Brain-wide 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

Abstract

Social aggression is an innate behavior that can aid an organism in securing access to resources[1]. Aggression can also disrupt group function and reduce survival under conditions of 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 brain-wide network. This network is organized by prominent theta (4-11Hz) and beta (14-30Hz) oscillations, leading from orbital frontal cortex and medial dorsal thalamus, and converging on ventral medial hypothalamus and medial amygdala. Activity in this network couples to brain-wide cellular firing, and the network is conserved in multiple contexts associated with aggression. Strikingly, network activity during social isolation encodes the trait aggressiveness of mice and causal cellular manipulations known to impact aggression bidirectionally regulate network activity. Finally, we show that the network plays a role in mediating aggressive behavior through causal mediation analysis. Thus, we define a brain-wide 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 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 question 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 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 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 during this interval correlated with trait aggressiveness on a mouse-by-mouse basis. 4) 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

93 developed a closed loop optogenetic stimulation approach that detected when the brain
94 transitioned into an aggressive state, on a second-by-second basis, and stimulated the
95 prefrontal cortex in a manner previously known to suppress attack behavior. We also employed
96 a designer receptor exclusively activated by designer drug (DREADD) based approach to
97 selectively activate cells in ventral medial hypothalamus in a manner that has been shown to
98 induce attack behavior. These manipulations bidirectionally impacted network activity. 5) As
99 this latter manipulation was also performed in mice on a different genetic background, we
100 further confirmed the generalizability of our network-level findings. Thus, we firmly establish a
101 network-level mechanism whereby the brain encodes an aggressive state in a manner that
102 generalizes across context and individuals.

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

104 We initially focused our analysis on the medial prefrontal cortex (i.e., prelimbic and infralimbic
105 cortex in mice), since this brain region had been implicated in social behaviors [29-31].
106 Moreover, prior work had shown that optogenetic stimulation of medial prefrontal cortex was
107 sufficient to suppress attack behavior and increase non-aggressive social behaviors in CD1
108 strain mice [21]. CD1 mice show periods of attack behavior, defined by biting, boxing
109 (kicking/clawing), or tussling behavior, when a male C57BL6/J (C57) mouse is introduced into
110 their home cage [32]. On the other hand, they exhibit periods of non-attack social interactions
111 such as sniffing, grooming, or resting (placing nose or forepaws against the subject mouse, but
112 not moving) when exposed to a female intruder [16].

113 We performed our optogenetic stimulation experiments during these social encounters using a
114 protocol modeled after prior work, where blue light stimulation is used to activate
115 channelrhodopsin in the medial prefrontal cortex for the entirety of a social encounter (473nm,
116 5mW, 20Hz, 3ms pulse width, Fig. 1a) [21]. Control experiments were performed using yellow
117 light stimulation (593.5nm, 5mW, 20Hz, 3ms pulse width), which does not activate
118 channelrhodopsin 2[33, 34]. Consistent with the prior report, medial prefrontal cortex
119 stimulation suppressed attack behavior towards C57 male mice and tended to increase non-
120 attack social behavior [N=8; t_7=3.43; P=0.0055, and t_7=-2.35; P=0.051 using a two-tailed paired

121 t-test, for attack and non-attack behavior, respectively, significance determined by a Benjamini-
122 Hochberg false discovery rate (FDR) correction, Fig. 1a]. However, cortical stimulation
123 decreased non-attack social behavior toward female intruders (t_7=3.647, P=0.0041, using paired
124 t-test, significance determined by FDR correction). Thus, medial prefrontal cortex stimulation
125 was unable to selectively suppress attack behavior. Rather, stimulation suppressed multiple
126 types of social behavior.

127 To probe for candidate brain regions that may exhibit such selectivity, we implanted CD1 mice
128 across multiple cortical and subcortical brain regions known to contribute to social behavior,
129 including infralimbic cortex, orbitofrontal cortex, prelimbic cortex, lateral septum, nucleus
130 accumbens, lateral habenula, mediodorsal thalamus, ventromedial hypothalamus, medial
131 amygdala, primary visual cortex, and ventral hippocampus. Following surgical recovery, we
132 recorded neural activity while the CD1 mice freely interacted with an intact male C57 mouse
133 and a female C57 mouse for 300 seconds each. We repeated these encounters over six
134 sessions, yielding a total of 1800 seconds of neural data and behavior for each exposure (Fig.
135 1B, see also Supplemental Fig. S1). We recorded a subset of these CD1 mice (N=9 of the 20
136 mice) as they interacted with a castrated male mouse intruder. Since CD1 mice do not generally
137 exhibit attack behavior towards castrated males [35], this encounter provided neurophysiology
138 data during additional non-attack social behaviors that were unpaired paired with female
139 sensory cues. We also acquired neural activity while the CD1 mice were isolated in their home
140 cage.

141 We first verified that each of the implanted brain regions encoded social behavior using
142 discriminative cross spectral factor analysis non-negative matrix factorization (dCSFA-NMF, see
143 Fig. 1e) [27]. dCSFA-NMF utilizes supervised machine learning to generate a statistical model
144 that is both descriptive [integrates brain local field potential (LFP) activity features across time]
145 and predictive (discovers networks that distinguish between types of external behavior). LFPs
146 reflect the activity of populations of neurons, and these signals can be consistently sampled
147 across mice. The electrical functional connectomes networks (*Electome Networks*) generated
148 from dCSFA-NMF integrate LFP power (oscillatory amplitude across 1-56 Hz; a correlate of
149 cellular and synaptic activity within brain regions), LFP synchrony (how two regions' LFP

150 frequencies synchronize across time; a correlate of brain circuit function), and LFP Granger
151 synchrony (Granger causality testing; a correlate of directional transfer of information across a
152 brain circuit). Furthermore, dCSFA-NMF generates electome network activity scores (an
153 indicator of the strength of each network) at a temporal resolution sufficient to capture brain
154 states underlying the external behavior under observation (in this instance, a resolution of one
155 second). The electome networks are designed to learn patterns that explain interpretable
156 correlates of neural activity whose expression relate to measured behavior, facilitating an
157 overall interpretable model [28]. Any given brain region can belong to multiple electome
158 networks and each electome network may incorporate any number of brain regions. dCSFA-
159 NMF thus integrates spatially distinct brain regions and circuits into electome networks that
160 encode behavior.

161 To explore whether there was a generalized activity pattern within individual regions that
162 encoded social behavior, we designed a series of dCSFA-NMF models based on local field
163 potential (LFP) oscillatory power in frequencies from 1-56Hz. Each model was trained using
164 observations pooled from twenty CD1 mice to separate periods where mice were socially
165 isolated from periods where they were engaged in social behavior (e.g., attack behavior
166 towards the intact males and non-attack social behavior towards male and female mice). The
167 number of networks for each model was chosen by a validation procedure to balance how
168 much brain activity is explained by the model, how much the model predicts behavior, and the
169 parsimoniousness of the model (see methods). To facilitate easy interpretation of the learned
170 networks, one network was encouraged to preferentially encode isolation from attack and non-
171 attack social behavior (i.e., supervision). We tested the accuracy of the models in new CD1
172 mice recorded under all three conditions (N=9 mice).

173 We observed high accuracy in decoding periods of isolation from social behavior for each
174 implanted brain area ($p<0.05$ using Wilcoxon rank-sum, significance determined by FDR
175 correction for 44 comparisons, Fig. 1c). Next, we designed a new series of models to separate
176 one of the three social behavioral states from the other three (e.g., attack behavior toward
177 intact males vs. female, intact male, and castrated male non-attack interactions). These models
178 were built using observations pooled from the same twenty CD1 mice, and again based on LFP

179 power. We then tested the accuracy of the models in the same nine hold-out mice. Thus, we
180 trained and tested the model's generalizability to decode each class of social behavior from the
181 other three for each of the 11 implanted brain areas (i.e., 33 additional models). Using this
182 approach, we found that five of the brain region-based statistical models decoded attack
183 behavior versus non-attack social behavior: infralimbic cortex, lateral habenula, ventral
184 hippocampus, medial amygdala, and medial dorsal thalamus. V1 successfully decoded the non-
185 attack male interaction from the other social conditions as well ($P<0.05$ using one-tailed
186 Wilcoxon rank-sum test, significance determined by FDR correction for 44 comparisons, see Fig.
187 1c). None of the other implanted brain regions showed this selectivity. Thus, five regions
188 independently encoded attack behavior vs. non-attack social behavior in a manner that
189 generalized across mice.

190 Our broad goal was to identify a neural mechanism that could be used to suppress aggression
191 while leaving other social behaviors intact. Thus, we tested whether the putative attack codes
192 we discovered for the five regions generalized to another context associated with aggression.
193 Specifically, urine from other male mice has been found to elicit aggressive and dominance
194 behavior in CD1 males [22, 36, 37]. As such, the most aggressive mice from the training and
195 testing groups (N=8) were allowed to freely explore a clean inanimate object or an object
196 covered in urine from another intact CD1 male mouse (seven sessions). We then tested
197 whether each of the five attack codes could distinguish periods where mice explored the clean
198 objects from those where mice explored objects covered in urine. Though only ventral
199 hippocampus model tended to decode behavior in this new context, none of the brain regions
200 showed statistically significant encoding at the mesoscale-level (LFP) following correction for
201 the five comparisons ($AUC=0.56\pm0.06$, $P=0.04$ for ventral hippocampus, using one-tailed
202 Wilcoxon rank-sum test, significance determined by FDR correction for 5 comparisons, Fig. 1d).

203 <u>An aggressive state is encoded at the network-level</u>

204 After failing to robustly decode attack behavior using LFPs independently from any of the 11
205 brain regions, we established that a brain network integrated information across all the
206 implanted brain regions to encode an aggressive state. This network-level encoding mechanism

generalized to multiple new contexts associated with aggression. Critically, the network also encoded attack behavior with a predictive efficacy that exceeded independent ventral hippocampus activity.

For this analysis, we trained a new dCFSA-NMF model using data from all the implanted brain regions. The model utilized eight electome networks (see methods), one of which was trained to encode periods of attack-behavior (positive class) from social behavior in castrated male and female social context (negative class). We also included non-attack social behavior towards intact males in the negative class to discourage the network from simply learning non-aggressive sensory cues specific to the intact male (i.e., supervision, electome network #1; see Fig. 1e). We then validated this network using the set of nine holdout CD1 mice from our single area coding test analysis. Again, none of these mice were used to train the electome networks.

We found that electome network #1 successfully discriminated between attack behavior and non-attack social behavior in the test mice (N=9, Wilcoxon signed rank, p=0.0020, Fig. 1e, middle). Thus, the supervised network generalized to new animals.

Attack behavior is indicative of an aggressive brain state. We also reasoned that it was possible for a mouse to be in an aggressive internal state though it was not actively exhibiting attack behavior. Since such a context was likely to be present immediately prior to or following attack behavior, we tested whether network activity pooled from the 6 seconds surrounding social behaviors encoded the distinct social conditions (Fig. 2a). Critically, these data windows were not used to train the network model since they did not contain attack or non-attack social behavior. Network activity was lower in the intervals surrounding attack behavior compared to non-attack social interactions with males or females ($F_{3,67}$=28.6, P<0.0001 using Friedman's test, followed by P<0.05 using Wilcoxon signed-rank test, significance determined by FDR correction, Fig. 2b). Strikingly, network activity during periods of isolation also negatively correlated with the time mice spent exhibiting attack behavior towards other males (R=-0.58; P=0.016 using Spearman's rank correlation, Fig. 2b), encoding aggression on a mouse-by-mouse basis. Thus, electome network #1 (hereafter referred to as *EN-Aggression Inhibition - AggINH*) represented a network that putatively inhibited aggression when its activity was highest. In contrast to the

235 models developed for each of the brain regions independently, *EN-AggINH* activity also
236 encoded the exposure to male urine (N=8, Network Activity = 9.1±1.0 and 8.3±1.1 for clean and
237 urine covered objects, respectively; P=0.012 using one railed Wilcoxon signed-rank test;
238 AUC=0.57±0.02, data not shown). Thus, *EN-AggINH* generalized to a second aggression context.

239 Next, we tested the model for ventral hippocampus since we observed a trend towards
240 decoding attack behavior in the urine context. This model failed to encode the aggressive state.
241 Specifically, activity from hippocampus was statistically indistinguishable between periods
242 surrounding attack behavior and non-attack social interactions with females and activity during
243 social isolation ($F_{3,67}$=7.9, P=0.047 using Friedman's test; P=0.36 and 0.72, respectively, using
244 Wilcoxon signed-rank test, significance determined by FDR correction, Fig. 2b). Moreover, there
245 was no relationship between hippocampal activity during isolation and the innate
246 aggressiveness of mice (P=0.60 using Spearman's rank, Fig. 2b). Thus, only network activity
247 encoded the aggressive internal state.

248 <u>*EN-Aggression Inhibition* maps to interpretable biology</u>

249 *EN-AggINH* was composed of prominent theta frequency activity (4-11 Hz) in medial amygdala
250 and beta frequency activity (14-30 Hz) in medial amygdala and prelimbic cortex (Fig. 3a-b).
251 Prominent synchrony and directionality were also observed in the theta and beta frequency
252 bands. Indeed, the network showed strong directionality that led from orbital frontal cortex
253 and primary visual cortex, relayed through medial dorsal thalamus, and infralimbic cortex, and
254 converged on medial amygdala and ventral hippocampus (Fig. 3c, and Supplementary Fig. S3).

255 We verified that the activity of *EN-AggINH* truly reflects biological activity, by relating the
256 electome network to neural firing, as in previous work [38]. To achieve this, we analyzed the
257 correlation between cellular activity across the implanted brain regions and the activity of *EN-*
258 *AggINH*, as cell activity is an undisputed measure of biological function. We then used a
259 permutation test to rigorously test our findings (Fig. 3e). Specifically, we shuffled cellular firing
260 within social behavioral conditions, maintaining the relationship between cell firing and
261 behavior. We then repeated this procedure 1000 times to generate a null distribution for which
262 only 5% of cells would be expected to exhibit firing coupled to network activity. We found that

263 ~18% of cells showed firing that was coupled to the activity of *EN-AggINH*, far more than could
264 be explained by chance. Specifically, of the 186 cells recorded, nine (4.8%) showed firing activity
265 that was positively correlated with *EN-AggINH* and 25 (13.4%) showed activity that was
266 negatively correlated (Fig. 3f). Thus, we confirmed that *EN-AggINH* activity reflects the
267 dynamics of cellular activity across the brain.

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

269 To further validate *EN-AggINH*, we established that activity in this network was modulated by
270 orthogonal biological conditions that have been shown to induce or suppress aggressive
271 behavior in mice. In most cases, we performed this analysis in new animals, which is considered
272 the gold standard of model validation in machine learning by experimental manipulation [39].
273 We transformed LFP data recorded from these new sessions into our original network model
274 (Fig. 4a).

275 We first performed a high-threshold validation experiment that involved testing whether our
276 network generalized to new mice. These mice were on different genetic background and
277 subjected to a new aggression context (Fig. 4b). Specifically, this approach also used a validated
278 cellular manipulation that causally induces aggression under a behavioral condition that would
279 otherwise not yield aggressive attack behavior (i.e., we used female social partners). We
280 expressed an excitatory DREADD (AAV-hsyn-DIO-hM3Dq) in the ESR1+ cells of ventromedial
281 hypothalamus, since it has been shown that direct excitation of these cells induces aggressive
282 behavior towards female mice [17, 40, 41]. Experiments were performed in the male F1
283 offspring of female CD1 strain mice crossed with ESR1-Cre male mice on a C57 strain
284 background. Subsequently, we implanted the mice with recording electrodes to target the same
285 brain regions as our initial experiment used to train the network model. Following recovery, we
286 performed behavioral and neural recording when mice were exposed to a female mouse. The
287 experimental mice were either treated with saline or CNO (Clozapine N-oxide, which activates
288 the excitatory DREADD), in a pseudorandomized order, prior to the repeated testing sessions.

289 As anticipated, treatment with CNO induced attack behavior towards the female mice (N=8; P=
290 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 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. 4e), again demonstrating the *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.

EN-Aggression Inhibition mediates attack behavior

We tested whether a cellular manipulation known to causally decrease aggression increased network activity. For these experiments, we stimulated medial prefrontal cortex when *EN-AggINH* was naturally suppressed in the brain (signaling the aggressive state). First, we built a closed loop system that estimated the activity of *EN-AggINH* in real time (Fig. 5a). This approach employed a new network encoded solely based on power and coherence measures (i.e., a reduced network, Fig. 5b). While this new network lacked the interpretive power of dCSFA-NMF, it enabled us to predict aggression states in real time (Fig. 5c). In principle, when the activity of *EN-AggINH* fell below an established threshold (signaling the onset of aggressive behavior), our closed loop approach would deliver a one second light stimulation (5mW, 20Hz, 3ms pulse width) to prefrontal cortex. To verify that this real-time estimation system worked as designed, we tested whether light stimulation was triggered by a decrease in *EN-AggINH* activity. Indeed, network activity was significantly lower one second prior to stimulation than it was two seconds prior to stimulation, demonstrating that our approach successfully identified when the *EN-AggINH* activity decreased below the threshold that signaled the onset of aggression (N=9; P<0.005 for within-subject comparison of *EN-AggINH* activity 1 vs 2 seconds prior to yellow light stimulation using one-tailed signed-rank test, Fig. 5d). Importantly, we found that prefrontal cortex stimulation acutely increased *EN-AggINH* activity (N=9, P<0.01 for comparison of *EN-AggINH* activity one second after blue vs. yellow stimulation, using one-tailed

320 signed-rank test, see Fig. 5d). Thus, we confirmed that two cellular manipulations that causality
321 regulated aggression also bidirectionally impacted *AggINH* activity.

322 Next, we wondered whether increasing *EN-AggINH* activity via prefrontal cortex stimulation as
323 the brain transitioned into a putative attack state would suppress aggressive behavior. We also
324 hypothesized that this approach would not alter non-attack social behavior during interactions
325 with females. Specifically, since *EN-AggINH* activity was not naturally suppressed during female
326 interactions we anticipated that the closed loop protocol would deliver less stimulation to the
327 brain (see Fig. 5e). Using this approach, we found that closed loop stimulation of prefrontal
328 cortex significantly suppressed attack behavior (N=9 mice that were not used to train the initial
329 model; t_8=6.1, P=0.0003, Fig. 5f) and increased non-attack behavior ([N=9; t_8=-2.3, P=0.049
330 comparing blue vs. yellow light stimulation using two-tailed paired t-test for attack behavior
331 and non-attack behavior], significance determined by FDR correction, Fig. 5f) towards the C57
332 intact males. No differences in non-attack social behavior were observed during exposure to
333 female mice (t_8=0.74, P=0.48 using two-tailed paired t-test, significance determined by FDR
334 correction, Fig. 5f).

335 Next, we used statistical methods to determine if the change in *EN-AggINH* activity was a
336 mechanistic route for how closed loop stimulation regulated behavior. We employed
337 mediation analysis [42], which is a framework to determine whether the impact of a
338 "treatment" (stimulation) on an outcome (aggressive behavior) is mediated by a change in an
339 intermediate variable (*EN-AggINH* activity). If so, the intermediate variable is viewed, at least in
340 part, as a mechanistic route (a mediator) for how the treatment impacts the outcome. We first
341 set out to use the classic Baron and Kenny approach [42] to determine whether *EN-AggINH*
342 activity mediates the effect of neurostimulation on aggressive behavior. According to this
343 statistical approach, there is a mediated effect of network activity on behavior if three
344 conditions are met: 1) stimulation modulates network activity, 2) network activity correlates
345 with behavior, and 3) modeling the behavior from network activity and stimulation together is
346 better than modeling behavior from stimulation alone. We previously identified a significant
347 direct effect of stimulation on attack behavior (P<0.005, Fig. 5f) and network activity (P<0.0005,
348 Fig. 5d). In order to match the conditions between the treatment and control cases as closely

349 as possible, we chose to use windows during the closed-loop stimulation procedure where the
350 laser was triggered, and then compared blue laser stimulation (treatment) to yellow laser
351 stimulation (control). A statistical model of behavior using network activity and stimulation (see
352 Fig. 5g, model 2) significantly outperformed the model using only stimulation (see Fig. 5g,
353 model 1; nested logistic regression models, $P<0.01$, likelihood ratio test), satisfying the
354 necessary conditions to show that *EN-AggINH* is a mediator.

355 After establishing that *EN-AggINH* activity mediated the impact of PFC stimulation on behavior,
356 we set out to evaluate the significance of the average causal mediation effect (ACME) and the
357 average direct effect (ADE) during the same stimulated closed-loop windows using causal
358 mediation analysis [43]. ACME is the causal effect of stimulation on behavior due to the change
359 in *EN-AggINH* activity (see. Fig 5g, model 3), and ADE is the causal effect on behavior from
360 prefrontal cortex stimulation not explained by the change in *EN-AggINH* activity. We found that
361 there was a significant ACME ($P<0.01$), but not a significant ADE ($P=0.48$). This analysis
362 suggested that *EN-AggINH* activation is the primary mechanism whereby prefrontal cortex
363 stimulation suppresses aggression.

364 We also performed an additional control experiment to verify that the suppression of
365 aggressive behavior observed in response to closed loop stimulation was due to the
366 synchronization of the light stimulation and endogenous *EN-AggINH* activity, and not simply
367 due to the dynamic pattern of stimulation delivered using this method. Here, we used the
368 stimulation patterns from our closed loop experiments to drive stimulation in a new group of
369 animals. Thus, for these sessions, prefrontal cortex stimulation occurred in a manner that
370 mirrored our closed loop stimulation experiments, except that stimulation was not fixed to
371 transitions into the aggressive internal state (i.e., open loop – nonsynchronous; Fig. 5h).
372 Strikingly, this pattern of stimulation did not impact aggressive behavior ($t_7=0.64$, $P=0.54$ using
373 paired t-test; see Fig. 5i). Together with our mediation analysis, these results suggest that *EN-*
374 *AggINH* activity plays a mechanistic role in mediating aggression.

375 Finally, since prior studies had also shown that prefrontal cortex stimulation induced
376 hyperactivity [44, 45], we tested whether the suppression of aggressive behavior that resulted

377 from prefrontal cortex stimulation was also associated with behavioral hyperactivity. While we
378 found that 20Hz tonic stimulation of prefrontal cortex induced behavioral hyperactivity in the
379 CD1 mice (N=10; T_9= 3.221 and P = 0.0052 for Velocity; T_9= 3.213 and P=0.0053 for distance
380 traveled, using one-tailed paired t-test; Fig. 5j), closed loop stimulation of prefrontal cortex had
381 no impact on gross locomotor behavior (N=9; T_8= 0.584 and P=0.29 for velocity; T_8= 0.86 and P=
382 0.21 for distance traveled, using one-tailed paired t-test; Fig. 5j). Thus, closed loop stimulation
383 suppressed aggressive behavior, but it did not induce the off-target effects of impaired pro-
384 social behavior and psychomotor activation that we observed with the classic 20Hz stimulation
385 paradigm.

386

387 **Discussion**

388 Behavior results from the integration of sensory information with the internal state of an
389 animal. While many cellular substates that contribute to social behavioral outputs have been
390 identified, less is known regarding the internal brain states that map sensory inputs to these
391 behaviors. Here, we set out to discover the internal state that regulates whether an animal will
392 exhibit aggressive or pro-social behavior. We reasoned that attack behaviors emerge from an
393 aggressive internal brain state. Thus, we used machine learning to discover the mesoscale
394 neural architecture of the brain when an animal exhibited attack vs. non attack social
395 behaviors. Like other well defined internal brain states, such as sleep, we found that the
396 network distinguishing attack behavior incorporated state dependent patterns of neural activity
397 across every brain region we measured. For multiple regions, differences were observed in local
398 oscillatory power, while others exhibited differences in oscillatory synchrony with a broader
399 collection of regions. Each brain region showed selectivity in the frequencies of oscillations that
400 contributed to the network. For example, prelimbic cortex showed strong activity in the beta
401 frequency range, while medial amygdala showed strong activity in the theta frequency range.
402 No brain region showed prominent activity contributions across all frequencies. We also
403 observed differences in the activity profile of a primary sensory region, V1, which may reflect a
404 change in encoding, or differences in sensory input observed during attack behavior. Critically,

405 the brain state identified during attack behavior was better captured by observing activity
406 across all the brain regions as an integrated network, than observing independent activity
407 within each brain region.

408 Though behavioral output has been classically utilized to infer the internal state of a brain, we
409 reasoned that an internal brain state was also likely present during intervals immediately
410 surrounding behavioral output. Thus, we tested whether the 'attack' network showed distinct
411 activity profiles in the time intervals surrounding attack and non-attack social behaviors.
412 Indeed, network activity was lower during interval surrounding attack behavior. Strikingly, we
413 also found that network activity when animals were isolated in their home cage encoded their
414 trait aggression. Thus, the network did not simply encode behavioral output since it was
415 observed outside of attack behavior. Rather the network encoded an aggressive internal brain
416 state.

417 Interesting, this aggressive brain state was encoded by decreased activity in the network. This
418 observation does not indicate that overall brain activity is suppressed during aggressive states.
419 Rather, our findings argue that the aggressive state is encoded by a network that decreases its
420 activity relative to when mice are socially isolated or engaged in pro-social behavior. Indeed,
421 our data suggested that several common regions/circuits were activated during aggressive and
422 pro-social behavior. These common circuits need not be reflected in our network since our
423 model was trained to differentiate attack vs. non-attack social behavior. Nevertheless, our
424 discovery of a network that decreased its activity during aggression raises the intriguing
425 hypothesis that this brain activity inhibits aggression during pro-social engagement. When
426 activity in this inhibition network is suppressed, aggressive behavior emerges.

427 This interpretation is supported by our validation experiments where we directly activated
428 ESR1+-Cre neurons in ventromedial hypothalamus. Our findings showed that direct activation
429 of these cells induced the aggressive brain state (suppressed *EN-AggINH* activity). When mice
430 treated with CNO were exposed to a stimulus that would generally produce non-attack social
431 behavior (i.e., a female mouse), attack behavior emerged. Thus, the presence of the aggressive
432 brain state changed the mapping between sensory input and behavior output. Similarly, direct

433 stimulation of medial prefrontal cortex biased mice towards exhibiting non-attack social
434 behavior when they were exposed to a stimulus that would generally induce attack behavior
435 (i.e., a male intruder). Our findings showed that medial prefrontal cortex stimulation decreased
436 the aggressive internal state (increased *EN-AggINH* network activity). Critically, our findings
437 using mediation analysis argue that the brain state represented by *EN-AggINH* is
438 mechanistically involved in the translation of medial prefrontal cortex stimulation to a
439 suppression of attack behavior.

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

453 Our closed loop stimulation approach was developed using a neural-network based
454 approximation technique that was substantially constrained relative to dCSFA-NMF.
455 Specifically, the reduced encoder was developed solely using power and coherence features.
456 This approach was necessary because the processing time to calculate granger directionality
457 was prohibitive for real time implementations. As such, the reduced encoder sacrifices some
458 predictive efficacy for speed. Nevertheless, we found that the reduced encoder was sufficient
459 to identify the precise time windows where the brain transitioned into aggression, as marked
460 by a decrease in *EN-AggINH* activity. In the future, novel approaches may allow for further
461 improvement in the precision of our real time stimulation approach. For example, future work

462 could exploit convolutional neural networks to bypass the feature extraction step. These neural
463 network encoders could be altered to predict both aggressive and pro-social states, such as the
464 generalized social appetitive network that we recently discovered [7]. By using both networks
465 concurrently to actuate a closed loop system, it may be possible to further suppress aggressive
466 behavior relative to pro-social behavior. Indeed, our current findings also pointed to a network
467 that exhibits increased activity during aggressive behavior (Electome Network 6, see Fig. 2b,
468 and Supplementary Fig. S4). This network showed strong encoding in the nine hold out test
469 mice (AUC=0.65±0.03, $P<5\times10^{-5}$ using Wilcoxon rank sum), and activity in the network also
470 generalized to the experiment in which we causally manipulated ventromedial hypothalamus
471 activity. Though the network failed to encode the urine paradigm, it is possible that it contains
472 activity that synergizes with *EN-AggINH* to encode aggressive social states more optimally. If
473 future studies demonstrate this potential, imitation encoders for both Electome Network 6 and
474 *EN-AggINH* could be integrated to further optimize closed loop approaches to selectively
475 suppress aggression.

476 Multiple neuropsychiatric disorders including mood disorders, psychotic disorders,
477 neurodevelopmental disorders, and neurodegenerative disorders are associated with deficits in
478 regulating social behavior, including aggression. While multiple pharmacological approaches
479 have been instituted to suppress aggressive behavior towards self and others, many of these
480 strategies act by sedating the individual and can disrupt aspects of pro-social function. Our
481 discovery of a brain network that encoded an aggressive state raises the potential for novel
482 approaches to suppress aggressive behavior that spare pro-social behavior. Indeed, compared
483 to a standard open loop stimulation protocol (20Hz stimulation) which suppressed both attack
484 and non-attack pro-social behavior, our closed loop stimulation approach spared non-attack
485 social behavior towards males or females. Intriguingly, the classic open loop stimulation
486 pattern (20Hz stimulation) also induced behavioral hyperactivity in experimental mice, while
487 our closed loop stimulation protocol did not. Thus, our findings also show that closed loop
488 stimulation may limit off target behavioral effects that are induced by classic stimulation
489 approaches.

Overall, our findings establish a generalized network-level mechanism 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 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 brain-wide 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. Figure shows 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 encoding of social states by individual brain regions (left). All implanted regions encoded social engagement; however, only five selectively encoded the attack behavior (right). Pink shading indicates P<0.05 with FDR correction. **d)** Attack codes discovered from the five brain regions failed to encode aggressive behavior induced by male urine (Gray shading indicates P<0.05 prior to but not following FDR correction). **e)** Schematic of machine-learning model used to discover network encoding attack

516 behavior (left). Encoding across eight learned networks (right). **f)** The supervised network
517 (purple, *EN-AggINH*) showed the strongest encoding. Data shown as mean ± SEM.

518 **Figure 2. *EN-Aggression Inhibition* encodes an aggressive internal state.** Neural activity was
519 sampled while mice were socially isolated and during intervals surrounding social behavior.
520 Network activity during this interval encoded attack behavior vs. male and female non-attack
521 social behaviors, while ventral hippocampal activity did not (P<0.05 using Freidman's test
522 followed by Wilcoxon sign-rank test). Network during isolation, but not ventral hippocampus
523 activity, encoded the subsequent total attack time of individual mice (P<0.05 using Spearman's
524 Rank Correlation).

525 **Figure 3. Dynamics and biological components of *EN-Aggression Inhibition*. a)** Prominent
526 oscillatory frequency bands composing *EN-AggINH* are highlighted for each brain region around
527 the rim of the circle plot. Prominent synchrony measures are depicted by lines connecting brain
528 regions through the center of the circle. The plot is shown at relative spectral energy of 0.4.
529 Theta (4-11 Hz) and beta (14-30 Hz) frequency components are highlighted in blue and green,
530 respectively. **b)** Example relative LFP spectral energy plots for three brain regions corresponding
531 to the circular plot in A (See Supplemental Figure S3 for full description of network features). **c)**
532 Granger offset measures were used to quantify directionality within *EN-AggINH*. Prominent
533 directionality was observed across the theta and beta frequency bands (shown at spectral
534 density threshold of 0.4 and a directionality offset of 0.3). Histograms quantify the number of
535 leading and lagging interactions between brain regions. **d)** Schematic depicting directionality
536 within *EN-AggINH*. **e-f)** *EN-AggINH* maps to cellular activity. e) A representative cell showing
537 firing activity that is negatively correlated with *EN-AggINH* activity. f) *EN-AggINH* activity
538 correlated with cellular firing across the brain across the brain.

539 **Figure 4. *EN-Aggression Inhibition* encodes distinct aggression contexts. a)** Strategy for
540 validating *EN-AggINH* using additional aggression paradigms. **b)** Experimental approach for
541 causally inducing aggression via direct cellular manipulation**. c-e)** Direct activation of ESR1+ cells
542 in ventromedial hypothalamus. c) Cellular activation induced attack behavior towards female
543 mice (P<0.001 using sign-rank test), d) decreased *EN-AggINH* activity across during social

544 interactions with female mice (P<0.01 using one-tailed Wilcoxon sign-rank test) and e) intervals
545 surrounding these interactions (P<0.05 using Wilcoxon sign-rank test).

546 **Figure 5. *EN-Aggression Inhibition* activity is causally related to aggression. a)** Schematic for
547 closed loop manipulation of *EN-AggINH* activity. **b)** Real-time estimation of aggression. Receiver
548 operating characteristic depicting detection of aggressive behavior in a mouse using *EN-AggINH*
549 activity vs. real-time reduced encoder is shown to the right. Dashed blue line corresponds to
550 the established detection threshold. **c)** Detection of aggression using reduced encoder vs. *EN-*
551 *AggINH* across mice (N=9; P=0.43 using two-tailed paired Wilcoxon sign-rank). **d)** *EN-AggINH*
552 activity relative to light stimulation during closed loop manipulation. Network activity
553 significantly decreased one second prior to yellow light stimulation (N=9, $^{\#\#}$P<0.005 using one-
554 tailed sign rank test; note that activity was normalized for each mouse to the average network
555 activity during isolation). Activity was also higher one second after stimulation with blue light
556 vs. yellow light (**P<0.01 using one-tailed signed-rank test). **e)** Portion of windows stimulated
557 during social behaviors (P=0.002 using one-tailed sign-rank test). **f)** Behavioral effects of closed
558 loop stimulation (**P<0.005 using paired t-test, significance determined using FDR correction).
559 **g)** Directed graph with the inferred modes of action derived from mediation analysis. There is a
560 causal relationship from stimulation to behavior and stimulation to *EN-AggINH* expression
561 (model 1; P<0.01 using signed rank and paired t-tests). *EN-AggINH* is a mediator from
562 stimulation to behavior (P<0.01 using nested logistic regression models, likelihood ratio test;
563 model 2), *EN-AggINH* activation is the primary mechanism whereby prefrontal cortex
564 stimulation suppresses aggression (P<0.01 using average causal mediation effect, model 3). **h)**
565 Schematic for nonsynchronous control stimulation. **i)** Nonsynchronous stimulation does not
566 impact aggressive behavior (P>0.05 using paired t-test). **j)** Impact of open loop (20Hz, fixed) and
567 closed loop stimulation on locomotor behavior (P**<0.01 using paired two-tailed t-test).

568

569 ## Author contributions

570 Conceptualization and Methodology – YSG, AT, NMG, SR, DEC, and KD; Formal Analysis – YSG,
571 AT, NMG, AJF, KKW, DEC, and KD; Investigation – YSG, AT, NMG, GET, AJF, KKW, SR, DEC, and

572 KD; Resources – YSG, AT, NMG, GET, AJF, KKW, SR, DEC, KD; Writing – Original Draft, YSG, AT,
573 NMG, SR, DEC, and KD; Writing – Review & Editing, YSG, AT, NMG, KKW, SR, DEC, and KD;
574 Visualization – YSG, AT, DEC, and KD; Supervision –KKW, DEC, and KD; Project Administration
575 and Funding Acquisition –SR, DEC, and KD; See Supplemental materials for detailed author
576 contributions.

577

578 **Declaration of Interests**

579 The authors declare no competing interests

580 **Materials and Methods**

581 Animals care and use

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

586 Six-month retired breeder male CD1 strain mice (Charles River Laboratories, Wilmington,
587 Massachusetts) were used to discover a network that encoded aggression, hereafter called *EN-*
588 *AggINH*. These CD1 mice were also used to probe the behavioral and network responses to
589 optogenetic stimulation. Mice were singly housed with enrichment. ESR1-Cre mice on a
590 C57/Bl6J background were provided by Scott Russo. These mice were crossed with CD1 females
591 in the Bryan Vivarium at Duke University to obtain F1 offspring. Fourteen-week-old virgin ESR1-
592 cre F1 male offspring were used to validate *EN-AggINH*. All F1 offspring were group-housed 2-5
593 mice per cage until they received viral injections in the ventromedial hypothalamus at 7-8
594 weeks. After surgery, these mice were singly housed with enrichment. All partner mice (intact
595 males, female, and castrated males) were 7-14 weeks old. These mice were purchased from
596 Jackson Laboratories (Bar Harbor, Maine) and maintained on a C57BL/6J strain background. All
597 stimulus mice were housed 5 per cage with enrichment. All behavior testing and
598 neurophysiological recordings occurred during the dark cycle.

599 Castration of C57 male mice

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

606 Electrode implantation surgery

607 Mice were anesthetized with 1% isoflurane and placed in a stereotaxic device. Grounding
608 screws were placed above the cerebellum, right parietal hemisphere, and anterior cranium. The
609 recording bundles were designed to target prelimbic cortex, infralimbic cortex, medial
610 amygdala, ventral hippocampus, primary visual cortex, mediodorsal thalamus, lateral habenula,
611 lateral septum nucleus, nucleus accumbens, ventrolateral portion of the ventromedial
612 hypothalamus, and orbitofrontal cortex were centered based on stereotaxic coordinates
613 measured from bregma. [Orbital frontal cortex: anterior/posterior (AP) 2.35mm, medial/lateral
614 (ML) 1.0mm, dorsal/ventral (DV) from dura -2.75mm; infralimbic cortex and prelimbic cortex:
615 AP 1.8mm, ML 0mm, DV -2.7mm from dura; medial amygdala: AP -1.25, ML 2.7mm, DV -4.3
616 from dura; lateral septum and nucleus accumbens: AP 1.0mm, ML 0mm, DV -4.0mm from dura;
617 ventromedial hypothalamus, lateral habenula, and medial dorsal thalamus: AP -1.47mm, ML
618 0mm, DV -5.4mm from dura; central hippocampus and primary motor cortex: AP -3.0mm, ML
619 2.6mm, DV -3.0mm from dura]. We targeted infralimbic cortex and prelimbic cortex by building
620 a 0.6mm DV stagger into the bundle. We targeted lateral septum and nucleus accumbens by
621 building a 0.3mm ML and 1.5mm DV stagger into the bundle. We targeted lateral habenula,
622 medial dorsal thalamus, and ventral medial hypothalamus by building a 0.3mm ML and
623 am1.9mm and 2.5mm DV stagger into our electrode bundle microwires. We targeted primary
624 motor cortex and ventral hippocampus using a 0.3mm ML and 2.5mm DV stagger in our
625 electrode bundle microwires. For optogenetic stimulation experiments, the addition of a Mono
626 Fiberoptic Cannula coupled to a 2.5mm metal ferrule (NA: 0.22, 100mm [inner diameter],
627 125mm buffer [outer diameter], MFC_100/125-0.22, Doric Lenses, Quebec) was built into the
628 prefrontal cortex bundle. The tip of the fiber was secured 300mm above the tip of the IL
629 microwire. Mice were allowed 10-15 days for to recovery from surgery before behavioral
630 testing.

631 Viral surgery

632 For optogenetic experiments [21], we used CD1 mice that showed an attack latency < 60s when
633 exposed to an intact C57 male. CD1 mice were anesthetized with 1% isoflurane and placed in a
634 stereotaxic device. The mice were unilaterally injected with AAV2-CamKII-Chr2-EYFP (purchased
635 from the Duke Viral Vector Core, Durham, NC; courtesy of K. Deisseroth), based on stereotaxic
636 coordinates from bregma (left Infralimbic cortex: AP 1.8mm, ML 0.3mm, DV -2.0mm from the
637 brain). A total of 0.5mL of virus was infused at the injection site at a rate of 0.1mL/min over five
638 minutes and the needle left in place for ten minutes after injection. For the open loop
639 stimulation experiment, CD1 mice were implanted with an optic fiber (Mono Fiberoptic Cannula
640 coupled to a 2.5mm metal ferrule (NA: 0.22, 100mm [inner diameter], 125mm buffer [outer
641 diameter], MFC_100/125-0.22, Doric Lenses, Quebec)) 0.3mm above the injection site
642 immediately after viral syringe was removed. These mice were allowed 3 weeks for recovery
643 prior to behavioral testing. For the closed-loop experiments, CD1 mice were allowed 3 weeks
644 for viral expression prior to implantation with an optrode.

645 For the ESR1-Cre validation experiment, F1 offspring were bilaterally injected with AAV2-hSyn-
646 DIO-GqDREADD (obtained from Addgene) based on stereotaxic coordinates measured from
647 bregma (AP -1.5mm, ML ±0.7mm, DV -5.7mm from the dura). A total of 0.3mL of virus was

infused bilaterally at a rate of 0.1mL/min and the needle left in place for five minutes after injection. Two weeks after viral infusion, F1 males were screened for aggressive behavior towards females. The F1 males received i.p. injections of CNO (1mg/kg) at the start of the screening session. Thirty-five minutes after injection, a novel C57 female was placed in the home cage for 5 minutes. Screening was repeated one week and two weeks later. Only F1 males who attacked females for at least two of the three screening sessions (9/13 mice) were implanted with electrodes [17]. The eight mice that showed good surgical recovery were subjected to further experiments.

Histological analysis

Histological analysis of implantation sites was performed at the conclusion of experiments to confirm recording sites and viral expression. Animals were perfused with 4% paraformaldehyde (PFA) and brains were harvested and stored for at least 96 hrs in PFA. Brains were cryoprotected with sucrose and frozen in OCT compound and stored at -80C. Brains were later sliced at 40mm. Brains from mice used to train and validate the network were stained using NeuroTrace fluorescent Nissl Stain (N21480, ThermoFisher Scientific, Waltham, MA) using standard protocol. Specifically, Nissl staining for brain tissue occurred on a shaker table at room temperature. Tissue was washed in PBST (0.1% Triton in phosphate-buffered saline solution) for 10 minutes. It was then washed for five minutes in PBS twice. The tissue was then protected from light for the remainder of the protocol. The tissue was incubated in 1:300 Nissl diluted in 2 mL PBS for 10 minutes. After the Nissl incubation, tissue was washed once in 0.1% PBST for 10 minutes, then twice in PBS for 5 minutes. Brains from ESR1 mice and mice used for 20 Hz or closed loop stimulation were mounted in Vectashield mounting medium containing DAPI (H-1200-10, Vector Laboratories, Newark, CA). Images were obtained at 10x using an Olympus fluorescent microscope. Of the 297 total implantation sites in the training and testing set of mice, 17 were mistargeted (~5.7% error rate). Of these mistargeted implants, 13 were within 200µm of the targeted structure. Given our prior work demonstrating high LFP spectral coherence (in the 1-55Hz frequency range) across microwires separated by 250um, in both cortical and subcortical brain regions [38], we chose to retain these animals in our analysis. The other four mistargeted implants were within 350µm of the targeted structure. The most reliably mistargeted site was ventral medial hypothalamus for which 4 animals were implanted within 200um of the target, and 2 animals were implanted within 350 µm of the target.

Machine learning analysis typically benefits from larger data sets. Thus, we concluded that maintaining a higher number of data points likely outweighed the effect of a small number of mistargeted brain regions, particularly since our LFP measures were robust to the targeting inaccuracies we observed histologically. As such, we pooled data from all 20 implanted animals to learn our initial model. We employed a similar strategy for our validation analysis, where an animal was only removed from the validation set if there was clear histological confirmation of mistargeting >200µm for any of the recorded regions. Specifically, presuming accurate targeting with 94.3% certainty and targeting within 200um at a higher certainty, we included animals with missing or damaged histological slices in our analysis. However, if there was clear histological confirmation of mistargeting for any of the recorded regions, the animal was removed from the validation testing. Critically, our validation procedure implies that the

690 machine learning models were robust regardless of any slight imprecision in the animals we
691 utilized for training.

692 Neurophysiological data acquisition

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

706 Behavioral recordings and analysis for training/testing models

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

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

724 After screening, mice were implanted, and data was acquired across 1-6 behavioral
725 testing/recording sessions following recovery. Sessions were separated by 5-7 days. Recordings
726 for all social encounters were performed in the home cages of the CD1 mice. Each behavioral
727 testing session began with 5 minutes of recording prior to introduction of the social stimulus.
728 All mice were subjected to encounters with an intact C57 male mouse and a female c57 mouse.
729 A subset of eighteen mice were also subjected to an encounter with a castrated male mouse,

730 and another subset of eighteen mice were subjected to exposure to objects covered in CD1
731 mouse urine. Six mice were recorded under all four conditions. Data observations (1 second
732 each) were pooled across eleven CD1 mice for training the network model. Object pairs
733 included yellow duplex blocks, curved red duplex blocks, weighted 5 mL conicals, glassware
734 tops, and objects assembled from black legos®. The CD1 mice were exposed to a different pair
735 of objects during each session. Order of exposure to stimulus mice and objects was shuffled for
736 every session.

737 For ESR1 male behavioral testing, mice were injected with either saline or CNO (1mg/kg, i.p.)
738 after the five-minute baseline recording. Thirty-five minutes after this injection, mice were
739 exposed to an intact male C57, a castrated male C57, and a female C57, presented in
740 pseudorandom order. Mice were subjected to six total recording sessions (three in which they
741 were treated with saline and three in which they were treated with CNO), again in
742 pseudorandom order. Sessions were separated by 5 days to allow an adequate washout of
743 CNO[46].

744 Behavior was scored for each second as an “attack”, “non-attack social interaction”, or “non-
745 interaction”. One second windows were identified as "aggressive" if the mouse was engaged in
746 biting, boxing (kicking/clawing), or tussling behavior [32]. Windows were labeled as "non-attack
747 social interaction" if the mouse had his nose or forepaws touching the stimulus mouse (intact
748 male/female/castrated male) or object, but was not biting, boxing, or tussling. Examples of
749 behaviors labeled "non-attack social behavior" included sniffing, grooming, or resting (placing
750 nose or forepaws against the subject mouse, but not moving). If the stimulus mouse had
751 his/her forepaws or nose on the CD1, but it was not reciprocated, this was labeled "non-
752 interaction". CD1 straight approach, sideways approach, and chasing of the stimulus mouse
753 could result in attack (biting/kicking/tousling), non-attack social behavior (nose or paw touch),
754 or withdrawal without any touch. Thus, while sideways approach and chasing are regularly
755 labeled as "aggressive" in the literature [32, 47, 48], and straight approach is regularly labeled
756 as "pro-social", these behaviors lacked consistent resolutions. Moreover, mice also
757 demonstrated these behaviors towards female and castrated mice (non-attack social context).
758 One second windows containing these behaviors were labeled "non-interaction". All other
759 timepoints not labeled "attack" or "non-attack social" were also labeled "non-interaction".
760 These behavioral criteria were selected to include ethologically aggression-related behaviors
761 and maximize the likelihood that the CD1 was aware of the presence of the stimulus mouse or
762 object during the behavioral window, while remaining confident in the classification of "attack"
763 and "non-attack social" window labels.

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

The videos used to generate the labels for training and testing our machine learning model was hand-scored by a trained researcher. Videos from ESR1 mice and optogenetic stimulation were automatically tracked using DeepLabCut [49, 50]. This information was then used for creating behavioral classifiers in SimBA [51].

LFP preprocessing and signal artifact removal

Each LFP signal was segmented into 1s non-overlapping windows. If there were multiple intact channels implanted in a region, they were averaged to produce a single signal. Windows with non-physiological noise were removed using an established automated heuristic [7]. Briefly, the envelope of the signal in each channel was estimated using the magnitude of the Hilbert transform. The Median Absolute Deviation (MAD) of the magnitude was then calculated on each channel of each recording. Signal was marked as non-physiological if the envelope exceeded a high threshold (5x MAD, which is roughly 4x the standard deviation for a normally distributed signal). Any data adjacent to non-physiological data that had an envelope value above a smaller threshold (0.167 MAD) was also considered non-physiological. All data marked in this way was ignored when averaging channels for each region. Any channels with standard deviation less than 0.01 was removed as well. If no channels were usable for a given region within a window, that whole window was removed from the data. This set of heuristics resulted in 34.7±5.1% of the data being excluded from analysis. After this, 60Hz line artifact was further removed using a series of Butterworth bandpass filters at 60Hz and harmonics up to 240Hz with a stopband width of 5Hz and stopband attenuation of -60dB. Finally, the signal was downsampled to 500Hz.

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

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

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

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

Spectral Granger causality features were estimated using the Multivariate Granger Causality (MVGC) MATLAB toolbox [52]. 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, 53] 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

808 feature is a ratio of total power to unexplained power. Exponentiated Granger feature values
809 were truncated at 10 to prevent implausible values.

810 Data for single region and network-level machine learning analyses

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

817 Single region decoding

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

831 Discriminative Cross-Spectral Factor Analysis – Nonnegative Matrix Factorization

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

837 We used Discriminative Cross-Spectral Factor Analysis – Nonnegative Matrix Factorization
838 (dCSFA-NMF) model [27]. This approach assumes each window of is an independent stationary
839 observation and examines dynamics in brain activity only at the scale of windows. A 1 second
840 window was chosen to balance capturing fine-grained transient behavior with sufficient length
841 to properly estimate spectral features [7]. Each window has associated spectral power,
842 coherence, and Granger causality features (in total $p = 9{,}586$ features), which is represented
843 as $x_i \in \mathbb{R}^p$ for the ith window. Each window was associated with a behavioral label that
844 identified which condition the CD1 mouse was subjected to during that window (intact male,
845 castrated male, or female) and whether the CD1 mouse was engaged in aggressive or non-
846 aggressive behavior during that window, and the aggressive behavior was coded as $y_i \in \{0,1\}$.

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

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

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

Previous work has also found that the reconstruction loss can reduce overfitting and make the learned predictions more robust [54]. To further reduce overfitting of the predictive aspect of the encoder, we applied an elastic net loss [55] on the encoder $Loss_{EN}$ with a weighting μ and the ratio between the L_1 and L_2 losses set to .5. The value for μ was set to a small value that had worked well previously. In this work, power features were also upweighted by a factor of 10 to accommodate that there were many more Granger features and truncated at 6 to prevent outliers from dominating the predictions.

These models and statistical analyses were implemented with Python 3.7 and Tensorflow version 2.4. Parameters were learned with stochastic gradient descent using the Nesterov accelerated ADAM optimizer [56]. Learning was performed for 30000 iterations, which was observed to be ample for parameter convergence. The learning rate and batch size were set to 1e-3 and 100 respectively, values that have empirically performed well in similar applications.

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

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.

Hyper-parameter selection

The dCSFA-NMF procedure requires selection of several settings in the algorithm. Specifically, we must choose the number of electome factors K, the importance of the supervised task λ,

the relative importance of the power features, coherence features, and Granger features, and the parameterization of the mapping function $f(x_i; \phi)$. Besides K, these settings were chosen to match previously used values or follow heuristics.

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

Validating Model Dimensionality

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

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

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

Decoder Information Content

The amount of information contained in the predictive model was quantified by the reduction in uncertainty. The associated formula for this reduction in uncertainty, known as the Bernoulli entropy, is 1-1*(p log_2(p) +(1-p)log_2(1-p)), where p is the accuracy of the model. At the extremes, an accuracy of 0.5 (random guessing) removes no uncertainty, whereas an accuracy of 1 or 0 completely eliminates uncertainty.

Single cell correlation to Electome Factor activity

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

Real-Time Encoder Approximation

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

Optogenetic stimulation

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

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

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

their individually matched closed-loop mouse. Light stimulation was delivered using the pattern implemented for the closed-loop partner mouse.

Mediation Analysis

For the Baron and Kenny approach [42] to establish that *EN-AggINH* expression mediated the behavioral effect of the neurostimulation, we first used two results previously described in the methods to establish that there was an effect from the neurostimulation on network expression and on behavior. We next limited the data used to the most relevant case, which is on the closed-loop stimulation. We limited ourselves to windows of LFP data during the closed-loop experiment where the laser was activated, either from the blue laser or the yellow laser to match the cases between the treatment and control as closely as possible. As only blue light should significantly manipulate neural activity, this is viewed as the treatment, and the yellow light is set as the control. We followed the procedures outlined in "LFP preprocessing and signal artifact removal" to preprocess the data and remove data with significant artifacts. *EN-AggINH* expression was calculated by projecting the data into the learned model. The remaining data was then fit into two logistic regression models to predict behavior using the statsmodel package in python [57]. The first model used only the stimulation to predict behavior (behavior ~ const + stimulation), and the second model used stimulation and network expression to predict behavior (behavior ~ const + network_expression + stimulation). These two models were compared by using a likelihood ratio test to evaluate whether the second model was significantly better.

For the causal mediation analysis, we again need to roughly balance treatment and control groups. We used the same data as described above in the classic mediation analysis. We define the treatment as blue versus yellow light stimulation, the mediator as *EN-AggINH* expression, and the outcome as aggressive versus non-aggressive behavior. Data was screened to remove windows with missing data. These data were then used in the causal mediation analysis approach proposed by Kosuke, Keele, and Tingley [43] by using the statsmodels package in python [57].

Statistics

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

References

1002 1. Lindenfors, P. and B.S. Tullberg, *Evolutionary aspects of aggression the importance of*
1003 *sexual selection.* Adv Genet, 2011. **75**: p. 7-22.

1004 2. Lee, S.C., T. Yamamoto, and S. Ueki, *Characteristics of aggressive behavior induced by*
1005 *nucleus accumbens septi lesions in rats.* Behav Neural Biol, 1983. **37**(2): p. 237-45.

1006 3. Haller, J., *The role of central and medial amygdala in normal and abnormal aggression:*
1007 *A review of classical approaches.* Neurosci Biobehav Rev, 2018. **85**: p. 34-43.

1008 4. Bak, M., et al., *The pharmacological management of agitated and aggressive behaviour:*
1009 *A systematic review and meta-analysis.* Eur Psychiatry, 2019. **57**: p. 78-100.

1010 5. Hashikawa, K., et al., *The neural circuits of mating and fighting in male mice.* Curr Opin
1011 Neurobiol, 2016. **38**: p. 27-37.

1012 6. Chen, P. and W. Hong, *Neural Circuit Mechanisms of Social Behavior.* Neuron, 2018.
1013 **98**(1): p. 16-30.

1014 7. Mague, S.D., et al., *Brain-wide electrical dynamics encode individual appetitive social*
1015 *behavior.* Neuron, 2022. **110**(10): p. 1728-1741 e7.

1016 8. Wong, L.C., et al., *Effective Modulation of Male Aggression through Lateral Septum to*
1017 *Medial Hypothalamus Projection.* Curr Biol, 2016. **26**(5): p. 593-604.

1018 9. Brayley, K.N. and D.J. Albert, *Suppression of VMH-lesion-induced reactivity and*
1019 *aggressiveness in the rat by stimulation of lateral septum, but not medial septum or*
1020 *cingulate cortex.* J Comp Physiol Psychol, 1977. **91**(2): p. 290-9.

1021 10. Couppis, M.H. and C.H. Kennedy, *The rewarding effect of aggression is reduced by*
1022 *nucleus accumbens dopamine receptor antagonism in mice.* Psychopharmacology (Berl),
1023 2008. **197**(3): p. 449-56.

1024 11. Flanigan, M., et al., *An emerging role for the lateral habenula in aggressive behavior.*
1025 Pharmacol Biochem Behav, 2017. **162**: p. 79-86.

1026 12. Golden, S.A., et al., *Basal forebrain projections to the lateral habenula modulate*
1027 *aggression reward.* Nature, 2016. **534**(7609): p. 688-92.

1028 13. Falkner, A.L., et al., *Decoding ventromedial hypothalamic neural activity during male*
1029 *mouse aggression.* J Neurosci, 2014. **34**(17): p. 5971-84.

1030 14. Falkner, A.L., et al., *Hypothalamic control of male aggression-seeking behavior.* Nat
1031 Neurosci, 2016. **19**(4): p. 596-604.

1032 15. Hashikawa, Y., et al., *Ventromedial Hypothalamus and the Generation of Aggression.*
1033 Front Syst Neurosci, 2017. **11**: p. 94.

1034 16. Lin, D., et al., *Functional identification of an aggression locus in the mouse*
1035 *hypothalamus.* Nature, 2011. **470**(7333): p. 221-6.

1036 17. Takahashi, A., et al., *Establishment of a repeated social defeat stress model in female*
1037 *mice.* Sci Rep, 2017. **7**(1): p. 12838.

1038 18. Shibata, S., T.Y. Yamamoto, and S. Ueki, *Differential effects of medial, central and*
1039 *basolateral amygdaloid lesions on four models of experimentally-induced aggression in*
1040 *rats.* Physiol Behav, 1982. **28**(2): p. 289-94.

1041 19. Dambacher, F., et al., *Reducing proactive aggression through non-invasive brain*
1042 *stimulation.* Soc Cogn Affect Neurosci, 2015. **10**(10): p. 1303-9.

1043 20. Choy, O., A. Raine, and R.H. Hamilton, *Stimulation of the Prefrontal Cortex Reduces*
1044 *Intentions to Commit Aggression: A Randomized, Double-Blind, Placebo-Controlled,*
1045 *Stratified, Parallel-Group Trial.* J Neurosci, 2018. **38**(29): p. 6505-6512.

1046 21. Takahashi, A., et al., *Control of intermale aggression by medial prefrontal cortex*
1047 *activation in the mouse.* PLoS One, 2014. **9**(4): p. e94657.

1048 22. Jones, R.B. and N.W. Nowell, *The effect of urine on the investigatory behaviour of male*
1049 *albino mice.* Physiol Behav, 1973. **11**(1): p. 35-8.

1050 23. Lischinsky, J.E. and D. Lin, *Neural mechanisms of aggression across species.* Nat
1051 Neurosci, 2020. **23**(11): p. 1317-1328.

1052 24. Nair, A., et al., *An approximate line attractor in the hypothalamus encodes an aggressive*
1053 *state.* Cell, 2023. **186**(1): p. 178-193 e15.

1054 25. de la Zerda, S.H., et al., *Social recognition in laboratory mice requires integration of*
1055 *behaviorally-induced somatosensory, auditory and olfactory cues.*
1056 Psychoneuroendocrinology, 2022. **143**: p. 105859.

1057 26. APA, *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR*, ed. APA. 2000,
1058 Washington, D.C.: American Psychiatric Association.

1059 27. Talbot, A., et al., *Supervised Autoencoders Learn Robust Joint Factor Models ofNeural*
1060 *Activity.* arxiv, 2020.

1061 28. Vu, M.T., et al., *A Shared Vision for Machine Learning in Neuroscience.* J Neurosci, 2018.

1062 29. Levy, D.R., et al., *Dynamics of social representation in the mouse prefrontal cortex.* Nat
1063 Neurosci, 2019. **22**(12): p. 2013-2022.

1064 30. Murugan, M., et al., *Combined Social and Spatial Coding in a Descending Projection from*
1065 *the Prefrontal Cortex.* Cell, 2017. **171**(7): p. 1663-1677 e16.

1066 31. Padilla-Coreano, N., et al., *Cortical ensembles orchestrate social competition through*
1067 *hypothalamic outputs.* Nature, 2022. **603**(7902): p. 667-671.

1068 32. Miczek, K.A., et al., *Aggressive behavioral phenotypes in mice.* Behav Brain Res, 2001.
1069 **125**(1-2): p. 167-81.

1070 33. Carlson, D., et al., *Dynamically Timed Stimulation of Corticolimbic Circuitry Activates a*
1071 *Stress-Compensatory Pathway.* Biol Psychiatry, 2017. **82**(12): p. 904-913.

1072 34. Mattis, J., et al., *Principles for applying optogenetic tools derived from direct*
1073 *comparative analysis of microbial opsins.* Nat Methods, 2011. **9**(2): p. 159-72.

1074 35. Novotny, M., et al., *Synthetic pheromones that promote inter-male aggression in mice.*
1075 Proc Natl Acad Sci U S A, 1985. **82**(7): p. 2059-61.

1076 36. Chen, A.X., et al., *Specific Hypothalamic Neurons Required for Sensing Conspecific Male*
1077 *Cues Relevant to Inter-male Aggression.* Neuron, 2020. **108**(4): p. 763-774 e6.

1078 37. Lee, D.L. and J.L. Wilson, *Urine from Sexually Mature Intact Male Mice Contributes to*
1079 *Increased Cardiovascular Responses during Free-Roaming and Restrained Conditions.*
1080 ISRN Vet Sci, 2012. **2012**.

1081 38. Hultman, R., et al., *Brain-wide Electrical Spatiotemporal Dynamics Encode Depression*
1082 *Vulnerability.* Cell, 2018. **173**(1): p. 166-180 e14.

1083 39. Roelofs, R., et al., *A Meta-Analysis of Overfitting in Machine Learning.* Advances in
1084 Neural Information Processing Systems 32 (NeurIPS 2019) 2019.

1085 40. Deonaraine, K.K., et al., *Sex-specific peripheral and central responses to stress-induced*
1086 *depression and treatment in a mouse model.* J Neurosci Res, 2020. **98**(12): p. 2541-2553.

1087 41. Li, L., et al., *Social trauma engages lateral septum circuitry to occlude social reward.*
1088 Nature, 2022.

1089 42. Baron, R.M. and D.A. Kenny, *The moderator-mediator variable distinction in social*
1090 *psychological research: conceptual, strategic, and statistical considerations.* J Pers Soc
1091 Psychol, 1986. **51**(6): p. 1173-82.

1092 43. Imai, K., L. Keele, and D. Tingley, *A general approach to causal mediation analysis.*
1093 Psychol Methods, 2010. **15**(4): p. 309-34.

1094 44. Anikeeva, P., et al., *Optetrode: a multichannel readout for optogenetic control in freely*
1095 *moving mice.* Nat Neurosci, 2011.

1096 45. Kumar, S., et al., *Cortical Control of Affective Networks.* J Neurosci, 2013. **33**(3): p. 1116
1097 –1129.

46. 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.

47. 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.

48. 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.

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

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

51. 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.

52. 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.

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

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

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

56. 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.

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

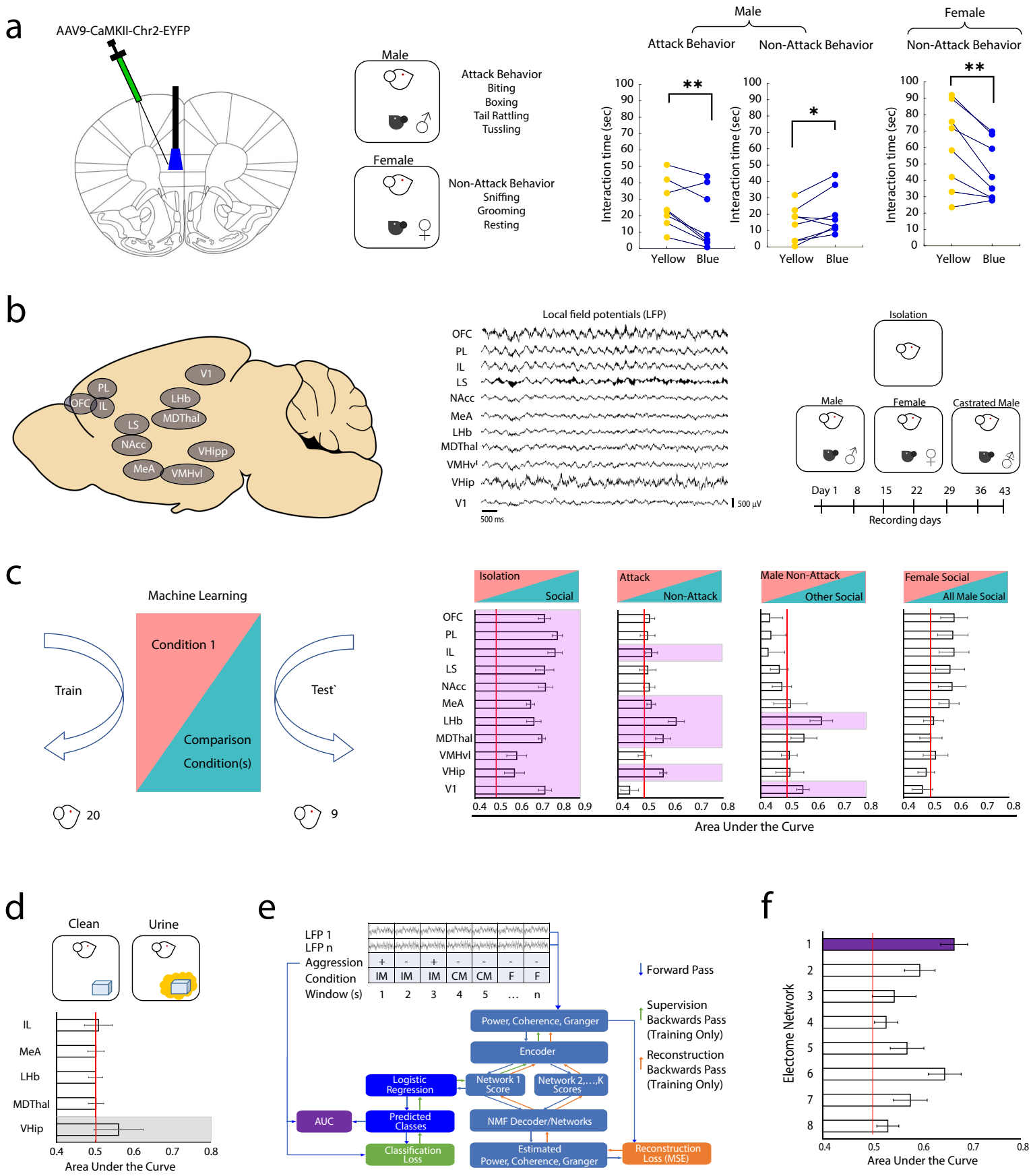
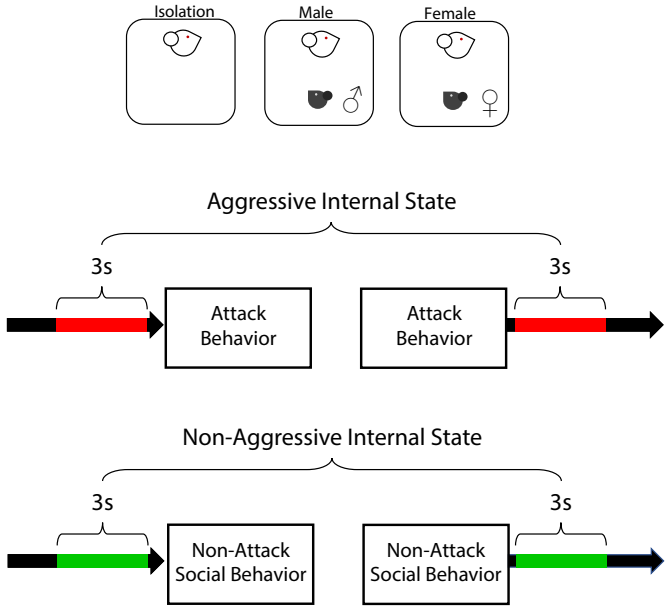

a
AAV9-CaMKII-Chr2-EYFP
Male
Attack Behavior
Biting
Boxing
Tail Rattling
Tussling
Female
Non-Attack Behavior
Sniffing
Grooming
Resting
Male
Attack Behavior
Non-Attack Behavior
Female
Non-Attack Behavior
Interaction time (sec)
Yellow
Blue
b
OFC
PL
IL
LS
NAcc
MeA
LHb
MDThal
VMHvl
VHipp
V1
Local field potentials (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

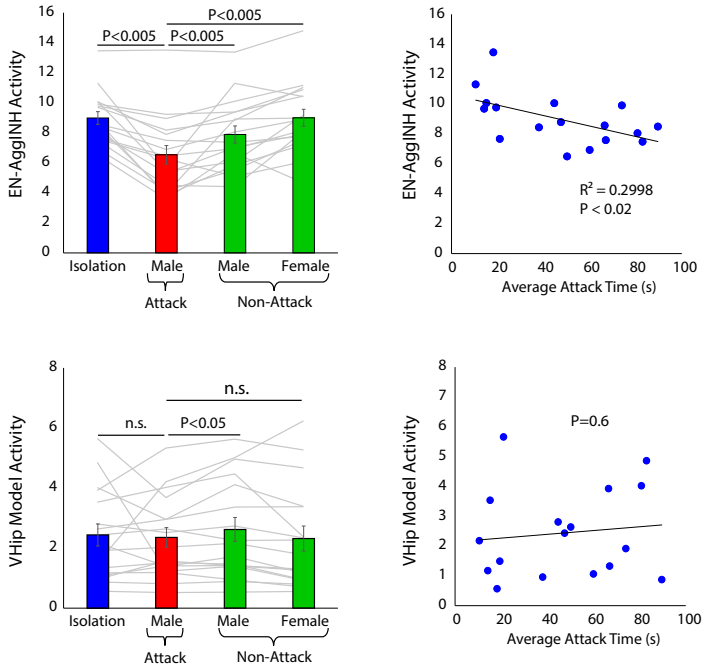
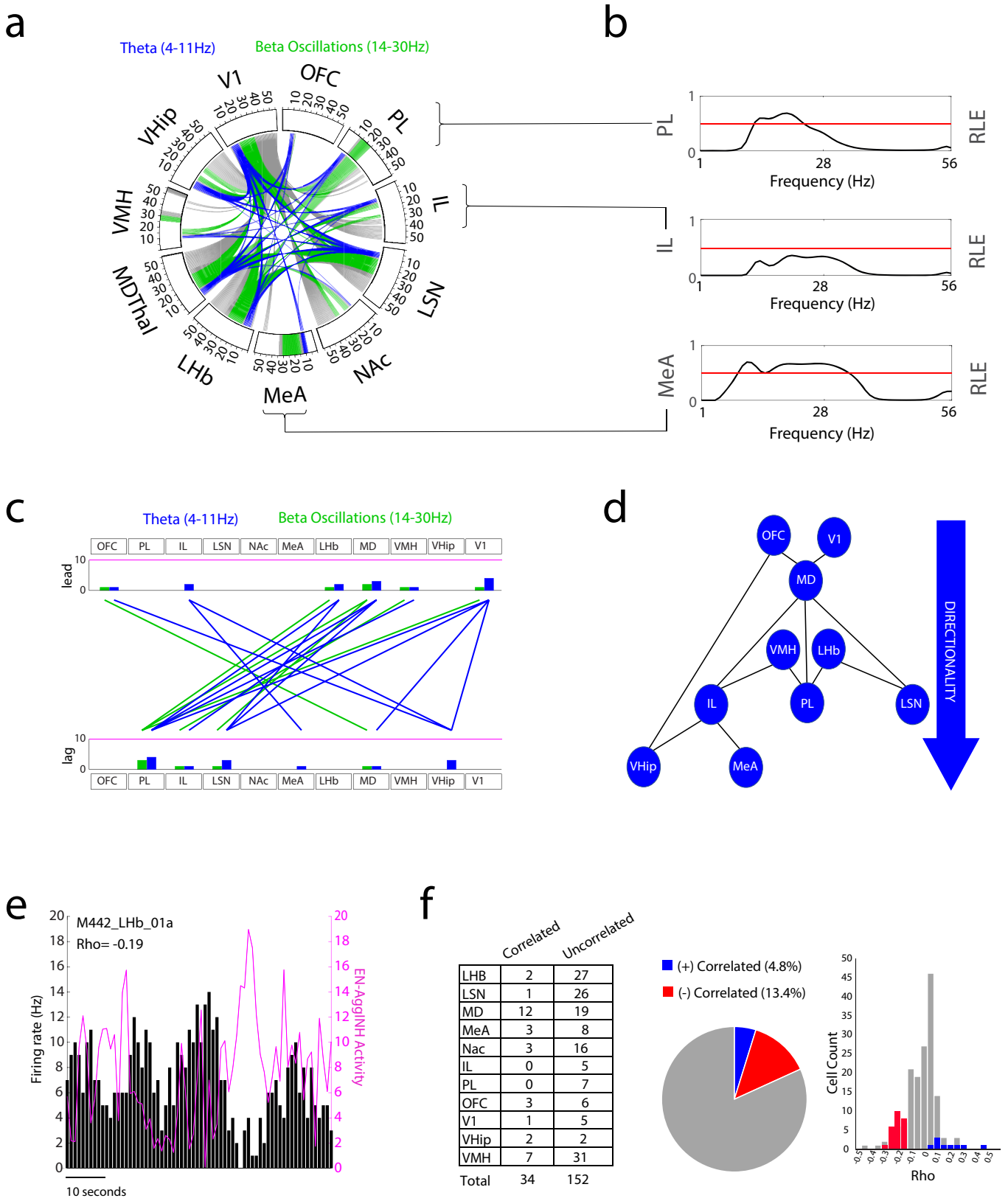

EN-AggINH Activity
P<0.005
P<0.005
P<0.005
Isolation
Male
Male
Female
Attack
Non-Attack
Average Attack Time (s)
R² = 0.2998
P < 0.02
VHip Model Activity
n.s.
n.s.
P<0.05
P=0.6

Figure 2



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

Figure 3

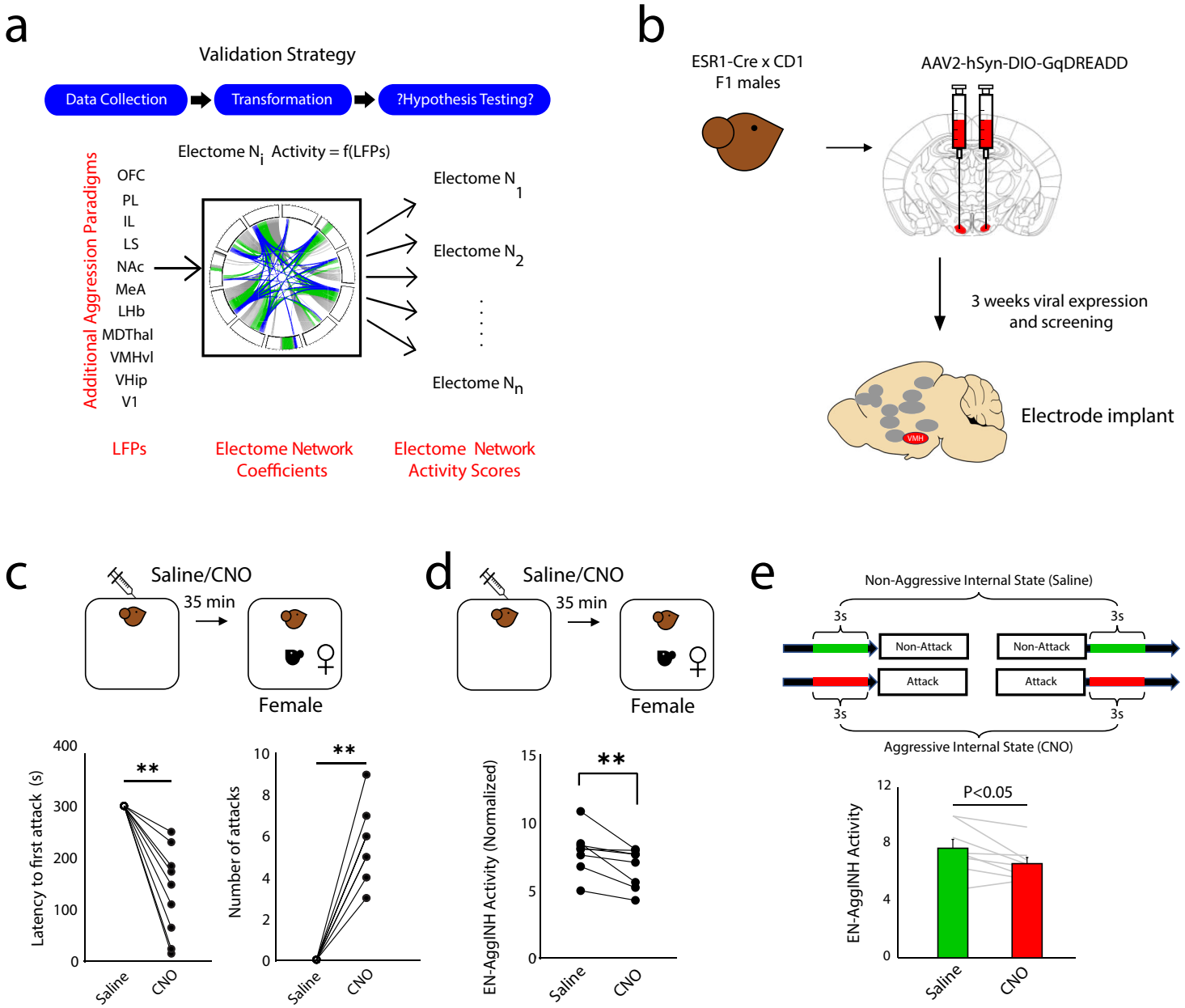

a
Validation Strategy
Data Collection
Transformation
?Hypothesis Testing?
Electome N_i Activity = f(LFPs)
Additional Aggression Paradigms
OFC
PL
IL
LS
NAc
MeA
LHb
MDThal
VMHvl
VHip
V1
Electome N_1
Electome N_2
Electome N_n
LFPs
Electome Network Coefficients
Electome Network Activity Scores
b
ESR1-Cre x CD1 F1 males
AAV2-hSyn-DIO-GqDREADD
3 weeks viral expression and screening
VMH
Electrode implant
c
Saline/CNO
35 min
Female
Latency to first attack (s)
Number of attacks
Saline
CNO
d
Saline/CNO
35 min
Female
EN-AggINH Activity (Normalized)
Saline
CNO
e
Non-Aggressive Internal State (Saline)
3s
Non-Attack
Attack
Aggressive Internal State (CNO)
P<0.05
EN-AggINH Activity
Saline
CNO

Figure 4

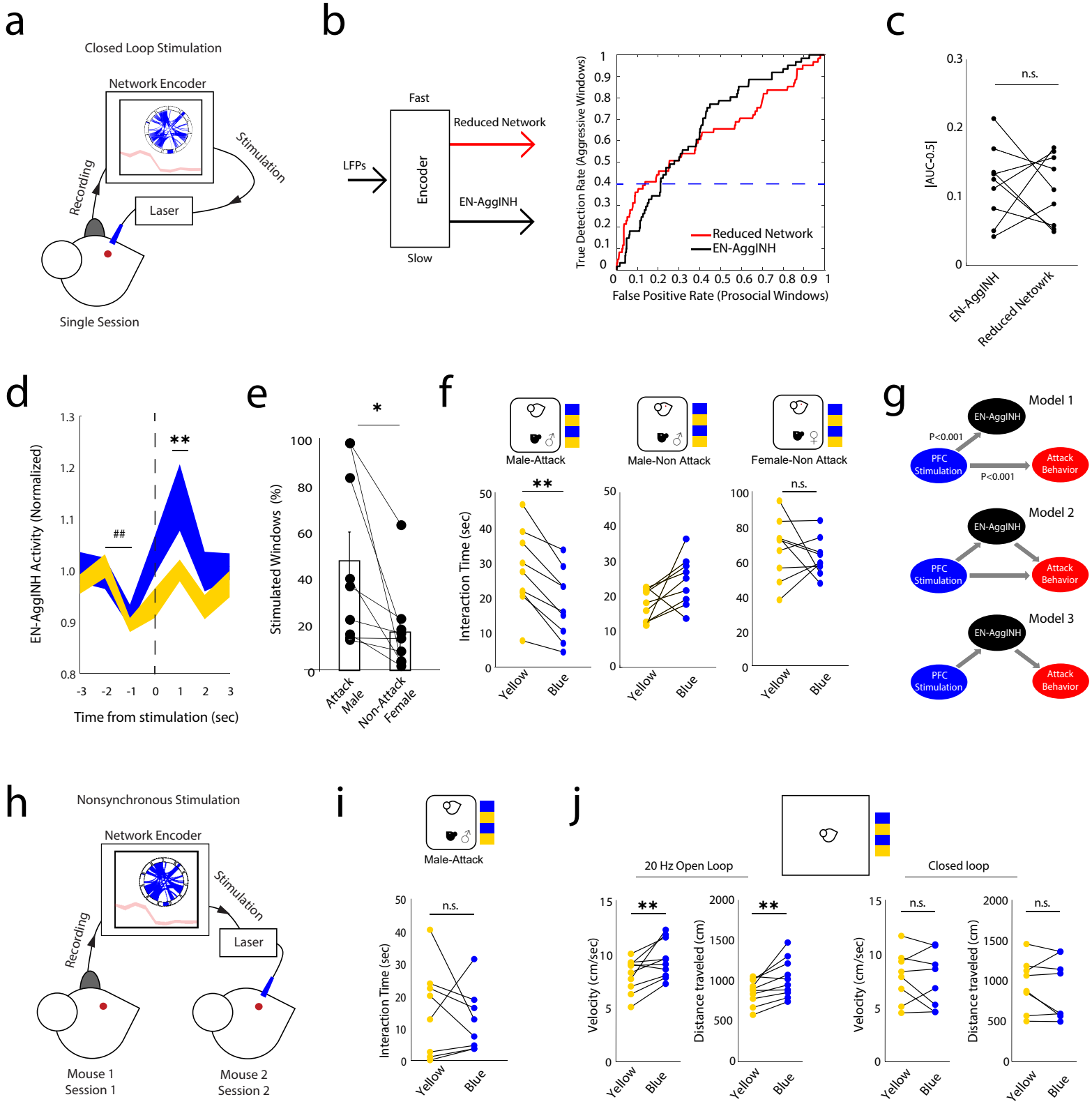

a
Closed Loop Stimulation
Network Encoder
Stimulation
Recording
Laser
Single Session
b
Fast
Reduced Network
LFPs
Encoder
EN-AggINH
Slow
True Detection Rate (Aggressive Windows)
False Positive Rate (Prosocial Windows)
Reduced Network
EN-AggINH
c
n.s.
|AUC-0.5|
EN-AggINH
Reduced Netowrk
d
EN-AggINH Activity (Normalized)
Time from stimulation (sec)
##
**
e
Stimulated Windows (%)
*
Attack Male
Non-Attack Female
f
Male-Attack
Male-Non Attack
Female-Non Attack
Interaction Time (sec)
**
n.s.
Yellow
Blue
g
Model 1
EN-AggINH
P<0.001
PFC Stimulation
Attack Behavior
P<0.001
Model 2
Model 3
h
Nonsynchronous Stimulation
Network Encoder
Stimulation
Recording
Laser
Mouse 1 Session 1
Mouse 2 Session 2
i
Male-Attack
n.s.
Interaction Time (sec)
Yellow
Blue
j
20 Hz Open Loop
Closed loop
Velocity (cm/sec)
Distance traveled (cm)
**
**
n.s.
n.s.
Yellow
Blue

Figure 5