ALIGNING HUMAN AND MOUSE EEG REPRESENTATIONS OF SLEEP STAGES WITH A QUADRUPLET SPLIT LATENT PERMUTATION AUTO-ENCODER

Javier G. Ciudad[1,*] **Anders V. Nørskov**[2] **Alexander N. Zahid**[2,3]

Birgitte R. Kornum[1,*] **Morten Mørup**[2,†]

[1]Department of Neuroscience, University of Copenhagen, Denmark
[2]Department of Applied Mathematics and Computer Science, Technical University of Denmark
[3]AI Center of Excellence, WS Audiology, Denmark
[*]{jgciudad,kornum}@sund.ku.dk, [†]mmor@dtu.dk

ABSTRACT

Rodents are the most widely used experimental animals in biomedical research, but human neurological disorders involve behaviors and mental states that are challenging to study in rodents. Techniques such as electroencephalography (EEG) monitor brain activity objectively, yet how brain states manifest across species remains unclear, limiting translatability. We use a Quadruplet Split Latent Permutation Autoencoder (QSLP-AE) to map human and mouse sleep EEG into a shared 2-d latent space. The QSLP-AE exchanges latent representations during reconstruction to produce aligned representations of mouse and human sleep stages, as a proxy for cross-species representations of mental states. Notably, QSLP-AE matches the performance of conventional contrastive learning using only the autoencoder reconstruction loss. Exploiting the 2-d space, we visualize and quantify the correspondence between species from the model perspective. These results demonstrate the potential of QSLP-AE to align neural representations and bridge the translational gap.

1 INTRODUCTION

Vital knowledge has been gained in neuroscience thanks to the use of rodents. However, the knowledge in rodents often translates poorly to clinical applications, with drugs for central nervous system disorders presenting very low success rates (Wilson et al., 2014; Goetghebeur & Swartz, 2016; Tian et al., 2017). Among the reasons involved, two aspects play a prominent role: (1) differences in brain anatomy and molecular mechanisms between humans and rodents despite substantial overlap; and (2) the difficulty of modeling complex human behaviors in animals (Azkona & Sanchez-Pernaute, 2022; Cavanagh et al., 2021). This points to a clear need: a readout that measures cognitive processes objectively and consistently across species (Barron et al., 2020).

Electroencephalography (EEG), which measures the electrical activity of neuronal populations, is a natural choice for identifying such common markers. Notably, sleep is one of the neural processes that can be most clearly defined from EEG, and in both humans and rodents, sleep presents distinct and clearly differentiable mental states known as sleep stages. In addition, sleep is altered in many neurological disorders (Wulff et al., 2010) and can be useful to understand the pharmacodynamics of psychoactive compounds, sometimes with similar effects in humans and rodents (Drinkenburg et al., 2016; Wilson & Danjou, 2015). However, there are still many gaps in methodology and sleep physiology that remain an obstacle for translating findings. This includes the number of sleep stages, currently defined by five stages in humans and three in mice (Rayan et al., 2022). Furthermore, some spectral features can present frequency shifts between species, such as theta band power (Jacobs, 2014), and sleep spindles (Maheshwari, 2020).

Consequently, there is a need for modeling procedures capable of mapping data from each species into a shared representational space, in which neural representations across species can be aligned

and more easily compared (Barron et al., 2020). Here, representation learning, provides a powerful framework for relating neural processes that might manifest slightly differently across species.

Whereas contrastive learning has been widely adopted to guide and align latent representations (Uelwer et al., 2025), we explore a recently proposed representation learning framework previously used for EEG signal conversion across human subjects (Nørskov et al., 2023). Specifically, we consider the quadruplet split latent permutation autoencoder (QSLP-AE) formulation introduced therein. Notably, the QSLP-AE only relies on an autoencoder loss to learn invariant representations. It achieves this by splitting the latent space into two separate latent representations, which account for neural state and individual-specific variability, and by including a permutation mechanism between latent representations. We here utilize the well-defined sleep stages as proxies for mental states, and adapt the QSLP-AE framework to map human and mouse EEG data into a two-dimensional shared space, investigating if species-aligned representations can be achieved. We systematically compare the QSLP-AE procedure with conventional contrastive learning to guide the latent representations towards species-agnostic representations of sleep stages, and we demonstrate that the QSLP-AE provides a promising alternative learning framework successfully aligning representations.

2 DATA

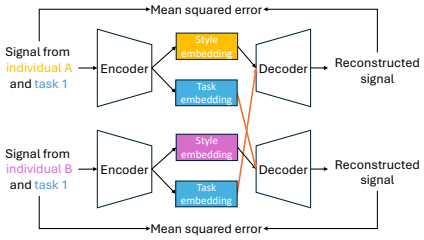
We use the Mouse Sleep Staging Validation dataset (MSSV) with EEG and EMG data of 92 mice from five laboratories (Rose et al., 2025). The dataset includes sleep scores in 4-second epochs assigned by a sleep expert to either Wake, REM (rapid-eye-movement sleep), or NREM (non rapid-eye-movement). For the human data, we use the SleepEDF (Kemp et al., 2000) and the Sleep Heart Health Study (SHHS) (Zhang et al., 2018) datasets, which include EEG and EMG recordings. SleepEDF has 77 subjects, and from SHHS (6,441 subjects) we randomly sample 120 subjects to match the human and mouse data. Both human datasets were originally scored according to the R&K standard in 30-second epochs, labeled as Wake, N1, N2, N3, N4 or REM (Rechtschaffen & Kales, 1968). We merge stages N1–N4 into a single NREM class to match mouse NREM, which is typically not subdivided in mice. A full overview of the datasets is available in Appendix A.1.

For the modeling, only one EEG and one EMG channel are used. Each EEG epoch is transformed into its power spectral density (PSD) vector $\boldsymbol{x} \in \mathbb{R}^{128}$, which represents power along 128 frequency bins. Unlike the EEG, only the total power of the EMG, rather than its frequency content, is relevant across sleep stages, so we use the epoch-wise root mean square (RMS) as the EMG input.

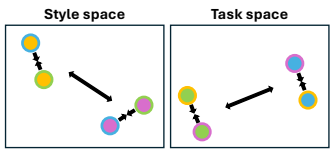
3 METHODS

Our model builds on the previously proposed Contrastive Split-Latent Permutation Autoencoder, considering the Quadruplet Split-Latent Permutation (QSLP-AE) approach described in Nørskov et al. (2023). QSLP-AE was originally designed to convert EEG signals between new, unseen subjects using EEG from event-related potentials (ERPs) (Kappenman et al., 2021). ERPs reflect the brain's electrical response to specific events, with a "task" referring to the event that elicits the response (e.g., visual or auditory cues). Thus, for a successful conversion, the model disentangles style information (i.e., subject-specific variability) from task information (i.e., stimulus-specific variability). In our goal of aligning cross-species EEG during sleep, we similarly need to disentangle individual-specific variability (style), which reflects the species, from the underlying sleep stage (task). Throughout, we use the word "individual" to denote a human or a mouse.

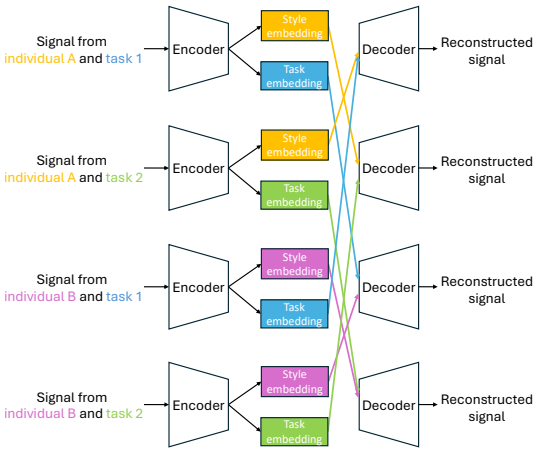
QSLP-AE has an autoencoder structure with two latent spaces that respectively encode style and task, and are ideally independent from each other. The model is trained via permutations of the latent representations between samples, which serve as the primary mechanism for the disentanglement. The process is shown graphically in Figure 1a. Two same-class instances are projected into their style and task embeddings. Same-class embeddings are permuted and decoded, and the weights are optimized to minimize the reconstruction error. Since the inputs are randomly sampled with respect to the style (i.e., individual and species), the model is trained to learn task representations invariant to species. An analogous swap enforces consistency in style embeddings across tasks (Figure 4b). However, since only one latent is permuted, the model may rely on the structural information of the unpermuted latent to minimize the error (Nørskov et al., 2023). Extending the permutation to four



(a) Same-task permutation



(b) Contrastive learning



(c) Quadruplet permutation

Figure 1: Training paradigms. **(a)** Same-task permutation. Same-class embeddings are permuted (orange lines). **(b)** Contrastive loss showing the same embeddings as in the quadruplet permutation, with the color of the circle indicating the class, and the edge of the circle indicating the secondary class (e.g., yellow circle with blue edge represents the style embedding of the input from individual A and task 1). **(c)** The quadruplet permutation removes all input-to-output paths.

input samples eliminates all direct paths from the input to the output, thereby preventing this issue and forming the quadruplet split latent permutation (QSLP) (see Figure 1c).

We modify the original architecture from Nørskov et al. (2023) to use PSDs instead of raw signals. We use a low-dimensional bottleneck with a 16-dimensional style latent space and a 2-dimensional task latent space to enable 2-d visualization of sleep stage representations and constrain cross-species variability. A baseline with as many dimensions as classes in the training set is included (199 individuals in the style space, 3 sleep stages in the task space). To compensate for the very low-dimensional bottleneck, we augment model capacity by increasing the number of layers and adding parallel encoders for the two latent representations, which completely separate the process of extracting style (i.e., individual and species specific information) and task features (i.e., sleep stage information). The full architecture is depicted in A.5.

3.1 Training paradigms

Quadruplet split latent permutation The quadruplet permutation removes all paths from input to output. The loss function is the average mean squared error (MSE) of the four samples used by the QSLP-AE. If $\boldsymbol{x}_{s,t} \in \mathbb{R}^{128}$ is the PSD of a sample with style s and task t, and $\hat{\boldsymbol{x}}_{s,t} \in \mathbb{R}^{128}$ is its reconstruction, the loss of the example in Figure 1c would be:

$$\mathcal{L}_{\text{QSLP}} = \frac{1}{4}(\|\boldsymbol{x}_{a,1} - \hat{\boldsymbol{x}}_{a,1}\|_2^2 + \|\boldsymbol{x}_{a,2} - \hat{\boldsymbol{x}}_{a,2}\|_2^2 + \|\boldsymbol{x}_{b,1} - \hat{\boldsymbol{x}}_{b,1}\|_2^2 + \|\boldsymbol{x}_{b,2} - \hat{\boldsymbol{x}}_{b,2}\|_2^2) \tag{1}$$

Contrastive learning The permutation paradigm relies on reconstruction loss in the original data space. Style–task disentanglement and cross-species alignment can alternatively be achieved using contrastive learning (CL) to promote latent alignment. Specifically, a contrastive loss in the task space explicitly pulls together representations of same-task pairs and pushes apart different-task representations (Figure 1b). The two input samples are again randomly selected with respect to species, so cross-species similarity is enforced. Likewise, a contrastive loss in the style space promotes similarity for same-individual, different-task pairs (Figure 1b). Both contrastive objectives minimize a temperature-scaled symmetric cross-entropy loss between two representations, with cosine similarity as similarity metric. This is equivalent to a NT-Xent-based CLIP loss (Chen et al., 2020; Radford et al., 2021). If we have $\mathbf{Z}'$ and $\mathbf{Z}''$ matrices in $\mathbb{R}^{D\times K}$ containing paired same-class representations for K classes in a latent space in $\mathbb{R}^D$, with τ as a trainable temperature parameter:

$$\mathcal{L}_{\text{NT-Xent}}(\boldsymbol{Z}', \boldsymbol{Z}'', k) = -\log \frac{\exp(\text{sim}(\boldsymbol{z}'_k, \boldsymbol{z}''_k)/\tau)}{\sum_{i=1}^{K} \mathbf{1}_{[\text{i}\neq\text{k}]} \exp(\text{sim}(\boldsymbol{z}'_k, \boldsymbol{z}''_i)/\tau)} \tag{2}$$

$$\mathcal{L}_{\text{CLIP}}(\boldsymbol{Z}', \boldsymbol{Z}'') = \frac{1}{K} \sum_{k=1}^{K} \Big(\mathcal{L}_{\text{NT-Xent}}(\boldsymbol{Z}', \boldsymbol{Z}'', k) + \mathcal{L}_{\text{NT-Xent}}(\boldsymbol{Z}'', \boldsymbol{Z}', k) \Big) \tag{3}$$

Standard autoencoder As a representation learning baseline without disentanglement promotion, we include a standard split latent autoencoder (equivalent to Figure 1a but without permutation).

3.2 Performance metrics

To assess the quality of the latent representations, we use classification accuracy on style and task prediction. We train four XGBoost probes on four classification tasks: style (i.e., individual) classification from style latents (S-on-S), task (i.e., sleep stage) classification from task latents (T-on-T), style classification from task latents (S-on-T), and task classification from style latents (T-on-S). K-nearest neighbors and multinomial regression probes are also shown in section A.6.

To quantify the similarity of sleep stages and species, we compute the distance between sleep stages in the 2-d task space. We use the Jensen-Shannon (JS) distance between the sleep stage distributions of different individuals (i.e., inter-individual distance). The JS distance is bounded between 0 and 1, with two identical distributions having a distance of 0. See sections A.7 and A.8 for further details.

4 Results and discussion

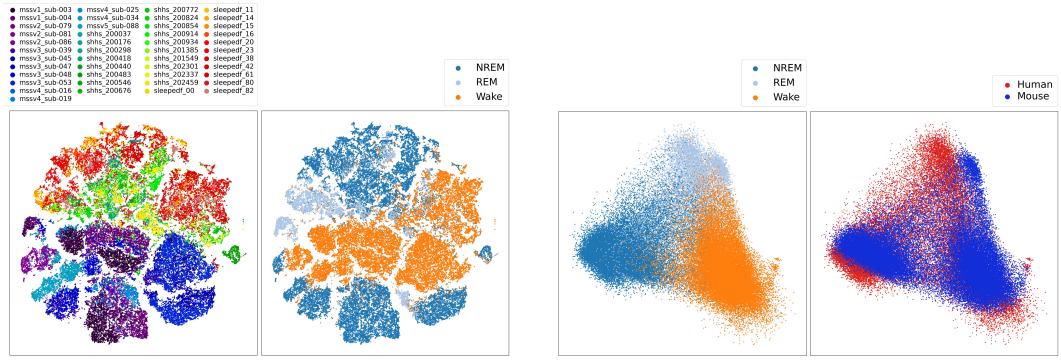
4.1 Quality of the latent representations

In Table 1, the baseline (reconstruction loss) shows low S-on-S and T-on-T accuracies, which is expected due to the absence of factors encouraging disentanglement. On the other hand, the contrastive and quadruplet losses achieve high style prediction performance (S-on-S), considering that there are 45 individuals in the test set. Sleep stage prediction is also considerably better than random (33%) in both losses, and comparable but below state-of-the-art (Rayan et al., 2022; Phan & Mikkelsen, 2022). Thus, the style and task spaces meaningfully represent their domains when using the quadruplet and contrastive losses. Remarkably, the contrastive and quadruplet losses show comparable performance, demonstrating that disentanglement and cross-species representations can arise from reconstruction loss alone, without explicit modeling of the latent space. The wide bottleneck ($d_s = 199, d_t = 3$) slightly improves S-on-S over the compact bottleneck ($d_s = 16, d_t = 2$), which in turn improves disentanglement (lower T-on-S and S-on-T), limits cross-species variability, and enables straightforward visualization of sleep stage representations in the 2-d latent space.

Regarding disentanglement, task information leaks into the style space (T-on-S in table 1), probably due to the fact that the task space is low-dimensional and the model uses the larger capacity in the style space to jointly encode task and style. However, the model's ability to effectively model task information in the task space is not prevented by this. Sleep stages are successfully encoded in the task space, but with species overlap instead. Figure 2 shows the style and task latent spaces. The style space, clearly structured in two human and mouse regions, encodes most of the human-mouse variability (Figure 2a, left), while most species-variability is removed from the task space (Figure 2b, right).

Table 1: Balanced accuracy of the four XGBoost probes described in 3.2, with narrow (16-d in style space, 2-d in task space) and wide (199-d and 3-d) bottlenecks. Mean and standard error of the mean across 4 folds of unseen individuals. Bold font indicates the best loss within the same bottleneck and probe. Arrows denote directionality (↑ = higher is better).

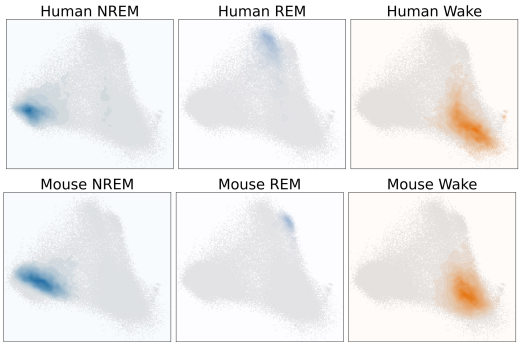
Losses	**Latent dims.**	**S-on-S ↑**	**T-on-T ↑**	**T-on-S ↓**	**S-on-T ↓**
Standard AE	$d_s = 16, d_t = 2$	35.36 ± 2.75	52.84 ± 2.92	$\mathbf{72.23 \pm 1.23}$	10.57 ± 0.72
	$d_s = 199, d_t = 3$	46.13 ± 4.56	50.81 ± 1.05	$\mathbf{87.59 \pm 0.25}$	$\mathbf{11.39 \pm 1.49}$
Contrastive Learning	$d_s = 16, d_t = 2$	$\mathbf{57.46 \pm 3.36}$	84.25 ± 0.61	83.97 ± 0.85	8.21 ± 0.74
	$d_s = 199, d_t = 3$	$\mathbf{63.12 \pm 3.58}$	85.15 ± 0.74	89.49 ± 0.21	11.86 ± 1.06
QSLP-AE	$d_s = 16, d_t = 2$	57.45 ± 5.89	$\mathbf{84.80 \pm 0.52}$	89.88 ± 0.19	$\mathbf{7.09 \pm 0.71}$
	$d_s = 199, d_t = 3$	60.51 ± 6.11	$\mathbf{85.37 \pm 0.32}$	90.31 ± 0.59	13.05 ± 1.25



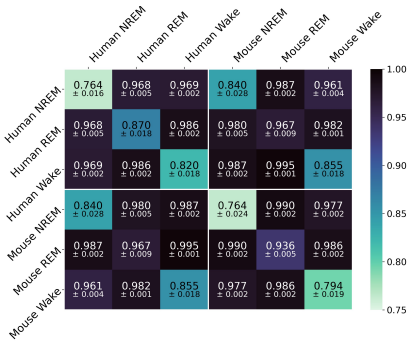
(a) 16-d style latent space (t-SNE reduction). Left: colored by individuals. Right: colored by sleep stage.

(b) 2-d task latent space. Left: colored by sleep stage. Right: colored by species.

Figure 2: Latent spaces. Test set of fold 1 in the QSLP-AE with compact bottleneck ($d_s = 16, d_t = 2$). In (a), individuals are colored in a gradient from mouse to human, with mice and humans generally colored in cold and warm colors respectively, with the corresponding dataset provided in the labels.



(a) Contour plots of the task space. Test set of fold 1 in the QSLP-AE model with compact bottleneck ($d_s = 16, d_t = 2$).



(b) Inter-individual JS distance between sleep stages in the QSLP-AE. Average and standard error of the mean across four folds of individuals unseen during training.

Figure 3: Visualization and quantification of sleep stage correspondence between species.

4.2 Cross-species representations

Figure 3a shows contour plots of the task latent space, demonstrating how corresponding stages occupy the same region independently of species. The distance between stages in the QSLP-AE is shown in Figure 3b. As expected, the inter-individual distance is smallest within the same stage and species, which is clearly seen in the diagonal in Figure 3b. Inter-individual variation within species is lowest for NREM sleep, intermediate for wake, and highest for REM, especially in mice. Regarding cross-species distance, correspondent stages between species present smaller distances, with NREM and wake showing similar distance between species, and REM presenting larger cross-species variability. This can also be seen in Figure 3a, where the REM clusters of both species are next to each other but do not overlap. JS distance for Standard AE and CL is shown in A.9.

5 Conclusion

This work illustrates how representation learning, and specifically the QSLP-AE framework, can be used to bridge neural representations across different species, even on unseen individuals. While the performance of the quadruplet loss is comparable to that of the contrastive, the quadruplet loss comes with the advantage of having a decoder to generate data from the latent space, which could be used to gain further insights. The most interesting application of this framework would be on data from neurological disorders, which would allow for understanding how disease models in animals

actually relate to human disease. A limitation is that sleep stages are more easily differentiated in the EEG, while other mental states of interest, such as attention level, cognitive processes or mood, do not manifest as clearly. Nevertheless, this framework presents itself as a promising avenue to address relevant questions in translational neuroscience.

ACKNOWLEDGMENTS

This work was supported by Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516)".

REFERENCES

Garikoitz Azkona and Rosario Sanchez-Pernaute. Mice in translational neuroscience: What R we doing? *Progress in Neurobiology*, 217:102330, October 2022. ISSN 0301-0082. doi: 10.1016/j.pneurobio.2022.102330. URL https://www.sciencedirect.com/science/article/pii/S0301008222001162.

Helen C. Barron, Rogier B. Mars, David Dupret, Jason P. Lerch, and Cassandra Sampaio-Baptista. Cross-species neuroscience: closing the explanatory gap. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1815):20190633, November 2020. doi: 10.1098/rstb.2019.0633. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2019.0633.

J.F. Cavanagh, D. Gregg, G.A. Light, S.L. Olguin, R.F. Sharp, A.W. Bismark, S.G. Bhakta, N.R. Swerdlow, J.L. Brigman, and J.W. Young. Electrophysiological biomarkers of behavioral dimensions from cross-species paradigms. *Translational Psychiatry*, 11(1), 2021. ISSN 2158-3188. doi: 10.1038/s41398-021-01562-w.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607. PMLR, November 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Wilhelmus H.I.M. Drinkenburg, Gé S.F. Ruigt, and Abdallah Ahnaou. Pharmaco-EEG Studies in Animals: An Overview of Contemporary Translational Applications. *Neuropsychobiology*, 72 (3-4):151–164, February 2016. ISSN 0302-282X. doi: 10.1159/000442210. URL https://doi.org/10.1159/000442210.

Dominik Endres and Johannes Schindelin. A new metric for probability distributions. *Information Theory, IEEE Transactions on*, 49:1858–1860, August 2003. doi: 10.1109/TIT.2003.813506.

Pascal JD Goetghebeur and Jina E Swartz. True alignment of preclinical and clinical research to enhance success in CNS drug development: a review of the current evidence. *Journal of Psychopharmacology*, 30(7):586–594, July 2016. ISSN 0269-8811. doi: 10.1177/0269881116645269. URL https://doi.org/10.1177/0269881116645269.

Joshua Jacobs. Hippocampal theta oscillations are slower in humans than in rodents: implications for models of spatial navigation and memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1635):20130304, February 2014. doi: 10.1098/rstb.2013.0304. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2013.0304.

Emily S. Kappenman, Jaclyn L. Farrens, Wendy Zhang, Andrew X. Stewart, and Steven J. Luck. ERP CORE: An open resource for human event-related potential research. *NeuroImage*, 225:117465, January 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117465. URL https://www.sciencedirect.com/science/article/pii/S1053811920309502.

B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, September 2000. ISSN 1558-2531. doi: 10.1109/10.867928. URL https://ieeexplore.ieee.org/document/867928.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs].

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, January 1991. ISSN 1557-9654. doi: 10.1109/18.61115. URL https://ieeexplore.ieee.org/document/61115/.

Atul Maheshwari. Rodent EEG: Expanding the Spectrum of Analysis. *Epilepsy Currents*, 20(3): 149–153, May 2020. ISSN 1535-7597. doi: 10.1177/1535759720921377. URL https://doi.org/10.1177/1535759720921377.

Anders Vestergaard Nørskov, Alexander Neergaard Zahid, and Morten Mørup. CSLP-AE: A Contrastive Split-Latent Permutation Autoencoder Framework for Zero-Shot Electroencephalography Signal Conversion. In *Thirty-seventh Conference on Neural Information Processing Systems*, November 2023. URL https://openreview.net/forum?id=G7Y145tm2F&noteId=Vv0DjbuHZG.

Huy Phan and Kaare Mikkelsen. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological Measurement*, 43(4):04TR01, April 2022. ISSN 0967-3334, 1361-6579. doi: 10.1088/1361-6579/ac6049. URL https://iopscience.iop.org/article/10.1088/1361-6579/ac6049.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, July 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Abdelrahman Rayan, Anjali Agarwal, Anumita Samanta, Eva Severijnen, Jacqueline Van Der Meij, and Lisa Genzel. Sleep scoring in rodents: Criteria, automatic approaches and outstanding issues. *European Journal of Neuroscience*, pp. ejn.15884, December 2022. ISSN 0953-816X, 1460-9568. doi: 10.1111/ejn.15884. URL https://onlinelibrary.wiley.com/doi/10.1111/ejn.15884.

Allan Rechtschaffen and Anthony Kales. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. University of California, Brain Information Service/BrainResearch Institute, Los Angeles, CA, USA, 1968.

Laura Rose, Alexander Neergaard Zahid, Louise Piilgaard, Christine Egebjerg, Frederikke Lynge Sørensen, Mie Andersen, Tessa Radovanovic, Anastasia Tsopanidou, Stefano Bastianini, Chiara Berteotti, Viviana Lo Martire, Micaela Borsa, Ryan K Tisdale, Yu Sun, Maiken Nedergaard, Alessandro Silvani, Giovanna Zoccoli, Antoine Adamantidis, Thomas S Kilduff, Noriaki Sakai, Seiji Nishino, Sébastien Arthaud, Christelle Peyron, Patrice Fort, Morten Mørup, Emmanuel Mignot, and Birgitte Rahbek Kornum. Probability estimation of narcolepsy type 1 in DTA mice using unlabeled EEG and EMG data. *SLEEP Advances*, 6(2):zpaf025, April 2025. ISSN 2632-5012. doi: 10.1093/sleepadvances/zpaf025. URL https://doi.org/10.1093/sleepadvances/zpaf025.

Yin Tian, Li Yang, Wei Xu, Huiling Zhang, Zhongyan Wang, Haiyong Zhang, Shuxing Zheng, Yupan Shi, and Peng Xu. Predictors for drug effects with brain disease: Shed new light from EEG parameters to brain connectomics. *European Journal of Pharmaceutical Sciences*, 110: 26–36, December 2017. ISSN 0928-0987. doi: 10.1016/j.ejps.2017.04.019. URL https://www.sciencedirect.com/science/article/pii/S0928098717302233.

Matthew M. Troester, Stuart F. Quan, Richard B. Berry, David T. Plante, Alexandre R. Abreu, and Mohammed Alzoubaidi. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 3.* American Academy of Sleep Medicine, Darien, IL, USA, 2023.

Tobias Uelwer, Jan Robine, Stefan Sylvius Wagner, Marc Höftmann, Eric Upschulte, Sebastian Konietzny, Maike Behrendt, and Stefan Harmeling. A survey on self-supervised methods for visual representation learning. *Machine Learning*, 114(4):111, March 2025. ISSN

1573-0565. doi: 10.1007/s10994-024-06708-7. URL https://doi.org/10.1007/s10994-024-06708-7.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Frederick J. Wilson and Philippe Danjou. Early Decision-Making in Drug Development: The Potential Role of Pharmaco-EEG and Pharmaco-Sleep. *Neuropsychobiology*, 72(3-4):188–194, 2015. ISSN 0302-282X, 1423-0224. doi: 10.1159/000382022. URL https://karger.com/article/doi/10.1159/000382022.

Frederick J. Wilson, Steven C. Leiser, Magnus Ivarsson, Søren R. Christensen, and Jesper F. Bastlund. Can pharmaco-electroencephalography help improve survival of central nervous system drugs in early clinical development? *Drug Discovery Today*, 19(3):282–288, March 2014. ISSN 1359-6446. doi: 10.1016/j.drudis.2013.08.001. URL https://www.sciencedirect.com/science/article/pii/S1359644613002626.

Katharina Wulff, Silvia Gatti, Joseph G. Wettstein, and Russell G. Foster. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nature Reviews Neuroscience*, 11(8):589–599, August 2010. ISSN 1471-0048. doi: 10.1038/nrn2868. URL https://www.nature.com/articles/nrn2868.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association: JAMIA*, 25(10):1351–1358, October 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy064.

A Appendix

A.1 Data

A summary of the datasets and the prevalence of each sleep stage is shown in table 2. REM stands for rapid eye movement sleep, a stage characterized by high brain activity, while NREM (non–rapid eye movement) sleep consists of deeper, restorative stages (Wulff et al., 2010).

Data splits Stratified 4-fold cross-validation (CV) is performed across individuals (humans and mice), with approximately 70%, 15% and 15% of the individuals of each dataset allocated to the training, evaluation and test sets respectively. Additionally, a given individual (human or mouse) can be in the evaluation or test set only once across all folds. All results tables show the mean and standard error of the mean across the test set of the 4 folds, and all latent spaces show test data (individuals unseen during training). Likewise, all probes are trained and tested on unseen individuals (see A.6).

Mouse epoch merging Transforming the raw signals from the time domain into PSDs gives human and mouse data the same shape, but the different epoch lengths (30 s vs 4s) still introduce human-mouse differences because shorter signals create noisier PSD estimates. To compensate, seven consecutive mouse epochs of the same sleep stage are merged into 28 s epochs.

Table 2: Datasets overview with ratio of each sleep stage per dataset (%). Humans: Although a single NREM class is used for training and reporting results, we show the full range of human sleep stages according to the standard of the American Academy of Sleep Medicine (Troester et al., 2023). Mice: Mouse epoch counts are based on the longer 28-second epochs.

Dataset	**Channels**	**Individuals**	**Epochs**	**Wake**	**NREM**			**REM**
					N1	**N2**	**N3**	
Human datasets								
SleepEDF	Fpz-Cz, Pz-Oz, submental EMG	77	135,855	0.33	0.11	0.36	0.07	0.13
SHHS	C3/A2, C4/A1, submental EMG	120	109,448	0.22	0.04	0.45	0.13	0.15
Total		197	245,303	0.28	0.08	0.40	0.10	0.14
Mouse datasets								
MSSV1	1 ipsilateral-frontoparietal, neck EMG	10	78,430	0.56		0.38		0.05
MSSV2	2 parietal, 2 frontal, neck EMG	17	23,304	0.49		0.46		0.05
MSSV3	1 parietal, 1 frontal, neck EMG	32	109,155	0.56		0.38		0.06
MSSV4	1 parietal, 1 frontal or 1 cerebellum, 1 frontal, neck EMG	27	9,589	0.27		0.66		0.07
MSSV5	1 parietal, 1 frontal, neck EMG	6	24,483	0.49		0.45		0.06
Total		92	244,961	0.54		0.41		0.05

A.2 SIGNAL PRE-PROCESSING PIPELINE

Some of the datasets provide more than one EEG channel. Since the model uses a single EEG and a single EMG channel, one EEG channel is randomly selected when multiple EEG channels are available.

Each EEG epoch is transformed into its power spectral density (PSD) vector $\boldsymbol{x} \in \mathbb{R}^{128}$, representing power across 128 frequency bins. This representation ensures a consistent input shape for human and mouse data despite differing epoch lengths (30 s vs. 4 s) and avoids the need for the autoencoder to reconstruct phase information, which is arbitrary and not informative for sleep staging given that epochs are randomly spaced in time. The PSD therefore constitutes the main EEG input to the model.

Unlike the EEG, sleep-stage information in the EMG is primarily conveyed by total power rather than frequency content. We therefore use the epoch-wise root mean square (RMS) as the EMG input.

Finally, the EEG PSD is normalized by the total power of each epoch. In addition, analogously to the EMG, we compute the RMS of the EEG signal and provide it as an auxiliary input to the model.

The pre-processing pipeline can be broken down as follows:

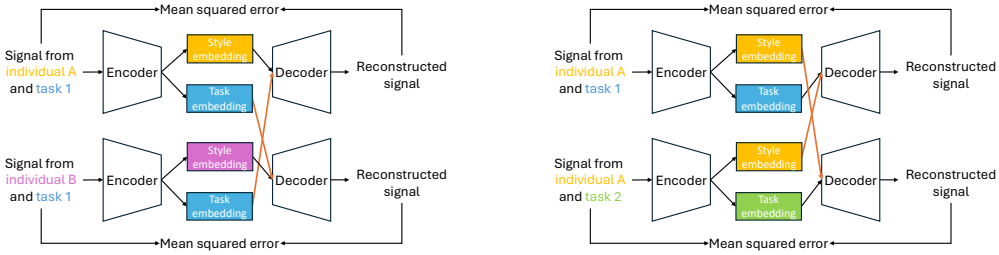
1. Resampling to 100 Hz with anti-alias filtering.
2. If more than one EEG channel available, a random channel is picked.
3. Filtering:
   - EEG: band-pass filtered in the range [0.5-25] Hz with a Butterworth filter of order 15.
   - EMG: high-pass filtered with a cutoff frequency of 10 Hz with a Butterworth filter of order 4.
4. Power spectral density (PSD) of the EEG.
   - Computation: Welch method with a 200-point window size, 100-point step size and 256-point FFT length. This yields a 128-dimensional vector that represents the power along 128 frequency bins.
   - Per-epoch normalization: normalization of the PSD by its total power, which is computed as the integral of the PSD computed with the Simpson method.
5. Root mean squared calculation (both in EEG and EMG). The RMS is z-scored using the mean and standard deviation across all epochs in the same recording.

A.3 TRAINING DETAILS

A batch size of 256 elements is used (128 pairs of samples in the contrastive loss, or 64 quadruplets of samples in the quadruplet loss). The Adam optimizer (Kingma & Ba, 2017) was used with a learning rate of 10^{-4}, decayed to 10^{-5} using cosine annealing starting at 50% of the training. Two XGBoost classifiers were trained and tested on the evaluation set, which contained different individuals than that from training. The XGBoost classifiers were trained to classify individuals and sleep stage from the style and task embeddings respectively. The harmonic mean of the balanced accuracy in individual and sleep stage classification was used to select the best model during training. Because of the conditions imposed by the quadruplet and contrastive losses, we consider an epoch has happened when the model has seen an amount of samples equivalent to the amount of samples in the training set, although samples from underrepresented classes are repeated because of the sampling constraints imposed by the contrastive and quadruplet losses. The models were trained for 220 epochs.

A.4 BASIC PERMUTATIONS

Basic permutations of task (Figure 4a) and individual embeddings (Figure 4b) are shown.



(a) Same-task permutation

(b) Same-individual permutation

Figure 4: Basic permutation logic in the QSLP framework. Two samples with either (a) same task or (b) same individual are sampled. The encoder projects the samples into style and task embeddings. Same-class embeddings are permuted (orange lines) and the input is reconstructed by the decoder. When individuals A and B are from different species, the permutation promotes cross-species representations of tasks (sleep stages).

A.5 ARCHITECTURE

The model (Figure 5) consists of two parallel encoders for style and task, which have 5 blocks of 1-dimensional convolutions in the frequency axis. Within each block, the 1-d convolutions are followed by a 1-d strided convolution that halves the frequency axis each time. Because the input PSD only contains relative spectral power, we concatenate the absolute RMS power of both EEG and EMG channels. Next, attention is conferred by 5 transformer layers (Vaswani et al., 2017). A last 1-d convolution with a larger kernel fully collapses the frequency dimension, so that only the latent dimension (i.e., the number of filters of the last convolution) is left. The style and task embeddings are later concatenated to create a unified embedding. Except when trained with the Contrastive Learning setting, the embeddings are propagated through a symmetric decoder, that reverses the transformations through transposed strided convolutions and yields the reconstructed PSD.

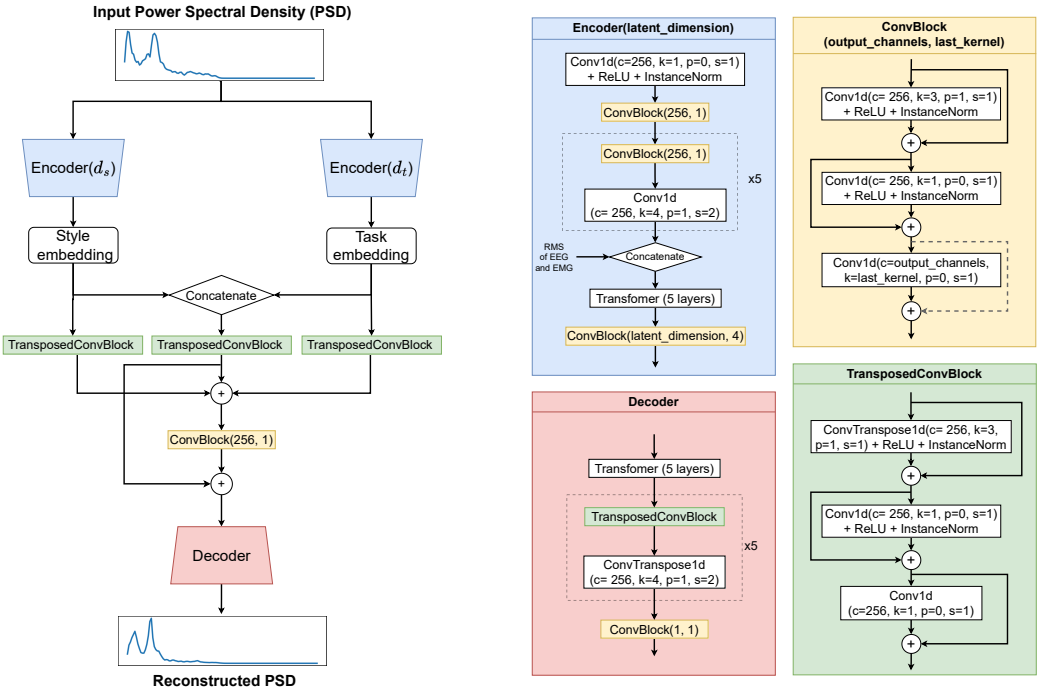


Figure 5: Architecture of the model. Encoder: d_s and d_t are the dimensions of the style and task vector embeddings respectively. 1-d convolutions: c is the number of channels of the convolution, k is the kernel size, p is the padding on both sides of the input, and s is the stride of the convolution.

A.6 LATENT SPACE PROBES

The quality of the learned latent spaces in encoding individuals and sleep stages is evaluated using three classifiers: XGBoost (to assess non-linearity), K-nearest neighbors with 200 neighbors (to assess local structure), and multinomial regression (to assess linearity). Only test-set individuals (unseen during the training of the model used for embedding extraction) are used to train and evaluate the classifiers. Within each of the 4 subject folds (see A.1), the test set is further divided into five sub-folds for cross-validation of the probes. To address strong class imbalance across individuals and sleep stages, each sub-fold's training set is undersampled to match the minority class. Balanced accuracy is used as the performance metric. Results for the losses in the main text and a combination of them are shown in tables 3, 4, 5 for the XGboost, KNN and MR classifiers respectively.

Table 3: Balanced accuracy of the XGBoost classifier, with narrow (16-d in style space, 2-d in task space) and wide (199-d and 3-d) bottlenecks. Mean and standard error of the mean across 4 folds of unseen individuals. Bold font indicates the best loss within the same bottleneck and probe. Arrows denote directionality (↑ = higher is better). S-on-S: style prediction from style latents. T-on-T: task prediction from task latents. T-on-S: task prediction from style latents. S-on-T: style prediction from task latents.

Losses	**Latent dims.**	**S-on-S ↑**	**T-on-T ↑**	**T-on-S ↓**	**S-on-T ↓**
Standard AE	$d_s = 16, d_t = 2$	35.36 ± 2.75	52.84 ± 2.92	$\mathbf{72.23 \pm 1.23}$	10.57 ± 0.72
	$d_s = 199, d_t = 3$	46.13 ± 4.56	50.81 ± 1.05	$\mathbf{87.59 \pm 0.25}$	11.39 ± 1.49
Contrastive Learning	$d_s = 16, d_t = 2$	57.46 ± 3.36	84.25 ± 0.61	83.97 ± 0.85	8.21 ± 0.74
	$d_s = 199, d_t = 3$	63.12 ± 3.58	85.15 ± 0.74	89.49 ± 0.21	11.86 ± 1.06
QSLP-AE	$d_s = 16, d_t = 2$	$\mathbf{57.45 \pm 5.89}$	$\mathbf{84.80 \pm 0.52}$	89.88 ± 0.19	$\mathbf{7.09 \pm 0.71}$
	$d_s = 199, d_t = 3$	60.51 ± 6.11	$\mathbf{85.37 \pm 0.32}$	90.31 ± 0.59	13.05 ± 1.25
Standard AE + CL	$d_s = 16, d_t = 2$	57.44 ± 3.33	84.44 ± 0.79	84.48 ± 0.21	8.52 ± 0.95
	$d_s = 199, d_t = 3$	$\mathbf{64.68 \pm 3.74}$	85.20 ± 0.36	90.49 ± 0.14	12.44 ± 1.78
QSLP-AE + CL	$d_s = 16, d_t = 2$	55.04 ± 4.96	84.27 ± 0.92	85.44 ± 0.43	8.37 ± 0.73
	$d_s = 199, d_t = 3$	63.13 ± 3.37	84.53 ± 0.78	89.95 ± 0.14	$\mathbf{9.77 \pm 1.16}$

Table 4: Balanced accuracy of the KNN classifier, with narrow (16-d in style space, 2-d in task space) and wide (199-d and 3-d) bottlenecks. Mean and standard error of the mean across 4 folds of unseen individuals. Bold font indicates the best loss within the same bottleneck and probe. Arrows denote directionality (↑ = higher is better). S-on-S: style prediction from style latents. T-on-T: task prediction from task latents. T-on-S: task prediction from style latents. S-on-T: style prediction from task latents.

Losses	**Latent dims.**	**S-on-S ↑**	**T-on-T ↑**	**T-on-S ↓**	**S-on-T ↓**
Standard AE	$d_s = 16, d_t = 2$	23.99 ± 4.43	54.60 ± 1.70	$\mathbf{64.68 \pm 0.93}$	10.33 ± 0.64
	$d_s = 199, d_t = 3$	28.17 ± 4.93	49.18 ± 1.24	75.53 ± 1.05	11.54 ± 1.67
Contrastive Learning	$d_s = 16, d_t = 2$	47.11 ± 5.56	83.61 ± 1.06	75.35 ± 0.61	6.84 ± 0.28
	$d_s = 199, d_t = 3$	51.40 ± 4.06	86.28 ± 0.82	$\mathbf{80.43 \pm 0.90}$	10.55 ± 1.21
QSLP-AE	$d_s = 16, d_t = 2$	37.86 ± 8.68	$\mathbf{84.46 \pm 0.47}$	87.06 ± 0.31	$\mathbf{6.59 \pm 0.35}$
	$d_s = 199, d_t = 3$	36.73 ± 9.23	85.66 ± 0.59	86.14 ± 0.81	10.68 ± 1.48
Standard AE + CL	$d_s = 16, d_t = 2$	$\mathbf{48.50 \pm 4.99}$	84.26 ± 1.30	73.21 ± 1.22	7.63 ± 0.60
	$d_s = 199, d_t = 3$	$\mathbf{51.47 \pm 5.35}$	85.99 ± 0.89	82.70 ± 0.76	10.73 ± 1.25
QSLP-AE + CL	$d_s = 16, d_t = 2$	43.72 ± 7.17	83.17 ± 1.66	76.99 ± 1.35	6.86 ± 0.39
	$d_s = 199, d_t = 3$	50.24 ± 4.93	$\mathbf{86.29 \pm 0.69}$	80.88 ± 0.90	$\mathbf{9.74 \pm 1.12}$

Table 5: Balanced accuracy of the Multinomial Regression classifier, with narrow (16-d in style space, 2-d in task space) and wide (199-d and 3-d) bottlenecks. Mean and standard error of the mean across 4 folds of unseen individuals. Bold font indicates the best loss within the same bottleneck and probe. Arrows denote directionality (↑ = higher is better). S-on-S: style prediction from style latents. T-on-T: task prediction from task latents. T-on-S: task prediction from style latents. S-on-T: style prediction from task latents.

Losses	**Latent dims.**	**S-on-S ↑**	**T-on-T ↑**	**T-on-S ↓**	**S-on-T ↓**
Standard AE	$d_s = 16, d_t = 2$	33.08 ± 2.10	50.18 ± 2.58	55.39 ± 2.46	10.50 ± 1.09
	$d_s = 199, d_t = 3$	48.79 ± 3.36	41.62 ± 2.50	85.64 ± 0.49	12.00 ± 1.02
Contrastive Learning	$d_s = 16, d_t = 2$	$\mathbf{56.42 \pm 0.93}$	84.17 ± 1.33	$\mathbf{51.90 \pm 1.58}$	5.27 ± 0.23
	$d_s = 199, d_t = 3$	68.13 ± 2.46	86.19 ± 0.99	$\mathbf{81.13 \pm 0.77}$	6.60 ± 0.08
QSLP-AE	$d_s = 16, d_t = 2$	44.31 ± 1.31	$\mathbf{85.30 \pm 0.72}$	82.76 ± 0.96	$\mathbf{4.03 \pm 0.25}$
	$d_s = 199, d_t = 3$	61.51 ± 3.94	85.57 ± 0.79	87.62 ± 0.54	6.89 ± 0.25
Standard AE + CL	$d_s = 16, d_t = 2$	55.19 ± 1.49	84.60 ± 1.03	53.57 ± 0.75	5.67 ± 0.45
	$d_s = 199, d_t = 3$	$\mathbf{70.36 \pm 2.29}$	86.18 ± 1.09	85.92 ± 0.53	$\mathbf{6.46 \pm 0.43}$
QSLP-AE + CL	$d_s = 16, d_t = 2$	52.68 ± 2.70	84.74 ± 1.48	58.43 ± 3.26	4.84 ± 0.32
	$d_s = 199, d_t = 3$	67.53 ± 2.20	$\mathbf{86.30 \pm 0.80}$	85.15 ± 0.30	6.51 ± 0.20

A.7 Average inter-individual distance between two groups

Suppose we have N individuals split in two groups:

$$\text{Group 1: } \{S_1^{(1)}, S_2^{(1)}, \ldots, S_{N_1}^{(1)}\}, \quad \text{Group 2: } \{S_{N_1+1}^{(2)}, S_{N_1+2}^{(2)}, \ldots, S_{N_2}^{(2)}\}$$

where $N = N_1 + N_2$. $S_i^{(1)}$ and $S_j^{(2)}$ represent the data from individual i in Group 1 and individual j in Group 2, respectively.

We randomly select $N_{\text{pairs}} = \min(N_1, N_2)$ pairs of individuals, with each pair consisting of one individual from each group. In within-species comparisons (where group 1 and 2 are the same), a individual can never be paired with itself.

$$\{(S_{i_k}^{(1)}, S_{j_k}^{(2)})\}_{k=1}^{N_{\text{pairs}}}, \quad i_k \neq j_k$$

The average inter-individual Jensen-Shannon distance between the two groups is then the average distance across the N_{pairs} pairs of individuals:

$$\overline{D}_{\text{JS}} = \frac{1}{N} \sum_{k=1}^{N_{\text{pairs}}} \text{JS-distance}(S_{i_k}^{(1)}, S_{j_k}^{(2)})$$

A.8 Histogram-based Jensen-Shannon distance

Let $X = \{x_1, x_2, \ldots, x_N\}$ and $Y = \{y_1, y_2, \ldots, y_M\}$ be two sets of points in 2D, where $x_i, y_j \in \mathbb{R}^2$. We divide the 2D space into 50 bins and count the number of points in each bin:

$$H_X[i,j] = \text{number of points from } X \text{ in bin } (i,j)$$
$$H_Y[i,j] = \text{number of points from } Y \text{ in bin } (i,j)$$

Each histogram is normalized so that the sum over all bins is 1:

$$P_X[i,j] = \frac{H_X[i,j]}{\sum_{k,l} H_X[k,l]}, \quad P_Y[i,j] = \frac{H_Y[i,j]}{\sum_{k,l} H_Y[k,l]}$$

A small constant $\varepsilon > 0$ is added to each bin to avoid taking the logarithm of zero, and the histograms are re-normalized:

$$P_X \leftarrow P_X + \varepsilon, \quad P_Y \leftarrow P_Y + \varepsilon$$

$$P_X \leftarrow \frac{P_X}{\sum_{i,j} P_X[i,j]}, \quad P_Y \leftarrow \frac{P_Y}{\sum_{i,j} P_Y[i,j]}$$

The average distribution is

$$M = \frac{1}{2}(P_X + P_Y)$$

The Jensen-Shannon divergence between P_X and P_Y is defined as:

$$D_{\mathrm{JS}}(P_X \parallel P_Y) = \frac{1}{2} \sum_{i,j} P_X[i,j] \log_2 \frac{P_X[i,j]}{M[i,j]} + \frac{1}{2} \sum_{i,j} P_Y[i,j] \log_2 \frac{P_Y[i,j]}{M[i,j]}$$

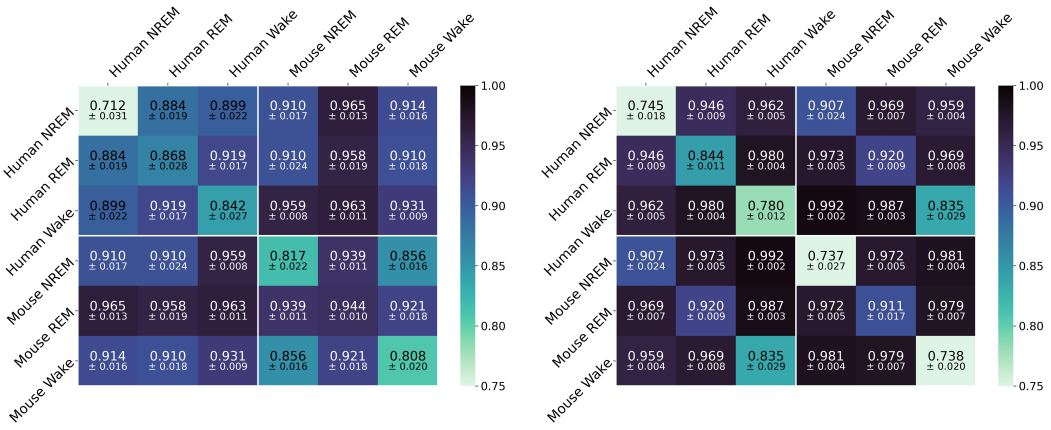
Finally, the square root of the divergence is applied:

$$\text{JS-distance}(X, Y) = \sqrt{D_{\mathrm{JS}}(P_X \parallel P_Y)}$$

A full derivation and the application of the square root to fulfill the mathematical definition of a metric can be found in Lin (1991) and Endres & Schindelin (2003).

A.9 Jensen-Shannon distance in CL and Standard AE

Inter-individual Jensen-Shannon distance matrices for the Standard Autoencoder and the Contrastive Learning models are shown in Figure 6. The Standard Autoencoder shows low overall distance between stages without any specific pattern. This stems from the fact that the task latent space is not effectively encoding stages (table 1. The Contrastive Learning model effectively achieves stage alignment, both within and cross-species.



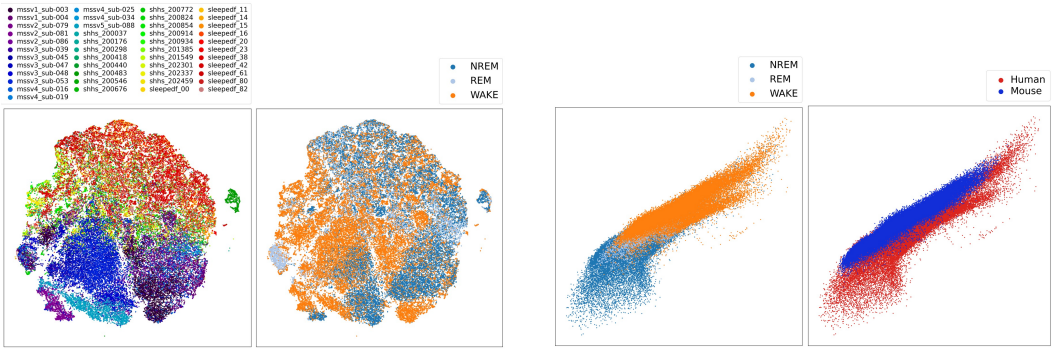
(a) Standard autoencoder

(b) Contrastive learning

Figure 6: Inter-individual Jensen-Shannon distance between sleep stages. Average and standard error of the mean across four folds of unseen individuals.

A.10 Latent space of Standard AE and CL

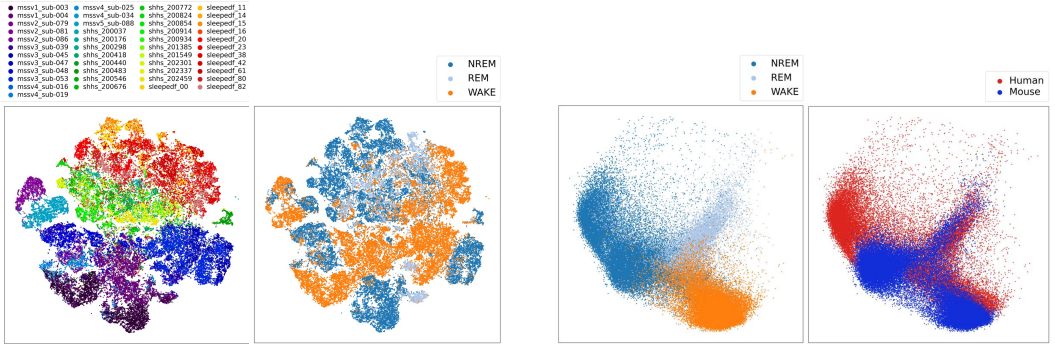
The latent spaces of the Standard Autoencoder model and the model trained with Contrastive Learning are shown in Figures 7 and 8 respectively.



(a) 16-d style latent space (t-SNE reduction). Left: colored by individuals. Right: colored by sleep stage.

(b) 2-d task latent space. Left: colored by sleep stage. Right: colored by species.

Figure 7: Latent spaces. Test set of fold 1 in Standard AE model with compact bottleneck ($d_s = 16, d_t = 2$). In (a), individuals are colored in a gradient from mouse to human, with mice and humans generally colored in cold and warm colors respectively, with the corresponding dataset provided in the labels.



(a) 16-d style latent space (t-SNE reduction). Left: colored by individuals. Right: colored by sleep stage.

(b) 2-d task latent space. Left: colored by sleep stage. Right: colored by species.

Figure 8: Latent spaces. Test set of fold 1 in Contrastive Learning model with compact bottleneck ($d_s = 16, d_t = 2$). In (a), individuals are colored in a gradient from mouse to human, with mice and humans generally colored in cold and warm colors respectively, with the corresponding dataset provided in the labels.

A.11 DATASET ALIGNMENT

A contour plot colored by dataset in shown in Figure 9.

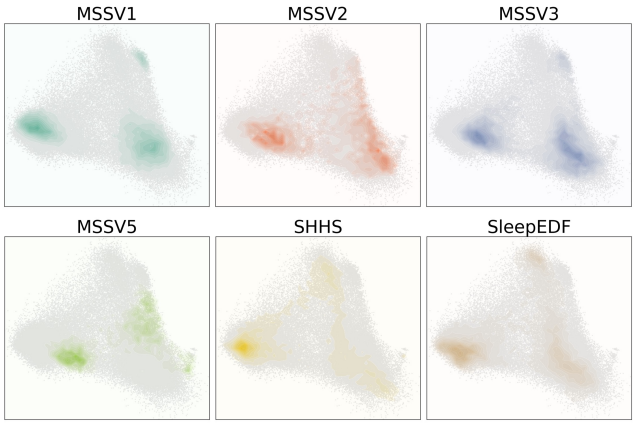


Figure 9: Contour plot of task space colored by dataset. Test set of fold 1 in the QSLP-AE with compact bottleneck ($d_s = 16, d_t = 2$). The smallest dataset (MSSV4) is omitted for figure geometry.

A.12 LLM USAGE DISCLOSURE

In accordance with the conference guidelines, we declare that LLMs have been used exclusively at the writing stage for removing errors and improving the readability of the manuscript. All research ideas, methodological development, and analysis of the results have been contributed by the authors exclusively.