Towards foundation models of naturalistic collective social-neural dynamics

Asaf Benjamin[1,2], **Sergey Anpilov**[1,2], **Omer Izhaki**[1,2], **Maya Sheffer**[1], **Tommaso Biagini**[1,2],
Yair Shemesh[1,2], **Alon Rubin**[1,2], **Inbar Saraf-Sinik**[1,2], **Mackenzie W. Mathis**[3],
Alexander Mathis[3], **Alon Chen**[1,2], **Ofer Yizhar**[1,2]

[1]Department of Brain Sciences, Weizmann Institute of Science, Rehovot, Israel
[2]Department of Molecular Neuroscience, Weizmann Institute of Science, Rehovot, Israel
[3]Brain Mind Institute, School of Life Sciences (SV),
École polytechnique fédérale de Lausanne (EPFL), Geneva, Switzerland

Abstract

Foundation models for the brain and body (FMBB) can transform neuroscience by improving generalization, standardization and reproducibility. However, building generally useful FMBB requires large-scale, high-quality datasets, which capture the complexity and variability of naturalistic social interactions. Unfortunately, measuring neural activity during such interactions is extremely challenging, especially with high spatiotemporal resolution. We have collected a large and unique dataset comprising 17 groups of three to four mice, freely interacting in an enriched environment under continuous video monitoring for one week. We used wireless neural loggers to electrophysiologically record medial prefrontal cortex (mPFC) spiking data from all the group members simultaneously, and we systematically perturbed this neural activity using wireless optogenetics to measure behavioral effects. To study different levels of behavior and their neural representations, we established an extensive, carefully curated and highly accurate preprocessing pipeline, including spike sorting, 3D pose estimation, interpretable behavioral feature extraction, high-level behavior classification, and social dominance hierarchy (SDH) extraction. We then trained foundation models using self-supervised representation learning with CEBRA on the behavioral data, pooled across sessions and animals. We devised a custom method of feature partitioning to make the contrastive learning task more challenging and show that using the learned embeddings instead of behavioral measurements as inputs to downstream models trained to predict neural activity significantly improves their performance. We then use feature attribution methods to show how this can complement classical analysis of neural tuning. Collectively, this work lays the groundwork for building FMBB for naturalistic social contexts and elucidating neural mechanisms during social behavior.

1 Introduction

A fundamental goal in neuroscience is to understand what neurons encode and how they represent this information[1–3]. Most studies approach this objective using reductionist paradigms with isolated individuals or simple pairwise interactions[4,5]. However, mammalian brains evolved to orchestrate complex and adaptive social behaviors, which are necessary for survival and reproduction[6–11]. We therefore aim to understand how the brain represents social information and behavior in a rich and naturalistic social context using groups of freely interacting individuals in an enriched environment (Fig. 1). We use mice as an animal model due to their rich and well-characterized social behavior,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

as well as the availability of powerful methods for genetic targeting and manipulation of neural activity[10,12]. We focus on the infralimbic part of the medial prefrontal cortex (mPFC; Fig. 1e), a brain region known to be crucial for (social) executive functions (e.g. working memory, decision making and planning), processing social information and coordinating social behavior[6,7,10,11,13–25].

To comprehensively characterize mPFC representations of behavior, we describe behavior at different levels using 3D pose estimation, interpretable low-level feature extraction, high-level behavior classification, and SDH extraction. We then utilize *neural encoding* models – models that predict neural activity based on interpretable behavioral variables – and SHAP-based feature attribution, which quantifies each behavioral variable's contribution to the predicted neural activity under rigorous definitions of *fair allocation*[26]. As we show below, this approach can complement classic, more direct single-variable analysis of neural tuning. However, since SHAP and similar methods explain the neurons' activity only through the model's predictions, they are inherently reliant on the accuracy of the models being used. Since high resolution recording of a particular neuron is typically time limited (1-2 h here and often shorter in electrophysiology experiments), leveraging the full power of large-scale machine learning to improve performance is challenging.

To overcome this limitation, we utilize recent breakthroughs in pretraining large-scale foundation models (FM) using self-supervised representation learning (SSRL) with contrastive learning and the recently proposed CEBRA library[27–29]. Since unlabeled data (e.g. behavioral data without neural recording or with neural recording from different session or animals) is abundant, we can use much larger and more expressive models to learn useful (rich, yet compressed) embeddings of behavior without overfitting. We then use these embeddings as inputs to downstream encoding models and show that this significantly improves performance. We then examine these improved models with SHAP values to show how this can complement classical analysis of neural tuning, by revealing when apparent single-feature tuning may be confounded by multivariate effects.

Related work: We first highlight recent work on mPFC codes for ethologically-relevant behavior in *trial-based competitions* in mice[6,7], and on *hippocampal* codes for naturalistic, self-motivated behavior in groups of flying *bats*[8,9]. Further, using large-scale foundation models for neural encoding and decoding is becoming increasingly popular[30–32], as is using machine learning (ML) and related methods to model animal behavior and/or its neural representations (often called *Computational Neuroethology*)[33–46]. In particular, Keypoint-MoSeq uses a *generative* model to *segment* sub-second behavioral syllables in pose dynamics, classifying solitary or social behaviors and capturing correlations with neural activity (e.g. striatal dopamine)[37]. Unlike Keypoint-MoSeq, we use a SOTA, contrastive learning framework, to learn useful representations without segmentation. LISBET uses a transformer model and four contrastive learning tasks specifically designed to segment (pairwise) social interactions, and correlate the behavioral motifs discovered with neural activity in the Ventral Tegmental Area[38]. Unlike LISBET and Keypoint-MoSeq, we aim to explain neural activity in terms of a comprehensive set of interpretable behavioral features, for both individual and *group-wise* behavior, using a more general (time-)contrastive learning task to minimize imposing potential priors or biases on the learned embeddings.

Our main contributions: (1) We collected a large-scale, naturalistic, group-wise, neural-behavioral dataset, with an extensive pipeline to extract interpretable behavioral features and events (1,472 neurons from 30 mice and 192 h of neural-behavioral recordings). (2) We train a self-supervised foundation model using CEBRA and introduce a custom feature-partitioning strategy to harden the contrastive task, demonstrating that the learned embeddings significantly improve neural predictions. (3) We show how these models can be used to gain new neuroscientific insights, e.g. by identifying when single-feature tuning may be confounded by multivariate effects.

2 Dataset collection and processing

2.1 Experimental setup and data collection

To study naturalistic, self-motivated, social and non-social behavior, we collected data from 17 groups of three or four outbred male mice in an enlarged (60x60cm) semi-naturalistic environment that contains all the resources needed for long-term housing (Fig. 1a; Appendix A.1). We continuously monitored the mice with an overhead camera (30 fps) for a week, minimizing experimenter interference. To study how the mPFC represents these naturalistic behaviors with minimal interference,

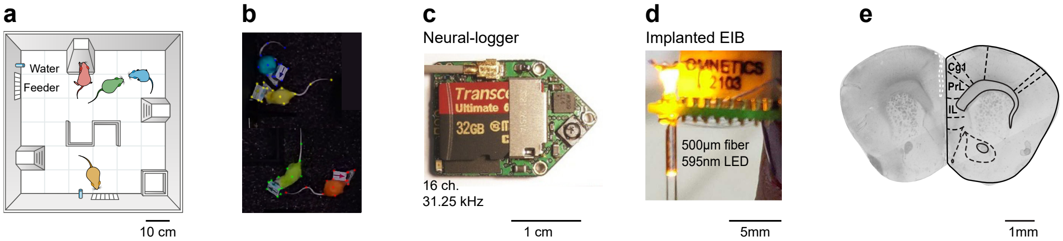

Figure 1: **Experimental setup**. **a.** Schematic of the enriched arena as viewed from the overhead camera. **b.** A representative video frame showing four colored mice freely interacting in the arena while connected to the wireless loggers. The loggers are covered by protective boxes marked with colored arrows on all sides to facilitate pose estimation. Asterisk marks represent the labeled key-points. **c.** A wireless head-mounted neural-logger. **d.** Implanted electrode interface board (EIB) with a 595 nm micro-LED and a 500 μm optic fiber for optogenetic manipulation, and two electrode bundles with eight electrodes each for electrophysiological recordings. **e.** Coronal section of a mouse brain showing mPFC subregions including the infralimbic cortex (IL), the electrode implantation site.

we used wireless head-mounted neural loggers (16 channels, 31.25 kHz) and 16-electrode bundle drives to electrophysiologically record mPFC activity with single-spike, single-unit resolution from all the group members simultaneously (Fig. 1c-e; Appendix A.1). We performed daily 1-2 h neural recording sessions in the arenas at the beginning of the dark phase, when the mice are most active. To study the causal role of mPFC neurons in regulating social behavior, we wirelessly optogenetically silenced mPFC neurons using a micro-LED and eOPN3, a virally targeted, highly sensitive opsin (Fig. 1c-e) while measuring the effect of these perturbations on behavior. Overall, we recorded 5,516 (manually spike-sorted) neurons from 59 mice, over 500 h of simultaneous neural-behavioral data, and over 1,000 h of additional behavior video-only data from the same mice. The analyses we present here include a representative subset of 1,472 neurons from 30 mice and 192 h of neural-behavioral data, which enables us to train large, expressive models without overfitting.

2.2 Behavioral and neural data processing

To comprehensively study behavior and its neural representations at different levels (spatiotemporal resolutions), we developed an extensive, meticulously curated and highly accurate processing pipeline.

2D pose estimation and 3D feature extraction: To accurately track the animals' bodyparts throughout the week, we manually labeled 17 key-points for each mouse in 2,439 video frames and used DeepLabCut (Fig. 1b; Appendix A.2)[47,48]. The model's 8-fold CV median error was 2.1 mm with 96% of errors below 1 cm. From the pose tracking data, we extracted a set of 36 low-level behavioral features for each animal, carefully chosen to balance *interpretability*, *comprehensiveness*, and low *redundancy*. We extracted each animal's 2D location and body orientation, 3D head direction, speed, acceleration and movement direction, as well as its distances, angles and relative movements w.r.t. each other animal (Appendix A.3). These features form the basis for our main neural prediction task.

Behavior classification and SDH: We also used these low-level behavioral features to accurately classify five high-level social behaviors, commonly studied due to their ethological importance: *Approach*, *Avoid*, *Attack*, and either *Aggressive* or *Non-aggressive Chase-Escape* (Appendix A.4). We manually labeled 3.3 h of video on a frame-by-frame basis and trained a gradient boosting classifier[49]. The 10-fold CV auROC (averaged across behaviors) was 0.98. We then used the *Aggressive Chases* predicted by the classifier to extract stable SDH (i.e. to assign social ranks to each group member) using the commonly used *David's Score*[50] (Appendix A.5).

We use each of these behavioral descriptions to study a different level of mPFC representation, though we focus here on the low-level features and social ranks.

Neural signal processing: We applied standard filtering and mean referencing followed by manual spike-sorting based on spike waveforms and timing to isolate single units (Appendix A.6).

3 Modeling: from supervised baselines to self-supervised foundation models

Main task formulation: We predict each mPFC neuron's activity in each recording session separately, based on one of two behavioral descriptions: (1) the 36 interpretable behavioral features above - for supervised learning (SL) baselines; (2) an embedding of these features learned by pretraining FMBB with SSRL (see below and Appendix A.7, Fig. 3). To mitigate performance overestimation due to temporal smoothness/autocorrelation, we used 5-fold CV with continuous 1-min segments (each segment is either in the train or test split) and random circular shuffling to compute chance level performance and statistical significance (Appendix A.9).

Pretraining with self-supervised representation learning: To leverage our abundant unlabeled behavioral data to improve downstream encoding performance, we concatenated the data along the time dimension across all animals and recording sessions. We pretrained a model using SSRL to learn a useful behavioral embedding from the pooled data while completely disregarding neural activity. Then we used that embedding as input to a simple SL model trained to predict each neuron's activity. This approach allows us to share statistical strength across sessions and individuals at the feature-learning stage, effectively building a foundation model of mouse behavior that can be applied to any session or subject.

We used the CEBRA library for SSRL, which provides a fast, easy to use implementation of this approach using contrastive learning, and achieves SOTA results on several datasets[27]. We used CEBRA's time-contrastive variant (fully self-supervised), where the only "supervisory signal" is that observations close in time should have similar embeddings. This encourages the embedding to capture the underlying behavioral state or dynamics that persist over short time scales. We used the 'offset10-model' (a temporal CNN with skip connections and a receptive field of 21 time steps – 2.1 s) with mostly default hyper-parameters (HP; see below and Appendix A.8, Fig. 3). For the final neural predictions, based on these learned embeddings, we used a regularized (ℓ_2, C=1) logistic regression (LR) model[51].

Making the contrastive learning task more challenging (e.g. using "hard negatives") can greatly boost performance on downstream tasks by preventing the model from trivially exploiting direct feature correlations[52]. We take a different approach to achieve a similar result – we first grouped the 36 behavioral features into 12 subsets of three related features each and trained 12 separate networks (one per feature subset/triplet; see Appendix A.8 for further rationale and details). We then concatenated the learned embeddings along the feature dimension to produce the final embedding used for downstream neural predictions. We used an output-dimension of 8 for each feature-triplet-network, yielding a concatenated output dimension of 96. We show below that this approach significantly improved performance over SL baselines, whereas pretraining without feature partitioning yielded no improvement (not shown for brevity).

Baseline supervised learning models: To quantify CEBRA's (potential) performance boost, we trained two baseline SL models on the original 36 behavioral features, instead of the embedding CEBRA learned. The first baseline model was based on the LR model above to compare the performance of an LR model when trained on these two different inputs. The second baseline model was based on the CEBRA model above to test if model expressiveness alone can explain any performance gain (see Appendix A.9). To fairly compare the baseline SL and SSRL models' performance, we also matched their HP and receptive fields (temporal window sizes). We then statistically compared between them and computed chance level performance and statistical significance using random circular shuffling of the neural signal relative to behavior (see Appendix A.9, Fig. 3).

Explaining neural predictions: To utilize these models to complement and improve classical neuroscience analysis, we used SHAP values to quantify the contribution of the original (interpretable) behavioral features to the neural predictions of the combined (CEBRA + LR) model. We wrapped the entire process – feature partitioning, embedding each feature triplet with the corresponding CEBRA model, concatenating the embeddings and neural prediction with the LR – into a single pipeline and used SHAP's model-agnostic *permutation* explainer to explain the predictions (see Appendix A.11).

4 Results: FMBB improves encoding and complements tuning analysis

CEBRA improves encoding performance: Using the SSRL behavioral embedding instead of the original features significantly improved neural encoding performance for the vast majority of mPFC

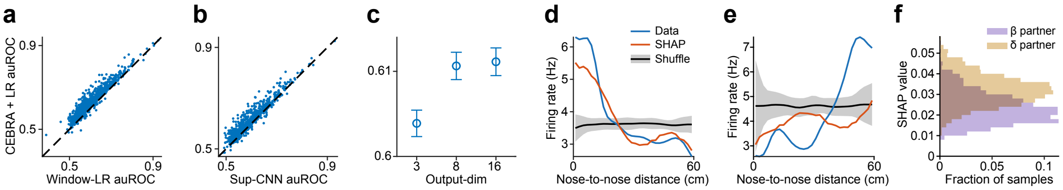

Figure 2: **FMBB improve encoding models and complement neural tuning analysis**. **a-b.** CV-performance (auROC) of baseline LR (a) and CNN (b) trained on the original 36 behavioral features (x-axis) vs LR trained on CEBRA embeddings (y-axis). Points above the diagonal indicate better performance with the CEBRA embeddings. **c.** Mean ± standard error (SE) of CV-performance (auROC) of LR trained on CEBRA embeddings for each CEBRA model output dimension. **d-e.** Two representative examples of a tuning curve (blue) and SHAP curve (orange), respectively showing the mean FR and SHAP values of an mPFC neuron as a function of the nose-to-nose distance between the subject and another group member. The black curve and gray shaded area respectively represent the mean and standard deviation (STD) of the shuffling distribution (see Appendix A.10). **f.** Distribution of SHAP values pooled across features for each of two partners interacting with the subject.

neurons for both the LR and CEBRA-supervised baselines (Fig. 2a-b; Wilcoxon signed-rank test across neurons: $p < 10^{-10}$ for both baselines). Importantly, even with our weaker LR baseline, we found that for 84% (1,233/1,472) of mPFC neurons, the model performed significantly above chance (p<0.05) with a mean z-score of 4.7 across significant neurons, reflecting strong baselines overall (Appendix A.9). These results indicate that a large-scale, expressive model trained on large amounts of (unlabeled) behavioral data can capture predictive latent structure/dynamics that may be missed by directly using the raw features with limited labeled data.

We further examined how the output (embedding) dimension of the CEBRA models affects the performance of downstream neural predictions. While reducing it from 8 to 3 significantly hurt performance (Fig. 2c; p=0.0079, one-way ANOVA followed by Bonferroni correction for multiple comparisons), increasing it to 16 yielded no significant difference (Fig. 2c; p>0.9).

CEBRA + SHAP for neural tuning analysis: We use SHAP values to explain the combined (CEBRA + LR) model predictions for two representative examples of mPFC neurons significantly tuned to the distance between the subject and another group member (Fig. 2d-e; see Appendix A.11). Note that both tuning curves (TC) are significantly different from the shuffling distribution. However, whereas the SHAP curve (SC) of the first neuron, which is similar to the TC, seems to further validate this tuning (Fig. 2d), the SC of the second neuron, which is more similar to the shuffle than to the TC, suggests that other features may be confounding this spurious correlation (Fig. 2e). This exemplifies the limitation of analyzing one or two variables at a time - the typical standard in neuroscience.

CEBRA + SHAP for comparing individuals: Finally, we demonstrate how our model can compare neural representations across social partners. For an 'alpha' mouse (most dominant in the group), we aggregated SHAP values for all features involving each of two specific partners and compared their distributions. Fig. 2f shows an example neuron that responds more to its 'delta' partner than to its 'beta' partner, indicating a stronger mPFC representation of the former.

5 Discussion

This work introduces a large-scale naturalistic neuro-behavioral social dataset and develops SSRL-pretrained foundation models that enhance neural predictions. A potential limitation of the work is that we pooled all the unlabeled data while ignoring potentially important information, such as the subjects' identity and social rank and the recording day, which could account for considerable behavioral-neural variance and improve performance. Moreover, we pretrained our models using strictly self-supervised learning (i.e. completely disregarding neural activity vs. joint neural-behavior modeling[27]). While this is conservative for downstream neural predictions, future work could explore various possibilities for utilizing neural activity to guide the representation learning process (e.g. as auxiliary variables). We also focused on single neurons; future work could extend our foundation model approach to population-level neural dynamics, investigating collective encoding in mPFC. Finally, while our dataset is large and diverse (many individuals freely interacting over long time periods), future work could extend it further to different experimental paradigms and/or brain regions.

Acknowledgments and Disclosure of Funding

A.C. is the incumbent of the Vera and John Schwartz Family Professorial Chair in Neurobiology at the Weizmann Institute of Science. This work was supported by Ruhman Family Laboratory for Research on the Neurobiology of Stress (to A.C.), research support from Bruno and Simone Licht the Perlman Family Foundation, founded by Louis L. and Anita M. Perlman (to A.C.), the Adelis Foundation (to A.C.), Sonia T. Marschak (to A.C.).

This work was supported by the European Research Council (ERC) Consolidator Grant 819496 (to O.Y.).

This work was supported by the WIS Data Science Fellowship (to A.B.).

We thank the WIS AI Hub for Scientific Discovery for feedback on preliminary work on a related problem and dataset, and all members of the Chen lab and Yizhar lab for continuous thoughtful feedback.

The authors declare no competing interests.

References

[1] Laurence F. Abbott and Peter Dayan. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 1st edition, 2001. ISBN 9780262041997. URL https://mitpress.mit.edu/9780262041997/theoretical-neuroscience/.

[2] Nikolaus Kriegeskorte and Xue Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience 2021 22:11*, 22:703–718, 9 2021. ISSN 1471-0048. doi: 10.1038/S41583-021-00502-3. URL https://www-nature-com/articles/s41583-021-00502-3.

[3] Mackenzie Weygandt Mathis, Adriana Perez Rotondo, Edward F. Chang, Andreas S. Tolias, and Alexander Mathis. Decoding the brain: From neural representations to mechanistic models. *Cell*, 187:5814–5832, 10 2024. ISSN 0092-8674. doi: 10.1016/J.CELL.2024.08.051. URL https://www.cell.com/action/showFullText?pii=S0092867424009802https://www.cell.com/action/showAbstract?pii=S0092867424009802https://www.cell.com/cell/abstract/S0092-8674(24)00980-2.

[4] John W. Krakauer, Asif A. Ghazanfar, Alex Gomez-Marin, Malcolm A. MacIver, and David Poeppel. Neuroscience needs behavior: Correcting a reductionist bias, 2 2017. ISSN 10974199.

[5] Nachum Ulanovsky. *Natural neuroscience : toward a systems neuroscience of natural behaviors*. The MIT Press, 2025. ISBN 0262044994.

[6] S. William Li, Omer Zeliger, Leah Strahs, Raymundo Báez-Mendoza, Lance M. Johnson, Aidan McDonald Wojciechowski, and Ziv M. Williams. Frontal neurons driving competitive behaviour and ecology of social groups. *Nature 2022 603:7902*, 603:661–666, 3 2022. ISSN 1476-4687. doi: 10.1038/S41586-021-04000-5. URL https://www-nature-com/articles/s41586-021-04000-5.

[7] Nancy Padilla-Coreano, Kanha Batra, Makenzie Patarino, Zexin Chen, Rachel R Rock, Ruihan Zhang, Sébastien B Hausmann, Javier C Weddington, Reesha Patel, Yu E Zhang, Hao-Shu Fang, Srishti Mishra, Deryn O LeDuke, Jasmin Revanna, Hao Li, Matilde Borio, Rachelle Pamintuan, Aneesh Bal, Laurel R Keyes, Avraham Libster, Romy Wichmann, Fergil Mills, Felix H Taschbach, Gillian A Matthews, James P Curley, Ila R Fiete, Cewu Lu, and Kay M Tye. Cortical ensembles orchestrate social competition through hypothalamic outputs. *Nature*, 603:667, 2022. doi: 10.1038/s41586-022-04507-5. URL https://doi.org/10.1038/s41586-022-04507-5.

[8] Saikat Ray, Itay Yona, Nadav Elami, Shaked Palgi, Kenneth W. Latimer, Bente Jacobsen, Menno P. Witter, Liora Las, and Nachum Ulanovsky. Hippocampal coding of identity, sex, hierarchy, and affiliation in a social group of wild fruit bats. *Science*, 387, 1 2025. ISSN 10959203. URL /doi/pdf/10.1126/science.adk9385.

[9] Angelo Forli and Michael M. Yartsev. Hippocampal representation during collective spatial behaviour in bats. *Nature*, 621:796–803, 9 2023. ISSN 14764687. doi: https://doi.org/10.1038/s41586-023-06478-7. URL https://www.nature.com/articles/s41586-023-06478-7.

[10] Lucy K. Bicks, Hiroyuki Koike, Schahram Akbarian, and Hirofumi Morishita. Prefrontal cortex and social cognition in mouse and man. *Frontiers in Psychology*, 6:1805, 11 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.01805. URL http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.01805/abstract.

[11] Ofer Yizhar and Dana R. Levy. The social dilemma: prefrontal control of mammalian sociability. *Current Opinion in Neurobiology*, 68:67–75, 6 2021. ISSN 0959-4388. doi: 10.1016/J.CONB.2021.01.007.

[12] Mathias Mahn, Inbar Saraf-Sinik, Pritish Patil, Mauro Pulin, Eyal Bitton, Nikolaos Karalis, Felicitas Bruentgens, Shaked Palgi, Asaf Gat, Julien Dine, Jonas Wietek, Ido Davidi, Rivka Levy, Anna Litvin, Fangmin Zhou, Kathrin Sauter, Peter Soba, Dietmar Schmitz, Andreas Lüthi, Benjamin R. Rost, J. Simon Wiegert, and Ofer Yizhar. Efficient optogenetic silencing of neurotransmitter release with a mosquito rhodopsin. *Neuron*, 109:1621–1635.e8, 5 2021. ISSN 0896-6273. doi: 10.1016/J.NEURON.2021.03.013.

[13] Dana Rubi Levy, Tal Tamir, Maya Kaufman, Ana Parabucki, Aharon Weissbrod, Elad Schneidman, and Ofer Yizhar. Dynamics of social representation in the mouse prefrontal cortex. *Nature Neuroscience 2019 22:12*, 22:2013–2022, 11 2019. ISSN 1546-1726. doi: 10.1038/S41593-019-0531-Z. URL https://www-nature-com/articles/s41593-019-0531-z.

[14] Malavika Murugan, Hee Jae Jang, Michelle Park, Ellia M Miller, Julia Cox, Joshua P Taliaferro, Nathan F Parker, Varun Bhave, Hong Hur, Yupu Liang, Alexander R Nectow, Jonathan W Pillow, and Ilana B Witten. Combined social and spatial coding in a descending projection from the prefrontal cortex. *Cell*, 171: 1663–1677.e16, 12 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.11.002.

[15] Nicholas A. Frost, Anna Haggart, and Vikaas S. Sohal. Dynamic patterns of correlated activity in the prefrontal cortex encode information about social behavior. *PLOS Biology*, 19:e3001235, 5 2021. ISSN 1545-7885. doi: 10.1371/JOURNAL.PBIO.3001235. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001235.

[16] Lyle Kingsbury, Shan Huang, Jun Wang, Ken Gu, Peyman Golshani, Ye Emily Wu, and Weizhe Hong. Correlated neural activity and encoding of behavior across brains of socially interacting animals. *Cell*, 178:429–446.e16, 7 2019. ISSN 10974172. doi: 10.1016/J.CELL.2019.05.022/ATTACHMENT/F22D5B55-2857-4DA7-B862-30ED625FD2B0/MMC2.MP4. URL http://www.cell.com/article/S0092867419305501/fulltext.

[17] Bo Liang, Lifeng Zhang, Giovanni Barbera, Wenting Fang, Jing Zhang, Xiaochun Chen, Rong Chen, Yun Li, and Da Ting Lin. Distinct and dynamic on and off neural ensembles in the prefrontal cortex code social exploration. *Neuron*, 100:700–714.e9, 11 2018. ISSN 0896-6273. doi: 10.1016/J.NEURON.2018.08.043.

[18] Fei Wang, Helmut W. Kessels, and Hailan Hu. The mouse that roared: neural mechanisms of social hierarchy. *Trends in Neurosciences*, 37:674–682, 11 2014. ISSN 0166-2236. doi: 10.1016/J.TINS.2014.07.005. URL https://www-sciencedirect-com/science/article/pii/S0166223614001210?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb.

[19] Fei Wang, Jun Zhu, Hong Zhu, Qi Zhang, Zhanmin Lin, and Hailan Hu. Bidirectional control of social hierarchy by synaptic efficacy in medial prefrontal cortex. *Science*, 334:693–697, 11 2011. ISSN 0036-8075. doi: 10.1126/SCIENCE.1209951.

[20] Tingting Zhou, Hong Zhu, Zhengxiao Fan, Fei Wang, Yang Chen, Hexing Liang, Zhongfei Yang, Lu Zhang, Longnian Lin, Yang Zhan, Zheng Wang, and Hailan Hu. History of winning remodels thalamo-pfc circuit to reinforce social dominance. *Science (New York, N.Y.)*, 357:162–168, 7 2017. ISSN 1095-9203. doi: 10.1126/science.aak9726. URL http://www.ncbi.nlm.nih.gov/pubmed/28706064.

[21] Tamara B Franklin, Bianca A Silva, Zinaida Perova, Livia Marrone, Maria E Masferrer, Yang Zhan, Angie Kaplan, Louise Greetham, Violaine Verrechia, Andreas Halman, Sara Pagella, Alexei L Vyssotski, Anna Illarionova, Valery Grinevich, Tiago Branco, and Cornelius T Gross. Prefrontal cortical control of a brainstem social behavior circuit. *Nature Neuroscience*, 20:260–270, 2 2017. ISSN 1097-6256. doi: 10.1038/nn.4470. URL http://www.nature.com/articles/nn.4470.

[22] Lisanne Michelle Jenkins, David Gordon Andrewes, Christian Luke Nicholas, Katharine Jann Drummond, Bradford Armstrong Moffat, Pramit Phal, Patricia Desmond, and Roy Peter Caspar Kessels. Social cognition in patients following surgery to the prefrontal cortex. *Psychiatry Research: Neuroimaging*, 224:192–203, 12 2014. ISSN 0925-4927. doi: 10.1016/J.PSCYCHRESNS.2014.08.007. URL https://www-sciencedirect-com/science/article/pii/S0925492714002078.

[23] Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 2018. doi: 10.1038/s41593-018-0147-8. URL https://doi.org/10.1038/s41593-018-0147-8.

[24] E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202, 11 2003. ISSN 0147006X. doi: 10.1146/ANNUREV.NEURO.24.1.167. URL https://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.24.1.167.

[25] Eunee Lee, Issac Rhim, Jong Won Lee, Jeong-Wook Ghim, Seungjoon Lee, Eunjoon Kim, and Min Whan Jung. Enhanced neuronal activity in the medial prefrontal cortex during social approach behavior. *Journal of Neuroscience*, 36:6926–6936, 6 2016. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0307-16.2016. URL `http://www.ncbi.nlm.nih.gov/pubmed/27358451`.

[26] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *31st Conference on Neural Information Processing Systems (NIPS)*, 2017. URL `https://github.com/slundberg/shap`.

[27] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617:360–368, 5 2023. ISSN 14764687. doi: 10.1038/S41586-023-06031-6;TECHMETA=14,69;SUBJMETA=114,116,1305,2394, 378,631;KWRD=MACHINE+LEARNING,NEURAL+DECODING. URL `https://www.nature.com/articles/s41586-023-06031-6`.

[28] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.*, 2020. URL `https://github.com/google-research/simclr`.

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv*, 2018.

[30] Mehdi Azabou, Georgia Tech, Krystal Xuejing Pan, Vinam Arora, Eva L Dyer, and Blake Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://poyo-plus.github.io`.

[31] Eric Y. Wang, Paul G. Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A. Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, Dat Tran, Jiakun Fu, Casey M. Schneider-Mizell, Nuno Maçarico da Costa, R. Clay Reid, Forrest Collman, Nuno Maçarico da Costa, Katrin Franke, Alexander S. Ecker, Jacob Reimer, Xaq Pitkow, Fabian H. Sinz, and Andreas S. Tolias. Foundation model of neural activity predicts response to new stimulus types. *Nature 2025 640:8058*, 640:470–477, 4 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08829-y. URL `https://www.nature.com/articles/s41586-025-08829-y`.

[32] Yizi Zhang, Yanchen Wang, Mehdi Azabou, Alexandre Andre, Zixuan Wang, Hanrui Lyu, The International Brain Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Neural encoding and decoding at scale. *arXiv*, 5 2025. URL `https://arxiv.org/pdf/2504.08201v3`.

[33] Sandeep Robert Datta, David J. Anderson, Kristin Branson, Pietro Perona, and Andrew Leifer. Computational neuroethology: A call to action. *Neuron*, 104:11–24, 2019. ISSN 08966273. doi: 10.1016/j.neuron.2019.09.038. URL `https://linkinghub.elsevier.com/retrieve/pii/S0896627319308414`.

[34] Jeffrey E. Markowitz, Winthrop F. Gillis, Celia C. Beron, Shay Q. Neufeld, Keiramarie Robertson, Neha D. Bhagat, Ralph E. Peterson, Emalee Peterson, Minsuk Hyun, Scott W. Linderman, Bernardo L. Sabatini, and Sandeep Robert Datta. The striatum organizes 3d behavior via moment-to-moment action selection. *Cell*, 174:44–58.e17, 6 2018. ISSN 10974172. doi: 10.1016/J.CELL.2018.04.019/ATTACHMENT/E43BEBB8-BFD1-4473-8F24-41C7F5672129/MMC3.MP4. URL `http://www.cell.com/article/S0092867418305129/fulltext`.

[35] Alexander B. Wiltschko, Matthew J. Johnson, Giuliano Iurilli, Ralph E. Peterson, Jesse M. Katon, Stan L. Pashkovski, Victoria E. Abraira, Ryan P. Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88:1121, 12 2015. ISSN 10974199. doi: 10.1016/J.NEURON.2015.11.031. URL `/pmc/articles/PMC4708087//pmc/articles/PMC4708087/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4708087/`.

[36] Eleanor Batty, Matthew R Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall, Winthrop Gillis, Jeffrey E Markowitz, Anne Churchland, John Cunningham, Sandeep Robert Datta, Scott W Linderman, and Liam Paninski. Behavenet: nonlinear embedding and bayesian neural decoding of behavioral videos. *33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[37] Caleb Weinreb, Jonah E. Pearl, Sherry Lin, Mohammed Abdal Monium Osman, Libby Zhang, Sidharth Annapragada, Eli Conlin, Red Hoffmann, Sofia Makowska, Winthrop F. Gillis, Maya Jay, Shaokai Ye, Alexander Mathis, Mackenzie W. Mathis, Talmo Pereira, Scott W. Linderman, and Sandeep Robert Datta. Keypoint-moseq: parsing behavior by linking point tracking to pose dynamics. *Nature Methods 2024 21:7*, 21:1329–1339, 7 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02318-2. URL `https://www.nature.com/articles/s41592-024-02318-2`.

[38] Giuseppe Chindemi, Benoit Girard, and Camilla Bellone. Lisbet: a machine learning model for the automatic segmentation of social behavior motifs, 11 2023. URL https://arxiv.org/pdf/2311.04069.

[39] Alexander I. Hsu and Eric A. Yttri. B-soid, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature Communications*, 12:1–13, 12 2021. ISSN 20411723. doi: 10.1038/S41467-021-25420-X;SUBJMETA=1647,1689,2198,2632,378,631;KWRD=BEHAVIOURAL+METHODS,DISEASES+OF+THE+NERVOUS+SYSTEM,MOTOR+CONTROL. URL https://www.nature.com/articles/s41467-021-25420-x.

[40] Kevin Luxem, Petra Mocellin, Falko Fuhrmann, Johannes Kürsch, Stephanie R. Miller, Jorge J. Palop, Stefan Remy, and Pavol Bauer. Identifying behavioral structure from deep variational embeddings of animal motion. *Communications Biology*, 5:1–15, 12 2022. ISSN 23993642. doi: 10.1038/S42003-022-04080-7;SUBJMETA=114,116,1647,2198,378,631;KWRD=BEHAVIOURAL+METHODS,COMPUTATIONAL+NEUROSCIENCE. URL https://www.nature.com/articles/s42003-022-04080-7.

[41] Joeri Bordes, Lucas Miranda, Maya Reinhardt, Sowmya Narayan, Jakob Hartmann, Emily L. Newman, Lea Maria Brix, Lotte van Doeselaar, Clara Engelhardt, Larissa Dillmann, Shiladitya Mitra, Kerry J. Ressler, Benno Pütz, Felix Agakov, Bertram Müller-Myhsok, and Mathias V. Schmidt. Automatically annotated motion tracking identifies a distinct social behavioral profile following chronic social defeat stress. *Nature Communications 2023 14:1*, 14:1–19, 7 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-40040-3. URL https://www.nature.com/articles/s41467-023-40040-3.

[42] Yaning Han, Ke Chen, Yunke Wang, Wenhao Liu, Zhouwei Wang, Xiaojing Wang, Chuanliang Han, Jiahui Liao, Kang Huang, Shengyuan Cai, Yiting Huang, Nan Wang, Jinxiu Li, Yangwangzi Song, Jing Li, Guo Dong Wang, Liping Wang, Yaping Zhang, and Pengfei Wei. Multi-animal 3d social pose estimation, identification and behaviour embedding with a few-shot learning framework. *Nature Machine Intelligence*, 6:48–61, 1 2024. ISSN 25225839. doi: 10.1038/S42256-023-00776-5;SUBJMETA=114,1305,18,601,631;KWRD=ANIMAL+BEHAVIOUR,MACHINE+LEARNING. URL https://www.nature.com/articles/s42256-023-00776-5.

[43] Cristina Segalin, Jalani Williams, Tomomi Karigo, May Hui, Moriel Zelikowsky, Jennifer J. Sun, Pietro Perona, David J. Anderson, and Ann Kennedy. The mouse action recognition system (mars): a software pipeline for automated analysis of social behaviors in mice. *bioRxiv*, page 2020.07.26.222299, 7 2020. doi: 10.1101/2020.07.26.222299. URL https://www.biorxiv.org/content/10.1101/2020.07.26.222299v1https://www.biorxiv.org/content/10.1101/2020.07.26.222299v1.abstract.

[44] Nastacia L. Goodwin, Jia J. Choong, Sophia Hwang, Kayla Pitts, Liana Bloom, Aasiya Islam, Yizhe Y. Zhang, Eric R. Szelenyi, Xiaoyu Tong, Emily L. Newman, Klaus Miczek, Hayden R. Wright, Ryan J. McLaughlin, Zane C. Norville, Neir Eshel, Mitra Heshmati, Simon R.O. Nilsson, and Sam A. Golden. Simple behavioral analysis (simba) as a platform for explainable machine learning in behavioral neuroscience. *Nature Neuroscience*, 27:1411–1424, 7 2024. ISSN 15461726. doi: 10.1038/S41593-024-01649-9;TECHMETA=60,64;SUBJMETA=116,2396,378,3919,631;KWRD=LEARNING+ALGORITHMS,SOCIAL+BEHAVIOUR. URL https://www.nature.com/articles/s41593-024-01649-9.

[45] Alessandro Marin Vargas, Axel Bisi, Alberto S. Chiappa, Chris Versteeg, Lee E. Miller, and Alexander Mathis. Task-driven neural network models predict neural dynamics of proprioception. *Cell*, 187:1745–1761.e19, 3 2024. ISSN 0092-8674. doi: 10.1016/J.CELL.2024.02.036. URL https://www.sciencedirect.com/science/article/pii/S0092867424002393.

[46] Lucas Stoffl, Andy Bonnetto, and Alexander Mathis. Elucidating the hierarchical nature of behavior with masked autoencoders. *ECVA*, 2024. URL https://github.com/amathislab/BehaveMAE.

[47] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21:1281–1289, 9 2018. ISSN 1097-6256. doi: 10.1038/s41593-018-0209-y. URL http://www.nature.com/articles/s41593-018-0209-y.

[48] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Steffen Schneider, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, Venkatesh N. Murthy, George Lauder, Catherine Dulac, Mackenzie Weygandt Mathis, and Alexander Mathis. Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods 2022 19:4*, 19:496–504, 4 2022. ISSN 1548-7105. doi: 10.1038/S41592-022-01443-0. URL https://www-nature-com/articles/s41592-022-01443-0.

[49] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016: 785–794, 3 2016. doi: 10.1145/2939672.2939785. URL `https://arxiv.org/pdf/1603.02754`.

[50] H A David. Ranking from unbalanced paired-comparison data. Technical report, 1987. URL `https://academic.oup.com/biomet/article-abstract/74/2/432/239730`.

[51] Fabian Pedregosa, Vincent Michel, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, David Cournapeau, Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Bertrand Thirion, Olivier Grisel, Vincent Dubourg, Alexandre Passos, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011. ISSN 1533-7928. URL `http://jmlr.org/papers/v12/pedregosa11a.html`.

[52] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 10 2021. URL `https://arxiv.org/pdf/2010.04592`.

[53] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 2 2018. URL `https://arxiv.org/pdf/1802.03426`.

[54] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005. ISBN 9780262256834. doi: 10.7551/MITPRESS/3206.001.0001. URL `https://direct.mit.edu/books/oa-monograph/2320/Gaussian-Processes-for-Machine-Learning`.

[55] Allen Institute. Spikeinterface/spikemetrics: Metrics for spike sorting validation/quality control, 2019. URL `https://github.com/SpikeInterface/spikemetrics`.

[56] William E Skaggs, Bruce L Mcnaughton, Katalin M Gothard, and Etan J Markus. An information-theoretic approach to deciphering the hippocampal code. In *Advances in Neural Information Processing Systems*, volume 5, 1992.

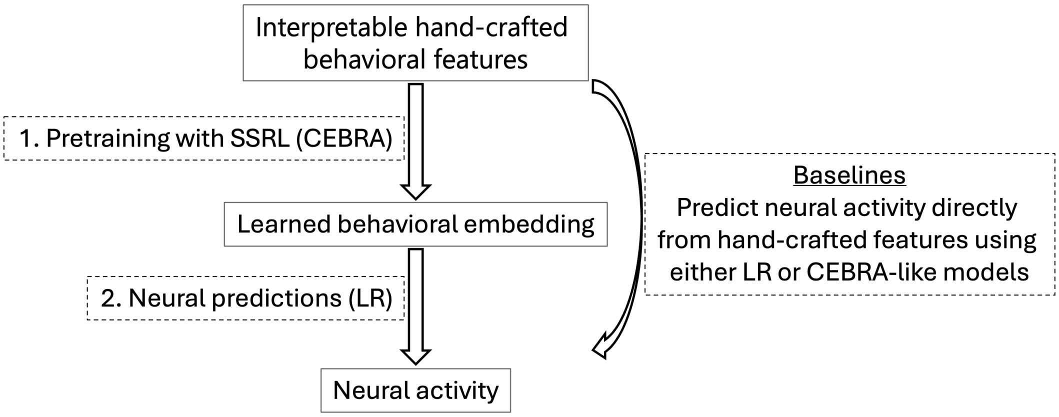
Interpretable hand-crafted behavioral features
1. Pretraining with SSRL (CEBRA)
Learned behavioral embedding
2. Neural predictions (LR)
Neural activity
Baselines
Predict neural activity directly from hand-crafted features using either LR or CEBRA-like models

Figure 3: Diagram illustrating our main modeling approach (left) and the related baselines (right). Text boxes with a continuous or dashed outline represent data representations and modeling steps, respectively. Our main proposed approach is to pretrain the CEBRA models on the interpretable hand-crafted behavioral features, pooled across sessions and animals, to obtain the learned behavioral embeddings. We then train our LR models to predict neural activity based on these learned embeddings. As baselines, we skip the pretraining and predict neural activity directly from the hand-crafted features, using either an LR or CEBRA-like model with HP matched to those used in our main proposed approach.

A Technical Appendices and Supplementary Material

Our main neural encoding approach and the corresponding baselines/controls are illustrated in Fig. 3.

A.1 Experimental setup and data collection

A.1.1 Animals

All the animals used for this study were adult (10–15 weeks old) male ICR mice ($n = 59$). Mice were kept on a 12-12 h light/dark cycle with free access to food and water. All procedures were approved by the Institutional Animal Care and Use Committee (IACUC).

A.1.2 Stereotaxic surgery and microwire array implantation

Mice were anesthetized with 4% isoflurane in an induction box. Once deeply anesthetized, they were immobilized in a stereotactic apparatus using ear bars, and isoflurane flow was provided via mask and reduced to $\sim$1.5%. To prevent postoperative pain, all mice received one dose of buprenorphine (0.1 mg/kg) injected subcutaneously. Microwire electrode arrays were implanted in the infralimbic region of the medial prefrontal cortex (distance from bregma according to the Allen Brain Atlas: $\mathrm{AP} + 1.97$; $\mathrm{ML}\ \pm 0.3$; $\mathrm{DV} - 3.0$) and secured to the skull using metabond and dental cement. To prevent the mice from chewing on each other's drives, quinine (a bitter solution) was applied on the drive using a brush. After surgery, all mice were allowed at least two weeks of recovery before any behavioral experiments began.

A.1.3 Animal color marking

The fur of each mouse was marked in a different color after the surgery and before any behavioral experiments, in order to facilitate automatic video color tracking. The painting was carried out under mild isoflurane anesthesia using commercially available semi-permanent hair dyes in the colors red, green, blue, and yellow (Tish & Snooky's NYC Inc., New York). The dyes were applied using a paintbrush and excess color was removed. After painting, each mouse was separated from the group until the dye dried and then reunited with their original cage-mates.

A.1.4 In-vivo electrophysiological recordings

The multi-electrode fixed drives consisted of a graded bundle of 16 microwires ($25\,\mu\mathrm{m}$ diameter straightened tungsten wires; Wiretronic Inc.), attached to an 18-pin dual-row connector (Omnetics). 16-channel neural loggers (MouseLog16, Deuteron Technologies) were used to wirelessly record extracellular neural activity with a sampling rate of $31.25\,\mathrm{kHz}$. The data were stored on a microSD card for later offline analysis. Each logger was used for continuous recordings of $\sim$60–120 min (81 min on average), depending on its battery capacity.

A.1.5 Protective boxes and 3D head tracking

A manually constructed 5.2×4 cm rectangle made from thin cardboard was folded 1 cm of each edge to create an open box that fits the neural-logger size. The logger's battery was glued (non-permanently) to the inner part of the box such that the 16 pins of the logger can be easily connected to the drive. In addition to the mechanical protection this box provides, we also suited it for 3D head orientation estimation. Each box has five equal and non-mobile sides that are covered with colorful arrow marks (Fig. 1b). As the mice moved their heads, different sides of the box (and the arrows) are exposed to the camera at different angles, making it easier to estimate the angles of the box relative to the floor or the mouse's body, and estimate the 3D orientation of the head.

A.1.6 Semi-naturalistic environment

The semi-naturalistic environment is designed for long-term (one week in this experiment) 24/7 housing of groups of mice under a 12-12 h light/dark cycle. The arena is 60×60 cm and is enriched with bedding, ramps, a labyrinth, a sheltered nest, feeders and water supply. The mice are free to interact with the environment and each other with minimal experimenter interruption or interference and are video-recorded 24/7 with an overhead camera (30 fps, ultra-high sensitivity 4K network camera with a 35 mm full-frame Exmor CMOS sensor, SNC-VB770, Sony) placed 2.6 m above the arena floor.

A.1.7 Synchronization of the cameras and neural-loggers

The neural-loggers synchronization accessory I/O box was used to activate a red LED and an IR light bulb placed within the recorded video frame and the corresponding logger timestamps were logged in the logger's event log file to enable alignment of the logger timestamps with the video frame that captured each light bulb activation. To make sure syncing would be possible even if some of the activations were not recorded by the camera, the light bulbs were activated at pseudo-randomized times throughout the experiment, which provides an unambiguous alignment in such a case.

A.1.8 Histology

Mice were deeply anesthetized using an intraperitoneal injection of Ketamine–Xylazine (160 mg/kg Ketamine, 20 mg/kg Xylazine), and the locations of implanted electrodes were marked with electrolytic lesions (unipolar 100 μA current for 30 s, for each polarity). The lesions enable us to identify the location of the electrode tip while observing the brain slice in the microscope. Twenty minutes following the lesion procedure, mice were further anesthetized using Pentobarbital (130 mg/kg) injected intraperitoneally. Later, transcardial perfusion was made using ice-cold phosphate-buffered saline (PBS, pH 7.4) followed by 4% paraformaldehyde solution (PFA). Brains were extracted, post-fixed overnight at $4\,^{\circ}\mathrm{C}$ in 4% PFA, and then moved to 30% sucrose solution for at least 48 h. Coronal sections ($35\,\mu\mathrm{m}$ thick) were acquired using a microtome (Leica Microsystems) and collected in a $1\times$ PBS solution. Sections were stained with a nucleic acid dye (DAPI, 1:10,000) to better visualize lesion location, and mounted on gelatin-coated slides, dehydrated, and embedded with mounting medium. Overview images ($4\times$) were acquired using a fluorescent microscope and electrode locations were recorded.

A.2 2D Pose estimation

To track each animal's 2D motion throughout the week, we used multi-animal DeepLabCut (v2.2 or v2.3), a commonly used deep learning-based pose estimation framework[47,48]. We defined 17 key-points per mouse: nose tip, ears, centroid of torso, hind legs, tail base, tail tip, tail middle, and the eight corners of the head-mounted logger's protective casing (which aids in estimating 3D head orientation). We manually labeled a large set of 2,439 video frames uniformly sampled from all the videos of all the animals for training. We trained the `dlcrnetms5` (a ResNet-50-based) architecture for 2×10^5 training iterations with augmentation method: `imgaug`, `identity_only: True`; `track_method:ellipse`. We evaluated the model using a standard 8-fold cross-validation scheme, where we partitioned the video frame data into 8 non-overlapping test sets of roughly equal size (up to a rounding to the nearest integer), and the rest of the data for each fold was the training set. We then calculated the median error (mean Euclidean distance between the predicted and manually annotated key-points per frame) across test frames, as well as the percent of errors within 1 cm. The 2D trajectories of these key-points over time form the basis of our behavioral analysis.

A.3 3D low-level behavioral feature extraction

We derived a set of 36 low-level behavioral features from the pose data, designed to comprehensively describe the behavior of the mice in an interpretable way while minimizing redundancy between features. Another guiding principle was to use features that have been studied in neuroscience (at least to some extent), to facilitate interpretation of neural correlates. These features can be naturally grouped into triplets of related features per individual and per pair (see below).

Since we analyze the neural activity of each mouse separately, it is useful to think of the extracted features for each mouse from its perspective, rather than as a general group description. Denote the currently analyzed mouse by A. For each animal A, we included (1) a subset of features that describe its own instantaneous behavior and (2) another subset that describes the behavior of each other group member B with respect to A. The former subset includes the following three triplets of features for each mouse:

1. 2D location (x,y coordinates of the centroid of the torso) and body orientation/azimuth between the tail base and the centroid of the torso. All three are calculated w.r.t. the arena, i.e., in an allocentric reference frame.
2. 3D head direction relative to the body (i.e., in an egocentric reference frame; see explanation below).
3. Speed, acceleration and movement direction of the centroid w.r.t. the arena.

The second subset of behavioral features includes for each animal A and each of the three partners B the following three triplets of (social) features:

1. Relative position and orientation, described in A's egocentric reference frame:
   (a) The nose-to-nose distance between A and B.
   (b) The angle between the azimuth of the line/vector "connecting" the noses of A and B and the body azimuth of A.
   (c) The angle between the body azimuth of B and that of A.
2. 3D head direction of B (see below).
3. Relative movement: B's speed, acceleration, and direction of movement w.r.t. A's location.

3D head direction: To extract the 3D head direction of each mouse relative to its body, we used the tracked 2D coordinates of the eight corners of the boxes protecting the neural loggers, which are rigidly connected to the head. To represent the orientation w.r.t. the mouse's body, we first centered the box coordinates w.r.t. one of the corners (the top-left-rear of the box) and then rotated the points around this point in 2D w.r.t. the body azimuth (described above), so that an angle of $0°$ corresponds to the head being in the same azimuth as the body. This yielded seven non-zero points.

We then used the UMAP algorithm to reduce the 14-dimensional data (7 corners, 2D) of these centered and rotated points to a 3D representation[53]. We used the `run_umap` function in MATLAB (R2023b). Besides the number of components, which was set to 3, we used the default parameters (e.g., `min_dist` $= 0.3$, `spread` $= 1$, `n_neighbors` $= 15$, metric = Euclidean).

We confirmed that neighboring points in the resulting 3D space tightly correspond to neighboring 3D directions of the protective boxes using a ground-truth dataset. We collected these data using a controllable, motorized pan-tilt device that we connected to the boxes and used to position the box in a wide range of 3D orientations with known Euler angles while video recording it in the experimental setup. We then analyzed the data using the same pipeline described above and plotted the UMAP embeddings in 3D, colored by each Euler angle separately, which showed clear clustering of the points by angle, as expected. We were also able to train a *Gaussian Process Regression* model to predict the Euler angles with similarly high CV accuracy when using either the 3D UMAP embeddings or the seven 2D points directly[54]. Thus, the 3D points in the UMAP space used here tightly correspond to 3D head directions.

Missing values imputation and standardization: We imputed missing values for each feature separately using linear interpolation across time, down-sampled the data from 30 fps to 10 Hz using the median of every non-overlapping 3-step window, and standardized each feature vector to have zero mean and unit variance. Together, this set of 36 features provides a rich description of each animal's instantaneous behavior at each moment, from basic locomotion and posture to spatial relations with others.

A.4 High-level behavior classification

For higher-level semantic behaviors, we selected five social behaviors of interest: *Approach, Avoid, Attack, Non-Aggressive Chase-Escape* and *Aggressive Chase-Escape* (see full definitions below). Using the video recordings, trained human annotators labeled occurrences of these behaviors on a frame-by-frame basis. We sampled 20 min of video from each of 10 groups (10 min of continuous video from a video with the loggers and 10 without), yielding 200 min (3.3 h) of annotation.

We augmented the low-level features (Section A.3) with a centered temporal window of length 65 (32 time-steps before the present step, 32 after) and used them as inputs to predict the occurrence of these behaviors in every time step. We trained a gradient boosted decision tree classifier (XGBoost v2.0.3). The model was trained in a one-vs-all fashion for each behavior. We used default XGBoost hyperparameters, except for `max_depth`, which was changed from 6 to 7, yielding a negligible improvement on a separate subset of data. Other regularization hyperparameters (e.g., `eta`) had little-to-no positive effect on performance and were therefore left unchanged.

Performance: The classifier was evaluated with 10-fold cross-validation, such that in each fold we held-out a different group of animals as a test group and trained the model on the other nine groups. This allows us to assess how well our model generalizes to new groups of animals in this paradigm. The model achieved high accuracy for all behaviors, with AUROC ≈ 0.97–0.99 per behavior:

Behavior	auROC
Approach	0.9867
Attack	0.9732
Avoid	0.9802
Aggressive Chase	0.9709
Non-Aggressive Chase	0.9922
Macro-average (5 behaviors)	0.9806

A.4.1 High-level behavior definitions

Approach: The initiator (A) orients toward the target (B) and moves toward B from an initial separation $\geq$ one body length (hindlimb translation $\geq$ half a body length). The event ends at contact (head/body/tail distance $\lesssim 2$ cm) or when B withdraws; if A then continues pursuing B, reclassify as *Chase*, and if A aborts and withdraws, reclassify as *Avoid*.

Avoid: Following close contact ($<$ one body length), one mouse withdraws $\geq$ two body lengths while the other remains approximately stationary (displacement $\leq$ half a body length) until the separation threshold is reached. Starts at head-turn/withdrawal; ends when the animal stops or enters the nest ($\geq$ half body inside). Exclude avoids that immediately follow a chase or nest-to-nest transitions.

Attack: Overt aggressive contact: bites or forceful pushes/kicks directed to the opponent's body (excluding the head). Brief feints without contact are not attacks; sniffing before/after is not part of the event. Discrete bouts separated by other behaviors are annotated as separate attacks.

Non-aggressive chase: Sustained pursuit in which the chaser follows the escaper for $\geq$ two body lengths on a similar path, with the chaser's head behind the escaper's body axis; both animals are in motion for part of the event. Mutual circling/sniffing without pursuit is not a chase. (Open-field / to-shelter / from-shelter variants apply the same core criteria.)

Aggressive chase A *Chase* that includes biting (physical harm) for part of the pursuit. If the interaction transitions into stationary, intensely aggressive wrestling/throwing, annotate that segment as *Attack*; chase segments before/after that still meet chase criteria remain *Chase*.

A.5 Social dominance hierarchy (SDH) extraction

To extract the SDH of each group, we used the model's predictions for *Aggressive Chase–Escape* on all the sessions of each group, a common approach in neuroscience and ethology. We used a standard dominance scoring (*David's Score*), which compares pairs of individuals based on "wins/losses" in such encounters[50]. Each mouse was assigned a dominance rank (Alpha = most dominant, then Beta, Gamma, Delta). These ranks were transitive and highly stable across days (not shown here).

A.6 Neural data preprocessing and unit quality control

All electrophysiology data underwent a standard preprocessing pipeline. We applied a band-pass filter from 300 Hz to 6 kHz to isolate spike frequency bands, then performed common-average referencing (subtracting the mean across channels) to reduce noise. We additionally removed segments with major artifacts (e.g., large amplitude transients visible on many channels simultaneously, often caused by mechanical disturbances).

Spike sorting: was done manually using Plexon Offline Sorter (v4). Putative spike waveforms were extracted using thresholding and then clustered based on raw waveform shapes, as well as principal component analysis and other low-dimensional representations of the data (e.g., waveform peak, valley, energy, etc.). The experimenter then refined clusters, ensuring each single unit had a consistent waveform and clear refractory period. To quantify unit isolation and stability more objectively and rigorously, we computed several quality metrics for each putative unit[55]. To qualify for the analyses reported here, we required that a unit pass three stringent criteria:

- **Refractory Period Violations (RPV):** we required that the proportion of spikes with an inter-spike interval below or equal to 1 ms was lower than 2%.
- **Signal-to-Noise Ratio (SNR):** defined as the ratio of the maximum amplitude of the mean spike waveform to the standard deviation of the background noise on one channel. We required SNR > 3.
- **Presence ratio:** the fraction of the recording in which the unit is firing, to ensure the electrode did not drift away from the unit or the unit otherwise ceased or started firing partway. We required $> 90\%$ for stable units.

Units failing any of these criteria were excluded from further analysis. In the curated subset of 1,472 neurons analyzed, each neuron meets these quality standards, typically surpassing them by a large margin.

Finally, spike trains were binned in 100 ms bins, matching the down-sampled behavioral feature data, and binarized to indicate whether the neuron spiked in each bin or not. We ran additional experiments with (Poisson) regression to the firing rates for some of the analyses and got similar results to the ones described here.

A.7 Main task formulation

We formulate the task as predicting each mPFC neuron's activity in each recording session separately, based on the behavioral features of interest. This is referred to as *encoding* in neuroscience (as opposed to *decoding*, which refers to predicting behavioral or external variables based on neural activity). The main advantages of this approach are as follows. First, as we demonstrated above, using feature attribution methods (SHAP values) on the trained models readily quantifies the extent to which different values of different interpretable behavioral features contribute to the neural predictions, which directly complements classical analysis of neural tuning. Previous work has shown that this kind of analysis is strictly more informative than other common kinds of analysis in neuroscience, in that neural tuning determines other properties (e.g. representational geometry and neural discriminability/decodability) while these properties do not determine neural tuning[2].

Moreover, the dimensionality of our behavioral data (36 features) is larger than that of our neural data (1-36 neurons per animal per recording session, $\sim$ 10 on average). Finally, pooling the behavioral data across recording sessions and animals is more straightforward than pooling the neural data, since behavioral features are individually meaningful and consistent across sessions, while sampled neuron identities are not. Thus, the amount of pooled behavioral data (192 h) is more than two orders of magnitude (143x) larger than the amount of neural data we record per neuron per session (1-2 h, which is typical or even long for electrophysiology experiments). Treating neural data as labels and behavioral data as features therefore allows us to utilize the powerful tools available for leveraging large amounts of unlabeled data to improve performance on downstream tasks – e.g. SSRL-based FMBB.

Cross-validation splits: When training all models to predict neural activity (both with and without the CEBRA embedding), we split each session's data into 5 folds for cross-validation. Because neural and behavioral time series are auto-correlated (smooth), which could lead to performance over-estimation, each session was divided into contiguous 1-minute segments, and entire segments were assigned to either training or testing sets, such that the ith fold's test set contained every ith minute of recording. This greatly reduces leakage of temporally adjacent points between train and test and ensures that each fold's test set covered the whole session duration in a distributed way (rather than, say, one long chunk at the end, which could be a different internal state).

A.8 Self-supervised learning with CEBRA

Rationale for feature partitioning to make the contrastive learning task more challenging: We reasoned that, in the full 36-dimensional space, contrasting pairs that are close or far in time could be trivial, since temporally distant, though behaviorally similar examples of the state of the entire group of animals would be extremely rare. In contrast, using only a small subset of related features (e.g., the 2D location and orientation of the torso, or the 3D head direction of a single animal; see subsection A.3) would make temporally distant, though behaviorally similar examples much more abundant. The model would then need to learn more subtle behavioral differences to minimize the contrastive learning loss.

Model and training: We used CEBRA (v0.4.0) in time-contrastive (purely self-supervised) mode to train the behavioral embedding model. We used the single-session mode, since we have already manually concatenated all the recording sessions into one large "pseudo-session". For each feature triplet, we trained a separate model with the same HP. We used the `offset10-model` (a temporal convolutional neural network) with an output dimension of 8 and 32 hidden units. The InfoNCE contrastive loss was computed using the cosine similarity and a positive pair consisting of two consecutive windows of data (`time_offsets` = 10 time-steps apart) and negatives drawn from different time points across the entire dataset. This yields a total receptive field of 21 time-steps (2.1 s). The model was trained for 10^4 epochs over all the data, using the Adam optimizer with a learning rate of 3×10^{-4}. After obtaining 12 subset embeddings, each of dimension 8, we concatenated them along the feature dimension to form the final 96-dimensional vector.

Full list of parameters for CEBRA

```
{’batch_size’: 4096,
 ’conditional’: None,
 ’criterion’: ’infonce’,
 ’delta’: None,
 ’device’: ’cuda_if_available’,
 ’distance’: ’cosine’,
 ’hybrid’: False,
 ’learning_rate’: 0.0003,
 ’max_adapt_iterations’: 500,
 ’max_iterations’: 10000,
 ’min_temperature’: 0.1,
 ’model_architecture’: ’offset10-model’,
 ’num_hidden_units’: 32,
 ’optimizer’: ’adam’,
 'optimizer_kwargs': (('betas', (0.9, 0.999)),
                      ('eps', 1e-08),
                      ('weight_decay', 0),
                      ('amsgrad', False)),
 'output_dimension': 8,
 'pad_before_transform': True,
 'temperature': 0.5,
 'temperature_mode': 'constant',
 'time_offsets': 10,}
```

Additional details on the implementation of CEBRA can be found in the original publication and documentation[27].

A.9 Baseline encoding models

The baseline models, trained to predict mPFC activity based on the original 36 behavioral features, were constructed to mimic the models used in the SSRL setting with CEBRA.

LR model and training: The SL baseline LR models were the same regularized (ℓ_2, C=1) models used to predict mPFC activity based on the CEBRA embeddings. However, to fairly compare the SL models with the ones trained on CEBRA embeddings, which have a receptive field (temporal window size) of 20 time steps, we complemented the original 36 behavioral features with a centered window of 21 time steps (10 step before the current step and 10 after, 2.1 s total) for the SL LR model, yielding a concatenated and flattened feature vector of length 756. We also tried training the model without temporal features, and with lower or higher values for C, which did not improve performance. The LR models were trained using scikit-learn (version 1.3.2) `LogisticRegression` with the default parameters except for `max_iter`, which was increased to 10^4 to resolve convergence issues on some of the units[51]. We included an intercept term in all models.

CEBRA-based supervised model and training: The CEBRA-based baseline SL models had the same architecture as the CEBRA encoder for a single window (out of a pair): convolutional networks with five time-convolutional layers. The first layer had a kernel size of two, and the next three had a kernel size of three and used skip connections. The final layer had a kernel size of three and mapped hidden dimensions to the output dimension, which in this case is one for binary classification[27]. Thus the receptive field (temporal window size) is 10 steps. We used 32 hidden units. We optimize the binary cross-entropy loss with logits for 10 epochs and a batch size of 4096 using the Adam optimizer with a learning rate of 3×10^{-4}, all in line with the CEBRA HP. Due to runtime constraints, for this comparison, we sampled seven neurons from each recording session (if there were < 7 units in a given session, we used all the units of that session) for a total of 920 (out of 1,472) representative sampled units.

Significance testing: To determine if a model's performance was above chance, we used circular shuffling of the neural data relative to the behavior data. That is, we randomly shifted the spike train time series relative to the behavior features by large (above 1 min), arbitrary offsets (wrap-around at the session end) to destroy any true correlation while preserving each signal's autocorrelation structure. We repeated this procedure 21 times to generate a null distribution of auROC values for each neuron–model. The p-value for the model was computed as the fraction of shuffled trials that achieved an auROC larger or equal to the real model's auROC. We also calculated a z-score for each model by taking the real auROC, subtracting the mean of the shuffled auROC, and dividing by their standard deviation.

A.10 Tuning curves and significance testing

To assess neuronal tuning to social distance, we constructed standard tuning curves by averaging the firing rate of each neuron as a function of the nose-to-nose distance between the subject and a specific social partner. Distances were discretized into uniform 2 cm bins, and mean firing rates were computed for each bin. To establish a null baseline, we performed circular shuffling of the spike trains relative to the behavioral time series (101 iterations), thereby preserving spike count and temporal auto-correlation while disrupting any real neural–behavior relationships. For each neuron, we computed the mean and standard deviation of the shuffled tuning curves, which were overlaid on the empirical tuning curve for visualization.

We quantified tuning strength using the information rate (IR) index commonly used to quantify tuning to spatial or other variables[56]. It measures the amount of information conveyed (reduction in uncertainty) about the variable of interest given the neural activity in bits per seconds or per spike. A neuron was classified as significantly tuned to nose-to-nose distance if its empirical IR exceeded the 95th percentile of the shuffled distribution. This shuffle-based significance criterion corrects for spurious correlations arising from finite sample sizes and temporal auto-correlations in the data.

A.11 Explaining predictions with SHAP values

We used SHAP values to quantify the contribution of the original (interpretable) behavioral features to the neural predictions of the combined (CEBRA + LR) model. We wrapped the entire prediction process—feature partitioning, embedding each feature triplet with the corresponding CEBRA model, concatenating the embeddings and neural prediction with the LR—into a single pipeline and used SHAP's model-agnostic permutation explainer to explain the predictions.

The background distribution was a uniform sample of up to 2,000 rows from the training design matrix, used with the Independent masker for imputation. Explanations were computed on a strided evaluation subset (every 50th sample) to reduce runtime, with a per-instance budget of `max_evals` $= 2 \times 36 + 2048$. We fixed the seeds for reproducibility.

The SHAP curves were computed in the exact same way as the tuning curves, except that we used the SHAP values in each time step instead of the original firing rates.

A.12 Compute hardware

All analysis was performed on a single local GPU (NVIDIA GeForce GTX 1080 Ti) installed on a standard laboratory work-station.