Precision editing of an aggression-encoding network relay suppresses violent action

Yael S. Grossman[2,*], Austin Talbot[2,7*], Neil M. Gallagher[2,3], Kathryn K. Walder-Christensen[2], Gwenaëlle E. Thomas[3], Alexandra Fink Skular[6], 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; medial prefrontal cortex

Abstract

Aggression is a psychological state characterized by intent to harm oneself or others, and often manifests as hostile, harmful, or violent action. Social aggression can aid an organism in securing access to resources[1,2], or it can impair group function and survival in behavioral pathology[3-5]. Since many brain regions contribute to multiple social behaviors[6-8], expanded knowledge of how the brain distinguishes between social states would enable the development of interventions that suppress violent action, while leaving other social behaviors intact. Here we showed that a murine aggressive internal state is encoded by a widespread network. This network is organized by prominent and synchronized theta (4-11 Hz) and beta (14-30 Hz) oscillations that relay through the medial prefrontal cortex, and couples to widespread cellular firing. Strikingly, network activity during social isolation was found to encode the trait aggressiveness of mice, and causal cellular manipulations known to impact aggression were found to modulate the network's activity. We next established that this network mediates aggression using closed-loop stimulation of medial prefrontal cortex and causal mediation analysis. Finally, we deployed 'Long-term integration of Circuits using connexins' (LinCx) to selectively edit a key circuit within the network from the medial prefrontal cortex to nucleus accumbens. Editing this circuit chronically suppressed violent action while leaving non-aggressive social behavior intact.

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) [9,10], nucleus accumbens (NAc) [3,11], lateral habenula (LHb) [12,13], the ventrolateral portion of ventromedial hypothalamus (VMHvl) [6,14-18], and medial amygdala (MeA) [4,19]. Medial prefrontal cortex stimulation has been shown to mitigate aggressive behavior in both humans [20,21] and rodents [22], implicating cortical regions in regulating aggression. Finally, sensory regions, such as those underlying olfaction[23,24], also contribute to aggression.

63 To appropriately regulate aggressive behavior, the brain must integrate information across
64 these and other cortical and subcortical regions, many of which also regulate non-aggressive
65 social behaviors [6,25]. Therefore, the brain must integrate information from overlapping regions
66 to distinguish aggression from other social behavioral states. Though efforts have revealed
67 several cellular-level processes within these regions that contribute to this mechanism [14,17,26],
68 the complementary network-level process that integrates information across regions to
69 distinguish aggressive states from pro-social states remains unknown. Addressing this
70 knowledge gap is of major importance as 1) mammals regularly select from a repertoire of
71 social behaviors based on external sensory cues to ensure their survival [7,27], and 2) a range of
72 psychiatric disorders are broadly marked by a failure to appropriately match behavior with
73 evolving social contexts [28]. Because independent changes arising from many distinct regions
74 may potentially converge to alter aggression regulation, the optimal circuit to target for
75 widescale therapeutic intervention remains unclear.

76 Direct stimulation of medial prefrontal cortex broadly suppresses social behaviors

77 We initially focused on modulating the medial prefrontal cortex (mPFC; i.e., prelimbic and
78 infralimbic cortex in mice) as a means to selectively suppress aggression, since this brain region
79 had been implicated in social behaviors [29-31]. Moreover, clinical studies have shown that direct
80 transcranial stimulation of mPFC decreased aggressive feelings in violent offenders [32] and
81 individuals with methamphetamine use disorder [33]. Thus, we performed optogenetic
82 stimulation of mPFC during social encounters using a protocol that was modeled after prior
83 work (Fig. 1a)[22].

84 To test the behavioral effects of mPFC stimulation, we examined CD1 mice in two established
85 social paradigms. When a male C57BL6/J (C57) mouse is introduced into their home cage, CD1
86 mice show periods of violent/aggressive behavior [biting, boxing (kicking/clawing), or tussling]
87 [34]. In contrast, when exposed to a female intruder, CD1 mice exhibit periods of non-aggressive
88 social interactions [sniffing, grooming, or resting (placing nose or forepaws against the subject
89 mouse, but not moving)] [17]. Importantly, the CD1 mice do not show violent behavior during this
90 latter social context.

Consistent with the prior reports, mPFC stimulation suppressed aggressive behavior towards C57 male mice intruders and tended to increase non-aggressive social behavior [N=8; t_7=3.48; P=0.01, and t_7=-2.35; P=0.051 using a two-tailed paired t-test, for aggressive and non-aggressive behavior, respectively, significance determined by a Benjamini-Hochberg false discovery rate (FDR) correction, Fig. 1a]. However, this suppression of social behavior towards intruders was not specific to aggressive behavior, as we found that mPFC stimulation decreased non-aggressive social behavior toward female intruders as well (t_7=3.647, P=0.008, using two-tailed paired t-test, significance determined by FDR correction). Thus, mPFC stimulation suppressed multiple types of social behavior, not solely aggressive behavior.

Individual brain regions fail to independently encode aggressive behavior across contexts

After failing to selectively suppress CD1 aggressive social behavior by targeting the mPFC, we set out to find brain regions that may exhibit such selectivity. We implanted CD1 mice across multiple cortical and subcortical brain regions known to contribute to social behavior, including infralimbic cortex[22], orbitofrontal cortex[35], prelimbic cortex[22], lateral septum [9,10], nucleus accumbens [3,11], lateral habenula [12,13], mediodorsal thalamus[36], ventromedial hypothalamus [6,14-18], medial amygdala [4,19], ventral hippocampus [37], and primary visual cortex.

Following surgical recovery, we recorded neural activity while the CD1 mice freely interacted with an intact male C57 mouse and a female C57 mouse for 300 seconds each. We repeated these encounters over six sessions, yielding a total of 1800 seconds of neural data and behavior for each exposure (Fig. 1b, see also Supplementary Fig. S1). We recorded a subset of these CD1 mice (N=9 of the 20 implanted mice) as they interacted with a castrated male mouse intruder. Since CD1 mice do not generally exhibit aggressive behavior towards castrated males [24], this encounter provided neurophysiology data during additional non-aggressive social behaviors that were not linked to female sensory cues. We also acquired neural activity while the CD1 mice were isolated in their home cage.

We first verified that each of the implanted brain regions encoded social behavior using discriminative cross spectral factor analysis nonnegative matrix factorization (dCSFA-NMF, see Fig. 1e) [38]. dCSFA-NMF is a statistical model that integrates brain local field potential (LFP)

activity features across time and discovers networks that distinguish between types of external behavior. LFPs reflect the activity of populations of neurons, and these signals can be consistently sampled across mice. The electrical functional connectome networks (electome networks) learned in dCSFA-NMF integrate LFP power (oscillatory amplitude across 1-56 Hz; a correlate of cellular and synaptic activity within brain regions), coherence (how two regions' LFP synchronize across time; a correlate of brain circuit function), and directionality (assessed via Granger causality testing; a correlate of directional transfer of information across a brain circuit). Furthermore, dCSFA-NMF produces electome network activity scores, which indicate the strength of each network at a temporal resolution sufficient to capture brain states underlying relevant external behavior (in this instance, a resolution of one second). Any given brain region can belong to multiple electome networks, and each electome network may incorporate any number of brain regions. dCSFA-NMF thus integrates spatially distinct brain regions and circuits into interpretable networks that encode measured behavior[39].

To explore whether there was a generalized activity pattern within individual regions that encoded social behavior, we fit dCSFA-NMF models based solely on 1-56Hz LFP oscillatory power in a single region. Each single-region model was trained using data pooled from 20 CD1 mice to separate periods when mice were socially isolated from periods when they were engaged in social behavior (e.g., aggressive behavior towards the intact males and non-aggressive social behavior towards intact and castrated male mice and female mice). We trained our models with one supervised network to discover the patterns of LFP activity that encoded social behavior, meaning that this one network was encouraged to relate to the behaviors of interest and the LFP features. We also trained three unsupervised networks that were encouraged to only explain the variance in LFP features (see methods for justification of hyperparameter selection). We tested the accuracy of these models in new hold-out CD1 mice exposed to intruder male and female mice (N=9 mice).

We observed high accuracy in distinguishing periods of isolation from social behavior, using the supervised network, for each implanted brain area ($p<0.05$ using Wilcoxon rank-sum, significance determined by FDR correction for 44 comparisons, Fig. 1c, right, 1st graph). Next, we designed a new series of models to separate one of the three social behavioral states from

148 the other two (e.g., aggressive behavior toward intact males vs. non-aggressive interactions
149 with female and non-aggressive interactions with intact/castrated male mice). These models
150 were built using data pooled from the same 20 CD1 mice and were again based on LFP power.
151 We then tested the accuracy of the supervised networks for these single-region models in
152 decoding within the same nine hold-out mice. Thus, we trained and tested the model's
153 generalizability to decode each class of social behavior from the other two for each of the 11
154 implanted brain areas (i.e., 33 additional models). Using this approach, we found that five of
155 the brain region-based statistical models decoded aggressive behavior versus non-aggressive
156 social behavior: infralimbic cortex, medial amygdala, lateral habenula, medial dorsal thalamus
157 and ventral hippocampus. V1 and lateral habenula successfully decoded the non-aggressive
158 male interaction from the other social conditions as well ($P<0.05$ using one-tailed Wilcoxon
159 rank-sum test, significance determined by FDR correction for 44 comparisons, see Fig. 1c, right,
160 2nd graph). None of the other implanted brain regions showed selectivity in its encoding for one
161 type of social behavior. Thus, five regions independently encoded violent action vs. non-
162 aggressive social behavior in a manner that generalized across mice in this particular social
163 context.

164 However, our overall goal was to identify a neural signature that could be used to suppress
165 aggression across multiple contexts, while leaving other social behaviors intact. Thus, we tested
166 whether these five putative single-region neural signatures would generalize to another context
167 associated with aggression. Specifically, urine from other male mice has been found to elicit
168 aggressive and dominance behavior in CD1 males [23,24,40,41]. As such, the most aggressive mice
169 from the training group (N=8 of the 20 training mice, see methods) were allowed to freely
170 explore a clean inanimate object or an object covered in urine from another intact CD1 male
171 mouse (seven sessions). We then tested whether the significant single-region models
172 distinguished periods when mice explored clean objects from those when mice explored
173 objects covered in urine. Though only the ventral hippocampus model tended to decode
174 behavior in this new context, none of the brain regions showed statistically significant encoding
175 at the mesoscale-level (LFP) following multiplicity correction ($P>0.05$ using one-tailed Wilcoxon
176 rank-sum test, significance determined by FDR correction for 5 comparisons, Fig. 1d).

177 Aggressive behavior is encoded at the network-level across mice and contexts

178 After failing to robustly decode aggressive behavior across contexts solely using LFPs from one
179 of the 11 brain regions, we established that a brain network integrated information across all
180 the implanted brain regions to encode aggressive behavior across mice. This network-level
181 encoding mechanism generalized to multiple new contexts associated with aggression.

182 For this analysis, we trained a new dCSFA-NMF model using data from all the implanted brain
183 regions. This model utilized LFP power for each region and the coherence and directionality
184 between regions. The model utilized one supervised network that was trained to encode
185 periods of aggressive behavior (positive class) vs. non-aggressive social behavior towards
186 castrated males and the female social context (negative class), in addition to capturing variance
187 in the neural activity. We included non-aggressive social behavior towards intact males in the
188 negative class to discourage the network from simply learning sensory cues specific to the
189 intact male. Based on our hyperparameter selection approach using the Bayesian Information
190 Criteria (see methods and Supplementary Fig. S2), seven additional unsupervised networks
191 were trained solely to account for the variance in neural activity. We then validated our
192 supervised network using the same set of nine hold-out CD1 mice as our single-region analysis.
193 Again, none of these mice were used to train the electome networks. We found that the
194 supervised network successfully discriminated between aggressive behavior and non-aggressive
195 social behavior in the test mice (electome network #1; N=9 hold-out test mice, AUC=0.67±0.03,
196 one-tailed Wilcoxon rank sum against chance, $P=2\times10^{-5}$, Fig. 1f and Supplementary Fig. S3-5).
197 Activity in this supervised network was lower when animals exhibited aggressive behavior,
198 while high activity in one of the unsupervised networks mapped directly to aggressive behavior
199 [Network #6; AUC=0.65±0.03, $P<5\times10^{-5}$ using one-tailed Wilcoxon rank sum against chance
200 (AUC=0.5, signifying the model's inability to distinguish between aggressive behavior and pro-
201 social behavior); 9 hold-out test mice, see Supplementary Fig. S3].

202 In contrast to the single-region models developed for each of the brain regions, electome
203 network #1 activity also encoded the exposure to male urine (N=8, Network Activity = 9.1±1.0
204 and 8.3±1.1 for clean and urine covered objects, respectively; P=0.012 using one-tailed

205 Wilcoxon signed-rank test to evaluate whether exposure to urine decreased network activity;
206 AUC=0.57±0.02, data not shown). Electome network #6 activity failed to generalize to this urine
207 context (AUC=0.48±0.01, data not shown). Thus, we focused our subsequent analyses on
208 electome network #1.

209 An aggressive psychological state encoded at the network-level

210 Aggressive behavior is indicative of the aggressive psychological state. Yet, we also reasoned
211 that it was possible for a mouse to be in an aggressive psychological state without being
212 actively engaged in aggressive behavior. Since such a context was likely to be present
213 immediately prior to or following the expression of aggressive behavior, we tested whether
214 network activity pooled from the 3 seconds preceding and 3 seconds following social behaviors
215 encoded the distinct social conditions (Fig. 2a; N=17 mice from the training set). These data
216 windows were not used to train the network model since they did not contain the behaviors of
217 interest. Activity of the supervised network (electome network #1) was lower in the intervals
218 surrounding aggressive behavior compared to isolation and both non-aggressive social
219 interactions with males and females ($F_{3,67}$=28.6, P<0.0001, Friedman test; post-hoc analysis
220 performed using two-tailed Wilcoxon signed-rank test, P<0.005 after FDR correction, Fig. 2b,
221 left). Strikingly, electome network #1 activity during periods of isolation also negatively
222 correlated with the time mice spent exhibiting aggressive behavior towards other males (R=-
223 0.58; P=0.016 using Spearman's rank correlation, Fig. 2b, right), encoding an aggressive
224 psychological state on a mouse-by-mouse basis. For comparison, we also performed similar
225 analysis for VHip, since this region had shown the largest AUC for the single-region models in
226 the urine context (Fig. 1d). VHip failed to encode the aggressive psychological state during the
227 intervals surrounding aggressive behavior, and VHip activity in the home cage also failed to
228 encode individual behavior (see Supplementary Fig. S4). Thus, electome network #1 uniquely
229 encoded aggressive behavior across multiple contexts and signaled a generalized aggressive
230 internal state that extended beyond the moments of overt aggressive behavior. Since
231 decreased electome network #1 activity corresponded with aggression, these results suggested
232 the network [hereafter referred to as *EN-Aggression Inhibition (EN-AggINH*)] putatively inhibits
233 aggression when it is active.

234 *EN-Aggression Inhibition (EN-AggINH)* activity couples to cellular firing

235 *EN-AggINH* was composed of prominent theta frequency activity (4-11 Hz) in medial amygdala
236 and beta frequency activity (14-30 Hz) in medial amygdala and prelimbic cortex (Fig. 3a-b).
237 Prominent coherence was also observed in the theta and beta frequency bands. Indeed, when
238 we quantified directionality across these synchronized bands via Granger causality, we saw that
239 activity flowed from orbital frontal cortex (OFC) and primary visual cortex (V1), relayed through
240 medial dorsal thalamus (MDThal) and then infralimbic cortex (IL), and flowed into medial
241 amygdala (MeA) and ventral hippocampus (VHip; Fig. 3c-d, and Supplementary Fig. S5-6).
242 Notably, these activity flow patterns involving a primary sensory region (i.e., V1) may reflect a
243 change in encoding that occurs during aggression, or unique patterns of visual sensory input
244 that occur during aggressive behavior.

245 Next, we verified that the activity of *EN-AggINH* truly reflects biological activity by relating the
246 electome network to neural firing, as in previous work [8,42]. To achieve this, we analyzed the
247 Spearman correlation between single and multiunit cellular activity across the implanted brain
248 regions and the activity of *EN-AggINH*, as cell activity is an undisputed measure of biological
249 function. We then used a permutation test to rigorously test the significance of our findings
250 (Fig. 3e). Specifically, we shuffled cellular firing within social behavioral conditions, maintaining
251 the relationship between cell firing and behavior. We then repeated this procedure 1000 times
252 to generate a null distribution for which only 5% of cells would be expected to exhibit firing
253 coupled to network activity. We found that ~18% of cells showed firing that was coupled to the
254 activity of *EN-AggINH*, far more than could be explained by chance ($Z=4.05$, $p<0.0001$ using
255 two-tailed Z-test). Specifically, of the 186 cells recorded, nine (4.8%) showed firing activity that
256 was positively correlated with *EN-AggINH* and 25 (13.4%) showed activity that was negatively
257 correlated (Fig. 3f). Thus, most cells that showed coupling to *EN-AggINH* were inhibited when
258 network activity increased. Notably, we also observed coupling between cellular firing and
259 network activity when we performed high-density single-unit recordings from the anterior-
260 medial axis of the brain using silicon probes (43 out of 473 total neurons, $Z=2.41$, $p=0.008$ using
261 a two tailed Z-test, see Supplementary Fig. S7). Together, these analyses confirmed that *EN-*
262 *AggINH* activity reflects the dynamics of cellular activity in the recorded brain regions.

263 *EN-Aggression Inhibition (EN-AggINH)* generalizes to new biological contexts related to
264 aggression

265 To further validate *EN-AggINH*, we established that activity in this network was modulated by
266 orthogonal biological conditions that have been shown to induce or suppress aggressive
267 behavior in mice. In most cases, we performed this analysis in new animals, which is necessary
268 to appropriately evaluate generalization to new subjects[43]. One-tailed tests were used when
269 prior literature provided a clear directional relationship between a given manipulation and
270 aggressive behavior. Specifically, having established that *EN-AggINH* activity decreases during
271 aggressive behavior, we hypothesized that causal manipulations which increased aggressive
272 behavior would also decrease EN-AggINH activity. Similarly, we hypothesized that
273 manipulations which decreased aggressive behavior would increase *EN-AggINH* activity.

274 We transformed LFP data recorded from these new sessions into *EN-AggINH* activity through
275 our original network model. For our first validation experiment, we tested whether our network
276 generalized to new mice on a different genetic background (Fig. 4a-b). Our approach was based
277 on chemogenetically activating the ESR1+ cells of ventrolateral nucleus of the ventromedial
278 hypothalamus (VMHvl), which has been shown to induce aggressive behavior towards female
279 mice [18,44,45] – within a social context that would not normally induce aggression. We crossed
280 CD1 strain mice with ESR1-Cre mice on a C57 strain background and expressed an excitatory
281 DREADD (AAV-hSyn-DIO-hM3Dq) in the VMHvl of their offspring. We then treated these
282 offspring with CNO (Clozapine N-oxide, which activates the excitatory DREADD expressed by
283 the ESR1+ cells), and selected the subset of mice (9 out of the 13 total mice) that showed
284 aggression toward females. Subsequently, we implanted these mice with recording electrodes
285 to target the same brain regions used to train the network model as in our initial experiment
286 (Fig. 4a). Following recovery, we performed behavioral and neural recordings when mice were
287 treated with saline or CNO, in a pseudorandomized order, and then exposed to a female
288 mouse.

289 As anticipated, treatment with CNO induced aggressive behavior towards the female mice in
290 the implanted mice (N=8; P=0.0039 for both attack latency and attack number using one-tailed

291 Wilcoxon signed-rank; Fig. 4b). When we probed neural activity across the entire exposure to
292 the female intruder, we found that treatment with CNO also suppressed *EN-AggINH* activity
293 (N=8, P=0.0039 using one-tailed Wilcoxon signed-rank; Fig. 4c). Thus, the network model
294 generalized to a second aggression context causally induced by a cellular manipulation, and it
295 was robust to different genetic backgrounds. Notably, network activity was also lower during
296 the time intervals surrounding aggressive/non-aggressive social behavior for the CNO vs. saline
297 treatment sessions (P=0.02 using one-tailed Wilcoxon signed-rank; Fig. 4d), again
298 demonstrating that *EN-AggINH* encoded an aggressive psychological state even while the mice
299 were not actively engaging in aggressive behavior. Importantly, these observations also
300 established that *EN-AggINH* does not simply encode sensory cues associated with male
301 intruders, since the network responses observed in the CNO-treated mice were induced by a
302 female intruder.

303 *EN-Aggression Inhibition (EN-AggINH)* mediates aggressive behavior

304 We used mediation analysis to determine whether *EN-AggINH* activity putatively played a
305 mechanistic role in suppressing aggressive behavior. Mediation analysis is a framework to
306 determine whether the impact of a "treatment" (manipulation) on an outcome (aggressive
307 behavior) is mediated by a change in an intermediate variable (*EN-AggINH* activity). If so, the
308 intermediate variable is viewed, at least in part, as a mechanistic route (a mediator) for how the
309 treatment impacts the outcome. Three components were necessary to optimally implement
310 our mediation analysis models: a manipulation that causally modulated 1) aggressive behavior
311 and 2) *EN-AggINH* activity, and 3) an approach to deliver the manipulation exactly when *EN-*
312 *AggINH* activity would predict the emergence of aggressive behavior, i.e. at low *EN-AggINH*
313 activity levels. We chose to build such an approach based on optogenetic stimulation of mPFC,
314 since we had previously found that such a manipulation causally suppressed aggressive
315 behavior [22].

316 First, we built a closed-loop system that estimated the activity of *EN-AggINH* in real time (Fig.
317 5a). This approach employed a new network encoded solely based on power and coherence
318 measures (i.e., a reduced network, Fig. 5b), because the processing time to calculate Granger

319 causality was prohibitive for real-time implementations. While this new network lacked the full
320 interpretive power of *EN-AggINH*, it enabled us to predict aggressive behavior in real time to a
321 similar degree as *EN-AggINH* did (N=9; P=0.43 using two-tailed paired Wilcoxon signed-rank;
322 Fig. 5c). In principle, when the activity of *EN-AggINH* fell below an established threshold
323 (signaling the onset of aggressive behavior, see methods section "Optogenetic stimulation"),
324 our closed-loop approach would deliver a one-second light stimulation (5mW, 20Hz, 3ms pulse
325 width) to mPFC. To verify that this real-time estimation system worked as designed, we tested
326 whether light stimulation corresponded with a decrease in *EN-AggINH* activity (i.e., the full
327 network). Indeed, network activity was significantly lower one second prior to stimulation than
328 it was two seconds prior to stimulation, demonstrating that our approach using the reduced
329 encoder successfully identified when the *EN-AggINH* activity decreased below the threshold
330 that signaled the onset of aggression (N=9; P<0.005 for within-subject comparison of *EN-*
331 *AggINH* activity 1 vs. 2 seconds prior to yellow light stimulation using one-tailed Wilcoxon
332 signed-rank test, Fig. 5d). Importantly, we found that mPFC stimulation acutely increased *EN-*
333 *AggINH* activity (N=9, P<0.01 for comparison of *EN-AggINH* activity one second after blue vs.
334 yellow stimulation, using one-tailed Wilcoxon signed-rank test, see Fig. 5d). Thus, our closed-
335 loop stimulation approach satisfied two of the components needed to implement our
336 mediation approach.

337 For the final component, we tested whether increasing *EN-AggINH* activity via mPFC
338 stimulation as the brain transitioned into a putative aggressive psychological state would
339 suppress aggressive behavior. Our closed-loop approach preferentially induced optogenetic
340 stimulation during aggressive interactions, compared to non-aggressive social interaction with
341 females, establishing its selectivity (P = 0.002 using one-tailed Wilcoxon signed rank test; Fig.
342 6a, left). Importantly, closed-loop stimulation significantly suppressed aggressive behavior (N=9
343 mice that were not used to train the initial model; t_8=6.1, P=0.0003, comparing blue vs. yellow
344 light stimulation using two-tailed paired t-test for aggressive behavior, significance determined
345 by FDR correction; Fig. 6a, right;). Thus, our closed-loop manipulation suppressed aggressive
346 behavior, satisfying the remaining component needed to implement our mediation analysis
347 approach.

348 We first used the classic Baron and Kenny approach [46] to determine whether *EN-AggINH*
349 activity mediates the effect of neurostimulation on aggressive behavior. According to this
350 statistical approach, there is a mediated effect of network activity on behavior if three
351 conditions are met: 1) stimulation modulates network activity, 2) network activity correlates
352 with behavior, and 3) modeling the behavior from network activity and stimulation together is
353 better than modeling behavior from stimulation alone. Indeed, we had identified a significant
354 direct effect of stimulation on aggressive behavior (see Fig. 6a, right) and network activity (see
355 Fig. 5d). To optimally match the analytical conditions between the treatment and control cases
356 for subsequent analysis, we used windows during the closed-loop stimulation procedure where
357 the laser was triggered and then compared blue laser stimulation (treatment) to yellow laser
358 stimulation (control). Thus, the data points used for our mediation analysis predicted imminent
359 or ongoing aggressive behavior, and network activity prior to the stimulation in both the control
360 (yellow light) and treatment (blue light) case were similar. A statistical model of behavior using
361 network activity and stimulation (see Fig. 5e, model 2) significantly outperformed the model
362 using only stimulation (see Fig. 5e, model 1; nested logistic regression models, $P<0.01$,
363 likelihood ratio test), satisfying the necessary conditions to show that *EN-AggINH* is a mediator.

364 After establishing that *EN-AggINH* activity mediated the impact of mPFC stimulation on
365 behavior, we set out to evaluate the significance of the average causal mediation effect (ACME)
366 and the average direct effect (ADE) during the same stimulated closed-loop windows using
367 causal mediation analysis [47]. ACME is the causal effect of stimulation on behavior due to the
368 change in *EN-AggINH* activity (see Fig. 5e, model 3), and ADE is the causal effect on behavior
369 from mPFC stimulation not explained by the change in *EN-AggINH* activity. We found that there
370 was a significant ACME ($P<0.01$), but not a significant ADE ($P=0.48$). This analysis suggested that
371 *EN-AggINH* activation is the primary mechanism whereby mPFC stimulation suppresses
372 aggression.

373 Next, we tested different mediation pathways to determine whether the suppressive effects of
374 stimulation on aggression are specifically mediated by *EN-AggINH* rather than alternative
375 neural processes. First, we evaluated whether theta power in any of the 11 different brain
376 regions could serve as a mediator in lieu of *EN-AggINH* activity (Fig. 5f, model 4). We chose this

377 frequency band since it was prominently featured in *EN-AggINH* and also in the electome
378 network we previously found to encode social appetitive behavior[8]. Across these 11 single-
379 region models, only OFC had a significant average causal mediation effect (*uncorrected* p-value
380 of 0.038, see Fig. 5f), but this model did not survive a correction for multiple comparisons.
381 Additionally, its ACME estimate was dwarfed by the size of the ACME estimate for *EN-AggINH*
382 (the mean estimate for the *EN-AggINH* model was 49.7% larger). This evidence suggests that
383 *EN-AggINH* is a much better mediator than any of these potential single-region 'biomarkers' by
384 themselves.

385 After failing to identify any significant mediation effect of theta activity within each of the
386 eleven brain regions, we tested whether including theta activity as an intermediary in our
387 causal graph would disrupt *EN-AggINH's* role as a mediator in aggressive behavior (Fig. 5g,
388 model 5). Here, we corrected for the role of theta power in the model of how *EN-AggINH*
389 changes as a function of stimulation, as well as correcting for theta power in forecasting
390 aggressive behavior. As such, this framework dictates that *EN-AggINH* cannot mediate behavior
391 that is already explained by changes in theta power in a specified region. When we ran eleven
392 models, one model for each brain region, we found that *EN-AggINH* still significantly mediated
393 aggressive behavior in all of them ($P<0.05$ for all models; see Fig. 5g, bottom). Thus, even after
394 accounting for these potential intermediate variables, our findings still supported *EN-AggINH* as
395 a mediator of aggressive behavior. Consistent with this finding, our mediation analysis
396 performed using data from the ESR1-Cre experiment showed that *EN-AggINH* also mediated
397 the impact of CNO treatment (see Supplementary Fig. S8). Thus, *EN-AggINH* reflects the
398 internal brain state that suppresses aggressive behavior.

399 <u>Validation of temporal activity and spatial spectral features of *EN-AggINH*</u>

400 We validated the temporal activity and spatial spectral features of *EN-AggINH* by establishing
401 that they could be utilized to selectively suppress aggression. Specifically, after determining
402 that our closed-loop manipulation suppressed aggression, we also quantified the impact of this
403 stimulation protocol on non-aggressive interactions with other male and female mice. We
404 found that closed-loop mPFC stimulation had no impact on non-aggressive behavior towards

405 the intact C57 males (N=9; t_8=-2.3, uncorrected P=0.049 comparing blue vs. yellow light
406 stimulation using two-tailed paired t-test for non-aggressive behavior, significance determined
407 by FDR correction, Fig. 6a, right, 2nd graph). No stimulation-induced differences in non-
408 aggressive behavior were observed during exposure to female mice (t_8=0.74, P=0.48 using two-
409 tailed paired t-test, significance determined by FDR correction, Fig. 6a, right, 3rd graph).
410 Interestingly, while the open-loop protocol induced hyperactivity during open field exploration
411 (Supplementary Fig. S9a, left), no differences in gross locomotor behavior were observed
412 during closed-loop mPFC stimulation (Supplementary Fig. S9a, right). Thus, closed-loop mPFC
413 stimulation selectively reduced aggression without affecting other social behaviors or gross
414 locomotor behavior.

415 Next, we verified that this selective modulation of aggression was due to synchronization of
416 light stimulation with endogenous *EN-AggINH* activity, and not simply due to the dynamic
417 pattern of stimulation delivered using this method. We performed an additional control
418 experiment where mPFC stimulation patterns in new animals were yoked to network activity
419 recorded from another mouse's brain, analogous to a "sham" in neurofeedback experiments
420 (i.e., nonsynchronous stimulation; Fig. 6b). Nonsynchronous stimulation failed to suppress
421 aggressive behavior ($F_{1,21}$=4.87, P =0.039 for light type × stimulation pattern effect for
422 exploratory analysis using a mixed effects, two-way ANOVA; t_{13}=0.09, P=0.93 for
423 nonsynchronous stimulation effects on aggressive behavior towards intruder male mice using
424 post-hoc two-tailed paired t-test, significance determined by FDR correction; see Fig. 6b, right,
425 1st graph), verifying that the suppression of aggressive behavior driven by closed-loop
426 stimulation was indeed due to delivery of stimulation timed to endogenous *EN-AggINH* activity.
427 Incidentally, nonsynchronous stimulation had no impact on non-aggressive social behavior
428 towards intact males or females (N=14; t_{13}=1.79, P=0.097; and t_{13}=0.54, P=0.60, for interaction
429 with males and females, respectively, comparing blue vs. yellow light stimulation using two-
430 tailed post-hoc paired t-tests, significance determined by FDR correction; see Fig. 6b, right, 2nd
431 and 3rd graphs). Thus, we validated the temporal activity component of *EN-AggINH.*

432 After establishing that we could selectively reduce aggression by temporally targeting mPFC
433 based on the activity level of *EN-AggINH*, we tested whether we could reduce aggression by

434 spatially targeting stimulation based on the sub-components of mPFC output circuity that
435 composed the network. Our goal was to identify sub-components of the *EN-AggINH* whose
436 information flow changes most as the system transitions between aggression states to find
437 promising targets for control. Our initial visualization of *EN-AggINH* was constrained to the
438 absolute information flow at the strongest synchronies (Fig. 3c-d). On the other hand, relative
439 LFP spectral energy (RLE) is the weight of each directionality feature of *EN-AggINH* normalized
440 to the sum of the weight of that feature across all the learned electome networks. As such, the
441 directionality RLEs quantify the magnitude to which a specific circuit increases its output as *EN-*
442 *AggINH* activity increases (Supplementary Fig. S6). We therefore identified target circuits by
443 looking at the RLE of the *EN-AggINH* directionality features that emerged from prelimbic (PL)
444 and infralimbic cortex (IL) cortex. Because *EN-AggINH* activity decreases during aggression, a
445 high directionality RLE also corresponds with the circuits within *EN-AggINH* that show the
446 strongest reduction in information flow when aggression emerges.

447 We focused our analysis on the directionality between mPFC [prelimbic (PL) and infralimbic
448 cortex (IL)] to nucleus accumbens (mPFC→NAc), medial amygdala (mPFC→MeA) and lateral
449 habenula (mPFC→LHb), since these circuits consisted of monosynaptic projections, enabling
450 direct targeting using optogenetics. We then quantified the 20Hz RLE of these mPFC output
451 circuits (Fig. 6c), since increasing mPFC activity at this frequency via optogenetic stimulation
452 was sufficient to suppress aggressive behavior (Fig. 1a) and increase *EN-AggINH* activity (Fig.
453 5d). *EN-AggINH's* RLE was highest for mPFC→NAc and then mPFC→MeA circuitry and lowest in
454 the mPFC→LHb circuit. Thus, a prominent decrease in information flow in mPFC→NAc and
455 mPFC→MeA circuitry was associated with aggression, while no such change was observed in
456 mPFC-LHb activity [$P<0.001$ for pairwise comparison of relative LFP spectral energy across these
457 six circuits (analysis performed separately for PL- and IL-dependent circuitry] using
458 bootstrapping, see methods)]. Given this observation, we reasoned that driving mPFC→NAc or
459 mPFC→MeA activity at 20Hz should selectively suppress aggression, while driving mPFC→LHb
460 activity should not.

461 We causally activated these three circuits at 20Hz and measured their impact on social
462 behaviors. To selectively stimulate the terminals of mPFC neurons in each target region (NAc,

463 MeA, or LHb), we injected mice with a retrograde AAV-Cre (rAAV-Cre) virus in one target region
464 and an AAV-DIO-channelrhodopsin-2 virus in mPFC. A stimulating fiber was placed above the
465 target region injected with rAAV-Cre. Social behavior was quantified during 20Hz stimulation
466 with yellow vs. blue light (5mW, 20Hz, 3ms pulse width). Blue light stimulation of mPFC→NAc
467 decreased aggression toward male C57 mice (t_8=5.5, P=0.0006 for blue vs. yellow light using
468 two-tailed paired t-test; N=9 mice, significance determined by FDR correction), but had no
469 impact on non-aggressive social behavior towards the male C57 mice (t_8=2.01, P=0.08 using a
470 two-tailed paired t-test) or social behavior towards female C57 mice (t_8=0.39, P=0.71 using a
471 two-tailed paired t-test; see Fig. 6d, right). Though mPFC→MeA tended to decrease aggression
472 toward male C57 mice, these differences did not reach statistical significance (t_8=2.24, P=0.056
473 for blue vs. yellow light using a two-tailed paired t-test; N=9 mice, significance determined by
474 FDR correction; see Fig. 6e, right, 1st graph). Again, no difference in non-aggressive social
475 behavior was observed towards male C57 mice (t_8=0.58, P=0.58 using a two-tailed paired t-test)
476 or females (t_8=0.80, P=0.45 using a two-tailed paired t-test; see Fig. 6e, right). Finally, blue light
477 stimulation of mPFC→LHb stimulation had no impact on aggression toward male C57 mice
478 (t_8=1.05, P=0.32; using two-tailed paired t-test, N=9 mice), non-aggressive social behavior
479 towards C57 males (t_8=1.42, P=0.19 using two-tailed paired t-test), or social interaction with
480 C57 females (t_8=0.45, P=0.67 using two-tailed paired t-test; see Fig. 6f, right). These results
481 demonstrated that directly stimulating the mPFC subcircuit which showed the greatest
482 decreases in aggression-related activity (i.e., mPFC → NAc) causally suppressed aggression. On
483 the other hand, stimulating a mPFC subcircuit with minimal activity changes during aggression
484 (i.e., mPFC → LHb) had no impact on social behavior towards male mice. Together, these
485 findings validated the spatial spectral features of *EN-AggINH.*

486 Having established that acute stimulation of the mPFC→NAc circuit selectively diminished
487 aggression, we tested whether editing this circuit using LinCx chronically suppressed aggressive
488 behavior. LinCx is composed of two gap junction hemichannels, $Cx34.7_{M1}$ and $Cx35_{M1}$, which
489 exclusively dock with each other[48]. Each of the hemichannels can be targeted to distinct cell
490 membranes that are in close proximity to yield an electrical synapse between the pair of cells
491 (Fig. 6g, left), increasing the coupling of neural activity between them[48]. After screening, we

492 assigned aggressive male CD1 mice to experimental and control groups in a manner that
493 balanced social behavior (aggressive and non-aggressive) between groups. After another
494 presurgical behavioral screening, experimental mice were injected with AAV-CaMKII-Cx34.7$_{M1}$-
495 2xHA-T2A-mEmerald in mPFC and AAV-CaMKII-Cx35$_{M1}$-FLAG-T2A-mCherry in NAc. Control mice
496 were injected with AAV-CaMKII-GFP in mPFC and AAV-CaMKII-mCherry in NAc (Fig. 6g, right).
497 We then quantified aggressive and non-aggressive social behavior serially at 2-, 5-, and 8-weeks
498 post-injection (Fig. 6h). Notably, strong trafficking of Cx34.7$_{M1}$ from mPFC to NAc (and thus
499 LinCx editing) is not observed until more than 3 weeks after the initial surgery[49]. Importantly,
500 the second pre-surgical screening of aggressive and non-aggressive behavior towards intact
501 males and females found no difference between the experimental and control groups (t_{19}=0.35,
502 P=0.73 for group comparison of aggressive behavior towards intact males, t_{19}=0.23, P=0.82 for
503 group comparison of non-aggressive social behavior towards intact males, and t_{19}=0.57, P=0.57
504 for group comparison of non-aggressive social behavior towards females, all using a two-tailed
505 t-test; N=10 LinCx-edited mice and N=11 control mice), demonstrating the robustness of our
506 group assignments (see Fig. 6h). Our analysis using a three-way ANOVA found a significant
507 group x social condition effect on social interaction time ($F_{4,168}$=2.097 and P=0.08 for group ×
508 social condition × week effect using mixed-effects model three-way ANOVA; $F_{2,168}$=7.38 and
509 P=0.0008 for group × social condition effect using mixed-effects model three-way ANOVA).
510 Indeed, editing the mPFC→NAc circuit using LinCx decreased aggressive behavior towards
511 intact males ($F_{2,38}$=2.85 and P=0.07 for group × week effect using mixed-effects model ANOVA;
512 $F_{1,19}$=6.82 and P=0.017 for group effect, $F_{2,38}$=8.6 and P<0.001 for week effect; significance
513 determined by a Benjamini-Hochberg false discovery rate (FDR) correction; T_{19}=2.92 and
514 P=0.009; T_{19}=3.13 and P=0.006 for post-hoc comparisons across groups at 5 and 8 weeks after
515 injection, respectively, using an unpaired two-tailed t-test; N=10 LinCx-edited mice and N=11
516 control mice; Fig. 6h, left). These mice showed increased non-aggressive social behavior
517 towards intact male intruders at 5, but not 8 weeks after injection ($F_{2,38}$=5.26 and P=0.01 for
518 group × week effect using mixed effects model ANOVA; significance determined by a Benjamini-
519 Hochberg false discovery rate (FDR) correction; T_{19}=2.46 and P=0.024; T_{19}=1.79 and P=0.090 for
520 post-hoc comparisons across groups at 5 and 8 weeks after injection, respectively, using an

521 unpaired two-tailed t-test; Fig. 6h, middle). Finally, we observed no difference in social behavior
522 towards females ($F_{1,19}$=0.02 and P=0.88 for group effect; $F_{2,38}$=2.05 and P=0.14 for group × week
523 effect using mixed effects model ANOVA; significance determined by a Benjamini-Hochberg
524 false discovery rate (FDR) correction; Fig. 6h, right). Consistent with our closed-loop mPFC
525 stimulation findings, LinCx editing did not induce hyperactivity during open field exploration
526 (Supplementary Fig. S9b). Taken together, these findings showed that editing the mPFC→NAc
527 chronically suppressed aggressive behavior, while leaving non-aggressive social behavior intact.

528

529 **Discussion**

530 We set out to identify a neural circuit that could be targeted to selectively decrease aggressive
531 behavior while leaving pro-social behavior intact. To achieve this outcome, we investigated the
532 neural mechanisms that underlie when a mouse engages in aggressive vs. pro-social behavior.
533 We reasoned that aggressive behaviors emerge when the brain is in an aggressive psychological
534 state, and that pro-social behaviors emerge when it is not. Thus, we used machine learning to
535 discover the mesoscale neural architecture of the brain when an animal expressed aggressive
536 behavior versus when it expressed non-aggressive social behaviors. The network we identified
537 incorporated distributed neural dynamics spanning all recorded regions. Distributed neural
538 dynamics are also observed in well-defined internal brain states such as sleep, which are
539 characterized by coordinated activity changes across widespread brain regions[50,51]. For multiple
540 regions, differences were observed in local oscillatory power, while others exhibited differences
541 in coherence between a broader collection of regions. No brain region showed prominent
542 activity contributions across all frequencies. Rather, each brain region showed selectivity in the
543 frequencies of oscillations that contributed to the network. For example, prelimbic cortex
544 showed strong activity in the beta frequency range, while medial amygdala showed strong
545 activity in the beta and theta frequency ranges. Most importantly, the brain state that was
546 identified across multiple contexts associated with aggression was better captured by an
547 integrated network composed of activity recorded across 11 brain regions, rather than the
548 independent activity within each brain region.

549 Interestingly, this aggressive brain state was encoded by decreased activity in the network.
550 Given that we identified more cells that increased their firing rates as network activity
551 decreased, the discovery of a network that decreases its activity during aggression does not
552 indicate that overall brain activity is suppressed during aggressive states. Rather, these findings
553 argue that the aggressive state is encoded by a network that decreases its activity relative to
554 when mice are socially isolated or engaged in pro-social behavior (Fig. 4B, left). Indeed, our
555 data suggested that several common regions/circuits were activated during aggressive and pro-
556 social behavior. These common regions/circuits need not be reflected in our network since our
557 model was trained to differentiate aggressive from non-aggressive social behavior.
558 Nevertheless, our discovery of a network that decreased its activity during aggression raises the
559 intriguing hypothesis that the brain actively inhibits aggression during pro-social engagement.
560 When activity in this inhibition network is suppressed, aggressive behavior emerges.

561 This interpretation is supported by our validation experiments where we chemogenetically
562 activated ESR1+ neurons in ventromedial hypothalamus, which induced the aggressive brain
563 state (i.e., suppressed *EN-AggINH* activity). When CNO-treated mice were then exposed to a
564 stimulus that would generally induce non-aggressive social behavior (i.e., a female mouse),
565 aggressive behavior emerged. Thus, the presence of the aggressive brain state changed the
566 mapping between sensory input and behavior output. Similarly, closed-loop optogenetic
567 stimulation of mPFC biased mice towards exhibiting non-aggressive social behavior when they
568 were exposed to a stimulus that would generally induce aggressive behavior (i.e., a male
569 intruder), as well as decreased the aggressive brain state (i.e., increased *EN-AggINH* network
570 activity). Finally, our mediation analyses suggest that the brain state represented by *EN-AggINH*
571 mediates the suppressive effects of mPFC stimulation on aggressive behavior.

572 Though behavioral output has been classically utilized to infer the internal state of a brain, we
573 reasoned that an internal aggressive psychological state was also likely present during intervals
574 immediately preceding and following aggressive behavior. Thus, we tested whether our
575 network showed distinct activity profiles in the time intervals surrounding aggressive behaviors.
576 Indeed, network activity was lower during intervals surrounding aggressive behavior compared
577 to intervals surrounding non-aggressive behavior. Strikingly, we also found that network activity

578 when animals were isolated in their home cage encoded their trait aggression. Thus, the
579 network did not simply encode aggressive action since its activity was observed separately from
580 behavior. Rather, the network encoded an internal aggressive psychological state.

581 Here, we framed internal brain states as a collection of functions that transform sensory input
582 into behavior. Indeed, we found that when *EN-AggINH* activity is suppressed, the brain
583 transforms both male and female social sensory cues into aggressive behavior. It is also widely
584 appreciated that internal and external sensory input can also cause the brain to transition from
585 one internal state to another. Along this line, we found that exposure to male mice could
586 promote an aggressive psychological internal state in CD1 mice even prior to aggressive
587 behavior, while exposure to a female mouse did not (under normal conditions). In this
588 framework, one would also anticipate that modulation approaches that decrease aggressive
589 behavior could mediate their effect by driving the brain out of the aggressive state, which is
590 represented by low *EN-AggINH* activity.

591 The ability to modulate aggressive behavior in this way could present a breakthrough for the
592 treatment of multiple neuropsychiatric disorders – including mood disorders, psychotic
593 disorders, neurodevelopmental disorders, and neurodegenerative disorders – with deficits in
594 regulating social behavior such as aggression. While pharmacological approaches have been
595 implemented to suppress aggressive behavior towards self and others, many of these strategies
596 act by sedating the individual and can disrupt aspects of pro-social function. Our discovery of a
597 brain network that encoded an aggressive state raises the potential for novel approaches to
598 suppress aggressive behavior that spare pro-social behavior. Indeed, compared to a standard
599 open-loop stimulation protocol which suppressed both aggressive behavior toward intact males
600 and non-aggressive pro-social behavior toward females (Fig. 1a), our closed-loop stimulation
601 approach spared non-aggressive social behavior towards females. Intriguingly, like other open-
602 loop mPFC stimulation studies [52,53], our 20Hz stimulation protocol induced behavioral
603 hyperactivity in experimental mice (Supplementary Fig. S9a, left). On the other hand, our
604 closed-loop stimulation protocol did not (Supplementary Fig. S9a, right). Thus, our findings also
605 show that closed-loop stimulation may limit off-target behavioral effects that are induced by
606 classic stimulation approaches.

607 Our closed-loop stimulation approach was developed using a neural network-based
608 approximation technique for which the features were substantially constrained relative to
609 dCSFA-NMF. Nevertheless, we found that the reduced encoder was sufficient to identify the
610 precise time windows when the brain transitioned into an aggressive psychological state, as
611 marked by a decrease in *EN-AggINH* activity. In the future, novel approaches may allow for
612 further improvement in the precision of our real-time stimulation approach, including fast
613 approximation by surrogate neural networks. Such an approach may enable the
614 implementation of multiple real-time networks that predict both aggressive and pro-social
615 states concurrently to actuate our closed-loop system and further suppress aggressive behavior
616 relative to pro-social behavior. Interestingly, our current findings also pointed to a network that
617 exhibits increased activity during aggressive behavior (electome network #6, see Fig. 1f and
618 Supplementary Fig. S10). Though the network failed to encode the urine paradigm and thus did
619 not generalize across aggressive contexts, it is possible that it contains activity that synergizes
620 with *EN-AggINH* to encode aggressive social states more optimally. If future studies
621 demonstrate this network's potential, a reduced encoder for electome network #6 could also
622 be integrated to further optimize our closed-loop approach to selectively suppress aggression.

623 Extending our mPFC stimulation studies, we used the spatiotemporal architecture of *EN-*
624 *AggINH* to identify an mPFC output circuit to NAc that was prominently suppressed during
625 aggression. Activating this mPFC→NAc circuit at 20Hz (open-loop optogenetic stimulation)
626 reduced aggression while sparing pro-social behavior towards intact males and females.
627 Stimulating another mPFC output circuit to LHb that was relatively less suppressed in *EN-*
628 *AggINH* failed to reduce aggression. Having identified this circuit specificity, we utilized LinCx to
629 edit the mPFC→NAc circuit, which chronically suppressed aggression and temporarily increased
630 pro-social behavior towards intact males. Additionally, this manipulation spared pro-social
631 behavior towards female mice. Thus, by mapping the distributed brain network that encodes
632 aggressive behavior and a generalized aggressive psychological state, performing multiple
633 causal manipulation experiments to validate the spatiotemporal features of this network, and
634 deploying targeted circuit editing using LinCx, we established a neural circuit that could be
635 targeted to diminish aggressive behavior while leaving pro-social behavior intact.

636 Overall, our findings establish a generalized network-level signature whereby the brain
637 suppresses aggression via active inhibition. Moreover, they highlight the exciting therapeutic
638 potential for state-specific (i.e., closed-loop) neuromodulation to regulate internal states and
639 for neural circuit editing to chronically and selectively suppress pathological social behavior.

640

641 ## Acknowledgements

642 We would like to thank Stephen Mague, Karim Abdelaal, and Ashleigh Rawls comments on this
643 work; Jean Mary Zarate for substantial guidance on revising this work; and Timothy Nyangacha
644 for technical support. This work was supported by WM Keck Foundation and Hope for
645 Depression Research Foundation grants to KD; NIH grants R01MH120158 and 1DP1MH132709 to
646 KD, 1R01EB026937 to DEC, and 1R01MH125430 to SJR, DEC, and KD. A special thanks to
647 Freeman Hrabowski, Robert and Jane Meyerhoff, and the Meyerhoff Scholarship Program.

648 ## Figure Legends

649 **Figure 1. A widespread network encodes aggressive behavior. a)** Direct stimulation of medial
650 prefrontal cortex suppresses social behavior. Schematic of optogenetic stimulation (left) and
651 social encounters utilized for testing (middle). Right: Medial prefrontal cortex stimulation
652 suppressed aggressive behavior, increased non-aggressive social behavior towards male mice,
653 and suppressed non-aggressive social behavior towards females (stimulation using blue -
654 473nm vs. yellow - 593.5nm light, 5mW, 20Hz, 3ms pulse width, *P<0.05 for each comparison
655 using two-tailed paired t-tests, n=8). **b)** Schematic for electrical recordings, showing targeted
656 brain regions (left) and representative local field potentials (middle) recorded on the days
657 indicated on the bottom timeline during isolation in the home cage, as well as during repeated
658 exposure to social contexts that produce aggressive (e.g., exposure to an intact male) and non-
659 aggressive (e.g., exposure to female or castrated male mouse) social behavior (right). **c)**
660 Framework to test individual brain regions' encoding of social states (left). All implanted regions
661 distinguished social engagement from isolation in test mice (N=9, right, 1st graph); however,
662 only five selectively encoded the aggressive behavior vs. non-aggressive behavior (right, 2nd

graph). Pink shading indicates P<0.05 with FDR correction. **d)** The most aggressive mice from the training group (N=8) were allowed to freely explore a clean inanimate object or an object covered in urine from another intact CD1 male mouse over seven sessions. Neural signatures for aggression discovered from the five individual brain regions in panel (c) 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 electome-based decoding and reconstruction of aggressive state from multi-region LFP activity. (Top) Multi-region LFP recordings are segmented into time windows aligned with labels of non-aggressive/aggressive behavior and the sex of the intruder. LFPs are transformed into features describing distributed neural activity. (Middle) These features and labels are then used to train a machine learning model that learns electome networks with corresponding network-level activity patterns across windows. (Bottom) The network activity from a single network is then used to predict aggressive behavior (left), while the learned network structure enables reconstruction of LFP features across brain regions (right), providing an interpretable representation of the distributed neural dynamics associated with aggression. **f)** Encoding across eight learned networks (1 supervised network and 7 unsupervised networks) in hold-out test mice (N=9). The supervised network (purple bar) showed the strongest encoding (AUC=0.67±0.03, one-tailed Wilcoxon rank sum, $P=2\times10^{-5}$). Data shown as mean ± SEM. The red line corresponds with an AUC of 0.5, which represents the model's inability to distinguish between aggressive and pro-social behavior.

Figure 2. *EN-Aggression Inhibition (EN-AggINH)* encodes an aggressive internal state. a) Neural activity was sampled while 17 mice were socially isolated (blue) and during intervals preceding and following their aggressive behavior towards intact male mice and non-aggressive social behavior towards female mice. These data windows were not used to train the network model since they did not contain the behaviors of interest. **b)** (Left) *EN-AggINH* activity during these preceding and following intervals encoded an aggressive psychological state (red) vs. male and female non-aggressive psychological states (green; N=17 mice from the training set; $F_{3,67}$=28.6, P<0.0001, Friedman test; post-hoc two-tailed Wilcoxon signed-rank test, P<0.005 after FDR correction). Individual animals are shown with gray lines. Data shown as mean ±

692 s.e.m. (Right) *EN-AggINH* activity during isolation negatively correlated with the time spent
693 exhibiting aggressive behavior towards other males (R=-0.58; P=0.016, two-tailed Spearman's
694 rank correlation).

695 **Figure 3. Dynamics and biological components of *EN-Aggression Inhibition (EN-AggINH)*. a)**
696 Prominent oscillatory frequency bands composing *EN-AggINH* are highlighted for each brain
697 region around the rim of the circle plot. Prominent coherence measures are depicted by lines
698 connecting brain regions through the center of the circle. The plot is shown at relative spectral
699 energy of 0.4, corresponding to the top 15% of predictors. Theta (4-11 Hz) and beta (14-30 Hz)
700 frequency components are highlighted in blue and green, respectively. **b)** Example relative LFP
701 spectral energy (RLE) plots (predictor weights) for three brain regions corresponding to the
702 circular plot in panel A (See Supplementary Fig. S5-6 for full description of network features).
703 The red line corresponds to the 0.4 RLE threshold utilized for the circular plot shown in A. **c)**
704 Directionality measures were used to depict directionality offsets (difference between A→B
705 and B→A) within *EN-AggINH*. Prominent directionality offsets were observed across the theta
706 and beta frequency bands (shown at spectral density threshold of 0.4 and a directionality offset
707 of 0.3). Histograms quantify the number of leading and lagging interactions between brain
708 regions. **d)** Schematic depicting directionality within *EN-AggINH*. **e-f)** *EN-AggINH* activity maps
709 to cellular activity. e) Three cells from lateral habenula (LHb), ventromedial hypothalamus
710 (VMH), and medial amygdala (MeA) showing firing activity that is negatively correlated with *EN-*
711 *AggINH* activity (red) and a VHip cell showing positively correlated firing (blue). f) *EN-AggINH*
712 activity correlated with cellular firing across the 11 recorded regions. Single- and multi-units
713 were used for analyses.

714 **Figure 4. *EN-Aggression Inhibition (EN-AggINH)* encodes distinct aggression contexts. a)**
715 Experimental approach for causally inducing aggression via chemogenetic activation of ESR1+
716 cells in ventromedial hypothalamus (VMH, region in red). After three weeks for viral expression,
717 we implanted recording electrodes to target the same brain regions used to train the network
718 model as in our initial experiment. Following recovery, we obtained behavioral and neural
719 recordings when mice were treated with saline or CNO. **b)** ESR1+ cell activation induced
720 aggressive behavior towards female mice (P<0.01 using Wilcoxon signed-rank test, N=8 mice),

721 and **c)** decreased *EN-AggINH* activity during social interactions with female mice (P<0.01 using
722 one-tailed Wilcoxon signed-rank test; N=8 mice) and **d)** during intervals surrounding these
723 interactions (P<0.05 using Wilcoxon signed-rank test). **P<0.01.

724 **Figure 5. *EN-Aggression Inhibition (EN-AggINH)* activity is causally related to aggression. a)**
725 Schematic for closed-loop manipulation of *EN-AggINH* activity. **b)** Real-time estimation of
726 aggression. Receiver operating characteristic depicting detection of aggressive behavior in a
727 mouse using *EN-AggINH* activity vs. real-time reduced encoder (i.e., based solely on power and
728 coherence features) is shown to the right. Dashed blue line corresponds to the established
729 detection threshold. **c)** Detection of aggression using reduced encoder vs. *EN-AggINH* across
730 mice (N=9; P=0.43 using two-tailed paired Wilcoxon signed-rank). **d)** *EN-AggINH* activity in
731 response to blue and yellow light stimulation during closed-loop manipulation. Network activity
732 significantly decreased one second prior to yellow light stimulation (N=9, ##P<0.005 using one-
733 tailed Wilcoxon signed-rank test for within-subject comparison of *EN-AggINH* activity 1 vs. 2
734 seconds prior to yellow light stimulation; note that activity was normalized to the average
735 network activity during isolation for each mouse). Activity was also higher one second after
736 stimulation with blue light vs. yellow light (**P<0.01 using one-tailed Wilcoxon signed-rank test).
737 **e)** Directed graph with the inferred modes of action derived from mediation analysis. Model 1:
738 There is a causal relationship from stimulation to behavior and from stimulation to *EN-AggINH*
739 expression (P<0.001 using Wilcoxon signed-rank for network activity and paired t-tests for
740 behavior). Model 2: *EN-AggINH* is a mediator from stimulation to behavior (P<0.01 using nested
741 logistic regression models, likelihood ratio test). Model 3: *EN-AggINH* activation is the primary
742 mechanism whereby mPFC stimulation suppresses aggression (P<0.01 using average causal
743 mediation effect). **f-g)** Directed graphs testing f) theta power within single regions as the
744 primary mechanism whereby mPFC stimulation suppresses aggression (model 4) and g) *EN-*
745 *AggINH* activation as the primary mechanism whereby mPFC stimulation suppresses aggression
746 when single-region power is included as an intermediary (model 5). The uncorrected P values
747 for each brain area in both models are shown below as -ln(P), where the red dashed line
748 corresponds to P=0.05.

749 **Figure 6. Validation of spatiotemporal features of *EN-Aggression Inhibition (EN-AggINH).* a)**
750 Percentage of windows stimulated during social behaviors using a closed-loop approach, which
751 preferentially induced optogenetic stimulation during aggressive interactions, compared to
752 non-aggressive social interaction with females (P=0.002 using one-tailed Wilcoxon signed-rank
753 test, left). Behavioral effects of closed-loop stimulation are shown on the right (blue and yellow
754 light stimulation compared using paired t-test, significance determined using FDR correction).
755 **P<0.01 for the entire panel. **b)** Schematic for nonsynchronous control stimulation (left).
756 Nonsynchronous stimulation does not impact aggressive or non-aggressive social behavior
757 towards males or females. **c)** Directionality for mPFC-dependent subcircuits within *EN-AggINH*
758 (shown as relative LFP spectral energy or RLE, see also Supplementary Fig. S6). **P<0.01 using
759 bootstrapping procedure over 1000 repeats. **d)** Viral targeting strategy (left) and behavioral
760 impact of PL→NAc circuit stimulation (right). **e)** Viral targeting strategy (left) and behavioral
761 impact of PL→MeA circuit stimulation (right). **f)** Viral targeting strategy (left) and behavioral
762 impact of PL→LH circuit stimulation (right). For all panels B-F: **P<0.01 using two-tailed paired
763 t-test. Order of blue and yellow light stimulation trials is shown next to the right of each social
764 condition diagram in panels A-B, D-F. **g)** Left: Schematic of electrical synapse formation for
765 circuit-selective LinCx editing. $Cx34.7_{M1}$ (green hemichannel) and $Cx35_{M1}$ (red hemichannel)
766 show heterotypic docking with each other, but not homotypic docking between two $Cx34.7_{M1}$
767 hemichannels. The synapse formed between them rectifies in the $Cx34.7_{M1}$ to $Cx35_{M1}$ direction.
768 Right: Schematic of $Cx34.7_{M1}$ and $Cx35_{M1}$ vs. control fluorophore targeting the mPFC and NAc
769 specifically. **h)** Impact of LinCx editing the mPFC-NAc circuit on aggressive (left) and non-
770 aggressive behavior towards intact males (middle), and non-aggressive behavior towards
771 females (right). Data is plotted as mean ± s.e.m. at 2, 5, and 8 weeks after viral injection
772 (*P<0.05 using repeated measures ANOVA followed by two-tailed unpaired t-test across
773 groups; N=10-11 mice per group). Box and whisker plots to the left of each plot show social
774 behavior in CD1 animals following group assignment, but prior to viral injection. No differences
775 were observed across groups for any of the three social behaviors (P>0.05 for all comparisons
776 across groups using an unpaired two-tailed t-test). For the box and whisker plots, the central
777 mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers

extend to the most extreme data points the algorithm does not consider to be outliers. Individual data points are shown as circles.

Author contributions

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

Declaration of Interests

The authors declare no competing interests.

Materials and Methods

Animal care and use

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

45 six-month-old, retired breeder male CD1 strain mice (Charles River Laboratories, Wilmington, Massachusetts) were used to discover a network that encoded aggression, hereafter called *EN-AggINH*. Another 226 male CD1 mice were used to examine behavioral and network responses to optogenetic stimulation. Mice were singly housed with enrichment.

802 ESR1-Cre male mice on a C57BL/6J background were provided by Scott Russo for the
803 chemogenetic experiments. These mice were crossed with female CD1 strain mice (Charles
804 River Laboratories, Wilmington, Massachusetts) in the Bryan Vivarium at Duke University to
805 obtain F1 offspring. 13 fourteen-week-old virgin ESR1-Cre+/CD1 male offspring were used to
806 validate *EN-AggINH*. All F1 offspring were group-housed 2-5 mice per cage until they received
807 viral injections in the ventromedial hypothalamus at 7-8 weeks. After surgery, these mice were
808 singly housed with enrichment.

809 60 six-month-old, retired breeder male CD1 strain mice were used for the LinCx editing
810 experiment. Mice were singly housed with enrichment

811 All intruder mice (C57BL/6J: 2-7 intact males, 2-7 females, and 2-7 castrated males per
812 experimental mouse) were 7-14 weeks old. These mice were purchased from Jackson
813 Laboratories (Bar Harbor, Maine) and were housed 5 per cage with enrichment. All behavioral
814 testing and neurophysiological recordings occurred during the dark cycle.

815 <u>Castration of C57BL/6J male mice</u>

816 18 male mice were anesthetized with 1% isoflurane. The scrotal sac was sanitized with betadine
817 and 70% ethanol. The testes were then moved into the sac by gently palpating the lower
818 abdomen. Next, an incision was made in the sac and the testes were extracted. After blood flow
819 was cut off to the testes using a thread tourniquet, the testes were severed from the mouse.
820 The remaining fatty tissue was placed back into the scrotum, which was then sutured. Mice
821 were allowed 10 days for recovery prior to experimental use.

822 <u>Electrode implantation surgery</u>

823 Mice were anesthetized with 1% isoflurane and placed in a stereotaxic device. Anchor screws
824 were placed above the cerebellum, right parietal hemisphere, and anterior cranium. The
825 recording bundles designed to target prelimbic cortex, infralimbic cortex, medial amygdala,
826 ventral hippocampus, primary visual cortex, mediodorsal thalamus, lateral habenula, lateral
827 septum nucleus, nucleus accumbens, ventrolateral portion of the ventromedial hypothalamus,
828 and orbital frontal cortex were centered based on stereotaxic coordinates measured from
829 bregma. [Orbital frontal cortex: anterior/posterior (AP) 2.35mm, medial/lateral (ML) 1.0mm,

830 dorsal/ventral (DV) from dura -2.75mm; infralimbic cortex and prelimbic cortex: AP 1.8mm, ML
831 0mm, DV -2.7mm from dura; medial amygdala: AP -1.25, ML 2.7mm, DV -4.3 from dura; lateral
832 septum and nucleus accumbens: AP 1.0mm, ML 0mm, DV -4.0mm from dura; ventromedial
833 hypothalamus, lateral habenula, and medial dorsal thalamus: AP -1.47mm, ML 0mm, DV -
834 5.4mm from dura; ventral hippocampus and primary visual cortex: AP -3.0mm, ML 2.6mm, DV -
835 3.0mm from dura]. We targeted infralimbic cortex and prelimbic cortex by building a 0.6mm DV
836 stagger into the bundle. We targeted lateral septum and nucleus accumbens by building a
837 0.3mm ML and 1.5mm DV stagger into the bundle. We targeted lateral habenula, medial dorsal
838 thalamus, and ventral medial hypothalamus by building a 0.3mm ML, and 1.9mm and 2.5mm
839 DV stagger into our electrode bundle microwires. We targeted primary visual cortex and ventral
840 hippocampus using a 0.3mm ML and 2.5mm DV stagger in our electrode bundle microwires. For
841 open- and closed-loop optogenetic stimulation experiments, the addition of a Mono Fiberoptic
842 Cannula coupled to a 2.5mm metal ferrule (NA: 0.22, 100µm [inner diameter], 125µm buffer
843 [outer diameter], MFC_100/125-0.22, Doric Lenses, Quebec) was built into the medial
844 prefrontal cortex bundle. The tip of the fiber was secured 300µm above the tip of the IL
845 microwires. Mice were allowed 10-15 days for recovery from surgery before behavioral testing.
846 For the medial prefrontal cortex projection-targeted physiological experiment, mice were
847 implanted in medial prefrontal cortex (mPFC), and one of the three projection-targeted regions
848 (nucleus accumbens [NAc], medial amygdala [MeA], or lateral habenula [LHb]) using the
849 coordinates described above. A 2.5mm metal ferrule (NA: 0.22, 100µm [inner diameter],
850 125µm buffer [outer diameter], MFC_100/125-0.22, Doric Lenses, Quebec) was built into the
851 microwire bundles targeting medial prefrontal cortex and the bundle targeting medial
852 prefrontal cortex terminals in the projecting region.

853 Viral surgery

854 For optogenetic experiments targeting mPFC soma, we used CD1 mice that showed an attack
855 latency < 60s when exposed to an intact C57BL/6J (C57) male. 40 CD1 mice were anesthetized
856 with 1% isoflurane and placed in a stereotaxic device. The mice were unilaterally injected with
857 AAV2-CamKII-ChR2-EYFP (purchased from the Duke Viral Vector Core, Durham, NC; courtesy of
858 K. Deisseroth), based on stereotaxic coordinates from bregma (left infralimbic cortex: AP

859 1.8mm, ML 0.3mm, DV -2.0mm from the brain). A total of 0.5µL of virus was infused at the
860 injection site at a rate of 0.1µL/min over 5 minutes, and the needle was left in place for 10
861 minutes after injection. For the open-loop stimulation experiment, CD1 mice were implanted
862 with an optic fiber (Mono Fiberoptic Cannula coupled to a 2.5mm metal ferrule (NA: 0.22,
863 100µm [inner diameter], 125µm buffer [outer diameter], MFC_100/125-0.22, Doric Lenses,
864 Quebec)) 0.3mm above the injection site immediately after viral syringe was removed. These
865 mice were allowed 3 weeks for recovery prior to behavioral testing. For the closed-loop
866 experiments, 11 CD1 mice were allowed 3 weeks for viral expression prior to implantation with
867 an optrode. For the nonsynchronous stimulation experiment, 19 CD1 mice were allowed 3
868 weeks for viral expression prior to implantation with an optic fiber with an electrode connector
869 attached.

870 For the ESR1-Cre validation experiment, 13 ESR1-Cre+/CD1 F1 offspring were bilaterally
871 injected with AAV2-hSyn-DIO-GqDREADD (obtained from Addgene) based on stereotaxic
872 coordinates measured from bregma (AP -1.5mm, ML ±0.7mm, DV -5.7mm from the dura). A
873 total of 0.3µL of virus was infused bilaterally at a rate of 0.1µL/min, and the needle was left in
874 place for 5 minutes after injection. Two weeks after viral infusion, F1 males were screened for
875 aggressive behavior towards females. The F1 males received i.p. injections of CNO (1mg/kg) at
876 the start of the screening session. 35 minutes after injection, a novel C57 female was placed in
877 the home cage for 5 minutes. Screening was repeated one week and two weeks later. Only F1
878 males who attacked females for at least two of the three screening sessions (9/13 mice) were
879 implanted with electrodes [18]. The eight mice that showed good surgical recovery were
880 subjected to further experiments.

881 For mPFC projection-targeted behavioral experiments, we used 46 male CD1 mice that showed
882 an attack latency <60s and initiated attacks at least three times within three minutes when
883 exposed to a C57 male mouse (categorized as aggressive mice). These mice were unilaterally
884 injected with AAV2-EF1a-DIO-ChR2-eYFP (obtained from Addgene) in the left mPFC based on
885 stereotaxic coordinates measured from bregma (AP 1.8mm, ML 0.3mm, DV -2.0mm from the
886 dura), and AAVrg-EF1a-Cre-mCherry (obtained from Addgene) was injected in the downstream
887 target region based on stereotaxic coordinates measured from bregma (NAc: AP 1.0mm, ML

888 0.9mm, DV -3.8mm from the dura; MeA: AP -1.25mm, ML 2.75mm, DV -4.3mm from the dura;
889 or LHb: AP -1.6mm, ML 0.4mm, DV -2.2mm from the dura). A total of 0.3µL of virus was infused
890 in the mPFC and 0.3µL of virus was infused in the downstream target region at a rate of
891 0.1µL/min. The needle was left in place for five minutes after injection. Immediately after the
892 viral syringe was removed from the downstream target region, mice were implanted with an
893 optic fiber (Mono Fiberoptic Cannula coupled to a 2.5mm metal ferrule (NA: 0.22, 100µm [inner
894 diameter], 125µm buffer [outer diameter], MFC_100/125-0.22, Doric Lenses, Quebec)) 0.3mm
895 above the downstream target region injection site. Four weeks after viral infusion/optic fiber
896 implantation, CD1 mice were tested for the behavioral effects of projection-targeted
897 stimulation.

898 For mPFC projection-targeted electrophysiological experiments, six male CD1 mice that did not
899 show aggression during screening were anesthetized with 1% isoflurane and placed in a
900 stereotaxic apparatus. Anchor screws were placed over the anterior cranium, parietal cortex,
901 and cerebellum. These mice were injected with AAV-CAG-DIO-ChR2-eYFP and AAVrg-EF1a-IRES-
902 Cre-mCherry, as described above, to target mPFC→NAc (N=2 mice), mPFC→MeA (N=2 mice), or
903 mPFC→LHb (N=2 mice). Optrodes targeting mPFC and the target regions were then implanted
904 as described.

905 For the LinCx circuit-editing experiment, 60 retired breeder male CD1 mice were screened for
906 aggressive behavior over three days for 5 minutes/day. Only mice that 1) attacked on days 2
907 and 3, 2) had an average total attack time >20 seconds, and 3) had an average latency to first
908 attack <60 seconds were included in the experiment. Anticlust in R was used to separate these
909 mice into two groups (LinCx-edited mice and control mice) with equivalent attack and non-
910 attack engagement times. Group assignments were performed independently across two
911 cohorts of mice (N=7 and 5 control mice, and N=7 and 6 LinCx-edited mice for cohort 1 and 2,
912 respectively). Mice in the LinCx-edited experimental group were then injected with 0.3µL AAV9-
913 CaMKII-Cx34.7$_{M1}$-T2A-mEmerald in left mPFC (AP 1.8 mm, ML 0.3 mm, DV -2.0 mm from dura)
914 and 0.3µL AAV9-CaMKII-Cx35$_{M1}$-T2A-mCherry in left NAc (AP 1.0 mm, ML 0.9 mm, DV -3.8 mm).
915 Virus was injected at 0.1µL/min using a microinjection syringe pump (WPI Micro2T SMARTouch
916 with UMP3 pumps). Syringe remained in place for at least 5 minutes after each injection.

917 Control mice were injected with AAV9-CaMKII-EGFP (Addgene viral prep # 50469-AAV9, a gift
918 from Bryan Roth) in mPFC and AAV9-CaMKII-mCherry (Addgene viral prep # 114469-AAV9, a
919 gift from Karl Deisseroth) in NAc using the same procedure noted for the LinCx-edited mice.

920 Histological analysis

921 Histological analysis of implantation sites was performed at the conclusion of experiments to
922 confirm recording sites and viral expression. Animals were perfused with 4% paraformaldehyde
923 (PFA), and brains were harvested and stored for at least 96 hrs in PFA. Brains were
924 cryoprotected with sucrose and frozen in OCT compound and stored at -80°C. Brains were later
925 sliced at 40µm. Brains from mice used to train and validate the network were stained using
926 NeuroTrace fluorescent Nissl Stain (N21480, ThermoFisher Scientific, Waltham, MA) using
927 standard protocol. Specifically, Nissl staining for brain tissue occurred on a shaker table at room
928 temperature. Tissue was washed in PBST (0.1% Triton in phosphate-buffered saline solution) for
929 10 minutes. It was then washed for 5 minutes in PBS twice. The tissue was then protected from
930 light for the remainder of the protocol. The tissue was incubated in 1:300 Nissl diluted in 2 mL
931 PBS for 10 minutes. After the Nissl incubation, tissue was washed once in 0.1% PBST for 10
932 minutes, then twice in PBS for 5 minutes. Brains from ESR1 mice and mice used for 20 Hz open-
933 or closed-loop stimulation were mounted in Vectashield mounting medium containing DAPI (H-
934 1200-10, Vector Laboratories, Newark, CA). Images were obtained at 10x using an Olympus
935 fluorescent microscope.

936 Neurophysiological data acquisition

937 Mice were connected to a Cerebus data acquisition system (Blackrock Microsystems, UT, USA)
938 while anesthetized with 1% isoflurane. Mice were allowed 60 minutes in their home cage prior
939 to behavioral and electrophysiological recordings. Local field potentials (LFPs) were bandpass
940 filtered at 0.5-250Hz and stored at 1000Hz. An online noise cancellation algorithm was applied
941 to reduce 60Hz artifact (Blackrock Microsystems, UT, USA). Neural spiking data was referenced
942 online against a channel recording from the same brain area that did not exhibit a SNR>3:1.
943 After recording, cells were sorted using an offline sorting algorithm (Plexon Inc., TX) to confirm
944 the quality of the recorded cells. Only cell clusters well-isolated compared to background noise,

945 defined as a Mahalanobis distance greater than 3 compared to the origin, were used for the
946 unit-electome network correlation analysis. We used both single (well-isolated clusters with
947 ISI<1.5) and multi-units (well-isolated clusters with ISI>1.5; N=186 total neurons) for our
948 analysis as our objective was to determine whether electome network activity was reflective of
949 cellular activity. Neurophysiological recordings were referenced to a ground wire connected to
950 anchor screws above the cerebellum and anterior cranium.

951 Behavioral recordings and analysis for training/testing models

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

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

970 After screening, 29 mice were implanted, and data was acquired across 1-6 behavioral
971 testing/recording sessions following recovery. Sessions were separated by 5-7 days. Recordings
972 for all social encounters were performed in the home cages of the CD1 mice. Each behavioral

973 testing session began with 5 minutes of recording prior to introduction of the social stimulus.
974 All mice were subjected to encounters with an intact C57 male mouse and a female C57 mouse.
975 A subset of 18 CD1 mice were also subjected to an encounter with a castrated male mouse. For
976 this selection, the aggressiveness of each mouse was defined as the median latency to an attack
977 an intact male intruder across the six initial recording sessions. The eight most aggressive mice
978 were also subjected to exposure to objects covered in CD1 mouse urine. Object pairs (one clean
979 and the other covered in urine) included yellow duplex blocks, curved red duplex blocks,
980 weighted 5 mL conicals, glassware tops, and objects assembled from black Legos®. The CD1
981 mice were exposed to a different pair of objects during each session. Order of exposure to
982 stimulus mice and objects was shuffled for every session. Six CD1 mice were recorded under all
983 four conditions. Data observations (1 second each) were pooled across 20 CD1 mice for training
984 the network model.
985 For ESR1-Cre+/CD1 male behavioral testing, eight mice were injected with either saline or CNO
986 (1mg/kg, i.p.) after the five-minute baseline recording. 35 minutes after this injection, mice
987 were exposed to an intact male C57, a castrated male C57, and a female C57, presented in
988 pseudorandom order. Mice were subjected to six total recording sessions, three in which they
989 were treated with saline and three in which they were treated with CNO, again in
990 pseudorandom order. Sessions were separated by 5 days to allow an adequate washout of
991 CNO[54].

992 Behavior was scored for each second as an “aggressive behavior”, “non-aggressive social
993 interaction”, or “non-interaction”. One-second windows were identified as "aggressive
994 behavior" if the mouse was engaged in biting, boxing (kicking/clawing), or tussling behavior [34].
995 Windows were labeled as "non-aggressive social interaction" if the mouse had his nose or
996 forepaws touching the stimulus mouse (intact male/female/castrated male) or object, but was
997 not biting, boxing, or tussling. Examples of behaviors labeled as "non-aggressive social
998 interaction" included sniffing, grooming, or resting (placing nose or forepaws against the
999 subject mouse, but not moving). If the stimulus mouse had his/her forepaws or nose on the
1000 CD1 but it was not reciprocated, this was labeled as "non-interaction". In our experimental
1001 settings, CD1 straight approach, sideways approach, and chasing of the stimulus mouse could

1002 result in aggressive behavior (biting/kicking/tousling), non-aggressive social interaction (nose or
1003 paw touch), or withdrawal without any touch. Thus, while sideways approach and chasing are
1004 regularly labeled as "aggressive" in the literature, and straight approach is not[34,55,56], these
1005 behaviors lacked consistent resolutions. Moreover, mice also demonstrated these behaviors
1006 towards female and castrated mice (i.e., in non-aggressive social behavior contexts). Therefore,
1007 one-second windows containing these behaviors were labeled as "non-interaction". All other
1008 timepoints not labeled as "aggressive behavior" or "non-aggressive social interaction" were also
1009 labeled as "non-interaction". These behavioral criteria were selected to include ethologically
1010 aggression-related behaviors and maximize the likelihood that the CD1 was aware of the
1011 presence of the stimulus mouse or object during the behavioral window, while remaining
1012 confident in the classification of "aggressive behavior" and "non-aggressive social interaction"
1013 window labels.

1014 Tail rattling is well-recognized in the literature as an aggressive behavior[34], and it was
1015 consistently only demonstrated by aggressive mice towards intact male mice in our
1016 experimental settings. Thus, we included this behavior in the “aggressive behavior” category.
1017 In our subset of 20 mice used for training the network, tail rattling was observed 8 ± 4s out of
1018 the 135 ± 26s “aggressive behavior” windows per mouse.

1019 The videos used to generate the behavioral labels for training and testing our machine learning
1020 model were hand-scored by a trained researcher. Videos from ESR1-Cre+/CD1 mice and
1021 optogenetic stimulation were automatically tracked using DeepLabCut [57,58]. This information
1022 was then used for creating behavioral classifiers in SimBA [59].

1023 <u>LFP preprocessing and signal artifact removal</u>

1024 Each LFP signal was segmented into 1s non-overlapping windows. If there were multiple intact
1025 channels implanted in a region, they were averaged to produce a single signal. Windows with
1026 non-physiological noise were removed using an established automated heuristic [8]. Briefly, the
1027 envelope of the signal in each channel was estimated using the magnitude of the Hilbert
1028 transform. The Median Absolute Deviation (MAD) of the magnitude was then calculated on
1029 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 a standard deviation less than 0.01 were removed as well. If no channels were usable for a given region within a window, that whole window was removed from the data. This set of heuristics resulted in 34.7±5.1% of the data windows 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, coherence, and 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 1-second window and 1-second steps at a 1Hz resolution. Coherence was estimated pairwise between all recorded 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.

Directionality features -- a correlate of directional transfer of information across a brain circuit -- were estimated using the Multivariate Granger causality (MVGC) MATLAB toolbox [60]. 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. These directionality features were processed identically to a previously reported approach [8]. Briefly, directionality features were exponentiated to approximately maintain the additivity assumption made implicitly by Nonnegative Matrix Factorization (NMF) models [8,61]

as, $\exp\left(f_{A\to B}(\omega)\right)$, where $f_{A\to B}(\omega)$ is the Granger causality at frequency ω from region A to region B. The exponentiated feature is a ratio of total power to unexplained power. Exponentiated directionality feature values were truncated at 10 to prevent implausible values.

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

We used 21460 seconds of behavioral/neural data, pooled across the 20 mice to train/validate our single-region and network models. This included a total of 4680 seconds while mice were socially isolated in their home cage, 14890 seconds when CD1 mice exhibited non-aggressive social behavior (3542 seconds towards intact males, 9067 seconds towards females, and 2281 seconds towards castrated), and 1890 seconds when mice exhibited aggressive behavior towards the intact males.

Discriminative Cross-Spectral Factor Analysis – Nonnegative Matrix Factorization

Our machine learning approach was designed to distinguish between behavioral windows when the CD1 mice showed aggressive behavior towards intact C57 males and windows when they exhibited pro-social behavior (interactions towards intact C57 males, castrated C57 males, or C57 females), and it was designed also to model neural variance. Here, we used neural and behavioral data from 29 mice to establish the final model, with a split of 20 and 9 mice for model training and testing, respectively.

Our approach used the Discriminative Cross-Spectral Factor Analysis – Nonnegative Matrix Factorization (dCSFA-NMF) model [38]. This approach assumes that each window is an independent stationary observation and examines dynamics in brain activity only at the scale of windows. In this case, the relevant dynamics happen at the scale of 1-second windows . A 1-second window was chosen to balance capturing fine-grained transient behavior with sufficient length to properly estimate spectral features [8]. Each window has associated spectral power, coherence, and directionality features (in total $p = 9{,}586$ features), which is represented as $x_i \in \mathbb{R}^p$ for the ith window. Each window (Fig. 1e, top; Supplementary Fig. S3a, top) was associated with a behavioral label that identified which condition the CD1 mouse was subjected to during that window (intact male, castrated male, or female) and whether the CD1 mouse

was engaged in aggressive or non-aggressive behavior during that window (coded as $y_i \in \{0,1\}$).

As a short description of the dCSFA-NMF model, all spectral power, coherence, and directionality features are described as an additive positive sum of K nonnegative electrical functional connectome (electome) networks. This model is learned using a supervised autoencoder. The objective we use to learn the parameters is

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

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 $W \in R^{p\times K}$, where each column will represent an electome network. This matrix can also be thought of as a decoder network that is limited to a single linear layer in an autoencoder framework. The electome network activities are given by the multi-output encoder function $f(x;\phi): R^p \rightarrow R^K$, where ϕ represent the parameters of the function. These activities correspond to how strongly expressed a network is within a window, allowing for variations in network strength over time. 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 networks' activities and the behavior. A sparsity constraint is enforced so that $d = [d_1, 0, \dots, 0]$, meaning that only a single electome is used to predict aggressive or non-aggressive behavior, simplifying interpretation (Fig. 1e, bottom left; Supplementary Fig. S3a, bottom). The first electome network is referred to as the supervised network, since it is used to predict the aggressive state, and the rest are referred to as unsupervised electome networks, as they are not directly linked to the behavior of interest. λ is a weighting parameter used to control the relative importance of prediction and reconstruction. In other words, this term varies how much the training values predicting aggressive or non-aggressive behavior compared to

explaining the LFP feature variance. We chose a value of $\lambda = 1$, which kept the two losses approximately equal during training.

Previous work has also found that the reconstruction loss can reduce overfitting and make the learned predictions more robust [62]. To further reduce overfitting of the predictive aspect of the network activities, we applied an elastic net loss[63], $Loss_{EN}$, on the encoder parameter A that produce the network activities. $Loss_{EN}$ has 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 [64]. In this work, power features were also upweighted by a factor of 10 to accommodate that there were many more coherence and directionality 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 [65]. 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 [64].

Predictive performance of the electome network was evaluated using data from new mice that were not included in learning the electome network. Network scores were estimated as an evaluation of the encoder learned during training of the dCSFA-NMF model.

<u>Hyper-parameter selection</u>

The dCSFA-NMF procedure requires selection of several settings in the algorithm. Specifically, we must choose: the number of electome networks K; the importance of the supervised task λ; the relative importance of the power features, coherence features, and directionality features; and the parameterization of the mapping function $f(x_i; \phi)$, which is referred to as the encoder. Besides K, these settings were chosen to match previously used values or to follow heuristics. Specifically, in our prior work, we demonstrated that the inferred model is highly insensitive to λ [64]. Thus, we chose a λ value to give roughly equal weight to the predictive and generative tasks. Similarly, since the number of power features grows linearly and the number of coherence and directionality features grows quadratically with the number of brain regions (as they are

pairwise), we chose to weight the power features to roughly match the coherence features. Since the decoder is chosen to be linear (given by the matrix W above), as noted above we also chose a linear mapping function for the encoder followed by a softplus function to ensure non-negativity. This approach served to limit complexity.

To choose the value of K, we evaluated the reconstruction error (Mean Squared Error) on the 20 training mice, which determined how well the electome networks described the neural features, how consistent the electome networks were, and how well they predicted the behavioral state (aggressive vs. non-aggressive) as we varied the total number of electome networks K while using a single supervised network (e.g., 1 supervised network and $K-1$ unsupervised networks). We did not expect the prediction quality to be especially dependent on the number of electome networks, as our previous work has demonstrated that $K$is not an especially important parameter in terms of predictive performance [64], but that it can impact reconstruction error and total model fit.

For our main selection criteria, we used the Bayesian Information Criterion (BIC) to select the number of unsupervised networks to use in the final network model. The BIC is defined as:

$$BIC(K) = Kp \log N - 2\log(\hat{L}),$$

where Kp is the number of model parameters (p is the number of spectral features), N is the number of samples, and $\hat{L}$ is the likelihood of the observed data using all estimated model parameters (ϕ, W, d). This criterion balances the model fit quantified by $\log(\hat{L})$ with the complexity quantified by $Kp \log N$. In this work, $-\log(\hat{L})$ is an L_2 loss, corresponding to a Gaussian observational likelihood, as used in the training above. Using the training animals above, the model parameter$s$ were learned from 80% of the data while model parameter $\hat{L}$ was evaluated on a 20% hold-out set to avoid overfitting. The BIC was evaluated for all dimensionalities from 1-20 networks, and the lowest value was selected as the best model. Since 7 unsupervised networks provided the best fit (a BIC of 5457701, see also Supplementary Fig. S2b), our final full network model utilized a total of 8 networks, 1 supervised and 7 unsupervised, across all 11 regions.

1165 For each single-region model, we trained 3 unsupervised networks and a single supervised
1166 network. Here, we reduced the number of networks compared to the full network model, given
1167 the dramatic reduction in the number of covariates considered by the model, as there are no
1168 longer pairwise features calculated between regions. Our objective was to compare the
1169 predictive performance of the single-region models against each other and the full network
1170 model. Since the predictive performance is driven by the supervised network[64], the smaller
1171 latent dimensionality of the single-region models had no impact on our final conclusions.

1172 Single-region decoding of social behavioral state

1173 To test the efficacy of any single brain region as a biomarker for aggression, we extracted
1174 power at 1-Hz frequency bins over 1-56 Hz from each region. One-second windows were
1175 pooled from the 20 CD1 training mice to generate a series of dCSFA-NMF models for each of
1176 the 11 brain regions (see above section for model hyperparameters). The models were trained
1177 to distinguish behavioral windows of one among three social states exhibited by CD1 mice: 1)
1178 male-directed aggressive behavior, 2) female non-aggressive social interactions, and 3)
1179 castrated male non-aggressive social interactions. We also developed a model to distinguish 4)
1180 periods when CD1 mice were isolated in their home cage from any of the three social states.
1181 Each model was then tested on data from a hold-out set of nine mice. The area under the
1182 receiver operating curve (AUC) was calculated for each hold-out mouse to determine the
1183 performance of the model. False discovery rate was used to correct for multiple hypothesis
1184 testing.

1185 Sensitivity of supervised electome network parameters to total number of electome networks

1186 A frequent concern of latent variable models, including dCSFA-NMF, is the dependence of the
1187 networks parameters W and encoder parameters A on the choice of latent dimensionality K,
1188 which determines the total number of electome networks in our models. In particular, in our
1189 case, we are most interested in how robustly we learn the encoder parameters and network
1190 parameters related to the single supervised network. To address this concern, we performed a
1191 sensitivity analysis to determine the extent to which the choice of K influenced the supervised
1192 electome network parameters (first column of W) and encoder (first row of A). In this

sensitivity analysis, we estimated a dCSFA-NMF model allowing the number of total networks to range from 2-20 while keeping a single supervised network. We then compared the similarity between each learned supervised encoder and supervised electome network parameters to our model with eight networks (the final, full model used in this work that has 1 supervised and 7 unsupervised networks). This was quantified using the cosine similarity, which measures the angle between two vectors, 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 Fig. S2a). We applied this to the vectors that define the supervised electome network (also referred to as a decoder due to the autoencoder structure; first column of W) and the supervised encoder (first row of A).

The supervised encoders were virtually identical across all the models except the one that utilized three networks, which learned a network that was positively associated with aggression (Supplementary Fig. S2a). We found that the supervised network parameters (decoder) maintained a strong consistency across most dimensions, particularly between 5-10 networks, as shown by the cosine similarities being greater than 0.95.

To evaluate the how large these cosine similarities are, we created a null distribution of the similarities across randomly chosen generative networks, where we randomly matched networks, meaning that we randomly matched the rows of W corresponding to network parameters (decoder) and columns of A corresponding to a single randomly chosen network (encoder). These similarities were substantially lower. In summary, across a range of tested latent dimensionalities, the supervised electome model converged on highly similar behavior-linked network patterns. In other words, varying the number of latent components did not materially change the structure, regional composition, or behavioral associations of the learned electome network. This suggests that the aggression-related network is a stable, data-driven feature of the recordings rather than an artifact of selecting a particular model dimensionality. Importantly, this conclusion applies to the supervised electome framework (that is, networks constrained by behavioral prediction), rather than to purely unsupervised decompositions.

<u>Single-cell correlation to electome network activity</u>

1221 Data acquired during the third behavioral testing session was from the 20 implanted training
1222 mice were used for cellular analysis. We used Spearman correlation to quantify the relationship
1223 between cellular firing windows and the activity of the electome network used to classify
1224 aggressive behavior. We performed 1000 permutations by randomly shuffling 1-second
1225 windows within each class for aggressive behavior and non-aggressive social interactions with
1226 male, female and castrated C57 mice. This approach maintained the relationship between
1227 network activity and behavior and the relationship between cell firing and behavior. We then
1228 calculated the Spearman correlation between network activity and cell firing for each
1229 permutation. A cell was deemed positively correlated if its unshuffled Spearman Rho was above
1230 97.5% of the permuted distribution and negatively correlated if it was below 2.5%.

1231 High-density single unit recordings

1232 Three mice which showed aggression during screening were anesthetized with isoflurane (1%)
1233 and placed in a stereotaxic device, and metal ground screws were secured to the anterior
1234 cranium and above cerebellum. Multi-wire electrodes were implanted to target OFC, MeA,
1235 VHip, V1, VMHvl, MDThal, and LHb based on the coordinates noted above, and a 1024-channel
1236 silicon probe (NeuroNexus, SINAPS_4S_1024) was implanted to target IL, PL, NAc, and LSN (-
1237 1.7mm AP, centered at the midline, measured from bregma; - 4mm DV from the dura).
1238 Individual shanks were spaced by 500µm on the probe such that the front two targeted medial
1239 prefrontal cortex bilaterally at ±0.25mm ML. The multi-wire electrode implant and silicon probe
1240 were secured to the same ground.

1241 Following surgical recovery, mice were habituated to the recording cables for 1-2 weeks and
1242 tested again for aggression. One of the implanted mice failed to show aggression and was
1243 removed from further testing. Neural data was recorded from the other two implanted mice
1244 during exposures to intact male, castrated male, and female mice using the Cerebus System
1245 (Blackrock Microsystems, UT, USA), as described above. Neural data was sampled concurrently
1246 from the implanted silicon probe at 20kHz using a SmartBox Pro acquisition system
1247 (NeuroNexus). An analog timing signal, consisting of 10ms pulses with a pseudorandomized
1248 inter-pulse-interval of 8-24 seconds, was generated using the analog output from the Cerebus
1249 system, and it was stored at 1000Hz alongside recorded neural data via the analog input

channel on the SmartBox Pro. This timing signal, recorded concurrently by both systems, was used to align the two sets of neural data for analysis.

To extract single unit activity, neural activity was converted to the *Neurodata Without Borders (NWB)* format, high-pass filtered at 300Hz, and the medial 1024 channels were automatically sorted with Kilosort4[66]. Single units were identified based on ISI violations < 0.5, presence ratio > 0.9, an amplitude cutoff < 0.1 (473/1052 total sorted cells)[67,68].

Real-time encoder approximation for closed-loop optogenetic experiments

Because calculating all the directionality features was 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 *EN-AggINH* to use in the closed-loop stimulation experiments. This 'fast' model was trained on the same data used to learn the full network model. The model was trained using regularized regression to best predict the output of the full encoder, $f(x_i; \phi)$, defined above. As such, this reduced encoder is also an affine transformation followed by a softplus activation with a smaller parameter set, $\phi_r = \{A_r, b_r\}$. The only difference from the full encoder is that the reduced encoder uses a smaller number of neural features and is otherwise the same mathematical form. This approximation explained a large component of the variance of the supervised network score derived from the hold-out validation mice ($R = 0.47$, p-value $<10^{-16}$).

Optogenetic stimulation

Mice were anesthetized with 1% isoflurane, then tethered to an optic patch cable placed over the optic fiber cannula. The mice were then allowed 60 minutes for recovery prior to behavioral testing and undergoing two stimulation sessions. For each session, mice were exposed to intact male and female mice. A subset of mice was also exposed to a castrated male mouse. During each condition, the CD1 mouse received two segments of alternating yellow (593.5nm, OEM Laser Systems, Model No. MGL-F-593.5/80mW) and blue (473nm, Crystal Laser LC, Reno, NV. Model No. DL473-025-O) light stimulation, for two minutes each. During the first session, mice were stimulated with yellow light first (yellow, blue, yellow, blue), and during the second session mice were stimulated with blue light first (blue, yellow, blue, yellow). CD1 mice received

1278 light stimulation for the entirety of the two-minute segment. CD1 mice received light
1279 stimulation for the entirety of the two-minute segment.

1280 For closed-loop optogenetic stimulation, mice anesthetized with 1% isoflurane were tethered
1281 to an optic patch cable and connected to the recording system. The mice were then allowed 60
1282 minutes for recovery prior to behavioral testing/neural recordings. Mice experienced three
1283 sessions of behavioral testing followed by two sessions of closed-loop stimulation. Stimulation
1284 sessions were separated by 5-7 days. For each of the behavioral testing sessions, CD1 mice
1285 were exposed to intact C57 males, females, and castrated male mice for 5 minutes each.
1286 Testing sessions two and three were used to determine a reduced network threshold at which
1287 40% of aggressive behavioral windows could be detected. For each stimulation session, mice
1288 were recorded for 3 minutes of baseline in their home cage, then during the three social
1289 encounters. The order of the three social encounters was shuffled for each session. Mice were
1290 recorded in an open field for 5 minutes after each session. During each social condition, the
1291 CD1 mouse went through two-minute segments of potential blue light stimulation (473nm,
1292 Crystal Laser LC, Reno, NV. Model No. DL473-025-O), then yellow light stimulation (593.5nm,
1293 OEM Laser Systems, Model No. MGL-F-593.5/80mW), and blue and yellow light stimulation
1294 again. During these two-minute segments, mice only received stimulation for one second when
1295 the reduced network score dropped below threshold. Behavior was averaged across the two
1296 stimulation sessions.

1297 For nonsynchronous stimulation, each CD1 mouse was pseudorandomly matched to a different
1298 mouse that had been used for closed-loop stimulation. Each nonsynchronous mouse was then
1299 subjected to the identical order of social conditions and blue and yellow light stimulation blocks
1300 as their individually matched closed-loop mouse. Light stimulation was delivered using the
1301 pattern implemented for the matched mouse receiving closed-loop stimulation.

1302 For the projection-targeted-stimulation behavioral experiment, CD1 males were exposed to
1303 four-minute segments of light stimulation in the following sequence: yellow, blue, yellow. Mice
1304 received light stimulation for the entirety of each segment. Within each stimulation segment,
1305 an intact male C57 and a female C57 were placed in the CD1 cage for two minutes each in
1306 pseudorandom order. A subset of the mice were also exposed to a castrated male.

1307 Immediately prior to experiments, light levels were calibrated using a power meter (ThorLabs,
1308 Model No. PM100D) and delivered using a Waveform generator (Agilent Technologies, Model
1309 No. 33210A) for the open-loop experiment and projection-targeted-stimulation behavioral
1310 experiment. For the closed-loop and nonsynchronous stimulation experiments, as well as the
1311 projection-targeted-stimulation physiological experiment, the laser was activated using analog
1312 output from the Cerebus recording system.

1313 Behavioral analysis for optogenetic experiments

1314 For open-loop, closed-loop, and nonsynchronous stimulation optogenetic stimulation
1315 experiments, mice were exposed to intact male mice, castrated male mice, and female mice in
1316 pseudorandomized order. Mice were stimulated across two testing sessions and the time mice
1317 engaged in aggressive and non-aggressive behaviors was averaged across the two yellow light
1318 stimulation segments and compared to the time mice engaged in these behaviors during the
1319 two blue light stimulation segments. For the mPFC projection-targeted experiment, mice were
1320 exposed to intact male mice and female mice in pseudorandomized order. A subset of mice was
1321 also exposed to a castrated male. Mice were stimulated during a single session. The amount of
1322 time mice spent engaged in aggressive and non-aggressive behaviors were averaged across the
1323 two yellow light stimulation segments and compared to the time mice engaged in these
1324 behaviors during the blue light stimulation segment.

1325 Mediation Analysis

1326 For the Baron and Kenny approach [46] to test whether *EN-AggINH* expression mediated the
1327 behavioral effect of neurostimulation, we first established the two prerequisite relationships: (i)
1328 that neurostimulation significantly altered *EN-AggINH* expression (Figure 5d) and (ii) that
1329 neurostimulation significantly reduced aggressive behavior (Figure 6a, right). We then tested
1330 whether inclusion of *EN-AggINH* expression attenuated the direct effect of stimulation on
1331 behavior, consistent with mediation. For the closed-loop optogenetic stimulation experiment,
1332 we focused on windows of LFP data when either the blue (treatment condition) or yellow laser
1333 (control condition) was activated to match the conditions between the treatment and control
1334 as closely as possible, which is necessary for clean statistical analysis. We followed the

1335 procedures outlined in "LFP preprocessing and signal artifact removal" to preprocess the data
1336 and remove data with significant artifacts. *EN-AggINH* expression was calculated by projecting
1337 the neural features into the learned model. The remaining data after the matching and
1338 preprocessing was then fit into two logistic regression models to predict aggressive vs. non-
1339 aggressive behavior using the statsmodel package in python [69]. The first model only used a
1340 constant and whether the subject had received blue or yellow light stimulation in that time
1341 window to predict aggressive vs. non-aggressive behavior in a logistic regression for each
1342 window of data, and the second model added *EN-AggINH* network expression during that
1343 window as an additional covariate. These two models were compared by using a likelihood
1344 ratio test to evaluate whether the second model was significantly better at predicting
1345 aggressive vs. non-aggressive behavior.

1346 For the causal mediation analysis, we used the same data as described above in the classic
1347 mediation analysis. We define the treatment as blue versus yellow light stimulation (control),
1348 the mediator as *EN-AggINH* expression, and the outcome as aggressive versus non-aggressive
1349 behavior. These data were then used in the causal mediation analysis approach proposed by
1350 Kosuke, Keele, and Tingley [47] by using the statsmodels package in python [69].

1351 <u>Physiological assessment of mPFC projection-targeted stimulation</u>

1352 Non-aggressive CD1 mice were implanted to target mPFC and one of three downstream regions
1353 as described above (NAc, MeA, or LHb). Mice were then connected to a fiberoptic patch cable
1354 in their home cage. Mice were subjected to two stimulation conditions: blue light delivered
1355 either to the mPFC soma or to the terminals of mPFC afferents within the downstream region
1356 for which each mouse was implanted. Mice were then stimulated with 40 light pulses (10ms,
1357 1mW) at each location with a pseudorandom inter-pulse-interval between 8-24 seconds. Neural
1358 data was collected during stimulation as described above and subjected to a 4th order
1359 Butterworth bandpass filter from 0.5Hz to 60Hz. LFP data was then aligned relative to each light
1360 pulse and averaged across the 40 light pulses to yield a light-evoked LFP average (i.e., evoked
1361 potential) for mPFC and each downstream site. As a negative control, experiments were also
1362 performed with 40 pulses of yellow light (10ms, 1mW).

1363 We quantified the evoked potential within the 1-60ms window and determined that light
1364 stimulation (blue or yellow) induced neural activity if the lowest value within this window was
1365 lower than activity within the 100-1000ms window prior to light stimulation and the 500-
1366 1000ms window following light stimulation. This corresponds to an alpha threshold of 0.05,
1367 Bonferroni-corrected for 60 comparisons. Blue light stimulation of mPFC soma and mPFC
1368 terminals induced neural activity for all three circuits tested. No significant evoked potentials
1369 were observed for any of the circuits in response to yellow light stimulation.

1370 Statistical comparison of directionality features within *EN-AggINH*

1371 To compare the extent to which PL/IL→NAc, PL/IL→MeA circuits and PL/IL→LHb circuits were
1372 suppressed during aggression, we quantified the relative LFP spectral energy (RLE) as a measure
1373 of relative directionality for each circuit within *EN-AggINH*. This RLE is the weight of each
1374 directionality feature of *EN-AggINH* normalized to the sum of the weight of that feature across
1375 all the learned electome networks. To make a reasonable approximation to the confidence
1376 intervals on the relative weights of PL/IL→NAc, PL/IL→MeA, and PL/IL→LHb at 20 Hz, as shown
1377 in Figure 6c, we performed the following bootstrapping procedure: first, we fixed the non-
1378 relevant components of the inferred full electome model, and then found the best fit on the
1379 *EN-AggINH* parameters by bootstrapping *over animals* for 1000 repeats. While this uncertainty
1380 approach does not hold for machine learning models in general, we note that this approach is
1381 reasonable here given the mathematical structure of the machine learning model. In particular,
1382 the model loss is biconvex over the parameters with penalty terms, meaning that if the network
1383 activities are fixed the network parameters are strongly convex and the optimization will yield a
1384 single, unique optimal solution to *EN-AggINH* parameters. Hence, this approach allowed us to
1385 get reliable and reproducible estimates of uncertainty through bootstrapping. Under this
1386 approximate procedure, we found that in all repeats that the PL/IL→NAc relationship was
1387 stronger than the PL/IL→MeA, which was in turn stronger than PL/IL→LHb. This implies a p-
1388 value of <1e-4 under this approximate procedure. Thus, both the NAc- and MeA-targeted
1389 circuits are statistically stronger than the LHb-targeted circuit.

1390 Data exclusions

1391 For the open-loop stimulation experiment, we behaviorally screened 30 CD1 mice. 10 mice that
1392 showed aggressive behavior towards intact males were implanted with stimulating fibers. Two
1393 mice were removed from behavioral analysis due to poor ChR2 expression.

1394 To discover *EN-AggINH*, we behaviorally screened 45 CD1 mice and implanted 31 of these mice
1395 that showed aggression. Two of these mice were removed from further analysis because they
1396 failed to show aggression post implantation. Of the 297 total implantation sites in the
1397 remaining 29 training/validation and testing set of mice, 17 were mistargeted (~5.7% error
1398 rate). Of these mistargeted implants, 13 were within 200µm of the targeted structure. Given
1399 our prior work demonstrating high LFP spectral coherence (in the 1-55Hz frequency range)
1400 across microwires separated by 250µm in both cortical and subcortical brain regions [42], we
1401 chose to retain these 13 implants in our analysis. The other four mistargeted implants were
1402 within 350µm of the targeted structure. The most reliably mistargeted site was ventral medial
1403 hypothalamus, for which 4 animals were implanted within 200µm of the target, and 2 animals
1404 were implanted within 350 µm of the target.

1405 Machine learning analysis typically benefits from larger data sets. Thus, we concluded that
1406 maintaining a higher number of data points likely outweighed the effect of a small number of
1407 mistargeted brain regions, particularly since our LFP measures were robust to the targeting
1408 inaccuracies we observed histologically. As such, we pooled data from all 20 implanted animals
1409 to learn our initial model (the training cohort). We employed a similar strategy for our
1410 validation analysis, where an animal was only removed from the validation set if there was clear
1411 histological confirmation of mistargeting >200µm for any of the recorded regions. Specifically,
1412 presuming accurate targeting with 94.3% certainty and targeting within 200µm at a higher
1413 certainty, we included animals with missing or damaged histological slices in our analysis.
1414 However, if there was clear histological confirmation of mistargeting for any of the recorded
1415 regions (as was the case for 1 mouse), the animal was removed from the validation testing.
1416 Importantly, our validation procedure implies that the machine learning models were robust,
1417 regardless of any slight imprecision in the animals we utilized for training.

1418 For the multi-wire and silicon probe recordings, we implanted three mice that showed
1419 aggressive behavior towards intact males. We removed one mouse because it showed poor

1420 aggression after implantation when tethered. For the ESR1-Cre+/CD1 experiment, we screened
1421 13 mice across three sessions following treatment with CNO. The three testing sessions were
1422 performed over 3 weeks. Only the nine male mice that attacked females during two of the
1423 three sessions were implanted with electrodes [18]. The eight mice that showed good surgical
1424 recovery were subjected to further experiments. All of these mice were retained for behavioral
1425 and physiological analysis.

1426 For the closed-loop stimulation experiment, we screened 30 CD1 mice and implanted the
1427 eleven mice that showed aggressive behavior towards intact males. Two mice were removed
1428 from behavioral/physiological analysis due to poor ChR2 expression or electrode mistargeting.
1429 For the nonsynchronous stimulation experiment, we screened 40 CD1 mice and implanted the
1430 19 mice that showed aggressive behavior towards intact males. Five animals were removed
1431 from final analysis due to poor tissue expression of ChR2.

1432 For the mPFC-projection-targeted stimulation experiment, we screened 120 CD1 mice. The 46
1433 animals that showed high aggression were injected with virus targeting mPFC→NAc (N=12),
1434 mPFC→MeA (N=16), or mPFC→LHb (N=18). 19 mice were removed from subsequent
1435 behavioral analysis because they showed low aggression during the first yellow light stimulation
1436 segment (i.e., less than 20s of aggressive behavior towards intact males), poor expression of
1437 ChR2, or mistargeting (mPFC→NAc: N=3/12; mPFC→MeA: N=7/16; mPFC→LHb: N=9/18). We
1438 used six mice that showed low aggression for the projection-targeted physiological
1439 experiments. Of these six mice, one mouse that was infected (virus) and implanted (electrodes)
1440 to target the mPFC→LHb circuit died prior to stimulation. One mouse that was infected (virus)
1441 and implanted (electrodes) to target the mPFC→MeA circuit was excluded because of poor
1442 ChR2 expression. To ensure balance observations across the three groups, we pseudorandomly
1443 removed one mouse from the mPFC→NAc circuit group.

1444 For the LinCx-editing experiment, we screened 60 CD1 mice. 25 mice that showed aggressive
1445 behavior towards intact males were pseudorandomly assigned to the LinCx-edited (N=13) or
1446 control groups (N=12) to balance aggressive behavior and non-aggressive social behaviors
1447 between the groups. Mice were behaviorally screened a second time to ensure the robustness
1448 of these group assignments. Mice were then injected with virus and subjected to behavioral

1449 testing at 2, 5, and 8 weeks after surgery. One control mouse and one LinCx-edited mouse were
1450 removed from the final behavioral analysis because they showed low aggression (i.e., less than
1451 20s of aggressive behavior towards intact males) during the second presurgical screening. Two
1452 additional LinCx-edited mice were removed from behavioral analysis: one mouse showed poor
1453 $Cx34.7_{M1}$ expression in the mPFC, and the other showed bilateral spread of $Cx35_{M1}$ to mPFC.
1454 We retained one control mouse that showed poor expression of mCherry in NAc in our analysis
1455 since we did not anticipate that mCherry expression in the NAc would alter social behavior.
1456 Importantly, presurgical aggressive and non-aggressive social behavior remained balanced
1457 between the LinCx-edited and control groups with these four mice removed (see Fig. 6h, N=10
1458 LinCx-edited mice and N=11 control mice).

1459 Statistics

1460 GraphPad Prism and MATLAB were used for statistical analyses of behavior and network
1461 activity. Parametric testing was used for comparison of behavioral measures and non-
1462 parametric tests were used for analysis comparing network activity. Paired t-tests were used for
1463 comparing within-subject behavioral response to optogenetic stimulation (two-tailed) or CNO
1464 application (one-tailed). We used a false discovery rate corrected alpha threshold for the three
1465 behavioral comparisons performed in each condition (aggressive behavior towards an intact
1466 male, non-aggressive interactions with an intact male, and non-aggressive interactions with a
1467 female). This correction was achieved using the Benjamini-Hochberg procedure. Because
1468 network activity is not normally distributed, we used a Friedman's tests to compare repeated
1469 measures of network activity. Post-hoc testing was performed with two-tailed Wilcoxon signed-
1470 rank tests to compare within-subject activity, and two-tailed rank sum testing was used to
1471 perform between-subject network activity. This analysis approach was used to compare
1472 network activity responses to optogenetic stimulation, stimulus mouse exposure and
1473 interaction, and CNO injection. We used directional tests for our validation experiments
1474 because our initial dCSFA-NMF model mapped low *EN-AggINH* activity to aggression. Thus, we
1475 used one-tailed tests to determine whether manipulations known to increase aggression would
1476 also decrease network activity, and whether manipulations known to decrease aggression
1477 would also increase network activity. We used a Spearman rank to determine the relationship

between network activity and aggressive behavior, and a Z test was used to determine if mice showed a greater proportion of cells that were coupled to network activity than would be expected by chance. We used a mixed effects model ANOVA to determine the impact of chronic mPFC→NAc LinCx-editing on behavior, and we used a false discovery rate-corrected alpha threshold to account for the three behavioral conditions. Data is presented as mean ± standard error of measurement, throughout the paper, unless otherwise specified.

Code and data availability

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

References

1 Blanchard, D. C. & Blanchard, R. J. What can animal aggression research tell us about human aggression? *Horm Behav* **44**, 171-177 (2003). https://doi.org/10.1016/s0018-506x(03)00133-8

2 Lorenz, K. *On aggression.*, (MJF Books, 1963).

3 Lee, S. C., Yamamoto, T. & Ueki, S. Characteristics of aggressive behavior induced by nucleus accumbens septi lesions in rats. *Behav Neural Biol* **37**, 237-245 (1983). https://doi.org/10.1016/s0163-1047(83)91254-2

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

5 Bak, M. *et al.* The pharmacological management of agitated and aggressive behaviour: A systematic review and meta-analysis. *Eur Psychiatry* **57**, 78-100 (2019). https://doi.org/10.1016/j.eurpsy.2019.01.014

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

7 Chen, P. & Hong, W. Neural Circuit Mechanisms of Social Behavior. *Neuron* **98**, 16-30 (2018). https://doi.org/10.1016/j.neuron.2018.02.026

1508 8 Mague, S. D. *et al.* Brain-wide electrical dynamics encode individual appetitive social
1509 behavior. *Neuron* **110**, 1728-1741 e1727 (2022).
1510 https://doi.org/10.1016/j.neuron.2022.02.016

1511 9 Wong, L. C. *et al.* Effective Modulation of Male Aggression through Lateral Septum to
1512 Medial Hypothalamus Projection. *Curr Biol* **26**, 593-604 (2016).
1513 https://doi.org/10.1016/j.cub.2015.12.065

1514 10 Brayley, K. N. & Albert, D. J. Suppression of VMH-lesion-induced reactivity and
1515 aggressiveness in the rat by stimulation of lateral septum, but not medial septum or
1516 cingulate cortex. *J Comp Physiol Psychol* **91**, 290-299 (1977).

1517 11 Couppis, M. H. & Kennedy, C. H. The rewarding effect of aggression is reduced by
1518 nucleus accumbens dopamine receptor antagonism in mice. *Psychopharmacology (Berl)*
1519 **197**, 449-456 (2008). https://doi.org/10.1007/s00213-007-1054-y

1520 12 Flanigan, M., Aleyasin, H., Takahashi, A., Golden, S. A. & Russo, S. J. An emerging role for
1521 the lateral habenula in aggressive behavior. *Pharmacol Biochem Behav* **162**, 79-86
1522 (2017). https://doi.org/10.1016/j.pbb.2017.05.003

1523 13 Golden, S. A. *et al.* Basal forebrain projections to the lateral habenula modulate
1524 aggression reward. *Nature* **534**, 688-692 (2016). https://doi.org/10.1038/nature18601

1525 14 Falkner, A. L., Dollar, P., Perona, P., Anderson, D. J. & Lin, D. Decoding ventromedial
1526 hypothalamic neural activity during male mouse aggression. *J Neurosci* **34**, 5971-5984
1527 (2014). https://doi.org/10.1523/JNEUROSCI.5109-13.2014

1528 15 Falkner, A. L., Grosenick, L., Davidson, T. J., Deisseroth, K. & Lin, D. Hypothalamic control
1529 of male aggression-seeking behavior. *Nat Neurosci* **19**, 596-604 (2016).
1530 https://doi.org/10.1038/nn.4264

1531 16 Hashikawa, Y., Hashikawa, K., Falkner, A. L. & Lin, D. Ventromedial Hypothalamus and
1532 the Generation of Aggression. *Front Syst Neurosci* **11**, 94 (2017).
1533 https://doi.org/10.3389/fnsys.2017.00094

1534 17 Lin, D. *et al.* Functional identification of an aggression locus in the mouse hypothalamus.
1535 *Nature* **470**, 221-226 (2011). https://doi.org/10.1038/nature09736

1536 18 Takahashi, A. *et al.* Establishment of a repeated social defeat stress model in female
1537 mice. *Scientific reports* **7**, 12838 (2017). https://doi.org/10.1038/s41598-017-12811-8

1538 19 Shibata, S., Yamamoto, T. Y. & Ueki, S. Differential effects of medial, central and
1539 basolateral amygdaloid lesions on four models of experimentally-induced aggression in
1540 rats. *Physiol Behav* **28**, 289-294 (1982).

1541 20 Dambacher, F. *et al.* Reducing proactive aggression through non-invasive brain
1542 stimulation. *Soc Cogn Affect Neurosci* **10**, 1303-1309 (2015).
1543 https://doi.org/10.1093/scan/nsv018

1544 21 Choy, O., Raine, A. & Hamilton, R. H. Stimulation of the Prefrontal Cortex Reduces
1545 Intentions to Commit Aggression: A Randomized, Double-Blind, Placebo-Controlled,
1546 Stratified, Parallel-Group Trial. *J Neurosci* **38**, 6505-6512 (2018).
1547 https://doi.org/10.1523/JNEUROSCI.3317-17.2018

1548 22 Takahashi, A., Nagayasu, K., Nishitani, N., Kaneko, S. & Koide, T. Control of intermale
1549 aggression by medial prefrontal cortex activation in the mouse. *PLoS One* **9**, e94657
1550 (2014). https://doi.org/10.1371/journal.pone.0094657

1551 23 Mugford, R. A. & Nowell, N. W. Pheromones and their effect on aggression in mice.
1552 *Nature* **226**, 967-968 (1970). https://doi.org/10.1038/226967a0

1553 24 Novotny, M., Harvey, S., Jemiolo, B. & Alberts, J. Synthetic pheromones that promote
1554 inter-male aggression in mice. *Proc Natl Acad Sci U S A* **82**, 2059-2061 (1985).
1555 https://doi.org/10.1073/pnas.82.7.2059

1556 25 Lischinsky, J. E. & Lin, D. Neural mechanisms of aggression across species. *Nat Neurosci*
1557 **23**, 1317-1328 (2020). https://doi.org/10.1038/s41593-020-00715-2

1558 26 Nair, A. *et al.* An approximate line attractor in the hypothalamus encodes an aggressive
1559 state. *Cell* **186**, 178-193 e115 (2023). https://doi.org/10.1016/j.cell.2022.11.027

1560 27 de la Zerda, S. H. *et al.* Social recognition in laboratory mice requires integration of
1561 behaviorally-induced somatosensory, auditory and olfactory cues.
1562 *Psychoneuroendocrinology* **143**, 105859 (2022).
1563 https://doi.org/10.1016/j.psyneuen.2022.105859

1564 28 APA. *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR*. (American
1565 Psychiatric Association, 2000).

1566 29 Levy, D. R. *et al.* Dynamics of social representation in the mouse prefrontal cortex. *Nat*
1567 *Neurosci* **22**, 2013-2022 (2019). https://doi.org/10.1038/s41593-019-0531-z

1568 30 Murugan, M. *et al.* Combined Social and Spatial Coding in a Descending Projection from
1569 the Prefrontal Cortex. *Cell* **171**, 1663-1677 e1616 (2017).
1570 https://doi.org/10.1016/j.cell.2017.11.002

1571 31 Padilla-Coreano, N. *et al.* Cortical ensembles orchestrate social competition through
1572 hypothalamic outputs. *Nature* **603**, 667-671 (2022). https://doi.org/10.1038/s41586-
1573 022-04507-5

1574 32 Molero-Chamizo, A., Martin Riquel, R., Moriana, J. A., Nitsche, M. A. & Rivera-Urbina, G.
1575 N. Bilateral Prefrontal Cortex Anodal tDCS Effects on Self-reported Aggressiveness in
1576 Imprisoned Violent Offenders. *Neuroscience* **397**, 31-40 (2019).
1577 https://doi.org/10.1016/j.neuroscience.2018.11.018

1578 33 He, J. *et al.* Effect of transcranial direct current stimulation over the left dorsolateral
1579 prefrontal cortex on the aggressive behavior in methamphetamine addicts. *Journal of*
1580 *psychiatric research* **164**, 364-371 (2023).
1581 https://doi.org/10.1016/j.jpsychires.2023.06.038

1582 34 Miczek, K. A., Maxson, S. C., Fish, E. W. & Faccidomo, S. Aggressive behavioral
1583 phenotypes in mice. *Behav Brain Res* **125**, 167-181 (2001).
1584 https://doi.org/10.1016/s0166-4328(01)00298-4

1585 35 Coccaro, E. F., McCloskey, M. S., Fitzgerald, D. A. & Phan, K. L. Amygdala and
1586 orbitofrontal reactivity to social threat in individuals with impulsive aggression. *Biol*
1587 *Psychiatry* **62**, 168-178 (2007). https://doi.org/10.1016/j.biopsych.2006.08.024

1588 36 Siegel, A., Edinger, H. & Koo, A. Suppression of attack behavior in the cat by the
1589 prefrontal cortex: role of the mediodorsal thalamic nucleus. *Brain research* **127**, 185-190
1590 (1977). https://doi.org/10.1016/0006-8993(77)90392-4

1591 37 Chang, C. H. & Gean, P. W. The Ventral Hippocampus Controls Stress-Provoked
1592 Impulsive Aggression through the Ventromedial Hypothalamus in Post-Weaning Social
1593 Isolation Mice. *Cell Rep* **28**, 1195-1205 e1193 (2019).
1594 https://doi.org/10.1016/j.celrep.2019.07.005

1595 38 Talbot, A., Dunson, D., Dzirasa, K. & Carlson, D. Supervised Autoencoders Learn Robust
1596 Joint Factor Models ofNeural Activity. *arxiv* (2020).

1597 39 Vu, M. T. *et al.* A Shared Vision for Machine Learning in Neuroscience. *J Neurosci* (2018).
1598 https://doi.org/10.1523/JNEUROSCI.0508-17.2018

1599 40 Chen, A. X. *et al.* Specific Hypothalamic Neurons Required for Sensing Conspecific Male
1600 Cues Relevant to Inter-male Aggression. *Neuron* **108**, 763-774 e766 (2020).
1601 https://doi.org/10.1016/j.neuron.2020.08.025

1602 41 Lee, D. L. & Wilson, J. L. Urine from Sexually Mature Intact Male Mice Contributes to
1603 Increased Cardiovascular Responses during Free-Roaming and Restrained Conditions.
1604 *ISRN Vet Sci* **2012** (2012). https://doi.org/10.5402/2012/185461

1605 42 Hultman, R. *et al.* Brain-wide Electrical Spatiotemporal Dynamics Encode Depression
1606 Vulnerability. *Cell* **173**, 166-180 e114 (2018). https://doi.org/10.1016/j.cell.2018.02.012

1607 43 Carlson, D. E., Chavarriaga, R., Liu, Y., Lotte, F. & Lu, B. L. The NERVE-ML (neural
1608 engineering reproducibility and validity essentials for machine learning) checklist:

ensuring machine learning advances neural engineering(). *J Neural Eng* **22** (2025). https://doi.org/10.1088/1741-2552/adbfbd

44 Deonaraine, K. K. *et al.* Sex-specific peripheral and central responses to stress-induced depression and treatment in a mouse model. *J Neurosci Res* **98**, 2541-2553 (2020). https://doi.org/10.1002/jnr.24724

45 Li, L. *et al.* Social trauma engages lateral septum circuitry to occlude social reward. *Nature* (2022). https://doi.org/10.1038/s41586-022-05484-5

46 Baron, R. M. & Kenny, D. A. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* **51**, 1173-1182 (1986). https://doi.org/10.1037//0022-3514.51.6.1173

47 Imai, K., Keele, L. & Tingley, D. A general approach to causal mediation analysis. *Psychol Methods* **15**, 309-334 (2010). https://doi.org/10.1037/a0020761

48 Ransey, E. *et al.* Long-term editing of brain circuits in mice using an engineered electrical synapse. *BioRxiv* (2025). https://doi.org/https://doi.org/10.1101/2025.03.25.645291

49 Ransey, E. *et al.* Long-term precision editing of neural circuits in mammals using engineered gap junction hemichannels. *BioRxiv* (2023). https://doi.org/https://doi.org/10.1101/2021.08.24.457429

50 Walder-Christensen, K. K. *et al.* Sleep-wake states are encoded across emotion-regulation regions of the mouse brain. *eNeuro* (2025). https://doi.org/10.1523/ENEURO.0291-25.2025

51 Gervasoni, D. *et al.* Global forebrain dynamics predict rat behavioral states and their transitions. *J. Neurosci.* **24**, 11137-11147 (2004).

52 Anikeeva, P. *et al.* Optetrode: a multichannel readout for optogenetic control in freely moving mice. *Nat Neurosci* (2011). https://doi.org/10.1038/nn.2992

53 Kumar, S. *et al.* Cortical Control of Affective Networks. *J Neurosci* **33**, 1116 –1129 (2013).

54 Jendryka, M. *et al.* Pharmacokinetic and pharmacodynamic actions of clozapine-N-oxide, clozapine, and compound 21 in DREADD-based chemogenetics in mice. *Scientific reports* **9**, 4522 (2019). https://doi.org/10.1038/s41598-019-41088-2

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

56 Kwiatkowski, C. C. *et al.* Quantitative standardization of resident mouse behavior for studies of aggression and social defeat. *Neuropsychopharmacology* **46**, 1584-1593 (2021). https://doi.org/10.1038/s41386-021-01018-1

57 Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**, 1281-1289 (2018). https://doi.org/10.1038/s41593-018-0209-y

58 Nath, T. *et al.* Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat Protoc* **14**, 2152-2176 (2019). https://doi.org/10.1038/s41596-019-0176-0

59 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). https://doi.org/10.1101/2020.04.19.049452

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

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

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

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

64 Talbot, A., Dunson, D., Dzirasa, K. & Carlson, D. Estimating a brain network predictive of stress and genotype with supervised autoencoders. *J R Stat Soc Ser C Appl Stat* **72**, 912-936 (2023). https://doi.org/10.1093/jrsssc/qlad035

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

66 Pachitariu, M., Sridhar, S., Pennington, J. & Stringer, C. Spike sorting with Kilosort4. *Nature methods* **21**, 914-921 (2024). https://doi.org/10.1038/s41592-024-02232-7

67 Hill, D. N., Mehta, S. B. & Kleinfeld, D. Quality metrics to accompany spike sorting of extracellular signals. *J Neurosci* **31**, 8699-8705 (2011). https://doi.org/10.1523/JNEUROSCI.0971-11.2011

68 Institute, A. <https://allensdk.readthedocs.io/en/latest/_static/examples/nb/ecephys_quality_metrics.html> (

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

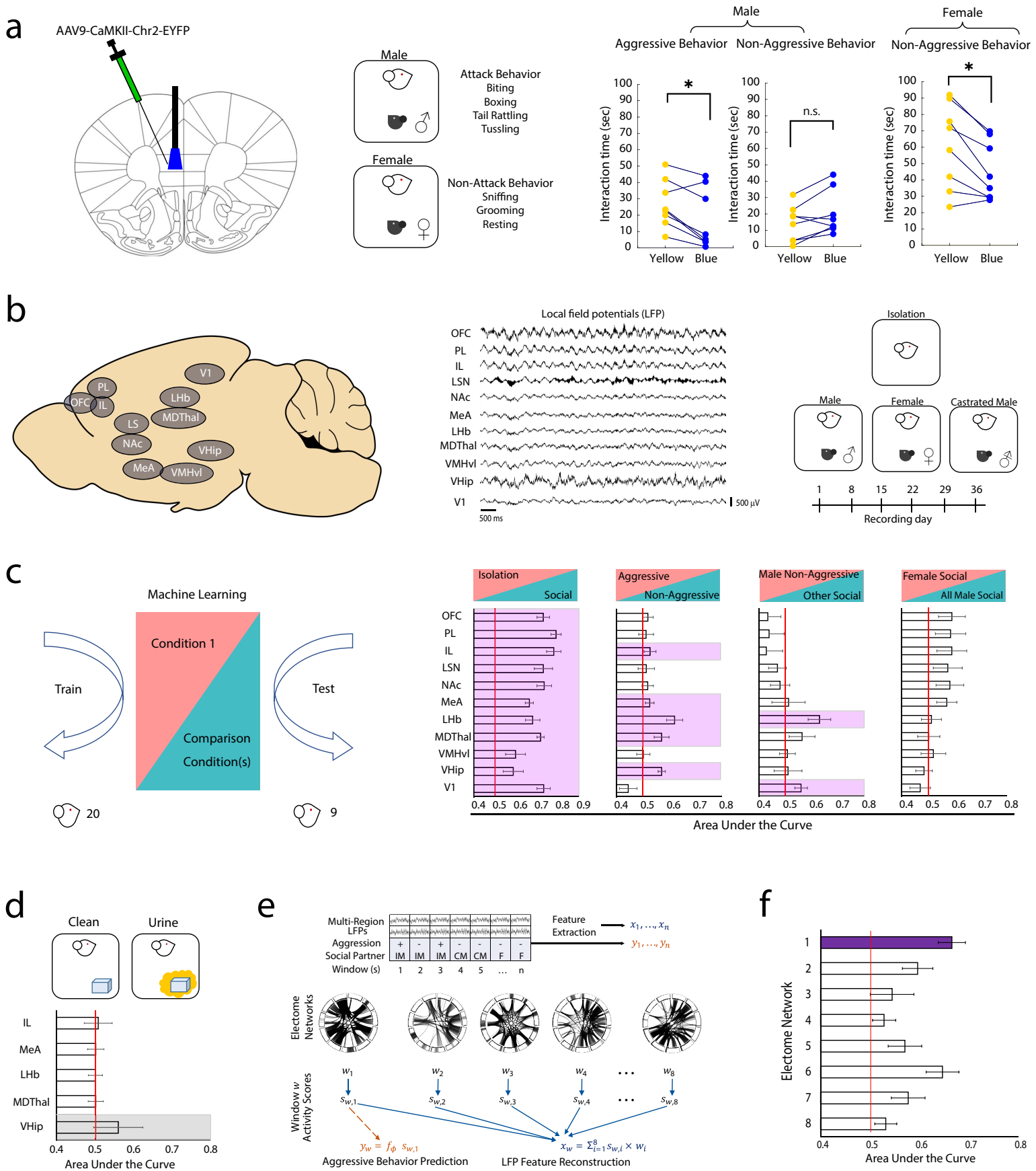
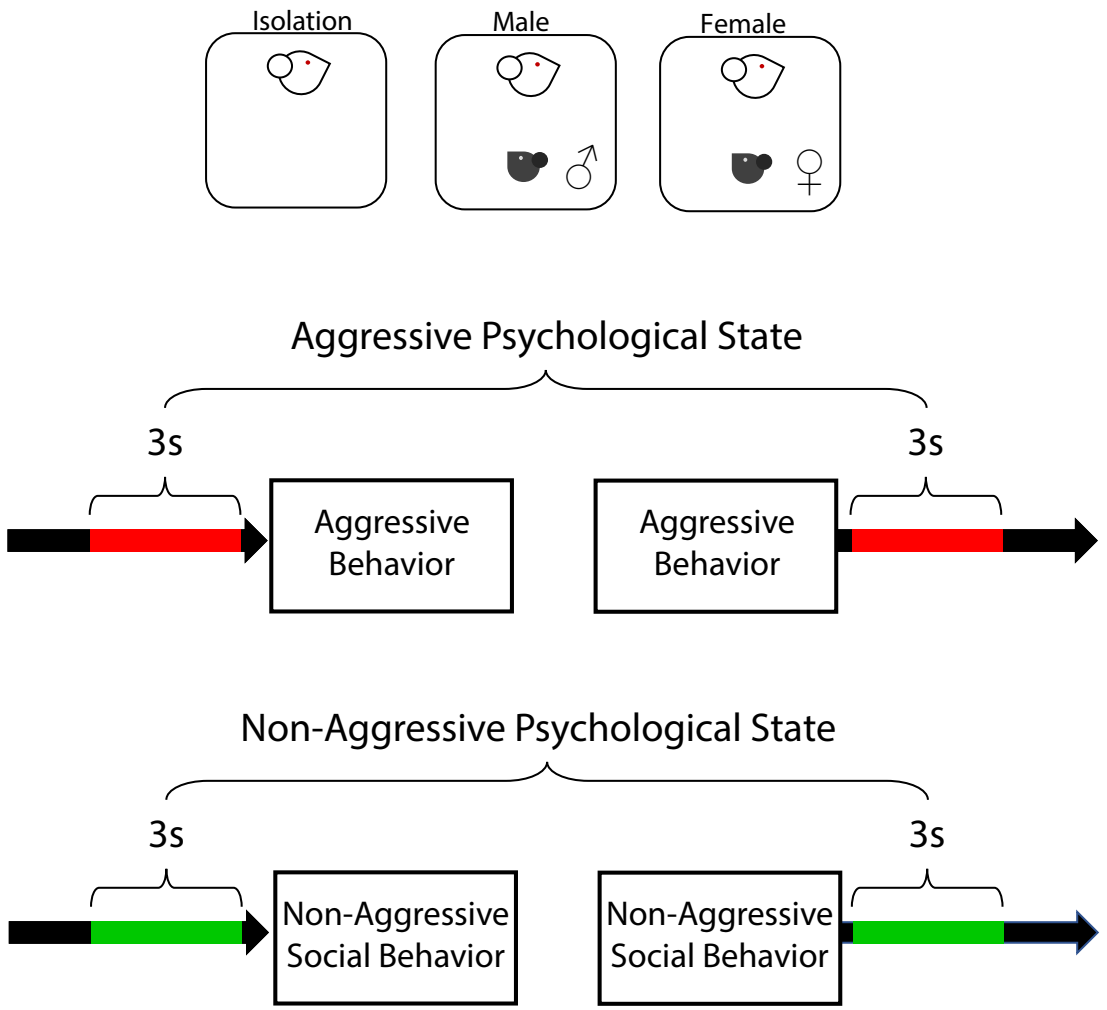

a
AAV9-CaMKII-Chr2-EYFP
Male
Attack Behavior
Biting
Boxing
Tail Rattling
Tussling
Female
Non-Attack Behavior
Sniffing
Grooming
Resting
Male
Aggressive Behavior
Non-Aggressive Behavior
Female
Non-Aggressive Behavior
Interaction time (sec)
Yellow
Blue
n.s.
b
V1
PL
OFC
IL
LHb
MDThal
LS
NAc
VHip
MeA
VMHvl
Local field potentials (LFP)
OFC
PL
IL
LSN
NAc
MeA
LHb
MDThal
VMHvl
VHip
V1
500 ms
500 µV
Isolation
Male
Female
Castrated Male
1
8
15
22
29
36
Recording day
c
Machine Learning
Condition 1
Comparison Condition(s)
Train
Test
20
9
Isolation
Social
Aggressive
Non-Aggressive
Male Non-Aggressive
Other Social
Female Social
All Male Social
Area Under the Curve
d
Clean
Urine
IL
MeA
LHb
MDThal
VHip
Area Under the Curve
e
Multi-Region LFPs
Aggression
Social Partner
Window (s)
Feature Extraction
Electome Networks
Window w Activity Scores
Aggressive Behavior Prediction
LFP Feature Reconstruction
f
Electome Network
Area Under the Curve

Figure 1

a


Isolation
Male
Female
Aggressive Psychological State
3s
3s
Aggressive Behavior
Aggressive Behavior
Non-Aggressive Psychological State
3s
3s
Non-Aggressive Social Behavior
Non-Aggressive Social Behavior

b

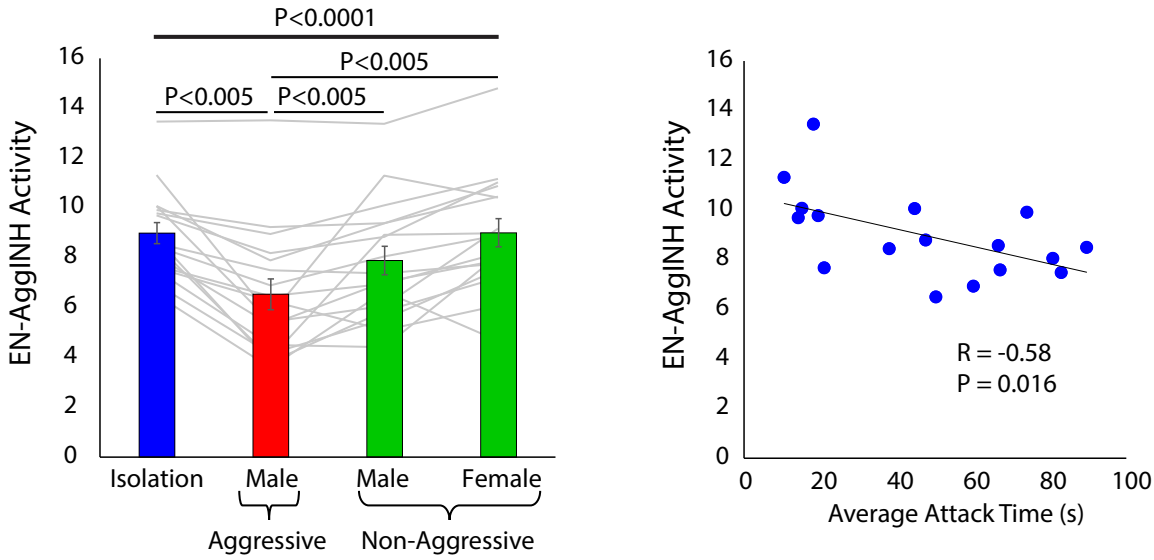

P<0.0001
P<0.005
P<0.005
P<0.005
EN-AggINH Activity
16
14
12
10
8
6
4
2
0
Isolation
Male
Male
Female
Aggressive
Non-Aggressive
EN-AggINH Activity
R = -0.58
P = 0.016
0
20
40
60
80
100
Average Attack Time (s)

Figure 2

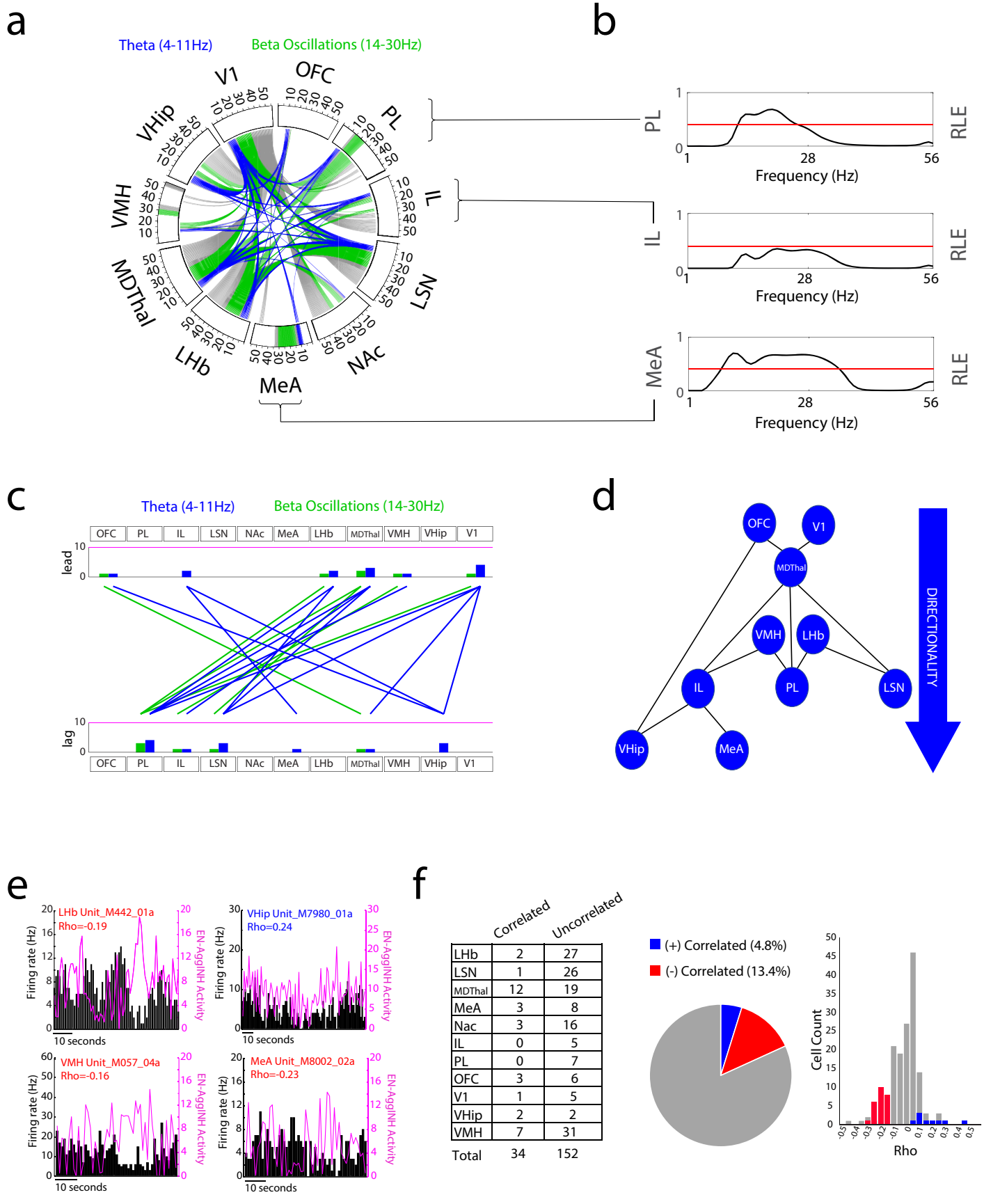

a
Theta (4-11Hz)
Beta Oscillations (14-30Hz)
V1
OFC
PL
IL
LSN
NAc
MeA
LHb
MDThal
VMH
VHip
b
PL
IL
MeA
RLE
Frequency (Hz)
c
Theta (4-11Hz)
Beta Oscillations (14-30Hz)
OFC PL IL LSN NAc MeA LHb MDThal VMH VHip V1
lead
lag
d
OFC
V1
MDThal
VMH
LHb
IL
PL
LSN
VHip
MeA
DIRECTIONALITY
e
LHb Unit_M442_01a
Rho=-0.19
VHip Unit_M7980_01a
Rho=0.24
VMH Unit_M057_04a
Rho=-0.16
MeA Unit_M8002_02a
Rho=-0.23
Firing rate (Hz)
EN-AggINH Activity
10 seconds
f
Correlated
Uncorrelated
LHb 2 27
LSN 1 26
MDThal 12 19
MeA 3 8
Nac 3 16
IL 0 5
PL 0 7
OFC 3 6
V1 1 5
VHip 2 2
VMH 7 31
Total 34 152
(+) Correlated (4.8%)
(-) Correlated (13.4%)
Cell Count
Rho

Figure 3

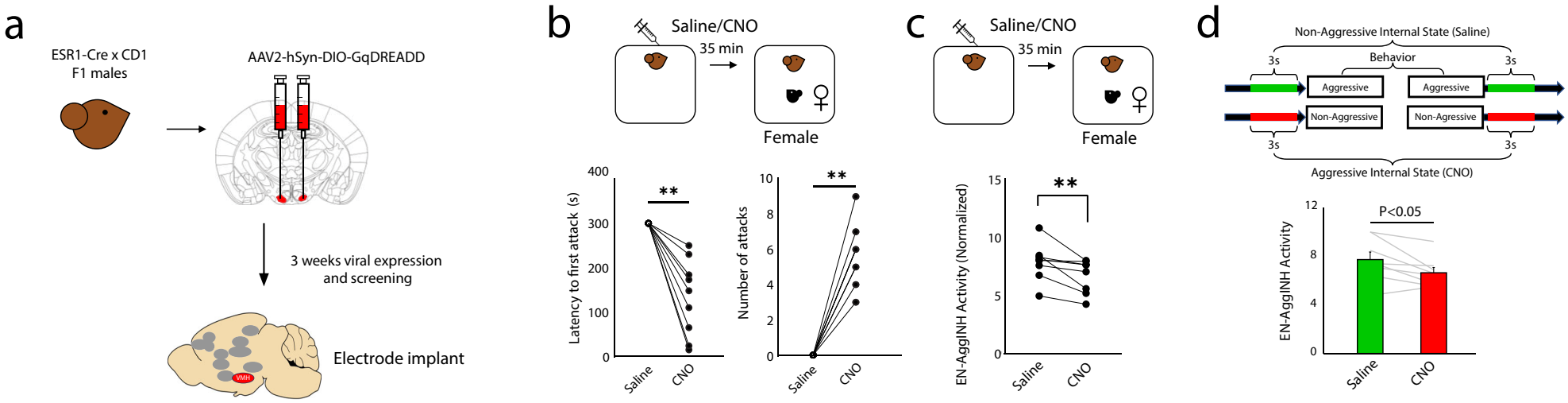

a
ESR1-Cre x CD1
F1 males
AAV2-hSyn-DIO-GqDREADD
3 weeks viral expression
and screening
Electrode implant
VMH
b
Saline/CNO
35 min
Female
Latency to first attack (s)
400
300
200
100
0
**
Saline
CNO
Number of attacks
10
8
6
4
2
0
**
Saline
CNO
c
Saline/CNO
35 min
Female
EN-AggINH Activity (Normalized)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

Figure 4

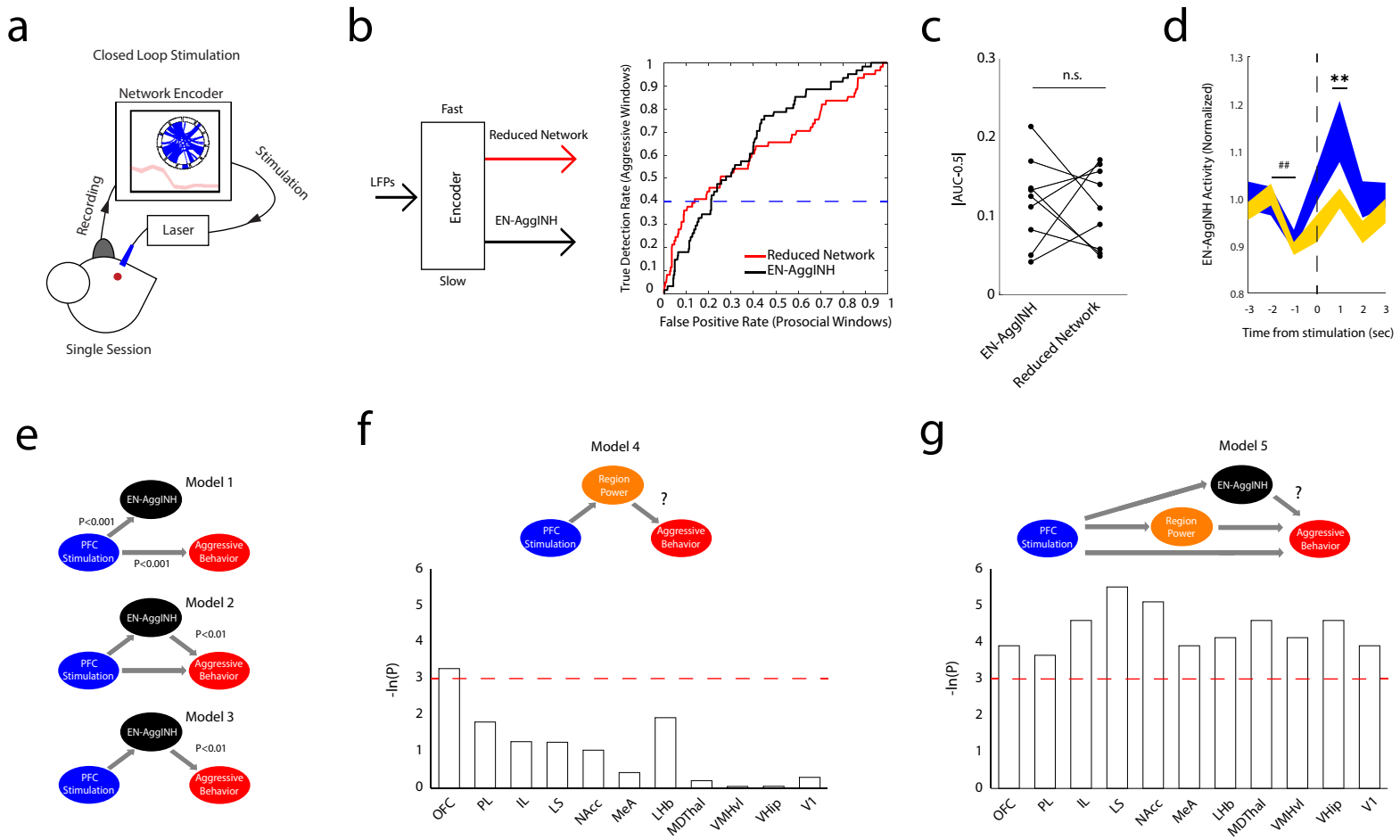

a
Closed Loop Stimulation
Network Encoder
Recording
Stimulation
Laser
Single Session
b
Fast
LFPs
Encoder
Reduced Network
EN-AggINH
Slow
True Detection Rate (Aggressive Windows)
False Positive Rate (Prosocial Windows)
Reduced Network
EN-AggINH
c
n.s.
|AUC-0.5|
EN-AggINH
Reduced Network
d
EN-AggINH Activity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

Figure 5

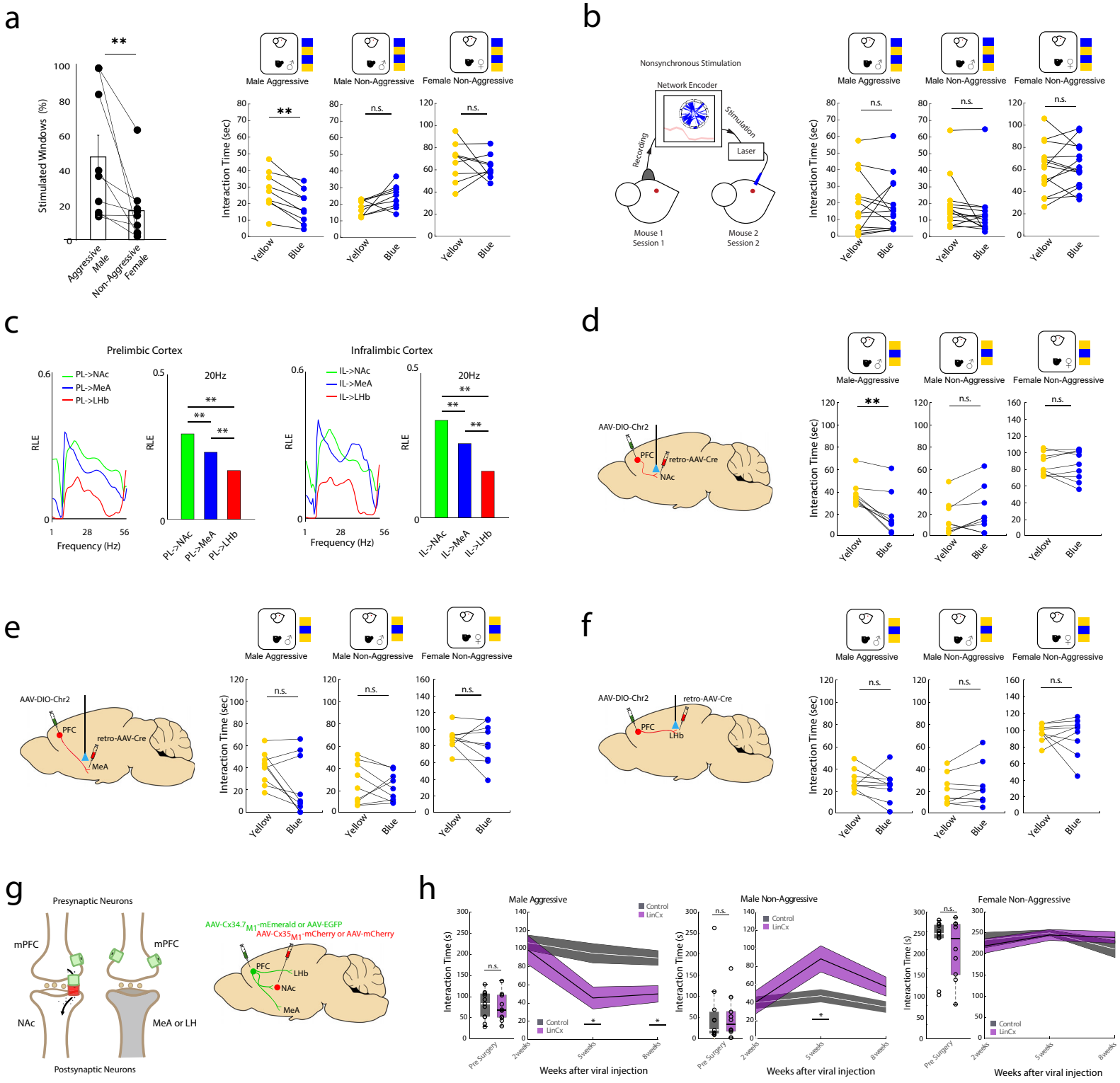

a
Stimulated Windows (%)
Aggressive Male
Non-Aggressive Female
Male Aggressive
Male Non-Aggressive
Female Non-Aggressive
Interaction Time (sec)
Yellow
Blue
n.s.
b
Nonsynchronous Stimulation
Network Encoder
Recording
Stimulation
Laser
Mouse 1
Session 1
Mouse 2
Session 2
c
Prelimbic Cortex
Infralimbic Cortex
PL->NAc
PL->MeA
PL->LHb
IL->NAc
IL->MeA
IL->LHb
20Hz
RLE
Frequency (Hz)
d
AAV-DIO-Chr2
PFC
retro-AAV-Cre
NAc
e
MeA
f
LHb
g
Presynaptic Neurons
mPFC
NAc
MeA or LH
Postsynaptic Neurons
AAV-Cx34.7M1-mEmerald or AAV-EGFP
AAV-Cx35M1-mCherry or AAV-mCherry
h
Male Aggressive
Male Non-Aggressive
Female Non-Aggressive
Control
LinCx
Interaction Time (s)
Pre Surgery
2weeks
5weeks
8weeks
Weeks after viral injection

Figure 6