoder* design: only the primary signal is passed through a parametric network, a temporal convolutional network (TCN) with learnable temperature (Schneider et al., 2023). Kinematics serve as a *sampling oracle* that defines which pairs of EMG windows should be close in the embedding space. KinEMbed instead uses a *dual-encoder* design: an MLP encoder processes EMG windows and a separate, shallower MLP encoder processes the corresponding DoA vectors. Both encoders are trained jointly, and their outputs are aligned in a shared embedding space.

**Positive pair construction.** CEBRA constructs positive pairs via a global kinematic nearest-neighbour lookup: for each anchor EMG window, it searches the *entire* dataset for the recording whose simultaneous kinematics most closely matches a sampled kinematic target state (Schneider et al., 2023). KinEMbed uses a simpler synchronous scheme: the EMG window at time $t$ is paired directly with the simultaneously recorded DoA vector $y_t$, with all other pairs in the mini-batch serving as negatives.

## D. Relationship to CPEP

A concurrent line of work, CPEP (Cui et al., 2025), applies cross-modal contrastive learning between surface EMG and hand pose in a broadly similar spirit to KinEMbed. Both methods train a dual-encoder architecture using synchronised EMG–kinematic pairs and discard the kinematic encoder at inference. Despite this surface similarity, the two methods differ fundamentally in task formulation, architectural design, loss construction, and intended clinical context. We detail these differences below.

**Task formulation.** The most consequential difference is the prediction target. KinEMbed is designed for *continuous regression*: the decoder produces real-valued estimates of five joint angles simultaneously, evaluated by coefficient of determination ($R^2$) on a held-out recording session. CPEP targets *discrete gesture classification*. This distinction is not superficial. *Continuous kinematic regression enables direct, proportional control*, which is essential for clinical settings as well as fluid interaction in virtual and augmented reality; discrete gesture recognition is a coarse proxy. Every subsequent design choice in each method follows from this task difference.

**The false-negative problem in regression.** Because KinEMbed targets a continuous output space, the standard NT-Xent objective introduces a structural problem that does not arise in classification: two windows recorded at different times but with nearly identical joint angles are treated as hard negatives and actively pushed apart in the embedding space, despite representing the same kinematic state. Although not used for final evaluation, we propose to address this with a *soft* NT-Xent variant that replaces one-hot positive targets with a Gaussian kernel over pairwise DoA distances, weighting each negative by its kinematic proximity to the anchor and using the median pairwise distance as a parameter-free bandwidth. Future work should explore the value of this loss further. CPEP does not address this problem, nor does it need to: in a discrete classification setting, two windows from different gesture classes are genuine negatives by construction.

**Architecture and feature representation.** KinEMbed uses lightweight MLP encoders operating on 256-dimensional EMG feature vectors (time-domain and spectral statistics per channel), producing 16-dimensional embeddings. A SimCLR-style projection head is appended during contrastive training and discarded at inference, decoupling the contrastive geometry from the downstream regression space. The downstream decoder is a Temporal Convolutional Network (TCN) trained on frozen embeddings to regress the five joint angles. CPEP uses Transformer encoders (four layers, $d$=256) operating on raw 2-second EMG waveforms pre-trained via Masked Autoencoder (MAE) (He et al., 2022), producing 256-dimensional embeddings. The downstream task being classification, no regression decoder is trained.

**Dataset scale and subject population.** KinEMbed is evaluated on NinaPro DB8 (Krasoulis et al., 2017), comprising 11 subjects (9 able-bodied, 2 with limb difference) across three recording sessions each. The inclusion of limb-different participants introduces substantial variability in neuromuscular structure and EMG signal characteristics, providing a challenging testbed for representation robustness under distribution shift. CPEP is evaluated on the emg2pose dataset (Salter et al., 2024), comprising 193 participants and approximately 370 hours of recording, and focuses entirely on able-bodied users. *The scale difference reflects different scientific goals*: CPEP demonstrates large-scale pre-training and zero-shot transfer across users and unseen gestures; KinEMbed demonstrates strong potential for contrastive kinematic alignment in the low-data, per-subject regime characteristic of EMG-based control.

**Evaluation protocol.** KinEMbed uses a strict cross-session evaluation protocol: models are trained on sessions 1 and 2 and evaluated on the independently recorded session 3, with no overlap. This directly tests the robustness of the embedding to electrode shift and inter-session variability – the primary failure mode for clinical EMG decoders. CPEP evaluates cross-user generalisation and zero-shot transfer to held-out gesture classes, which is the relevant generalisation axis for a large-scale consumer application.

**Summary.** KinEMbed and CPEP share the cross-modal contrastive pre-training paradigm but diverge in every dimension that matters for their respective applications. KinEMbed is a *continuous regression framework*, operating in the low-data regime with limb difference subjects and evaluated by joint angle prediction accuracy. CPEP is a discrete *classification* method for large-scale consumer gesture recognition, operating with orders of magnitude more data and evaluated by gesture identification accuracy. The surface resemblance in training objective reflects the generality of the cross-modal contrastive framework; the differences in task, loss, architecture, data, and evaluation reflect the distinct requirements of the two applications.