A widespread oscillatory network encodes an aggressive internal state

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 J. 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

Keywords

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

Abstract

Social aggression is an innate behavior that can aid an organism in securing access to resources[1], or it can impair group function and survival in behavioral pathology[2-4]. Since many brain regions contribute to multiple social behaviors[5-7], expanded knowledge of how the brain distinguishes between social states would enable the development of interventions that suppress aggression, while leaving other social behaviors intact. Here we show that a murine aggressive internal state is encoded by a widespread network. This network is organized by prominent and synchronized theta (4-11Hz) and beta (14-30Hz) oscillations that relay through the prefrontal cortex, and couples to widespread cellular firing. Strikingly, network activity during social isolation encodes the trait aggressiveness of mice, and causal cellular manipulations known to impact aggression can bidirectionally regulate the network's activity. Finally, we use closed-loop stimulation of prefrontal cortex and causal mediation analysis to establish that the network is a mediator of aggressive behavior. Thus, we define a widespread network that encodes an aggressive internal state within and across mice.

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) [8, 9], nucleus accumbens [2, 10], lateral habenula [11, 12], the ventrolateral portion of ventromedial hypothalamus [5, 13-17], and medial amygdala [3, 18]. Prefrontal cortex stimulation has been shown to mitigate aggressive behavior in both humans [19, 20] and rodents [21], implicating cortical regions in regulating aggression. Finally, sensory regions, such as those underlying olfaction, also contribute to aggression [22].

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 [5, 23], the brain must ultimately utilize information from overlapping regions to segregate aggression from other social behavioral states. Though efforts have revealed several cellular-level processes within one of these regions that contribute to this mechanism [13, 16, 24], the complementary network level process that

integrates information across regions to distinguish aggressive states from prosocial states remains unknown. Addressing this knowledge gap is of major importance as 1) mammals regularly select from a repertoire of social behaviors based on external sensory cues to ensure their survival [6, 25], and 2) a range of psychiatric disorders are broadly marked by a failure to appropriately match behavior with evolving social contexts [26].

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 attack behavior and non-attack social behavior. After confirming that statistical models based on single brain regions could independently differentiate attack behavior from non-attack social behaviors, we asked whether this code was also reflected at a network level, where millisecond-timescale information was integrated across all the 11 brain areas. For this analysis, we used a machine learning approach that models variations in natural patterns of activity within and between implanted brain regions across seconds of time (a timescale that we reasoned would allow us to capture socially-relevant internal states)[27]. Our approach also tuned the model to optimally encode attack vs. non-attack social behavior.

We then applied multiple levels of validation of increasing generalization to our single-region and network-level models [28]. 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 third level of validation. 4) Next, we tested whether these two models encoded an aggressive state, more broadly, and not simply attack behavior. Only the network model encoded the social context of mice during periods immediately prior to or following aggressive and prosocial interactions. Moreover, network activity prior to and following attacks correlated with trait aggressiveness on a mouse-by-mouse basis. 5) Finally, we subjected the network to two high levels of experimental validation, each based on causal cellular manipulations, that bidirectionally impact attack behavior. Specifically, we developed a close-loop optogenetic stimulation approach that

93 detected when the brain transitioned into an aggressive state, on a second-by-second basis,
94 and stimulated the prefrontal cortex in a manner previously known to suppress attack behavior.
95 We also employed a designer receptor exclusively activated by designer drug (DREADD)-based
96 approach to selectively activate cells in ventral medial hypothalamus in a manner that has been
97 shown to induce attack behavior. These manipulations bidirectionally impacted network
98 activity. 6) As this latter manipulation was also performed in mice on a different genetic
99 background, we further confirmed the generalizability of our network-level findings. Thus, we
100 firmly establish a network-level architecture whereby the brain encodes an aggressive state in a
101 manner that generalizes across context and individuals.

102 Direct stimulation of prefrontal cortex broadly suppresses social behaviors

103 We initially focused our efforts to selectively suppress aggressive behavior on modulating the
104 medial prefrontal cortex (i.e., prelimbic and infralimbic cortex in mice), since this brain region
105 had been implicated in social behaviors [29-31]. Prior work had shown that optogenetic
106 stimulation of prefrontal cortex was sufficient to suppress attack behavior and increase non-
107 aggressive social behaviors in CD1 strain mice [21]. Further highlighting the translational
108 potential of such an approach, clinical studies had shown that direct transcranial stimulation of
109 prefrontal cortex decreased aggressive feelings in violent offenders [32], and in individuals with
110 methamphetamine use disorder [33].

111 We performed our optogenetic stimulation experiments during these social encounters using a
112 protocol modeled after prior work, where blue light stimulation is used to activate
113 channelrhodopsin 2 in the medial prefrontal cortex for the entirety of a social encounter
114 (473nm, 5mW, 20Hz, 3ms pulse width, Fig. 1a) [21]. Control experiments were performed using
115 yellow light stimulation (593.5nm, 5mW, 20Hz, 3ms pulse width), which does not activate
116 channelrhodopsin 2[34, 35]. CD1 mice show periods of attack behavior, defined by biting,
117 boxing (kicking/clawing), or tussling behavior, when a male C57BL6/J (C57) mouse is introduced
118 into their home cage [36]. On the other hand, they exhibit periods of non-attack social
119 interactions such as sniffing, grooming, or resting (placing nose or forepaws against the subject
120 mouse, but not moving) when exposed to a female intruder [16]. Consistent with the prior

121 report, medial prefrontal cortex stimulation suppressed attack behavior towards C57 male mice
122 and tended to increase non-attack social behavior [N=8; t_7=3.43; P=0.0055, and t_7=-2.35;
123 P=0.051 using a two-tailed paired t-test, for attack and non-attack behavior, respectively,
124 significance determined by a Benjamini-Hochberg false discovery rate (FDR) correction, Fig. 1a].
125 To ensure that this suppression of aggression was selective, we also tested mice in a second
126 social paradigm in which they interacted with a female. Importantly, the CD1 mice do not show-
127 attack behavior during this social context. Here, we found that cortical stimulation decreased
128 non-attack social behavior toward female intruders (t_7=3.647, P=0.0041, using paired t-test,
129 significance determined by FDR correction). Thus, medial prefrontal cortex stimulation was
130 unable to selectively suppress attack behavior. Rather, stimulation suppressed multiple types of
131 social behavior.

132 Multiple brain regions fail to independently encode attack behavior across mice and contexts

133 After failing to selectively suppress CD1 aggressive social behavioral by targeting the prefrontal
134 cortex, we set out to probe for brain regions that may exhibit such selectivity. We implanted
135 CD1 mice across multiple cortical and subcortical brain regions known to contribute to social
136 behavior, including infralimbic cortex[21], orbitofrontal cortex[37], prelimbic cortex[21], lateral
137 septum [8, 9], nucleus accumbens [2, 10], lateral habenula [11, 12], mediodorsal thalamus[38],
138 ventromedial hypothalamus [5, 13-17], medial amygdala [3, 18], ventral hippocampus [39], and
139 primary visual cortex.

140 Following surgical recovery, we recorded neural activity while the CD1 mice freely interacted
141 with an intact male C57 mouse and a female C57 mouse for 300 seconds each. We repeated
142 these encounters over six sessions, yielding a total of 1800 seconds of neural data and behavior
143 for each exposure (Fig. 1b, see also Supplemental Fig. S1). We recorded a subset of these CD1
144 mice (N=9 of the 20 mice) as they interacted with a castrated male mouse intruder. Since CD1
145 mice do not generally exhibit attack behavior towards castrated males [40], this encounter
146 provided neurophysiology data during additional non-attack social behaviors that were
147 unpaired with female sensory cues. We also acquired neural activity while the CD1 mice were
148 isolated in their home cage.

149 We first verified that each of the implanted brain regions encoded social behavior using
150 discriminative cross spectral factor analysis non-negative matrix factorization (dCSFA-NMF, see
151 Fig. 1e) [41]. dCSFA-NMF utilizes supervised machine learning to generate a statistical model
152 that is both descriptive [integrates brain local field potential (LFP) activity features across time]
153 and predictive (discovers networks that distinguish between types of external behavior). LFPs
154 reflect the activity of populations of neurons, and these signals can be consistently sampled
155 across mice. The electrical functional connectome networks (*Electome Networks*) generated
156 from dCSFA-NMF integrate LFP power (oscillatory amplitude across 1-56 Hz; a correlate of
157 cellular and synaptic activity within brain regions), LFP synchrony (how two regions' LFP
158 frequencies synchronize across time; a correlate of brain circuit function), and LFP Granger
159 synchrony (Granger causality testing; a correlate of directional transfer of information across a
160 brain circuit). Furthermore, dCSFA-NMF generates electome network activity scores (an
161 indicator of the strength of each network) at a temporal resolution sufficient to capture brain
162 states underlying the external behavior under observation (in this instance, a resolution of one
163 second). The electome networks are designed to learn patterns that explain interpretable
164 correlates of neural activity whose expression relate to measured behavior, facilitating an
165 overall interpretable model [28]. Any given brain region can belong to multiple electome
166 networks, and each electome network may incorporate any number of brain regions. dCSFA-
167 NMF thus integrates spatially distinct brain regions and circuits into electome networks that
168 encode behavior.

169 To explore whether there was a generalized activity pattern within individual regions that
170 encoded social behavior, we designed a series of dCSFA-NMF models based on LFP oscillatory
171 power in frequencies from 1-56Hz. Each single-region model was trained using observations
172 pooled from 20 CD1 mice to separate periods when mice were socially isolated from periods
173 when they were engaged in social behavior (e.g., attack behavior towards the intact males and
174 non-attack social behavior towards male and female mice). We trained our models with one
175 supervised network to discover the patterns of LFP activity that encoded social behavior. We
176 also trained three unsupervised networks to account for the remaining variance in neural
177 activity that was not directly related to social behavior (see methods for justification of

178 hyperparameter selection). We tested the accuracy of the models in new CD1 mice recorded
179 under all three conditions (N=9 mice).

180 We observed high accuracy in decoding periods of isolation from social behavior, using the
181 supervised network, for each implanted brain area ($p<0.05$ using Wilcoxon rank-sum,
182 significance determined by FDR correction for 44 comparisons, Fig. 1c). Next, we designed a
183 new series of models to separate one of the three social behavioral states from the other three
184 (e.g., attack behavior toward intact males vs. non-attack interactions with female, intact male,
185 and castrated male mice). These models were built using observations pooled from the same
186 20 CD1 mice, and again based on LFP power. We then tested the accuracy of the supervised
187 networks in the models for the same nine hold-out mice. Thus, we trained and tested the
188 model's generalizability to decode each class of social behavior from the other three for each of
189 the 11 implanted brain areas (i.e., 33 additional models). Using this approach, we found that
190 five of the brain region-based statistical models decoded attack behavior versus non-attack
191 social behavior: infralimbic cortex, lateral habenula, ventral hippocampus, medial amygdala,
192 and medial dorsal thalamus. V1 successfully decoded the non-attack male interaction from the
193 other social conditions as well ($P<0.05$ using one-tailed Wilcoxon rank-sum test, significance
194 determined by FDR correction for 44 comparisons, see Fig. 1c). None of the other implanted
195 brain regions showed this selectivity. Thus, five regions independently encoded attack behavior
196 vs. non-attack social behavior in a manner that generalized across mice.

197 Our broad goal was to identify a neural signature that could be used to suppress aggression
198 while leaving other social behaviors intact. Thus, we tested whether the putative aggression
199 codes we discovered for the five regions generalized to another context associated with
200 aggression. Specifically, urine from other male mice has been found to elicit aggressive and
201 dominance behavior in CD1 males [22, 42, 43]. As such, the most aggressive mice from the
202 training and testing groups (N=8) were allowed to freely explore a clean inanimate object or an
203 object covered in urine from another intact CD1 male mouse (seven sessions). We then tested
204 whether each of the five regions' putative aggression codes could distinguish periods where
205 mice explored the clean objects from those where mice explored objects covered in urine.
206 Though only ventral hippocampus model tended to decode behavior in this new context, none

207 of the brain regions showed statistically significant encoding at the mesoscale-level (LFP)
208 following multiplicity correction (AUC=0.56±0.06, P=0.04 for ventral hippocampus, using one-
209 tailed Wilcoxon rank-sum test, significance determined by FDR correction for 5 comparisons,
210 Fig. 1d).

211 An aggressive state is encoded at the network-level

212 After failing to robustly decode attack behavior using LFPs independently from any of the 11
213 brain regions, we established that a brain network integrated information across all the
214 implanted brain regions to encode an aggressive state. This network-level encoding mechanism
215 generalized to multiple new contexts associated with aggression. Critically, the network also
216 encoded attack behavior with a predictive efficacy that exceeded independent ventral
217 hippocampus activity.

218 For this analysis, we trained a new dCFSA-NMF model using data from all the implanted brain
219 regions. This model utilized LFP power for each region, and the coherence and Granger
220 directionality between them. The model utilized one supervised network that was trained to
221 encode periods of attack-behavior (positive class) vs. social behavior in castrated male and
222 female social context (negative class). We also included non-attack social behavior towards
223 intact males in the negative class to discourage the network from simply learning non-
224 aggressive sensory cues specific to the intact male (i.e., supervision, electome network #1; see
225 Fig. 1e). Based on our hyperparameter selection approach using the Bayesian Information
226 Criteria (see methods and Supplemental Fig. S2), seven additional unsupervised networks were
227 trained to account for the variance in neural activity that was not related to attack vs. non-
228 attack social behavior. We then validated our supervised network using the set of nine holdout
229 CD1 mice from our single area coding test analysis. Again, none of these mice were used to
230 train the electome networks. We found that the supervised network (network #1) successfully
231 discriminated between attack behavior and non-attack social behavior in the test mice (N=9,
232 Wilcoxon signed rank, p=0.0020, Fig. 1f). Incidentally, we also observed that one of the
233 unsupervised networks (Network #6) showed strong encoding in the nine hold-out test mice
234 (AUC=0.65±0.03, $P<5\times10^{-5}$ using Wilcoxon rank sum).

235 Attack behavior is indicative of an aggressive brain state. We also reasoned that it was possible
236 for a mouse to be in an aggressive internal state, even though it was not actively exhibiting
237 attack behavior. Since such a context was likely to be present immediately prior to or following
238 attack behavior, we tested whether network activity pooled from the 3 seconds preceding and
239 3 seconds following social behaviors encoded the distinct social conditions (Fig. 2a). Critically,
240 these data windows were not used to train the network model since they did not contain attack
241 or non-attack social behavior. Activity of the supervised network (network #1) was lower in the
242 intervals surrounding attack behavior compared to non-attack social interactions with males or
243 females ($F_{3,67}$=28.6, P<0.0001 using Friedman's test, followed by P<0.05 using Wilcoxon signed-
244 rank test, significance determined by FDR correction, Fig. 2b, top left). Strikingly, network #1
245 activity during periods of isolation also negatively correlated with the time mice spent
246 exhibiting attack behavior towards other males (R=-0.58; P=0.016 using Spearman's rank
247 correlation, Fig. 2b, top right), encoding aggression on a mouse-by-mouse basis. Thus, electome
248 network #1 [hereafter referred to as *EN-Aggression Inhibition (EN-AggINH*)] represented a
249 network that putatively inhibited aggression when its activity was highest. In contrast to the
250 models developed for each of the brain regions independently, *EN-AggINH* activity also
251 encoded the exposure to male urine (N=8, Network Activity = 9.1±1.0 and 8.3±1.1 for clean and
252 urine covered objects, respectively; P=0.012 using one tailed Wilcoxon signed-rank test;
253 AUC=0.57±0.02, data not shown). Network #6 activity failed to generalize to this urine context
254 (AUC=0.48±0.01, data not shown). Thus, only *EN-AggINH* generalized to this second aggression
255 context.

256 Next, we tested the model for ventral hippocampus since we observed a trend towards
257 decoding attack behavior in the urine context. This model failed to encode the aggressive state.
258 Specifically, activity from ventral hippocampus was statistically indistinguishable between
259 periods surrounding attack behavior and non-attack social interactions with females and
260 activity during social isolation ($F_{3,67}$=7.9, P=0.047 using Friedman's test; P=0.36 and 0.72,
261 respectively, using Wilcoxon signed-rank test, significance determined by FDR correction, Fig.
262 2b, bottom left). Moreover, there was no relationship between ventral hippocampal activity

263 during isolation and the innate aggressiveness of mice (P=0.60 using Spearman's rank, Fig. 2b,
264 bottom right). Thus, only network activity encoded the aggressive internal state.

265 *EN-Aggression Inhibition* activity couples to cellular firing

266 *EN-AggINH* was composed of prominent theta frequency activity (4-11 Hz) in medial amygdala
267 and beta frequency activity (14-30 Hz) in medial amygdala and prelimbic cortex (Fig. 3a-b).
268 Prominent synchrony was also observed in the theta and beta frequency bands. Indeed, when
269 we quantified directionality across these synchronized bands, we saw that activity flowed from
270 orbital frontal cortex and primary visual cortex, relayed through medial dorsal thalamus, and
271 infralimbic cortex, and flowed into medial amygdala and ventral hippocampus (Fig. 3c-d, and
272 Supplemental Fig. S3).

273 We verified that the activity of *EN-AggINH* truly reflects biological activity, by relating the
274 electome network to neural firing, as in previous work [44]. To achieve this, we analyzed the
275 correlation between cellular activity across the implanted brain regions and the activity of *EN-*
276 *AggINH*, as cell activity is an undisputed measure of biological function. We then used a
277 permutation test to rigorously test our findings (Fig. 3e). Specifically, we shuffled cellular firing
278 within social behavioral conditions, maintaining the relationship between cell firing and
279 behavior. We then repeated this procedure 1000 times to generate a null distribution for which
280 only 5% of cells would be expected to exhibit firing coupled to network activity. We found that
281 ~18% of cells showed firing that was coupled to the activity of *EN-AggINH*, far more than could
282 be explained by chance (χ^2=16.4, p=0.00005). Specifically, of the 186 cells recorded, nine
283 (4.8%) showed firing activity that was positively correlated with *EN-AggINH* and 25 (13.4%)
284 showed activity that was negatively correlated (Fig. 3f). Thus, most cells that showed coupling
285 to *EN-AggINH* were inhibited when network activity increased. These analyses confirmed that
286 *EN-AggINH* activity reflects the dynamics of cellular activity across the brain.

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

288 To further validate *EN-AggINH*, we established that activity in this network was modulated by
289 orthogonal biological conditions that have been shown to induce or suppress aggressive

behavior in mice. In most cases, we performed this analysis in new animals, which is considered the gold standard of model validation in machine learning [45]. We transformed LFP data recorded from these new sessions into our original network model.

For our first gold-standard validation experiment, we tested whether our network generalized to new mice on a different genetic background engaging in a new aggression context (Fig. 4a-b). Specifically, this approach also used a validated cellular manipulation that causally induces aggression under a behavioral condition that would otherwise not yield aggressive attack behavior (i.e., we used female social partners). 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 [17, 46, 47]. 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, prior to the repeated testing sessions.

As anticipated, treatment with CNO induced attack behavior towards the female mice (N=8; P= 0.0039 for both attack latency and attack number using one-tailed Wilcoxon sign rank; Fig. 4c). When we probed neural activity across the entire exposure to the female intruder, we found that treatment with CNO also suppressed *EN-AggINH* activity (N=8, P=0.0039 using one-tailed Wilcoxon sign-rank; Fig. 4d). Thus, the network model generalized to a second aggression context induced by a cellular manipulation, and was robust to different genetic backgrounds. Critically, network activity was also lower during the time intervals surrounding attack/non-attack social behavior for the CNO vs. saline treatment sessions (P=0.02 using one-tailed Wilcoxon sign-rank; Fig. 4d), again demonstrating that *EN-AggINH* encoded an aggressive state. Our observations also established that *EN-AggINH* does not simply encode sensory cues associated with male intruders, since the network responses observed in the CNO treated mice were induced by a female intruder.

319 *EN-Aggression Inhibition* mediates attack behavior

320 We used mediation analysis to determine whether *EN-AggINH* activity putatively played a
321 mechanistic role in suppressing attack behavior. Mediation analysis is a framework to
322 determine whether the impact of a "treatment" (manipulation) on an outcome (attack
323 behavior) is mediated by a change in an intermediate variable (*EN-AggINH* activity). If so, the
324 intermediate variable is viewed, at least in part, as a mechanistic route (a mediator) for how the
325 treatment impacts the outcome. Three components were necessary to optimally implement
326 test our mediation analysis models: a manipulation that causally modulated 1) attack behavior
327 and 2) *EN-AggINH* activity, and 3) an approach to deliver the manipulation during levels of *EN-*
328 *AggINH* activity that would predict the emergence of attack behavior. We chose to build such
329 an approach based on prefrontal cortex optogenetic stimulation, since we had previously found
330 that such a manipulation causally suppressed attack behavior [21].

331 Specifically, we set out to preferentially stimulate medial prefrontal cortex when *EN-AggINH*
332 was naturally suppressed in the brain (signaling the onset of attack behavior). First, we built a
333 closed-loop system that estimated the activity of *EN-AggINH* in real time (i.e., within 200ms,
334 Fig. 5a). This approach employed a new network encoded solely based on power and coherence
335 measures (i.e., a reduced network, Fig. 5b), because the processing time to calculate Granger
336 directionality was prohibitive for real-time implementations. While this new network lacked the
337 interpretive power of dCSFA-NMF, it enabled us to predict attack behavior in real time (Fig. 5c).
338 In principle, when the activity of *EN-AggINH* fell below an established threshold (signaling the
339 onset of attack behavior), our closed-loop approach would deliver a one-second light
340 stimulation (5mW, 20Hz, 3ms pulse width) to prefrontal cortex. To verify that this real-time
341 estimation system worked as designed, we tested whether light stimulation was triggered by a
342 decrease in *EN-AggINH* activity. Indeed, network activity was significantly lower one second
343 prior to stimulation than it was two seconds prior to stimulation, demonstrating that our
344 approach successfully identified when the *EN-AggINH* activity decreased below the threshold
345 that signaled the onset of aggression (N=9; $P<0.005$ for within-subject comparison of *EN-*
346 *AggINH* activity 1 vs. 2 seconds prior to yellow light stimulation using one-tailed signed-rank
347 test, Fig. 5d). Importantly, we found that prefrontal cortex stimulation acutely increased *EN-*

348 *AggINH* activity (N=9, P<0.01 for comparison of *EN-AggINH* activity one second after blue vs.
349 yellow stimulation, using one-tailed signed-rank test, see Fig. 5d). Thus, our closed-loop
350 stimulation approach satisfied two of the components needed to implement our mediation
351 approach. Next, we tested whether increasing *EN-AggINH* activity via prefrontal cortex
352 stimulation as the brain transitioned into a putative attack state would suppress aggressive
353 behavior. We found our closed-loop stimulation approach significantly suppressed attack
354 behavior (see Fig. 6a; N=9 mice that were not used to train the initial model; t_8=6.1, P=0.0003,
355 comparing blue vs. yellow light stimulation using two-tailed paired t-test for attack behavior,
356 significance determined by FDR correction). Thus, our closed-loop manipulation suppressed
357 attack behavior, satisfying the remaining component needed to implement our mediation
358 analysis approach.

359 We first used the classic Baron and Kenny approach [48] to determine whether *EN-AggINH*
360 activity mediates the effect of neurostimulation on aggressive behavior. According to this
361 statistical approach, there is a mediated effect of network activity on behavior if three
362 conditions are met: 1) stimulation modulates network activity, 2) network activity correlates
363 with behavior, and 3) modeling the behavior from network activity and stimulation together is
364 better than modeling behavior from stimulation alone. Indeed, we had identified a significant
365 direct effect of stimulation on attack behavior (P<0.005, see Fig. 6a) and network activity
366 (P<0.0005, Fig. 5d). To optimally match the conditions between the treatment and control
367 cases, we used windows during the closed-loop stimulation procedure where the laser was
368 triggered, and then compared blue laser stimulation (treatment) to yellow laser stimulation
369 (control). Thus, the data points used for our mediation analysis predicted imminent or ongoing
370 attack behavior, and network activity prior to the stimulation in both the control (yellow light)
371 and treatment (blue light) case were similar. A statistical model of behavior using network
372 activity and stimulation (see Fig. 5e, model 2) significantly outperformed the model using only
373 stimulation (see Fig. 5e, model 1; nested logistic regression models, P<0.01, likelihood ratio
374 test), satisfying the necessary conditions to show that *EN-AggINH* is a mediator.

375 After establishing that *EN-AggINH* activity mediated the impact of PFC stimulation on behavior,
376 we set out to evaluate the significance of the average causal mediation effect (ACME) and the

377 average direct effect (ADE) during the same stimulated closed-loop windows using causal
378 mediation analysis [49]. ACME is the causal effect of stimulation on behavior due to the change
379 in *EN-AggINH* activity (see. Fig 5e, model 3), and ADE is the causal effect on behavior from
380 prefrontal cortex stimulation not explained by the change in *EN-AggINH* activity. We found that
381 there was a significant ACME (P<0.01), but not a significant ADE (P=0.48). This analysis
382 suggested that *EN-AggINH* activation is the primary mechanism whereby prefrontal cortex
383 stimulation suppresses aggression.

384 Next, we tested models where *EN-AggINH* activity functioned as a biomarker, rather than a
385 mediator of attack behavior. In these models, the manipulation modifies another neural
386 process, which in turn, simultaneously impacts attack behavior and *EN-AggINH* activity (Fig. 5g,
387 model 5). First, we evaluated whether theta power in 11 different brain regions could serve as a
388 mediator in lieu of *EN-AggINH* activity (Fig. 5f, model 4). We chose this frequency band since it
389 was prominently featured in *EN-AggINH* and within the network we previously found to encode
390 social appetitive behavior[7]. Across these 11 models, only orbitofrontal cortex had a significant
391 average causal mediation effect (*uncorrected* p-value of 0.038, see Fig. 5f). Critically, this model
392 did not survive a correction for multiple comparisons, and its ACME estimate was dwarfed by
393 the size of the ACME estimate for *EN-AggINH* (the estimate for the EN-AggINH model was
394 49.7% larger). This evidence suggests that EN-AggINH is a much better mediator than any of
395 these other potential 'biomarkers' by themselves.

396 After failing to identify any significant mediation effect of theta activity within each of the
397 eleven brain regions, we tested whether including theta activity as an intermediary in our
398 causal graph would disrupt *EN-AggINH's* role as a mediator in attack behavior (Fig. 5g, model 5).
399 Here, we corrected for the role of theta power in the model of how EN-AggINH changes as a
400 function of stimulation, as well as correcting for theta power in forecasting attack behavior. As
401 such, this framework dictates that *EN-AggINH* cannot mediate behavior that is already
402 explained by changes in theta power in a specified region. When we ran eleven models, one
403 model for each brain region, we found that *EN-AggINH* still significantly mediated attack
404 behavior in all of them (P<0.05 for all models; see Fig. 5g, bottom). Thus, even after accounting

405 for these potential intermediate variables, our findings still supported *EN-AggINH* as a mediator
406 of attack behavior.

407 Validation of temporal activity and spatial spectral features of *EN-AggINH*

408 We validated the temporal activity and spatial spectral features of *EN-AggINH* by establishing
409 that they could be utilized to selectively suppress aggression. Specifically, after determining
410 that our closed-loop manipulation suppressed aggression, we also quantified the impact of this
411 stimulation protocol on non-aggressive interactions with other male and female mice. We
412 found that closed-loop PFC stimulation increased non-attack behavior towards the intact C57
413 males [N=9; t_8=-2.3, P=0.049 comparing blue vs. yellow light stimulation using two-tailed paired
414 t-test for attack behavior and non-attack behavior, significance determined by FDR correction,
415 Fig. 6a). No differences in non-attack social behavior were observed during exposure to female
416 mice (t_8=0.74, P=0.48 using two-tailed paired t-test, significance determined by FDR correction,
417 Fig. 6a). Thus, closed-loop PFC stimulation selectively reduced aggression.

418 To verify that this selective modulation of aggression was due to synchronization of the light
419 stimulation with endogenous *EN-AggINH* activity, and not simply due to the dynamic pattern of
420 stimulation delivered using this method, we performed an additional control experiment where
421 we used the stimulation patterns from our closed-loop experiments to drive stimulation in a
422 new group of animals (e.g., randomly copying patterns from another mouses brain, analogous
423 to a "sham" in neurofeedback experiments). Thus, for these sessions, prefrontal cortex
424 stimulation occurred in a manner that mirrored our closed-loop stimulation experiments,
425 except that stimulation was not fixed to endogenous *EN-AggINH* activity (i.e., open loop –
426 nonsynchronous; Fig. 6b). Nonsynchronous stimulation failed to suppress aggressive behavior
427 ($F_{1,21}$=4.87, P =0.039 for light type × stimulation pattern effect for post-hoc analysis using a
428 mixed effects model two-way ANOVA; t_{13}=0.09, P=0.93 for nonsynchronous stimulation using
429 paired t-test; see Fig. 6b), verifying that the suppression of attack behavior driven by closed-
430 loop stimulation was indeed due to delivery of stimulation timed to endogenous *EN-AggINH*
431 activity. Incidentally, nonsynchronous stimulation had no impact on non-attack social behavior
432 towards intact males or females (N=14; t_{13}=1.79, P=0.097; and t_{13}=0.54, P=0.60, for interaction

433 with males and females, respectively, comparing blue vs. yellow light stimulation using two-
434 tailed paired t-test). Thus, we validated the temporal activity component of *EN-AggINH.*

435 After establishing that we could selectively reduce aggression by temporally targeting PFC
436 based on the activity state of *EN-AggINH*, we tested whether we could reduce aggression by
437 spatially targeting stimulation based on the sub-components of PFC output circuity that
438 composed the network. We identified potential spatially specific targets by looking at the
439 relative LFP spectral Granger directionality from prefrontal cortex that occurred in the
440 aggressive internal state. Our initial visualization of *EN-AggINH* was constrained to the absolute
441 information flow at the strongest synchronies (Fig 3c-d). On the other hand, the relative
442 measures provide a measure of which circuits decrease their information flow prior to and
443 during attack behavior since *EN-AggINH* activity decreases during aggression (see Fig. 6c). In
444 other words, the relative Granger directionality measures quantified information flow pathways
445 that decreased the most during aggression. We focused our analysis on the Granger
446 directionality between PFC [prelimbic (PL) and infralimbic cortex (IL)] to nucleus accumbens
447 (PFC→NAc), medial amygdala (PFC→MeA) and lateral habenula (PFC→LHb), since *EN-AggINH's*
448 relative LFP spectral energy was highest for PFC→NAc and PFC→MeA circuitry and lowest in
449 the PFC→LHb circuit. Thus, a prominent decrease in information flow in PFC→NAc and
450 PFC→MeA circuitry was associated with aggression, while no such change was observed in PFC-
451 LH activity. Critically, all three circuits consisted of monosynaptic projections, enabling direct
452 targeting using optogenetics. We next quantified the relative spectral energy of these circuits at
453 20Hz since stimulating PFC at this frequency was sufficient to suppress the aggressive internal
454 state (Fig. 5d) and attack behavior (Fig. 1a). Given their representation in *EN-AggINH*, we
455 reasoned that driving PFC→NAc or PFC→MeA activity at 20Hz should selectively suppress
456 aggression, while driving PFC→LHb activity should not.

457 We causally activated these three circuits at 20Hz and measured their impact on social
458 behaviors. To selectively stimulate the terminals of PFC neurons in each target region (NAc,
459 MeA, or LHb), we injected mice with a retrograde AAV-Cre (rAAV-Cre) virus in one target region
460 and an AAV-DIO-channel rhodopsin-2 virus in PFC (N=8-9 per group). A stimulating fiber was

placed above the target region injected with rAAV-Cre. Social behavior was quantified during 20Hz stimulation with yellow vs. blue light (5mW, 20Hz, 3ms pulse width).

Blue light stimulation of PFC→NAc or PFC→MeA decreased aggression (t_8=2.4, P=0.04; t_7=5.9, P=0.001 for NAc and MeA stimulation, respectively for blue vs. yellow light using two tailed-paired t-test; N=8-9 mice per group, see Fig. 6d-e). This stimulation also increased non-attack social behavior towards the male C57 mice (t_8=3.1, P=0.015; t_7=3.8, P=0.007 for NAc and MeA stimulation, respectively). Neither of these stimulation protocols impacted social behavior towards female C57 mice (t_8=1.2, P=0.27; t_7=0.8, P=0.46 for NAc and MeA stimulation, respectively). On the other hand, PFC→LHb stimulation had no impact on aggression (t_7=0.38, P=0.71; using two-tailed paired t-test, N=7 mice, see Fig. 6f), or non-attack social behavior towards C57 males (t_7=0.24, P=0.82 using two-tailed paired t-test). Though this stimulation protocol tended to increase social interaction with C57 females, these results did not reach statistical significance (t_7=2.2, P=0.06 using two-tailed paired t-test). These results demonstrated that directly stimulating the PFC subcircuits that normally showed the greatest decreases in aggression-related activity causally and selectively suppressed aggression. On the other hand, stimulating a PFC subcircuit with minimal activity changes during aggression had no impact on social behavior towards male mice. Thus, these findings validated the spatial spectral features of *EN-AggINH.*

Discussion

Here, we set out to discover the internal state that regulates whether an animal will exhibit aggressive or pro-social behavior. We reasoned that attack behaviors emerge from an aggressive internal brain state. Thus, we used machine learning to discover the mesoscale neural architecture of the brain when an animal exhibited attack vs. non attack social behaviors. Like other well-defined internal brain states, such as sleep, we found that the network distinguishing attack behavior incorporated state-dependent patterns of neural activity across every brain region we measured. For multiple regions, differences were observed in local oscillatory power, while others exhibited differences in oscillatory synchrony with a broader collection of regions. Each brain region showed selectivity in the frequencies of

489 oscillations that contributed to the network. For example, prelimbic cortex showed strong
490 activity in the beta frequency range, while medial amygdala showed strong activity in the beta
491 and theta frequency range. No brain region showed prominent activity contributions across all
492 frequencies. We also observed differences in the activity profile of a primary sensory region,
493 V1, which may reflect a change in encoding, or differences in visual sensory input observed
494 during attack behavior. Critically, the brain state identified during attack behavior was better
495 captured by the activity across all recorded brain regions as an integrated network, rather than
496 the independent activity within each brain region.

497 Though behavioral output has been classically utilized to infer the internal state of a brain, we
498 reasoned that an internal brain state was also likely present during intervals immediately
499 preceding and following behavioral output. Thus, we tested whether the aggression network
500 showed distinct activity profiles in the time intervals surrounding attack and non-attack social
501 behaviors. Indeed, network activity was lower during interval surrounding attack behavior.
502 Strikingly, we also found that network activity when animals were isolated in their home cage
503 encoded their trait aggression. Thus, the network did not simply encode behavioral output
504 since it was observed separately from attack behavior. Rather, the network encoded an
505 aggressive internal brain state.

506 Interestingly, this aggressive brain state was encoded by decreased activity in the network.
507 Given that we identified more cells that increased their firing rates as network activity
508 decreased, the discovery of a network that decreases its activity during aggression does not
509 indicate that overall brain activity is suppressed during aggressive states. Rather, these findings
510 argue that the aggressive state is encoded by a network that decreases its activity relative to
511 when mice are socially isolated or engaged in pro-social behavior. Indeed, our data suggested
512 that several common regions/circuits were activated during aggressive and pro-social behavior.
513 These common circuits need not be reflected in our network since our model was trained to
514 differentiate attack vs. non-attack social behavior. Nevertheless, our discovery of a network
515 that decreased its activity during aggression raises the intriguing hypothesis that the brain
516 actively inhibits aggression during pro-social engagement. When activity in this inhibition
517 network is suppressed, aggressive behavior emerges.

518 This interpretation is supported by our validation experiments where we directly activated
519 ESR1+-Cre neurons in ventromedial hypothalamus. Our findings showed that direct activation
520 of these cells induced the aggressive brain state (suppressed *EN-AggINH* activity). When mice
521 treated with CNO were exposed to a stimulus that would generally produce non-attack social
522 behavior (i.e., a female mouse), attack behavior emerged. Thus, the presence of the aggressive
523 brain state changed the mapping between sensory input and behavior output. Similarly, direct
524 stimulation of medial prefrontal cortex biased mice towards exhibiting non-attack social
525 behavior when they were exposed to a stimulus that would generally induce attack behavior
526 (i.e., a male intruder). Our findings showed that medial prefrontal cortex stimulation decreased
527 the aggressive internal state (increased *EN-AggINH* network activity). Critically, our findings
528 using mediation analysis argue that the brain state represented by *EN-AggINH* contributes to
529 the mediation of medial prefrontal cortex stimulation to a suppression of attack behavior.
530 Supporting this finding, our mediation analysis performed using data from the ESR1-Cre
531 experiment showed that *EN-AggINH* also mediated the impact of CNO treatment (see
532 Supplemental Fig. S5). Thus, *EN-AggINH* reflects the internal brain state that suppresses basal
533 aggression.

534 Here, we framed internal brain states as a collection of functions that transform sensory input
535 into behavior. Indeed, we found that when *EN-AggINH* activity is suppressed, the brain
536 transforms both male and female social sensory cues into attack behavior. It is also widely
537 appreciated that sensory input can also cause the brain to transition from one internal state to
538 another. For example, a loud sound can cause an animal to transition from sleeping to a hyper
539 aroused internal state. Along this line, we found that exposure to male mice could promote an
540 aggressive internal state in CD1 mice even prior to attack behavior, while exposure to a female
541 mouse did not (under normal conditions). In this framework, one would also anticipate that
542 many modulatory strategies that regulate attack behavior could mediate their effect by driving
543 the brain out of the state represented by low *EN-AggINH* activity. Indeed, we predict that
544 delivering a bright visual cue or a strong sensory cue (i.e., air puff) timed to decreases in *EN-*
545 *AggINH* activity could also potentially be used suppress attack behavior, since many circuits and
546 sensory inputs likely converge onto the internal state represented by *EN-AggINH*.

547 Our closed-loop stimulation approach was developed using a neural-network based
548 approximation technique for which the features were substantially constrained relative to
549 dCSFA-NMF. Nevertheless, we found that the reduced encoder was sufficient to identify the
550 precise time windows when the brain transitioned into aggression, as marked by a decrease in
551 *EN-AggINH* activity. In the future, novel approaches may allow for further improvement in the
552 precision of our real-time stimulation approach. For example, future work could exploit
553 convolutional neural networks to bypass the feature extraction step. These neural network
554 encoders could be altered to predict both aggressive and pro-social states, such as the
555 generalized social appetitive network that we recently discovered [7]. By using both networks
556 concurrently to actuate a closed-loop system, it may be possible to further suppress aggressive
557 behavior relative to pro-social behavior. Indeed, our current findings also pointed to a network
558 that exhibits increased activity during aggressive behavior (Electome Network #6, see Fig. 1f,
559 and Supplemental Fig. S6). Though the network failed to encode the urine paradigm, it is
560 possible that it contains activity that synergizes with *EN-AggINH* to encode aggressive social
561 states more optimally. If future studies demonstrate this potential, imitation encoders for both
562 Electome Network 6 and *EN-AggINH* could be integrated to further optimize closed-loop
563 approaches to selectively suppress aggression.

564 Multiple neuropsychiatric disorders including mood disorders, psychotic disorders,
565 neurodevelopmental disorders, and neurodegenerative disorders are associated with deficits in
566 regulating social behavior, including aggression. While multiple pharmacological approaches
567 have been instituted to suppress aggressive behavior towards self and others, many of these
568 strategies act by sedating the individual and can disrupt aspects of pro-social function. Our
569 discovery of a brain network that encoded an aggressive state raises the potential for novel
570 approaches to suppress aggressive behavior that spare pro-social behavior. Indeed, compared
571 to a standard open-loop stimulation protocol (20Hz stimulation) which suppressed both attack
572 and non-attack pro-social behavior, our closed-loop stimulation approach spared non-attack
573 social behavior towards males or females. Intriguingly, like other open-loop PFC stimulation
574 studies [50, 51], our 20Hz stimulation protocol induced behavioral hyperactivity in experimental
575 mice (see Supplemental Fig. S7). On the other hand, our closed-loop stimulation protocol did

not (see Supplemental Fig. S7). Thus, our findings also show that closed-loop stimulation may limit off-target behavioral effects that are induced by classic stimulation approaches.

Overall, our findings establish a generalized network-level signature whereby the brain suppresses aggression via active inhibition. Moreover, they highlight the exciting potential for state-specific neuromodulation to regulate internal states.

Acknowledgements

We would like to thank Stephen Mague, Karim Abdelaal, and Ashleigh Rawls and Jean M. Zarate 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. A widespread network encodes attack behavior. a) Direct stimulation of prefrontal cortex suppresses social behavior. Schematic of optogenetic stimulation (left) and social encounters utilized for testing (middle). Prefrontal cortex stimulation suppressed attack behavior, increased non-attack social behavior towards male mice, and suppressed non-attack social behavior towards females (*P<0.05 for each comparison). **b)** Schematic for electrical recordings, showing targeted brain regions (left), and representative local field potentials (middle) recorded during repeated exposure to social contexts that produce attack and non-attack social behavior (right). **c)** Framework to test individual brain regions' encoding of social states (left). All implanted regions encoded social engagement; however, only five selectively encoded the attack behavior vs. non-attack behavior (right). Pink shading indicates P<0.05 with FDR correction. **d)** Attack codes discovered from the five brain regions failed to encode

602 aggressive behavior induced by male urine (gray shading indicates P<0.05 prior to but not
603 following FDR correction). **e)** Schematic of machine-learning model used to discover network
604 encoding attack behavior (left). The inputs to the model included LFP activity from the 11 brain
605 regions, the aggression class (+/-), and the social condition (IM-Intact male, CM-Castrated Male,
606 F-Female) for each 1-second data window. **f)** Encoding across eight learned networks. The
607 supervised network (purple, *EN-AggINH*) showed the strongest encoding. Data shown as mean
608 ± SEM.

609 **Figure 2. *EN-Aggression Inhibition* encodes an aggressive internal state. a)** Neural activity was
610 sampled while mice were socially isolated (blue) and during intervals preceding and following
611 social behavior. **b)** Network activity during these intervals encoded attack behavior vs. male and
612 female non-attack social behaviors, while ventral hippocampal activity did not (P<0.05 using
613 Friedman's test followed by Wilcoxon sign-rank test). During isolation (blue) Network activity,
614 but not ventral hippocampus activity, encoded the subsequent total attack time of individual
615 mice (P<0.05 using Spearman's Rank Correlation).

616 **Figure 3. Dynamics and biological components of *EN-Aggression Inhibition*. a)** Prominent
617 oscillatory frequency bands composing *EN-AggINH* are highlighted for each brain region around
618 the rim of the circle plot. Prominent synchrony measures are depicted by lines connecting brain
619 regions through the center of the circle. The plot is shown at relative spectral energy of 0.4.
620 Theta (4-11 Hz) and beta (14-30 Hz) frequency components are highlighted in blue and green,
621 respectively. **b)** Example relative LFP spectral energy plots for three brain regions corresponding
622 to the circular plot in A (See Supplemental Fig. S3-4 for full description of network features). **c)**
623 Granger offset measures were used to quantify directionality within *EN-AggINH*. Prominent
624 directionality was observed across the theta and beta frequency bands (shown at spectral
625 density threshold of 0.4 and a directionality offset of 0.3). Histograms quantify the number of
626 leading and lagging interactions between brain regions. **d)** Schematic depicting directionality
627 within *EN-AggINH*. **e-f)** *EN-AggINH* maps to cellular activity. e) Three cells from LHb, VMH, and
628 MeA showing firing activity that is negatively correlated with *EN-AggINH* activity (red) and a
629 VHip cell showing positively correlated firing (blue). f) *EN-AggINH* activity correlated with
630 cellular firing across the brain across the brain. Single- and multi-units were used for analyses.

631 **Figure 4. *EN-Aggression Inhibition* encodes distinct aggression contexts. a)** Experimental
632 approach for causally inducing aggression via direct activation of ESR1+ cells in ventromedial
633 hypothalamus. b) Cellular activation induced attack behavior towards female mice (P<0.001
634 using sign-rank test), c) decreased *EN-AggINH* activity during social interactions with female
635 mice (P<0.01 using one-tailed Wilcoxon sign-rank test) and d) intervals surrounding these
636 interactions (P<0.05 using Wilcoxon sign-rank test).

637 **Figure 5. *EN-Aggression Inhibition* activity is causally related to aggression. a)** Schematic for
638 closed-loop manipulation of *EN-AggINH* activity. **b)** Real-time estimation of aggression. Receiver
639 operating characteristic depicting detection of aggressive behavior in a mouse using *EN-AggINH*
640 activity vs. real-time reduced encoder is shown to the right. Dashed blue line corresponds to
641 the established detection threshold. **c)** Detection of aggression using reduced encoder vs. *EN-*
642 *AggINH* across mice (N=9; P=0.43 using two-tailed paired Wilcoxon sign-rank). **d)** *EN-AggINH*
643 activity relative to light stimulation during closed-loop manipulation. Network activity
644 significantly decreased one second prior to yellow light stimulation (N=9, $^{\#\#}$P<0.005 using one-
645 tailed sign rank test; note that activity was normalized for each mouse to the average network
646 activity during isolation). Activity was also higher one second after stimulation with blue light
647 vs. yellow light (**P<0.01 using one-tailed signed-rank test). **e)** Directed graph with the inferred
648 modes of action derived from mediation analysis. There is a causal relationship from
649 stimulation to behavior and from stimulation to *EN-AggINH* expression (model 1; P<0.01 using
650 signed rank and paired t-tests). *EN-AggINH* is a mediator from stimulation to behavior (P<0.01
651 using nested logistic regression models, likelihood ratio test; model 2), *EN-AggINH* activation is
652 the primary mechanism whereby prefrontal cortex stimulation suppresses aggression (P<0.01
653 using average causal mediation effect, model 3). **f-g)** Directed graph testing f) local theta power
654 as the primary mechanism whereby prefrontal cortex stimulation suppresses aggression (model
655 4) and g) *EN-AggINH* activation as the primary mechanism whereby prefrontal cortex
656 stimulation suppresses aggression when local power is included as an intermediary (model 5).
657 The uncorrected P values for each brain area in both models are shown below as -log(P).

658 **Figure 6. Validation of spatiotemporal features of *EN-Aggression Inhibition.* a)** Portion of
659 windows stimulated during social behaviors using a closed-loop approach (P=0.002 using one-

660 tailed sign-rank test, left). Behavioral effects of closed-loop stimulation, right). **b)** Schematic for
661 nonsynchronous control stimulation (left). Nonsynchronous stimulation does not impact
662 aggressive or non-attack social behavior towards males or females. **c)** Granger Coherence for
663 PFC-dependent subcircuits within *EN-AggINH* (shown as relative spectral energy, see also
664 Supplemental Figure S4) **d)** Viral targeting strategy (left) and behavioral impact of PrL→NAc
665 circuit stimulation (right). **e)** Viral targeting strategy (left) and behavioral impact of PrL→MeA
666 circuit stimulation (right). **f)** Viral targeting strategy (left) and behavioral impact of PrL→LH
667 circuit stimulation (right). **P<0.005, *P<0.05 using two tailed paired t-test. Order of blue and
668 yellow light stimulation trials is show next to social condition diagrams.

669

670

671 ## Author contributions

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

679

680 ## Declaration of Interests

681 The authors declare no competing interests

682 ## Materials and Methods

683 <u>Animal care and use</u>

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

688 Twenty-nine six-month retired breeder male CD1 strain mice (Charles River Laboratories,
689 Wilmington, Massachusetts) were used to discover a network that encoded aggression,
690 hereafter called *EN-AggINH*. Another fifty-five CD1 mice were used to probe the behavioral and
691 network responses to optogenetic stimulation. Mice were singly housed with enrichment.
692 ESR1-Cre mice on a C57/Bl6J background were provided by Scott Russo. These mice were
693 crossed with CD1 females in the Bryan Vivarium at Duke University to obtain F1 offspring. Eight
694 fourteen-week-old virgin ESR1-cre F1 male offspring were used to validate *EN-AggINH*. All F1
695 offspring were group-housed 2-5 mice per cage until they received viral injections in the
696 ventromedial hypothalamus at 7-8 weeks. After surgery, these mice were singly housed with
697 enrichment. All partner mice (C57BL/6J: two to seven intact males, two to seven females, and
698 two to seven castrated males per experimental mouse) were 7-14 weeks old. These mice were
699 purchased from Jackson Laboratories (Bar Harbor, Maine). All stimulus mice were housed 5 per
700 cage with enrichment. All behavioral testing and neurophysiological recordings occurred during
701 the dark cycle.

702 <u>Castration of C57 male mice</u>

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

709 <u>Electrode implantation surgery</u>

710 Mice were anesthetized with 1% isoflurane and placed in a stereotaxic device. Anchor screws
711 were placed above the cerebellum, right parietal hemisphere, and anterior cranium. The

712 recording bundles designed to target prelimbic cortex, infralimbic cortex, medial amygdala,
713 ventral hippocampus, primary visual cortex, mediodorsal thalamus, lateral habenula, lateral
714 septum nucleus, nucleus accumbens, ventrolateral portion of the ventromedial hypothalamus,
715 and orbitofrontal cortex were centered based on stereotaxic coordinates measured from
716 bregma. [Orbitofrontal cortex: anterior/posterior (AP) 2.35mm, medial/lateral (ML) 1.0mm,
717 dorsal/ventral (DV) from dura -2.75mm; infralimbic cortex and prelimbic cortex: AP 1.8mm, ML
718 0mm, DV -2.7mm from dura; medial amygdala: AP -1.25, ML 2.7mm, DV -4.3 from dura; lateral
719 septum and nucleus accumbens: AP 1.0mm, ML 0mm, DV -4.0mm from dura; ventromedial
720 hypothalamus, lateral habenula, and medial dorsal thalamus: AP -1.47mm, ML 0mm, DV -
721 5.4mm from dura; central hippocampus and primary motor cortex: AP -3.0mm, ML 2.6mm, DV -
722 3.0mm from dura]. We targeted infralimbic cortex and prelimbic cortex by building a 0.6mm DV
723 stagger into the bundle. We targeted lateral septum and nucleus accumbens by building a
724 0.3mm ML and 1.5mm DV stagger into the bundle. We targeted lateral habenula, medial dorsal
725 thalamus, and ventral medial hypothalamus by building a 0.3mm ML, and 1.9mm and 2.5mm
726 DV stagger into our electrode bundle microwires. We targeted primary motor cortex and
727 ventral hippocampus using a 0.3mm ML and 2.5mm DV stagger in our electrode bundle
728 microwires. For optogenetic stimulation experiments, the addition of a Mono Fiberoptic
729 Cannula coupled to a 2.5mm metal ferrule (NA: 0.22, 100mm [inner diameter], 125mm buffer
730 [outer diameter], MFC_100/125-0.22, Doric Lenses, Quebec) was built into the prefrontal
731 cortex bundle. The tip of the fiber was secured 300mm above the tip of the IL microwire. Mice
732 were allowed 10-15 days for recovery from surgery before behavioral testing.

733 Viral surgery

734 For optogenetic experiments targeting PFC soma [21], we used CD1 mice that showed an attack
735 latency < 60s when exposed to an intact C57 male. Thirty-five CD1 mice were anesthetized with
736 1% isoflurane and placed in a stereotaxic device. The mice were unilaterally injected with AAV2-
737 CamKII-ChR2-EYFP (purchased from the Duke Viral Vector Core, Durham, NC; courtesy of K.
738 Deisseroth), based on stereotaxic coordinates from bregma (left Infralimbic cortex: AP 1.8mm,
739 ML 0.3mm, DV -2.0mm from the brain). A total of 0.5mL of virus was infused at the injection
740 site at a rate of 0.1mL/min over five minutes, and the needle was left in place for ten minutes

741 after injection. For the open-loop stimulation experiment, CD1 mice were implanted with an
742 optic fiber (Mono Fiberoptic Cannula coupled to a 2.5mm metal ferrule (NA: 0.22, 100mm
743 [inner diameter], 125mm buffer [outer diameter], MFC_100/125-0.22, Doric Lenses, Quebec))
744 0.3mm above the injection site immediately after viral syringe was removed. These mice were
745 allowed 3 weeks for recovery prior to behavioral testing. For the closed-loop experiments, CD1
746 mice were allowed 3 weeks for viral expression prior to implantation with an optrode.

747 For the ESR1-Cre validation experiment, thirteen F1 offspring were bilaterally injected with
748 AAV2-hSyn-DIO-GqDREADD (obtained from Addgene) based on stereotaxic coordinates
749 measured from bregma (AP -1.5mm, ML ±0.7mm, DV -5.7mm from the dura). A total of 0.3mL
750 of virus was infused bilaterally at a rate of 0.1mL/min, and the needle was left in place for five
751 minutes after injection. Two weeks after viral infusion, F1 males were screened for aggressive
752 behavior towards females. The F1 males received i.p. injections of CNO (1mg/kg) at the start of
753 the screening session. Thirty-five minutes after injection, a novel C57 female was placed in the
754 home cage for 5 minutes. Screening was repeated one week and two weeks later. Only F1
755 males who attacked females for at least two of the three screening sessions (9/13 mice) were
756 implanted with electrodes [17]. The eight mice that showed good surgical recovery were
757 subjected to further experiments.

758 For PFC projection-targeting experiments, we used forty-four male CD1 mice that showed an
759 attack latency <60s and initiated attacks at least three times within three minutes when
760 exposed to an C57 male mouse. These mice were unilaterally injected with AAV2-EF1a-DIO-
761 ChR2-eYFP (obtained from Addgene) in the left prefrontal cortex based on stereotaxic
762 coordinates measured from bregma (AP 1.8mm, ML 0.3mm, DV -2.0mm from the dura), and
763 AAVrg-EF1a-Cre-mCherry (obtained from Addgene) was injected in the downstream target
764 region based on stereotaxic coordinates measured from bregma (NAc: AP 1.0mm, ML 0.9mm,
765 DV -3.8mm from the dura; MeA: AP -1.25mm, ML 2.75mm, DV -4.3mm from the dura; or LHb:
766 AP -1.6mm, ML 0.4mm, DV -2.2mm from the dura). A total of 0.3mL of virus was infused in the
767 prefrontal cortex and 0.3mL of virus was infused in the downstream target region at a rate of
768 0.1mL/min. The needle was left in place for five minutes after injection. Immediately after the
769 viral syringe was removed from the downstream target region, mice were implanted with an

770 optic fiber (Mono Fiberoptic Cannula coupled to a 2.5mm metal ferrule (NA: 0.22, 100mm
771 [inner diameter], 125mm buffer [outer diameter], MFC_100/125-0.22, Doric Lenses, Quebec))
772 0.3mm above the downstream target region injection site. Three weeks after viral
773 infusion/optic fiber implantation, CD1 mice were screened for aggression. Thirty-three
774 implanted mice that continued to show an attack latency <60s and initiated attacks at least
775 three times within three minutes were used for testing effects of projection targeted
776 stimulation on aggressive behavior.

777 Histological analysis

778 Histological analysis of implantation sites was performed at the conclusion of experiments to
779 confirm recording sites and viral expression. Animals were perfused with 4% paraformaldehyde
780 (PFA), and brains were harvested and stored for at least 96 hrs in PFA. Brains were
781 cryoprotected with sucrose and frozen in OCT compound and stored at -80C. Brains were later
782 sliced at 40µm. Brains from mice used to train and validate the network were stained using
783 NeuroTrace fluorescent Nissl Stain (N21480, ThermoFisher Scientific, Waltham, MA) using
784 standard protocol. Specifically, Nissl staining for brain tissue occurred on a shaker table at room
785 temperature. Tissue was washed in PBST (0.1% Triton in phosphate-buffered saline solution) for
786 10 minutes. It was then washed for five minutes in PBS twice. The tissue was then protected
787 from light for the remainder of the protocol. The tissue was incubated in 1:300 Nissl diluted in 2
788 mL PBS for 10 minutes. After the Nissl incubation, tissue was washed once in 0.1% PBST for 10
789 minutes, then twice in PBS for 5 minutes. Brains from ESR1 mice and mice used for 20 Hz or
790 closed-loop stimulation were mounted in Vectashield mounting medium containing DAPI (H-
791 1200-10, Vector Laboratories, Newark, CA). Images were obtained at 10x using an Olympus
792 fluorescent microscope. Of the 297 total implantation sites in the training and testing set of
793 mice, 17 were mistargeted (~5.7% error rate). Of these mistargeted implants, 13 were within
794 200µm of the targeted structure. Given our prior work demonstrating high LFP spectral
795 coherence (in the 1-55Hz frequency range) across microwires separated by 250um, in both
796 cortical and subcortical brain regions [44], we chose to retain these animals in our analysis. The
797 other four mistargeted implants were within 350µm of the targeted structure. The most

798 reliably mistargeted site was ventral medial hypothalamus for which 4 animals were implanted
799 within 200um of the target, and 2 animals were implanted within 350 µm of the target.

800 Machine learning analysis typically benefits from larger data sets. Thus, we concluded that
801 maintaining a higher number of data points likely outweighed the effect of a small number of
802 mistargeted brain regions, particularly since our LFP measures were robust to the targeting
803 inaccuracies we observed histologically. As such, we pooled data from all 20 implanted animals
804 to learn our initial model. We employed a similar strategy for our validation analysis, where an
805 animal was only removed from the validation set if there was clear histological confirmation of
806 mistargeting >200µm for any of the recorded regions. Specifically, presuming accurate targeting
807 with 94.3% certainty and targeting within 200um at a higher certainty, we included animals
808 with missing or damaged histological slices in our analysis. However, if there was clear
809 histological confirmation of mistargeting for any of the recorded regions (as was the case for 1
810 mouse), the animal was removed from the validation testing. Critically, our validation
811 procedure implies that the machine learning models were robust regardless of any slight
812 imprecision in the animals we utilized for training.

813 Neurophysiological data acquisition

814 Mice were connected to a data acquisition system (Blackrock Microsystems, UT, USA) while
815 anesthetized with 1% isoflurane. Mice were allowed 60 minutes in their home cage prior to
816 behavioral and electrophysiological recordings. Local field potentials (LFPs) were bandpass
817 filtered at 0.5-250Hz and stored at 1000Hz. An online noise cancellation algorithm was applied
818 to reduce 60Hz artifact (Blackrock Microsystems, UT, USA). Neural spiking data was referenced
819 online against a channel recording from the same brain area that did not exhibit a SNR>3:1.
820 After recording, cells were sorted using an offline sorting algorithm (Plexon Inc., TX) to confirm
821 the quality of the recorded cells. Only cell clusters well-isolated compared to background noise,
822 defined as a Mahanalobis distance greater than 3 compared to the origin, were used for the
823 unit-electome network correlation analysis. We used both single (well isolated clusters with
824 ISI<1.5) and multi-units (well isolated clusters with ISI<1.5; N=186 total neurons) for our
825 analysis as our objective was to determine whether electome network activity was reflective of

826 cellular activity. Neurophysiological recordings were referenced to a ground wire connected to
827 anchor screws above the cerebellum and anterior cranium.

828 Behavioral recordings and analysis for training/testing models

829 The CD1 mice used for training and testing the electome model were first subjected to
830 screening to assess their basal level of aggressiveness. Screening occurred once a day for three
831 consecutive days prior to surgical implantation. Animals were screened in cohorts. For each
832 screening session, an intact male C57 was placed in the CD1's home cage for 5 minutes and the
833 latency to first attack was recorded. To ensure that our network generalized broadly across CD1
834 mice, we used a training and testing set for which ~50% of the mice showed high aggression
835 during screening (i.e., latency to attack < 60s), and ~50% of the mice showed low to moderate
836 aggression (i.e., latency to attack > 60s). Animals that showed no aggression during screening
837 (16/45 mice) were excluded from further experiments.

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

847 After screening, twenty-nine mice were implanted, and data was acquired across 1-6 behavioral
848 testing/recording sessions following recovery. Sessions were separated by 5-7 days. Recordings
849 for all social encounters were performed in the home cages of the CD1 mice. Each behavioral
850 testing session began with 5 minutes of recording prior to introduction of the social stimulus.
851 All mice were subjected to encounters with an intact C57 male mouse and a female C57 mouse.
852 A subset of eighteen CD1 mice were also subjected to an encounter with a castrated male
853 mouse, and another subset of eighteen mice were subjected to exposure to objects covered in

854 CD1 mouse urine. Object pairs included yellow duplex blocks, curved red duplex blocks,
855 weighted 5 mL conicals, glassware tops, and objects assembled from black legos®. The CD1
856 mice were exposed to a different pair of objects during each session. Order of exposure to
857 stimulus mice and objects was shuffled for every session. Six of the CD1 mice were recorded
858 under all four conditions. Data observations (1 second each) were pooled across eleven CD1
859 mice for training the network model. Object pairs included yellow duplex blocks, curved red
860 duplex blocks, weighted 5 mL conicals, glassware tops, and objects assembled from black
861 legos®. The CD1 mice were exposed to a different pair of objects during each session. Order of
862 exposure to stimulus mice and objects was shuffled for every session.
863 For ESR1 male behavioral testing, eight mice were injected with either saline or CNO (1mg/kg,
864 i.p.) after the five-minute baseline recording. Thirty-five minutes after this injection, mice were
865 exposed to an intact male C57, a castrated male C57, and a female C57, presented in
866 pseudorandom order. Mice were subjected to six total recording sessions (three in which they
867 were treated with saline and three in which they were treated with CNO), again in
868 pseudorandom order. Sessions were separated by 5 days to allow an adequate washout of
869 CNO[52].

870 Behavior was scored for each second as an “attack”, “non-attack social interaction”, or “non-
871 interaction”. One-second windows were identified as "aggressive" if the mouse was engaged in
872 biting, boxing (kicking/clawing), or tussling behavior [36]. Windows were labeled as "non-attack
873 social interaction" if the mouse had his nose or forepaws touching the stimulus mouse (intact
874 male/female/castrated male) or object, but was not biting, boxing, or tussling. Examples of
875 behaviors labeled "non-attack social behavior" included sniffing, grooming, or resting (placing
876 nose or forepaws against the subject mouse, but not moving). If the stimulus mouse had
877 his/her forepaws or nose on the CD1, but it was not reciprocated, this was labeled "non-
878 interaction". CD1 straight approach, sideways approach, and chasing of the stimulus mouse
879 could result in attack (biting/kicking/tousling), non-attack social behavior (nose or paw touch),
880 or withdrawal without any touch. Thus, while sideways approach and chasing are regularly
881 labeled as "aggressive" in the literature [36, 53, 54], and straight approach is regularly labeled
882 as "pro-social", these behaviors lacked consistent resolutions. Moreover, mice also

883 demonstrated these behaviors towards female and castrated mice (non-attack social context).
884 One-second windows containing these behaviors were labeled "non-interaction". All other
885 timepoints not labeled "attack" or "non-attack social" were also labeled "non-interaction".
886 These behavioral criteria were selected to include ethologically aggression-related behaviors
887 and maximize the likelihood that the CD1 was aware of the presence of the stimulus mouse or
888 object during the behavioral window, while remaining confident in the classification of "attack"
889 and "non-attack social" window labels.

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

896 The videos used to generate the labels for training and testing our machine learning model
897 were hand-scored by a trained researcher. Videos from ESR1 mice and optogenetic stimulation
898 were automatically tracked using DeepLabCut [55, 56]. This information was then used for
899 creating behavioral classifiers in SimBA [57].

900 <u>LFP preprocessing and signal artifact removal</u>

901 Each LFP signal was segmented into 1s non-overlapping windows. If there were multiple intact
902 channels implanted in a region, they were averaged to produce a single signal. Windows with
903 non-physiological noise were removed using an established automated heuristic [7]. Briefly, the
904 envelope of the signal in each channel was estimated using the magnitude of the Hilbert
905 transform. The Median Absolute Deviation (MAD) of the magnitude was then calculated on
906 each channel of each recording. Signal was marked as non-physiological if the envelope
907 exceeded a high threshold (5x MAD, which is roughly 4x the standard deviation for a normally
908 distributed signal). Any data adjacent to non-physiological data that had an envelope value
909 above a smaller threshold (0.167 MAD) was also considered non-physiological. All data marked
910 in this way was ignored when averaging channels for each region. Any channels with standard

deviation less than 0.01 were 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 steps. 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 [58]. 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 [7]. Briefly, Granger features were exponentiated to approximately maintain the additivity assumption made implicitly by NMF models [7, 59] 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 21460 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, 14890 seconds where CD1 mice exhibited non-attack social behavior (3542 seconds towards intact males, 9067 seconds towards females, and 2281 seconds towards castrated), and 1890 seconds where mice exhibited attack behavior towards the intact males.

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-nine 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 [41]. This approach assumes each window of is an independent stationary observation and examines dynamics in brain activity only at the scale of windows. A one-second window was chosen to balance capturing fine-grained transient behavior with sufficient length to properly estimate spectral features [7]. 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 (electome) networks. This model is learned using a supervised autoencoder. The objective we use to learn the parameters is

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

963 Here, $Loss_X$ is the reconstruction loss of the features derived from electrophysiology, which for
964 our work was an L_2 loss. Our predictive loss $Loss_Y$ is the cross-entropy loss commonly used for
965 binary classification. Each of the K networks is represented in vector form and combined to
966 make a matrix $W \in R^{p \times K}$. The electome network scores are given by the multi-output function
967 $f(x; \phi): R^p \rightarrow R^K$, where ϕ represent the parameters of the function. In our model, the multi-
968 output function was an affine transformation of $Ax + b$ followed by a softplus rectification,
969 defined as $softplus(x) = \log(1 + \exp x)$, thus $\phi = \{A, b\}$. Parameters $d \in R^k$ represent the
970 relationship between the electome networks and the behavior. A sparsity constraint is enforced
971 so that $d = [d_1, 0, \ldots, 0]$, meaning that only a single electome is used to predict behavior,
972 simplifying interpretation. λ is a weighting parameter used to control the relative importance
973 of prediction. We chose a value that kept the two losses approximately equal during training,
974 which corresponded to 1.

975 Previous work has also found that the reconstruction loss can reduce overfitting and make the
976 learned predictions more robust [60]. To further reduce overfitting of the predictive aspect of
977 the encoder, we applied an elastic net loss [61] on the encoder $Loss_{EN}$ with a weighting μ and
978 the ratio between the L_1 and L_2 losses set to .5. The value for μ was set to a small value that
979 had worked well previously. In this work, power features were also upweighted by a factor of
980 10 to accommodate that there were many more Granger features and truncated at 6 to
981 prevent outliers from dominating the predictions.

982 These models and statistical analyses were implemented with Python 3.7 and Tensorflow
983 version 2.4. Parameters were learned with stochastic gradient descent using the Nesterov
984 accelerated ADAM optimizer [62]. Learning was performed for 30000 iterations, which was
985 observed to be ample for parameter convergence. The learning rate and batch size were set to
986 1e-3 and 100 respectively, values that have empirically performed well in similar applications.

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

990 <u>Hyper-parameter selection</u>

The dCSFA-NMF procedure requires selection of several settings in the algorithm. Specifically, we must choose the number of electome networks 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. Specifically, in our prior work, we demonstrated that the inferred model is highly insensitive to λ [27]. Thus, we chose a λ value to give roughly equal weight to the predictive and generative tasks. Similarly, since the former task grows linearly with brain regions and the latter task grows quadratically, we chose to weight the power features to rough match the coherence features. Since the decoder is also linear, we chose a linear mapping function followed by a softplus to ensure non-negativity. This approach served to limit complexity.

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.
Specifically, our previous work has demonstrated that latent dimensionality is not an important parameter in terms of predictive performance [27]. Thus, we trained one supervised network for all the models tested in this study. We also trained multiple unsupervised networks for each model to explain variance in brain activity that was not directly related to predictive performance. Since our previous work had found that the supervised network has relatively low variance, we used the Bayesian Information Criterion (BIC) to select the number of unsupervised networks (latent factors d) to use in the final network model. The BIC is defined as:

$$BIC(d) = k \log N - 2 \log(\hat{L}),$$

where $k = p \cdot d$ is the number of model parameters (p is the number of spectral features), N is the number of samples, and $\hat{L} = p(X|\hat{\theta})$ is the likelihood of the observed data using the estimated model parameters. This criterion balances the model fit quantified by $\hat{L}$ with the

complexity quantified by $k \log N$. In this work, $-\log(\hat{L})$ is an L_2 loss, corresponding to a Gaussian observational likelihood. The model parameter $\hat{\theta}$ was estimated on 80% of the data while model parameter $\hat{L}$ was evaluated on a 20% holdout set to avoid overfitting. The BIC was evaluated for all dimensionalities from 1-20 networks, and the lowest value was selected as the best model. Since 7 unsupervised networks provided the best fit (a BIC of 5457701, see also Supplemental Fig. S2), our final network model utilized a total of 8 networks, 1 supervised and 7 unsupervised, across all 11 regions.

For each single-region model, we trained 3 unsupervised networks and a single supervised network. Here, we reduced the number of networks as compared to the full network model, given the dramatic reduction in the number of covariates considered by the model. Critically, our objective was to compare the predictive performance of the single-region models against each other and the full network model. Since the predictive performance is driven by the supervised network[27], the smaller latent dimensionality of the single region models had no impact on our final conclusions.

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. One-second windows were pooled from the twenty CD1 mice and used to generate a series of dCSFA-NMF models for each of the 11 brain regions. The models were trained to distinguish behavioral windows of one social state exhibited by CD1 mice, from windows of two other social states. These three social states included 1) male-directed attack, 2) female non-attack social interactions, and 3) castrated male non-attack social interactions. We also developed a model to distinguish 4) periods where CD1 mice 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.

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 aggression electome network 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 (Supplemental 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 across 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 for both 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.

<u>Decoder Information Content</u>

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.

1073 Single-cell correlation to Electome Network activity

1074 Data acquired during the third behavioral testing session was from the twenty implanted mice
1075 were used for cellular analysis. We used Spearman correlation to quantify the relationship
1076 between cellular firing windows and the activity of the electome network used to classify attack
1077 behavior. We performed 1000 permutations of randomly shuffling 1 second windows within
1078 each class for attack and non-attack social interactions with male, female and castrated C57
1079 mice. This approach maintained the relationship between network activity and behavior and
1080 the relationship between cell firing and behavior. We then calculated the Spearman correlation
1081 between network activity and cell firing for each permutation. A cell was deemed positively
1082 correlated if its unshuffled Spearman Rho was above 97.5% of the permutated distribution and
1083 negatively correlated if it was below 2.5%.

1084 Real-Time Encoder Approximation

1085 Because Granger Causality features were too computationally demanding for real-time
1086 calculation, we developed a 'fast' dCSFA-NMF model that relied only on power and coherence
1087 features for estimation of aggressive state to use in the closed-loop stimulation experiments.
1088 This 'fast' model was trained on the same data. The model was trained using regularized
1089 regression to best predict the output of the full encoder. As such, this reduced encoder is also
1090 an affine transformation followed by a SoftPlus activation with a smaller parameter set,
1091 $\phi_r = \{A_r, b_r\}$. This approximation explained a large component of the variance of the
1092 supervised network score on the hold-out validation mice ($R = 0.47$, p-value $<10^{-16}$).

1093 Optogenetic stimulation

1094 Mice were anesthetized with 1% isoflurane, then tethered to an optic patch cable placed over
1095 the optic fiber cannula. For closed-loop experiments, nine mice were also connected to the
1096 recording system. The mice were then allowed 60 minutes for recovery prior to session
1097 recording. For the fiber-only optogenetic stimulation experiments, CD1 mice experienced two
1098 stimulation sessions. For closed-loop optogenetic stimulation, CD1 mice experienced three
1099 sessions of behavioral screening followed by two sessions of closed-loop stimulation.
1100 Stimulation sessions were separated by 5-7 days between sessions. For behavioral screening,

1101 CD1 mice were exposed to intact C57 males, females, and castrated male mice for 5 minutes
1102 each. Screening sessions two and three were used to determine a reduced network threshold
1103 at which 40% of aggressive behavioral windows could be detected. For each session, mice were
1104 recorded for 3 minutes of baseline in their home cage, then during the three social encounters.
1105 Mice were recorded in an open field for 5 minutes after each session. The order of the three
1106 social encounters were shuffled for each session. During each condition, the CD1 mouse
1107 received segments of alternating blue (473nm, Crystal Laser LC, Reno, NV. Model No. DL473-
1108 025-O) and yellow (593.5nm, OEM Laser Systems, Model No. MGL-F-593.5/80mW) light
1109 stimulation, for two minutes each.

1110 For open-loop stimulation targeting PFC soma, CD1 mice received light stimulation for the
1111 entirety of the two-minute segment. For closed-loop, mice received stimulation for one second
1112 when the reduced network score dropped below threshold.

1113 For nonsynchronous stimulation, each of the fourteen CD1 mouse was psuedorandomly
1114 matched to a different mouse that had been used for closed-loop stimulation. Each non-
1115 synchronous mouse was then subjected to the identical order of conditions and yellow and blue
1116 light stimulation blocks as their individually matched closed-loop mouse. Light stimulation was
1117 delivered using the pattern implemented for the closed-loop partner mouse.

1118 For projection targeting stimulation testing, CD1 males were exposed to one testing session
1119 composed of three six-minute blocks of light stimulation. The sequence of light stimulation was
1120 a yellow light stimulation segment, a blue light stimulation segment, then a final segment of
1121 yellow light stimulation. Within each six-minute stimulation block, an intact male C57 and a
1122 female C57 were sequentially placed in the CD1 cage for three-minutes each.

1123 Immediately prior to experiments, light levels were calibrated using a power meter (ThorLabs,
1124 Model No. P0025297 and 11070530), and delivered using a Waveform generator (Agilent
1125 Technologies, Model No. 33210A) for the open-loop experiment. For closed-loop and
1126 nonsynchronous stimulation experiments, the laser was activated using analog output from the
1127 Cerebus recording system.

1128

1129 Mediation Analysis

1130 For the Baron and Kenny approach [48] to establish that *EN-AggINH* expression mediated the
1131 behavioral effect of the neurostimulation, we first used two results previously described in the
1132 methods to establish that there was an effect from the neurostimulation on network
1133 expression and on behavior. We next constrained the data used to the most relevant case,
1134 which is on the closed-loop stimulation. Specifically, we focused on windows of LFP data during
1135 the closed-loop experiment when either the blue or yellow laser was activated to match the
1136 cases between the treatment and control as closely as possible. As only blue light should
1137 significantly manipulate neural activity, this is viewed as the treatment, and the yellow light is
1138 set as the control. We followed the procedures outlined in "LFP preprocessing and signal
1139 artifact removal" to preprocess the data and remove data with significant artifacts. *EN-AggINH*
1140 expression was calculated by projecting the data into the learned model. The remaining data
1141 was then fit into two logistic regression models to predict behavior using the statsmodel
1142 package in python [63]. The first model used only the stimulation to predict behavior (behavior
1143 ~ const + stimulation), and the second model used stimulation and network expression to
1144 predict behavior (behavior ~ const + network_expression + stimulation). These two models
1145 were compared by using a likelihood ratio test to evaluate whether the second model was
1146 significantly better.

1147 For the causal mediation analysis, we again need to roughly balance treatment and control
1148 groups. We used the same data as described above in the classic mediation analysis. We define
1149 the treatment as blue versus yellow light stimulation, the mediator as *EN-AggINH* expression,
1150 and the outcome as aggressive versus non-aggressive behavior. These data were then used in
1151 the causal mediation analysis approach proposed by Kosuke, Keele, and Tingley [49] by using
1152 the statsmodels package in python [63].

1153

1154 Statistics

1155 GraphPad Prism and Matlab were used for statistical analyses of behavior and network activity.
1156 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.

Code and data availability

This learning algorithm is publicly available code at https://github.com/carlson-lab/encodedSupervision. Data will be made available for replication purposes and pre-agreed upon scientific extensions with a material transfer agreement.

References

1. Lindenfors, P. and B.S. Tullberg, *Evolutionary aspects of aggression the importance of sexual selection.* Adv Genet, 2011. **75**: p. 7-22.

2. Lee, S.C., T. Yamamoto, and S. Ueki, *Characteristics of aggressive behavior induced by nucleus accumbens septi lesions in rats.* Behav Neural Biol, 1983. **37**(2): p. 237-45.

3. Haller, J., *The role of central and medial amygdala in normal and abnormal aggression: A review of classical approaches.* Neurosci Biobehav Rev, 2018. **85**: p. 34-43.

4. Bak, M., et al., *The pharmacological management of agitated and aggressive behaviour: A systematic review and meta-analysis.* Eur Psychiatry, 2019. **57**: p. 78-100.

5. Hashikawa, K., et al., *The neural circuits of mating and fighting in male mice.* Curr Opin Neurobiol, 2016. **38**: p. 27-37.

6. Chen, P. and W. Hong, *Neural Circuit Mechanisms of Social Behavior.* Neuron, 2018. **98**(1): p. 16-30.

7. Mague, S.D., et al., *Brain-wide electrical dynamics encode individual appetitive social behavior.* Neuron, 2022. **110**(10): p. 1728-1741 e7.

8. Wong, L.C., et al., *Effective Modulation of Male Aggression through Lateral Septum to Medial Hypothalamus Projection.* Curr Biol, 2016. **26**(5): p. 593-604.

9. Brayley, K.N. and D.J. Albert, *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, 1977. **91**(2): p. 290-9.

1187 10. Couppis, M.H. and C.H. Kennedy, *The rewarding effect of aggression is reduced by*
1188 *nucleus accumbens dopamine receptor antagonism in mice.* Psychopharmacology (Berl),
1189 2008. **197**(3): p. 449-56.

1190 11. Flanigan, M., et al., *An emerging role for the lateral habenula in aggressive behavior.*
1191 Pharmacol Biochem Behav, 2017. **162**: p. 79-86.

1192 12. Golden, S.A., et al., *Basal forebrain projections to the lateral habenula modulate*
1193 *aggression reward.* Nature, 2016. **534**(7609): p. 688-92.

1194 13. Falkner, A.L., et al., *Decoding ventromedial hypothalamic neural activity during male*
1195 *mouse aggression.* J Neurosci, 2014. **34**(17): p. 5971-84.

1196 14. Falkner, A.L., et al., *Hypothalamic control of male aggression-seeking behavior.* Nat
1197 Neurosci, 2016. **19**(4): p. 596-604.

1198 15. Hashikawa, Y., et al., *Ventromedial Hypothalamus and the Generation of Aggression.*
1199 Front Syst Neurosci, 2017. **11**: p. 94.

1200 16. Lin, D., et al., *Functional identification of an aggression locus in the mouse*
1201 *hypothalamus.* Nature, 2011. **470**(7333): p. 221-6.

1202 17. Takahashi, A., et al., *Establishment of a repeated social defeat stress model in female*
1203 *mice.* Sci Rep, 2017. **7**(1): p. 12838.

1204 18. Shibata, S., T.Y. Yamamoto, and S. Ueki, *Differential effects of medial, central and*
1205 *basolateral amygdaloid lesions on four models of experimentally-induced aggression in*
1206 *rats.* Physiol Behav, 1982. **28**(2): p. 289-94.

1207 19. Dambacher, F., et al., *Reducing proactive aggression through non-invasive brain*
1208 *stimulation.* Soc Cogn Affect Neurosci, 2015. **10**(10): p. 1303-9.

1209 20. Choy, O., A. Raine, and R.H. Hamilton, *Stimulation of the Prefrontal Cortex Reduces*
1210 *Intentions to Commit Aggression: A Randomized, Double-Blind, Placebo-Controlled,*
1211 *Stratified, Parallel-Group Trial.* J Neurosci, 2018. **38**(29): p. 6505-6512.

1212 21. Takahashi, A., et al., *Control of intermale aggression by medial prefrontal cortex*
1213 *activation in the mouse.* PLoS One, 2014. **9**(4): p. e94657.

1214 22. Jones, R.B. and N.W. Nowell, *The effect of urine on the investigatory behaviour of male*
1215 *albino mice.* Physiol Behav, 1973. **11**(1): p. 35-8.

1216 23. Lischinsky, J.E. and D. Lin, *Neural mechanisms of aggression across species.* Nat
1217 Neurosci, 2020. **23**(11): p. 1317-1328.

1218 24. Nair, A., et al., *An approximate line attractor in the hypothalamus encodes an aggressive*
1219 *state.* Cell, 2023. **186**(1): p. 178-193 e15.

1220 25. de la Zerda, S.H., et al., *Social recognition in laboratory mice requires integration of*
1221 *behaviorally-induced somatosensory, auditory and olfactory cues.*
1222 Psychoneuroendocrinology, 2022. **143**: p. 105859.

1223 26. APA, *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR,* ed. APA. 2000,
1224 Washington, D.C.: American Psychiatric Association.

1225 27. Talbot, A., et al., *Estimating a brain network predictive of stress and genotype with*
1226 *supervised autoencoders.* J R Stat Soc Ser C Appl Stat, 2023. **72**(4): p. 912-936.

1227 28. Vu, M.T., et al., *A Shared Vision for Machine Learning in Neuroscience.* J Neurosci, 2018.

1228 29. Levy, D.R., et al., *Dynamics of social representation in the mouse prefrontal cortex.* Nat
1229 Neurosci, 2019. **22**(12): p. 2013-2022.

1230 30. Murugan, M., et al., *Combined Social and Spatial Coding in a Descending Projection from*
1231 *the Prefrontal Cortex.* Cell, 2017. **171**(7): p. 1663-1677 e16.

1232 31. Padilla-Coreano, N., et al., *Cortical ensembles orchestrate social competition through*
1233 *hypothalamic outputs.* Nature, 2022. **603**(7902): p. 667-671.

1234 32. Molero-Chamizo, A., et al., *Bilateral Prefrontal Cortex Anodal tDCS Effects on Self-*
1235 *reported Aggressiveness in Imprisoned Violent Offenders.* Neuroscience, 2019. **397**: p.
1236 31-40.

1237 33. He, J., et al., *Effect of transcranial direct current stimulation over the left dorsolateral*
1238 *prefrontal cortex on the aggressive behavior in methamphetamine addicts.* J Psychiatr
1239 Res, 2023. **164**: p. 364-371.

1240 34. Carlson, D., et al., *Dynamically Timed Stimulation of Corticolimbic Circuitry Activates a*
1241 *Stress-Compensatory Pathway.* Biol Psychiatry, 2017. **82**(12): p. 904-913.

1242 35. Mattis, J., et al., *Principles for applying optogenetic tools derived from direct*
1243 *comparative analysis of microbial opsins.* Nat Methods, 2011. **9**(2): p. 159-72.

1244 36. Miczek, K.A., et al., *Aggressive behavioral phenotypes in mice.* Behav Brain Res, 2001.
1245 **125**(1-2): p. 167-81.

1246 37. Coccaro, E.F., et al., *Amygdala and orbitofrontal reactivity to social threat in individuals*
1247 *with impulsive aggression.* Biol Psychiatry, 2007. **62**(2): p. 168-78.

1248 38. Siegel, A., H. Edinger, and A. Koo, *Suppression of attack behavior in the cat by the*
1249 *prefrontal cortex: role of the mediodorsal thalamic nucleus.* Brain Res, 1977. **127**(1): p.
1250 185-90.

1251 39. Chang, C.H. and P.W. Gean, *The Ventral Hippocampus Controls Stress-Provoked*
1252 *Impulsive Aggression through the Ventromedial Hypothalamus in Post-Weaning Social*
1253 *Isolation Mice.* Cell Rep, 2019. **28**(5): p. 1195-1205 e3.

1254 40. Novotny, M., et al., *Synthetic pheromones that promote inter-male aggression in mice.*
1255 Proc Natl Acad Sci U S A, 1985. **82**(7): p. 2059-61.

1256 41. Talbot, A., et al., *Supervised Autoencoders Learn Robust Joint Factor Models ofNeural*
1257 *Activity.* arxiv, 2020.

1258 42. Chen, A.X., et al., *Specific Hypothalamic Neurons Required for Sensing Conspecific Male*
1259 *Cues Relevant to Inter-male Aggression.* Neuron, 2020. **108**(4): p. 763-774 e6.

1260 43. Lee, D.L. and J.L. Wilson, *Urine from Sexually Mature Intact Male Mice Contributes to*
1261 *Increased Cardiovascular Responses during Free-Roaming and Restrained Conditions.*
1262 ISRN Vet Sci, 2012. **2012**.

1263 44. Hultman, R., et al., *Brain-wide Electrical Spatiotemporal Dynamics Encode Depression*
1264 *Vulnerability.* Cell, 2018. **173**(1): p. 166-180 e14.

1265 45. Roelofs, R., et al., *A Meta-Analysis of Overfitting in Machine Learning.* Advances in
1266 Neural Information Processing Systems 32 (NeurIPS 2019) 2019.

1267 46. Deonaraine, K.K., et al., *Sex-specific peripheral and central responses to stress-induced*
1268 *depression and treatment in a mouse model.* J Neurosci Res, 2020. **98**(12): p. 2541-2553.

1269 47. Li, L., et al., *Social trauma engages lateral septum circuitry to occlude social reward.*
1270 Nature, 2022.

1271 48. Baron, R.M. and D.A. Kenny, *The moderator-mediator variable distinction in social*
1272 *psychological research: conceptual, strategic, and statistical considerations.* J Pers Soc
1273 Psychol, 1986. **51**(6): p. 1173-82.

1274 49. Imai, K., L. Keele, and D. Tingley, *A general approach to causal mediation analysis.*
1275 Psychol Methods, 2010. **15**(4): p. 309-34.

1276 50. Anikeeva, P., et al., *Optetrode: a multichannel readout for optogenetic control in freely*
1277 *moving mice.* Nat Neurosci, 2011.

1278 51. Kumar, S., et al., *Cortical Control of Affective Networks.* J Neurosci, 2013. **33**(3): p. 1116
1279 –1129.

52. Jendryka, M., et al., *Pharmacokinetic and pharmacodynamic actions of clozapine-N-oxide, clozapine, and compound 21 in DREADD-based chemogenetics in mice.* Sci Rep, 2019. **9**(1): p. 4522.

53. Miczek, K.A. and J.M. O'Donnell, *Intruder-evoked aggression in isolated and nonisolated mice: effects of psychomotor stimulants and L-dopa.* Psychopharmacology (Berl), 1978. **57**(1): p. 47-55.

54. Kwiatkowski, C.C., et al., *Quantitative standardization of resident mouse behavior for studies of aggression and social defeat.* Neuropsychopharmacology, 2021. **46**(9): p. 1584-1593.

55. Mathis, A., et al., *DeepLabCut: markerless pose estimation of user-defined body parts with deep learning.* Nat Neurosci, 2018. **21**(9): p. 1281-1289.

56. Nath, T., et al., *Using DeepLabCut for 3D markerless pose estimation across species and behaviors.* Nat Protoc, 2019. **14**(7): p. 2152-2176.

57. Nilsson, S.R., et al., *Simple Behavioral Analysis (SimBA): an open source toolkit for computer classification of complex social behaviors in experimental animals.* bioRxiv, 2020.

58. Barnett, L. and A.K. Seth, *The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference.* J Neurosci Methods, 2014. **223**: p. 50-68.

59. Gallagher, N.M., et al., *Cross-Spectral Factor Analysis* Advances in Neural Information Processing Systems, 2017. **30**.

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

61. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society, 2005(67): p. 301-320.

62. Tato, A. and R. Nkambou, *Infusing Expert Knowledge Into a Deep Neural Network Using Attention Mechanism for Personalized Learning Environments.* Front Artif Intell, 2022. **5**: p. 921476.

63. Seabold, S., Perktold, J., *Statsmodels: Econometric and Statistical Modeling with Python.* PROC. OF THE 9th PYTHON IN SCIENCE CONF., 2010.

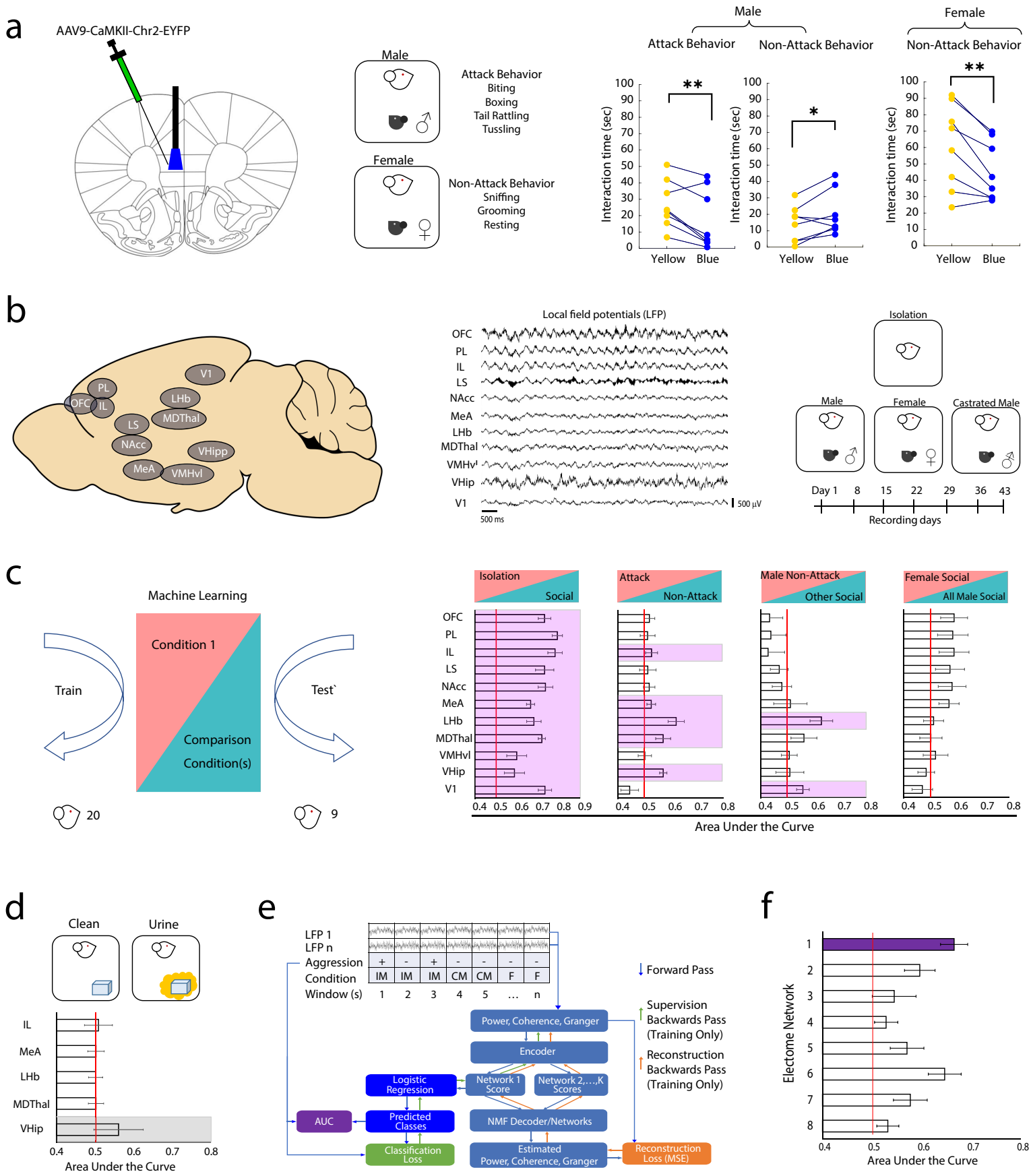
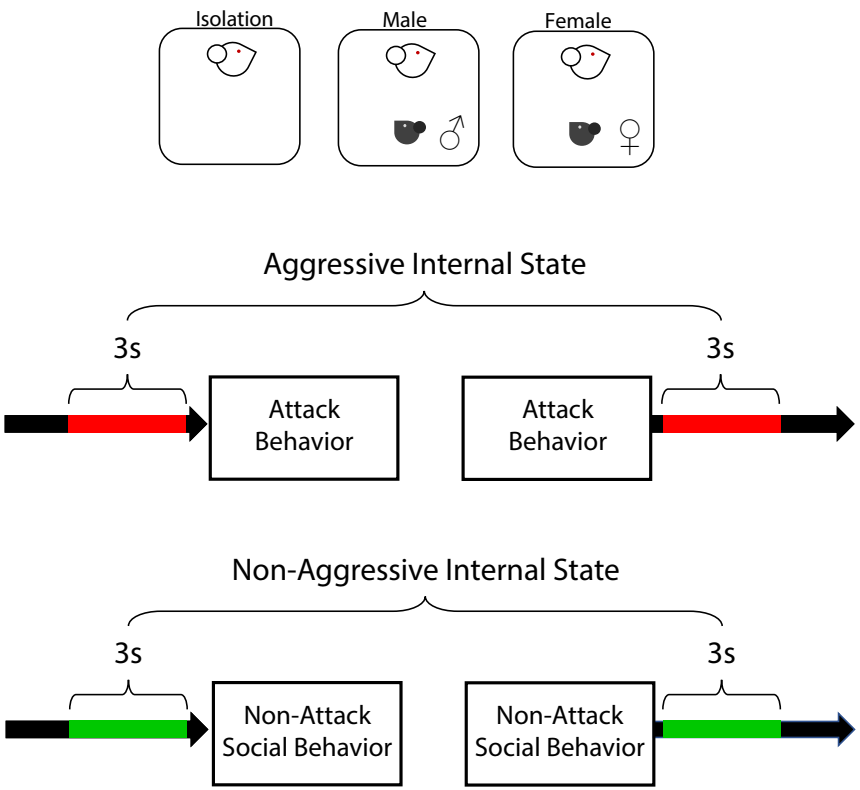

a
AAV9-CaMKII-Chr2-EYFP
Male
Attack Behavior
Biting
Boxing
Tail Rattling
Tussling
Female
Non-Attack Behavior
Sniffing
Grooming
Resting
Male
Attack Behavior
Non-Attack Behavior
Female
Non-Attack Behavior
Interaction time (sec)
Yellow
Blue
b
OFC
PL
IL
LS
NAcc
MeA
LHb
MDThal
VMHvl
VHipp
V1
Local field potentials (LFP)
VMHvl
VHip
500 ms
500 µV
Isolation
Male
Female
Castrated Male
Day 1
8
15
22
29
36
43
Recording days
c
Machine Learning
Condition 1
Comparison Condition(s)
Train
Test`
20
9
Isolation
Social
Attack
Non-Attack
Male Non-Attack
Other Social
Female Social
All Male Social
Area Under the Curve
d
Clean
Urine
IL
MeA
LHb
MDThal
VHip
Area Under the Curve
e
LFP 1
LFP n
Aggression
Condition
Window (s)
IM
CM
F
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)
f
Electome Network
Area Under the Curve

Figure 1

a


Isolation
Male
Female
Aggressive Internal State
3s
3s
Attack Behavior
Attack Behavior
Non-Aggressive Internal State
3s
3s
Non-Attack Social Behavior
Non-Attack Social Behavior

b

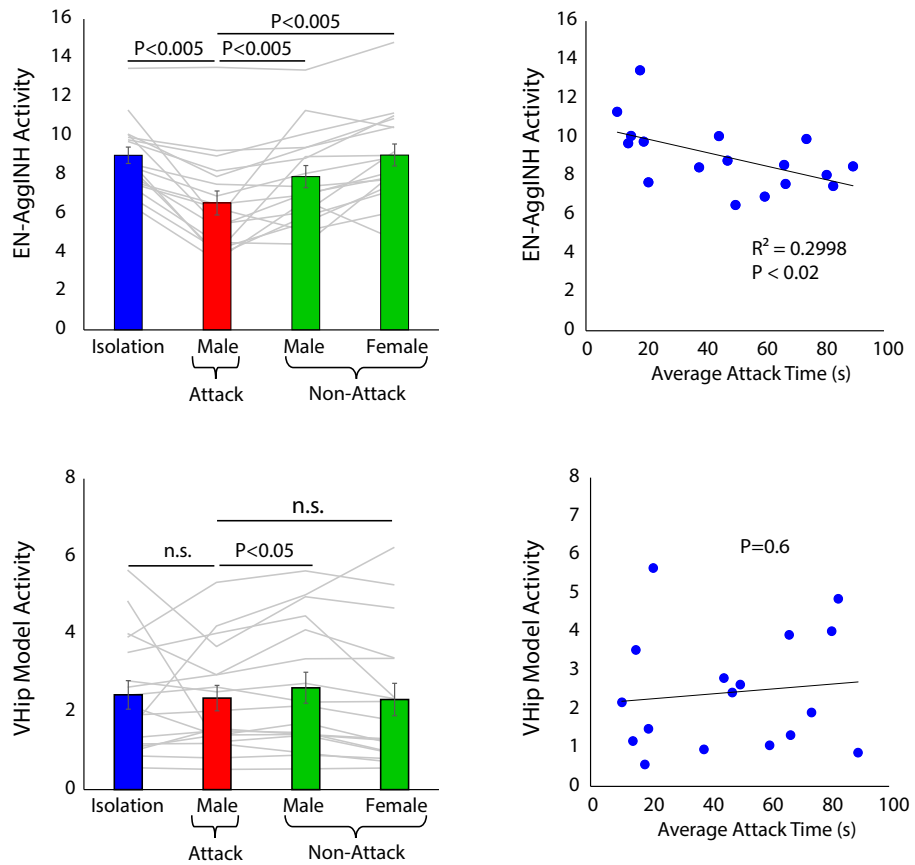

EN-AggINH Activity
P<0.005
P<0.005
P<0.005
Isolation
Male
Male
Female
Attack
Non-Attack
EN-AggINH Activity
$R^2 = 0.2998$
$P < 0.02$
Average Attack Time (s)
VHip Model Activity
n.s.
n.s.
P<0.05
Isolation
Male
Male
Female
Attack
Non-Attack
VHip Model Activity
P=0.6
Average Attack Time (s)

Figure 2

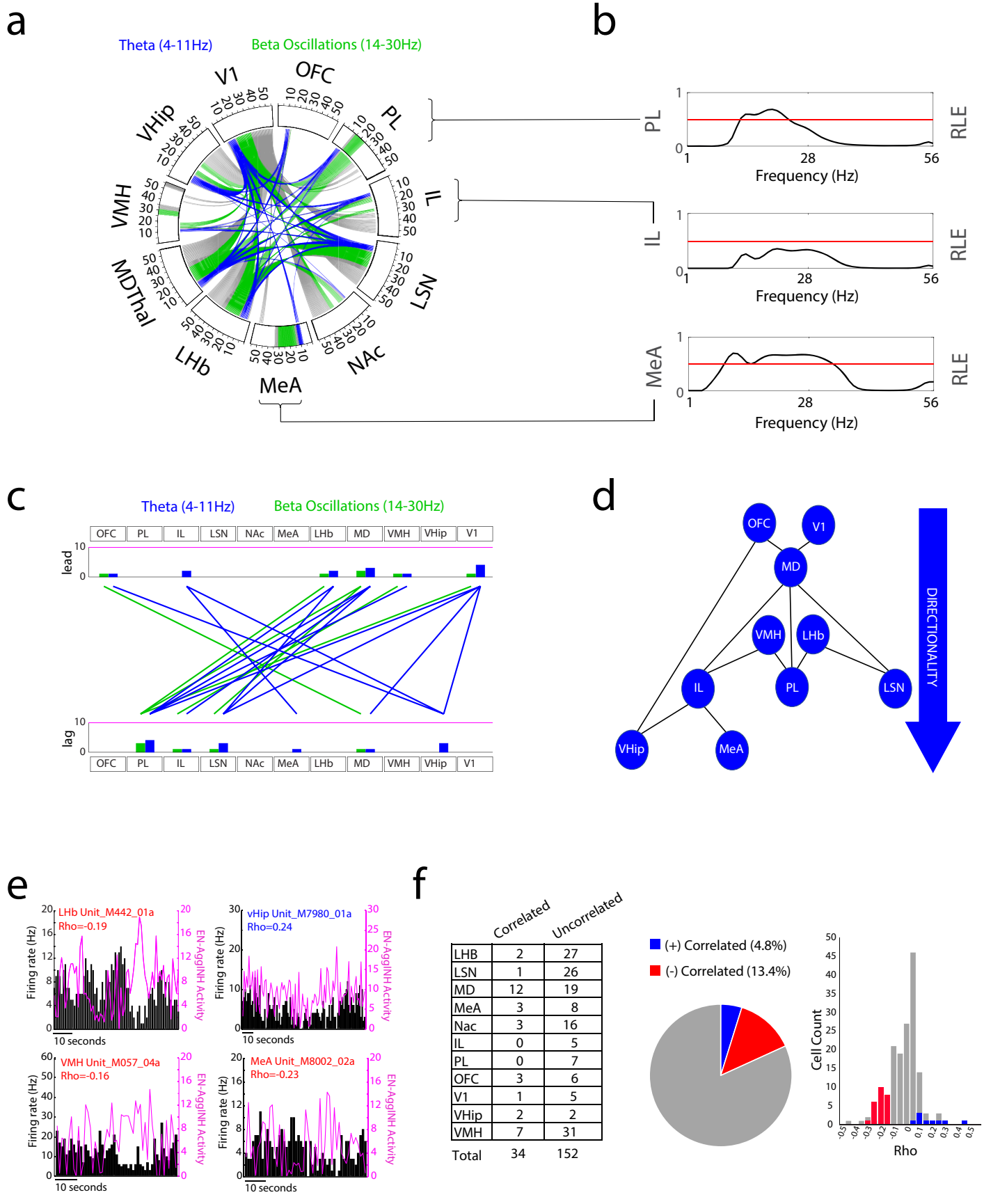

a
Theta (4-11Hz)
Beta Oscillations (14-30Hz)
V1
OFC
PL
IL
LSN
NAc
MeA
LHb
MDThal
VMH
VHip
b
PL
IL
MeA
RLE
Frequency (Hz)
c
Theta (4-11Hz)
Beta Oscillations (14-30Hz)
OFC PL IL LSN NAc MeA LHb MD VMH VHip V1
lead
lag
d
OFC
V1
MD
VMH
LHb
IL
PL
LSN
VHip
MeA
DIRECTIONALITY
e
LHb Unit_M442_01a
Rho=-0.19
vHip Unit_M7980_01a
Rho=0.24
VMH Unit_M057_04a
Rho=-0.16
MeA Unit_M8002_02a
Rho=-0.23
Firing rate (Hz)
EN-AggINH Activity
10 seconds
f
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
(+) Correlated (4.8%)
(-) Correlated (13.4%)
Cell Count
Rho

Figure 3

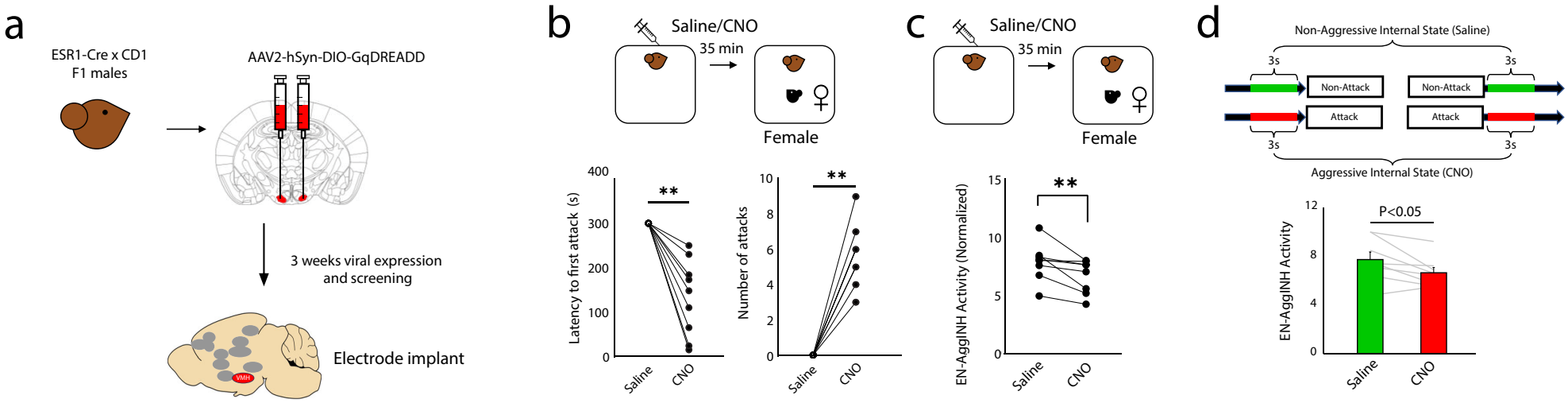

a
ESR1-Cre x CD1
F1 males
AAV2-hSyn-DIO-GqDREADD
3 weeks viral expression
and screening
VMH
Electrode implant
b
Saline/CNO
35 min
Female
Latency to first attack (s)
400
300
200
100
0
**
Saline
CNO
Number of attacks
10
8
6
4
2
0
**
Saline
CNO
c
Saline/CNO
35 min
Female
EN-AggINH Activity (Normalized)
15
10
5
0
**
Saline
CNO
d
Non-Aggressive Internal State (Saline)
3s
Non-Attack
Attack
Aggressive Internal State (CNO)
EN-AggINH Activity
12
8
4
0
P<0.05
Saline
CNO

Figure 4

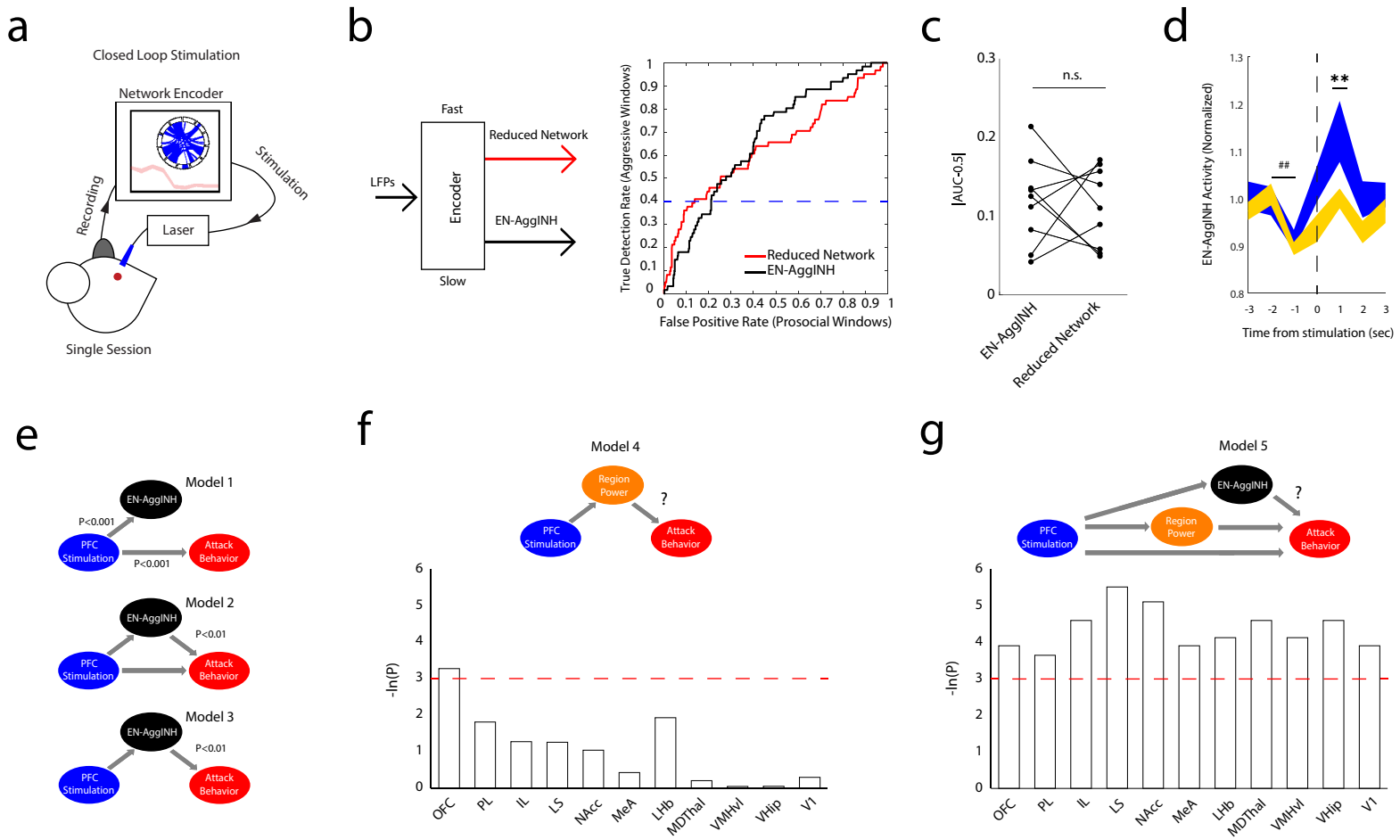
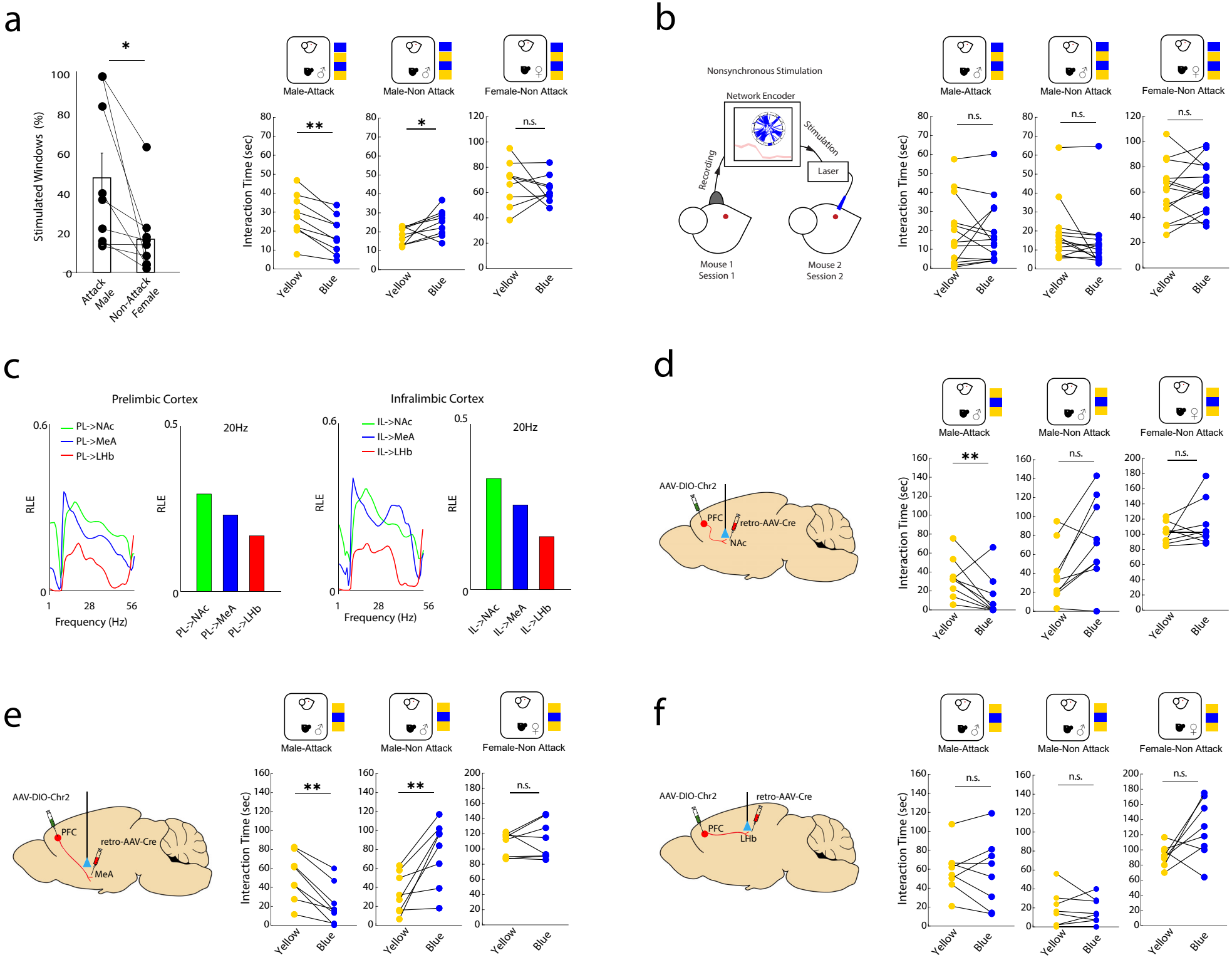

a
Closed Loop Stimulation
Network Encoder
Recording
Stimulation
Laser
Single Session
b
Fast
LFPs
Encoder
Reduced Network
EN-AggINH
Slow
True Detection Rate (Aggressive Windows)
False Positive Rate (Prosocial Windows)
Reduced Network
EN-AggINH
c
n.s.
|AUC-0.5|
EN-AggINH
Reduced Network
d
EN-AggINH Activity (Normalized)
##
**
Time from stimulation (sec)
e
Model 1
EN-AggINH
P<0.001
PFC Stimulation
P<0.001
Attack Behavior
Model 2
EN-AggINH
P<0.01
PFC Stimulation
Attack Behavior
Model 3
EN-AggINH
P<0.01
PFC Stimulation
Attack Behavior
f
Model 4
Region Power
?
PFC Stimulation
Attack Behavior
-ln(P)
OFC
PL
IL
LS
NAcc
MeA
LHb
MDThal
VMHvl
VHip
V1
g
Model 5
EN-AggINH
?
PFC Stimulation
Region Power
Attack Behavior
-ln(P)
OFC
PL
IL
LS
NAcc
MeA
LHb
MDThal
VMHvl
VHip
V1

Figure 5


a
Stimulated Windows (%)
Attack Male
Non-Attack Female
Male-Attack
Male-Non Attack
Female-Non Attack
Interaction Time (sec)
Yellow
Blue
n.s.
b
Nonsynchronous Stimulation
Network Encoder
Recording
Stimulation
Laser
Mouse 1 Session 1
Mouse 2 Session 2
c
Prelimbic Cortex
Infralimbic Cortex
PL->NAc
PL->MeA
PL->LHb
IL->NAc
IL->MeA
IL->LHb
20Hz
RLE
Frequency (Hz)
d
AAV-DIO-Chr2
PFC
retro-AAV-Cre
NAc
e
MeA
f
LHb