Brain-wide oscillatory network selectively encodes aggression

Yael S. Grossman[2,*], Austin Talbot[2,7*], Neil M. Gallagher[2,3], Gwenaëlle E. Thomas[3], Alexandra J. Fink[6], Kathryn K. Walder-Christensen[2] , Scott Russo[6], David E. Carlson[5,8,†], Kafui Dzirasa[1,2,3,4,9,†].

[1]Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA; [2]Dept. of Psychiatry and Behavioral Sciences, [3]Dept. of Neurobiology, [4] Dept. of Neurosurgery, [5]Dept. of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina 27710, USA; [6]Fishberg Department of Neuroscience and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York, *10029* USA; [7]Department of Statistical Science, [8]Dept. of Civil and Environmental Engineering, [9]Dept. of Biomedical Engineering, Duke University, Durham North Carolina 27708, USA;

*These authors contributed equally

†Senior authors; Contributed equally

Correspondence should be sent to:

Kafui Dzirasa, M.D. Ph.D.
Dept. of Psychiatry and Behavioral Sciences
Duke University Medical Center
Durham, NC 27710, USA
Email: kafui.dzirasa@duke.edu
Twitter: @KafuiDzirasa

or

David E. Carlson, PhD
Department of Biostatistics and Bioinformatics
Department of Civil and Environmental Engineering
Duke University
Durham, NC 27708, USA
Email: david.carlson@duke.edu

Summary

Social aggression is an innate behavior that can aid an organism in securing access to resources. Aggression can also disrupt group function and reduce survival under conditions of behavioral pathology. Since many brain regions contribute to multiple social behaviors, expanded knowledge of how the brain encodes distinct social states would enable the development of interventions that suppress aggression, while leaving other social behaviors intact. Here we show that murine aggression is optimally encoded by a brain-wide network. This network is organized by prominent theta (4-11Hz) and beta (14-30Hz) oscillations, leading from orbital frontal cortex and medial dorsal thalamus, and converging on ventral medial hypothalamus and medial amygdala. Activity in this network is coupled to brain-wide cellular firing, and the network is conserved in new mice from multiple genetic backgrounds, and in multiple contexts. Finally, we develop a closed loop stimulation protocol based on network activity levels. This protocol suppressed aggression, but not pro-social behavior. Thus, we defined a causal brain-wide network that selectively encodes aggressive behavior across mice and established a new approach for state specific control of affective behavior.

Keywords

aggression; social behavior; electome network; closed loop stimulation; machine learning

Introduction

Social behavior reflects the integration of sensory information with internal affective states. Many subcortical brain regions contribute to aggressive behavior in mammals including lateral septum (LSN) (Brayley and Albert, 1977; Wong et al., 2016), nucleus accumbens (Couppis and Kennedy, 2008; Lee et al., 1983), lateral habenula (Flanigan et al., 2017; Golden et al., 2016), the ventrolateral portion of ventromedial hypothalamus (Falkner et al., 2014; Falkner et al., 2016; Hashikawa et al., 2016; Hashikawa et al., 2017; Lin et al., 2011; Takahashi et al., 2017), and medial amygdala (Haller, 2018; Shibata et al., 1982). Prefrontal cortex stimulation has been shown to mitigate aggressive behavior in both humans (Choy et al., 2018; Dambacher et al., 2015) and rodents (Takahashi et al., 2014), implicating cortical regions in regulating aggression. Finally, sensory regions, such as those underlying olfaction, also contribute to aggression (Jones and Nowell, 1973).

To appropriately regulate aggressive behavior, the brain must integrate information across these and other cortical and subcortical regions. Moreover, since many of these regions also regulate non-aggressive social behaviors (Hashikawa *et al.*, 2016; Lischinsky and Lin, 2020), the brain must ultimately utilize overlapping sources of information to segregate aggression from other social behavioral states. While this points to the existence of a broader brain-network level process underlying aggression, such a mechanism that generalizes across animals remains unknown. Addressing this question is of major importance as 1) mammals regularly select from a repertoire of social behavioral states based on external sensory cues to ensure their survival (Chen and Hong, 2018; de la Zerda et al., 2022), and 2) a range of psychiatric disorders are broadly marked by a failure to appropriately match behavior with evolving social contexts (APA, 2000).

To address this question, we implanted mice with microwire electrodes across eleven brain regions implicated in regulating complex social behavior. We then recorded electrical activity from these brain regions, concurrently, as mice engaged in social encounters that induced aggressive and pro-social behavior. After confirming that statistical models based on single brain regions could independently differentiate aggressive and pro-social behaviors, we asked

whether this code was also reflected at a network level, where millisecond timescale information was integrated across all the brain areas. For this analysis, we used a cutting-edge machine learning approach that modeled variations in natural patterns of activity within and between implanted brain regions across seconds of time (Talbot et al., 2020). This approach also tuned the model to optimally predict aggressive vs. pro-social behavior.

We then applied multiple levels of validation of increasing generalization to our single region and network level models (Vu et al., 2018). 1) We tested whether the models generalized to data collected from time periods that had not been used to train our model (i.e., hold out sessions). 2) We tested whether our models generalized to new animals, and 3) we tested an orthogonal behavioral context associated with aggression. Only the model that was based solely on activity from ventral hippocampus and the network model survived this level of validation. Critically, the network model was substantially more predictive than the model based on independent activity from ventral hippocampus. 4) Finally, we subjected the network to two high levels of experimental validation, each based on causal cellular manipulations, that bidirectionally impacted network activity and behavior. Specifically, we developed a closed loop optogenetic stimulation approach that detected when the brain transitioned into an aggressive state, on a second-by-second basis, and stimulated the brain into a more pro-social state. We also employed a designer receptor exclusively activated by designer drug (DREADD) based approach to selectively activate cells in ventral medial hypothalamus in a manner that has been shown to induce aggression. 5) As this latter manipulation was also performed in mice on a different genetic background, we further confirmed the generalizability of our network-level findings. Thus, we firmly establish a network-level mechanism whereby the brain selectively encodes aggression in a manner that generalizes across context and individuals.

Results

Only ventral hippocampus selectively encodes aggression across mice and contexts

To uncover neural mechanisms that could selectively be used to suppress aggressive behavior, we initially focused on the medial prefrontal cortex (i.e., prelimbic and infralimbic cortex in mice). This brain region had been implicated in social behaviors (Levy et al., 2019; Murugan et al., 2017; Padilla-Coreano et al., 2022), and prior work had also shown that direct stimulation of medial prefrontal cortex was sufficient to suppress aggressive behavior (Takahashi *et al.*, 2014). Mirroring this prior work, we set out to optogentically stimulate the infralimbic of CD1 mice. We chose male CD1 retired breeders for our experiments given their propensity to exhibit aggressive behavior characterized by tail rattling, biting, boxing, or tussling when exposed to male C57BL6/J (C57) mice (Miczek et al., 2001). Following surgical recovery, CD1 mice were allowed to freely interact with an intact male C57 mouse. We also allowed mice to interact with a castrated male C57 mouse, and a female C57 mouse. Consistent with previous reports (Lin *et al.*, 2011), exposure to the intact male, but not the castrated male or intact female mouse, induced attack behavior from a majority of CD1 mice. We performed our optogenetic stimulation experiments during these social encounters using a protocol modeled after prior work (473nm, 5mW, 20Hz, 3ms pulse width) (Takahashi *et al.*, 2014). Control experiments were performed using yellow light stimulation (593.5nm, 5mW, 20Hz, 3ms pulse width), which does not activate channel rhodopsin 2. Though optogenetic stimulation successfully suppressed aggression in the CD1 mice [N=8; t_7=3.428, p=0.0055 using a paired t-test, significance determined by a Benjamini-Hochberg false discovery rate (FDR) correction], we found that this manipulation also suppressed the total time mice engaged in pro-social interactions across the three conditions (t_7=3.10, p=0.017 using paired t-test). Indeed, we observed decreased pro-social interactions with castrated C57 male mice and female C57 mice (N=8; t_7=2.733, P=0.0146; t_7=3.647, P=0.0041 for castrated male and female mouse encounters, respectively using paired t-test, significance determined by FDR correction; Fig. 1A). Thus, infralimbic cortex stimulation was unable to selectively suppress aggression.

To probe for candidate brain regions that may exhibit such selectivity, we implanted CD1 mice across multiple cortical and subcortical brain regions known to contribute to social behavior, including infralimbic cortex, orbitofrontal cortex, prelimbic cortex, lateral septum, nucleus accumbens, lateral habenula, mediodorsal thalamus, ventromedial hypothalamus, medial

amygdala, primary visual cortex, and ventral hippocampus. Following surgical recovery, we recorded neural activity while the CD1 mice freely interacted with an intact male C57 mouse, a castrated male C57 mouse, and a female C57 mouse for 300 seconds. We repeated these encounters over six sessions, yielding a total of 2100 seconds of neural data and behavior for each exposure (Fig. 1B, see also Supplemental Fig. S1). Data was also acquired while the CD1 mice were isolated in their home cage.

We first verified that all implanted brain regions encoded social behavior. To explore whether there was a generalized activity pattern in individual regions that encoded social behavior, we designed a series of linear models based on local field potential (LFP) oscillatory power in frequencies from 1-56Hz. LFPs reflect the activity of populations of neurons, and these signals can be consistently sampled across mice. Each model was trained using observations pooled from twenty CD1 mice to separate periods where mice were socially isolated from periods where they were engaged in the social behavior (aggression or pro-social social behavior). We then tested the accuracy of the models in new CD1 mice (N=9). We observed high accuracy in decoding social periods for each implanted brain area ($p<0.05$ using Wilcoxon rank-sum, , significance determined by FDR correction for 11 comparisons, see Fig. 1C).

After verifying that all the regions we implanted broadly encoded social behavior, we probed whether they selectively encoded different types of social behavior. We designed a new series of linear models, again based on LFP power. These models were built using observations pooled from the same twenty CD1 mice to separate one of the three social behavioral states from the other two (e.g., aggression vs. female and castrated male non-aggressive interactions). We then tested the accuracy of the models in the same nine hold-out mice. Thus, we trained and tested the model's generalizability to decode each social state from the other two for each of the 11 implanted brain areas (i.e., 33 models). Using this approach, we found that four of the brain region-based statistical models decoded aggression versus the pro-social states: primary visual cortex, lateral habenula, ventral hippocampus, and medial dorsal thalamus. Lateral habenula successfully decoded the female interaction from the other two social states as well ($P<0.05$ using Wilcoxon rank-sum test, significance determined by FDR correction for 33 comparisons, see Fig. 1C). None of the other implanted brain regions showed selectivity. Thus, primary visual

cortex, lateral habenula, ventral hippocampus, and medial dorsal thalamus were the only regions that encoded aggression vs. pro-social states in a manner that generalized across mice.

Our broad goal was to identify a neural mechanism that could be used to suppress aggression while leaving other social behaviors intact. Thus, we tested whether the putative aggression-selective codes we discovered for primary visual cortex, lateral habenula, ventral hippocampus, and medial dorsal thalamus generalized to another context associated with aggression. Specifically, urine from other male mice has been found to illicit aggressive and dominance behavior in CD1 males (Chen et al., 2020; Jones and Nowell, 1973; Lee and Wilson, 2012). As such, the most aggressive mice from the training and testing groups (N=8) were allowed to feely explore a clean inanimate object or an object covered in urine from another intact CD1 male mouse (seven sessions). We then tested whether the neural aggression-selective codes we identified using data from any of the four brain regions could distinguish periods where mice explored the clean objects from those where mice explored objects covered in urine. We found that only the ventral hippocampus model successfully decoded aggression in this new context (AUC=0.54±0.02, P=0.40; AUC=0.51±0.01, P=0.40; AUC=0.55±0.01, P<0.01; and AUC=0.53±0.02, P=0.08 for primary visual cortex, lateral habenula, ventral hippocampus, and medial dorsal thalamus, respectively, using one-tailed Wilcoxon rank-sum test, significance determined by FDR correction). Thus, taken together with our initial analyses, these findings established that ventral hippocampus independently encoded aggression in a manner that generalized across animals and contexts.

Aggression is optimally encoded at the network level

After determining that only ventral hippocampus independently and selectively encoded aggression, we established that a brain network could integrate information across all the implanted brain regions to encode the brain state. This network-level encoding mechanism generalized to multiple new contexts associated with aggression. Critically, the network also encoded aggression with a predictive efficacy that exceeded independent ventral hippocampus activity. To achieve this, we employed discriminative cross spectral factor analysis non-negative matrix factorization (dCSFA-NMF, Fig. 2A) (Talbot *et al.*, 2020). dCSFA-NMF utilizes supervised

machine learning to generate a statistical model that is both descriptive (integrates brain activity features across time) and predictive (discovers networks that distinguish between types of external behavior). The electrical functional connectomes networks (*Electome Networks*) generated from dCSFA-NMF integrate LFP power (oscillatory amplitude across 1-56 Hz; a correlate of cellular and synaptic activity within brain regions), LFP synchrony (how two regions' LFP frequencies synchronize across time; a correlate of brain circuit function), and LFP Granger synchrony (Granger causality testing; a correlate of directional transfer of information across a brain circuit). Furthermore, dCSFA-NMF generates electome network activity scores (an indicator of the strength of each network) at a temporal resolution consistent with the external behavior under observation (in this instance, a resolution of one second). The electome networks are designed to learn patterns that explain interpretable correlates of neural activity whose expression relate to measured behavior, facilitating an overall interpretable model (Vu *et al.*, 2018). Any given brain region can belong to multiple electome networks and each electome network may incorporate any number of brain regions. dCSFA-NMF thus integrates spatially distinct brain regions and circuits into electome networks that encode behavior.

We used data from the initial group of twenty CD1 mice to train our dCFSA-NMF model. We trained eight electome networks, which was chosen by a validation procedure to balance how much brain activity is explained by the model, how much the model predicts behavior, and the parsimoniousness of the model (see methods and Supplemental Fig. S2). To facilitate easy interpretation of the learned networks, one network was encouraged to preferentially encode periods of aggression versus the other social behavioral states (i.e., supervision, electome network #1; see Fig. 2A). We then validated this network using the set of nine holdout CD1 mice from our single area coding test analysis. Again, none of these mice were used to train the electome networks. Our findings showed that electome network #1 successfully discriminated between aggressive and non-aggressive social behavior in the test mice (N=9, Wilcoxon signed rank, p=0.0020, Fig. 2B). Thus, the supervised network generalized to new animals. When we compared the coding strength of electome network #1 to the model we developed from ventral hippocampus activity, we found that the network showed significantly stronger encoding (|AUC-0.5|=0.17±0.03 vs. 0.08±0.03, for electome network #1 and ventral hippocampus,

respectively; P= 0.004 using signed rank test, see Fig. 2C). This corresponded to 13 times more information in the network than ventral hippocampus (under an information theoretic framework -see methods). Thus, aggression was optimally encoded at the network level in the murine brain. Surprisingly, electome network #1 was negatively correlated with aggression. Closer examination the network's temporal dynamics showed that *activity sharply* decreased at the initiation of attack behavior (Fig. 3A). This network activity returned to baseline after the cessation of attack behavior. Interestingly, network activity was also lower in the intervals surrounding aggressive social interactions compared to pro-social interactions with castrated males or females (P<0.05 using Wilcoxon rank sum test, significance determined by FDR correction). Thus, electome network #1 (hereafter referred to as *EN-Aggression Inhibition - AggINH*) represented a network that putatively inhibited aggression when its activity was highest.

EN-Aggression Inhibition maps to interpretable biology

EN-AggINH was composed by prominent theta frequency activity (4-11 Hz) in MeA and beta frequency activity (14-30 Hz) in medial amygdala and prelimbic cortex (Fig. 3B). Prominent synchrony and directionality were also observed in the theta and beta frequency bands. Indeed, the network showed strong directionality that led from orbital frontal cortex and primary visual cortex, relayed through medial dorsal thalamus, and infralimbic cortex, and converged on medial amygdala and ventral hippocampus. Interestingly, local activity in ventral hippocampus did not contribute to the network (Fig. 3C, and Supplementary Fig. S3).

We verified that the activity of *EN-AggINH* truly reflects biological activity, by relating the electome network to neural firing, as in previous work (Hultman et al., 2018). To achieve this, we analyzed the correlation between cellular activity across the implanted brain regions and the activity of *EN-AggINH*, as cell activity is an undisputed measure of biological function. We then used a permutation test to rigorously test our findings (Fig. 4A). Of the 186 cells recorded, nine (4.8%) showed firing activity that was positively correlated with *EN-AggINH* and 25 (13.4%) showed activity that was negatively correlated (Fig. 4B). Overall, ~18% of cells showed a

relationship with *EN-AggINH*, far more than could be explained by chance. Thus, we confirmed that *EN-AggINH* activity reflects the dynamics of cellular activity across the brain.

EN-Aggression Inhibition generalizes to new biological contexts related to aggression

To further validate *EN-AggINH*, we established that activity in this network was modulated by orthogonal biological conditions that have been shown to induce or suppress aggressive behavior in mice. In several cases, we performed this analysis in new animals, which is considered the gold standard of model validation in machine learning by experimental manipulation (Roelofs et al., 2019). First, we probed activity of the network in the urine paradigm —in which the most aggressive mice from the training and testing groups (N=8 total) feely explored a clean inanimate object or an object covered in urine from another intact CD1 male mouse— that we had used to examine the extent to which our individual brain region-based models generalized. To test whether exposure to urine altered *EN-AggINH* activity in our experimental mice, we transformed LFP data recorded from these new sessions into our original network model (Fig. 5A). We found that exposure to the urine significantly decreased *EN-AggINH* activity (N=8, P=0.0078 using Wilcoxon signed-rank test, see Fig. 5B). Thus, these findings showed that the network model generalized to a new aggression context.

Next, we performed a high-threshold validation experiment that involved testing whether our network generalized to new mice on a different genetic background (Fig. 5C). This approach also used a validated cellular manipulation that causally induces aggression, and it was performed under a behavioral condition that would otherwise not yield aggressive attack behavior (i.e., we used female social partners). Specifically, we expressed an excitatory DREADD (AAV-hsyn-DIO-hM3Dq) in the ESR1+ cells of ventromedial hypothalamus, since it has been shown that direct excitation of these cells induces aggressive behavior towards female mice (Deonaraine et al., 2020; Li et al., 2022; Takahashi *et al.*, 2017). Experiments were performed in the male F1 offspring of female CD1 strain mice crossed with ESR1-Cre male mice on a C57 strain background. Subsequently, we implanted the mice with recording electrodes to target the same brain regions as our initial experiment used to train the network model. Following recovery, we performed behavioral and neural recording when mice were exposed to a female

mouse. The experimental mice were either treated with saline or CNO (Clozapine N-oxide, which activates the excitatory DREADD), in a pseudorandomized order, across repeated test sessions. After confirming *EN-AggINH* encoded aggression towards intact male mice in the mice treated with saline (N=7, p=0.0078 using one-tailed Wilcoxon sign rank, Fig. 5D), we explored whether treatment with CNO impacted behavior and *EN-AggINH*. As anticipated, treatment with CNO induced aggressive behavior towards the female mice (N=8; T_7= 5.398 and P= 0.0005; T_7= 8.615 and P=<0.0001 for attack latency and attack number, respectively, using one-tailed paired t-test; see Fig. 5E). Treatment with CNO also suppressed *EN-AggINH* activity during the female encounter (N=8, P=0.0039 using one-tailed Wilcoxon sign-rank; see Fig. 5F), confirming that a second experimental manipulation that drives aggression successfully suppressed *EN-AggINH*. Taken together, these findings showed that our network model generalized to a second aggression context across sex, and to different genetic backgrounds.

Temporally precise manipulation of EN-Aggression Inhibition selectively *suppresses aggression*

We developed and employed a closed loop state modulator to regulate *EN-AggINH* and establish its causality. Since our initial single area decoding analysis and optogenetic manipulation study suggested that networks underlying multiple social behaviors share circuit components with *EN-AggINH*, we recognized a potential confound to establishing the causality of *EN-AggINH*. Specifically, even if causally manipulating a circuit within *EN-AggINH* impacted aggression, these effects may be secondary to a mechanism that impacts social behavior more broadly. To address this possibility, we implemented an approach that had previously been utilized to establish the causality of another neural signature in behavior (Girardeau et al., 2009). Namely, we chose to selectively stimulate *EN-AggINH* activity when the network was naturally suppressed in the brain and monitor the impact of this manipulation on behavior (Fig. 6A).

Since our initial results demonstrated that direct activation of infralimbic cortex suppressed aggressive behavior, we verified that stimulation of this region increased *EN-AggINH* activity as well (N=9, P<0.01 for comparison of *EN-AggINH* activity one second after blue vs. yellow stimulation, using one-tailed signed-rank test, see Fig. 6D). We then built a closed loop system

that estimated the activity of *EN-AggINH* in real time. Here, we employed a new network encoded solely based on power and coherence measures (i.e., a reduced network, Fig. 6B). While this approach lacked the interpretive power of dCSFA-NMF, it enabled us to predict aggression states in real time (Fig. 6C). In principle, when the activity of *EN-AggINH* fell below an established threshold (signaling the onset of aggressive behavior), this approach would deliver a one second light stimulation (5mW, 20Hz, 3ms pulse width) to infralimbic cortex. To verify that this real-time estimation system worked as designed, we tested whether light stimulation was triggered by a decrease in *EN-AggINH* activity. Indeed, network activity was significantly lower one second prior to stimulation than it was two seconds prior to stimulation, demonstrating that our approach successfully identified when the *EN-AggINH* activity decreased below the threshold that signaled the onset of aggression (N=9; $P<0.005$ for within-subject comparison of *EN-AggINH* activity 1 vs 2 seconds prior to yellow light stimulation using one-tailed signed-rank test, Fig. 6D).

Next, we tested whether increasing *EN-AggINH* as the brain transitioned into a putative aggressive state would suppress aggressive behavior in new CD1 mice (N=9 mice that were not used to train the initial model). We also hypothesized that this approach would not alter the pro-social behavior of the implanted CD1 mice during interactions with a castrated male or a female mouse, since *EN-AggINH* was not naturally suppressed during the pro-social interactions (see Fig. 6E). Using this approach, we found that closed loop stimulation of infralimbic cortex based on *EN-AggINH* activity significantly suppressed aggressive behavior towards the C57 intact males (N=9; $t_8=6.1$, $P=0.0003$ comparing blue vs. yellow light stimulation using two-tailed paired t-test, significance determined by FDR correction). No differences in pro-social behavior were observed from the stimulated CD1 mice during exposure to female mice ($t_8=0.74$, $P=0.48$ using two-tailed paired t-test, significance determined by FDR correction). Though closed loop stimulation tended to reduce pro-social behavior towards the castrated males (N=9; $t_8=2.67$, $P=0.03$ using two-tailed paired t-test; not significant as determined by FDR correction); it also tended to increase pro-social behavior towards the intact male mice. As such no differences in total time spent engaging in pro-social behavior were exhibited by CD1 mice in response to blue

light stimulation (t_8=0.81, P=0.44 using paired t-test). Taken together, closed loop stimulation preferentially impacted aggression vs. pro-social behavior.

To ensure that this suppression of aggressive behavior was due to the synchronization of the light stimulation and endogenous *EN-AggINH* activity, and not simply stimulation at the dynamic pattern inherent to *EN-AggINH* activity, we performed an additional control experiment. Here, we used the stimulus patterns from our closed loop experiments to drive stimulation in a new group of animals. Thus, for these sessions, infralimbic stimulation occurred in a manner that mirrored our closed loop stimulation experiments, except it was not synchronized to transitions into the aggressive state (i.e., open loop – nonsynchronous; Fig. 6G). Strikingly, this pattern of stimulation did not impact aggressive behavior (t_7=0.64, P=0.54 using paired t-test; see Fig. 6H). Together these results established that *EN-AggINH* activity was indeed causal, since manipulating activity of the network during brain state transitions selectively suppressed aggressive behavior.

Finally, since prior studies had also shown that medial prefrontal cortex stimulation induced hyperactivity (Anikeeva et al., 2011; Kumar et al., 2013), we tested whether the suppression of aggressive behavior that resulted from infralimbic cortex stimulation was also associated with behavioral hyperactivity. While we found that 20Hz fixed frequency stimulation of infralimbic cortex induced behavioral hyperactivity in the CD1 mice (N=10; T_9= 3.221 and P = 0.0052 for Velocity; T_9= 3.213 and P=0.0053 for distance traveled, using one-tailed paired t-test; Fig. 6I), closed loop stimulation of infralimbic cortex had no impact on gross locomotor behavior (N=9; T_8= 0.584 and P=0.29 for velocity; T_8= 0.86 and P= 0.21 for distance traveled, using one-tailed paired t-test; Fig. 6I). Thus, closed loop stimulation suppressed aggressive behavior, but it did not induce the off-target effects of impaired pro-social behavior and psychomotor activation that we observed with the classic 20Hz stimulation paradigm.

Discussion

Many brain regions have been implicated in complex social behavior such as aggression; nevertheless, the mechanisms whereby the brain integrates activity across these regions to appropriately match individual behavioral responses to evolving social context remains to be clarified. Here, we implanted recording wires into eleven cortical and subcortical brain structures implicated in regulating aggressive behavior. We then recorded local LFP electrical oscillations while mice were socially isolated and while they were exposed to a series of social encounters that induced pro-social behavior or aggression. With this approach, we hypothesized that we could discover a network pattern of electrical activity that differentiated aggressive states from pro-social states. Importantly, since LFPs reflect the activity of populations of neurons, we also hypothesized that this network pattern would generalize to encode aggression in new animals and contexts. This second level of testing was critical for establishing that the network we ultimately uncovered was mechanistically linked to aggression, rather than simply reflecting other features of the environment that accompanied aggressive social encounters in our testing context.

We initially explored whether we could distinguish periods of social isolation from periods where animals were engaged in social behavior. These analyses, all validated in new mice, revealed that all the regions we implanted were engaged during social behavior. This observation was consistent with causal manipulation studies which showed that direct stimulation of infralimbic cortex suppressed both aggressive and pro-social behaviors. Indeed, since prior literature suggests that ventral medial hypothalamus plays a role in both aggression and sexual behavior (Yang et al., 2013), it is likely that many of the aggression-related brain regions we identify here play a broader role in shaping multiple types of social behavior, rather than aggression alone.

When we tested whether we could decode aggression vs. pro-social behavior using independent brain regions, we found that activity in primary visual cortex, lateral habenula, medial dorsal thalamus, and ventral hippocampus each independently distinguished the two social brain states. Importantly, since different patterns of movement in both the experimental and partner mice (i.e., changes in the visual scene) were observed during aggressive vs. pro-social behaviors we were not surprised to discover that activity in visual cortex could be used to

differentiate the two states. Indeed, we would have anticipated that activity in brain regions responsible for audition and olfaction could be used to decode aggression vs. pro-social behavioral intervals as well. To address this limitation, we designed a validation context in which aggression was induced in the absence of a partner mouse. Specifically, we used an object covered in the urine of another CD1 mouse. We reasoned that this exposure would drive an aggressive brain state in a manner that was decoupled from the changes in the visual scene that accompanied aggressive encounters with an intact male partner mouse. We then tested our encoding models developed using behavior from social encounters on the data that we acquired during exposure to male urine. While ventral hippocampus generalized to this new context, activity from visual cortex, lateral habenula, and medial dorsal thalamus failed to encode the urine induced aggressive encounter. Thus, out of the eleven areas we tested, only ventral hippocampus encoded aggressive behavior in a manner that generalized to a new aggression context in new mice.

We also probed whether aggression was optimally encoded under a network framework using machine learning. Our approach integrated activity across all our implanted brain regions into a model that both predicted aggressive behavior and revealed how activity patterns naturally coupled together in the mouse brain. Importantly, this network model better encoded aggression compared to the model solely based on ventral hippocampal activity. The network model also generalized across mice and to multiple new contexts associated with aggression. Specifically, the electome network successfully encoded the aggressive state observed when mice were exposed to the object covered in male urine, and it encoded the aggressive state that was induced by direct exogenous stimulation of ventromedial hypothalamus. Critically, this latter causal manipulation study was performed in mice on a different genetic background, further highlighting the generalizability of the network mechanism.

The network was organized by prominent theta and beta synchrony between many of the brain regions recorded, including the primary visual cortex. While the model also contained local power in primary visual cortex, this feature was not strongly represented in the network. Rather, high primary visual cortex activity was observed in the other learned networks that showed less coupling to aggressive behavior. These findings suggest that while primary visual

cortex activity is indeed part of the broader aggression related network that we learned, the vast share of activity observed in primary visual cortex is not coupled to aggressive states vs. pro-social states (Supplemental Fig. S4). Critically, while we discovered *EN-AggINH* using a behavioral context that included visual information, *EN-AggINH* also generalized to the urine context in which aggression was not paired to the same visual context. These findings demonstrate its robustness. Nevertheless, as *EN-AggINH* was learned from a single experimental context, future experiments in which aggression networks are generated using multiple behavioral contexts, including those that do not include visual information, may help to further refine the prominence of circuits involving primary visual cortex in shaping aggressive behavior.

The observation that *EN-AggINH* activity was suppressed during aggression does not indicate that overall brain activity is suppressed during aggressive states. Rather, our findings argue that the aggressive state was optimally encoded by a network that decreases its activity relative to when mice are socially isolated or engaged pro-social behavior. Indeed, our data suggested that several common circuits were activated during aggressive and pro-social behavior. These common circuits need not be reflected in our network since our model was trained to differentiate aggressive states from pro-social states. Nevertheless, our discovery of a network that decreased its activity during aggression raises the intriguing hypothesis that this brain activity inhibits aggression during pro-social engagement. When activity in this inhibition network is suppressed, aggressive behavior emerges.

Interestingly acute stimulation of infralimbic cortex increased activity in the network. We exploited this observation to develop and employ a closed loop approach. This approach stimulated infralimbic cortex activity as network activity dropped below an established threshold, signaling the emergence of aggression. Our closed loop stimulation strategy suppressed aggressive behavior, whereas mice stimulated in a manner that was not timed to network activity (e.g., nonsynchronous) showed no change in aggression. Thus, increasing the activity of *EN-AggINH* suppressed aggression. Combined with the findings from our ventromedial hypothalamus DREADD experiment, these findings demonstrate that *EN-AggINH*

can be bidirectionally modulated using two causal cellular manipulations that regulate aggression.

Our approach represents a powerful framework for probing how manipulating cell type specific activity at critical network nodes impacts the activity of other circuits in the network. Specifically, we used machine learning and gold standard validation methods to identify/reduce a network parameter space defined by interconnected brain circuits. Though our optogenetic and DREADD manipulations occurred at two different network sites, both converged to impact network activity and behavior. It is likely that these independent manipulations impacted distinct sub circuits within *EN-AggINH.* Though the current study was underpowered to address this critical question, future work using this approach has the potential to uncover the principled manner whereby activity at one site of a network modulates the activity of interconnected circuits. Ultimately, this approach may also reveal the mechanism whereby convergent network-level activity emerges from interdependent, yet distinct, neural circuits.

Multiple neuropsychiatric disorders including mood disorders, psychotic disorders, neurodevelopmental disorders, and neurodegenerative disorders are associated with deficits in regulating social behavior, including aggression. While multiple pharmacological approaches have been instituted to suppress aggressive behavior towards self and others, many of these strategies act by sedating the individual and can disrupt aspects of pro-social function. Our discovery of a brain network that causally encodes aggression raises the potential for novel approaches to suppress aggressive behavior that spare pro-social behavior. Indeed, compared to a standard open loop stimulation protocol (20Hz stimulation) which suppressed both aggressive and pro-social behavior, our closed loop stimulation approach had no impact in total time mice spent engaged in pro-social behavior. Intriguingly, the classic open loop stimulation pattern (20Hz stimulation) induced behavioral hyperactivity in experimental mice, while our closed loop stimulation protocol did not. Thus, our findings also show that closed loop stimulation may limit off target behavioral effects that are induced by classic stimulation approaches.

The closed loop stimulation approach was developed using a neural-network based approximation technique that was substantially constrained relative to dCSFA-NMF. Specifically, the reduced encoder was developed solely using power and coherence features. This approach was necessary because the processing time to calculate granger directionality was prohibitive for real time implementations. As such, the reduced encoder sacrifices some predictive efficacy for speed. Nevertheless, we found that the reduced encoder was sufficient to identify the precise time windows where the brain transitioned into aggression, as marked by a decrease in *EN-AggINH* activity. This approach enabled us to regulate aggression while leaving other social behaviors intact. In the future, novel approaches may allow for further improvement in the precision of our real time stimulation approach. For example, future work could exploit convolutional neural networks to bypass the feature extraction step. These neural network encoders could be altered to predict both aggressive and pro-social states, such as the generalized social appetitive network that we recently discovered (Mague et al., 2022). By using both networks concurrently to actuate a closed loop system, it may be possible to further suppress aggressive behavior relative to pro-social behavior. Indeed, our current findings also pointed to a network that exhibits increase activity during aggressive behavior (Electome Network 6, see Fig 2B, and Supplementary Fig. S4). This network showed strong encoding in the nine hold out test mice (AUC=0.65±0.03, $P<5\times10^{-5}$ using Wilcoxon rank sum), and activity in the network also generalized to the experiment in which we causally manipulated ventromedial hypothalamus activity. Though the network failed to encode the urine paradigm, it is possible that the network contains activity that synergizes with *EN-AggINH* to encode aggressive social states more optimally. If future studies demonstrate this potential, imitation encoders for both Electome Network 6 and *EN-AggINH* could be integrated to further optimize closed loop approaches to selectively suppress aggression.

Overall, our findings establish a generalized network-level mechanisms whereby the brain selectively encodes the inhibition of aggressive behavior. Moreover, they highlight the exciting potential for state-specific neuromodulation to regulate affect.

Acknowledgements

We would like to thank Stephen Mague, Karim Abdelaal, and Ashleigh Rawls for comments on this work; and Timothy Nyangacha for technical support. This work was supported by WM Keck Foundation and Hope for Depression Research Foundation grants to KD; NIH grants R01MH120158 to KD, 1R01EB026937 to DEC, and 1R01MH125430 to S.R., DEC, and KD. A special thanks to Freeman Hrabowski, Robert and Jane Meyerhoff, and the Meyerhoff Scholarship Program.

Figure Legends

Figure 1. Brain regions independently encode aggression A) Direct stimulation of infralimbic cortex suppresses social behavior. Schematic of optogenetic stimulation (left) and social encounters utilized for testing (middle). Infralimbic cortex stimulation suppressed social behavior in all three contexts (*P<0.05). **B)** Schematic for electrical recordings. Figure shows targeted brain regions (left), and representative local field potentials (middle) recorded during repeated exposure to aggressive and pro-social contexts (right). **C)** Framework to test encoding of social states by individual brain regions (left). All implanted regions encoded social engagement; however, only four selectively encoded the aggressive social encounter (middle/right). Pink shading indicates P<0.05 with FDR correction.

Figure 2. Aggression is optimally encoded at the network level. A) Schematic of machine-learning model used to discover network encoding aggression. **B)** Encoding across eight learned networks. The supervised network (red, *EN-AggINH*) showed the strongest encoding. **C)** *EN-AggINH* showed stronger encoding that independent activity from ventral hippocampus (P= 0.004 using two-tailed Wilcoxon sign-rank test). Data shown as mean ± SEM.

Figure 3. Dynamics and biological components of *EN-Aggression Inhibition*. A) *EN-AggINH* activity at the onset and termination of social engagement in mice utilized to train the network model. Gray highlights behavioral windows utilized for training (left). Average activity across 3 seconds prior to/following social interactions (right). Network activity is shown normalized to activity observed during social isolation. **B)** Prominent oscillatory frequency bands composing *EN-AggINH* are highlighted for each brain region around the rim of the circle plot. Prominent synchrony measures are depicted by lines connecting brain regions through the center of the

circle. The plot is shown at relative spectral density 0.4. Theta (4-11 Hz) and beta (14-30 Hz) frequency components are highlighted in blue and yellow, respectively. **C)** Granger offset measures were used to quantify directionality within *EN-AggINH*. Prominent directionality was observed across the theta and beta frequency bands (shown at spectral density threshold of 0.4 and a directionality offset of 0.3). Histograms quantify the number of leading and lagging interactions between brain regions. **D)** Schematic depicting directionality within *EN-AggINH*.

Figure 4. *EN-Aggression Inhibition* activity is coupled to cellular activity. E-F) *EN-AggINH* maps to cellular activity. A) A representative cell showing firing activity that is negatively correlated with *EN-AggINH* activity. B) *EN-AggINH* activity correlated with cellular firing across the brain across the brain.

Figure 5. *EN-Aggression Inhibition* encodes distinct aggression contexts. A) Strategy for validating *EN-AggINH* using additional aggression paradigms. **B)** Exposure to the urine of another CD1 mouse decreased *EN-AggINH* activity (N=8 mice; P<0.02 using one-tailed Wilcoxon sign-rank). **C)** Experimental approach for causally inducing aggression via direct cellular manipulation. **D)** Network activity decreased in ESR1-Cre×CD1 F1 male mice during aggression (N=8, P<0.01 using one-tailed Wilcoxon sign-rank). **E-F)** Direct activation of ESR1+ cells in ventromedial hypothalamus. E) Cellular activation induced aggression towards female mice (P<0.001 using one tailed t-test) and F) decreased *EN-AggINH* activity (P<0.01 using one-tailed Wilcoxon sign-rank test).

Figure 6. *EN-Aggression Inhibition* activity is causally related to aggression. A) Schematic for closed loop manipulation of *EN-AggINH* activity. **B)** Real-time estimation of aggression. Receiver operating characteristic depicting detection of aggressive behavior in a mouse using *EN-AggINH* activity vs. real-time reduced encoder is shown to the right. Dashed blue line corresponds to the established detection threshold. **C)** Detection of aggression using reduced encoder vs. *EN-AggINH* across mice (N=9; P=0.43 using two-tailed paired Wilcoxon sign-rank). **D)** *EN-AggINH* activity relative to light stimulation during closed loop manipulation. Network activity significantly decreased one second prior to yellow light stimulation (N=9, $^{\#\#}$P<0.005 using one-tailed sign rank test). Activity was also higher one second after stimulation with blue light vs.

yellow light (**P<0.01 using one-tailed signed-rank test). **E)** Portion of windows stimulated during social conditions (F_8= 4.453, P=0.023 using one-way ANOVA; *P<0.05 for comparisons vs. aggression using t-test with FDR correction). **F)** Behavioral effects of closed loop stimulation (**P<0.005 using paired t-test with FDR correction). **G)** Schematic for nonsynchronous control stimulation. **H)** Nonsynchronous stimulation does not impact aggressive behavior (P>0.05 using paired t-test). **I)** Impact of open loop (20Hz, fixed) and closed loop stimulation on locomotor behavior (P**<0.01 using paired two-tailed t-test).

Author contributions

Conceptualization and Methodology – YSG, AT, NMG, SR, DEC, and KD; Formal Analysis – YSG, AT, NMG, AJF, KKW, DEC, and KD; Investigation – YSG, AT, NMG, GET, AJF, KKW, SR, DEC, and KD; Resources – YSG, AT, NMG, GET, AJF, KKW, SR, DEC, KD; Writing – Original Draft, YSG, AT, NMG, SR, DEC, and KD; Writing – Review & Editing, YSG, AT, NMG, KKW, SR, DEC, and KD; Visualization – YSG, AT, DEC, and KD; Supervision –KKW, DEC, and KD; Project Administration and Funding Acquisition –SR, DEC, and KD; See Supplemental materials for detailed author contributions.

Declaration of Interests

The authors declare no competing interests

Materials and Methods

Animals care and use

All procedures were approved by the Duke University Institutional Animal Care and Use Committee in compliance with National Institute of Health (NIH) Guidelines for the Care and Use of Laboratory Animals. Mice were maintained on a reverse 12-hr light cycle with *ad libitum* access to food and water.

Six-month retired breeder male CD1 strain mice (Charles River Laboratories, Wilmington, Massachusetts) were used to discover a network the encoded aggression, hereafter called *EN-AggINH*. These CD1 mice were also used to probe the behavioral and network responses to optogenetic stimulation. Mice were singly housed with enrichment. ESR1-Cre mice on a C57/Bl6J background were provided by Scott Russo. These mice were crossed with CD1 females

in the Bryan Vivarium at Duke University to obtain F1 offspring. Fourteen-week-old virgin ESR1-cre F1 male offspring were used to validate *EN-AggINH*. All F1 offspring were group-housed 2-5 mice per cage until they received viral injections in the ventromedial hypothalamus at 7-8 weeks. After surgery, these mice were singly housed with enrichment. All partner mice (intact males, female, and castrated males) were 7-14 weeks old. These mice were purchased from Jackson Laboratories (Bar Harbor, Maine) and maintained on a C57BL/6J strain background. All stimulus mice were housed 5 per cage with enrichment. All behavior testing and neurophysiological recordings occurred during the dark cycle.

Castration of C57 male mice

Male mice were anesthetized with 1% isoflurane. The scrotal sac was sanitized with betadine and 70% ethanol. The testes were then moved into the sac by gently palpating the lower abdomen. Next, an incision was made in the sac and the testes were extracted. After blood flow was cut off to the testes using a thread tourniquet, the testes were removed. The remaining fatty tissue was placed back into the scrotum, which was then sutured. Mice were allowed 10 days for recovery prior to experimental use.

Electrode implantation surgery

Mice were anesthetized with 1% isoflurane and placed in a stereotaxic device. Grounding screws were placed above the cerebellum, right parietal hemisphere, and anterior cranium. The recording bundles were designed to target prelimbic cortex, infralimbic cortex, medial amygdala, ventral hippocampus, primary visual cortex, mediodorsal thalamus, lateral habenula, lateral septum nucleus, nucleus accumbens, ventrolateral portion of the ventromedial hypothalamus, and orbitofrontal cortex were centered based on stereotaxic coordinates measured from bregma. [Orbital frontal cortex: anterior/posterior (AP) 2.35mm, medial/lateral (ML) 1.0mm, dorsal/ventral (DV) from dura -2.75mm; infralimbic cortex and prelimbic cortex: AP 1.8mm, ML 0mm, DV -2.7mm from dura; medial amygdala: AP -1.25, ML 2.7mm, DV -4.3 from dura; lateral septum and nucleus accumbens: AP 1.0mm, ML 0mm, DV -4.0mm from dura; ventromedial hypothalamus, lateral habenula, and medial dorsal thalamus: AP -1.47mm, ML 0mm, DV -5.4mm from dura; central hippocampus and primary motor cortex: AP -3.0mm, ML 2.6mm, DV -3.0mm from dura]. We targeted infralimbic cortex and prelimbic cortex by building a 0.6mm DV stagger into the bundle. We targeted lateral septum and nucleus accumbens by building a 0.3mm ML and 1.5mm DV stagger into the bundle. We targeted lateral habenula, medial dorsal thalamus, and ventral medial hypothalamus by building a 0.3mm ML and am1.9mm and 2.5mm DV stagger into our electrode bundle microwires. We targeted primary motor cortex and ventral hippocampus using a 0.3mm ML and 2.5mm DV stagger in our electrode bundle microwires. For optogenetic stimulation experiments, the addition of a Mono Fiberoptic Cannula coupled to a 2.5mm metal ferrule (NA: 0.22, 100mm [inner diameter], 125mm buffer [outer diameter], MFC_100/125-0.22, Doric Lenses, Quebec) was built into the PFC bundle. The tip of the fiber was secured 300mm above the tip of the IL microwire. Mice were allowed 10-15 days for to recovery from surgery before behavioral testing.

Viral surgery

For optogenetic experiments (Takahashi *et al.*, 2014), we used CD1 mice that showed an attack latency < 60s when exposed to an intact C57 male. CD1 mice were anesthetized with 1% isoflurane and placed in a stereotaxic device. The mice were unilaterally injected with AAV2-CamKII-Chr2-EYFP (purchased from the Duke Viral Vector Core, Durham, NC; courtesy of K. Deisseroth), based on stereotaxic coordinates from bregma (left Infralimbic cortex: AP 1.8mm, ML 0.3mm, DV -2.0mm from the brain). A total of 0.5mL of virus was infused at the injection site at a rate of 0.1mL/min over five minutes and the needle left in place for ten minutes after injection. For the open loop stimulation experiment, CD1 mice were implanted with an optic fiber (Mono Fiberoptic Cannula coupled to a 2.5mm metal ferrule (NA: 0.22, 100mm [inner diameter], 125mm buffer [outer diameter], MFC_100/125-0.22, Doric Lenses, Quebec)) 0.3mm above the injection site immediately after viral syringe was removed. These mice were allowed 3 weeks for recovery prior to behavioral testing. For the closed-loop experiments, CD1 mice were allowed 3 weeks for viral expression prior to implantation with an optrode.

For the ESR1-Cre validation experiment, F1 offspring were bilaterally injected with AAV2-hSyn-DIO-GqDREADD (obtained from Addgene) based on stereotaxic coordinates measured from bregma (AP -1.5mm, ML ±0.7mm, DV -5.7mm from the dura). A total of 0.3mL of virus was infused bilaterally at a rate of 0.1mL/min and the needle left in place for five minutes after injection. Two weeks after viral infusion, F1 males were screened for aggressive behavior towards females. The F1 males received i.p. injections of CNO (1mg/kg) at the start of the screening session. Thirty-five minutes after injection, a novel C57 female was placed in the home cage for 5 minutes. Screening was repeated one week and two weeks later. Only F1 males who attacked females for at least two of the three screening sessions (9/13 mice) were implanted with electrodes (Takahashi *et al.*, 2017). The eight mice that showed good surgical recovery were subjected to further experiments.

Histological analysis

Histological analysis of implantation sites was performed at the conclusion of experiments to confirm recording sites and viral expression. Animals were perfused with 4% paraformaldehyde (PFA) and brains were harvested and stored for at least 96 hrs in PFA. Brains were cryoprotected with sucrose and frozen in OCT compound and stored at -80C. Brains were later sliced at 40mm. Brains from mice used to train and validate the network were stained using NeuroTrace fluorescent Nissl Stain (N21480, ThermoFisher Scientific, Waltham, MA) using standard protocol. Specifically, Nissl staining for brain tissue occurred on a shaker table at room temperature. Tissue was washed in PBST (0.1% Triton in phosphate-buffered saline solution) for 10 minutes. It was then washed for five minutes in PBS twice. The tissue was then protected from light for the remainder of the protocol. The tissue was incubated in 1:300 Nissl diluted in 2 mL PBS for 10 minutes. After the Nissl incubation, tissue was washed once in 0.1% PBST for 10 minutes, then twice in PBS for 5 minutes. Brains from ESR1 mice and mice used for 20 Hz or closed loop stimulation were mounted in Vectashield mounting medium containing DAPI (H-1200-10, Vector Laboratories, Newark, CA). Images were obtained at 10x using an Olympus fluorescent microscope. Of the 297 total implantation sites in the training and testing set of mice, 3 were mistargeted (~1% error rate). As machine learning analysis typically benefit from larger data sets, we concluded that maintaining a higher number of data points likely outweighed the effect of a small number of inaccurate targets. Thus, we pooled data from all

20 implanted animals to learn our initial model. We employed a similar strategy for our validation analysis, where an animal was only removed from the validation set if there was clear histological confirmation of mistargeting for any of the recorded regions. Specifically, presuming accurate targeting with 99% certainty, we included animals with missing or damaged histological slices in our analysis. However, if there was clear histological confirmation of mistargeting for any of the recorded regions, the animal was removed from the validation testing. Critically, our validation procedure implies that the machine learning models were robust regardless of any slight imprecision in the animals we utilized for training.

Neurophysiological data acquisition

Mice were connected to a data acquisition system (Blackrock Microsystems, UT, USA) while anesthetized with 1% isoflurane. Mice were allowed 60 minutes in their home cage prior to behavioral and electrophysiological recordings. Local field potentials (LFPs) were bandpass filtered at 0.5-250Hz and stored at 1000Hz. An online noise cancellation algorithm was applied to reduce 60Hz artifact (Blackrock Microsystems, UT, USA). Neural spiking data was referenced online against a channel recording from the same brain area that did not exhibit a SNR>3:1. After recording, cells were sorted using an offline sorting algorithm (Plexon Inc., TX) to confirm the quality of the recorded cells. Only cell clusters well-isolated compared to background noise, defined as a Mahanalobis distance greater than 3 compared to the origin, were used for the unit-Electome Factor correlation analysis. We used both single and multi-units for our analysis as our objective was to determine whether Electome Network activity was reflective of cellular activity. Neurophysiological recordings were referenced to a ground wire connected to ground screws above the cerebellum and anterior cranium.

Behavioral recordings and analysis for training/testing models

The CD1 mice used for training and testing the electome model were first subjected to screening to assess their basal level of aggressiveness. Screening occurred once a day for three consecutive days prior to surgical implantation. For each screening session, an intact male C57 was placed in the CD1's home cage for 5 minutes and the latency to first attack was recorded. To ensure that our network generalized broadly across CD1 mice, we used a training and testing set for which ~50% of the mice showed high aggression (i.e., latency to attack < 60s), and ~50% of the mice showed low to moderate aggression (i.e., latency to attack > 60s). Animals were screened in cohorts, and we excluded 16/45 mice from further experiments.

All screening/testing occurred within the home cage of mice except for the quantification of cortical stimulation induced gross locomotor activity. These latter experiments were performed in a 44cm × 44cm × 35cm (L×W×H) open field arena. Subject mice (CD1 and ESR1 males) were acclimated to the recording tether for three days prior to the first recording session. Each acclimation session involved anesthetizing the mouse with 1% isoflurane, tethering the subject mouse, allowing 60 minutes to recover from isoflurane, then placing a male C57 in the home cage for 5 minutes. Mice were then anesthetized with isoflurane again and detached from the tether. The aggressive tone of experimental mice was determined based on average latency to attack partner mice during the second and third acclimation sessions.

After screening, mice were implanted, and data was acquired across 1-6 behavioral testing/recording sessions following recovery. Sessions were separated by 5-7 days. Recordings for all social encounters were performed in the home cages of the CD1 mice. Each behavioral testing session began with 5 minutes of recording prior to introduction of the social stimulus. All mice were subjected to encounters with an intact C57 male mouse and a female c57 mouse. A subset of eighteen mice were also subjected to an encounter with a castrated male mouse, and another subset of eighteen mice were subjected to exposure to objects covered in CD1 mouse urine. Six mice were recorded under all four conditions. Data observations (1 second each) were pooled across eleven CD1 mice for training the network model. Object pairs included yellow duplex blocks, curved red duplex blocks, weighted 5 mL conicals, glassware tops, and objects assembled from black legos®. The CD1 mice were exposed to a different pair of objects during each session. Order of exposure to stimulus mice and objects was shuffled for every session.

For ESR1 male behavioral testing, mice were injected with either saline or CNO (1mg/kg, i.p.) after the five-minute baseline recording. Thirty-five minutes after this injection, mice were exposed to an intact male C57, a castrated male C57, and a female C57, presented in pseudorandom order. Mice were subjected to six total recording sessions (three in which they were treated with saline and three in which they were treated with CNO), again in pseudorandom order. Sessions were separated by 5 days to allow an adequate washout of CNO(Jendryka et al., 2019).

Behavior was scored for each second as an “aggressive interaction”, “pro-social interaction”, or “non-interaction”. One second windows were identified as "aggressive" if the mouse was engaged in biting, boxing (kicking/clawing), tail rattling, or tussling behavior (Miczek *et al.*, 2001). Windows were labeled as "pro-social interaction" if the mouse had his nose or forepaws touching the stimulus mouse (intact male/female/castrated male) or object, but was not biting, boxing, or tussling. Examples of behaviors labeled "pro-social" included sniffing, grooming, or resting (placing nose or forepaws against the subject mouse, but not moving). If the stimulus mouse had his/her forepaws or nose on the CD1, but it was not reciprocated, this was labeled "non-interaction". CD1 straight approach, sideways approach, and chasing of the stimulus mouse could result in attack (biting/kicking/tousling), pro-social (nose or paw touch), or withdrawal without any touch. Thus, while sideways approach and chasing are regularly labeled as "aggressive" in the literature (Kwiatkowski et al., 2021; Miczek *et al.*, 2001; Miczek and O'Donnell, 1978), and straight approach is regularly labeled as "pro-social", these behaviors lacked consistent resolutions. Moreover, mice also demonstrated these behaviors towards female and castrated mice (pro-social context). One second windows containing these behaviors were labeled "non-interaction". All other timepoints not labeled "aggressive" or "pro-social" were also labeled "non-interaction". These behavioral criteria were selected to include ethologically aggression-related behaviors and maximize the likelihood that the CD1 was aware of the presence of the stimulus mouse or object during the behavioral window, while remaining confident in the classification of "aggressive" and "pro-social" window labels.

While tail rattling is not an attack behavior like the other behaviors that were labeled as "aggressive", it was consistently only demonstrated by aggressive mice towards intact male mice. Moreover, tail rattling is well-recognized in the literature as an aggressive behavior. In our subset of 20 mice used for training the network, tail rattling was observed 8 ± 4s out of the 135 ± 26s 'aggressive windows' per mouse.

The videos used to generate the labels for training and testing our machine learning model was hand-scored by a trained researcher. Videos from ESR1 mice and optogenetic stimulation were automatically tracked using DeepLabCut (Mathis et al., 2018; Nath et al., 2019). This information was then used for creating behavioral classifiers in SimBA (Nilsson et al., 2020).

LFP preprocessing and signal artifact removal

Each LFP signal was segmented into 1s non-overlapping windows. If there were multiple intact channels implanted in a region, they were averaged to produce a single signal. Windows with non-physiological noise were removed using an established automated heuristic (Mague *et al.*, 2022). Briefly, the envelope of the signal in each channel was estimated using the magnitude of the Hilbert transform. The Median Absolute Deviation (MAD) of the magnitude was then calculated on each channel of each recording. Signal was marked as non-physiological if the envelope exceeded a high threshold (5x MAD, which is roughly 4x the standard deviation for a normally distributed signal). Any data adjacent to non-physiological data that had an envelope value above a smaller threshold (0.167 MAD) was also considered non-physiological. All data marked in this way was ignored when averaging channels for each region. Any channels with standard deviation less than 0.01 was removed as well. If no channels were usable for a given region within a window, that whole window was removed from the data. This set of heuristics resulted in 34.7±5.1% of the data being excluded from analysis. After this, 60Hz line artifact was further removed using a series of Butterworth bandpass filters at 60Hz and harmonics up to 240Hz with a stopband width of 5Hz and stopband attenuation of -60dB. Finally, the signal was downsampled to 500Hz.

Estimation of LFP oscillatory power, cross-spectral coherence, and Granger directionality.

Signal processing was performed using Matlab (The MathWorks, Inc. Natick MA). For LFP power, spectral power was estimated using Welch's method using a 250-millisecond window and 125 millisecond step. Windows were zero-padded to give a 1Hz resolution. Cross-spectral coherence was estimated pairwise between all regions using Welch's method and magnitude-squared coherence defined as

$$C_{AB}(\omega) = \frac{|Psd_{AB}(\omega)|^2}{Psd_{AA}(\omega)Psd_{BB}(\omega)},$$

where A and B are two regions and $Psd_{AA}(\omega)$ and $Psd_{AB}(\omega)$ are the power and cross spectra at a given frequency ω, respectively.

Spectral Granger causality features were estimated using the Multivariate Granger Causality (MVGC) MATLAB toolbox (Barnett and Seth, 2014). To get stable Granger causality estimates, a 6th order highpass Butterworth filter – with a stopband at 1Hz and a passband starting at 4 Hz – was applied to the data using the *filtfilt* function (MATLAB, The MathWorks, Inc. Natick MA). Granger causality values for each window were estimated with a 20 order AR model at 1 Hz

intervals to align with the power and coherence features. Granger features were processed identically to a previously reported approach (Mague *et al.*, 2022). Briefly, Granger features were exponentiated to approximately maintain the additivity assumption made implicitly by NMF models (Gallagher et al., 2017; Mague *et al.*, 2022) as, $\exp\left(f_{A\to B}(\omega)\right)$, where $f_{A\to B}(\omega)$ is the Granger causality at frequency ω from region A to region B. The exponentiated feature is a ratio of total power to unexplained power. Exponentiated Granger feature values were truncated at 10 to prevent implausible values.

Data for single region and network-level machine learning analyses

We used 21,460 seconds of data, pooled across the twenty mice, to train/validate our single region and network models. This included a total of 4680 seconds while mice were socially isolated in their home cage, 14,890 seconds where CD1 mice exhibited pro-social behavior (3542 seconds towards intact males, 9067 seconds towards females, and 2,281 seconds towards castrated), and 1,890 seconds where mice exhibited aggression towards the intact males.

Single region decoding

To test the efficacy of any single brain region as a biomarker for aggression, we extracted power at 1 Hz frequency bins over 1-56 Hz from each region. We then used the *fitglm* function (MATLAB, The MathWorks, Inc. Natick MA) for binary outcomes with the logit link function to build a predictive model from the power across frequencies. One second windows were pooled from the twenty CD1 mice and used to generate a series of logistic models for each of the 11 brain regions. The models were trained to distinguish behavioral windows from one social state exhibited by CD1 mice, from two other social states. These three social states included 1) male-directed aggression, 2) female pro-social interactions, and 3) castrated male pro-social interactions. We also developed a model to distinguish 4) periods where CD1 were isolated in their home cage from any of the three social states. Each model was then tested on data from a holdout set of nine mice. The area under the receiver operating curve (AUC) was calculated for each holdout mouse to determine the performance of the model. False discovery rate was used to correct for multiple hypothesis testing.

Discriminative Cross-Spectral Factor Analysis – Nonnegative Matrix Factorization

The network was trained to distinguish between behavioral windows when the CD1 mice showed aggressive behavior towards intact C57 males, and windows where they exhibited pro-social behavior. These latter windows comprised pro-social interactions towards intact C57 males, castrated C57 males, or C57 females. Here, we used data from twenty-seven mice to learn the final model, with a split of 20 and 7 for model training and internal validation.

We used Discriminative Cross-Spectral Factor Analysis – Nonnegative Matrix Factorization (dCSFA-NMF) model (Talbot *et al.*, 2020). This approach assumes each window of is an independent stationary observation and examines dynamics in brain activity only at the scale of windows. A 1 second window was chosen to balance capturing fine-grained transient behavior with sufficient length to properly estimate spectral features (Mague *et al.*, 2022). Each window has associated spectral power, coherence, and Granger causality features (in total $p = 9{,}586$ features), which is represented as $x_i \in \mathbb{R}^p$ for the ith window. Each window was associated

with a behavioral label that identified which condition the CD1 mouse was subjected to during that window (intact male, castrated male, or female) and whether the CD1 mouse was engaged in aggressive or non-aggressive behavior during that window, and the aggressive behavior was coded as $y_i \in \{0,1\}$.

As a short description of the dCSFA-NMF model, the features are described as an additive positive sum of K non-nonnegative electrical functional connectome networks. This model is learned using a supervised autoencoder. The objective we use to learn the parameters is

$$\min_{W,d,\phi} \sum_{i=1}^{N} Loss_X\big(x_i, Wf(x_i;\phi)\big) + \lambda\, Loss_Y\big(y_i, df(x_i;\phi)\big) + \mu\, Loss_{EN}(A).$$

Here, $Loss_X$ is the reconstruction loss of the features derived from electrophysiology, which for our work was an L_2 loss. Our predictive loss $Loss_Y$ is the cross-entropy loss commonly used for binary classification. Each of the K networks is represented in vector form and combined to make a matrix $W \in R^{p\times K}$. The electome factor scores are given by the multi-output function $f(x;\phi): R^p \to R^K$, where ϕ represent the parameters of the function. In our model, the multi-output function was an affine transformation of $Ax + b$ followed by a softplus rectification, defined as $softplus(x) = \log(1 + \exp x)$, thus $\phi = \{A, b\}$. Parameters $d \in R^k$ represent the relationship between the electome factors and the behavior. A sparsity constraint is enforced so that $d = [d_1, 0, \dots, 0]$, meaning that only a single electome is used to predict behavior, simplifying interpretation. λ is a weighting parameter used to control the relative importance of prediction. We chose a value that kept the two losses approximately equal during training, which corresponded to 1.

Previous work has also found that the reconstruction loss can reduce overfitting and make the learned predictions more robust (Le et al., 2018). To further reduce overfitting of the predictive aspect of the encoder, we applied an elastic net loss (Zou and Hastie, 2005) on the encoder $Loss_{EN}$ with a weighting μ and the ratio between the L_1 and L_2 losses set to .5. The value for μ was set to a small value that had worked well previously. In this work, power features were also upweighted by a factor of 10 to accommodate that there were many more Granger features and truncated at 6 to prevent outliers from dominating the predictions.

These models and statistical analyses were implemented with Python 3.7 and Tensorflow version 2.4. Parameters were learned with stochastic gradient descent using the Nesterov accelerated ADAM optimizer (Tato and Nkambou, 2022). Learning was performed for 30000 iterations, which was observed to be ample for parameter convergence. The learning rate and batch size were set to 1e-3 and 100 respectively, values that have empirically performed well in similar applications. This learning algorithm is publicly available code at https://github.com/carlson-lab/encodedSupervision.

Predictive performance was evaluated in new mice not involved in learning the network. Given processed data from the new mice, factor scores were estimated as an evaluation of the encoder learned during training of the dCSFA-NMF model.

Hyper-parameter selection

The dCSFA-NMF procedure requires selection of several settings in the algorithm. Specifically, we must choose the number of electome factors K, the importance of the supervised task λ, the relative importance of the power features, coherence features, and Granger features, and the parameterization of the mapping function $f(x_i; \phi)$. Besides K, these settings were chosen to match previously used values or follow heuristics.

To choose the value of K, we evaluated the reconstruction error (Mean Squared Error) on the seven internal validation mice, which evaluates how well the Electome networks describe the neural measurements. As the goal for our analysis was to maintain high reconstruction and effectively predict the behavior, an elbow analysis was used to choose the number of electome networks K after which we observed minimal gains in explaining the data.

Validating Model Dimensionality

A frequent concern of latent variable models (including dCSFA-NMF) is the dependence of the networks and encoder on the choice of latent dimensionality. To address this concern, we performed a sensitivity analysis on the supervised network to determine the extent to which the choice of this dimensionality influenced the learned aggressive network electome and encoder. In this sensitivity analysis, we estimated a dCSFA-NMF model allowing the number of total networks to range from two to twenty. We then compared the similarity between each learned encoder and decoder to our model with eight networks (the final model used in this work). This was quantified using the cosine similarity, which measures the angle between two networks (or encoders), ranging from -1 to 1. A value of 1 indicates perfect alignment (pointing in the same direction), 0 is orthogonal, and -1 indicates that the vectors point in opposite directions (Supplementary Figure S2).

We found that the supervised network maintained a strong consistency across most dimensions, particularly between 5-10 networks, as shown by the cosine similarities being greater than 0.95. The supervised encoders were virtually identical all the models except the one that utilized three networks. This model learned a network that was positively associated with aggression.

To evaluate the robustness of the similarities across most of the supervised networks, we created a null distribution of the similarities across randomly chosen generative networks. These later similarities were substantially lower both for the network composition and the encoder. This indicates that as far as the supervised electome and encoder are concerned, latent dimensionality is not particularly influential on the resulting network, and by extension the biological interpretation.

Decoder Information Content

The amount of information contained in the predictive model was quantified by the reduction in uncertainty. The associated formula for this reduction in uncertainty, known as the Bernoulli entropy, is 1-1*(p log_2(p) +(1-p)log_2(1-p)), where p is the accuracy of the model. At the

extremes, an accuracy of 0.5 (random guessing) removes no uncertainty, whereas an accuracy of 1 or 0 completely eliminates uncertainty.

Single cell correlation to Electome Factor activity

Data acquired during the third behavioral testing session was from the twenty implanted mice were used for cellular analysis. We used Spearman correlation to quantify the relationship between cellular firing windows and the activity of the electome network used to classify aggression. We performed 1000 permutations of randomly shuffling 1 second windows within each class for aggressive and pro-social interactions with male, female and castrated C57 mice. We then calculated the spearman correlation of each cell for each permutation. A cell was deemed positively correlated if it exhibited a spearman Rho above 97.5% of the permutated distribution and negatively correlated if it was below 2.5%.

Real-Time Encoder Approximation

Because Granger causality features were too computationally demanding for real-time calculation, we developed a 'fast' dCSFA-NMF model that relied only on power and coherence features for estimation of aggressive state to use in the closed-loop stimulation experiments. This 'fast' model was trained on the same data. The model was trained using regularized regression to best predict the output of the full encoder. As such, this reduced encoder is also an affine transformation followed by a SoftPlus activation with a smaller parameter set, $\phi_r = \{A_r, b_r\}$. This approximation explained a large component of the variance of the supervised network score on the hold-out validation mice ($R = 0.47$, p-value $<10^{-16}$).

Optogenetic stimulation

Mice were anesthetized with 1% isoflurane, then tethered to for electrophysiological recording and had an optic patch cable places over the optic fiber cannula. The mice were then allowed 60 minutes for recovery prior to session recording. For the fiber only optogenetic stimulation experiments, CD1 mice experienced two stimulation sessions. For closed loop optogenetic stimulation, CD1 mice experienced three sessions of behavioral screening, then two sessions of closed-loop, then two sessions of open-loop, then one session of random stimulation. Stimulation sessions were separated by 5-7 day between sessions. For behavioral screening, CD1 mice were exposed to intact C57 males, females, and castrated male mice for 5 minutes each. Screening session two and three were used to determine reduced network threshold at which 40% of aggressive behavioral windows could be detected. For each session, mice were recorded for 3 minutes of baseline in their home cage, then during the three social encounters. Mice were recorded in an open field for 5 minutes after each session. The order of the three social encounters were shuffled for each session. During each condition, the CD1 mouse received two segments of alternating blue (473nm) and yellow (593.5nm) light stimulation, for two minutes each.

For open-loop, CD1 mice received light stimulation for the entirety of the two-minute segment. For closed-loop, mice received stimulation for one second when reduced network score dropped below threshold.

For nonsynchronous stimulation, each CD1 mouse was randomly matched to a different mouse that had been used for closed-loop stimulation. Each non-synchronous mouse was then subjected to the identical order of conditions and yellow and blue light stimulation blocks as their individually matched closed-loop mouse. Light stimulation was delivered using the pattern implemented for the closed-loop partner mouse.

Statistics

GraphPad Prism and Matlab were used for statistical analyses of behavior and network activity. Paired T-tests were used for comparing within subject behavioral response to optogenetic stimulation or CNO application and corrected for false discovery rate for multiple hypothesis testing through the Benjamini-Hochberg procedure. One tailed Wilcoxon signed rank tests were used to compare within-subject mean network score responses to optogenetic stimulation, stimulus mouse exposure and interaction, and CNO injection. Data is presented as mean ± standard error of measurement, throughout the paper, unless otherwise specified.

References

Anikeeva, P., Andalman, A.S., Witten, I., Warden, M., Goshen, I., Grosenick, L., Gunaydin, L.A., Frank, L.M., and Deisseroth, K. (2011). Optetrode: a multichannel readout for optogenetic control in freely moving mice. Nat Neurosci. 10.1038/nn.2992.

APA. (2000). Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR (American Psychiatric Association).

Barnett, L., and Seth, A.K. (2014). The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. Journal of neuroscience methods *223*, 50-68. 10.1016/j.jneumeth.2013.10.018.

Brayley, K.N., and Albert, D.J. (1977). Suppression of VMH-lesion-induced reactivity and aggressiveness in the rat by stimulation of lateral septum, but not medial septum or cingulate cortex. J Comp Physiol Psychol *91*, 290-299.

Chen, A.X., Yan, J.J., Zhang, W., Wang, L., Yu, Z.X., Ding, X.J., Wang, D.Y., Zhang, M., Zhang, Y.L., Song, N., et al. (2020). Specific Hypothalamic Neurons Required for Sensing Conspecific Male Cues Relevant to Inter-male Aggression. Neuron *108*, 763-774 e766. 10.1016/j.neuron.2020.08.025.

Chen, P., and Hong, W. (2018). Neural Circuit Mechanisms of Social Behavior. Neuron *98*, 16-30. 10.1016/j.neuron.2018.02.026.

Choy, O., Raine, A., and Hamilton, R.H. (2018). Stimulation of the Prefrontal Cortex Reduces Intentions to Commit Aggression: A Randomized, Double-Blind, Placebo-Controlled, Stratified, Parallel-Group Trial. J Neurosci *38*, 6505-6512. 10.1523/JNEUROSCI.3317-17.2018.

Couppis, M.H., and Kennedy, C.H. (2008). The rewarding effect of aggression is reduced by nucleus accumbens dopamine receptor antagonism in mice. Psychopharmacology (Berl) *197*, 449-456. 10.1007/s00213-007-1054-y.

Dambacher, F., Schuhmann, T., Lobbestael, J., Arntz, A., Brugman, S., and Sack, A.T. (2015). Reducing proactive aggression through non-invasive brain stimulation. Soc Cogn Affect Neurosci *10*, 1303-1309. 10.1093/scan/nsv018.

de la Zerda, S.H., Netser, S., Magalnik, H., Briller, M., Marzan, D., Glatt, S., Abergel, Y., and Wagner, S. (2022). Social recognition in laboratory mice requires integration of behaviorally-induced somatosensory, auditory and olfactory cues. Psychoneuroendocrinology *143*, 105859. 10.1016/j.psyneuen.2022.105859.

Deonaraine, K.K., Wang, Q., Cheng, H., Chan, K.L., Lin, H.Y., Liu, K., Parise, L.F., Cathomas, F., Leclair, K.B., Flanigan, M.E., et al. (2020). Sex-specific peripheral and central responses to stress-induced depression and treatment in a mouse model. J Neurosci Res *98*, 2541-2553. 10.1002/jnr.24724.

Falkner, A.L., Dollar, P., Perona, P., Anderson, D.J., and Lin, D. (2014). Decoding ventromedial hypothalamic neural activity during male mouse aggression. J Neurosci *34*, 5971-5984. 10.1523/JNEUROSCI.5109-13.2014.

Falkner, A.L., Grosenick, L., Davidson, T.J., Deisseroth, K., and Lin, D. (2016). Hypothalamic control of male aggression-seeking behavior. Nat Neurosci *19*, 596-604. 10.1038/nn.4264.

Flanigan, M., Aleyasin, H., Takahashi, A., Golden, S.A., and Russo, S.J. (2017). An emerging role for the lateral habenula in aggressive behavior. Pharmacol Biochem Behav *162*, 79-86. 10.1016/j.pbb.2017.05.003.

Gallagher, N.M., Ulrich, K., Dzirasa, K., Carin, L., and Carlson, D.E. (2017). Cross-Spectral Factor Analysis Advances in Neural Information Processing Systems *30*.

Girardeau, G., Benchenane, K., Wiener, S.I., Buzsaki, G., and Zugaro, M.B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. Nat Neurosci *12*, 1222-1223. 10.1038/nn.2384.

Golden, S.A., Heshmati, M., Flanigan, M., Christoffel, D.J., Guise, K., Pfau, M.L., Aleyasin, H., Menard, C., Zhang, H., Hodes, G.E., et al. (2016). Basal forebrain projections to the lateral habenula modulate aggression reward. Nature *534*, 688-692. 10.1038/nature18601.

Haller, J. (2018). The role of central and medial amygdala in normal and abnormal aggression: A review of classical approaches. Neurosci Biobehav Rev *85*, 34-43. 10.1016/j.neubiorev.2017.09.017.

Hashikawa, K., Hashikawa, Y., Falkner, A., and Lin, D. (2016). The neural circuits of mating and fighting in male mice. Current opinion in neurobiology *38*, 27-37. 10.1016/j.conb.2016.01.006.

Hashikawa, Y., Hashikawa, K., Falkner, A.L., and Lin, D. (2017). Ventromedial Hypothalamus and the Generation of Aggression. Front Syst Neurosci *11*, 94. 10.3389/fnsys.2017.00094.

Hultman, R., Ulrich, K., Sachs, B.D., Blount, C., Carlson, D.E., Ndubuizu, N., Bagot, R.C., Parise, E.M., Vu, M.T., Gallagher, N.M., et al. (2018). Brain-wide Electrical Spatiotemporal Dynamics Encode Depression Vulnerability. Cell *173*, 166-180 e114. 10.1016/j.cell.2018.02.012.

Jendryka, M., Palchaudhuri, M., Ursu, D., van der Veen, B., Liss, B., Katzel, D., Nissen, W., and Pekcec, A. (2019). Pharmacokinetic and pharmacodynamic actions of clozapine-N-oxide, clozapine, and compound 21 in DREADD-based chemogenetics in mice. Scientific reports *9*, 4522. 10.1038/s41598-019-41088-2.

Jones, R.B., and Nowell, N.W. (1973). The effect of urine on the investigatory behaviour of male albino mice. Physiology & behavior *11*, 35-38. 10.1016/0031-9384(73)90119-4.

Kumar, S., Black, S.J., Hultman, R., Szabo, S.T., DeMaio, K.D., Du, J., Katz, B.M., Feng , G., Covington, H.E., 3rd, and Dzirasa, K. (2013). Cortical Control of Affective Networks. J Neurosci *33*, 1116 –1129.

Kwiatkowski, C.C., Akaeze, H., Ndlebe, I., Goodwin, N., Eagle, A.L., Moon, K., Bender, A.R., Golden, S.A., and Robison, A.J. (2021). Quantitative standardization of resident mouse behavior for studies of aggression and social defeat. Neuropsychopharmacology *46*, 1584-1593. 10.1038/s41386-021-01018-1.

Le, L., Patterson, A., and White, M. (2018). Supervised autoencoders: Improving generalization performance with unsupervised regularizers. NeuroIPS.

Lee, D.L., and Wilson, J.L. (2012). Urine from Sexually Mature Intact Male Mice Contributes to Increased Cardiovascular Responses during Free-Roaming and Restrained Conditions. ISRN Vet Sci *2012*. 10.5402/2012/185461.

Lee, S.C., Yamamoto, T., and Ueki, S. (1983). Characteristics of aggressive behavior induced by nucleus accumbens septi lesions in rats. Behav Neural Biol *37*, 237-245.

Levy, D.R., Tamir, T., Kaufman, M., Parabucki, A., Weissbrod, A., Schneidman, E., and Yizhar, O. (2019). Dynamics of social representation in the mouse prefrontal cortex. Nat Neurosci *22*, 2013-2022. 10.1038/s41593-019-0531-z.

Li, L., Durand-de Cuttoli, R., Aubry, A.V., Burnett, C.J., Cathomas, F., Parise, L.F., Chan, K.L., Morel, C., Yuan, C., Shimo, Y., et al. (2022). Social trauma engages lateral septum circuitry to occlude social reward. Nature. 10.1038/s41586-022-05484-5.

Lin, D., Boyle, M.P., Dollar, P., Lee, H., Lein, E.S., Perona, P., and Anderson, D.J. (2011). Functional identification of an aggression locus in the mouse hypothalamus. Nature *470*, 221-226. 10.1038/nature09736.

Lischinsky, J.E., and Lin, D. (2020). Neural mechanisms of aggression across species. Nat Neurosci *23*, 1317-1328. 10.1038/s41593-020-00715-2.

Mague, S.D., Talbot, A., Blount, C., Walder-Christensen, K.K., Duffney, L.J., Adamson, E., Bey, A.L., Ndubuizu, N., Thomas, G.E., Hughes, D.N., et al. (2022). Brain-wide electrical dynamics encode individual appetitive social behavior. Neuron *110*, 1728-1741 e1727. 10.1016/j.neuron.2022.02.016.

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., and Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci *21*, 1281-1289. 10.1038/s41593-018-0209-y.

Miczek, K.A., Maxson, S.C., Fish, E.W., and Faccidomo, S. (2001). Aggressive behavioral phenotypes in mice. Behav Brain Res *125*, 167-181. 10.1016/s0166-4328(01)00298-4.

Miczek, K.A., and O'Donnell, J.M. (1978). Intruder-evoked aggression in isolated and nonisolated mice: effects of psychomotor stimulants and L-dopa. Psychopharmacology (Berl) *57*, 47-55. 10.1007/BF00426957.

Murugan, M., Jang, H.J., Park, M., Miller, E.M., Cox, J., Taliaferro, J.P., Parker, N.F., Bhave, V., Hur, H., Liang, Y., et al. (2017). Combined Social and Spatial Coding in a Descending Projection from the Prefrontal Cortex. Cell *171*, 1663-1677 e1616. 10.1016/j.cell.2017.11.002.

Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., and Mathis, M.W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nat Protoc *14*, 2152-2176. 10.1038/s41596-019-0176-0.

Nilsson, S.R., Goodwin, N.L., Choong, J.J., Hwang, S., Wright, H.R., Norville, Z.C., Tong, X., Lin, D., Bentzley, B.S., Eshel, N., et al. (2020). Simple Behavioral Analysis (SimBA): an open source toolkit for computer classification of complex social behaviors in experimental animals. bioRxiv. 10.1101/2020.04.19.049452.

Padilla-Coreano, N., Batra, K., Patarino, M., Chen, Z., Rock, R.R., Zhang, R., Hausmann, S.B., Weddington, J.C., Patel, R., Zhang, Y.E., et al. (2022). Cortical ensembles orchestrate social competition through hypothalamic outputs. Nature *603*, 667-671. 10.1038/s41586-022-04507-5.

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., and Schmidt, L. (2019). A Meta-Analysis of Overfitting in Machine Learning. Advances in Neural Information Processing Systems 32 (NeurIPS 2019)

Shibata, S., Yamamoto, T.Y., and Ueki, S. (1982). Differential effects of medial, central and basolateral amygdaloid lesions on four models of experimentally-induced aggression in rats. Physiol Behav *28*, 289-294.

Takahashi, A., Chung, J.R., Zhang, S., Zhang, H., Grossman, Y., Aleyasin, H., Flanigan, M.E., Pfau, M.L., Menard, C., Dumitriu, D., et al. (2017). Establishment of a repeated social defeat stress model in female mice. Scientific reports *7*, 12838. 10.1038/s41598-017-12811-8.

Takahashi, A., Nagayasu, K., Nishitani, N., Kaneko, S., and Koide, T. (2014). Control of intermale aggression by medial prefrontal cortex activation in the mouse. PLoS One *9*, e94657. 10.1371/journal.pone.0094657.

Talbot, A., Dunson, D., Dzirasa, K., and Carlson, D. (2020). Supervised Autoencoders Learn Robust Joint Factor Models ofNeural Activity. arxiv.

Tato, A., and Nkambou, R. (2022). Infusing Expert Knowledge Into a Deep Neural Network Using Attention Mechanism for Personalized Learning Environments. Front Artif Intell *5*, 921476. 10.3389/frai.2022.921476.

Vu, M.T., Adali, T., Ba, D., Buzsaki, G., Carlson, D., Heller, K., Liston, C., Rudin, C., Sohal, V., Widge, A.S., et al. (2018). A Shared Vision for Machine Learning in Neuroscience. J Neurosci. 10.1523/JNEUROSCI.0508-17.2018.

Wong, L.C., Wang, L., D'Amour, J.A., Yumita, T., Chen, G., Yamaguchi, T., Chang, B.C., Bernstein, H., You, X., Feng, J.E., et al. (2016). Effective Modulation of Male Aggression through Lateral Septum to Medial Hypothalamus Projection. Curr Biol *26*, 593-604. 10.1016/j.cub.2015.12.065.

Yang, C.F., Chiang, M.C., Gray, D.C., Prabhakaran, M., Alvarado, M., Juntti, S.A., Unger, E.K., Wells, J.A., and Shah, N.M. (2013). Sexually dimorphic neurons in the ventromedial hypothalamus govern mating in both sexes and aggression in males. Cell *153*, 896-909. 10.1016/j.cell.2013.04.017.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society (*Part 2*), 301-320.

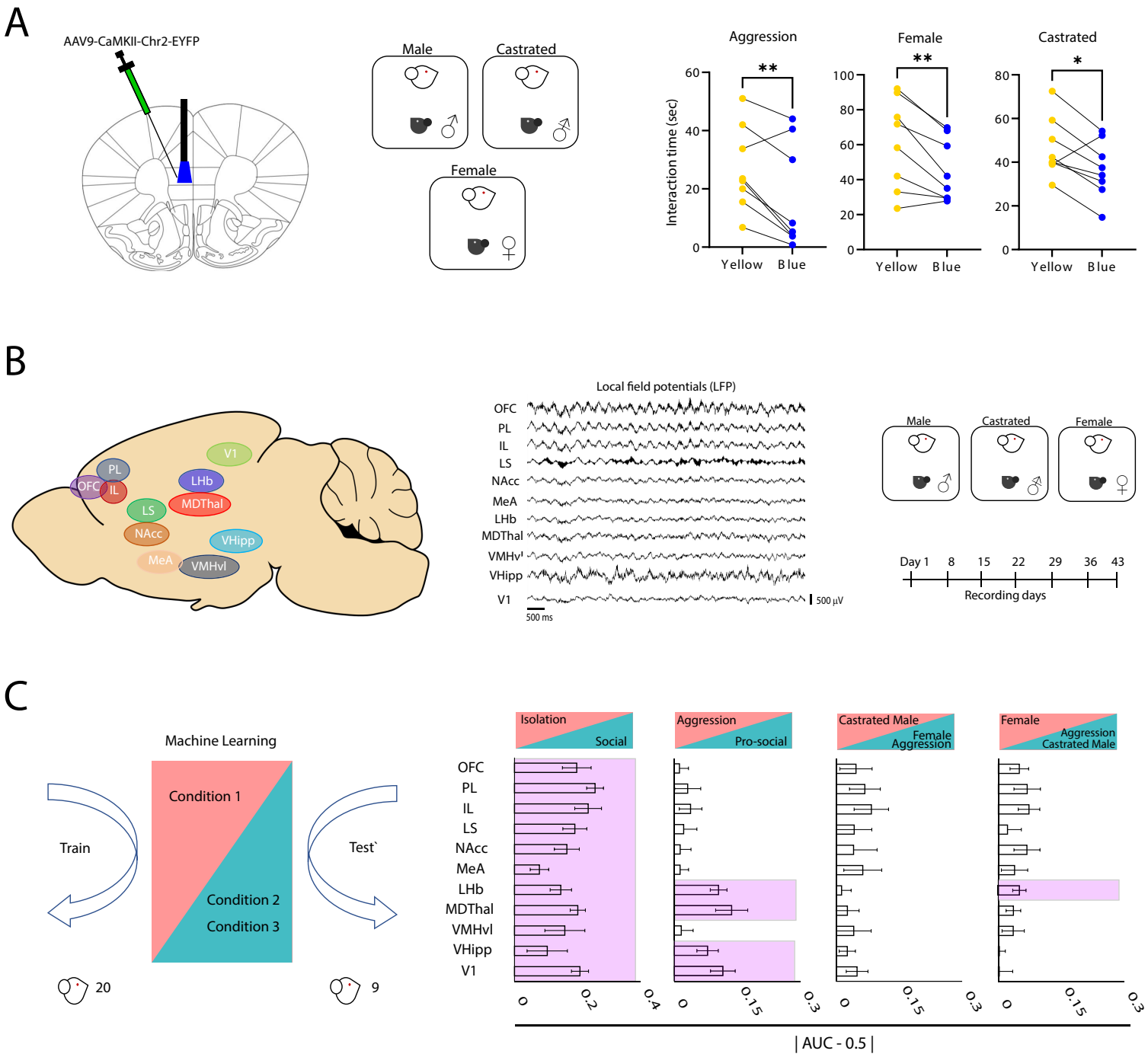

A
AAV9-CaMKII-Chr2-EYFP
Male
Castrated
Female
Aggression
Female
Castrated
Interaction time (sec)
Yellow
Blue
B
Local field potentials (LFP)
OFC
PL
IL
LS
NAcc
MeA
LHb
MDThal
VMHvl
VHipp
V1
500 ms
500 µV
Day 1
8
15
22
29
36
43
Recording days
C
Machine Learning
Condition 1
Condition 2
Condition 3
Train
Test`
20
9
Isolation
Social
Aggression
Pro-social
Castrated Male
Female Aggression
Female
Aggression Castrated Male
| AUC - 0.5 |

Figure 1

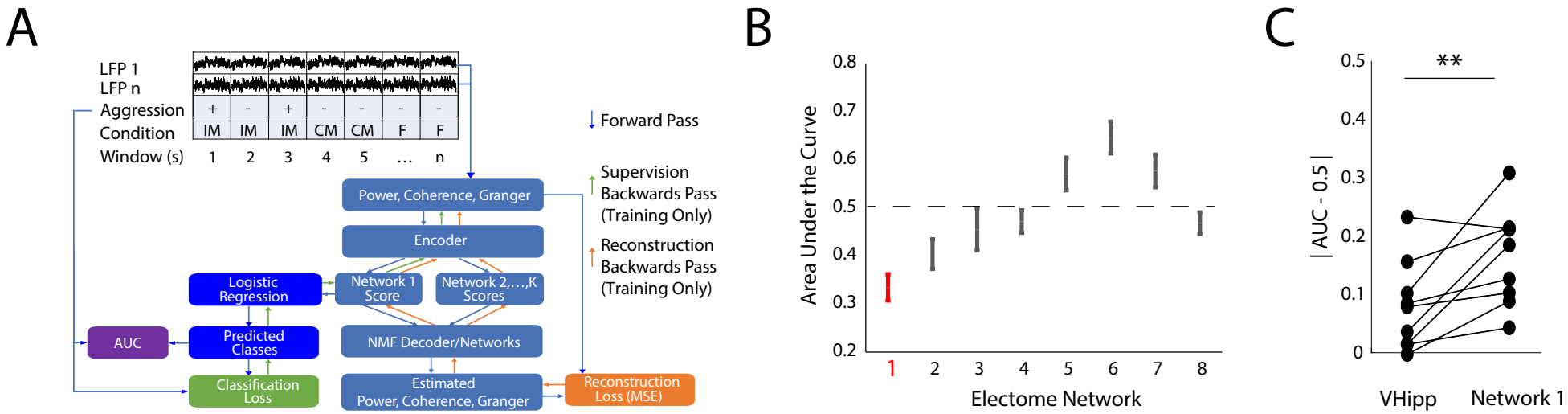

A
LFP 1
LFP n
Aggression
Condition
Window (s)
+ - + - - - -
IM IM IM CM CM F F
1 2 3 4 5 … n
Forward Pass
Supervision Backwards Pass (Training Only)
Reconstruction Backwards Pass (Training Only)
Power, Coherence, Granger
Encoder
Logistic Regression
Network 1 Score
Network 2,…,K Scores
AUC
Predicted Classes
NMF Decoder/Networks
Classification Loss
Estimated Power, Coherence, Granger
Reconstruction Loss (MSE)
B
Area Under the Curve
0.8 0.7 0.6 0.5 0.4 0.3 0.2
1 2 3 4 5 6 7 8
Electome Network
C
| AUC - 0.5 |
0.5 0.4 0.3 0.2 0.1 0
**
VHipp
Network 1

Figure 2

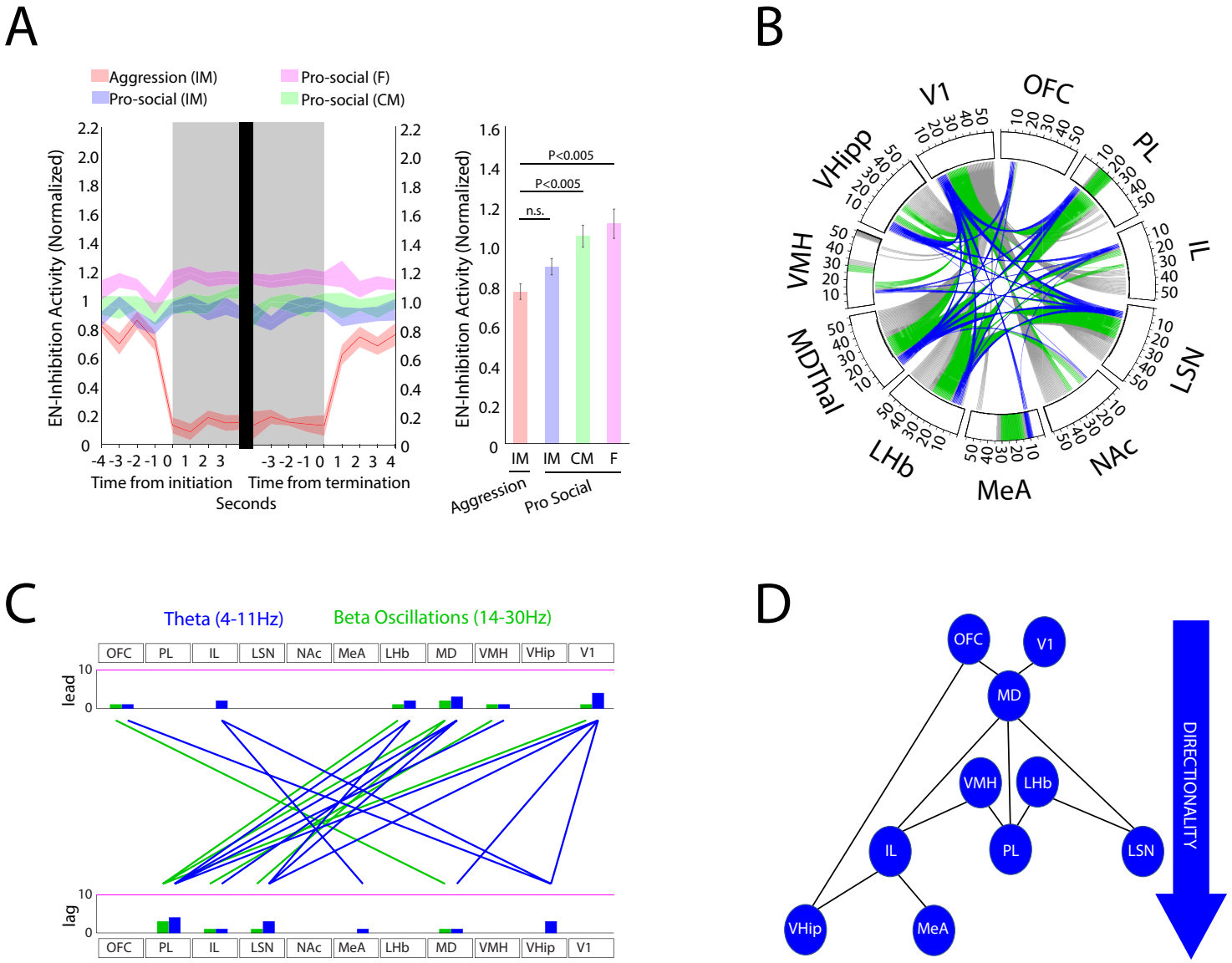
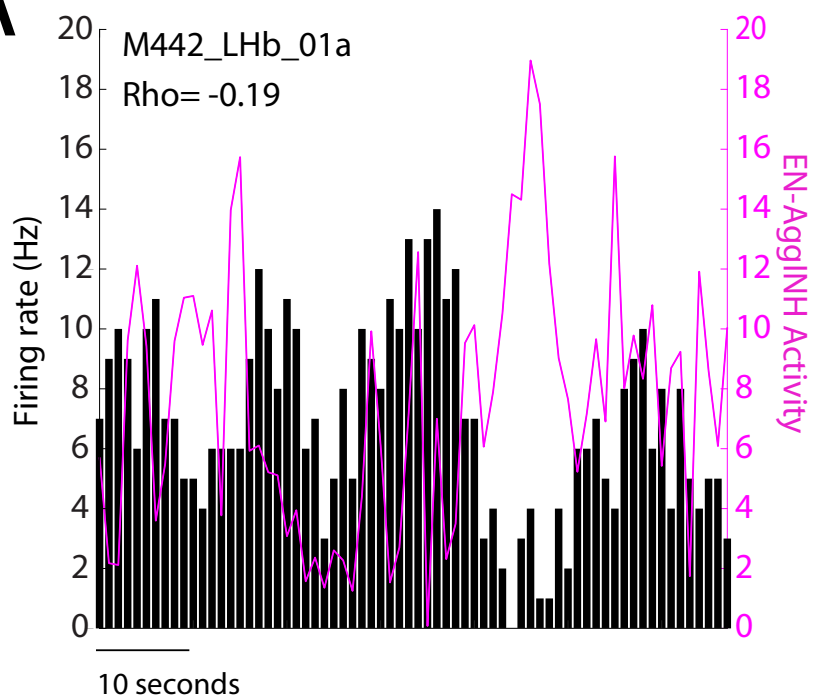

A
Aggression (IM)
Pro-social (IM)
Pro-social (F)
Pro-social (CM)
EN-Inhibition Activity (Normalized)
Time from initiation
Time from termination
Seconds
P<0.005
P<0.005
n.s.
IM
IM
CM
F
Aggression
Pro Social
B
V1
OFC
PL
IL
LSN
NAc
MeA
LHb
MDThal
VMH
VHipp
C
Theta (4-11Hz)
Beta Oscillations (14-30Hz)
OFC
PL
IL
LSN
NAc
MeA
LHb
MD
VMH
VHip
V1
lead
lag
D
OFC
V1
MD
VMH
LHb
IL
PL
LSN
VHip
MeA
DIRECTIONALITY

Figure 3

A



B

	Correlated	Uncorrelated
LHB	2	27
LSN	1	26
MD	12	19
MeA	3	8
Nac	3	16
IL	0	5
PL	0	7
OFC	3	6
V1	1	5
Vhip	2	2
VMH	7	31
Total	34	152

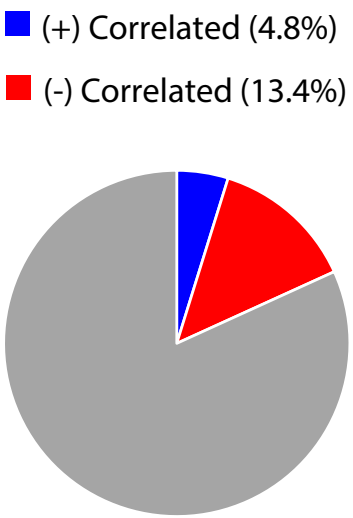


Figure 4

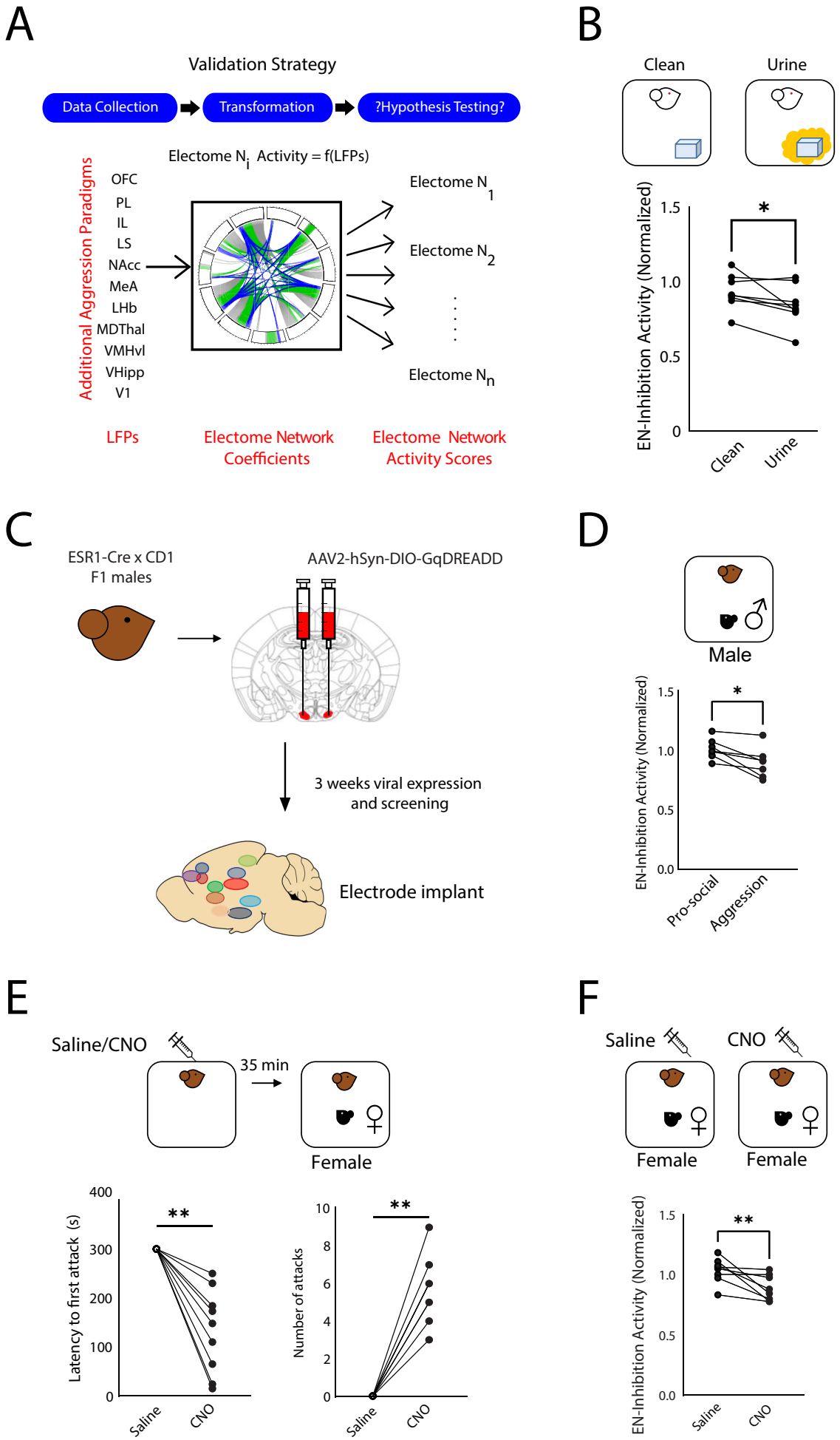

A
Validation Strategy
Data Collection
Transformation
?Hypothesis Testing?
Electome N_i Activity = f(LFPs)
Additional Aggression Paradigms
OFC
PL
IL
LS
NAcc
MeA
LHb
MDThal
VMHvl
VHipp
V1
Electome N_1
Electome N_2
Electome N_n
LFPs
Electome Network Coefficients
Electome Network Activity Scores
B
Clean
Urine
EN-Inhibition Activity (Normalized)
1.5
1.0
0.5
0
*
Clean
Urine
C
ESR1-Cre x CD1 F1 males
AAV2-hSyn-DIO-GqDREADD
3 weeks viral expression and screening
Electrode implant
D
Male
EN-Inhibition Activity (Normalized)
1.5
1.0
0.5
0.0
*
Pro-social
Aggression
E
Saline/CNO
35 min
Female
Latency to first attack (s)
400
300
200
100
0
**
Saline
CNO
Number of attacks
10
8
6
4
2
0
**
Saline
CNO
F
Saline
CNO
Female
Female
EN-Inhibition Activity (Normalized)
1.5
1.0
0.5
0.0
**
Saline
CNO

Figure 5

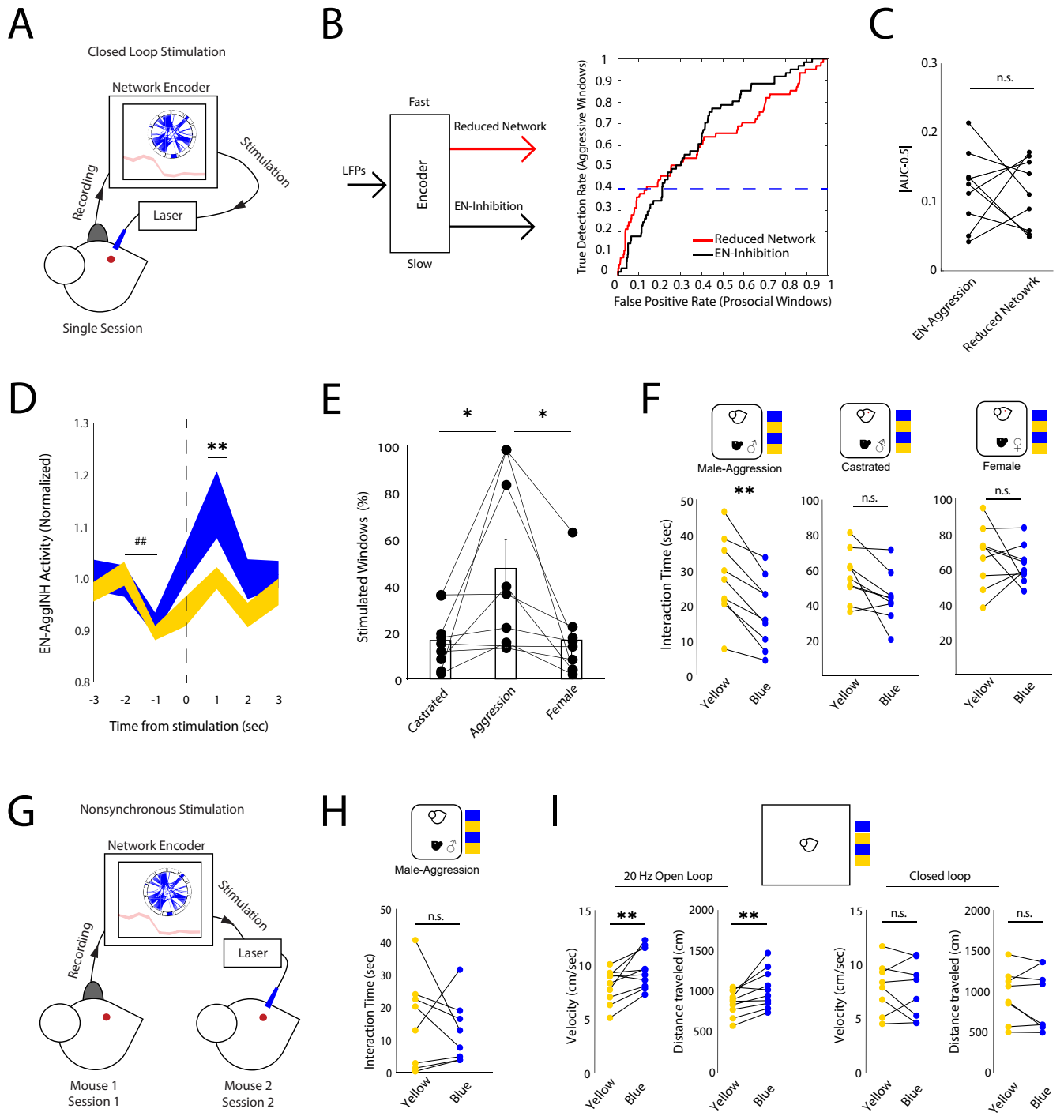

A
Closed Loop Stimulation
Network Encoder
Recording
Stimulation
Laser
Single Session
B
Fast
LFPs
Encoder
Reduced Network
EN-Inhibition
Slow
True Detection Rate (Aggressive Windows)
False Positive Rate (Prosocial Windows)
Reduced Network
EN-Inhibition
C
n.s.
|AUC-0.5|
EN-Aggression
Reduced Netowrk
D
EN-AggINH Activity (Normalized)
Time from stimulation (sec)
##
**
E
Stimulated Windows (%)
Castrated
Aggression
Female
*
*
F
Male-Aggression
Castrated
Female
Interaction Time (sec)
Yellow
Blue
**
n.s.
n.s.
G
Nonsynchronous Stimulation
Network Encoder
Recording
Stimulation
Laser
Mouse 1
Session 1
Mouse 2
Session 2
H
Male-Aggression
Interaction Time (sec)
n.s.
Yellow
Blue
I
20 Hz Open Loop
Closed loop
Velocity (cm/sec)
Distance traveled (cm)
**
**
n.s.
n.s.
Yellow
Blue

Figure 6