Reconstructing Circuit Connectivity from *in vivo* Spike Trains Using Deep Domain-Adaptive Matching

Kaiwen Sheng[1,2,3,5], Shanghang Zhang[2], Maxime Beau[5], Peng Qu[4], Le Yang[4], Xiaofei Liu[2], Liuyuan He[2], Youhui Zhang[4], Lei Ma[2,3*], Kai Du[1*]

[1]Institute for Artificial Intelligence, Peking University, Beijing, China

[2]School of Computer Science, Peking University, Beijing, China

[3]Beijing Academy of Artificial Intelligence, Beijing, China

[4]Department of Computer Science and Technology, Tsinghua University, Beijing, China

[5]Wolfson Institute for Biomedical Research, University College London, London, United Kingdom

[*]Correspondence: lei.ma@pku.edu.cn (L.M.) and kai.du@pku.edu.cn (K.D.)

Abstract

Inferring the connectivity of neural circuits from *in vivo* experimental data is essential for understanding the neural architectures that underpin behavior and cognition. However, advanced machine learning approaches in this domain, particularly deep learning, are significantly constrained by incomplete *in vivo* observations. This limitation restricts the ability to assign 'ground-truth' labels to *in vivo* data, compelling researchers to rely on synthetic data generated by biophysical neural models for initial training. Yet, this strategy introduces a widely perceived insurmountable challenge: 'model mismatch', where synthetic model dynamics fail to accurately match true neural dynamics. To tackle these interrelated challenges, we introduce DeepDAM (Deep Domain-Adaptive Matching), a robust, adaptive, and versatile framework that fundamentally transforms the training of deep neural networks (DNNs) for inference tasks using both synthetic data and unlabeled *in vivo* data. DeepDAM fine-tunes DNNs on a fusion of synthetic and unlabeled *in vivo* data, adaptively aligning the DNN's feature space with *in vivo* neural dynamics, thereby effectively mitigating the model mismatch challenge. Impressively, we validated the framework with extracellular recordings in the hippocampal CA1 region of freely behaving mice, which achieves an exemplary Matthews correlation coefficient of 0.97-1.0 in connectivity inference accuracy, significantly outperforming existing methods (~0.6-0.7). Importantly, our framework can well adapt to diverse experimental conditions and a broad spectrum of neural properties and scales, demonstrating its high generalizability across various *in vivo* and *ex vivo* scenarios. This marks a significant step towards the accurate and comprehensive reconstruction of functional mammalian brains using data-driven methodologies.

Introduction

Neural circuit connectivity underlies information processing and transmission within the brain, fundamentally shaping network dynamics, behavior, and cognition (1-4). Therefore, accurate reconstruction of this connectivity from *in vivo* recordings is essential for deciphering the computational mechanisms that facilitate the brain's functioning. However, the intrinsic complexity and heterogeneity of neural networks, which are characterized by a diverse array of neuron types and intricate synaptic connections, pose significant challenges in this field. A primary issue is that even state-of-the-art *in vivo* techniques (5-7) can only capture a limited subset of neurons embedded within a vast neural network, leading to incomplete or ambiguous experimental observations. Moreover, researchers often lack sufficient knowledge of the system under investigation.

Addressing these incomplete observations, model-based inference is a crucial technique for inferring neural connections from experimental measurements. This inference paradigm relies on a hypothesized model that simulates (synthesizes) similar neural activities observed in experimental data for inferring connectivity (8). Traditionally, Unsupervised model-based methods like the inverse Ising inference model (9) and Generalized Linear Model (10, 11) are extensively used in this field. Although these approaches are effective in simpler contexts like early sensory areas, Das and Fiete's evaluation using synthetic datasets revealed an inevitable limitation in brain regions with dense recurrent connections (8). This limitation, referred to as the 'model mismatch' problem, specifically arises from discrepancies between the synthetic models used during training and testing.

In contrast, deep learning (DL)-based inference methods are increasingly important due to their natural suitability for large datasets. Specifically, these methods train deep neural networks (DNNs) exclusively on synthetic data generated by the hypothesized biophysical model (12). This training strategy effectively circumvents the need for complete 'ground-truth' connectivity labeling, which is often unattainable due to the incomplete and ambiguous nature of *in vivo* data. Our preliminary results suggest that the DNNs with LSTM-ResNet architecture outperform traditional unsupervised model-based inference methods on synthetic datasets (13). However, when applied to *in vivo* datasets, we observed a significant degradation in performance, which motivates us to conduct a deeper investigation.

In this work, we initially uncovered a significant gap in the feature distributions between synthetic and real *in vivo* datasets, highlighting a critical 'model mismatch' issue. This discovery prompted a fundamental shift in our approach, leading us to develop Deep Domain-Adaptive Matching (DeepDAM), a model-based deep learning framework specifically tailored for circuit connectivity inference tasks that utilize both unlabeled *in vivo* spike trains and synthetic neural dynamics. The key idea of our approach is to reconceptualize the traditional model mismatch problem as a 'data

mismatch' or an 'out-of-distribution' (OOD) challenge—a well-established concern in machine learning but novel in its application to this domain in computational neuroscience.

DeepDAM begins with domain adaptation (14, 15), effectively aligning DNN feature spaces with the true dynamics of neural data, followed by self-training phases that refine DNNs to more accurately adapt to real-world neural recordings (16-20). This strategy not only mitigates the intrinsic discrepancies but also enhances the predictive accuracy and applicability of our models to diverse experimental scenarios. Given that *in vivo* datasets with reliable labels authenticated by the dataset's creators are extremely rare, our application of these methods on such a dataset is pioneering. To the best of our knowledge, we are the first to employ this rigorously labeled dataset for model validation, and therefore among the first to address the model mismatch issue from an out-of-distribution (OOD) perspective in this specific domain.

In this unique dataset, our framework has demonstrated the ability to accurately infer monosynaptic connectivity from *in vivo* multi-electrode extracellular recordings in the hippocampal CA1 region of freely running mice—a region known for its complexity due to limited prior knowledge and the presence of strongly recurrent connections that introduce additional confounds (**Fig. 1a**) (8). By integrating LSTM-ResNet architecture with DeepDAM strategy, our framework achieves an outstanding Matthews correlation coefficient (MCC) of 0.97~1.0 in terms of connectivity inference accuracy, exceeding other popular model-based inference methods (10, 12), which typically score around ~0.6-0.7 in this task. Importantly, the robustness of our framework has been validated across various conditions and datasets, each dataset characterized by different levels of data ambiguity. This demonstrates its reliability and adaptability to a range of experimental scenarios, particularly in noisy and less controlled experimental conditions.

Additionally, when the knowledge of the studied neural system is more comprehensive, DeepDAM can be applied to a wide spectrum of traditional model-fitting tasks, from single neurons to microcircuits, demonstrating its exceptional generality. This adaptability has been showcased through its successful application in inferring biophysical properties from massive intracellular recordings in single neurons in mice (21, 22) and from extracellular recordings in a stomatogastric microcircuit in Cancer Borealis. In these scenarios, our approach has consistently outperformed widely-used model-based inference methods in both efficiency and accuracy (23), underscoring its broad applicability across various experimental contexts.

In summary, the DeepDAM framework represents a paradigm shift in neural circuit connectivity inference, rather than just a specific instance of a 'domain-adaptation' or 'self-training' techniques. It presents a robust solution for leveraging synthetic data to effectively address *in vivo* tasks, thus mitigating the longstanding issue of model mismatch and expanding the utility of computational models across diverse *in vivo* and *ex vivo* contexts. The ultimate aim of our research is the ambitious

reconstruction of entire mammalian and human brains in silicon, directly linking neural circuits to behavior and cognition. We are confident that with future integration of advanced techniques such as meta-learning (24), OOD generalization (25), and various domain adaptation methods, our framework will unlock unprecedented opportunities for analyzing petabyte-scale data and incorporating multi-modal experimental techniques (26). This will significantly enhance our capacity to understand the brain's intricate mechanisms, paving the way for major breakthroughs in neuroscience.

Results

Addressing model mismatch in neural connectivity inference as an out-of-distribution problem

There are two primary challenges arising in the efforts of inferring monosynaptic connectivity from *in vivo* spike trains: (1) Incomplete observations, where the interaction between the recorded and unrecorded neurons can falsely enhance spike-timing correlations, leading to potential misinterpretations (**Fig. 1a**); and (2) Unknown network complexity, where computational models cannot fully cover the intricate nature of brain networks, such as cell-type-specific biophysics and recurrent connections. These challenges manifest themselves as significant 'model mismatch', i.e., discrepancies between the response of the computational model and the actual dynamics of the neural network (as shown in **Fig. 1c**). We further highlight this mismatch through t-distributed stochastic neighbor embedding (t-SNE) visualizations (27), revealing a clear misalignment between model-generated cross-correlograms (CCGs) from synthetic data and those directly obtained from the extracellular multi-electrode recordings in CA1 (28) when displayed in a two-dimensional space (**Fig. 1c**).

The basic idea of solving the model mismatch problem is to consider it as an OOD problem in machine learning, i.e., the performance is significantly degraded due to the mismatch in distribution between the model-generated dynamics and the actual neural dynamics. Domain adaptation techniques are commonly used to deal with the OOD problem by effectively bridging the gap between different data distributions (29). Based on recent advances in unsupervised domain adaptation techniques, instead of narrowing such gap in the original data space, we focus on the shared and informative features of synthetic and experimental data (i.e., feature space) (14, 30, 31). Accordingly, we developed a comprehensive, DNN-based framework for inferring monosynaptic connectivity from unlabeled *in vivo* extracellular recordings:

1. **Model formulation**: A multiple-timescale adaptive threshold (MAT) network (**Fig. 1b**), is formulated to mimic the dynamics of the underlying neural network. This MAT network is a recurrent neural network with excitatory and inhibitory synaptic weights that follow lognormal distributions (**Fig. S1b**) to generate firing rate distributions similar to hippocampal neural recordings (**Fig. S1a**) (12).

2. **Training dataset synthetization**: Exploiting the transparency of this computational model, a synthetic dataset is produced, providing ground-truth labels for cross-correlograms (CCGs) of neuron pairs (**Fig. 1b**). These CCGs, a series of time-lagged correlations between two spike trains, serve as crucial tools for understanding neural connectivity patterns (10, 12, 28, 32, 33).

3. **Feature alignment**: Features that are shared between the synthetic CCGs and experimental CCGs and informative of the underlying connectivity are extracted from DNN-based feature extractors using domain adaptation (**Fig. 1b**).

4. **Classifier construction**: The DNN's classifier is first trained on labeled synthetic data via supervised learning. This step complements feature alignment by making sure that the features are informative about neural connectivity (**Fig. 1b**).

5. **Classifier refinement**: To address potential overfitting of the DNN's classifier to synthetic data, we refine the classifier through self-training using unlabeled experimental data. This is achieved by assigning pseudo labels to the experimental data (**Fig. 1b**).

Our primary contribution to this field is the transformation of the conventional three-step model-based inference method (comprising steps 1, 2, and 4) into a five-step, DNN-based framework ('DeepDAM'). Traditional inference methods, lacking feature alignment and classifier refinement steps, are prone to training biases, as both the feature extractor and classifier are trained exclusively on labeled synthetic data, ignoring the experimental dataset. This results in the classifier overfitting to synthetic features and skewing inferences when applied to experimental data (**Fig. 1d**, **Fig. S2a** & **b**). In contrast, our construction of DeepDAM, which includes Feature alignment and Classifier refinement, can effectively address these shortcomings, thereby minimizing the gap in feature space, preventing the classifiers from overfitting the synthetic data, and ensuring more accurate inferences on experimental data (**Fig. 1e** & **Fig. S2c**).

It is noteworthy that, unlike traditional model-fitting methods, which typically focus on adjusting model parameters to fit data, our approach emphasizes 'model-generation', where the MAT model is used as a generative tool to create synthetic training datasets without fine-tuning the MAT model parameters. This shifts from direct model optimization to 'model-generation' is crucial, as fine-tuning the biophysical network model is often impractical due to the complexities of the underlying biological mechanisms and the extensive number of model parameters (see Discussions).

In the following sections, we will first elaborate on key design elements of the DeepDAM framework's implementation, highlighting its inherent flexibility for diverse applications. Next, we will further validate the performance and robustness of the DeepDAM in comparison to traditional inference methods, using real *in vivo* spike data for inferring monosynaptic connectivity. Finally, we will illustrate the versatility of DeepDAM framework by extending its application to traditional model-fitting tasks, such as inferring biophysical properties of large-scale single neurons from *in vitro* and a microcircuit from *ex vivo* data.

Key implementations of DeepDAM

In our DeepDAM framework, we strategically incorporate and customize two core machine learning strategies—domain adaptation and self-training—to address the OOD challenges inherent in inferring monosynaptic connectivity from *in vivo* recordings. This fusion involves several critical designs, which we will elaborate on below.

The first key design element of DeepDAM is the formulation of composite loss functions for the DNN. In addition to the standard loss function for assessing regular connectivity accuracy, we introduce a composite loss function L comprising the domain adaptation loss (L_{DA}) and the self-training loss (L_{ST}) given by Equation 1. Here, L_{DA} facilitates Feature Alignment, while L_{ST} is crucial for Classifier Refinement.

$$L = L_{DA} + L_{ST} \#(1)$$

The domain adaptation loss in our DeepDAM framework is derived from unsupervised domain adaptation techniques (14, 30, 31) to reduce the distributional discrepancy between synthetic data (x_{syn}) and unlabeled experimental data (x_{exp}) in the feature space, as indicated in Equation 2 and 3:

$$L_{syn} = \frac{1}{N_{syn}} \sum_{i=1}^{N_{syn}} \mathcal{L}\left(g\left(f\left(x_{syn}^{(i)}\right)\right), y_{syn}^{(i)}\right) \#(2)$$

$$L_{DA} = L_{syn} + \lambda_{dis}\, L_{dis}\left(f\left(x_{syn}\right), f\left(x_{exp}\right)\right) \#(3)$$

where $\mathcal{L}$ is the cross-entropy loss function, $y_{syn}^{(i)}$ is the label for $x_{syn}^{(i)}$, g is the classifier and λ_{dis} is the scaling factor of the discrepancy loss. Next, the synthetic and experimental features are categorized into clusters through the spherical K-means clustering algorithm, based on synthetic labels and pairwise feature similarity (Methods) (14). Our aim is to strategically minimize the distance L_{dis} between features within the same cluster and maximize it between different clusters (**Fig. 2b**; Methods). At last, the feature extractor f and the classifier g are co-trained via the DNN to ensure high performance on labeled synthetic data. Overall, the domain adaptation loss (L_{DA}) facilitates the transfer of this high performance from synthetic to experimental data.

However, even with domain adaptation, the classifier is not optimized on experimental features and the feature alignment process cannot perfectly align these two feature distributions, potentially making the classifier overfit on synthetic features. Therefore, we integrated self-training to explicitly refine the classifier to the experimental features (**Fig. 2b**) (16-19). Self-training applies "supervised learning" on unlabeled experimental data by tagging pseudo-labels to them, which are given by the synthetic data labels within the same cluster (**Fig. 2b** & **d**; Equation 4).

$$L_{ST} = \frac{\sum_{i=1}^{N_{exp}} G_{ST}\left(x_{exp}^{(i)}\right) \mathcal{L}\left(g\left(f\left(x_{exp}^{(i)}\right)\right), y_{pseudo}\left(x_{exp}^{(i)}\right)\right)}{\sum_{i=1}^{N_{exp}} G_{ST}\left(x_{exp}^{(i)}\right)} \#(4)$$

where $y_{pseudo}\left(x_{exp}^{(i)}\right)$ is the pseudo label of experimental datum $x_{exp}^{(i)}$ and $G(\cdot)$ is a binary function for gating $x_{exp}^{(i)}$ to the self-training process.

A second key design element in our DeepDAM framework is the introduction of a gating mechanism. At the beginning of the training process, the features extracted and the pseudo-labels assigned may be noisy and inaccurate. To address this, we introduce a gating mechanism that filters out experimental data if their distances to the designated cluster center exceed a predefined threshold. This filtration occurs during the domain adaptation and self-training phases of the current training iteration (as illustrated in **Fig. 2a;** $G_{ST}(\cdot)$ in Equation 4 and $G_{DA}(\cdot)$ in Equation 7), which ensures that only the most relevant and reliable data contribute to the model training at each step. The above steps, including the gating mechanism, are repeated iteratively until the model converges. In each training step, backpropagation (34) is employed to refine the parameters of both the feature extractor and the classifier, guided by the composite loss function described earlier (Methods and SI Appendix).

For clarity and reference, the pseudo code encapsulating the entire training process of our framework is concisely summarized in **Fig. 2c**.

At last, it should be noted that we utilized all available experimental data for domain adaptation and self-training rather than dividing it into validation and testing sets, but we should also emphasize that none of the ground-truth were used throughout the training process. Our primary concern for such design is to reflect the real-world conditions of neuroscience research, where the goal is to infer connectivity from all available data without relying on ground-truth labels. Nevertheless, we have also conducted cross-animal validation experiments to evaluate our model's generalization capabilities (see Discussions for details).

Visualization of feature space evolution during training

Next, to gain insights into the underlying mechanisms of our DeepDAM framework, we visualized the evolution of the feature space during the training process, comparing to the conventional deep learning approach lacking domain adaptation and self-training (i.e. *vanilla* deep learning method). The synthetic dataset was generated by the MAT neural network and the experimental data were a subset of extracellular recordings from the CA1 region of freely-running mice with ground-truth labels (28) (**Fig. 3a**; SI Appendix). After training, both methods were capable of accurately distinguishing

between synthetic connection and non-connection data (**Fig. 3b** bottom right) but only DeepDAM can successfully classify experimental data.

For the vanilla deep learning method, the experimental data were inappropriately positioned between the synthetic data labeled as 'connection' and 'non-connection' in the feature space (**Fig. 3b** bottom right). This misplacement was due to the persistent distributional discrepancy between synthetic and experimental data in the input space, a consequence of model mismatch. As a result, the classifier tended to overfit to the synthetic data, leading to inaccurate classification of the experimental features.

In stark contrast, our method effectively utilized domain adaptation to cluster synthetic and experimental data of the same class together in the feature space (**Fig. 3b** top right). This strategic grouping allowed the classifier, which was trained on synthetic data via supervised learning and on experimental data via self-training, to accurately classify both datasets. As the training proceeds, the pseudo labels for self-training and overall prediction performance gradually stabilized (**Fig. S3**). To quantitatively evaluate the discrepancy between synthetic and experimental feature distributions, we computed the distance between these distributions (**Fig. 3c**). The results revealed that our approach successfully differentiated connection and non-connection data in the feature space, enhancing the classifier's ability to distinguish them accurately (**Fig. 3c** top). On the contrary, the vanilla deep learning method struggled to differentiate between synthetic connection data and experimental non-connection data, and between experimental non-connection and connection data, due to model mismatch and the classifier's overfitting to synthetic data (**Fig. 3c** bottom).

To conclude, our DeepDAM framework offers an effective solution to the model mismatch challenge, ensuring the transfer of precise neural connectivity inference from synthetic data to experimental data.

Evaluation of DeepDAM framework's performance on *in vivo* dataset

In this section, we systematically evaluated the performance of our framework on real *in vivo* spike data with ground-truth labels, comparing with existing popular methods. Specifically, we benchmarked against three methods: the vanilla deep learning approach, CoNNECT (12), and GLMCC (10).

Briefly, CoNNECT is essentially a deep learning method without domain adaptation and self-training, similar to the vanilla deep learning method but with a normal convolution neural network architecture. The authors of CoNNECT provided a DNN model trained solely on synthetic data generated from the same MAT network; we directly utilized their published model for performance comparison (12). GLMCC, in contrast, employed a Generalized Linear Model to fit CCGs based on

synthetic data, which was then applied to rat hippocampal data. In our evaluation, we adopted their model with the original hyperparameters (10), as these were specifically tuned for hippocampal data, ensuring a fair comparison (SI Appendix).

It is essential to highlight the fundamental challenge in neuroscience research: the scarcity of *in vivo* datasets with reliable ground-truth connectivity labels. Direct measurement of synaptic connections in living brains is extraordinarily difficult with current technologies, making ground-truth connectivity largely unknown in most cases. This unique *in vivo* dataset is, to our knowledge, one of the few available with "inferred labels" provided by the dataset creators (28).

Specifically, the dataset's labels were established using two distinct labeling methods, resulting in two subsets: (1) a smaller subset with reliable labels, validated by the consistency between the two labeling methods. This subset allows us to rigorously test the efficacy of our method on real *in vivo* data, providing a dependable benchmark for our techniques; (2) a larger subset with less reliable labels, characterized by inconsistencies between the labeling methods. This subset represents a more common scenario in real-world research, where data ambiguity is prevalent.

Our performance evaluation was based on the smaller subset with reliable labels (28). It is important to emphasize that we did not use any ground-truth labels for training our model; these labels were solely employed to assess the model's prediction accuracy on the experimental data.

We chose MCC as our performance metric, as it is consistent with previous literature (10, 12) and also suitable for imbalanced classification problems (35). The results showed that these three existing methods exhibited noticeable limitations. Predominantly, the vanilla deep learning method tended towards false positives, while both GLMCC and CoNNECT generated a mix of false positives and negatives (**Fig. 4**). In contrast, our method overcame these limitations, achieving an MCC of 0.91 (SD=0.08), in stark contrast to vanilla deep learning (0.61, SD=0.15), GLMCC (0.62), and CoNNECT (0.58).

Interestingly, we observed some performance fluctuations across different random seeds during the training process (**Fig. 4b**). To address this variability and ensure the robustness of our method, we introduced an aggregated version of our method, where we trained the same framework across various random initializations, generating a unique inference prediction for each experimental data point with each random seed. Subsequently, we employed a majority vote algorithm to aggregate the results from all random seeds. Remarkably, our aggregated method successfully eliminated noisy inference results and achieved an outstanding MCC of 0.97 (SD=0.04; **Fig. 4b**; Methods). Our aggregated method is fairly robust which achieved an MCC of 0.91-0.98 across a wide range of hyperparameters (**Error! Reference source not found.**) and such high performance can only be achieved with both self-training and domain adaptation (**Fig. S6**). Similarly, the hyperparameter combination we selected was also based on the consistency across random seeds (SI Appendix).

Evaluation of DeepDAM framework's robustness on *in vivo* dataset

A hallmark of a robust inference algorithm is its consistent performance across varied experimental scenarios. For example, the length of experimental recordings has been shown to significantly influence inference accuracy (10). Therefore, we assessed our method's resilience by subsampling experimental recordings to varying lengths. As recording length increased, our method's performance consistently improved (**Fig. 5b**). This is likely due to longer recordings capturing more nuanced neuron interactions, whereas shorter lengths can obscure critical temporal patterns (**Fig. 5a**). Impressively, even when restricted to 20-minute recordings, our method surpassed other frameworks tested on full-length recordings (0.15-0.27 higher MCC score) (**Fig. 4b** & **Fig. 5b**).

Furthermore, considering the variability in computational models employed in different studies, we tested our method's robustness by training it using three MAT neural networks of different sizes ($N_{neuron} = 250, 500, 1000,$ respectively). The consistent outcome was a robust MCC of 0.99 for $N_{neuron} = 250$ (SD=0.02), 0.97 for $N_{neuron} = 500$ (SD=0.03), 0.97 for $N_{neuron} = 1000$ (SD=0.03) across all models (**Fig. 5d**), highlighting our framework's adaptability across diverse research scenarios. Interestingly, the highest performance, though marginally, was achieved at $N_{neuron} = 250$, suggesting that this size of network model might best mimic the dynamics of the recorded neural ensemble.

We conducted a detailed analysis of the failure cases of all tested methods. Notably, our techniques only resulted in false positives, which were potentially caused by three different reasons. The first reason is the covariation between pairs of neurons (blue examples in **Fig. S4**), where the center peak starts before 0s and the amplitude of the peak is not significantly larger than the baseline fluctuation. Secondly, insufficient number of spikes can lead to false negatives (purple examples in **Fig. S4**). The third reason is that the CCGs for these cases exhibited small peaks just after zero lag (orange examples in **Fig. S4**), which likely led DeepDAM to misclassify them as connections. These concerned CCGs displayed slight peaks after zero, likely causing DeepDAM to incorrectly classify them (**Fig. S4**). Furthermore, these CCGs generally had fewer spikes compared to the clearer CCGs in **Fig. 4**, supporting our observation that longer recording durations tend to enhance performance. We highlighted that the aggregated DeepDAM only failed four CCGs in all 100 aggregations, demonstrating the robustness of our approach. Conversely, other methods displayed a mix of false positives and false negatives. For the false positives of all other methods, some of them are also due to the lack of spikes (purple examples in **Fig. S4**). While for other mistakes, interestingly, they are largely affected by covariations between pairs of neurons, which widely exist in the neural system, such neural oscillations and common inputs. Strikingly, all other methods made false negatives even if there is a very clear peak after 0s in the CCGs (right two columns from 3rd to 6th rows of **Fig. S4**).

This suggests that methods developed solely on synthetic data have limited generalization capabilities on real experimental data.

To conclude, our DeepDAM framework, which synergizes domain adaptation and self-training, effectively mitigates the model mismatch problem and delivers outstanding performance in inferring monosynaptic connectivity from *in vivo* spike trains. Its demonstrated robustness, evident in handling varying recording lengths and accommodating different model structures, establishes the framework as a versatile and powerful tool, offering significant potential for broader applications in the field.

Application to a broader *in vivo* spike dataset with low signal-to-noise ratios

As we transition our method to real-world scenarios, its capability to effectively handle uncertain or noisy data becomes increasingly important. In the previous section, our framework was tested on the subset from English et al. (28), where ground-truth labels were consistently established by two labeling methods, suggesting that the CCGs from this subset typically exhibit high signal-to-noise ratios. To further evaluate the robustness of our framework, we expanded our analysis to include a broader set of *in vivo* spike data characterized by lower signal-to-noise ratios. This larger dataset, also sourced from English et al.'s study (28), contained CCGs with inconsistent or uncertain ground-truth labels according to the two tagging methods, which is a scenario likely more common in *in vivo* experiments (**Fig. 6a**; SI Appendix).

Interestingly, in the enlarged experimental dataset, our aggregated method demonstrated greater consistency with the second ground-truth tagging method, which tended to identify connections more frequently than the first method (**Fig. 6b & c**; Discussions). Yet, discrepancies still existed between our method and both tagging methods, requiring further investigation to determine the most effective approach. Our previous evaluations suggest that increasing the recording time, reducing CCG noise and enhancing prediction reliability could be a viable solution (**Fig. 5b**). Additionally, we encountered cases where neither tagging method could definitively classify certain CCGs as connections or non-connections due to insignificant statistical tests. Our inference method also exhibited uncertainty in these cases (**Fig. 6a** bottom row), primarily due to these CCGs having an insufficient number of spikes for the corresponding neuron pairs (**Fig. 6a** bottom row), underscoring the need for longer recording times for neurons with lower firing rates to yield convincing inferences (**Fig. 5b**) (10).

In conclusion, our findings emphasize the need for comprehensive evaluations using large-scale datasets with accurate ground truths to refine and enhance connectivity inference methods. Improving the ground-truth labeling process and expanding the dataset size are essential steps to enhance the robustness of the inference methods.

Versatility in inferring biophysical properties across neural scales

In this final section, we briefly introduced the versatility of our DeepDAM framework, showcasing its adaptability for traditional model-fitting tasks. Based on the foundational Equation 1-4, the framework can be easily extended to infer biophysical properties of both single neurons and microcircuits. This extension involves making subtle modifications to the discrepancy function L_{dis}, details of which are elaborated in the Supplementary Information. Unlike the previous described model-generation paradigm for connectivity inference, in model-fitting tasks, model parameters are iteratively fine-tuned so that the model's responses closely align with experimental observations in terms of selected biophysical features.

As for the single-neuron inference, our framework inferred biophysical properties of single neurons of mice using intracellular recordings *in vitro*, including passive membrane dynamics, background noise magnitude, and attributes of ion channels. Such properties align with the parameters set by the single-compartment Hodgkin-Huxley model, known for capturing the vast dynamics of cortical neurons (**Fig. S8a**; SI Appendix) (36). We assessed the z-scored mean absolute error (MAE) between biophysical patterns of synthetic model responses and experimental data, including spiking statistics and membrane potential behaviors (**Fig. S8d**). Impressively, our method efficiently inferred biophysical patterns of 574 neurons across 14 brain regions in the Allen Cell Types Dataset (22) (**Fig. S9a** & **b**), with consistent inference accuracy across different regions. The results revealed that both spike and subthreshold patterns, which were major distinguishable patterns across different neuron types, were well predicted, suggesting that our inference results could potentially facilitate studying subcellular differences among different neuron types. Among all biophysical patterns, the half-spike width is the most poorly tuned biophysical pattern (**Fig. S9c** & **d**). One potential reason is that the single-compartment HH model is relatively simple to fit this biophysical pattern, while increasing the number of (sodium) channels could give a better fitting accuracy (21). Moreover, when compared with evolutionary search (ES) algorithm, perhaps the most popular method for single-neuron biophysics inference (23), our approach demonstrated superior performance, achieving an inference score 1.89 times higher (i.e., number of good inferences given the same sample size; our framework: 574, ES: 303, total: 1052) (**Fig. S9e**; SI Appendix).

Furthermore, our framework inferred 31 distinct properties of the intricate stomatogastric ganglion microcircuit of the *Cancer Borealis* (**Fig. S10a**), including nuances of enriched ion channels and diverse synapses, which aims to replicate the circuit's characteristic pyloric patterns (**Fig. S10b**, **c** & **d**; SI Appendix). To compare the inference performance, we evaluated it against the state-of-the-art sequential neural posterior estimation (SNPE) method, specifically SNPE-C (37) (performance given by the pretrained model from Gonçalves et al. (23); SI Appendix). Our framework demonstrated a 2.54 times smaller inference error for all biophysical patterns than SNPE-C (MAE of our framework:

0.0136, SD=0.0024; MAE of SNPE-C: 0.0346, SD=0.0048) (**Fig. S10e**). In summary, the above results underscore our framework's robustness and adaptability in neural analysis across multiple scales.

Discussion

The present study introduced a novel model-based deep learning inference framework to tackle the challenging problem of neural circuit connectivity reconstruction from *in vivo* recordings, which could potentially bridge the gap between neural circuitry functions and its architecture (1-4, 38, 39). It has been demonstrated that our framework overcame the limitations of existing model-based methods (10, 12) and accurately inferred monosynaptic connectivity in the hippocampal CA1 region of freely running mice. The key innovation of our approach lies in its ability to address the challenges posed by the model mismatch problem, which often leads to biased inference outcomes (8). By integrating domain adaptation and self-training with computational models, our framework adapts to discrepancies between synthetic and experimental data and establishes a mapping from paired spike trains to their corresponding connectivity, significantly improving the accuracy of connectivity inference. We have also demonstrated the broad applicability of our framework in inferring biophysical properties of single neurons and microcircuits. By extending the framework to classic model fitting tasks, we successfully inferred a range of biophysical properties of single neurons in the mouse primary visual cortex and the stomatogastric ganglion microcircuit of the *Cancer Borealis*. In both cases, our approach outperformed existing methods in accuracy and efficiency (21, 23), highlighting its potential for advancing the study of neural computation and function across various observational scales.

One of the major challenges in inferring neural connectivity from *in vivo* recordings is the partial observation of neurons, leading to incomplete knowledge of the network. As a result, computational models for the biological network often fail to capture the repertoire of the real biological dynamics and could produce substantial model artefacts. Eliminating such mismatch based on prior knowledge is difficult due to the intricate complexity and insufficient understanding of the underlying biological mechanisms. Consequently, this unrecognized mismatch can introduce bias in connectivity inference when solely relying on the model as the reference (8). Our framework effectively addressed these obstacles by learning domain-invariant and informative features through domain adaptation and fine-tuning the connectivity classifier using experimental data through self-training. As a result, our method achieved 0.97-1.0 MCC to infer monosynaptic connectivity of the mice CA1 region on a dataset with high signal-to-noise ratio, significantly exceeding current methods. Notably, these existing methods all demonstrated near-perfect inference results on synthetic data (10, 12) but failed to transfer such excellent performance to real experimental data. Similar trend was seen in model-based inference framework in other domains, such as inferring biophysical properties of single neurons and microcircuits (13, 40). Our study not only formulated this issue as the OOD problem in machine learning and proposed a unified solution, but also emphasized the importance of evaluating model-based inference techniques using real experimental data.

Another straightforward and widely used method for tackling the mismatch problem in robotics and autonomous driving is domain randomization (41-43). This technique involves varying model parameters to enhance the diversity of synthetic data from different domains. This could potentially enable the machine learning model to learn domain-invariant features that can generalize to real-world applications. Therefore, to test whether increasing the diversity and size of the synthetic dataset can lead to better inference, we simulated 27 different MAT networks by altering three keys: background noise, connectivity ratio, and network size to generate a synthetic dataset that was 16.5 times larger than the previous one (SI Appendix). Our tests revealed that domain randomization increased the MCC by 0.11 for the vanilla deep learning method and by 0.03 for its aggregated version. Despite these improvements, the performance still lagged behind that of DeepDAM and its aggregated version by 0.17 and 0.23 MCC respectively (**Fig. S5**). These results suggest that while domain randomization can enhance the performance of standard deep learning approaches, the benefits are relatively modest, likely due to the high-dimensional nature of the model and the complexity of the neural structure involved. Interestingly, domain randomization can also improve the performance of DeepDAM to an MCC score of 1.0, suggesting that combining DeepDAM with other advanced machine learning techniques is a promising future direction.

The development of large-scale datasets with ground-truth connectivity is urgently needed to further assess and compare these methods, especially because different methods yield varying inference predictions on ambiguous data (**Fig. 6**) and the performance of our method has already saturated on this available dataset. For these ambiguous data, our aggregated method was generally more consistent with the second modified label tagging method (**Fig. 6d**). We introduced this modified label tagging method because we observed that the experimentally validated ground-truth tagging method was excessively strict. It demanded the peak of the CCG to significantly surpass the slowly co-modulated baseline by a considerable margin (p_{fast} threshold: 0.001) and to be significantly higher than the peak in the anti-causal direction (**Fig. 6b**) (28). We relaxed the p_{fast} threshold to 0.01 for three consecutive time bins and eliminated the requirement in the anti-causal direction (i.e., no p_{causal} prerequisites; **Fig. 6b**; SI Appendix), so the adjusted tagging method leaned toward identifying more connections. The fact that our aggregated inference method aligned closer with the modified tagging method (**Fig. 6c** top row) suggests that the original labeling criterion often led to the classification of numerous CCGs with relevant connectivity information as indeterminate, such as bidirectional connections or those coexisting with network modulations (**Fig. 6c** top row). Mitigating the impact of these ambiguous CCGs can be accomplished in several ways. Firstly, extending the recording duration (**Fig. 5b**) can be a straightforward approach, as many experimental methods support chronic recordings of the same group of neurons (7). Alternatively, incorporating probabilistic methods (44), ensemble methods (45) or noisy pseudo labels (16, 18, 19) into our existing framework can explicitly model the uncertainty in our inference results of such ambiguous cases.

In this study, we tailored the DeepDAM training process to better address real-world experimental requirements. We included all experimental data for domain adaptation and self-training without using any ground-truth labels during training. We chose not to segregate data from different animals for evaluation primarily for two reasons: 1) Ideally, in real-world settings, we aim to analyze all available data rather than just a subset; 2) Variations in data across subjects can reduce inference accuracy, similar to the mismatch observed between synthetic and experimental data. Addressing the issue of cross-subject generalization, which remains a challenge in the field, we assessed the generalization ability of our DeepDAM framework using a "leave-one-animal-out" approach across 8 mice. The generalization performance of both DeepDAM (MCC: 0.88, SD=0.15) and its aggregated version (MCC: 0.95, SD=0.07) was remarkably close to their initial performance (DeepDAM MCC: 0.91, SD=0.08; DeepDAM aggregated MCC: 0.97, SD=0.03) and substantially better than other methods, including those employing domain randomization (**Fig. S7**). These results underscore the adaptability and effectiveness of DeepDAM in handling the inherent variability found in real-world experimental settings.

Our results also demonstrated the robustness and general applicability of the DeepDAM framework, which is designed to be agnostic to brain areas, species and other diversity of the neural system. The framework depends on a mechanistic model capable of approximating the dynamics of the underlying systems, such as the firing rate distribution. Therefore, once such simple mechanistic model has been obtained in another species, such as the visual cortex in primates (46), the DeepDAM framework can be easily applied. While it stands to gain from advancements in more precise mechanistic models, its potential significance lies in advancing the mechanistic modeling of large-scale neural circuits. This is crucial because commonly used techniques like artificial recurrent neural networks (47) and multicompartmental neuronal networks (48, 49) often encounter significant model mismatch issues, which can also be conceptualized as an OOD problems (50). Therefore, we anticipate that DeepDAM, in conjunction with these modeling techniques, will promote a closed-loop, iterative process for designing and refining mechanistic models to directly deduce the system's architecture and biophysics from experimental data.

Furthermore, DeepDAM can be easily extended to the classic model-fitting task for biophysical inference of single neurons and microcircuits through changing the discrepancy loss and achieved high accuracy and efficiency in both cases. One limitation of DeepDAM on model fitting is that it can only output one set of feasible neural properties to match the experimental data, while many studies have demonstrated there can be multiple sets of feasible properties (23, 51-56). This could be potentially achieved by replacing the L_{syn} with the loss function used in SNPE methods (23).

In summary, our framework substantially alleviates the OOD problem in inferring monosynaptic connectivity and biophysical properties from experimental data. This indicates that our framework can

potentially facilitate the interaction between experimental and modeling studies to uncover the relationship between low-level neural properties and high-level functions. With continued advancements in experimental techniques and machine learning methodologies, our framework paves the way for a more profound understanding of the brain's intricate mechanisms, leading to transformative insights into neural computation and cognition. For example, our framework explores the mechanistic formation of neural representations underlying behavior and cognition from a bottom-up perspective, providing a valuable complement to current statistical and geometrical analysis on large-scale neural representations from a top-down perspective (57-59). Furthermore, our framework is in line with the rising next-generation ML techniques that utilize simulators and synthetic data to learn richer representations of real world (42). By tightly fusing computational modeling and ML techniques, our framework may inspire ground-breaking methods that use enriched computational models and synthetic data to address cutting-edge challenges in multiple fields (42, 60), such as behavioral analysis (61), protein function prediction (62), and medicine and healthcare (63).

Methods

Here we describe the training details of the DeepDAM framework and leaves other methodological details in SI Appendix.

Training configurations to infer monosynaptic connectivity

As our method is purely unsupervised and does not use any ground-truth labels, the entire experimental dataset was used for domain adaptation and self-training during the training process. The batch size is 800. The initial learning rate is 0.0001 and adapted by the Adam optimizer (64) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The discrepancy loss is defined as the estimated contrastive domain discrepancy at the feature space (14), given by Equation 5-7:

$$L_{dis}\left(f\left(D_{syn}\right), f\left(D_{exp}\right)\right) \triangleq \frac{1}{2}\left(\underbrace{\sum_{c=0}^{1} \widehat{D}^{cc}\left(f\left(D_{syn}^{c}\right), f\left(D_{exp}^{c}\right)\right)}_{intra\ class\ distance} - \underbrace{\sum_{c=0}^{1}\sum_{\substack{c'=0 \\ c'\neq c}}^{1} \widehat{D}^{cc'}\left(f\left(D_{syn}^{c}\right), f\left(D_{exp}^{c'}\right)\right)}_{inter\ class\ distance}\right) \#(5)$$

$$D_{syn}^{c} = \left\{x_{syn} | c = y_{syn}, \left(x_{syn}, y_{syn}\right) \in D_{syn}\right\} \#(6)$$

$$D_{exp}^{c} = \left\{x_{exp} | c = y_{pseudo}\left(x_{exp}\right), x_{exp} \in D_{exp}, G_{DA}\left(x_{exp}\right) = 1\right\} \#(7)$$

where $G_{DA}(\cdot)$ is the binary gating function for domain adaptation and $\widehat{D}^{cc'}\left(f\left(D_{syn}^{c}\right), f\left(D_{exp}^{c}\right)\right)$ $(c, c' \in \{0,1\})$ is the empirical kernel mean embedding which describes the distributional distance at the feature space between synthetic data labeled as class c and experimental data pseudo-labeled as class c'. The empirical kernel mean embeddings of the contrastive domain discrepancy are given by Equation 8-12:

$$D^{pq}\left(f\left(D_{syn}\right), f\left(D_{exp}\right)\right) = e_1 + e_2 - 2e_3, p, q \in \{0, 1\} \#(8)$$

$$e_1 = \frac{\sum_{i=1}^{N_{syn}} \sum_{j=1}^{N_{syn}} \mu_{pq}\left(y_{syn}^{i}, y_{syn}^{j}\right) \sum_{l=1}^{5} k^{l}\left(f\left(x_{syn}^{i}\right), f\left(x_{syn}^{j}\right)\right)}{\sum_{i=1}^{N_{syn}} \sum_{j=1}^{N_{syn}} \mu_{pq}\left(y_{syn}^{i}, y_{syn}^{j}\right)} \#(9)$$

$$e_2 = \frac{\sum_{i=1}^{N_{exp}} \sum_{j=1}^{N_{exp}} \mu_{pq}\left(y_{pseudo}^{i}, y_{pseudo}^{j}\right) \sum_{l=1}^{5} k^{l}\left(f\left(x_{exp}^{i}\right), f\left(x_{exp}^{j}\right)\right)}{\sum_{i=1}^{N_{exp}} \sum_{j=1}^{N_{exp}} \mu_{pq}\left(y_{pseudo}^{i}, y_{pseudo}^{j}\right)} \#(10)$$

$$e_3 = \frac{\sum_{i=1}^{N_{syn}} \sum_{j=1}^{N_{exp}} \mu_{pq}\left(y_{syn}^{i}, y_{pseudo}^{j}\right) \sum_{l=1}^{5} k^{l}\left(f\left(x_{syn}^{i}\right), f\left(x_{exp}^{j}\right)\right)}{\sum_{i=1}^{N_{syn}} \sum_{j=1}^{N_{exp}} \mu_{pq}\left(y_{syn}^{i}, y_{pseudo}^{j}\right)} \#(11)$$

$$\mu_{pq}\left(y^{i}, y^{j}\right) = \begin{cases} 1, \text{if } y^{i} = p \text{ and } y^{j} = q \\ 0, \text{otherwise} \end{cases} \#(12)$$

where k^l is a Gaussian kernel with a variance of 2^l normalized by the number of samples and total distance (14). The discrepancy scaling factor is set as 0.001. The gating threshold for eliminating experimental data from domain adaptation is 0.01 and the gating threshold for self-training is 0.01.

Pseudo labels of experimental data are obtained by applying the spherical K-means clustering algorithm at the feature space among synthetic data and experimental data. The initial cluster center of class c is the synthetic cluster center, given by Equation 13. Then the algorithm repeats the following two steps until convergence or reaching the maximum iterations: 1) assigning each experimental datum a pseudo label based on the nearest cluster center; 2) updating the cluster center based on experimental data, given by Equation 14.

$$O_0^c = \sum_{i=1}^{N_{syn}} I\left(y_{syn}^{(i)} = c\right) \frac{f\left(x_{syn}^{(i)}\right)}{\left\|f\left(x_{syn}^{(i)}\right)\right\|}, c \in \{0, 1\} \#(13)$$

$$O_t^c = \sum_{i=1}^{N_{exp}} I\left(y_{pseudo}\left(x_{exp}^{(i)}\right) = c\right) \frac{f\left(x_{exp}^{(i)}\right)}{\left\|f\left(x_{exp}^{(i)}\right)\right\|}, c \in \{0, 1\} \#(14)$$

where O_0^c is the initial cluster center for class c, O_t^c is the cluster center for class c at training step t and I is the indicator function. Crucially, synthetic data labels ensure that synthetic data with different labels are not grouped in the same cluster.

As the nature of the pseudo labels is noisy, we incorporated the generalized cross-entropy loss (Equation 15) which has been demonstrated to be noise-robust (20).

$$L_{ST} = \frac{(1-c)(1-(1-p_c)^q) + c\left(1-p_c^q\right)}{q} \#(15)$$

Where $p_c = g\left(f\left(x_{exp}\right)\right)$ is the output probability for class c of DeepDAM. In this study, we chose $q = 1$.

For the aggregated method, we trained the DeepDAM framework for 10 different random seeds. The we ensembled the predictions from 5 randomly selected seeds and used the mean probability for each class as the aggregated probability to compute our performance metrics. This procedure was repeated 100 times.

References

1. A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, L. F. Abbott, Optimal degrees of synaptic connectivity. *Neuron* **93**, 1153-1164 e1157 (2017).

2. A. Litwin-Kumar, S. C. Turaga, Constraining computational models using electron microscopy wiring diagrams. *Curr Opin Neurobiol* **58**, 94-100 (2019).

3. N. Brunel, Is cortical connectivity optimized for storing information? *Nat Neurosci* **19**, 749-755 (2016).

4. K. D. Harris, T. D. Mrsic-Flogel, Cortical connectivity and sensory coding. *Nature* **503**, 51-58 (2013).

5. G. Buzsaki, Large-scale recording of neuronal ensembles. *Nat Neurosci* **7**, 446-451 (2004).

6. J. J. Jun, N. A. Steinmetz, J. H. Siegle, D. J. Denman, M. Bauza, B. Barbarits, A. K. Lee, C. A. Anastassiou, A. Andrei, C. Aydin, M. Barbic, T. J. Blanche, V. Bonin, J. Couto, B. Dutta, S. L. Gratiy, D. A. Gutnisky, M. Hausser, B. Karsh, P. Ledochowitsch, C. M. Lopez, C. Mitelut, S. Musa, M. Okun, M. Pachitariu, J. Putzeys, P. D. Rich, C. Rossant, W. L. Sun, K. Svoboda, M. Carandini, K. D. Harris, C. Koch, J. O'Keefe, T. D. Harris, Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232-236 (2017).

7. N. A. Steinmetz, C. Aydin, A. Lebedeva, M. Okun, M. Pachitariu, M. Bauza, M. Beau, J. Bhagat, C. Bohm, M. Broux, S. Chen, J. Colonell, R. J. Gardner, B. Karsh, F. Kloosterman, D. Kostadinov, C. Mora-Lopez, J. O'Callaghan, J. Park, J. Putzeys, B. Sauerbrei, R. J. J. van Daal, A. Z. Vollan, S. Wang, M. Welkenhuysen, Z. Ye, J. T. Dudman, B. Dutta, A. W. Hantman, K. D. Harris, A. K. Lee, E. I. Moser, J. O'Keefe, A. Renart, K. Svoboda, M. Hausser, S. Haesler, M. Carandini, T. D. Harris, Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science* **372**, (2021).

8. A. Das, I. R. Fiete, Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nat Neurosci* **23**, 1286-1296 (2020).

9. E. Schneidman, M. J. Berry, 2nd, R. Segev, W. Bialek, Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007-1012 (2006).

10. R. Kobayashi, S. Kurita, A. Kurth, K. Kitano, K. Mizuseki, M. Diesmann, B. J. Richmond, S. Shinomoto, Reconstructing neuronal circuitry from parallel spike trains. *Nat Commun* **10**, 4468 (2019).

11. W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, E. N. Brown, A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J Neurophysiol* **93**, 1074-1089 (2005).

12. D. Endo, R. Kobayashi, R. Bartolo, B. B. Averbeck, Y. Sugase-Miyamoto, K. Hayashi, K. Kawano, B. J. Richmond, S. Shinomoto, A convolutional neural network for estimating synaptic connectivity from spike trains. *Sci Rep* **11**, 12087 (2021).

13. K. Sheng, P. Qu, L. Yang, X. Liu, L. He, Y. Zhang, L. Ma, K. Du, A general LSTM-based deep learning method for estimating neuronal models and inferring neural circuitry. *bioRxiv*, 2021.2003.2014.434027 (2021).

14. G. Kang, L. Jiang, Y. Yang, A. G. Hauptmann, Contrastive adaptation network for unsupervised domain adaptation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4888-4897 (2019).

15. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17**, 2096–2030 (2016).

16. A. Iscen, G. Tolias, Y. Avrithis, O. Chum, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (2019), pp. 5070-5079.

17. D.-H. Lee, in *Workshop on challenges in representation learning, ICML*. (Atlanta, 2013), vol. 3, pp. 896.

18. E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, K. McGuinness, in *2020 International Joint Conference on Neural Networks (IJCNN)*. (IEEE, 2020), pp. 1-8.

19. W. Shi, Y. Gong, C. Ding, Z. M. Tao, N. Zheng, in *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018), pp. 299-315.

20. Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems* **31**, (2018).

21. N. W. Gouwens, J. Berg, D. Feng, S. A. Sorensen, H. Zeng, M. J. Hawrylycz, C. Koch, A. Arkhipov, Systematic generation of biophysically detailed models for diverse cortical neuron types. *Nat Commun* **9**, 710 (2018).

22. Allen Cell Types Database. (2015).

23. P. J. Gonçalves, J. M. Lueckmann, M. Deistler, M. Nonnenmacher, K. Ocal, G. Bassetto, C. Chintaluri, W. F. Podlaski, S. A. Haddad, T. P. Vogels, D. S. Greenberg, J. H. Macke, Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife* **9**, (2020).

24. T. Hospedales, A. Antoniou, P. Micaelli, A. Storkey, Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence* **44**, 5149-5169 (2021).

25. Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, P. Cui, Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, (2021).

26. N. L. Turner, T. Macrina, J. A. Bae, R. Yang, A. M. Wilson, C. Schneider-Mizell, K. Lee, R. Lu, J. Wu, A. L. Bodor, A. A. Bleckert, D. Brittain, E. Froudarakis, S. Dorkenwald, F. Collman, N. Kemnitz, D. Ih, W. M. Silversmith, J. Zung, A. Zlateski, I. Tartavull, S. C. Yu, S. Popovych, S. Mu, W. Wong, C. S. Jordan, M. Castro, J. Buchanan, D. J. Bumbarger, M. Takeno, R. Torres, G. Mahalingam, L. Elabbady, Y. Li, E. Cobos, P. Zhou, S. Suckow, L. Becker, L. Paninski, F. Polleux, J. Reimer, A. S. Tolias, R. C. Reid, N. M. da Costa, H. S. Seung, Reconstruction of neocortex: Organelles, compartments, cells, circuits, and activity. *Cell* **185**, 1082-1100 e1024 (2022).

27. L. Van der Maaten, G. Hinton, Visualizing data using t-SNE. *Journal of machine learning research* **9**, (2008).

28. D. F. English, S. McKenzie, T. Evans, K. Kim, E. Yoon, G. Buzsaki, Pyramidal cell-interneuron circuit architecture and dynamics in hippocampal networks. *Neuron* **96**, 505-520 e507 (2017).

29. M. Wang, W. Deng, Deep visual domain adaptation: A survey. *Neurocomputing* **312**, 135-153 (2018).

30. S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. W. Vaughan, A theory of learning from different domains. *Mach Learn* **79**, 151-175 (2010).

31. S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems* **19**, 137–144 (2006).

32. A. M. Aertsen, G. L. Gerstein, Evaluation of neuronal connectivity: sensitivity of cross-correlation. *Brain Res* **340**, 341-354 (1985).

33. S. Ostojic, N. Brunel, V. Hakim, How connectivity, background activity, and synaptic properties shape the cross-correlation between spike trains. *J Neurosci* **29**, 10234-10253 (2009).

34. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors. *Nature* **323**, 533-536 (1986).

35. D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* **21**, 1-13 (2020).

36. M. Pospischil, M. Toledo-Rodriguez, C. Monier, Z. Piwkowska, T. Bal, Y. Fregnac, H. Markram, A. Destexhe, Minimal Hodgkin-Huxley type models for different classes of cortical and thalamic neurons. *Biol Cybern* **99**, 427-441 (2008).

37. D. Greenberg, M. Nonnenmacher, J. Macke, Automatic Posterior Transformation for Likelihood-Free Inference. *Proceedings of the 36th International Conference on Machine Learning* **97**, 2404--2414 (2019).

38. T. A. Jarrell, Y. Wang, A. E. Bloniarz, C. A. Brittin, M. Xu, J. N. Thomson, D. G. Albertson, D. H. Hall, S. W. Emmons, The connectome of a decision-making neural network. *Science* **337**, 437-444 (2012).

39. K. L. Briggman, M. Helmstaedter, W. Denk, Wiring specificity in the direction-selectivity circuit of the retina. *Nature* **471**, 183-188 (2011).

40. R. Ben-Shalom, J. Balewski, A. Siththaranjan, V. Baratham, H. Kyoung, K. G. Kim, K. J. Bender, K. E. Bouchard, Inferring neuronal ionic conductances from membrane potentials using CNNs. *bioRxiv*, 727974 (2019).

41. J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. (IEEE, 2017), pp. 23-30.

42. C. M. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, J. Hodgins, Next-generation deep learning based on simulators and synthetic data. *Trends Cogn Sci* **26**, 174-187 (2022).

43. X. B. Peng, M. Andrychowicz, W. Zaremba, P. Abbeel, in *2018 IEEE international conference on robotics and automation (ICRA)*. (IEEE, 2018), pp. 3803-3810.

44. Y. Gal, Z. Ghahramani, in *international conference on machine learning*. (PMLR, 2016), pp. 1050-1059.

45. B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* **30**, (2017).

46. M. Schrimpf, J. Kubilius, M. J. Lee, N. A. Ratan Murty, R. Ajemian, J. J. DiCarlo, Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron* **108**, 413-423 (2020).

47. D. Durstewitz, G. Koppe, M. I. Thurm, Reconstructing computational system dynamics from neural data with recurrent neural networks. *Nat Rev Neurosci* **24**, 693-710 (2023).

48. Y. N. Billeh, B. Cai, S. L. Gratiy, K. Dai, R. Iyer, N. W. Gouwens, R. Abbasi-Asl, X. Jia, J. H. Siegle, S. R. Olsen, C. Koch, S. Mihalas, A. Arkhipov, Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. *Neuron* **106**, 388-403 e318 (2020).

49. H. Markram, E. Muller, S. Ramaswamy, M. W. Reimann, M. Abdellah, C. A. Sanchez, A. Ailamaki, L. Alonso-Nanclares, N. Antille, S. Arsever, G. A. Kahou, T. K. Berger, A. Bilgili, N. Buncic, A. Chalimourda, G. Chindemi, J. D. Courcol, F. Delalondre, V. Delattre, S. Druckmann, R. Dumusc, J. Dynes, S. Eilemann, E. Gal, M. E. Gevaert, J. P. Ghobril, A. Gidon, J. W. Graham, A. Gupta, V. Haenel, E. Hay, T. Heinis, J. B. Hernando, M. Hines, L. Kanari, D. Keller, J. Kenyon, G. Khazen, Y. Kim, J. G. King, Z. Kisvarday, P. Kumbhar, S. Lasserre, J. V. Le Be, B. R. Magalhaes, A. Merchan-Perez, J. Meystre, B. R. Morrice, J. Muller, A. Munoz-Cespedes, S. Muralidhar, K. Muthurasa, D. Nachbaur, T. H. Newton, M. Nolte, A. Ovcharenko, J. Palacios, L. Pastor, R. Perin, R. Ranjan, I. Riachi, J. R. Rodriguez, J. L. Riquelme, C. Rossert, K. Sfyrakis, Y. Shi, J. C. Shillcock, G. Silberberg, R. Silva, F. Tauheed, M. Telefont, M. Toledo-Rodriguez, T. Trankler, W. Van Geit, J. V. Diaz, R. Walker, Y. Wang, S. M. Zaninetta, J. DeFelipe, S. L. Hill, I. Segev, F. Schurmann, Reconstruction and simulation of neocortical microcircuitry. *Cell* **163**, 456-492 (2015).

50. W. Qian, J. A. Zavatone-Veth, B. S. Ruben, C. Pehlevan, Partial observation can induce mechanistic mismatches in data-constrained models of neural dynamics. *bioRxiv*, 2024.2005. 2024.595741 (2024).

51. A. A. Prinz, D. Bucher, E. Marder, Similar network activity from disparate circuit parameters. *Nat Neurosci* **7**, 1345-1352 (2004).

52. M. G. Metzen, M. J. Chacron, Stimulus background influences phase invariant coding by correlated neural activity. *Elife* **6**, (2017).

53. E. Marder, M. L. Goeritz, A. G. Otopalik, Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms. *Curr Opin Neurobiol* **31**, 156-163 (2015).

54. G. J. Gutierrez, T. O'Leary, E. Marder, Multiple mechanisms switch an electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators. *Neuron* **77**, 845-858 (2013).

55. L. M. Alonso, E. Marder, Visualization of currents in neural models with similar behavior and different conductance densities. *Elife* **8**, (2019).

56. E. Marder, J. M. Goaillard, Variability, compensation and homeostasis in neuron and network function. *Nat Rev Neurosci* **7**, 563-574 (2006).

57. A. E. Urai, B. Doiron, A. M. Leifer, A. K. Churchland, Large-scale neural recordings call for new insights to link brain and behavior. *Nat Neurosci* **25**, 11-19 (2022).

58. M. Jazayeri, S. Ostojic, Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Curr Opin Neurobiol* **70**, 113-120 (2021).

59. S. Chung, L. F. Abbott, Neural population geometry: An approach for understanding biological and artificial neural networks. *Curr Opin Neurobiol* **70**, 137-144 (2021).

60. N. Sapoval, A. Aghazadeh, M. G. Nute, D. A. Antunes, A. Balaji, R. Baraniuk, C. J. Barberan, R. Dannenfelser, C. Dun, M. Edrisi, R. A. L. Elworth, B. Kille, A. Kyrillidis, L. Nakhleh, C. R. Wolfe, Z. Yan, V. Yao, T. J. Treangen, Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun* **13**, 1728 (2022).

61. L. A. Bolanos, D. Xiao, N. L. Ford, J. M. LeDue, P. K. Gupta, C. Doebeli, H. Hu, H. Rhodin, T. H. Murphy, A three-dimensional virtual mouse generates synthetic training data for behavioral analysis. *Nat Methods* **18**, 378-381 (2021).

62. C. Wan, D. T. Jones, Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. *Nat Mach Intell* **2**, 540-550 (2020).

63. R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, F. Mahmood, Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* **5**, 493-497 (2021).

64. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, (2014).

Acknowledgements

Funding

This work is supported by National Key R&D Program of China (N0. 2020AAA0130400).

Author contributions

Conceptualization and methodology: KS, KD

Investigation and analysis: KS, SZ, MB, PQ, LY

Discussion: KS, SZ, MB, PQ, LY, XL, LH, YZ, LM, KD

Supervision: KD, LM

Writing—original draft: KS, SZ, MB, LM, KD

Competing interests

Authors declare that they have no competing interests.

Data and materials availability

The source code of our framework in this work will be available upon acceptance. All data are available in the main text or the supplementary materials.

Figures

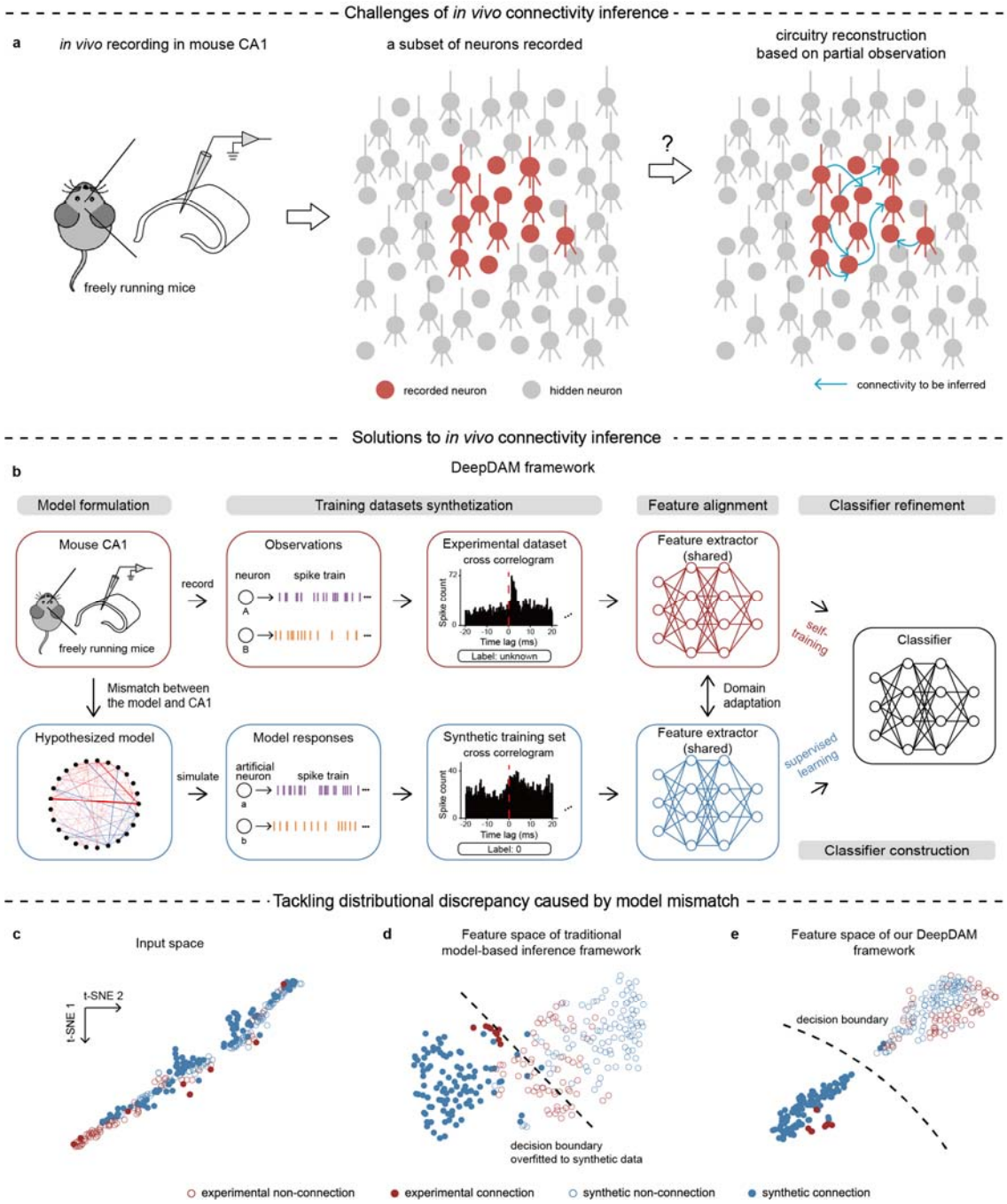

Fig. 1. DeepDAM inference framework for monosynaptic connectivity inference in CA1 networks. (**a**) Challenges associated with inferring connectivity, including incomplete observations and strong recurrent connections. (**b**) DeepDAM framework. The gray boxes outlined the five major steps of the framework. Note that both the feature extractor and the classifier are shared between synthetic and experimental data. (**c**) Distributional discrepancy of experimental and synthetic data

caused by model mismatch in the model construction step. (**d**) Synthetic and experimental feature distributions after training the traitional model-based inference framework. (**e**) Synthetic and experimental feature distributions after training our DeepDAM framework.

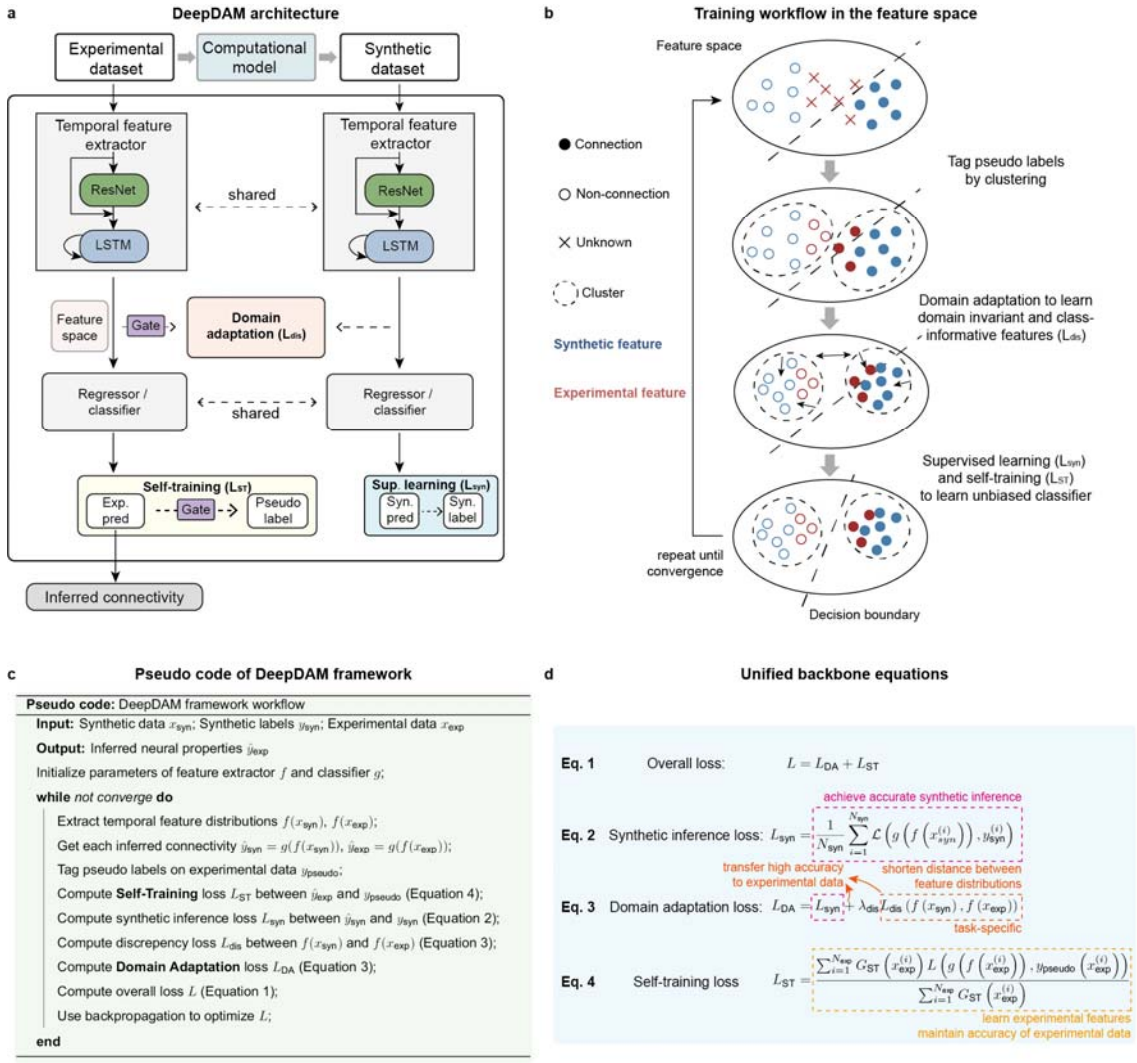

Fig. 2. DeepDAM framework architecture. (**a**) Workflow of the DeepDAM framework. The temporal feature extractor and regressor/classifier are shared between synthetic data and experimental data. Firstly, temporal features of synthetic data and experimental data are extracted by the shared temporal feature extractor. The output space from the temporal feature extractor is the feature space. Secondly, the shared regressor or classifier processes the temporal features and outputs the inferred neural properties. Sup. learning stands for supervised learning. (**b**) The training workflow of our inference framework. Each dot or cross represented a sample feature. **(c)** Pseudo code. The equation IDs correspond to the IDs in the main text and **d**. **(d)** The function of each unified backbone equation.

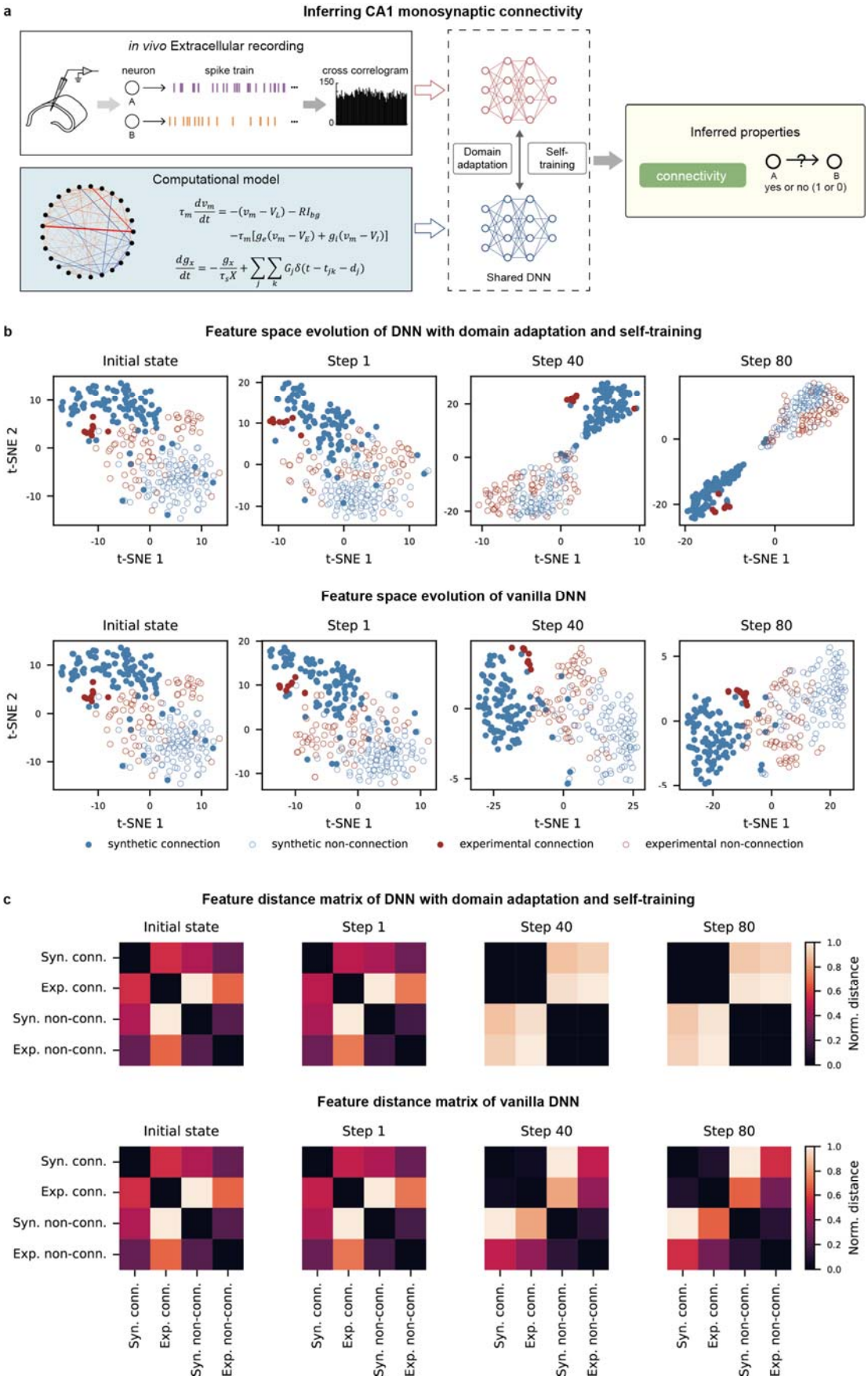
a
Inferring CA1 monosynaptic connectivity
in vivo Extracellular recording
neuron
spike train
cross correlogram
Computational model
Domain adaptation
Self-training
Shared DNN
Inferred properties
connectivity
yes or no (1 or 0)
b
Feature space evolution of DNN with domain adaptation and self-training
Initial state
Step 1
Step 40
Step 80
t-SNE 1
t-SNE 2
Feature space evolution of vanilla DNN
synthetic connection
synthetic non-connection
experimental connection
experimental non-connection
c
Feature distance matrix of DNN with domain adaptation and self-training
Feature distance matrix of vanilla DNN
Syn. conn.
Exp. conn.
Syn. non-conn.
Exp. non-conn.
Norm. distance

Fig. 3. Training workflow of our inference framework. (**a**) The experimental data and computational model for inferring the existence of monosynaptic connectivity between pairs of neurons. (**b**) Evolution of feature representations across the training process for both our inference method and the vanilla deep learning method without self-training and domain adaptation. The same fifty randomly selected synthetic connection and non-connection samples are shown in each subplot. (**c**) The evolution of distance matrix of synthetic and experimental features across the training process for both our inference method and the vanilla deep learning method without self-training and domain adaptation. Distance between two groups was calculated as the average Euclidean distance between each pair of features.

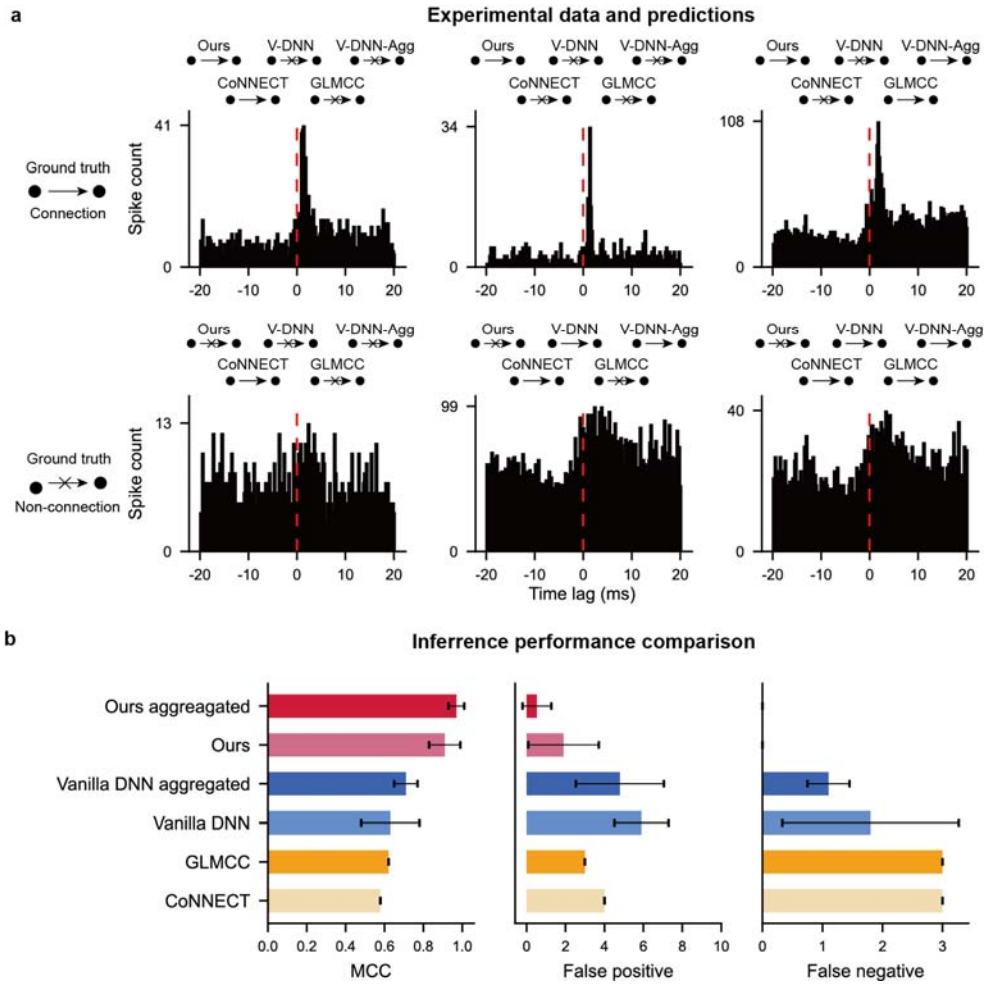

Fig. 4. Performance comparison on experimental dataset with ground-truth labels. (**a**) Example experimental CCGs and predictions by different methods. Ours: our aggregated version. V-DNN: the vanilla deep learning method. V-DNN-Agg: the aggregated vanilla deep learning method. CCGs were plotted from [-20, 20] ms based on the neuron spike timing with the time bin of 0.2 ms. (**b**) Performance comparison between 6 different methods. Error bar: standard deviation across 10 random seeds.

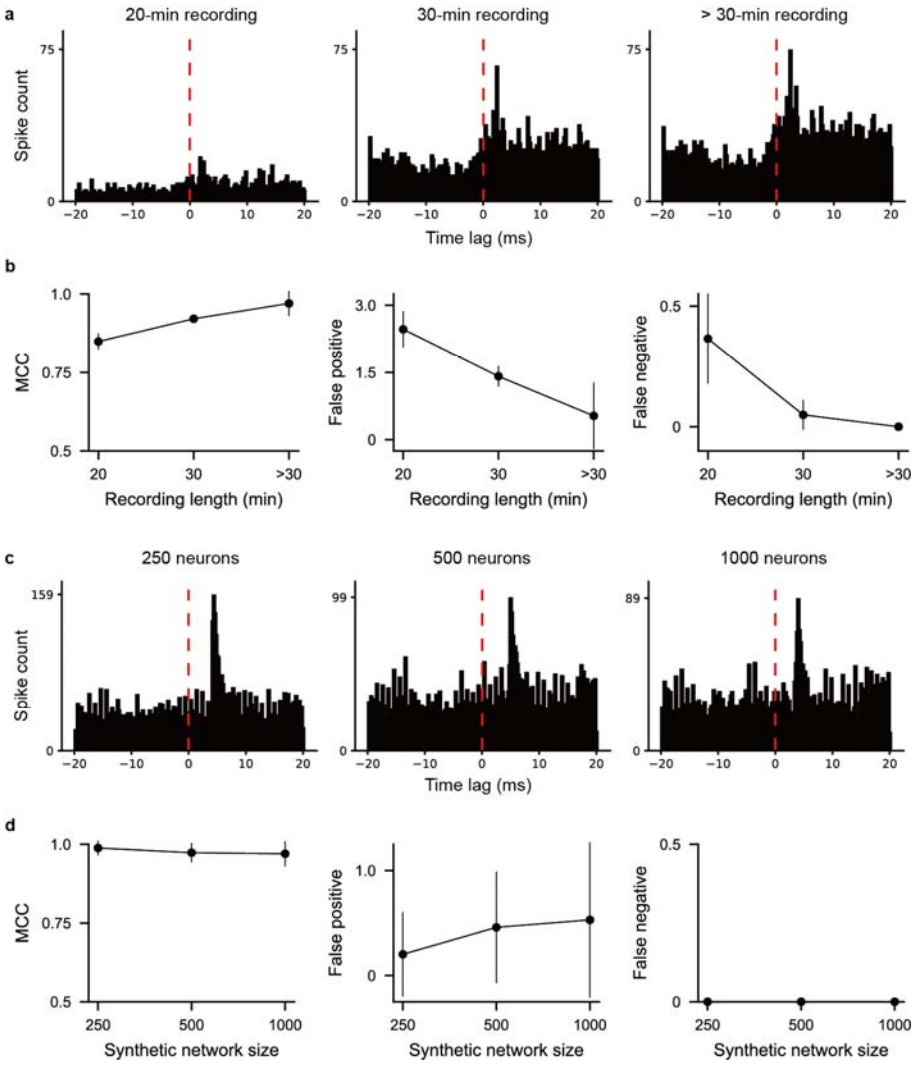

Fig. 5. Performance stability across different experimental conditions and computational model structures. (**a**) Experimental CCG of the same pair of neurons at different recording lengths. (**b**) MCC, false positives, false negatives of our aggregated method at different recording lengths. Error bar: standard deviation across 10 random seeds. (**c**) Synthetic CCGs generated from MAT neural networks of different sizes. (**d**) MCC, false positives, false negatives of our aggregated method at different sizes of MAT neural networks. Error bar: standard deviation across 10 random seeds.

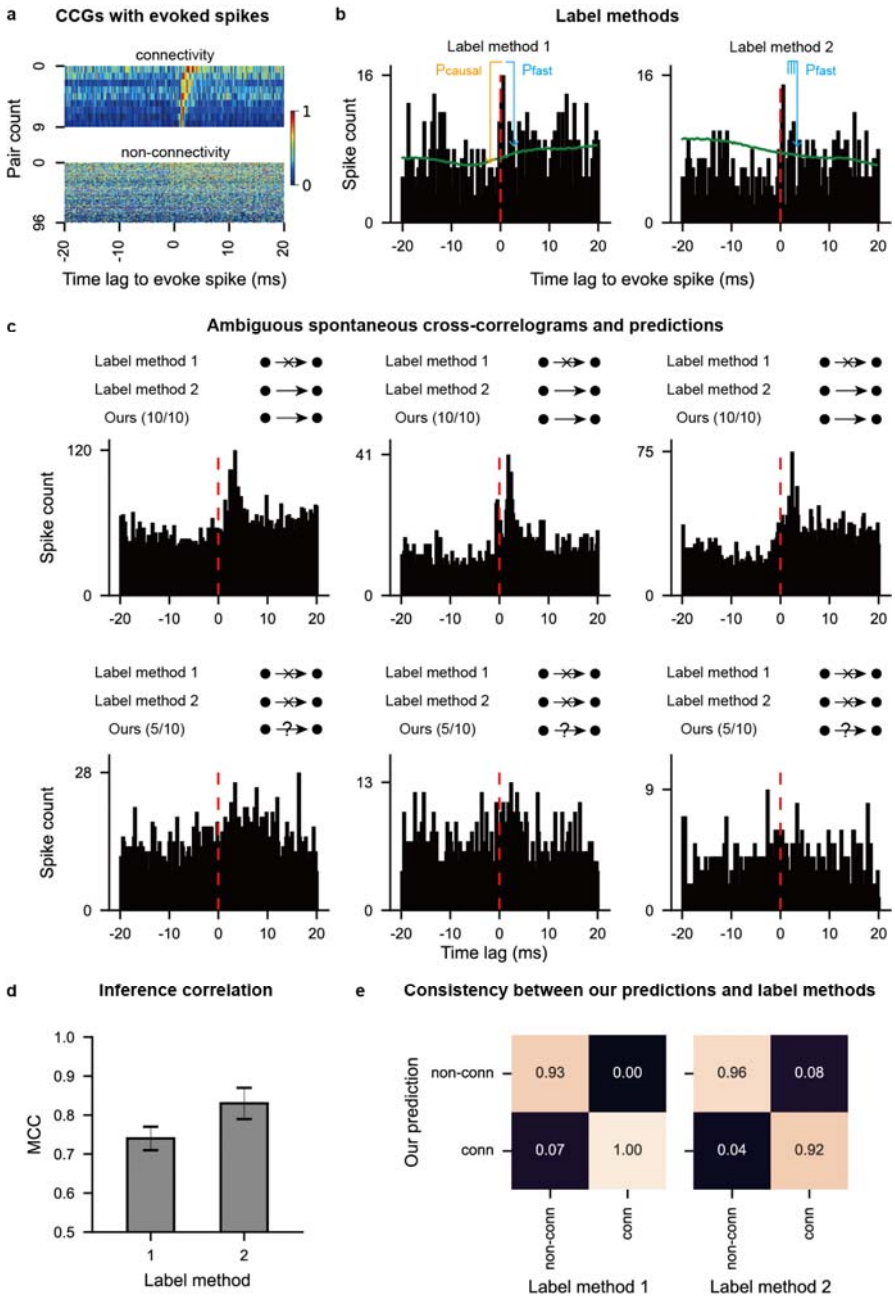

Fig. 6. Consistency between the prediction of our aggregated method and two label methods. (**a**) All CCGs with evoked spikes in the large dataset. For labeling, the CCGs were calculated based on 'presynaptic' spikes that were evoked via juxtacellular stimulation. Spontaneous spikes of the 'presynaptic' neuron were discarded. The label is given by label method 1. Color bar denotes the spike count, normalized by the maximal value of all CCGs. (**b**) Illustration of two label methods. The green line represents the slowly co-modulated baseline (SI Appendix). Briefly, label method 1 tests the lowest p-value of time bins in both directions (p_{causal} and p_{fast}) while label method 2 tests the p-value of three consecutive time bins in one direction (SI Appendix). (**c**) Experimental CCGs and

predictions of our aggregated method and two label methods. For inference, CCGs were calculated using spontaneous spikes of the 'presynaptic' neuron, while evoked spikes and the corresponding stimulus window was discarded from the CCGs. The numbers in the parentheses indicate the count of random seeds out of 10 that resulted in the same inference outcome. (**d**) MCC between our aggregated method and two label methods on the amibigous dataset. (**e**) Consistency matrix between our aggregated method and two ground-truth tagging methods on the amibigous dataset. For example, the upper-left grid in the left panel represented the percentage of experimental CCGs in the dataset that both our aggregated method and label method 1 inferred as non-connections.