Rage against the machine: advancing aggression ethology through machine learning

Nastacia L. Goodwin

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2024

Reading Committee:
Sam Golden, Chair
Susan Ferguson
Larry Zweifel

Program Authorized to Offer Degree:
Neuroscience

University of Washington

Abstract

Rage against the machine: advancing aggression ethology through machine learning

Nastacia L. Goodwin

Chair of the Supervisory Committee:
Sam Golden
Biological Structure

Aggression is a highly conserved behavior and exists along a spectrum from adaptive to maladaptive. Adaptive aggression can serve to protect mates, territory, and resources. Maladaptive aggression, however, can present as escalated and uncontrolled, and can occur comorbid with neuropsychiatric disorders including autism spectrum disorders, post-traumatic stress disorder, and intermittent explosive disorder. Inappropriate aggression seeking is detrimental to both individuals and society, and current treatment options are largely ineffective, or associated with significant side effects (Coccaro et al. 2009; Carlson et al. 2010; Frogley et al. 2012; Khushu and Powney 2016). In the clinical literature, aggression is typically delineated into instrumental, reactive (fight or flight), and appetitive (rewarding) phenotypes. Preclinically, there is a long history of research involving reactive aggression, but a much smaller body of work only in males examining the neurobiology of appetitive aggression. The goal of this dissertation was to further develop preclinical models of appetitive aggression in mice by understanding the different behavioral and whole-brain activation patterns between the sexes, and by directly comparing appetitive and reactive aggression phenotypes. A significant portion of my work in this arena has involved developing a machine learning based platform for high throughput and consistent scoring of aggression behaviors - Simple Behavioral Analysis (SimBA). Importantly, I posit that machine learning based behavioral detection paired with artificial intelligence explainability techniques allows users to objectively quantify and share behavioral classifiers in an RRID-like fashion. Using this platform, I have discovered that while both males and females

exhibit reactive aggression, males but not females show appetitive aggression. I examined the neural correlates of this behavioral sex difference using whole-brain c-fos activity mapping, identifying a potential network inhibiting appetitive aggression in females. In males, I further identified the lateral septum as a potential locus of differential control of reactive and appetitive aggression. Ultimately, this dissertation indicates that reactive and appetitive aggression are neurally dissociable processes, with an inhibitory network in females gating appetitive aggression.

Table of Contents

Table of Figures:

Table of Tables:

Gratitudes

Thank you to Annaliese, who knew that I was a scientist before I ever realized it was an option.

Thank you to Sam, who gave me the reins and let me run.

Thank you to my dad, Denise, Ashley, Carly, Taylor, and my grandparents for always being there when I needed it, and never making me feel guilty when I was too busy to be around.

Thank you to my committee members, who were patient when I didn't understand what they were asking.

Thank you to Carlee, Eric, Kevin, Yi, and Jovana for being great lab mates and better friends.

Thank you to my cohort, who has always been there.

Thank you to Sierra, for everything.

Thank you to Katie, Nikki, Allison, and Meredith for keeping me sane and full of laughter.

Thank you to the Smith College donors who funded my undergraduate education.

Thank you to Jordan, for being excited about my fiber photometry data when I wanted to throw it in the trash.

Thank you to Rocket and Jackson, for being such good dogs.

Thank you to Zazu, for being kind of a bad cat, but still very loveable.

Most of all, thank you to my army of undergraduates. I am so proud of every one of you, and so grateful for your help. This dissertation truly would not have been possible if not for all of you dedicating your time and energy to volunteering.

Chapter 1 : Introduction

Aggression is an adaptive and highly conserved behavior which can serve to protect mates, territory, and resources. Maladaptive aggression, however, can present as a symptom of neuropsychiatric disorders including autism spectrum disorders, post-traumatic stress disorder, and intermittent explosive disorder. Inappropriate aggression seeking is detrimental to both individuals and society, and current treatment options are largely ineffective, or associated with significant side effects (Coccaro et al. 2009; Carlson et al. 2010; Frogley et al. 2012; Khushu and Powney 2016). In the clinical literature, aggression is typically delineated into instrumental (aggressive behavior meant to achieve a specific goal), reactive (fight or flight), and appetitive (rewarding) phenotypes. Preclinically, there is a long history of research involving reactive aggression, but a much smaller body of work only in males examining the neurobiology of appetitive aggression. The goal of this dissertation was to further develop preclinical models of appetitive aggression in mice by understanding the different behavioral and whole-brain activation patterns between the sexes, and by directly comparing appetitive and reactive aggression phenotypes. An integral component of this work has been developing Simple Behavioral Analysis (SimBA) for automated behavioral detection in multi- animal videos.

Machine learning solutions to behavioral annotation

Complex social behavior is notoriously difficult to hand score due to observer drift, a lack of clear ethological definitions, and the rapidity with which these behavioral sequences take place. A single aggression bout in rodents can include bites, kicks, lunges, lateral threats, tail rattles, and male-male mounting within a matter of seconds. Furthermore, many researchers are beginning to move away from single-animal measures into dyadic social behavior, examining the interplay between both aggressor and subordinate. For example, species typical mouse behavior includes decreases in aggression when a subordinate shows submission behavior, and one measure of maladaptively escalated aggression can be continued attacks despite subordination. The expertise and time required to hand-annotate aggression videos severely limits throughput, and in many occasions reduces outcome measures to simple latency to attack.

The recent introduction of reliable open-source multi-animal pose tracking in video such as SLEAP (Pereira et al. 2022) and DeepLabCut (DLC, (Mathis et al. 2018a)) has revolutionized

our ability to use machine learning techniques for behavioral detection. These programs detect frame-by-frame cartesian coordinates of animal body parts of interest, which can then be fed into various machine learning platforms to identify or detect behaviors of interest, as discussed extensively in Chapters 2 and 3. We created SimBA as an open-source program which takes in these frame-by-frame body part coordinates and calculates hundreds of features between them (distance between ears, etc.) and feeds these values into supervised random forest classifiers. These classifiers, which users can construct or share across labs, then provide a frame-by-frame probability of behaviors of interest occurring. The creation of this program has allowed me to rapidly analyze millions of frames of video for multiple behaviors of interest for both members of my dyads. Furthermore, we have included easy access to post-hoc machine learning explainability tools in SimBA to allow users to understand why their models are producing particular results.

Explainable machine learning

Explainable artificial intelligence is a burgeoning field concerned with understanding how and why machine learning classifiers produce their results. Many machine learning tools currently in use are so called 'black box' models, which move from data input to answer output with little user understanding of the processes occurring in between. While there are a plethora of inherently explainable models available for use, their adoption has been stymied by a fallacy in the artificial intelligence field that there is an inherent trade-off between explainability and performance (Rudin 2019). The development of post-hoc explainability tools has allowed more users to gain insight into how specific feature values impact their model outcome, typically via some process of removing input features and examining the impact on the model output.

Gaining this understanding of feature contributions via in-built or post-hoc explainability methods has several advantages for behavioral neuroscience, including the standardization of behavioral definitions, the quantitative description of specialized behavioral studies, and the explainability of black-box models. These concepts are discussed at length in Chapter 3. Shapley Additive Explanations (SHAP) scores are a post-hoc game-theory based method for understanding the contribution of particular features to the final outcome probability of a behavior occurring. A major advantage of SHAP scores is their additivity axiom, allowing users to bin features into biologically relevant bins to understand the contribution of these larger feature categories. This feature binning is vital to the current state of computational

neuroethology, as it is extremely common for labs to use different pose estimation techniques and thereby feature set sizes, prohibiting the one-to-one comparison of raw features. Importantly, SHAP scores can be calculated for any behavioral classifier and reported as a quantifiable metric of the phenomenon being measured – rather than reporting 1-2 sentences in the methods as is the current standard (i.e. "attacks consist of physically antagonistic bites and kicks by the resident"). Throughout my experimental work, I have used SHAP scores to understand how animals differentially express attack and other behaviors across sex, environment, and aggression phenotype, as discussed in the chapters below.

Table 1.1. Aggression phenotypes and associated assays.

Aggression phenotype	Assays (*developed by NLG)
Reactive	• Home cage resident intruder tests • Chronic social defeat assays • Non-contingent operant administration*
Appetitive	• Conditioned place preference • T-maze, shock grids • Contingent operant self-administration • Sensory contact assays

Using these tools in conjunction with whole-brain c-fos mapping, I was then able to identify networks and regions differentially associated with appetitive aggression across the sexes. Reactive aggression phenotypes in mice are typically measured in non-volitional assays such as the home-cage resident intruder test, in which a stimulus mouse is placed into the test animal's cage to freely interact. Appetitive aggression has typically been studied using either conditioned place preference tests, which measure preferences for aggression in non-volitional social situations, or using operant self-administration tasks during which a test animal can self-administer an intruder to freely interact with. In my prior work, I had found that testing environment significantly impacts aggression behavior (Chapter 4) and aggression motifs (Chapter 6), and therefore needed to design an assay to directly compare appetitive and reactive aggression in the same arenas. I therefore designed a 'quasi-yoked' paradigm, in which one set of animals contingently self-administered themselves an intruder, and an associated group was non-contingently administered an intruder with the same frequency as the contingent group. We could not directly yoke the groups because only ~70% of animals in the contingent groups learn the task, and direct yoking

led to unacceptable rates of attrition. With these 'contingent' and 'non-contingent' assays designed, I was able to directly compare aggression phenotypes between the sexes.

Sex differences in appetitive aggression

Female appetitive aggression is grossly understudied. In one of the first studies of female appetitive aggression seeking, I unexpectedly found that female mice who demonstrate reactive aggression in resident intruder tests go on to self-administer intruders, but, unlike males, do not typically attack them (Aubry*, Burnett*, Goodwin* et al., 2022, Chapter 5). This dissociation between males and females in appetitive aggression phenotype drove me to elucidate the neural networks responsible for suppressing appetitive aggression in females but not males (Figure 1). In the following experiments, I conducted whole-brain c-fos activity mapping between males and females during contingent self-administration, and between males undergoing either contingent or non-contingent administration (Figure 1a, Chapter 6). Whole-brain c-fos activity mapping serves as a proxy for strongly activated neurons and can identify regions and networks of regions

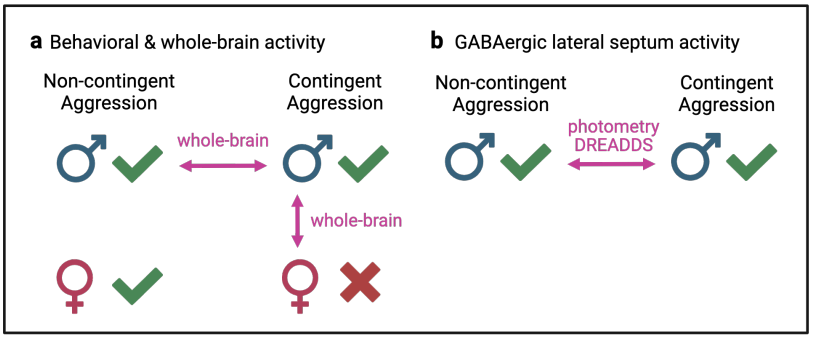


Figure 1.1 Experimental groups for whole-brain and region-specific analysis. *Females demonstrate reactive but not appetitive aggression phenotypes (Aubry*, Burnett*, Goodwin* et al., 2022).*

that are activated by aggressive experiences. Furthermore, I was awarded an F31 NRSA to study the role of the lateral septum (LS) in driving these aggression phenotypes due to its historical role in reactive aggression and 'septal rage', as discussed below.

Appetitive aggression circuitry

Several decades of work have begun to unravel the complex circuitry of aggression (Figure 1A), yet new nodes continue to be found (Zhu et al. 2021), highlighting the need for an unbiased whole-brain view of aggression circuitry. When examining specific nodes, the use of viral and genetic techniques have been instrumental in identifying regions influencing appetitive

aggression seeking (Figure 1B). The interplay between the lateral septum (LS), the ventral

tegmentum area (VTA), and the ventromedial ventrolateral hypothalamus (VMHvl) is of particular interest due to the lateral septum's known role in reactive aggression and potential influences on appetitive aggression (Figure 1C). While there are ~20 regions identified as influencing aggressive behavior, I will provide background here on regions with the most evidence for involvement in appetitive aggression, namely the lateral septum (LS), the ventral tegmental area (VTA), the nucleus accumbens (NAc), subnuclei of the hypothalamus including the medial preoptic area (MPOA) and the ventromedial ventrolateral hypothalamus (VMHvl), and the lateral habenula (LHb).

The lateral septum (LS) is an integral hub for reactive aggression (Figure 1C). Lesioning studies in the 1970s and 1980s resulted in 'septal rage', in which animals indiscriminately attack any available target, indicating that LS activity typically gates aggression behaviors (Slotnick and McMullen 1972; Miley and Baenninger 1972; Stark and Henderson 1972; Albert and Richmond 1975; Lau and Miczek 1977; Caroline Blanchard et al. 1977; Wallace and Thorne 1978; Blanchard et al. 1979; Potegal et al. 1981; Kishore and Desiraju 1990). The CA2 and CA3 regions of the hippocampus send excitatory projections to the caudal and rostral LS (LSr) which increase reactive aggression (Leroy et al. 2018) by exciting inhibitory microcircuitry within the LS (Wang et al. 2019), thereby reducing overarching septal activity. The LS canonically acts as a 'brake' on the ventrolateral ventromedial hypothalamus (VMHvl), inhibiting reactive aggression (Wong et al. 2016); thus lesioning the LS disinhibits the VMHvl, permitting reactive aggression. More recent work has begun to elucidate the interconnections of the LS that place it in a position to influence both appetitive and reactive aggression. Excitatory CA3 neurons project to the caudal LS (LSc), which in turn sends GABAergic projections to the VTA, which primarily synapse on local inhibitory GABAergic neurons thereby releasing inhibition within the VTA leading to increased downstream striatal dopamine levels (Vega-Quiroga et al. 2018; Soden et al. 2020; Wang et al. 2021). The VTA in turn sends dopaminergic projections back to the LSr which promote reactive aggression (Yu et al. 2014a). The LS also sends inhibitory projections to the lateral habenula (LHb), and inhibiting LHb firing via a GABAergic basal forebrain projection has been shown to increase conditioned place preferences to aggression associated contexts (Golden et al. 2016). The lateral septum's central placement in this network of reactive and appetitive aggression-related nodes highlights its potential as a key region for the differential

control of aggression phenotypes.

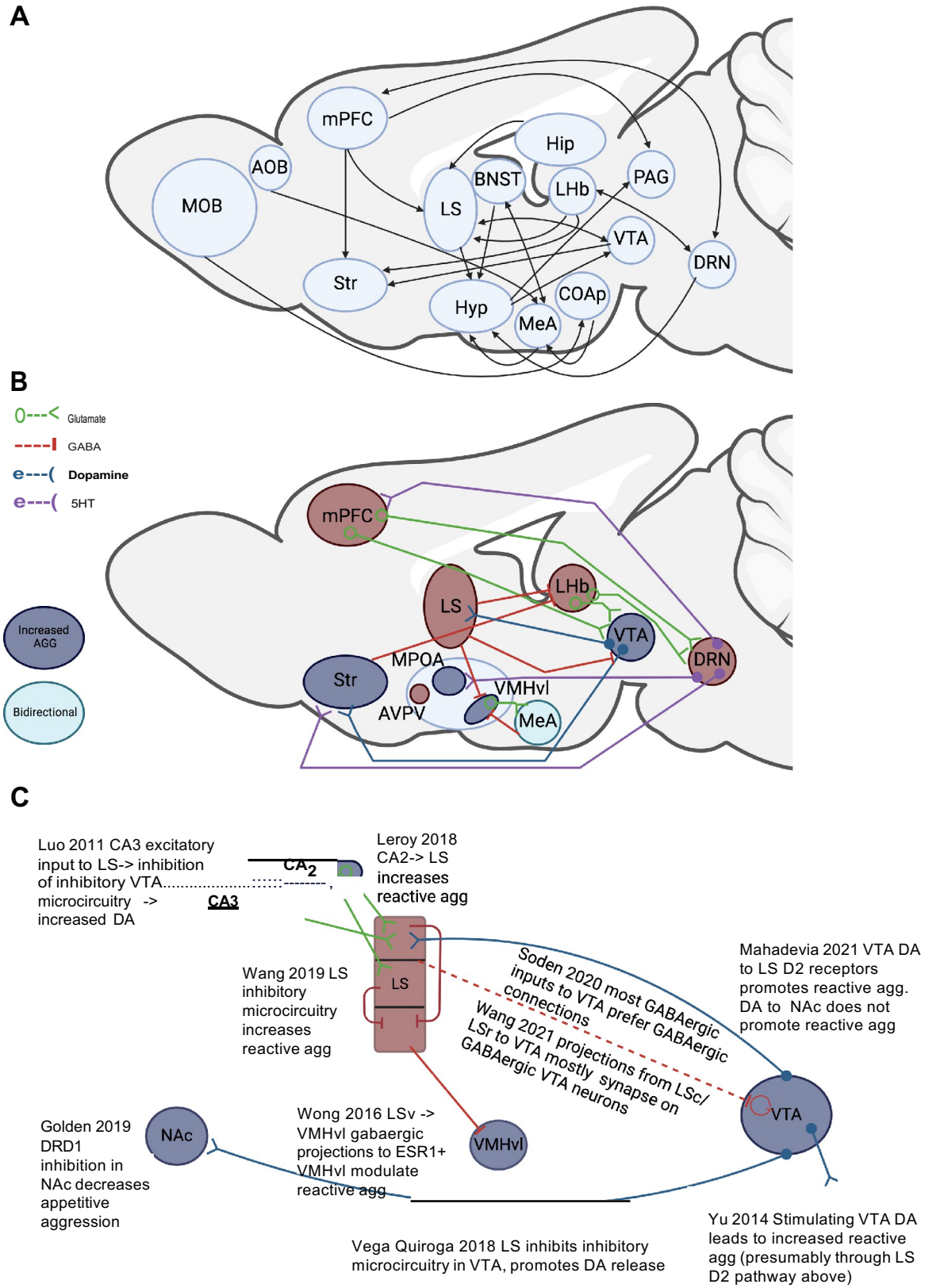


Figure 1.2. Aggression circuitry in the mouse brain. *A) An overview of the immense body of work studying reactive aggression. B) Circuitry associated with appetitive aggression seeking. C) Lateral septal involvement in aggression.*

Electrical stimulation and lesion studies have highlighted the role of the ventromedial hypothalamus (VMH) in the expression of attack across species (Putkonen 1966; Chi and Flynn 1971; Lipp and Hunsperger 1978; Kruk et al. 1979, 1983, 1990; Albert et al. 1979; Fuchs et al. 1981, 1985; Siegel and Pott 1988; Lammers et al. 1988; Siegel et al. 1999). More recently, the ventrolateral subregion of the VMH (VMHvl) has been identified as a region which bidirectionally controls reactive attack behavior (Lin et al. 2011). The VMHvl contains a large amount of hormone receptors, with estrogen receptor alpha (Esr1) and the progesterone receptor (PR) co-expressing in nearly 100% of cells (Hashikawa et al. 2018). These cells are key populations for mediating aggression in both males and females (Yang et al. 2013, 2017; Lee et al. 2014; Hashikawa et al. 2017), and a subpopulation of these neurons are active in males both when participating in and witnessing aggression (Yang et al. 2023). Many studies are moving into examining VMHvl connectivity with regard to aggression, and have found that a ventral hippocampus to VMHvl pathway bidirectionally controls impulsive attack behavior (Chang and Gean 2019), while a VMHvl to periaqueductal gray glutamatergic pathway connects to neurons that polysynaptically target the jaw and time-lock to biting behavior (Falkner et al. 2020). Despite these advances in preclinical models, however, there is little evidence that the VMHvl is involved in the elicitation of aggression in humans, and imaging studies have rather implicated upstream nuclei (Haller et al. 2006; Fanning et al. 2017).

The MPOA appears to primarily regulate aggression intensity. Disrupting MPOA connections to the VTA leads to decreases in maternal aggression (Numan and Callahan 1980; Numan and Smith 1984), and bilateral lesions inhibit reactive aggression severity in both male rats and female hamsters (Hammond and Rowe 1976; Bermond 1982; Albert et al. 1986). This may be partially driven by oxytocin, as the infusion of oxytocin into the MPOA in female Syrian hamsters significantly reduces the duration of aggression (Harmon et al. 2002). Increased aggression in pair-bonded male prairie voles toward strangers is also associated with increased c-fos activity in the MPOA (Gobrogge et al. 2007). In male mice, cMPOA^{ESR1+} projections to the VMHvl respond to the perceived fighting capability of an opponent, and increased activation of this pathway suppresses reactive aggression behavior in male mice (Wei et al. 2023). Additionally, 1′fosB but not 1′JunD overexpression in the MPOA resulted in female-directed

aggression from male mice (McHenry et al. 2016), indicating that there may be a special role for 1′fosB in aggression behaviors.

The lateral habenula (LHb) is comprised primarily of glutamatergic projection neurons, and serves to inhibit the release of dopamine from midbrain regions in order to signal negative valence (Matsumoto and Hikosaka 2007; Kawai et al. 2015; Li et al. 2021). Inhibiting LHb activity leads to increased CPP for aggression related settings (Golden et al. 2016), and LHb vGlut2+ neurons reduce their activity during biting phases of aggression, while GAD2+ neurons increase their activity. Furthermore, these GAD2 neurons exert local inhibitory control over the LHb as a whole, driven by orexin positive projections from the lateral hypothalamus (Flanigan et al. 2020). Glutamatergic projections from the LHb to the dorsal raphe nucleus are required for heightened aggression following instigated aggression, but not basal aggression (Takahashi et al. 2022). In clinical populations, lower global efficiency in the left LHb is highly associated with increased reactive aggression in patients with intermittent explosive disorder (Gan et al. 2019).

While it would be convenient if only appetitive aggression phenotypes involved reward circuitry, this is not the case. An important phenomenon of note is the ‘winner effect’, which is deeply intertwined with common experimental methods in aggression research. The winner effect will be discussed in more length in Chapter 2, but in short: winning fights, including reactive fights, can be reinforcing and makes an animal both more likely to engage in and win future fights (Ginsburg and Allee 1942; Oyegbile and Marler 2005, 2006; Kudryavtseva et al. 2011). In order to study aggression behavior rather than defeat behavior, resident intruder tests are frequently rigged so that the resident rodent wins the encounter, typically by introducing a younger, smaller, or previously defeated intruder. This rigged methodology is essential for appetitive aggression testing, as a resident animal will stop operantly responding for an intruder if the intruder begins to win fights. In the future, as automated behavioral detection is further integrated with high throughput electrophysiological methods and calcium imaging, it would be more ethologically relevant (and interesting, in my opinion) to revise resident intruder methodology to include a mixture of dominant and subordinate intruder animals to better understand dynamic changes in aggression and reward circuitry.

Thus, delving into reward circuitry, the VTA interacts with the LS and MPOA as discussed above to influence both aggression phenotypes. Optogenetically stimulating dopaminergic neurons in the VTA directly increases reactive aggression (Yu et al. 2014b)

through projections to the LS, not through projections to the nucleus accumbens (Mahadevia et al. 2021). In hamsters, fos reactivity in the VTA is positively correlated with reactive aggression and dominance behaviors in males (Gil et al. 2013), and male mice who repeatedly win reactive aggressive encounters experience increased tyrosine hydroxylase and dopamine transporter mRNA levels of the VTA, which may contribute to the winner effect (Filipenko et al. 2001). Furthermore, injection of GABA agonists injected into the caudal and rostral parts of the VTA led to increased reactive aggression in male rats (Arnt and Scheel-Krüger 1979).

While stimulation of dopaminergic projections from the VTA to the nucleus accumbens (NAc) does not influence reactive aggression in mice (Mahadevia et al. 2021), NAc deep brain stimulation reduces aggression in clinical populations (Park et al. 2017; Harat et al. 2021). Antagonism of dopamine receptor 1 (DRD1) decreases appetitive aggression seeking, while DRD2 antagonism leads to either no change or decreased seeking (Couppis and Kennedy 2008; Golden et al. 2019a). The NAc also shows considerable remodeling following repeated aggressive experiences. In male mice, 1′fosB induction following repeated aggression encounters increases DRD1 expressing medium spiny neurons (D1-MSNs) in the NAc, and overexpression of 1′fosB in existing D1-MSNs leads to intensified aggression behavior but no change in aggression conditioned place preference (Aleyasin et al. 2018). In female hamsters, repeated experience significantly increases spine density within the NAc core, but not in the shell or caudate putamen (Staffend and Meisel 2012), and leads to an increase in the expression of postsynaptic density AMPARs and group I metabotropic glutamate receptors, and an increase in the strength of the association between postsynaptic proteins and glutamate receptors (Borland et al. 2020). Dopamine increases in the NAc following, but not during resident intruder aggression in male rats (Erp and Miczek 2000).

Whole-brain specific experiments

These region and circuit specific results indicate that there are likely larger networks involved in controlling reactive and appetitive aggression. In order to better understand these networks, in addition to searching for novel regions implicated in aggression control, I next conducted whole-brain c-fos activity mapping in i) males and females undergoing contingent self-administration and ii) males undergoing contingent versus non-contingent aggression administration.

Between the sexes, social behaviors are grossly similar except for the moderate aggression shown by males. Preliminary whole-brain data indicate that a handful of regions significantly differ between both males and females, including several sensory regions and the bed nucleus of the stria terminalus, the caudate putamen, and the paraventricular nucleus of the hypothalamus. In males, non-contingently exposed animals show slightly longer attack duration with more attack bouts than contingent animals. Neurally, the lateral septum differs significantly, as well as the caudate putamen and several thalamic nuclei. Whole-brain c-fos mapping serves as a proxy for strongly activated neurons and network analysis can be used to identify regions and networks of regions that are activated by aggression experiences. In collaboration with Michelle Jin at Columbia University, we are currently performing network analysis on these datasets.

Conclusions

Appetitive aggression research is regaining popularity after many decades of work primarily focused on reactive aggression in males. The development of operant self-administration techniques has lowered the barriers to access for these types of studies and allows animals volitional control over intruder administration. Using these behavioral tools, we can now directly compare reactive and appetitive aggression phenotypes, which have long been considered separable in clinical literature, but have been understudied in the preclinical realm. My hypothesis throughout my dissertation work has been that reactive and appetitive aggression phenotypes are neurally separable phenomena controlled by overlapping networks of regions with differential activity in key nuclei, including the lateral septum. Working toward testing this hypothesis, I have worked as part of a team to develop a platform for the high throughput automated detection of animal behavior from recorded video and have published multiple papers calling for a reconceptualization of behavioral classifiers as objective, RRID-like reagents within the field. SimBA is now used in labs across the world. Further, I have directly tested my hypothesis via high-throughput behavioral testing and whole-brain c-fos activity mapping. My results indicate that while aggression behavior is grossly similar between phenotypes, this behavior is governed by different whole-brain activity networks.

Chapter 2 : Rage Against the Machine: Advancing the study of aggression ethology via machine learning.

Nastacia L. Goodwin[†1,2], Simon RO Nilsson[†1], Sam A. Golden[1,2,3]

[1] University of Washington, Department of Biological Structure, Seattle, Washington, USA
[2] University of Washington, Graduate Program in Neuroscience, Seattle, Washington, USA
[3] University of Washington, Center of Excellence in Neurobiology of Addiction, Pain, and Emotion (NAPE), Seattle, Washington, USA
[†] Denotes equivalent contribution

1. Introduction

There is an increased risk for abnormal or pathological aggression in individuals suffering from neuropsychiatric disorders. A growing literature indicates that certain types of aggression, namely compulsive aggression seeking, may be a distinct externalizing pathology that potentially functions through dysregulation of reward processing in a manner akin to drug addiction (Blair 2016; Chester and DeWall 2016). This is supported by the common finding that, in humans, some neuropsychiatric disorders present comorbid with inappropriate aggression, including depression and substance abuse (Tyrer et al. 2015; Fazel et al. 2015), and recidivism rates of violent offenders closely mimic relapse rates for drug addicted individuals (Sinha 2011; Ducrose et al. 2014). We and others have hypothesized that studying the motivational component of aggression, termed appetitive aggression, may provide novel therapeutic approaches to the treatment of maladaptive aggression presenting both within and outside of neuropsychiatric comorbidities (Miczek et al. 2015, 2017; Golden et al. 2019b; Covington et al. 2019). Unfortunately, from a preclinical perspective, there are relatively few established models for mechanistically studying the neurobiological basis of maladaptive appetitive aggression (Hashikawa et al. 2018; Flanigan and Russo 2019; Golden et al. 2019b).

Aggression motivation has long been a focus of preclinical research (Thompson 1963), invariably examined through the lens of ethological analysis, but often stymied due to the tremendous effort and durations required to manually score complex social behavior with sampling frequencies and accuracies that match modern neuroscience techniques. Several recent reviews have highlighted the power of machine learning approaches for creating automated behavioral classifiers to study social behavior, including aggression (Anderson and Perona 2014; Egnor and Branson 2016; Robie et al. 2017; Gris et al. 2017a; Brown and de Bivort 2018; Akay and Hess 2019; Mathis and Mathis 2019a; Datta et al. 2019). The ability to recognize and categorize common behaviors in model species is an integral component of model reproducibility and extendibility, and these observations can be combined with computational neuroethology to circumvent several inherent issues with manual annotation - most notably including observer drift and bias, long analysis times, and inter-rater-reliability (Anderson and Perona 2014; Egnor and Branson 2016; Datta et al. 2019). Additionally, both supervised and unsupervised machine learning algorithms have uncovered previously unknown behavioral repertoires in model organisms and have confirmed foundational assumptions of ethology

(Vogelstein et al. 2014; Wiltschko et al. 2015; Rudolf et al. 2019). Several machine learning-based open source packages have been developed that can track one or multiple animals during freely moving behavior, while others, including one recently released by our lab (Nilsson et al. 2020), can be used to classify behaviors based on pose-estimation tracking (Table 2). These approaches have led to a renaissance in the study of ethology, which is poised to catapult classical behavioral neuroscience into the realm of "big data".

However, even as these techniques have removed several hurdles impeding high throughput behavioral analysis, the use of automated behavioral classifiers has proven oddly difficult to adopt and generalize across labs. Implementation of these techniques is often slowed by a lack of computational knowledge, the need for specialized and expensive equipment, and the high computational expense to adequately train new classifiers. Due to the complexity of these programs, even the initial installation can be intimidating or difficult regardless of previous programming experience. In this perspective, we will briefly review the state of the appetitive aggression literature (also see Golden and Shaham 2018; Golden et al. 2019b), and then within this context, provide a primer on how machine learning approaches (regarding both acquisition/tracking and predictive classifier analysis) may be incorporated into future studies. Definitions of commonly used terms in computational neuroethology are included in Table 1. We propose, predominantly thanks to efforts of numerous labs in developing and advancing machine learning methods for behavioral tracking (He et al. 2017; Mathis et al. 2018a; Graving et al. 2019a; Pereira et al. 2019a), that currently available approaches are sufficient to overcome the main limitations preventing wide adoption of machine learning for scoring complex social behavior within the context of pre-clinical aggression research.

2. Appetitive aggression

2A. Winners like to win: revisiting aggression reward

Sixty years of behavioral research have shown that the opportunity for, or experience of, an aggressive encounter with a conspecific can be reinforcing in many species, including select mammals. Early work established the propensity of Siamese fighting fish (*Beta splendens*) to perform operant tasks for the opportunity to attack a static (Thompson 1963; Thompson and Sturm 1965) or animate representation of a conspecific (Craft et al. 2003, 2007; Elcoro et al. 2008). Similar behaviors were also demonstrated in homing pigeons (Cole and Parker 1971) and male fighting cocks (Thompson 1964). Although this pioneering work showed that animals will

seek the opportunity for aggression (once thought to be a uniquely human trait), these studies did not explore the relationship between seeking aggression and the actual experience of an aggressive encounter.

Subsequent work allowing physical attacks against conspecifics has highlighted the importance of dominance in solidifying aggression reward seeking. Among many species, those animals that win their first aggressive encounter are more likely to win subsequent bouts, while those that lose are more likely to lose subsequent bouts. Such "winner" and "loser" effects (Ginsburg and Allee 1942; Oyegbile and Marler 2005, 2006; Kudryavtseva et al. 2011) have been demonstrated in meadow voles (Vlautin and Ferkin 2013),crayfish (Momohara et al. 2016), flour beetles (Okada et al. 2019), *Drosophila* (Trannoy et al. 2015; Kim et al. 2018), lobster cockroaches (Kou et al. 2019), Syrian hamsters (Schwartzer et al. 2013), and many species and strains of mice (Oyegbile and Marler 2006). Some mammals do not display the winner effect, however, and species such as the white-footed mouse only do so following experimental manipulations of testosterone levels (Fuxjager et al. 2011). Aggressive encounters may be sought prior to exposure, but winning often matters in maintaining the drive to seek aggression and in determining future performance.

To minimize the effects of bout outcome on behavioral measures, many mouse assays for aggression use a variation on the resident-intruder task where a smaller subordinate intruder is introduced into the home-cage of a larger, older resident. Early studies using these procedures found that mice will cross electric grids (Lagerspetz 1964) and navigate T-mazes (Tellegen and Horn 1972; Legrand 1978) and runways (Legrand 1970) to attack a subordinate. Furthermore, physical aggression alone is sufficient to condition mice to prefer aggression-paired contexts (Potegal 1979; Taylor 1979). While these early influential studies demonstrate the utility of mouse models in the study of appetitive aggression, subsequent work shifted toward using resident intruder (Miczek and O'Donnell 1978; Brain et al. 1981) or sensory contact (Kudryavtseva et al. 1991) assays where pairings are repeated and the subsequent interactions are recorded and scored for a variety of behavioral measures. However, such procedures may be biased toward reactive aggression due to their inescapable and involuntary nature (Kudryavtseva et al. 2011, 2014). To overcome these limitations, recent work has focused on behavioral procedures that examine both reactive and appetitive aggression using operant conditioning

tasks, and this approach is instrumental for understanding the neurobiological differences underlying different types of aggression.

2B. Individual variability in inbred and outbred lines

A technical confound associated with preclinical aggression research is the relative lack of innately expressed aggression exhibited by commonly used inbred mouse strains relative to outbred strains (Jones and Brain 1987). Unlike outbred strains, inbred mice often require significant experimenter manipulation, either in the form of extended social isolation (Banerjee 1971) or repeated social instigation (Kudryavtseva et al. 2014; Covington et al. 2018), to exhibit significant levels of aggression towards conspecifics. To overcome this confound, preclinical aggression research, and especially studies of appetitive aggression, has focused on the use of outbred mice (Chia et al. 2005) that exhibit a spectrum of innate aggression behavior (Golden et al. 2016). While inbred mice are typically preferred for many research applications, recent meta-analysis shows that inbred strains do not have greater trait stability than outbred mice, and data from outbred mice may be more generalizable across conditions and populations (Tuttle et al. 2018).

We and others have used CFW or CD-1 outbred mice as these strains display several naturally occurring aggression phenotypes ranging from compulsive addictive-like aggression seeking to aggression avoidance (Golden et al. 2016, 2017b). Unfortunately, the use of outbred strains precludes the use of genetically defined Cre-recombinase based methods. To overcome this, we have introduced a hybrid breeding strategy using the F1 hybrid offspring derived from an inbred strain of interest and an outbred CD-1 (Golden et al. 2017a; Aleyasin et al. 2018). Following behavioral phenotyping for baseline reactive and appetitive aggression levels between inbred and hybrid populations, as well as molecular phenotyping for appropriate transgene expression, this approach introduces genetic selectivity to innately aggressive mice populations. We propose that this approach, in combination with more traditional but time-consuming backcrossing of Cre lines onto outbred strains or the impending development of outbred CRISPR transgenic lines, will provide a strong foundation for aggression research moving forward.

2C. Unconditioned vs. conditioned aggression

The conditioned place preference (CPP) procedure has historically been used to evaluate the rewarding effects of drugs and alcohol (Beach 1957; Mucha et al. 1982). In this procedure, one

distinct context is paired with the conditioned stimulus while another context is paired with the unconditioned stimulus. During a subsequent stimulus-free test, the laboratory animal chooses to spend time within the conditioned or unconditioned context. An increase in preference for the conditioned stimulus paired context is indicative of rewarding effects (Bardo and Bevins 2000). More recently, CPP procedures have been developed to assess the relative reward of affiliative social interactions (Panksepp and Lahvis 2007; Dölen et al. 2013; Goodwin et al. 2018).

Similarly, based on studies in female Syrian hamsters (Meisel and Joppa 1994) and male outbred OF-1 mice (Martínez et al. 1995), we have adapted a CPP procedure in combination with the resident-intruder social defeat procedure (Miczek et al. 1982; Kudryavtseva et al. 1991; Golden et al. 2011) to study aggression reward in CD-1 mice (Golden et al. 2016). Using this method, we first categorized unconditioned reactive aggression in adult CD-1 male mice through repeated daily resident-intruder assays with adolescent submissive C57BL/6J intruder mice. Mice that attacked the intruders during these screening assays (70%) were termed aggressors, while mice that did not attack (30%) were termed non-aggressors. We then evaluated conditioned aggression motivation using the aggression CPP assay. Mice that displayed aggression during the initial screening tests developed aggression CPP, while those that did not attack the intruders demonstrated conditioned place aversion (Golden et al. 2016). In a series of follow-up experiments, we parametrically explored the aggression CPP phenomenon and observed several key findings. First, based on the observation that unconditioned reactive aggression falls along a continuum in CD-1 mice, we characterized individual differences in aggression CPP by testing a third phenotype, termed "variable aggressors," composed of mice that performed inconsistently when repeatedly exposed to the resident-intruder procedure (Golden et al. 2017a). The variable aggressive mice exhibited significant, although weaker, aggression CPP, suggesting that repeated unconditioned aggression experiences can transform non-rewarding aggressive encounters into a rewarding experience. Second, aggression CPP is a learned phenomenon that can be acquired even by initially non-aggressive mice. Specifically, we exposed a large cohort of non-aggressors to 10 days of repeated resident-intruder testing and found that 50% transitioned to variable aggressors and exhibited aggression CPP. Lastly, aggression CPP is persistent, lasting several weeks following the final condition session (Golden et al. 2017a).

Notably, the portion of CD-1 mice that fail to show aggression is small and has not been the focus of study due to the difficulty in screening animals and filling group sizes. Advances in

automated behavioral tracking, however, may be able to distinguish these phenotypes at an earlier age, alleviating these restrictions, as will be discussed in the second part of this review.

2D. Addiction-like aggression behavior and relapse

Beyond measures of reactive aggression and conditioned aggression reward, several groups have developed operant tasks that measure appetitive aggression. The Miczek group has designed an operant conditioning panel including active and inactive nose-poke ports that can be introduced into the home cage of outbred CFW mice. The resident mice are trained to nose poke on fixed ratio and fixed interval schedules of reinforcement to attack intruders (Fish et al. 2002, 2005; Bannai et al. 2007). The development of operant aggression tasks has allowed the pharmacological decoupling of aggression *seeking* versus aggression *consumption* behaviors. The Miczek group reported that the $GABA_a$ positive allosteric modulator allopregnanolone increases operant response rates at lower doses than are required for increased attack behaviors (Fish et al. 2002), but that the effects are inhibited by the rise in corticosterone that are necessary for both operant responding and escalated aggression behaviors (Fish et al. 2005). While alcohol administration increases the motivation to fight, these effects are distinct from fighting performance and were not impacted by the antagonism of NMDA or AMPA receptors (Covington et al. 2018). 5-HT_{1b} receptor agonism was found to decrease attack intensity without changing operant responding (Bannai et al. 2007).

The Kennedy group subsequently replicated and extended these results, finding reliable operant aggression self-administration in mice under progressive ratio, differential reinforcement of low rate behavior, and variable ratio reinforcement schedules (Couppis and Kennedy 2008; May and Kennedy 2009). Local nucleus accumbens dopamine receptor 1 or 2 antagonism in Swiss Webster mice inhibited both operant responding and select attack behaviors (Couppis and Kennedy 2008), and we have found similar results following chemogenetic inhibition of dopamine receptor (Drd) type 1, but not type 2, in hybrid F1 CD-1 x Drd1-Cre or Drd2-Cre mice (Golden et al. 2019). Extensions of these operant procedures have also shown that animals rapidly cease aggression self-administration when confronted with a non-submissive intruder (Falkner et al. 2016a), further highlighting the necessity of winning in promoting aggression reward.

Pathological aggression in humans mimics cardinal features of drug addiction. Aggressive encounters are often sought despite severe negative consequences, pathological aggression

develops only in a minority of individuals (Lacourse et al. 2002; Provencal et al. 2015), and recidivism rates for violent offenders who are incarcerated for repeat violent offenses are similar to the relapse rates of individuals who take addictive drugs (Anthony et al. 1994; Ducrose et al. 2014). Such cardinal features have been reverse-translated from the clinic to create animal models of addiction-like behavior. Deroche-Gamonet (2004) used a combination of operant procedures including fixed and progressive ratio tasks, as well as fixed ratio with cue-contingent shock punishment, within a rodent model of cocaine addiction that has high face validity to the DSM IV criteria (Deroche-Gamonet 2004). In a cohort of rats that initially showed equal levels of cocaine self-administration and sensitization, 17% of animals went on to show addiction-like indicators including difficulty stopping or limiting intake, high motivation for access, and continued use despite negative consequence. These measures also correlated with relapse propensity (Deroche-Gamonet 2004; Piazza and Deroche-Gamonet 2013). Based on the above considerations, we have developed a modified operant chamber to test relapse to aggression seeking following (*i*) home-cage forced abstinence (Pickens et al. 2011), (*ii*) voluntary choice-based abstinence (Caprioli et al. 2015), and (*iii*) punishment-induced abstinence (Krasnova et al. 2014; Marchant et al. 2019).

Using this approach, we have shown that preclinical addiction models can be used to identify the neural mechanisms controlling appetitive aggression and relapse, as well as pathological or compulsive manifestations of aggression (Golden et al. 2017b). About 70% of aggressive mice learn to lever-press for aggressive interactions, and using several gold-standard models derived from the preclinical addiction literature, we observed aggression relapse after forced abstinence, punishment-induced abstinence, or choice-based voluntary abstinence that persists long after the last aggressive act. Through cluster analysis of the aggression-related measures we also identified a subset of mice that met criteria previously developed to denote compulsive addiction in rodent models (Deroche-Gamonet 2004). Specifically, the cluster analysis identified a subset of compulsive addiction-like aggressive mice (~19%) that exhibited intense operant-reinforced attack behavior, decreased likelihood to select an alternative palatable food reward over aggression, heightened relapse vulnerability and progressive ratio responding, and resilience to punishment-induced suppression of aggression-reinforced operant responding.

Importantly, this study found that contingent punishment is effective for suppressing aggression-seeking behavior in the majority of aggressive mice, but not in those exhibiting

compulsive aggression-seeking behaviors, and ultimately fails to prevent spontaneous recovery of aggression-seeking following extended abstinence in nearly all aggressive mice. These data are especially interesting in light of reports that, in rodents, footshock elicits both unconditioned aggression (O'kelly and Steckle 1939; Azrin et al. 1967) and Pavlovian conditioned aggression to a footshock-paired tone (Vernon and Ulrich 1966). However, work in non-human primates (Ulrich et al. 1969; Azrin 1970) and rats (Baenninger and Grossman 1969; Roberts and Blase 1971) have shown that mechanical pain or shock-induced aggression is suppressed by contingent, but not non-contingent, shock punishment. Further, non-contingent footshock inhibits aggression in dominant but not subordinate mice (Frischknecht et al. 1985). Together, and within the common context of 'punishment as a tool to prevent aggression', these data suggest greater nuance and present areas for future study within the context of adaptive and maladaptive compulsive aggressive behavior.

2E. Conclusions

The recent renaissance in aggression research has been assisted by the integration of carefully designed behavioral assays, derived from animal models traditionally used to study compulsive drug use and relapse in the addiction field, and modern neuroscience techniques such as chemogenetics (Coward et al. 1998; Armbruster et al. 2007) and optogenetics (Boyden et al. 2005). These manipulations have begun to help us understand the neural regions driving attack behavior and aggression salience (Miczek et al. 2001; Falkner et al. 2016a; Han et al. 2017; Stagkourakis et al. 2018; Golden et al. 2019a), but the timescale of neural manipulations and recordings necessitate higher resolution behavioral scoring than hand-scoring is able to consistently and reliably provide. Both aggression seeking and consumption phases of behavioral assays will benefit from more rapid, reliable and non-subjective quantification of behavior. Within the field of pharmacology, real-time quantification and behavioral prediction are necessary for closed-loop manipulations that can causally differentiate the behavioral sequence constituents of reactive and appetitive aggressive behavior. Specific to current operant procedures, such approaches will allow a detailed understanding of the sequala of behavioral events that occur after an operant action is contingently reinforced with an intruder. All of these time locked events, at previously unfeasible time scales and objectivity, may provide additional information on the motivational state of the aggressor and the neural mechanisms guiding this behavior.

Currently, aggression consumption outcome measures are typically restricted to latency to first attack, or proportion of attack versus non-attack behaviors. Such simplifications leave valuable ethological data unexplored, but these analytical omissions have been necessary due to the complexity of aggression behavior and the length of time and training required to accurately manually score assays. Species typical attack behavior in mice generally includes bites directed toward the back and flanks of intruder mice, and decreased attack behavior upon displays of submission by an intruder. Additionally, shifts toward more damaging bite locations or continued attack despite submission can both be indicative of escalated aggression (Takahashi and Miczek 2014; Newman et al. 2018). Automated behavioral classifiers can simultaneously measure these and other behaviors, of both the resident and intruder animal, allowing for a richer ethological continuum. The classifiers - once validated – can be curated and disseminated through online repositories, which can eliminate observer variability and present significant opportunities for cross-site standards within behavioral analyses. Accessible classifiers, that are fast, operational across labs - and interpretable and explainable (Table 1) - also has obvious benefits for transparency and scientific rigor.

We and others propose that open source machine learning techniques will allow for rapid, high-throughput explorations of the incredible nuance of these behaviors without sacrificing accuracy. The integration of these automated behavioral classifiers will overcome the hurdle of hand scoring which has bottlenecked the field since its emergence.

3. Machine learning

3A. Embracing machine learning

Automated video assessments can exceed human performance (Gris et al. 2017a), and behavioral classifiers increase throughput and consistency (Schaefer and Claridge-Chang 2012) in addition to reducing human bias and anthropomorphism in scoring (Robie et al. 2017). The development of specialized machine learning behavioral classifiers have uncovered previously unknown behavioral repertoires in animals including mice, drosophila, bats, and *C. elegans* (Vogelstein et al. 2014; Wiltschko et al. 2015; Rudolf et al. 2019; Zhang and Yartsev 2019). Markerless pose estimation and behavioral classification algorithms are rapidly improving, and many of the requisites are met for expanding machine learning approaches to classify complex social behavior. A key component to the adoption and success of these efforts is promoting open

source packages that can be easily and widely adopted throughout the behavioral neuroscience community.

There are several requirements for the ready adoption of machine learning approaches:

(1) Ease of use: The field of computational neuroethology is advancing at rapid pace, such that the ability of many behavioral neuroscientists to take advantage of these pipelines are curtailed by their skill behind a command line. Many of the recent additions to the field such as DeepLabCut (Mathis et al. 2018a) and DeepPoseKit (Graving et al. 2019a), however, provide graphic user interphases (GUIs) that allow the user to avoid setting hyperparameters or organizing projects via command line entries. Programs with easy installation and approachable graphical user interfaces will open doors for labs without easy access to computationally inclined individuals.

(2) Generalizability: Behavioral assays within and between labs can often be filmed under different lighting conditions, with animals and backgrounds shifting in hues, using variable hardware and acquisition parameters such as frame rate, resolution, color scales, and video format. Many current implementations of machine learning for animal behavior (Table 2) require specialized hardware setups that can be prohibitively expensive or challenging to implement and scale-up for high-throughput behavioral assays, and often do not generalize well from one preparation to another (Anderson and Perona 2014). Machine learning approaches that can use standard, readily accessible recording hardware without the need for specialized builds are vital for generalizability and comprehensive implementation.

3) Cost management. Current pose estimation and behavioral classifier programs benefit from the use of depth cameras, multiple filming angles, specialized acquisition hardware, and proprietary software. Access to high-end analytical platforms, either in the form of cloud-based or local solutions, are a requirement for generating both tracking and classifier models. These costs propel computational neuroethology out of the reach of many behavioral neuroscience labs. However, the behavioral assays commonly utilized for the study of aggressive behavior are well defined and often recorded with a single camera, making it possible to generate and share ‘base’ tracking and classifier models that preclude the need for individual labs to heavily invest in specialized hardware or commercial software.

4) Accuracy. At a minimum, automated classifiers need to meet the accuracy of manual annotation conducted by highly trained observers. Regardless of the adoption of machine

learning techniques, the ability to consistently recognize behaviors of interest in model species is integral to behavioral research, and the experimenter aptitude to recognize significant behavioral events determines the success in machine model generation, tuning and validation. We have found, however, that even observers trained to high rates of interrater reliability on strict operational definitions of behaviors (Table 3) may occasionally miss shorter behavioral bouts when scoring long video recordings (e.g., 20k+ video frames). Machine learning approaches provide the advantage of never tiring, and in our experience, classifiers consistently detect shorter behavioral bouts missed by human observers. While ideal algorithms for specific tasks are typically dependent on data structure and the computational knowledge within a lab, we and other labs have found that supervised approaches (Nilsson et al. 2020), such as random forest machine learning algorithms, provide an appropriate balance of ease of use, interpretability, explainability, and accuracy that often exceeds human annotation accuracy.

5) Easy expansion of training sets. Ethological datasets provide an interesting challenge for machine learning. Training data provided to algorithms must be representative of the testing data, but behavioral assays are often conducted in discrete cohorts with unique properties. Training sets should contain a balanced mix of control and treatment videos, and diverse and unstable environmental influences (cage change schedules, seasonally noisy HVAC systems, etc.) must be considered as potential confounds if not represented in the original classifier training data. Specific to pharmacological research, when individual animal movement characteristics and social behaviors may be altered by pharmacological interventions, it is not feasible to create representative training sets from the outset. We and others have approached this challenge by creating the ability to easily add additional hand annotated videos (several short videos are typically sufficient) to training sets, which can then rapidly be used to create updated, iterative, behavior classifiers and animal tracking models.

6) Interpretability. Interpretable methods are of outmost importance for reproducibility, transparency and rigor (Table 1). Methods for creating accurate classification algorithms in behavioral neuroscience should provide transparent processes allowing for clear understanding of how classifications are made and, ideally through GUI control, the capacity to titrate model parameters to operationalized standards. This may include the ability to visualize decision paths and the importances of individual independent variables or features for the classification result and how the data and features are balanced within the model. These concepts have however –

traditionally – not always been favored within computer and data sciences (Rudin, 2019) and may introduce additional challenges when creating accessible methods within preclinical social behavior and aggression research.

Aggression behavioral assays provide excellent insight into the current capabilities and limitations of machine learning algorithms due to the numerous technical challenges they present. Behavioral components often include a very rapid succession of bites, lunges, pursuits, anogenital sniffing, tail rattles, lateral threats, boxing, and occasionally head shakes and prosocial behaviors such as grooming, among many others. These sequences can be further reduced to levels of granularity which may or may not convey important information (e.g., face bite versus flank bite). Furthermore, on video, animals frequently occlude each other and significantly change their body shapes as they aggress, creating difficulties for pose estimators. To our knowledge, two groups have published work using automated classifiers for evaluating aggression behaviors. Hong et al., 2015 was able to classify attack behaviors in male mice of different colors using a multi-camera apparatus, and Burgos-Artizzu et al., 2012 achieved accuracy rates of ~61% using a multi-camera apparatus with two unmarked mice of the same color. Our lab has recently released Simple Behavioral Analysis (SimBA) (Nilsson et al. 2020), an automated behavioral classification technique for aggressive, submissive, and social behaviors, that was built upon pose-estimation and the six principles outlined above.

The remainder of this review will discuss the current state of machine learning for social behavior classification in rodents in terms of aggression behavior, and the progress of the field toward meeting the goals discussed above.

3B. Supervised versus unsupervised learning

Many machine learning techniques for behavioral analysis first identify the position of animals frame-by-frame in a video via pose estimation or background subtraction. In particular, accurate and rapid pose-estimation via recently developed and generalizable convolutional neural network architectures - accessible through packages such as DeepLabCut (Mathis et al. 2018a), DeepPoseKit (Graving et al. 2019a), and LEAP (Pereira et al. 2019a) - provide a platform for generating the pose-estimation data needed to create rich feature sets required for machine classification of complex social behavior. For example, the predicted poses may be used to estimate distances between animals, their velocities, angles and accelerations across rolling windows and this may correlate with human annotated instances of aggressive behavior and used

in further modelling techniques (Table 4). These positional data are then analyzed to cluster statistically similar images that the program either identifies as a predefined behavior (supervised learning), or the cluster is studied by the experimenter who then adds behavioral labels post-hoc (unsupervised learning).

Unsupervised learning has successfully classified behaviors in tunicates, drosophila, individual mice, and pairs of mice after cropping the animals into individual videos (Vogelstein et al. 2014; Wiltschko et al. 2015; Klibaite et al. 2017; Rudolf et al. 2019; Dolensek et al. 2020). These techniques can be tremendously powerful in identifying the inherent structure present in behavior. Because the user does not predefine behaviors of interest, these algorithms have been proposed to be less biased than supervised techniques. Thus far, unsupervised techniques have resulted in important advances in mouse ethology including the identification of facial expressions corresponding to neuronally separable emotion states (Dolensek et al. 2020), and novel sub-second behavioral structures (Wiltschko et al. 2015). The program B-SOiD has also combined unsupervised t-SNE clustering and Support Vector Machines to great success in classifying naturally occurring behaviors in single mice, within a pipeline that is readily accessible to non-specialists (Hsu and Yttri 2019).The power to identify new behavioral repertoires is alluring and has the potential to significantly advance the field of ethology. Unsupervised machine learning tools are notoriously difficult to tune and interpret, however, and we propose that they should not be a first line option for generalization to labs without significant computational experience.

Training an unsupervised behavioral classifier generally involves feeding many unlabeled video frames into a user-defined algorithm which then identifies and separates images into behavioral clusters based on user-defined mathematical requirements for differences and similarities. The de Bivort lab has conducted one of the only large-scale direct comparisons of unsupervised learning techniques in classifying animal behavior (the leg movement of flies) and provides an excellent overview of unsupervised learning techniques (Todd et al. 2017). A typical pipeline for clustering often involves data pre-processing, dimensionality reduction, and cluster assignment (Todd et al. 2017; Datta et al. 2019), though see (Wiltschko et al. 2015).

Each of these steps are modular and can be performed using any of several different algorithms. Pre-processing can include transformations such as time frequency analysis, vector normalization, and wavelet transformation, while dimensionality reduction is often accomplished

via techniques such as principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE). The final clustering algorithms, such as k-means clustering or Gaussian mixture modeling, are used to group the data into clusters that researchers can apply behavioral labels towards (Todd et al. 2017). The flexibility of unsupervised learning approaches at each step makes them very powerful in adapting to different datasets, but tuning parameters requires careful consideration.

Importantly, unsupervised learning techniques do not have built in performance metrics. Tuning them to high perceived accuracy (e.g., all user defined attacks are also defined by the algorithm as an attack) is in essence training the model to meet particular output benchmarks, i.e., creating a supervised algorithm (Brown and de Bivort 2018). Some classes produced by an unsupervised approach may also be difficult for users to define as recognizable behaviors. Todd et al., 2017 propose evaluating networks via parameters such as minimum dwell time within a behavior class and the reliability of the algorithm when trained multiple times on the same data set (Todd et al. 2017; Brown and de Bivort 2018). These parameters acknowledge the real-world constraints on behavior and the need for replication while imposing less bias onto the model in terms of expected output. We propose that any implementation of algorithmic features during unsupervised training should be accompanied by a description of the experimental objectives driving their necessity.

Supervised learning algorithms require greater user oversight in that users are required to annotate training data that the classification system can use to learn to correctly label images. Several pose estimators are validated for use in drosophila (Branson 2014; Berman et al. 2014; Günel et al. 2019), and tools such as DeepLabCut (Mathis et al. 2018a), LEAP (Pereira et al. 2019a), and DeepPoseKit (Graving et al. 2019a) have been validated in mice and other non-human mammals. DeepLabCut and LEAP use variations on the supervised learning technique of deep neural networks to estimate animal pose, while the DeepPoseKit uses a novel Stacked DenseNet algorithm. For in-depth explanations of these estimators, see (Mathis et al. 2018a; Graving et al. 2019a; Pereira et al. 2019a; Nath et al. 2019; Mathis and Mathis 2019a). In many of these packages, users label body-parts at set coordinates, and these training labels are used to generalize computed labeling rules to future images. Pose estimators are powerful tools in identifying finer points of animal activity, and are able to identify phenotypes not detected with simply trajectory tracking (Hong et al. 2015). Notably, new packages can include project

management and image annotation GUIs, which greatly enhances the user experience and has resulted in more widespread adoption.

Supervised learning techniques can be used for both pose estimation and behavioral classification, but unlike unsupervised algorithms, can only classify predetermined behaviors and cannot identity novel behavioral repertoires. Training accurate supervised machine learning classifiers requires the careful, frame-by-frame annotation of a behavior of interest. Human variability and annotation mistakes, such as erroneous labelling of aggression events as non-aggression events, can generate significant noise that propagate to future machine analyses (Frenay and Verleysen 2014). It is therefore important to have precise behavioral operational definitions that encompass and exclude the behavioral events and non-events, respectively, and we present examples of how such precise operational definitions may look in Table 3. While such precise descriptions are not current standard laboratory practice, and functionally translate to extra time hand-annotating videos, they can promote the introduction of standardized cross-laboratory definitions through shareable classifier repositories that ultimately increase replicability as discussed previously.

Several behavioral classifiers using supervised machine learning techniques have been validated in mice, including JAABA (Kabra et al. 2013), Autotyping (Jhuang et al. 2010), and SimBA (Nilsson et al. 2020). Each uses simple, single camera setups, with significant differences in underlying machine learning approaches to creating predictive classifiers. JAABA classifies walking and following behavior in groups of same colored mice using a boosting ensemble algorithm (Kabra et al. 2013), and Autotyping uses a combination of Hidden Markov Model and Support Vector Machines to classify a variety of home cage behaviors in single-housed mice (Jhuang et al. 2010). SimBA uses random forest classifiers to classify aggressive, defensive, and other social behaviors in mice following pose-estimation (Nilsson et al. 2020). Some programs also use multiple cameras, depth cameras, or other specialized setups for behavioral classification. The Anderson Lab has developed a random forest classification technique which can identify attack, mounting, and investigation behaviors in differently colored mice using multiple depth sensing cameras (Hong et al. 2015). Boosting ensemble algorithms have also been used to classify social movement, attack, copulation, exploratory, and drinking and eating behaviors with integrated top and side view filming (Burgos-Artizzu et al. 2012a).

Whether implementing supervised or unsupervised machine learning techniques, there is a strong argument for using algorithms that are easily interpretable (Rudin 2019). Black box models have been helpful in image analysis and uncovering new behavioral patterns, but the inability to readily understand the underlying assumptions of algorithms is a detriment. Black box models - not unlike prevalent behavioral protocols - can appear to be valid and accurate (strong face validity) whilst measuring factors unrelated to the phenomenon it claims to measure (weak construct validity). Examples of such fallacies, with more trivial consequences, would include convolutional neural networks that 'successfully' discriminate Husky dogs and wolfs based predominantly on only the presence or absence of snow in the image (Ribeiro et al. 2016a), or networks generating 'successful' gender classifications from iris texture images through the presence or absence of mascara (Kuehlkamp et al. 2017). More seriously, black box models are notorious for aggregated racial and social-economic biases (Mittelstadt et al. 2016; Rudin 2019; Obermeyer et al. 2019); while clearly a different class of concern than what is anticipated in preclinical behavioral computational neuroethology, the cautionary tale remains relevant and important. Explainable and interpretable models should thus be adopted as best scientific practice (Rudin 2019) to allow for user oversight of the parameters guiding decision processes.

While supervised learning techniques require users to predefine behaviors of interest, they are often much easier to tune and interpret than unsupervised learning techniques, particularly for labs without statisticians or computer scientists on staff. In the spirit of ease of use, generalizability, and accuracy, we propose that supervised learning techniques are a good starting point for the automated examination of aggressive behaviors.

3C. Common classifying algorithms for supervised learning

Within supervised learning, there are many different algorithms from which to choose, and no method is universally superior to the others (Hand 2006). The appropriate selection depends on the structure, noise, and biases within each dataset, as well as the threshold for acceptable training duration and willingness to troubleshoot parameters (Hand 2006; Anderson and Perona 2014; Egnor and Branson 2016; Gris et al. 2017a; Akay and Hess 2019). Common techniques for behavioral and image classification include variations of Neural Networks, Support Vector Machines, Gradient Boosting Machines, Random Forests, and Hidden Markov Models. All of these algorithms have potential benefits and may outperform others on specific datasets, but we

and others have found that random forest classifiers provide high accuracy classification (Breiman 2001; Liaw and Wiener 2002; Nilsson et al. 2020) with the added benefit of interpretability, being easy to tune, and robust against overfitting. Here we describe several of these algorithms and highlight the potential pros and cons they possess regarding the classification of aggression-related behaviors.

Neural Networks (Hopfield 1982; LeCun et al. 1989)

Neural networks or convoluted neural networks are typically used for large datasets with a high number (tens of thousands) of features and observations, as well as noise. Features may include the trajectory of individual body parts, Euclidean distances, body part movement over a small frame of time, etc. Users provide an input layer of data to the neural network and define the desired categories in the output layer. The network then computes additional hidden layers between the input and output layers to correctly classify input data into the appropriate output categories or classifiers. Hidden layers are advantageous as specific features are not user-selected for inclusion within the algorithm, but the cost is an absence of information on what the resulting classifications are ultimately based on. Neural networks can be challenging to interpret and tune, and similar performance can be achieved with other methods (Lietman et al. 1999; Nitze et al. 2012; Liu et al. 2013). Many programs, including the pose estimators described above, successfully use deep neural networks for pose and/or behavioral classifications (Karpathy et al. 2014; Krizhevsky et al. 2017; Mathis et al. 2018a; Pereira et al. 2019a). Due to the lack of information regarding feature weights, we propose that more transparent algorithms which perform at similar levels are preferable for behavioral classification derived from tracked features.

Random Forests (Breiman 2001; Liaw and Wiener 2002)

Random forests are a type of ensemble algorithm. Ensemble algorithms are composed of many independently trained weak models which are combined to make strong predictions. Random forests rely on the bootstrapping of both (i) subsets of data, and (ii) predefined features to make a powerful forest of decision trees than can then vote on the classification of a behavior. For example, a decision tree in a random forest starts with a subset of data (e.g., 500 frames of 50,000) then uses a subset of the features (e.g., 10 out of 100 features) to split the data into yes/no classifications. Random forests are useful because each branch is created using the feature

which provides the most information, and by combining these data from the forest, it is possible to determine which features were most important for creating the behavioral classification. This technique is computationally rapid as the trees are independent and can be built in parallel, and they are unlikely to overfit data based on noise (Breiman 2001).

One weakness of random forests is their inability to natively support biased datasets. These datasets are common in behavioral videos, in which most frames do not contain the behavior of interest. For example, in a five-minute video filmed at 80FPS with 10 total seconds of tail rattle, this would result in 23,200 frames with no tail rattle, and 800 frames with tail rattle. Constructing a random forest with these data as-is would mostly construct trees with no or very few instances of tail rattle frames. In order to train the random forests more robustly, over-sampling and/or under-sampling techniques (Chawla et al. 2002; Batuwita and Palade 2013) can be used to balance the data.

Gradient Boosting Machines (Freund and Schapire 1997; Friedman et al. 2000; Friedman 2001)

Gradient boosting machines build decision trees iteratively, attempting to fix erroneous classifications after each node split by finding a different feature which better classifies incorrectly classified data. Gradient boosting can be more robust than random forests if tuned correctly but are more prone to overfitting data and incorrectly identifying noise, such as pose misestimation, as legitimate behavioral data. While pose estimation is often highly accurate, there are outliers that may cause overfitting when fed into a gradient boosting machine. With more accurate pose estimation and highly consistent filming conditions across labs, gradient boosting machines may provide superior classification than random forests, but it is difficult to currently meet these conditions and maintain generality of video recording conditions. While the use of gradient boosting machines is a good goal, their widespread adoption may depend on how well ‘base’ models, and their underlying recording acquisition parameters, are standardized across laboratory recording environments.

Support Vector Machines (Cortes and Vapnik 1995)

Support vector machines are frequently used, and work best with, small datasets containing few outliers (Chih-Wei Hsu and Chih-Jen Lin 2002; Xu 2006). Using multiple tuning parameters (kernel, regularization, gamma, and margin), support vector machines identify clusters of similar and dissimilar data in multidimensional space and find the regression line that best separates

clusters based on the training data. Understanding linear algebra and principle component analysis, in addition to having appreciable patience with tuning parameters, is helpful when using support vector machines. Like random forests, support vector machines do not work well natively with biased data sets, which can similarly be overcome through re-balancing and over- and under-sampling techniques.

There are few comparative studies of classification algorithms due to lack of generalization to other datasets, but in those that exist, support vector machines do not typically perform as well as random forests or gradient boosting machines (Caruana and Niculescu-Mizil 2006; Caruana et al. 2008). Due to the larger number of tuning parameters involved in support vector machines, and the similar or superior classification performance by random forest classifiers, we suggest that random forests may be a better starting point for creating generalizable and easily adopted classifiers.

Hidden Markov Models (HMM) (Rabiner and Juang 1986)

HMMs can be integrated with supervised or unsupervised techniques and excel at finding patterns in small sequences of data. These models have recently been extended for use in behavioral classifications due to the similarities from frame to frame of video, and the inherent structure of behavior (Carola et al. 2011; Wiltschko et al. 2015; Arakawa et al. 2017). HMMs use hidden weights to predict the probability of a transition from one state to another. For example, if a resident is biting the flank of an intruder in one frame, it is unlikely to be grooming its face in the next frame. As indicated by the name of the algorithm, behavioral classifications depend on a hidden state. Anderson & Perona, 2014 raise an interesting point that HMMs may be valuable in objectively measuring behaviors resulting from the "emotion state" of an animal. HMMs have been used in rodent aggression studies to investigate the effects of context (intruder behavior) on the aggression behavior of a home cage resident to parse species typical versus escalated aggression phenotypes (Haccou et al. 1988; Natarajan et al. 2009), and may provide an interesting path forward for understanding contextual aggression. While the hidden states can potentially be found via maximum a posteriori state estimations (Allahverdyan and Galstyan 2009), we propose that more inherently interpretable algorithms are preferable for initial behavioral classification.

3D. Promising directions

We strongly reiterate that there is no single correct pipeline for machine learning classification of animal behavior, but rather that the most successful approaches will likely use multiple pipelines in parallel while taking advantages of their pros and attempting to diminish their cons. The powerful techniques discussed above are summarized in Table 5, and the choice between them often revolves around the applicable knowledge within a lab and which algorithm best classifies data during pilot testing (Table 5). To facilitate generalizability to new labs, however, random forests seem like an excellent option due to their relatively few input parameters, robustness to noise, relative interpretability and explainability, and high performance on a wide array of data. A future goal of automated behavioral analysis is long-term real-time tracking of individuals within groups in a naturalistic setting, and many groups are currently working with these and other techniques to solve parts of this problem.

Distinguishing similarly looking animals has historically been a challenge for automated pose estimators, and ultimately problematic for the computational study of social behavior as many experiments require the use of nearly identical looking animals. Although not a problem when using outbred CD-1 mice as residents and C57 mice as intruders, due to their difference in coat color, this does preclude the generality of machine learning approaches to many rodent social behavioral assays. There are several recent programs that work to track the identity of individual and similarly appearing animals in groups over time, including ToxID, MoST, and idTracker.ai (Thanos et al. 2017; Rodriguez et al. 2017; Romero-Ferrero et al. 2019). Additionally, social tracking extensions of the program LEAP (sLEAP) and DeepLabCut have recently been released, and provide pose estimation for multiple interacting animals (Mathis et al. 2018a; Pereira et al. 2019a) by taking advantage of algorithms such as optical flow (Brox et al. 2006), part affinity fields (Cao et al. 2017), and deep-learning. The program idTracker uses separate segmentation and identification networks to individually identify animals and their trajectories and back propagation techniques that identify animals during partial occlusions (Pérez-Escudero et al. 2014).

Beyond image detection, there are several groups working to integrate RFID tracking and video tracking (Weissbrod et al. 2013; de Chaumont et al. 2019; Peleh et al. 2019). Live Mouse Tracker (de Chaumont et al. 2019) uses a combination of RFID tracking and depth imaging and has been validated with groups of up to 4 mice in a semi-natural environment over time. These

programs are highly influential in allowing for the long-term study of group dynamics in large, enriched home cage environments and currently provide the state-of-art tracking solutions for mice of similar appearance. A commercial version of this program, RFID-Assisted SocialScan, is also available and integrates a large behavioral arena with attached nest boxes (Peleh et al. 2019). Together, these programs can track the identity of individuals while examining social hierarchies over long periods of time without excessive manual oversight or behavioral scoring. However, these approaches require significant investments in highly specialized hardware.

Several groups pioneering pose estimation are also continually working to increase the speed of their algorithms to achieve real-time results (Graving et al. 2019a; Mathis and Mathis 2019a). Currently, ToxID and Sensory Orientation Software are able to provide real time tracking of individual mice and their trajectories (Gomez-Marin et al. 2012; Rodriguez et al. 2017), and DeepLabCut has recently been used for real-time pose estimation (Forys et al. 2018). Extending these platforms for in depth real-time behavioral classification in mice will allow for the evaluation of behavior on the timescale of neural activity and allow fully "closed-loop" recording and manipulation studies. The Stytra package is currently validated in zebrafish larvae, and is able to both analyze pose and activity and rapidly provide stimuli to the experimental animal based on its activity (Stih et al. 2019).

These open source packages are continuing to improve and emerge, and integration between them is becoming easier. Pose estimators that can rapidly track individual animals despite multiple occlusions is an essential first step for the automation of aggression assays, and our lab has focused heavily on exploring random forest classifiers for the classification of aggression behaviors. Integrating the pose estimators above with high-accuracy and easy to use classifiers in an open source GUI allows for the generalization of these techniques across labs, while maintaining the ability to take advantage of newer tracking technologies as they emerge.

For example, our lab has recently developed an open-source graphical workflow (for more information, see the SimBA GitHub repository) for creating supervised decision ensembles from body-part tracking data generated through convolutional neural networks in accordance with the considerations and principles discussed herein (Nilsson et al. 2020). The package incorporates accessible menus for pre-processing and annotating videos, generating / deploying classification algorithms, with advanced machine learning validation, tuning and visualization tools. The models created through this software – together with extensive documentation and tutorials - are

available for the scientific community through online repositories and we hope that this can spur further data sharing efforts that boosts the accuracy and scope of machine classification techniques in pre-clinical aggression research.

4. Conclusion

Aggression is an immutable force that contributes to the suffering and death for millions of people around the world (Sumner et al. 2015). While aggression can be highly rewarding and pursued despite adverse consequences (Chester and DeWall 2016; Gan et al. 2019), as well as sought after lengthy abstinence (Ducrose et al., 2014), aggression is not generally viewed in the medical profession or by the public as a component of neuropsychiatric disease states (Golden et al., 2017a; Golden and Shaham, 2018). Indeed, in the recent formulation of Research Domain Criteria **(**RDoC**)** of NIMH the term aggression does not appear under any of the research domains. Not surprisingly, no progress has been made in the treatment of pathological aggression, and little progress has been made using aggression as a diagnostic tool for neuropsychiatric comorbidity (Martin et al. 2013). We believe that drawing greater attention to the neurobiological mechanisms of maladaptive appetitive aggression within the context of neuropsychiatric disease will both act to (*i*) de-stigmatize aggression comorbidity in a manner similar to current approaches to substance abuse treatments and (*ii*) support the identification of novel mechanistic therapeutic approaches, which have presently been relegated to neuroleptic dopamine antagonists like haloperidol (Ostinelli et al. 2017) that exert their clinical efficacy through the neuroleptic's sedative effects (Calver et al. 2015). Therefore, we propose that neurobiological and behavioral tools used to study drug seeking and relapse should be used to study brain mechanisms of appetitive aggression, both preclinically and clinically. Further, expanding beyond reactive aggression to incorporate the full spectrum of ethologically relevant aggressive behaviors is an essential step in increasing the utility and translation of ongoing preclinical research. Machine learning tools enabling high throughput automated behavioral analysis are key to pursuing these lines of inquiry at a pace and depth necessary for modern neuroscience methods.

Significant progress has been made in developing open source software packages that are capable of tracking social behavior over long periods of time and across diverse acquisition parameters (Table 5), and we propose that a focus on increasing ease of use, generalizability, cost-consciousness, accuracy, and easy expansion of training sets will allow for their wider

adoption. Such automated scoring promises to remove long-standing bottlenecks within aggression research and allow for the high-throughput experiments required for mapping the nuance of aggression phenotypes while simultaneously building a common language of aggression phenotypes through shared machine models across labs.

Table 2.1. Definitions of common terms in computational neuroethology.

Term	Description
Pose	The relationships and location of body-part 'key-points' (e.g., individual joints/digits in motor control tasks; or the nose, ears, flanks, centroid and tail base in an open-field task) that are relevant for determining an individual's activity in a video frame.
Pose-estimation	Computer-generated predictions of the individual(s) pose in a video frame.
Behavior classification	A computer-generated categorical prediction describing the presence or absence of a behavior (e.g., attack vs. no attack) in a video frame.
Features	Independent variables derived from pose-estimation (e.g., animal velocities and the distances between animals) that are used to predict the presence and absence of a behavior in a video frame.
Feature weight / importance	The relevance of a specific feature for accurately classifying the presence and absence of a behavior in a video frame. For example, the distance between the first animal's nose and the second animal's tail base has importance or weight for accurately classifying anogenital sniffing.
Ensemble algorithm	The grouping of multiple learning algorithms based on different subsets of features. The individually weak learning algorithms are combined to create a strong or accurate behavior classification (e.g., random forest algorithm).
Hidden states	Unobservable factors determining the probability for ensuing expression of a behavior given the current behavior.
Model interpretability	The extent to which a machine model's decision processes can be understood by humans (e.g., *how* does the algorithm discriminate between attack and non-attack events?).
Model explainability	The extent to which a machine learning system allows a qualitative comprehension of the association between the features of a specific observation and its prediction (e.g., *why* is the video frame classified as containing an attack event?). Note that an explainable model also is an interpretable model.
'Black-box' model	A metaphor applied to machine learning algorithms with decision processes that are not readily explainable or interpretable.

Table 2.2 Currently available open source tracking or behavioral classification software. * Next to software name indicates multiple cameras or specialized equipment necessary.

Ctrax (Branson et al. 2009)	Drosophila	Yes	Behavior	Walk, stop, sharp turn, crabwalk, backup, touch, chase	Computer vision	http://ctrax.sourceforge.net/
CADABRA (Dankert et al. 2009)	Drosophila	Yes	Behavior	Lunging, tussling, wing threat/extension, circling, chasing, copulation	Supervised	http://www.vision.caltech.edu/cadabra/
Autotyping (Jhuang et al. 2010)	Mice	No	Behavior	Home-cage behaviors: drink, eat, groom, hang, micromovement, rear, rest, walk	Supervised	http://cbcl.mit.edu/software-datasets/mouse/
JAABA (Kabra et al. 2013)	Drosophila, mice	Yes	Behavior	Flies: walk, stop, crabwalk, backup, touch, chase, jump, copulation, wing flick/grooming/extension, righting, center pivot, tail pivot. Mice: follow, walk.	Supervised	https://sourceforge.net/projects/jaaba/files/
MiceProfiler (de Chaumont et al. 2012)	Mice	Yes, if no obstruction	Behavior	Head-head contact, anogenital contact, side by side, approach, leave, follow, chase	Physics engine	http://icy.bioimageanalysis.org/plugin/mice-profiler-tracker/
Unnamed* (Burgos-Artizzu et al. 2012b)	Mice	Yes	Behavior	Approach, walk away, circle, chase, attack, copulation, drink, eat, clean, sniff, rear	Supervised	
Unnamed* (Matsumoto et al. 2013)	Rats	Yes	Behavior	Rearing, head-head and head – hip contact, approach, leave, follow, mount, intromission, ejaculation	Physics engine	See supplementary material
MotionMapper (Berman et al. 2014)	Flies	Yes	Behavior	Grooming: wing, leg, abdomen, head. Wing waggle, running	Unsupervised	https://github.com/gordonberman/MotionMapper
Unnamed* (Hong et al. 2015)	Mice	If distinct	Behavior	Attack, close investigation, mounting	Supervised	

MoSeq* (Wiltschko et al. 2015)	Mice	No	Behavior	Dart, micromovement, pause, rear, walk	Unsupervised	MTA and private repository
Unnamed (Unger et al. 2017)	Mice	Yes	Behavior	Direct contact, nose to nose, anogenital sniffing, following, mating, self-grooming	Unsupervised	
DuoMouse (Arakawa et al. 2017)	Mice	Yes	Behavior	Sniffing, following, indifferent	Unsupervised	http://www.mgrl-lab.jp/DuoMouse.html
B-SOiD (Hsu & Yttri, 2019)	Mice	No	Behavior	Pause, rear, groom, sniff, orient left/right, locomotion	Unsupervised	https://github.com/YttriLab/B-SOID
SimBA (Nilsson et al., 2020)	Mice, rats	Yes	Behavior	Attack, pursuit, lateral threat, anogenital sniff, mounting, upright submissive, allogrooming, flee, scramble, boxing, approach, avoidance, drink, clean, eat, rear, walk away, circle	Supervised	https://github.com/sgoldenlab/simba
Stytra (Stih et al. 2019)	Zebrafish larvae	No	Closed loop tracking & stimulus admin. (real time)		Computer vision	https://github.com/portugueslab/stytra

ToxID (Rodriguez et al. 2017; also see ToxTrac Rodriguez et al. 2018)	Fish, ants, mice	Yes	Individual tracking in groups (real time)		Computer vision	https://sourceforge.net/projects/toxtrac/
MoST (Thanos et al. 2017)	Mice	Yes	Individual tracking in groups		Computer vision	
idtracker.ai (Romero-Ferrero et al. 2019, also see idTracker Pérez-Escudero et al. 2014)	Drosophila, fish, ants, mice	Yes	Individual tracking in groups		Unsupervised	http://idtracker.ai/
LocoMouse* (Machado et al. 2015)	Mice	No	Kinematics		Supervised	https://github.com/careylab/LocoMouse
DeepLabCut (Mathis et al. 2018)	Drosophila, mice, horses, humans, fish	Yes	Pose		Supervised	http://www.mousemotorlab.org/deeplabcut
LEAP (Pereira et al. 2019) Successor: Social LEAP Estimates Animal Pose (SLEAP)	Drosophila, mice	Yes, through SLEAP	Pose		Supervised	https://github.com/talmo/leap https://github.com/murthylab/sleap
DeepFly3d* (Günel et al. 2019)	Drosophila, humans	No	Pose		Unsupervised	https://github.com/NeLy-EPFL/DeepFly3D
DeepPoseKit (Graving et al. 2019)	Drosophila, locusts, zebras	If distinct	Pose (real time)		Supervised	https://github.com/jgraving/DeepPoseKit

Sensory Orientation Software (Gomez-Marin et al. 2012)	Drosophila, mice, fish	No	Posture & trajectory (real time)		Computer vision	https://sourceforge.net/projects/sos-track/

Table 2.3. Operational definitions of behaviors

Classifier	Description	Start frame	Duration of behavior	End frame
Attack	Clear physical antagonistic interaction initiated by the Resident mouse. Brief pauses can occur (equal or less than 133ms, or 4 frames at 30fps) as long as the Resident mouse remains oriented toward the Intruder.	First frame when the Resident mouse makes physical antagonistic contact with the Intruder. Typically, this is characterized by outstretched Resident forepaw(s) contacting the Intruder, while the Resident has an open mouth to initiate a bite. Can also be characterized by the first frame of a slap or quick bite without the forepaws being outstretched.	Short breaks (equal or less than 133ms, or 4 frames at 30fps) may be present in attack behavior, as long as Resident is still oriented toward Intruder. Attacks can include tussling, biting, boxing, and corralling as part of the attack bout.	First frame when Resident mouse orients away from Intruder. Typically, this is a slight turning of the head to look in a different direction, followed by a relaxation of the body and moving away from the Intruder.
Anogenital sniffing	Resident mouse is sniffing the anogenital region of the Intruder. Resident must be sniffing at base of tail, not further up on back or on legs.	First frame when the Resident mouse is clearly sniffing anogenital region of Intruder, rather than side, back, or leg.	Uninterrupted sniffing of anogenital region.	First frame when Resident mouse moves head away from anogenital region, either to move away from Intruder or to sniff non-anogenital region.
Lateral threat	Resident mouse is in close proximity (typically less than one body length away from the Intruder) to face of Intruder with back arched and side displayed toward Intruder. Ears are often pinned with shoulder and side of face nearest to Intruder mouse tilted slightly toward ceiling.	First frame when Resident mouse orients side to Intruder and tilts front of body toward Intruder.	Resident will often circle Intruder or move front half side to side in front of Intruder, feigning attacks prior to actual attack.	First frame when lateral threat posture is dropped. Animal will shift head away to look at other target or will begin an attack. If animal does not attack, the back posture will relax.
Pursuit	Resident mouse is following in the Intruders path as the Intruder moves away from the Resident.	First frame when the Resident is moving toward Intruder as Intruder moves away. Typically, this is characterized by the Intruder running away after an attack ends and the Resident follows directly after, or when the Intruder walks past the Resident and the Resident markedly changes directions to pursue the Intruder in the Intruders path.	The Intruder is specifically moving away from the Resident. The Intruder is not sniffing the arena, foraging, etc. Intruder typically moves in a straight line until it reaches the edge of the area, at which point it will turn to face and watch the Resident. Resident is moving along the same path as the Intruder without sniffing or foraging in the arena.	First frame when either mouse stops moving, or the Resident deviates from path of the Intruder.
Tail rattle	Resident's tail is curved at two or more points (typically in an "S" shape) and showing rapid back and forth movement.	Defined as rapid back and forth movement of tail. In Resident mice this manifests as curling of tail into "S" shape, where next frames show inversion of shape and rapid whipping back and forth of the tail.	The tail continues to move rapidly back and forth and can occasionally straighten if tail is moving fast enough. Tail side to side movement is continuous.	First frame when the curled side to side movement of the tail stops. The tail typically relaxes and straightens out.

Table 2.4. Recommended workflow and hardware specifications for performing machine classification of complex social behaviors in experimental animals using open-source software. *For more information, see the GitHub repositories of the Sam Golden lab.*

(i) **Record and pre-process videos**	Record videos at a quality where behaviors of interest are apparent to trained human observers. Crop / clip video, down-sample resolution, and re-sample frame rates, if required.	Camera allowing recordings at ≥ 30fps and resolutions ≥640x480 (i.e., standard webcam).	FFmpeg OpenCV
(ii) **Track animals**	Estimate the location of the animals and their body-parts in each frame of the videos with convolutional neural network.	CPU supported by graphical processing unit (GPU), through a cloud solution (e.g., Google Colaboratory, AWS SageMaker, Microsoft Azure) or locally (e.g., NVIDEA Rtx2080Ti).	DeepLabCut DeepPoseKit LEAP YOLO mask RCNN
(iii) **Supervised decision ensembles**	Extract feature sets, annotate behaviors, build/tune algorithms, evaluate classification performance.	Processing time benefits from higher-end CPUs and solid state drives (SSDs). We recommend a CPU with ≥8 cores (e.g., Intel i9-9900K).	scikit-learn imblearn OpenCV Shapely
(iv) **Visualize/analyze classification results**	Visually evaluate machine classifications. Perform descriptive statistics of behavior sequences, their durations and patterns.	Processing time benefits from higher-end CPUs and solid state drives (SSDs). We recommend a CPU with ≥8 cores (e.g., Intel i9-9900K).	OpenCV FFmpeg matplotlib standard python library
Note. Accurate behavior classifications require arenas and recording conditions that allow detailed animal tracking of relevant body-parts through convolutional neural network architectures (workflow Step ii). Generally, if the recording conditions allow the behaviors of interest to be reliably observed by human annotators in recorded videos, then the behaviors of interest can also be accurately annotated by machine classification techniques.			

Table 2.5. Overview of advantages and limitations of select opensource options for tracking and behavioral classification.

Technique	Open -source alternative(s)	Advantages	Limitations
Segmentation / identification networks for 'fingerprinting'	idtracker.ai	• Can track large groups of unmarked individuals • Accessible GUI	• Requires GPU support • No live tracking • Only provides trajectory of animal centroid(s) • Not validated in mouse/rat video with high amounts of aggression • Users must generate their social behavior classifiers from tracking data, which may also require additional layer of pose-estimation processing.
RFID / depth camera system	Live Mouse Tracker	• Built-in battery of social behavior classifiers • Group analysis of unmarked animals (≤ 4 mice) • No model training required • Live tracking (30Hz) • Accessible SQL data format • True longitudinal analysis (days)	• Manufacture cost (~$2k) • Manufacturing time • Dedicated platform (cannot be used to analyze historical datasets). • Arena / RFID implementation require *Animal Study Protocol* submissions / revision • Validated for mice only
Pose-estimation	DeepLabCut LEAP DeepPoseKit	• Accurate tracking of user-defined body-parts • Very large user-base • Active code development • Accessible GUI's (LEAP / DLC) • Accepts most recording environments/ and video formats.	• CNN training time • Requires GPU access • Group and live tracking features not yet available • Users generate their social behavior classifiers from tracking data
Behavioral classification	SimBA JAABA	• Accessible GUI's • User-defined social behavior classifiers • Can accept any tracking data as input	• • Requires third-party tracking (and GPU support) • No live classifications • Social classification restricted by the inability of third-party tracking solutions to track multiple similarly coat-colored animals

Note. The techniques listed above are not mutually exclusive, and the most successful behavioral analyses likely depend on synergistic implementations of disparate computational methods.

Chapter 3 : Towards the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience.

Nastacia L. Goodwin[†1,2], Simon RO Nilsson[†1], Jia Jie Choong[1,4], Sam A. Golden[1,2,3]

[1] University of Washington, Department of Biological Structure, Seattle, Washington, USA

[2] University of Washington, Graduate Program in Neuroscience, Seattle, Washington, USA

[3] University of Washington, Center of Excellence in Neurobiology of Addiction, Pain, and Emotion (NAPE), Seattle, Washington, USA

[4] University of Washington, Department of Electrical and Computer Engineering, Seattle, Washington, USA

[†] Denotes equivalent contribution

Introduction

Ethology, the study of behavior and social organization from a biological perspective, is a central tenant of modern behavioral neuroscience research. Invariably, ethological observation and analysis is plagued by the significant effort and time required to manually annotate complex behavior. These careful efforts are rewarded, however, by the ability to identify and categorize common behaviors in model species, as well as variations due to disease or experimental perturbation. This ability is an integral component of behavioral model reproducibility and extendibility. More recently, there has been a concerted effort to combine these ethological observations with machine learning-based approaches for automated behavioral classification, in combination with the study of brain function, under the umbrella of computational neuroethology (Datta et al. 2019).

The combination of automated classification with ethological analysis circumvents several issues inherent to manual annotation, namely observer drift and bias, long analysis times, and poor inter-rater-reliability both within and between research groups (Anderson and Perona 2014; Egnor and Branson 2016; Datta et al. 2019). Frame-by-frame behavioral analysis also allows for the acquisition of behavioral annotation with sampling frequencies and accuracies that match modern neural recording and manipulation techniques. Thus far, machine learning algorithms have uncovered previously unknown behavioral repertoires in model organisms and have confirmed and extended foundational assumptions of ethology through the incorporation of neural data (Vogelstein et al. 2014; Wiltschko et al. 2015; Rudolf et al. 2019).

With the increased adoption of new open-source platforms for automated behavioral classification, machine learning techniques are now accessible to a wide array of behavioral neuroscience labs despite variations in budget and computational experience. A major advantage of these machine learning approaches is the identification of behavioral latent motifs - the individual behavioral components that make up a gestalt behavioral definition. The machine learning models that predict these motifs and behaviors can be shared between and iterated upon by research groups to develop a common understanding of behavior. In this current opinion we focus on three main consequences of this last statement, and their applications towards behavioral neuroscience (Fig. 1).

First, despite repeated calls by researchers for more transparent and operationalized reporting (Landis et al. 2012, 2016a) and an equally strong edict by the National Institutes of

Health (NIH) in the form of enhanced reproducibility through rigor and transparency (NIH 2015), behavioral neuroscience has yet to enter a 'golden era' of behavioral definition standardization. Pragmatically, it is common for observational behavioral metrics to be reported as loose subjective definitions within a sentence or two in the methods. More recently, predominantly because of journals adopting the NIH rigor and reproducibility standards, observational behavioral metrics are reported with more objective operationalized definitions including information regarding a behavior's timing, transitions, and constraints**.** Despite these efforts, however, the challenges of manual behavioral annotation inherently prevent consistent use and transfer of operationalized behavior definitions. Here, we propose that the widespread adoption of machine learning-based automated behavior classification will circumvent issues inherent to manual annotation, allowing for more standardized behavioral definitions both within and across research groups. More specifically, we highlight the importance of developing and reporting explainability metrics when using these tools.

Second, while the standardization of assays and behavioral definitions is considered integral to reproducibility, studying underlying mechanisms often and seemingly paradoxically requires deviation from standardized behavioral procedures. Indeed, with the use of increasingly complex neural recording and manipulation methods, there is an equal need to repeat that "the neural basis of behavior cannot be properly characterized without first allowing for independent detailed study of the behavior itself" (Krakauer et al. 2017). While it is pragmatically attractive to operationally define a standardized behavior as a proxy for its neural correlates and control, this can introduce bias when expanding a study population. The need for extensions of standardized behaviors are highlighted by recent observations considering sex as a biological variable in stress-related procedures, which have revealed for example, that female rats demonstrate "darting", a novel active fear coping behavior not typically seen in males (Gruene et al. 2015), while female California mice show increased social vigilance (Greenberg et al. 2014). Similarly, driven by technological considerations, the transition from head-fixed neural recording to freely moving neural recording has identified divergent mechanisms between the same constrained or unconstrainted behaviors (Meyer et al. 2020). Questioned directly, as neural data increases in fidelity and size and as behavioral procedures become more specialized, how can the field of behavioral neuroscience ensure accurate reporting of these behavioral procedures in a way that ensures reproducibility? The need for behavioral specialization is at odds with the

movement towards more generalized standardization - but aligns well with approaches in machine learning explainability.

Third, to negate the above concerns, significant effort has been spent developing unsupervised approaches for behavioral analysis that may be especially suited for identifying previously unnoticed behavioral motifs in model species. However, these unsupervised models are often opaque, and users cannot readily determine the reasoning that guides their decisions. In commercial sectors, such a lack of transparency is unacceptable by both end users and regulatory agencies. To retain scientific rigor and reproducibility, behavioral neuroscience should be at the forefront of demanding transparency via explainable models when possible, or via post-hoc explainability methods when black-box models cannot be avoided.

Here, we propose the adoption of Shapley values using Shapley Additive Explanations (SHAP) as an open-source resource for the explainability of human annotation as well as supervised and unsupervised behavioral classification results in preclinical behavioral neuroscience. We present several examples of the utility for using Shapley values within the context of the above three concerns. Shapley values are, of course, only one possible solution and we provide a summary of alternatives that can be used when adopting both supervised and unsupervised machine learning into the analytical portfolio of behavioral neuroscience labs.

A brief machine learning primer

Prior to diving into the details of explainability, transparency, and universality, a brief primer on machine learning provides valuable context. Firstly, critically, accurate behavioral classification requires equally accurate and precise animal tracking data. Typically, frame-by-frame positional data is obtained via 2D or 3D pose estimation (Mathis et al. 2018b; Graving et al. 2019a; Pereira et al. 2019a; Dunn et al. 2021; Karashchuk et al. 2021a), algorithms such as optical flow (Bohnslav et al. 2021), or histograms of oriented gradients (Dolensek et al. 2020). The fields of pose estimation and video tracking are rapidly expanding and beyond the scope of this article (but see (Mathis and Mathis 2019b; Pereira et al. 2020) for excellent reviews on the state of the field). Ultimately, it is important for users to understand that with any machine learning application, the quality of the input (tracking and video annotations) directly impacts the quality of the output. The adage “garbage in, garbage out” is especially accurate in this case. Significant improvements to accessibility and use of open-source software have occurred in the last several years ensuring that users, with care and validation, can generate consistently

acceptable pose-estimation models using open-source tools. Once quality tracking has been obtained, users can move on to using machine learning algorithms that typically fall into supervised or unsupervised classes, which can be used entirely separately or in combination for different analytical purposes.

Supervised machine learning techniques are often sufficient for basic behavioral analysis, and can more simply tell researchers when and how often a predefined behavior of interest is occurring. For each individual frame of a video, users provide the algorithm with (*i*) many calculated features (i.e., nose to nose distance), and (*ii*) the "answer" as to whether the behavior of interest is occurring. Different supervised methods such as support vector machines, random forests, and gradient boosting machines then look for mathematical rules which best delineate the features of negative versus positive frames. For more detailed explanations of their application, there are numerous reviews available (Anderson and Perona 2014; Datta et al. 2019; Goodwin et al. 2020a).

When using unsupervised techniques, users provide no "answer" for the algorithm, but rather set rules for similarities and differences within the data. The art of tuning these hyperparameters can be difficult, and generally requires a deeper knowledge of statistics and the models being used. When wielded correctly, unsupervised algorithms output different clusters (which can be adjusted to consist of individual frames or entire behavioral bouts). Researchers then select and look at several examples from each cluster to try to establish definitions that differentiate the clusters behaviorally. Importantly, users must be careful to tune hyperparameters based on objective metrics like realistic physiological parameters, to prevent any biases regarding expected behaviors and predictions that fall outside of explainable reasoning.

These efforts are championed by the burgeoning field of Explainable Artificial Intelligence (XAI), which calls for the use of explainable methods for both high- and low-stakes decision making. Succinctly, XAI proposes that if a model lacks validation, or if a greater understanding of how a model works is required, XAI provides tools to let you do so. Recent reviews provide detailed explanations of the predominant concepts and challenges within the XAI field (Doshi-Velez and Kim 2017; Vu et al. 2018; Rudin 2019; Lundberg et al. 2020; Das and Rad 2020; Shahroudnejad 2021). Here, we will focus our discussion on explainability methods that are either intrinsic (built into the model) or applied post-hoc, and we define

"explainable" as how readily users can understand the relationship between feature values and the model's decisions.

Critically, the principles of behavioral neuroscience are not changed by the incorporation of these concepts and techniques. While most of these algorithms were not designed with the field in mind, great strides have been made to modify canonical machine learning tools for use in ethology. Common principles of behavioral design remain – training sets should consist of individuals from each experimental group, and training a classifier on the behavior of a very small number of animals will not provide a generalizable model. Just as machine learning algorithms rely on good tracking data, they also rely on solid experimental design with construct and predictive validity towards the biology they are examining. More specifically and pragmatically, as more labs begin adopting machine learning tools for behavioral classification, we note the opportunity to set a precedent for the incorporation of accessible explainability metrics.

Looks can be deceiving

A scientific approach solicits face, construct, and predictive validity within its methods and results (Markou et al. 2009). In pose-estimation and behavior classification, face and predictive validity are immediately achieved through intuitive visualizations with overlays and other performance metrics, while construct validity typically isn't a requisite. Construct validity is essential, however, as the models and human observers may study different features to reach the same conclusion. For example, behavioral classification models can be sensitive to minor perturbations in experimental set-ups between training and testing (changes in animal sizes, age, sex, bedding material, presence/absence of experimental equipment such as head stages, wires, and operant modules). Explainability metrics can be used as a diagnostic tool to identify confounding variables and avoid such pitfalls a priori.

Tools for transparency and explainability

The use of XAI is rapidly expanding through the inclusion of both inherently explainable algorithms (Rudin 2019) and post-hoc explainability methods(1953; Lundberg and Lee 2017). Here, we support the use of post-hoc explainability methods for computational neuroethology in behavioral neuroscience due to their relative simplicity of use, flexibility, and generalizability to diverse platforms. Primarily, despite a decade of calls for increased training, only ~15% of

neuroscience graduate programs in the US require computational coding courses (Pevzner and Shamir 2009; Grisham et al. 2016; Society for Neuroscience 2016; Goldman and Fee 2017; Juavinett 2022). Within the field of behavioral neuroscience, this generally functionally translates into the development of machine learning platforms that are lab specific. Second, once a platform is created and has face validity there is little incentive – or funding support - to further develop and incorporate explainable algorithms. Optimally, newly developed platforms should be developed with inherent explainability in mind. Functionally, incorporation of post-hoc explainability methods that do not require de-novo platform development are more likely to be widely adopted within the field.

Explainability methods aim to provide accurate and verbalizable accounts for how known feature values contribute to model decisions. The output from explainability calculations enable researchers to communicate and compare the results of disparate classifiers and supports experimenters in making informed decisions for machine model implementation and use. Interpretability is an active and emergent research area, and several initiatives in academic and commercial sectors are focused on devising new algorithms. These initiatives focus primarily on explaining decisions, promoting human/AI interactions, and the user's ability to evaluate and improve the trustworthiness and generalizability of different models.

Explainability methods can be deterministic or have gradations of stochasticity, and provide local (e.g., individual video frames) and/or global explanations (aggregate feature weights within a model). The methods may be algorithm-specific or agnostic, and be intrinsic to the model or require additional processing. The core methodologies of most interpretability approaches will nevertheless seem familiar to experimentalists: features undergo some class of iterative ablation or permutations, the change in performance after such manipulation is evaluated, and the manipulation's impact is succinctly summarized (Covert et al. 2021). How the permutation is performed, evaluated, and summarized differs between methods with varying relatedness to human processes and intuition for complex outcomes (Miller 2019).

Although the advancement of behavior research may necessitate divergent experimental protocols and computational methods, any practical explainability methods obligate a certain level of cross-lab standardization. The most meaningful explainability metrics would also provide intuitive information on strength, directionality, and potentially linear and non-linear relationships between the known feature values and model decisions at the level of individual

observations. Meaningful explanations in behavioral science also strongly benefit from methods that are ubiquitously employed, transparent and open source. Accessible explainability methods also compute features contributions in ways that agree with human intuition. Recent XAI methods achieve this by exploiting solution concepts from collaborative game theory, using axioms defined by *fair* and *rational* payoffs in N-person games (Covert et al. 2021).The axioms vary by solution concept, but some are more desirable and generally accepted within the context of machine learning classification. For example, it may be desirable that:

(i) Features with identical effects on classification probabilities are assigned identical importance scores (symmetry),

(ii) Features without effect on classification probabilities are allocated zero importance (null or dummy features),

(iii) The contribution of a feature within an ensemble model should equal the sum of its contributions to each model in the ensemble (additivity)

(iv) The grand total of feature contributions should equal the grand total of the model output (efficiency).

The Shapley value is a solution concept that uniquely satisfies these rules (Osborne and Rubinstein 1994) and is therefore an appealing tool for understanding and presenting explainability metrics within machine learning applications.

In brief, the Shapley value is a coalition game theoretical approach for fairly distributing the proceeds of a game to its players. To understand the contribution of feature X in a machine learning model, the output of a model involving features S is subtracted from the output of a model involving features $S + X$. This calculation is subsequently repeated for all possible permutations of features in the model. Feature X is subsequently allocated the proceeds according to its contribution within each model permutation. The Shapley value holds theoretical properties that can make interpretations readily understandable, including additivity among features in relation to a model's prediction. This is known as the Shapley additivity axiom.

Shapley values are a method for presenting intuitive explainability metrics while maintaining agnosticism to the model's structure. Its use in machine learning, however, has been prohibited by its computational costs; calculating Shapley values for all feature permutations necessitates exponential time and unfeasible runtimes for even smaller dataset. However, through a suite of approximation algorithms, the SHAP (Shapley Additive Explanations) library

(Lundberg et al. 2020) performs Shapley value calculations using model agonistic approaches (KernelSHAP), tree-based models (TreeSHAP (Lundberg et al. 2019)), linear models (LinearSHAP), deep neural networks (DeepSHAP (Lundberg and Lee 2017)) in polynomial time making it accessible to machine learning applications and behavioral neuroscience. The optimal SHAP algorithm is determined by the input model; TreeSHAP, LinearSHAP and DeepSHAP are optimized for decision trees, linear model, and neural networks, respectively, while KernelSHAP remain model agnostic.

Shapley values in behavioral analysis

Computational developments (McKinney 2011; Lam et al. 2015, 2018) have made it possible to efficiently calculate comprehensive feature batteries for accurate and generic machine learning pipelines in behavioral neuroscience. Within such generic applications, feature batteries are comprised of hundreds of keypoint-to-keypoint metric distances, angles, and movement velocities and path tortuosities in various temporal and spatial windows. At this resolution, individual feature values have limited relevance for explaining human annotator conduct and classification probabilities, and presenting feature explainability metrics for each individual feature does not make the model's decisions explainable.

Here, the Shapley value additivity axiom can help make model decisions more explainable (Fig. 2). Prior to the evaluation of individual feature contributions for a binary behavioral classifier, there are two known values. First, we know the average probability (*expected value*) that the behavior is occurring in a video frame. For example, the expected value is 5% if the behavior was annotated as present in 5% of the video frames used to create the classifier. Second, we also know the outcome prediction for an individual frame. For example, the model predicts that the behavior is present, in a previously unseen frame, with 85% probability. When using TreeSHAP and Shapley values, the difference between these two numbers (the *outcome probability* minus the *expected value* equals 80%) is attributed to the combination of feature values, and this difference value is distributed amongst the features according to their contribution within the ensemble. The Shapley value output in this example is a vector which length equals the number of features, and which sum equals 80.

To increase interpretability and universality of Shapley values representing a given supervised classification model, we can collapse Shapley value contributions of features that measure similar, general, and often colinear characteristics of the behavior of interest into interpretable biologically based categories while maintaining consistency and accuracy. To understand random forest classifiers, we propose the aggregation of Shapley values, for example derived from the open-source TreeSHAP package, into different behavioral classes and sampling frequency sub-classes (Fig. 3). In Figure 3, for example, we have selected classes and sampling frequencies for explaining supervised predictions of complex social behavior in two interacting mice. Ultimately, these classes and sampling frequencies are dictated by the selection of model organism and behaviors of interest. The opportunity to bin feature importance scores into user-defined superordinate classes - while maintaining explainability - is a result of the mathematically proven Shapley value additivity and efficiency axioms (Osborne and Rubinstein 1994). These axioms ensure that whichever Shapley values are collapsed, their aggregated grand total is a valid and interpretable representation of their combined contribution within the classification scenario. This attribute is appealing for behavioral neuroscience use-cases, and other experimental scenarios, that require a high-degree of flexibility.

Importantly, the use of these categories ensures that classifiers that target the same behavior, but use different feature sets, can be directly compared. A common reason that classifiers use different feature sets is a consequence of the initial pose-estimation scheme selected by experimenters. For example, two research groups interested in the same behavior may use pose-estimation schemes that identify 5 versus 8 body parts, respectively, due to their experimental needs; regardless, by binning their classification features within the same biologically determined classes and sampling frequencies, the use of Shapley values allows for direct comparisons between their classifiers.

We argue that explainability scores, such as Shapley values, can address several general concerns pertaining both supervised machine classifications and human annotations in behavioral neuroscience settings.

(i) As Shapley values can be used in a generalized manner, without needing identical pose-estimation key-points or feature sets, they provide a quantifiable, rather than qualitative, description of an operationalized behavioral definition that can be used across labs, institutes, and annotators.

(ii) Shapley values may reveal construct validity issues, where machine models produce accurate classifications through features that are independent correlates of the true features of interest.

(iii) Shapley values are widely accepted, understood, and under continual development and scrutiny in the greater computer science and AI fields.

In addition to its utility for supervised classification explainability, Shapley values can also be employed a priori or post-hoc in unsupervised learning contexts to investigate or explain latent structures within embeddings vectors (Fig. 4). For example, to understand the discriminating features of different behavioral clusters in dimensionality reduced space, the cluster assignments can function as target variables for supervised classifiers. For example, the cluster "A" classifier dataset would contain all bouts for clusters A, and B, with cluster A bouts annotated as positive, and all other bouts annotated as negative. Users can then objectively compare which features drive the classification of a behavior as A or B. These supervised classifiers are straightforward to make with current open-source tools and are optimal for Shapley value calculations using the TreeSHAP library. Shapley values are then calculated for each cluster, and this analysis gives quantitative data of cross-cluster differences that can support and supplement qualitative visual analysis of the different behavioral events. Conversely, Shapley values have also been used as input to dimensionality reduction and clustering algorithms which, as opposed to raw feature values, intrinsically represent the importance of the features for the target variable and may thus produce more insightful cluster representations (Lundberg et al. 2020). A further caveat in unsupervised analysis of raw feature values is the independent scales across the different feature. This shortcoming is addressed when clustering Shapley values as all features are standardized to a single unit scale, which is the probability of the behavior (Lundberg et al. 2020).

Alternatives to Shapley values

Shapley values are one paradigm amongst a continuously growing number of approaches for XAI. The fundamental priority is not choosing a specific algorithm, but rather that such algorithms become regularly used in behavioral neuroscience as machine learning approaches to behavior analysis become more common. Here, we have promoted the use of Shapley values because of its wide use (Lundberg and Lee 2017) and highly active development (Lundberg 2022), in conjunction with its relative ease of implementation and generalizability of post-hoc

explainability to machine learning platforms in behavioral neuroscience. However, promising methods that address annotator differences within inherently interpretable models through program synthesis and wavelets are also being developed and are likely to significantly enhance inter-annotator model explainability (Sun et al. 2021b; Tjandrasuwita et al. 2021).

Currently, popular and computationally inexpensive methods for interpretability in classification scenarios are aggregate gini impurity and entropy measures that provide information gain following inclusion of the feature into decision ensembles. These measures are intrinsically used to create decision trees and therefore often calculated by default. Most prevalent machine learning libraries provide aggregate feature importance's through gini impurity and entropy measures, highlighting their popularity and common use (*Scikit*: (Pedregosa et al., 2011); MLlib (Meng et al.); *Xgboost:* (Chen and Guestrin 2016); Tensorflow (Abadi et al. 2016)).

Conversely, permutation importance calculations provide a global explainability measure by calculating information loss following the scrambling of features (Breiman 2001). LIME is another prevalent approach that provides local explainability metrics by fitting inherently interpretable linear regression models on perturbations around the original feature values (Ribeiro et al. 2016a). LIME is closely related to SHAP, but does not satisfy game theory axioms [32]. Another increasingly popular approach for local explainability that explores causality use counterfactuals. Here, feature values are titrated and a causal relationship between the feature value and the outcome is concluded when concomitant changes in model decisions are observed (Verma et al. 2020).

Inherently explainable algorithms also can provide increased explainability for machine learning applications (Rudin 2019). The conceptual definitions of "inherently explainable" vary, but typically involve low-dimensional datasets (i.e., few features) that are analyzed using decomposable algorithms that can be verbalized without further advanced statistical analysis (Lipton 2017).. Notably, a decision tree algorithm can be accurate and considered "inherently explainable" if carefully crafted using a reduced number of select and highly predictive features. Select features could be identified using experimenter domain-knowledge, or through pre-processing using advanced statistical and machine learning methods (Sun et al. 2021b).The inclusion of inherently explainable models should be a goal for developers. However, we

propose a more general use of post-hoc explainability for behavioral classification for several pragmatic reasons:

(i) The output of prevalent machine learning toolkits for behavior analysis (Kabra et al. 2013; Wiltschko et al. 2015; Hsu and Yttri 2019; Nilsson et al. 2020; Luxem et al. 2020; Graving and Couzin 2020) are all directly amenable to post-hoc explainability methods. These tools, however, can currently not be used out-of-box to produce inherently explainable models.

(ii) A carefully curated feature-set can provide inherent interpretability for supervised applications. However, the same curated feature-set could bias and obstruct detection of experimenter-undefined novel behaviors in unsupervised applications. This disconnect could be addressed by using different feature-sets for supervised versus unsupervised learning.

(iii) Two inherently explainable machine learning models, that classify different behaviors, may not be directly compared as they will rely on different input data and feature sets.

Notably - although smaller feature-sets can produce both accurate and inherently explainable models - this does not equal computational savings for behavioral neuroscience use-cases. Many features computed by generic classification tools (Kabra et al. 2013; Nilsson et al. 2020) are not necessary for accurate classification, but are required for post-hoc applications such as unsupervised learning and aggregate and conditional descriptive statistics. Thus, although fewer features are used by inherently explainable models, the users of inherently explainable models may still have to perform the same extensive calculations as those used by less transparent approaches.

Future Directions

Ongoing efforts aim to develop, catalog, and promote XAI for deep learning and other machine learning methods (Das and Rad 2020; Shahroudnejad 2021) .While such work tackles technical topics of comprehensive importance, our own efforts explore the potential of well-developed XAI methods for practicable and wide use in behavioral neuroscience experiments. New XAI methods are developed and documented at pace, and the behavioral neuroscience field should be ready to adapt as superior and accessible methods become available.

We have focused on supervised applications, but there is a pressing demand to explain the biological relevance of behaviors detected through unsupervised algorithms. Unsupervised

applications parse behavior with reduced or no experimenter oversight or ground-truth, and labels are then assigned by experimenters post-hoc (Hsu and Yttri 2019; Luxem et al. 2020; Graving and Couzin 2020). The assignment of behavior labels may depend on qualitative assessments through visualizations and inspection of a reduced number of features. Future unsupervised applications would benefit from accurate, in-depth, and verbalizable statistical explanations for why the algorithm dissociated the clustered behaviors and what each clustered behavior represents, in addition to the qualitative appraisals and metrics on their validity (Moulavi et al. 2014). This will enhance experimenter confidence in the algorithms and increase their validity and reproducibility within the greater behavioral neuroscience community.

Another ongoing direction, and area of active development, for both pose estimation pipelines and behavioral classification (DLStream (Lab 2022); DLC (Kane et al. 2020);Google ML Kit (2018), Tensorflow Lite (Abadi et al., 2015)) are real-time applications. Real-time behavioral classification, based on real-time pose-estimation or other behavioral tracking, is a critical component for impending closed-loop recording and stimulation platforms that predict current and future behaviors. Within such applications, an in-depth understanding of the workings of real-time classifiers, partly through explainability measures, is critical as experiments depend on one-shot inference without opportunities for post-hoc re-analysis or model retraining. Further, from the technical perspective of computational expense during real-time predictions, enhanced explainability metrics allow for the winnowing of unneeded features without the loss of needed predictive power.

Lastly, supervised classification scenarios are becoming more complex and varied as they are more widely adopted. Future supervised classification scenarios will likely involve shifting multi-animal, multi-object environments, as well as user-defined spatial regions of interest and experimental subjects that share partly overlapping features (age, sex, strain) (Winters et al. 2022a).To understand the generalizability of machine learning models in such complex, less stable settings, enhanced explainability may help.

Conclusion

The use of explainability metrics in combination with both supervised and unsupervised methods have clear implications for the transparency and universality of behavioral classification across research groups. Explainability metrics, if reported when using machine learning

classification, provide discrete and quantifiable descriptions of behavior and remove the subjectivity of standard operationalized definitions. Shapley values, as described, provide a simple and effective metric that is easy to understand and apply in non-specialized settings.

Moving forward, for this purpose, efforts must be made to establish portals like the Research Resource Identification (RRID) database (Bandrowski et al. 2015). RRID reporting is now widely accepted, and often required for publication, due to its alignment with the Materials Design Analysis Reporting (MDAR) Checklist for Authors (Chambers et al. 2019) that sets the minimum reporting criteria for studies in the life sciences and is endorsed by major journal publishers. Already spanning antibodies, reagents, model organisms, and even software, it is equally important to begin reporting behavioral classification metrics.

Early efforts to organize such a system have been led by open-source organizations such as the OpenBehavior (White et al. 2019) project, which provides a logistical framework to begin developing and supporting the hosting and dissemination of behavior-centric open-source pipelines. Alternatively, organizations like The Jackson Laboratory (JAX), already world-leaders in genetic databasing across inbred and outbred mice strains, have also turned their attention to behavioral phenotyping. The Mouse Phenome Database (MPD) provides primary phenotype data for the laboratory mouse and provides tools to analyze and visualize these data in relationship to JAX genetic databases (Bogue et al. 2020). Recently, the MPD has housed data generated using unsupervised analysis of grooming behavior (Geuther et al. 2021). Platforms like OpenBehavior.org and MPD are primed to not only host this behavioral analysis, but as machine learning becomes more common, also the provide access to the explainability metrics paired with these data.

Open-source initiatives now enable most behavioral neuroscience labs to perform automated and accurate detection of both experimenter-defined and novel behaviors in video-recordings. How these algorithms produce their predictions are typically not quantified, leading to a lack of transparency for cross-lab comparisons and machine model construct validity. Parallel strides in AI explainability have produced open-source algorithms that permit accurate and stable calculations of detailed contributions of individual features and their interactions for machine model decisions at the level of individual observations. These explainability algorithms have begun to be incorporated into many tools aimed at behavioral neuroscientists (Mathis et al.

2018b; Nilsson et al. 2020) and we anticipate that metrics based on Shapley values or other post-hoc explainability metrics will be an important part of these efforts.

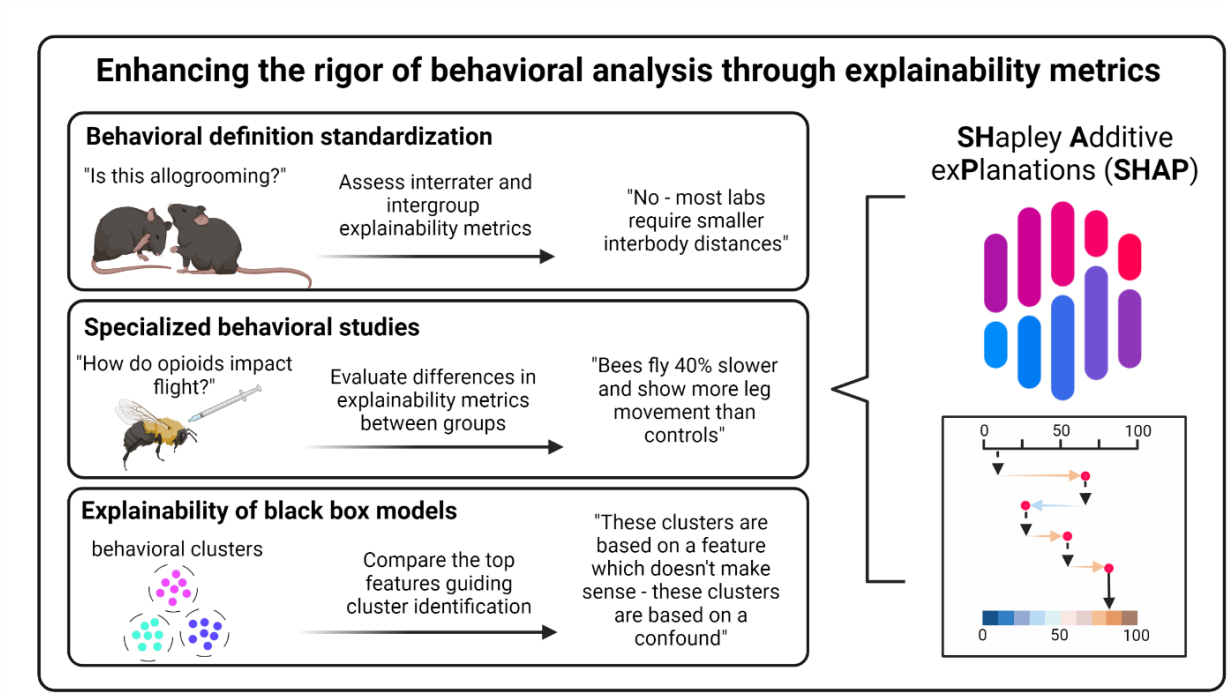


Figure 3.1. Enhancing the rigor of behavioral analysis through explainability metrics. *SHapley Additive exPlanations (SHAP) is an opensource explainability platform under continuous development by the AI field. SHAP has several advantages for computational neuroethology, including promoting standardized behavioral definitions, providing objective and specific metrics of changes in behavior between groups, and in increasing the transparency of unsupervised clustering results.*

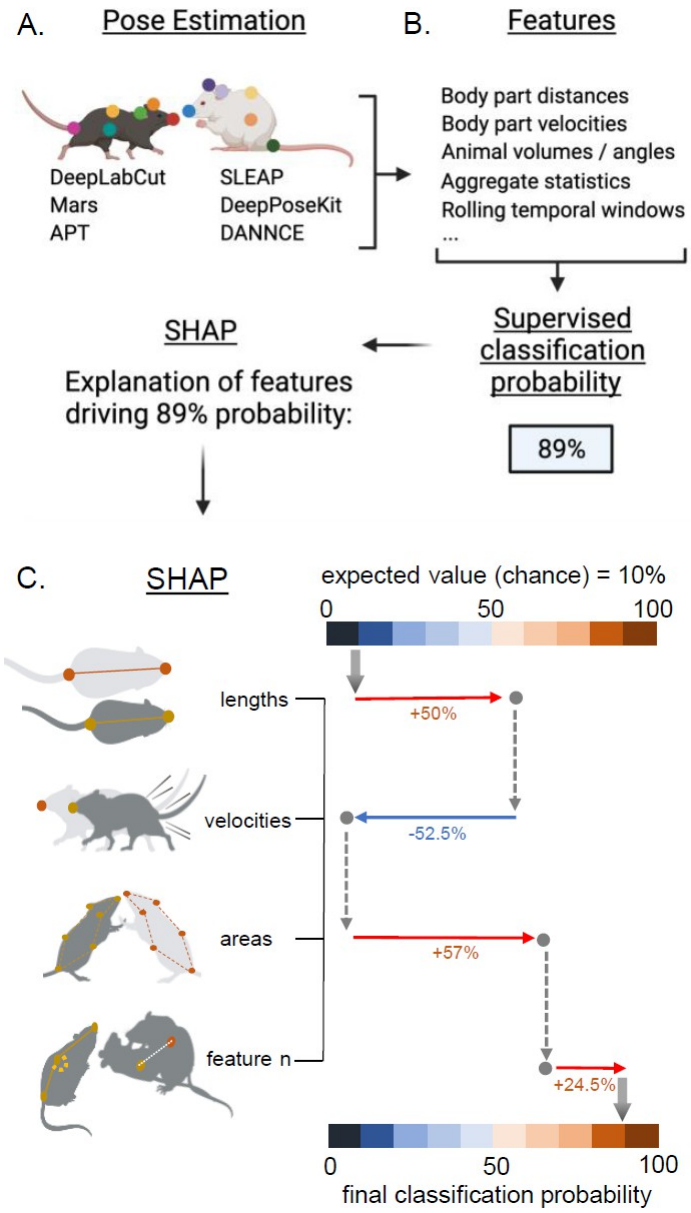


Figure 3.2. Shapley value explainable artificial intelligence pipeline for behavioral neuroscience*. (A) Many automated behavior analysis platforms rely on pose estimation data from raw video as input, followed by the calculation of feature batteries based upon the position of animals within and across frames. (B) These feature batteries are used as input to supervised machine learning algorithms, which output a probability score of the behavior of interest occurring in a specific video frame. (C) The contribution of the different features values towards the output probability score can be decomposed post-hoc into a verbalizable description. For example, features measuring animal lengths may contribute to an increase in the behavior classification probability, while features measuring animal velocities contributes to a decrease the classification probability of the same behavior. Shapley values ensures that the combined description of all feature contributions is accurate, rational, and related to the final classification probability through its additivity and efficiency axioms.*

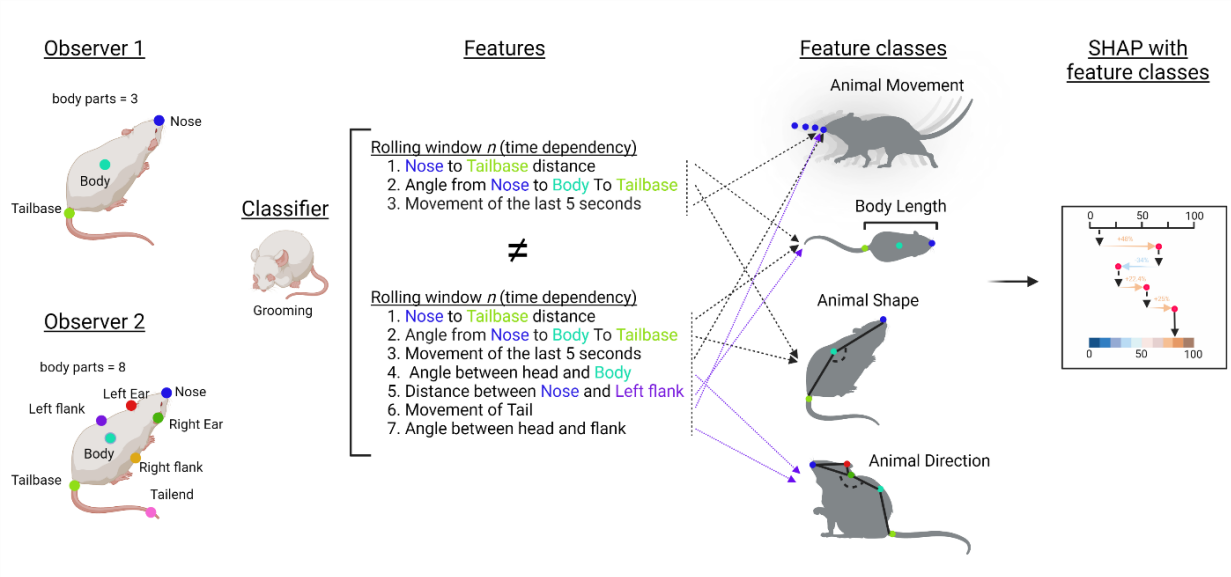


Figure 3.3. Adapting SHAP to behavioral neuroscience. *Standardization in behavioral neuroscience remains elusive, and labs may differ in pose estimation techniques as well as specific features that they calculate per video frame. Despite these differences, binning features into broader feature categories allows for the direct comparison of these classifiers across labs.*

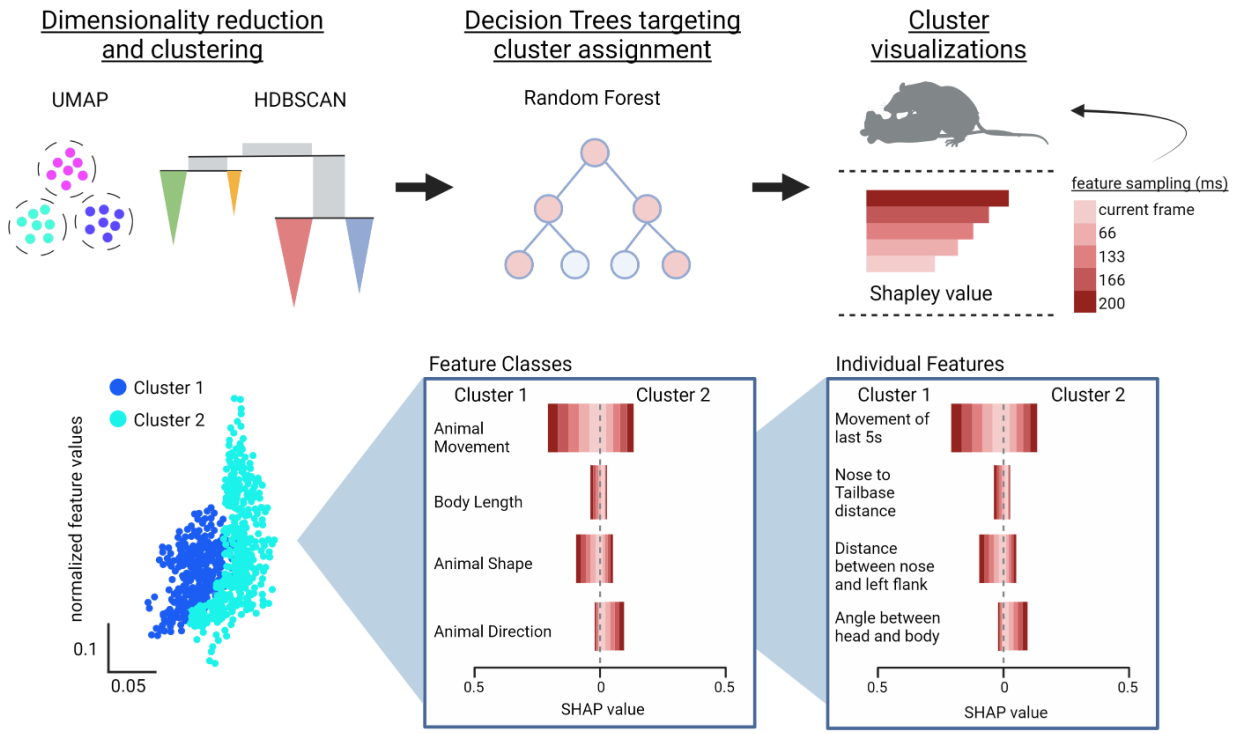


Figure 3.4. Supervising the unsupervised*. Current methods for understanding unsupervised clusters are subjective – namely watching clustered bouts and trying to find and describe the differences. SHAP provides the opportunity for an objective description of feature differences contributing to cluster alignment via the creation and analysis of supervised classifiers for each cluster.*

Table 3.1. Definitions of interest (indicated with red in main text)

Term	Description
Coalition game	Coalition game theory deals with situations where the outcome is known, and rules dictate how actors interacted to achieve the outcome, and how the proceeds of the game should be divided amongst the actors. In the context of machine learning, the outcome is the models' decision, and the actors are the features.
Counterfactual	The minimum change within a feature value that produce an alternative decision by a machine learning model
Deterministic vs stochastic	Stochastic algorithms produce different results when implemented twice on the same dataset. Deterministic algorithms give the same results when implemented twice on the same dataset. Stochasticity is favorable on larger datasets and is present in most prevalent machine learning techniques in behavioral neuroscience (e.g., decision trees, deep neural networks). Algorithm stochasticity can typically be reduced by increasing the number of observations (i.e., behavior events). Stochastic methods may be considered disadvantages in regulated settings, as the same algorithm can produce different outcomes for the same observation when implemented twice.
Ensemble model	The grouping of multiple learning algorithms. The individually weak learning algorithms are combined to create a strong or accurate behavior classification (e.g., random forest algorithm).
Entropy and Gini impurity	Measurements that evaluate the randomness of target labels after a feature split. For example, if the behavior of interest requires a minimum velocity, features that measure body-part velocities will decrease entropy and impurity and be associated and concluded to have greater importance classification.
Feature battery	Numerical measurements that together describe the activities and morphologies of the actor(s) in video frames. A battery can be composed of several hundred values that represent data from both a single video frame (e.g., the volume or rotation of an actor) and aggregate data that include preceding video frames (e.g., the velocity or path of a body-part).
Feature classes	A subgroup of features in the feature battery that measure similar activities or morphologies in the video frame. For example, two values that represent the velocities of two different body-parts both belong to a class of velocity features.
Feature value	A single metric value representing a feature in a feature battery. For example, a feature value can be the millimeter distance between two body-parts.
Interpretability calculation	A calculation which output attempts to make the machine decision processes understood by humans. An interpretability calculation can provide a numerical output that makes the decision on a single video frame interpretable (also referred to as 'local interpretability' or 'explainability'), or an aggregate output that make the decisions on all video frames interpretable (also referred to as 'global interpretability').
Model decisions	The final output of a machine learning model. A classifier decision may be presented as a single value, or a vector of values, ranging between 0-1. For example, the probabilities that a behavior is present in an image or that a body-part exists in different parts of an image are model decisions.
Permutation importance	The decrease in model performance when the values associated with a feature is randomly shuffled.
Shapley additivity axiom	Dictates that the explanation algorithms' output value relates to the decision value. Within a decision tree ensemble algorithm, the additivity axiom ensures that the grand total contribution of a feature is the sum of its contribution to each tree.

Shapley value	In machine learning, Shapley values is a method for summarizing the cooperation and contribution of the features to the model decision.
Solution concept	Formal rules within fame theory, that dictate the strategies of the players in order to produce rational game outcomes

Chapter 4 : Simple behavioral analysis (SimBA) as a platform for explainable machine learning in behavioral neuroscience

Nastacia L. Goodwin[1,2,3], Jia J. Choong[1,4], Sophia Hwang[1], Kayla Pitts[1], Liana Bloom[1], Aasiya Islam[1], Yizhe Y. Zhang[1,2,3], Eric Szelenyi[1,3], Xiaoyu Tong[11], Emily L. Newman[6], Klaus Miczek[7], Hayden R. Wright[8,9], Ryan J. McLaughlin[8,9], Neir Eshel[10], Mitra Heshmati[1,2,3,5], Simon R.O. Nilsson[†1], Sam A. Golden[†1,2,3]

[1] University of Washington, Department of Biological Structure, Seattle, Washington, USA
[2] University of Washington, Graduate Program in Neuroscience, Seattle, Washington, USA
[3] University of Washington, Center of Excellence in Neurobiology of Addiction, Pain, and Emotion (NAPE), Seattle, Washington, USA
[4] University of Washington, Department of Electrical and Computer Engineering, Seattle, Washington, USA
[5] University of Washington, Department of Anesthesiology and Pain Medicine, Seattle, Washington, USA
[6] Department of Psychiatry, Harvard Medical School, McLean Hospital, Belmont, MA 02478, USA
[7]Tufts University, Department of Psychology, Medford. Massachusetts, USA
[8] Washington State University, Department of Integrative Physiology and Neuroscience, Pullman, WA, USA
[9] Washington State University, Graduate Program in Neuroscience, Pullman, WA, USA
[10] Stanford University, Department of Psychiatry and Behavioral Sciences, Stanford, CA, USA
[11] New York University, Neuroscience Institute, New York, NY, USA

INTRODUCTION

Behavioral neuroscience requires detailed behavior(Krakauer et al. 2017), but the notoriously painstaking process of hand-annotating live or recorded assays poses a significant bottleneck preventing comprehensive behavioral analysis. The manual approach can be arduous, non-standardized, and susceptible to confounds produced by observer drift, long analysis times, and poor inter-rater-reliability (Anderson and Perona 2014; Egnor and Branson 2016; Datta et al. 2019). These caveats prevent the detailed study of complex social repertoires in larger datasets, and notably provide lower temporal resolution than most modern methodologies such as *in vivo* electrophysiological, fiber photometry, and single-cell calcium endomicroscopy recordings (Gunaydin et al. 2014; Kim et al. 2015; Ferenczi et al. 2016; Falkner et al. 2016b). Computational neuroethology[4] – the marriage of traditional neuroscience techniques, ethological observation, and machine learning – is heralded as one potential solution toward deeper behavioral analysis in more ethologically relevant settings. Furthermore, these data are collected at sampling frequencies that match modern neural recording and manipulation techniques.

The recent rapid development of open-source pipelines for markerless animal pose estimation, which allow for accurate tracking of experimenter-defined body-parts in noisy and variable environments (Mathis et al. 2018c; Graving et al. 2019b; Pereira et al. 2019b; Geuther et al. 2019), provide a framework for automated machine-learning based behavioral analyses. Using patterns in animal pose over sliding temporal windows, supervised algorithms are trained to find predefined behaviours of interest. These automated behavioral assessments often exceed human performance(Gris et al. 2017b), increase throughput and consistency(Schaefer and Claridge-Chang 2012) and reduce human bias and anthropomorphism within scoring(Robie et al. 2017). Therefore, open-source pipelines for pose estimation and behavioral analysis are increasingly focused on improving computational accessibility to non-specialists via graphical user interfaces (GUIs), easier installation processes, and extensive documentation and tutorials. But as more labs (and manuscript and grant reviewers) adopt these techniques as the *de-facto* expected standard, it is increasingly important to focus on model explainability and behavioral nuance.

Explainability methods in behavioral neuroscience(Vu et al. 2018) aim to determine why and how machine learning models are coming to conclusions, allowing researchers to (*i*) standardize behavioral definitions if desired, (*ii*) precisely describe and report specialized, non-standard,

behavioral variations, and (*iii*) more objectively quantify differences in unsupervised behavioral clusters and scrutinize their biological relevance(Goodwin et al. 2022). More precisely, computing and sharing explainability metrics is an essential step in reconceptualizing behavioral classifiers as objective and shareable reagents akin to the commonly used Research Reagent Identifiers (RRIDs) system for wet lab reagents. As researchers, we are already expected to report behavioral features such as sex, time of day of testing, light cycles, and other experimental details, yet the operational definitions of behaviors themselves are often relegated to one to two sentences in the methods. Incorporation of explainability metrics allows for objective and complete reporting of behavioral classifiers, leading to enhanced reproducibility. This is not an argument for field-wide standardization of behavioral algorithms, but rather an opportunity to precisely and objectively capture and report metrics of computer-aided behavioral analyses between experiments and research groups.

Here, we present Simple Behavioral Analysis (SimBA) and introduce accessible tools for validation and explainability of supervised behavioral classifications. SimBA is an open-source, primarily GUI-based program built in a modular fashion to increase non-specialized user access to automated behavioral analysis via supervised machine learning techniques. Two parallel and integrated branches of SimBA allow (*i*) the generation of non-machine learning based descriptive statistics of movement and region of interest (ROI) analyses and (*ii*) supervised machine learning based behavioral classification. SimBA is agnostic to the choice of animal species or number of experimental subjects and has been used to classify fish(Newton et al. 2021), wasp (Jernigan et al. 2022), moth(Dahake et al. 2022), mouse(Yamaguchi et al. 2020; Rigney et al. 2021; Murphy et al. 2021; Kwiatkowski et al. 2021; Cui et al. 2021; Chen et al. 2021; Dawson et al. 2022; Winters et al. 2022b; Neira et al. 2022, 2023; Baleisyte et al. 2022; Miczek et al. 2022; Cruz-Pereira et al. 2022; Linders et al. 2022; Hon et al. 2022; Slivicki et al. 2023; Nygaard et al. 2023; Ojanen et al. 2023), rat(Barnard et al. 2023; Lapp et al. 2023), and bird(Ausra et al. 2021) behavior. Further, this approach promotes the wider dissemination of classifiers and associated explainability metrics between research groups, and is compatible with new or historical videos annotated by open-source packages such as BORIS(Friard and Gamba 2016) and commercial packages like Noldus Observer or EthovisionXT(Spink et al. 2001).

Importantly, SimBA introduces several machine learning interpretability tools and seamlessly incorporates the application of Shapley Additive exPlanations(Lundberg 2022) (SHAP) scores,

which are one possible solution to providing explainability and transparency of behavioral classifiers. SHAP is a widely cited open-source and post-hoc explainability method which can be applied and compared across any of the current machine learning platforms regardless of pose estimation scheme. Furthermore, Shapley values are widely accepted, understood, and under continual development and scrutiny in the greater computer science and artificial intelligence fields. It is proposed that explainable AI is critical for the future of neuroscience(Vu et al. 2018), and for a deeper discussion into the usefulness of explainability approaches in computational neuroethology see Goodwin, Nilsson et al., 2022(Goodwin et al. 2022).

Here, we highlight the functionality and importance of this approach in multiple datasets to demonstrate the utility of explainability metrics in capturing and describing subtle behavioral variations across sex and environment in freely moving social behavior. Specifically, we examine five behaviors (attack, anogenital sniffing, pursuit, escape, and defensive behavior) across males and females in chronic social defeat stress (CSDS) assays, and across male rodents in varied contextual environments across resident intruder or CSDS assays. To understand the relationship between such behavioral differences between laboratories, we perform SHAP analysis on aggressive social behaviors from different research groups and compare the specific timescales and feature bins characterizing behaviors. We show that this approach provides quantitative descriptions of behavior, allowing for the use of behavior as an objective shareable reagent.

We present a platform for the rapid frame-by-frame supervised analysis of animal behavior, in conjunction with explainability tools which rapidly and accessibly capture behavioral subtleties often left out of standard behavioral metrics. We propose that behavioral classifiers, in combinations explainability metrics, can be publicly shared and used in the fashion of RRIDs to increase reportability and reproducibility within behavioral neuroscience.

Results:

Accessible machine learning for behavioral neuroscientists

SimBA provides accessible machine learning tools to non-specialized users using standard computational hardware via a simple, single-line installation, a graphical user interface (GUI), and extensive documentation. Following raw video acquisition (Fig. 1), users can pre-process their videos (ex. cropping, trimming, changing resolution and contrast) in SimBA prior to

performing pose estimation in their open-source program of choice such as DeepLabCut(Mathis et al. 2018c; Lauer et al. 2021), SLEAP(Pereira et al. 2022) DeepPoseKit(Graving et al. 2019b), and MARS(Segalin et al. 2021). Following GUI-guided import of animal tracking data (Fig. 1A) , SimBA calculates relationships between body parts across static and dynamic time windows (*features*) that are used to train supervised random forest machine learning classifiers for behavioral predictions (Fig. 1B). SimBA – by default – computes explainable feature representations of movements, angles, paths, velocities, distances, and sizes within individual frames and as rolling time-window aggregates. To provide flexibility for advanced users, the SimBA library includes a larger battery of runtime-optimized feature calculators covering frequentist and circular statistics, anomaly scores, temporal and spectral analyses, relationships between pose-estimation and user-defined regions-of-interests, bounding-box methods and other ML distribution comparison techniques(Newton et al. 2021; Winters et al. 2022b; Lapp et al. 2023). All of these can be deployed within user-defined time-windows in use-cases where default feature calculators are insufficient. Finally, SimBA is highly flexible and can function as a programmatic platform allowing fully customized user-composed feature extraction classes. These hundreds of features per individual video frame can then be used to either train supervised machine learning classifiers or fed to trained classifiers which create frame-by-frame predictions of the probability of a behavior of interest occurring.

Training supervised machine learning algorithms requires human annotations of a subset of video frames as either positive or negative for the behavior of interest, which algorithms learn to differentiate using the associated feature values. We have streamlined the training set construction process by building in-line behavioral annotation tools including raw video annotation, machine-assisted annotation where the user verifies annotations produced by a prior behavioral model, and methods for importing existing behavioral labels (Fig. 1B). This process is supported by video batch post-processing tools that can be used to create targeted video clips containing the behaviors of interest that maximize biological replicates and contain comparable amounts of positive and negative frames (Fig. 2C), precluding the need to annotate every frame of individual videos. Due to their ease of adoption for new users, interpretability, and robustness to overfitting(Breiman 2001; Liaw and Wiener 2002; Goodwin et al. 2020b), our pipeline uses random forest supervised machine learning algorithms; this further allows for the calculation of

classical machine learning performance metrics and supports the creation of multiple visualizations and other hands-on validation tools for individual classifiers (Fig. 1B).

Classifier construction and performance

In addition to accommodating shared classifiers and pooled annotation sets, pre-existing classifiers can be adapted to new experimental cohorts and conditions via GUI-assisted thresholding, with limited additional classifier training(Nygaard et al. 2023). Random forest classifiers output a probability per frame of a behavior of interest occurring. As new cohorts of animals are screened and recording contexts are altered, the classifier certainty may decrease for positive frames, but the probability of non-event frames typically stays extremely low. As such, achieving appropriate performance on new samples is supported by dynamic control of discrimination thresholds using an interactive thresholding tool for positive versus negative behavior events (Fig. 1B, 2C). Precise thresholds can also be calculated via precision-recall curves (Fig. 2A). Most labs will eventually perform manipulations that alter behavior setups to the point of classifier failure. This is solved by the expansion and incorporation of new training sets. Using the video batch pre-processing interface, new experimental videos may be annotated, manually or with the assisted scoring platform, and easily added to prior models. Thus, new behavioral frames can be rapidly added to the training sets to improve performance, allowing groups to iteratively create updated behavioral classifiers (Fig. 2C).

Standard machine learning performance metrics for all classifiers are provided (Fig. 2A-B, Fig. S7-8; but see Discussion for comments on data leakage). For most classifiers, performance improvement occurs within the first ~20k positively annotated frames, equivalent to ~11 minutes using standard 30fps video acquisition. Classifier performance increases with additional annotations, higher clarity of operational definitions, and further iterations for targeted misclassification correction. Rat, CRIM, and mouse classifiers achieved F1 performance on behavior present frames of > 0.91, 0.73, and 0.77 respectively (Fig. 2A-B, Fig S7-8). Within SimBA, classifier training includes multiple checks of performance on novel videos to assess generalizability. Using the built-in visualization and validation tools, new unannotated behavioral videos can be analyzed to assess and subsequently adjust performance as necessary. Importantly, hand versus machine scoring results – comparing frame by frame classifications on independent videos – indicate high classifier performance both within labs (Attack Pearson's $R^2 = 0.91$ on 16 independent videos, Fig. S12) and across labs for multiple classifiers (Pearson's 0.936 to 0.998

R^2 across 8 classifiers(Miczek et al. 2022); Pearson's = 0.77, 0.94, 0.98 R^2, (Neira et al. 2022); confusion matrix accuracy = 98.6%, 99.3%, 85.0%(Winters et al. 2022b)).

For each classifier, the maximal F1 score is impacted by pose estimation performance, behavior distinctiveness, and training set construction. For example, in the case of pose-estimation performance, a ~4mm median error in 'tail end' tracking results in failed classification of tail rattles, while ~1-2.5mm median error in 'body hull' points do not affect classifications of other behaviors at F1 > 0.74 (Figs. S5, S7A-C). For behaviors associated with distinct pose estimation signatures, such as drinking, small training sets are sufficient for high performance (<2 minutes of positive frames, F1 = 0.965, Figs. S8 & S11). Conversely, to accurately classify multiple behaviors that share similar pose estimation signatures, such as attack and mounting, larger and more diverse training sets are required (attack: >30 minutes of positive frames, F1 = 0.921, Fig. 2). Comparing the 'chase' classifier from the CRIM datasets (~2 min of positive frames, F1 = 0.717) and the 'pursuit' classifier from the University of Washington (~1 minute of positive frames, F1 = 0.853) reveals the influence of training set construction on F1 score. Ultimately, users will reach an asymptotic point at which further training does not improve classification performance, which depends largely on tracking performance, the set of used features, and the distinctiveness of the behavior of interest (Fig. 2A).

SHAP calculations reveal similarities and differences between annotators, species, and behaviors

Explanations for how machine learning models reach their decisions can help researchers communicate and compare the results of disparate classifiers and support researchers in making informed decisions for machine model implementation and use(Ribeiro et al. 2016b; Sundararajan et al. 2017; Lundberg et al. 2020; Hatwell et al. 2020). One recent influential and accessible approach for generating explainable metrics of tree-based classifiers is SHAP (Shapley Additive exPlanations)(Lundberg and Lee 2017). We chose to use SHAP because it is one of the most widely adopted open source explainability methods available, in addition to being well-documented and under active development and support(Lundberg 2022). Shapley values are just one paradigm amongst a continuously growing number of approaches for explainable artificial intelligence (XAI); alternate options include LIME(Ribeiro et al. 2016b) and counterfactuals(Verma et al. 2020), all with different pros and cons. While it is preferable for new pipelines to be built with inherently explainable algorithms(Rudin 2019), we propose that

post-hoc explainability metrics are highly compatible with the current state of the field due to the diversity of ML packages being developed.

In essence, the Shapley value presents a game theory-based method for equitably distributing a game's earnings among its players. To illustrate, consider a scenario where individual feature contributions need to be assessed for a binary behavioral classifier. There are two key pieces of information available beforehand. Firstly, there's the average probability (expected value) indicating the likelihood of the behavior occurring in a video frame, such as 10% based on the number of positive annotations from the training data. Secondly, the model predicts the presence of the behavior in a new frame with 75% probability. By employing Shapley values, the disparity between these values (in this case, 65%) is attributed to the combination of feature values. This difference is then allocated to the features based on their contributions within the ensemble. The resulting Shapley value output is a vector of the same length as the number of features, with a total sum of 65 (for a detailed mathematical explanation of SHAP and its application in behavioral neuroscience, refer to Supplementary Methods).

SimBA calculates many hundreds of features (Table 2) that individually may not be very informative. While SHAP values can be calculated for each of these features if desired, the additivity axiom allows users to bin features into ethologically relevant feature categories (Table 3) that are related to the biology of their model system and/or behaviors of interest. Another key advantage is that classifiers generated for the same behavior, but using different pose estimation schemas (e.g., one model with 5 body parts versus a model with 8 body parts), may still be directly compared by such bins, which would not be possible using individual feature importance's (Fig. S17). SHAP is built into SimBA classifier construction to encourage its use and accessibility, but more-so we propose that the adoption of any type of explainability paradigm is more important than the specific algorithm selected.

To specifically demonstrate SHAP's potential to enhance behavioral reportability within preclinical behavioral neuroscience, we collected and compared independently created rodent attack classifiers from expert annotators at institutions across the United States. These classifiers and annotators used data from different recording environments, with different experimental protocols, strains, sex, video formats and pose-estimation models (Fig. 3A, Statistical Appendix A). The classifiers all showed high accuracy in their respective environments.

SHAP analysis of attack classifiers revealed the relative importance of feature category bins(Goodwin et al. 2022) for discerning both attack and non-attack frames. Across labs, animal distances and movement were important for identifying attack, while animal shapes were least important. In addition to identifying these common patterns, a stark difference in the influence of resident shape was seen between the Stanford attack classifier and classifiers from other sites (Fig. 3C). To ensure that this was not due to incongruent operational definitions of attack behavior, we manually annotated the Standford dataset based on our operational definition of attack. Manual annotations were highly consistent between sites (Fig. S16A, $R^2 = 0.998$; Fig. S16B, gantt plot). SHAP analysis showed that UW annotations relied on resident body shape over longer durations than Stanford. This demonstrates that while the annotators have high inter-rater reliability, SHAP determined they rely on different features for attack scoring. This observation is likely based on the biological differences between the two sites datasets, where Stanford used C57 resident mice rather than outbred strains like other research locations.

Since SHAP values do not convey directional relationships, we plotted the 32k observations within each category (8k from each site) with the normalized feature values on the y-axis and SHAP value on the x-axis (Fig. 3D, Statistical Appendix A). Here the left-most panel shows that as the distance between animals decreases, the attack classification probability increases across the sites, with shorter animal distances having a stronger positive impact on attack classification probabilities at CRIM/Caltech than the classifiers annotated at other sites.

Rat versus mouse attack behavior

Next, we analyzed SHAP values for the rat resident intruder classifier and compared the values with the five classifiers generated by the mouse consortium of attack classifiers (Fig. 4A). Rat attack events were recognized primarily by the shape of the resident and the intruder, and distances between the resident and the intruder. Rat attacks differed significantly by feature category bin, but not feature sliding window size. All feature bin correlations were significant, with intruder shape being most negatively correlated with attack probability in rats and intruder movement most positively correlated in mice (Fig. 4B).

Automated behavioral analysis of reactive aggression across sex

Female assays for the study of aggression in mice have historically focused on maternal aggression due to the stated low propensity for these mice to display aggression during reactive

aggression tasks (resident intruder or chronic social defeat stress tests) typically used for males(Takahashi et al. 2017). Recently, studies have revisited reactive aggression in females and have found that a subset of female Swiss Webster (CFW) mice attack same-sex intruders when either virgin and singly housed(Hashikawa et al. 2017), or cohoused with a castrated male(Newman et al. 2019). While CFW females do not appear to show aggression reward by standard measures including aggression conditioned place preference and aggression self-administration(Aubry et al. 2022), this still presents an opportunity to directly compare male and female reactive aggression behavior.

Following screening females for resident-intruder aggression, we conducted female(Newman et al. 2019; Aubry et al. 2022) and male(Golden et al. 2011) chronic social defeat stress (CSDS) assays (see Supplemental Methods). We calculated the total duration, number of bouts, mean bout duration and interval for attack, pursuit, anogenital sniffing, escape, and defensive behaviors (SHAP values, Fig. S2, S4), both per day of testing and averaged across all five testing days. Males and females showed significant differences in all five assayed behaviors (Fig. 5D), with females showing higher average total durations and number of bouts in attack, pursuit, and escape behaviors ($p < 0.001$ for all), while males had higher levels of anogenital sniffing and defensive behaviors (duration: $p = 0.0461$, 0.0374; bouts: $p = 0.0298$, 0.0146). There were no differences for any of the behaviors in mean bout interval or in attack or anogenital mean bout duration. Females showed longer behavioral bouts for pursuit, defensive, and escape behaviors ($p < 0.001$ for all). Only three metrics were significantly affected by day: escape duration and bouts (duration: interaction $p = 0.0122$, day $p = 0.3700$, sex $p < 0.001$; bouts: interaction $p = 0.0415$, day $p = 0.3947$, sex $p < 0.001$) and number of pursuit bouts (interaction $p = 0.0404$, day $p = 0.1724$, sex $p < 0.001$).

To understand how males and females performed behaviors predicted by the same supervised behavioral classes, we calculated SHAP values for 10k positive behavioral frames per behavior and sex (Fig. 5D, right). There were significant differences in SHAP values across sexes for all behaviors (Appendix A), with attack and pursuit behaviors differing most significantly in intruder shape, anogenital sniffing in intruder movement, defensive behavior in combined animal movement, and escape behavior in resident shape (Appendix A).

Automated behavioral analysis of male reactive aggression across environments

Resident-intruder assays are typically conducted in an animal's home cage to measure the reactive or territorial aggression of the resident animal. Frequently, these assays examine resident males across their first exposures to established subordinate intruders. Alternately, CSDS testing is typically performed in thinner subdivided hamster cages, where the "resident" animal is an established aggressor(Golden et al. 2011) and the intruder is a smaller mouse that experiences up to 10 consecutive days of defeat. Directly comparing RI and CSDS datasets allows us to gain a preliminary understanding of the intersection of aggression experience and testing environment for driving social behaviors.

RI and CSDS male behavior differed significantly by day, total duration across all five behaviors, number of bouts in attack, anogenital, defensive, and escape behaviors, and in the mean bout interval between anogenital and escape events (Fig. 6C, Appendix A). RI males showed a marked decrease in anogenital sniffing duration across days (interaction $p = 0.0123$, day $p = 0.0478$, environment $p < 0.0071$), with concomitant increases in attack (interaction $p < 0.001$, day $p = 0.0204$, environment $p = 0.9408$) and pursuit behaviors (interaction $p < 0.0258$, day $p = 0.0295$, environment $p < 0.001$), indicating a shift from exploratory to aggressive behaviors as males gained aggression experience. SHAP comparisons of classifiers revealed differing behavioral motifs between environments. Attack differed most significantly in intruder movement, while anogenital and pursuit behaviors differed in resident shape, defensive behavior by animal distances, and escape by intruder shape (Fig. 6C, Appendix A).

Discussion

There is a vibrant and growing ecosystem of computational behavioral tools designed specifically for the behavioral neuroscience community, that have directly impacted scientific directions and methods(Shemesh and Chen 2023) (see Table 1 for examples). SimBA has been heavily influenced by many of these packages(Kabra et al. 2013; Mathis et al. 2018c; Graving et al. 2019b; Pereira et al. 2022), and now influences others(Bordes et al. 2022; Winters et al. 2023; Lapp et al. 2023). Since its release, numerous independent labs have used SimBA to address varied scientific questions across diverse model systems, providing critical feedback on SimBA and its ongoing development. The key feedback has consistently focused on introducing easily accessible explainability metrics that are useful to non-specialized users and allow for generalizable use of classifiers and compatibility of disparate datasets.

Here, we used SimBA to analyze data including male and female chronic social defeat stress (CSDS) behavior, and male CSDS and resident intruder (RI) behaviors and describe the use of explainability metrics. Between sexes we found that male and female intruders differ significantly in their coping strategies, with females defending themselves less than males. These differences may partially explain differences in resilience to social stress between males and females. Between CSDS and RI assays, which are performed in differently sized arenas, we found differences in resident aggressive and intruder coping behaviors between males. This demonstrates clear cross-protocol behavioral differences and emphasizes a need for standardization and quantification of experimental variation within social behavior research in general and aggression research specifically, and we provide a tool to do this.

We demonstrate the utility of SHAP values for uncovering differences within supervised machine learning scenarios. We quantify significant differences in features and time bins defining female and male aggression as well as aggressive behaviors across distinct testing environments. For example, SHAP values revealed and quantified potential cross-lab classifier confounds, and aided comparisons between male and female attack events.

Although strong arguments remain for using inherently explainable models in machine learning(Rudin 2019), many maturing behavioral neuroscience tools still rely on black-box methods. The post-hoc nature of SHAP allows independent labs an algorithmic freedom while maintaining options for cross-model explainability comparisons through game theory and the additivity axiom. A significant drawback of SHAP, however, is the computational and runtime cost associated with analyzing larger datasets. SimBA depends on the TreeSHAP(Lundberg et al. 2019) algorithm to calculate SHAP values. A challenge when calculating SHAP values through TreeSHAP is the computational complexity and run-times that scale polynomially with the number of estimators and features, and can interfere with the analyses of larger datasets(Covert et al. 2021). To negate this, SimBA includes multiprocessing options that allows users to linearly reduce their run-times with the count of their available CPU-cores. SHAP also represents a post-hoc approach when inherently explainable models (e.g., using a few carefully selected relevant features joined with low algorithm complexity) may suffice for low-complexity behaviors(Goodwin et al. 2022). Furthermore, as a correlative approach, users should avoid interpreting causal relationships between SHAP-inferred feature importance's and classification probabilities. Ongoing efforts in SimBA involve parallelization and just-in-time compilation

making the full catalogue of methods available on standard behavioral neuroscience hardware at the maximum speeds allowed by the machine. Notably, the SimBA platform has additional in-built access to other prevalent model interpretability methods (e.g., partial dependencies, permutation importance's and gini/entropy-based measures) when needed.

Manual behavioral analyses typically depend on nondescript qualitative operational definitions, posing challenges for standardization across individuals and laboratories, and potentially contributing to issues with reproducibility downstream. In contrast, as demonstrated here, the incorporation of ML-based behavioral analysis coupled with explainability metrics produce comprehensive quantitative operational definitions of behavior. This allows us to re-conceptualize behavioral analysis though precise and verbalizable statistical component rules that are applied when scoring behaviors of interest. These quantified definitions can be shared as resources, akin to RRID-like reagents, enhancing transparency and reportability and facilitating cross-study comparisons and reproducibility.

As a supervised learning tool, a weakness of our approach is the cost of collecting the human behavioral annotations required for fitting reliable downstream models. Other prominent platforms may circumvent such costs through active learning, task-programming, or semi-supervised statistical techniques(Lorbach et al. 2019; Sun et al. 2021b; Whiteway et al. 2021; Schweihoff et al. 2022). Despite our publicly available classifiers and annotations, users will typically have to add a few short, varied, and representative annotations to their classifier training sets due to the inherent variation in individual laboratory experimental setups. However, the behavioral neuroscience community has an exceptionally rich historical bounty of carefully documented animal behaviors in video recordings. To lessen user burden, SimBA includes several tools to accommodate use of such historical annotations for supervised machine learning purposes. Further parallel initiatives, including MABe (Sun et al. 2021a), the OpenBehavior project (2016b), and Jackson Laboratory Mouse Phenome Database (2001), may also address these limitations through the distribution of larger repositories encompassing diverse video recordings with associated human annotations.

A further challenge for the field is appreciating how to accurately judge the performance of machine learning models, which can be biased towards subsets of animals and recording environment and that depend on time-series data that is vulnerable to leakage (Kapoor and Narayanan 2022). Maintaining true independence across training and hold-out sets is a standing

challenge within machine learning which can be particularly difficult in behavioral neuroscience use-cases where insufficient amounts of fully independent annotations are available. Fundamentally, although low performance metrics are indicative of an inadequate classifier, high metrics may not be sufficiently indicative of reliable and generalizable out-of-sample performance. We recommend that users include smaller held-out and hand-scored representative validation sets for validation purposes.

In conclusion, SimBA provides a user-centric, modular, and accessible platform for machine learning analysis of behavior with the primary aim of promoting an understanding of the behavioral nuances that we value as neuroethologists. Here, we demonstrate the utility of SimBA as a behavioral analysis tool by consolidating cross-site behavioral datasets with machine learning methods to quantify and describe complex behaviors across experimental contexts.

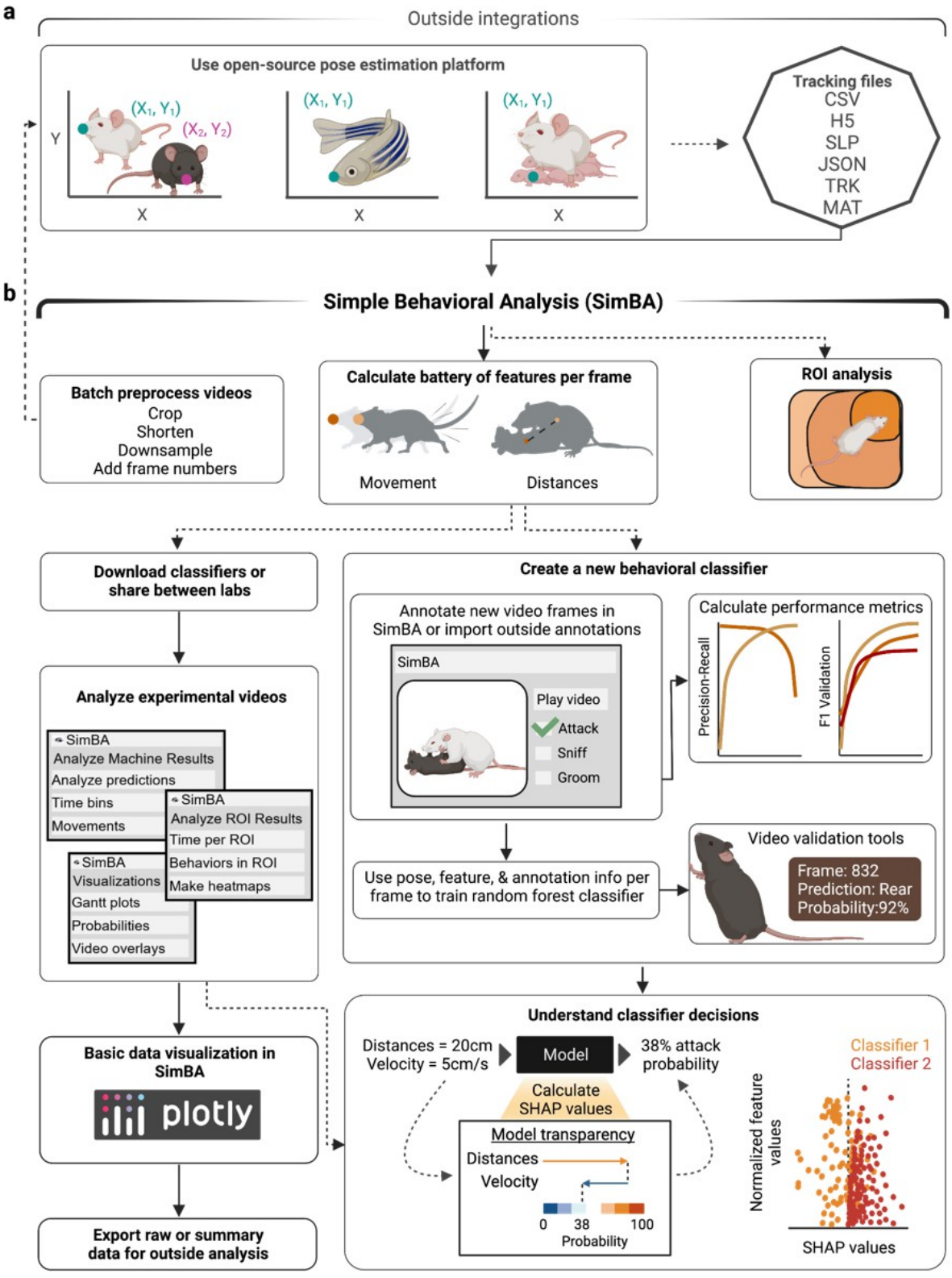
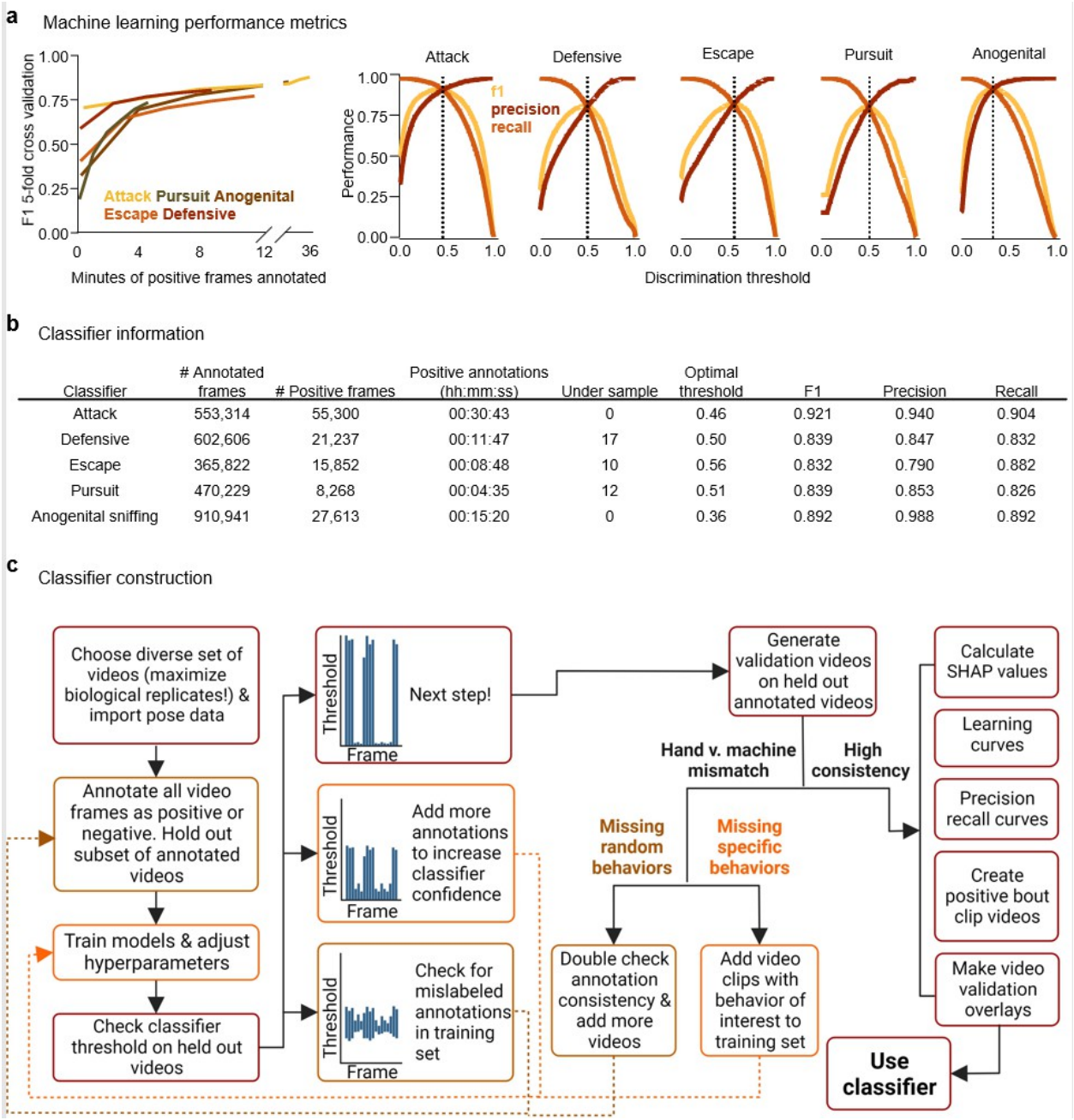


Figure 4.1. SimBA workflow and outside integrations. *SimBA is an open-source, graphical user interface-based program built in a modular fashion to address many of the specific analysis needs of behavioral neuroscientists. SimBA contains a suite of video*

editing options to prepare raw experimental videos for markerless pose tracking, behavior classifications and visualizations. Once users have analyzed their videos for animal pose data via common open-source pipelines (a), the data is imported to SimBA for subsequent analysis (b). Within SimBA, users have the option to perform pose estimation outlier corrections, interpolation and smoothing methods, or use uncorrected pose data in any SimBA module. To perform supervised behavioral classification, users can download premade classifiers from our OSF repository, request classifiers from collaborators, or create classifiers by annotating new videos in the scoring interface. Users can also use historical lab annotations created in programs such as Noldus ObserverXT, Ethovision, or BORIS. A variety of tools are provided for evaluating classifier performance, including calculating standard machine learning metrics and visualization tools for easy hands-on qualitative validation. Following behavioral classification, users can perform a batch analyses' and extract behavioral measures. To understand the decision processes of classifiers, we encourage users to calculate and report explainability metrics, including SHAP values. We provide extensive documentation, tutorials and step-by-step walkthroughs for all SimBA functionality.



Classifier	# Annotated frames	# Positive frames	Positive annotations (hh:mm:ss)	Under sample	Optimal threshold	F1	Precision	Recall
Attack	553,314	55,300	00:30:43	0	0.46	0.921	0.940	0.904
Defensive	602,606	21,237	00:11:47	17	0.50	0.839	0.847	0.832
Escape	365,822	15,852	00:08:48	10	0.56	0.832	0.790	0.882
Pursuit	470,229	8,268	00:04:35	12	0.51	0.839	0.853	0.826
Anogenital sniffing	910,941	27,613	00:15:20	0	0.36	0.892	0.988	0.892

Figure 4.2. Classifier construction workflow and classifier performance metrics. (a) *Machine learning performance metrics for the classifiers used in Figures 5-6 (See Figs. S7-8 and 10-12 for in-depth classifier performance data). Left: F1 5-fold cross validation learning curves plotted against minutes of positive frames annotated (30 frames per second). Right: Precision-recall curves plotted against discrimination threshold for five classifiers, which can be used in combination with the SimBA interactive thresholding visualization tool to determine the most appropriate detection threshold for classifiers and specific datasets.* ***(b)*** *Extended information for the training sets for each of the five classifiers.* ***(c)*** *Workflow for creating high fidelity and generalizable supervised behavioral classifiers. The dotted lines indicate optional loops for iteratively improving classifier performance. Behavioral operational definitions, and classifier SHAP values, are shown in Figures S1-4.*

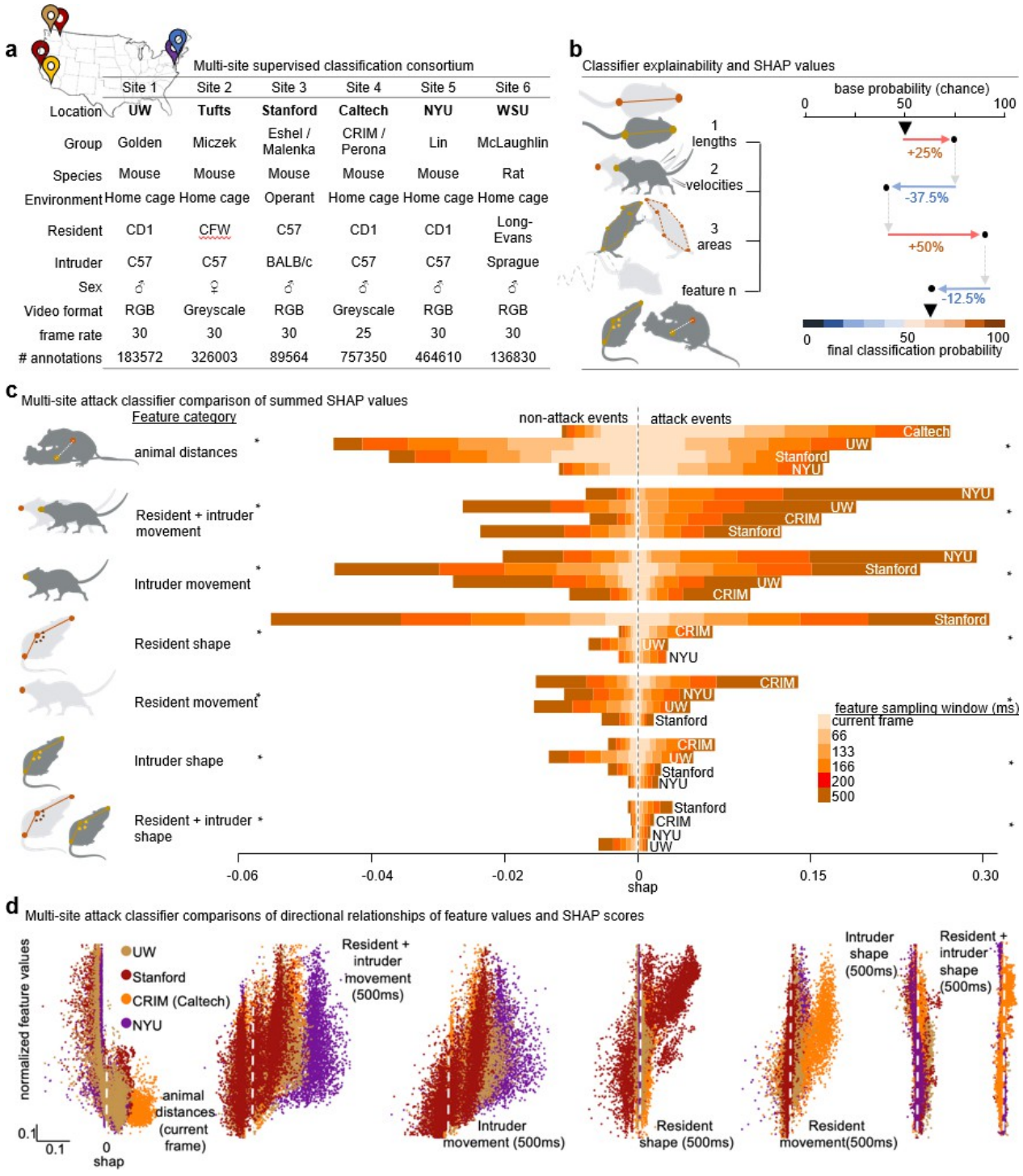


Figure 4.3. Attack consortium data. ***(a)*** *Description of the consortium dataset used for the cross-site attack classifier comparisons.* ***(b)*** *Schematic description of SHAP values, where the final video frame classification probability is divided among the individual features according to their contribution.* ***(c)*** *Comparison of summed feature SHAP values, collapsed into seven behavioral feature categories for four different mouse attack classifiers. We divided each category into six further sub-categories that represented features within the categories with different frame sampling frequencies (1 frame – 500ms) and are denoted by shaded colors. Asterisks denote significant main effect of consortium site, $p < 0.0001$. See Appendix A for full statistical analysis.*

(d) *Scatter plots showing the directional relationships between normalized feature values and SHAP scores in four mouse resident-intruder attack classifiers and seven feature sub-categories. The dots represent 32k individual video frames (8k from each sites dataset), and color represents the consortium site where the annotated dataset was generated.*

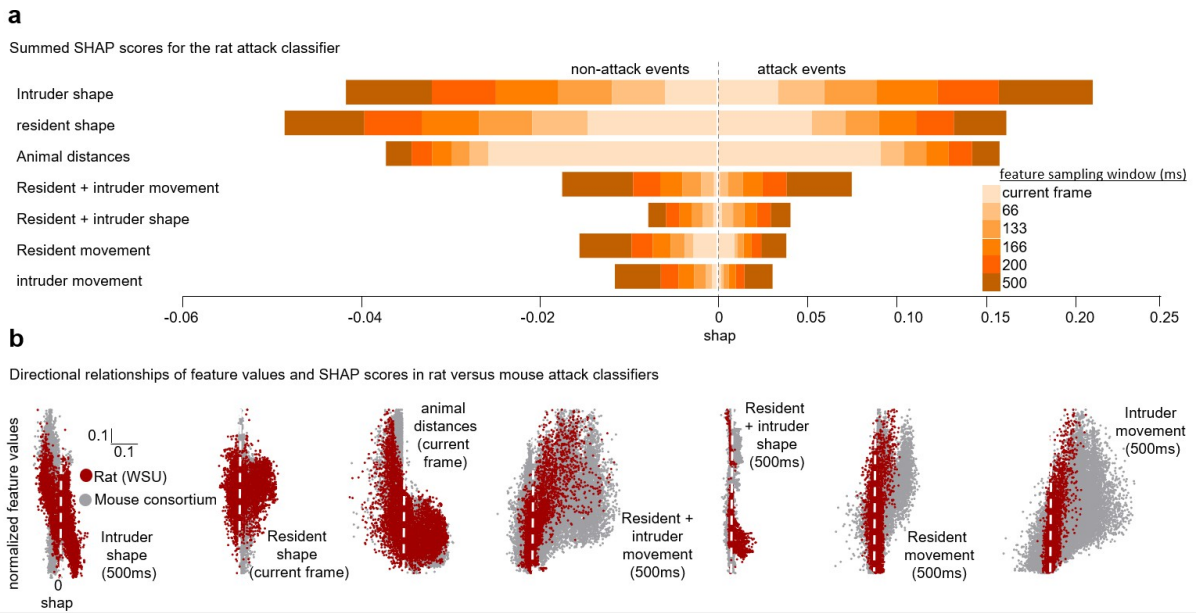


Figure 4.4. SHAP cross-species attack classifier data. *Explainable classification probabilities in the rat resident-intruder attack classifier using SHAP.* ***(a)*** *Summed SHAP values, collapsed into seven behavioral feature categories for the rat random forest attack classifier. Colors denote sliding window duration as in Figure 3.* ***(b)*** *Scatter plots showing the directional relationships between normalized feature values and SHAP scores in seven feature sub-categories of the rat resident intruder attack classifier. The rat attack classifier is shown in red. For comparison, the SHAP values for the mouse attack classifiers (from Figure 4), are shown in grey. Dots represent individual video frames. See Appendix A for full statistical analysis.*

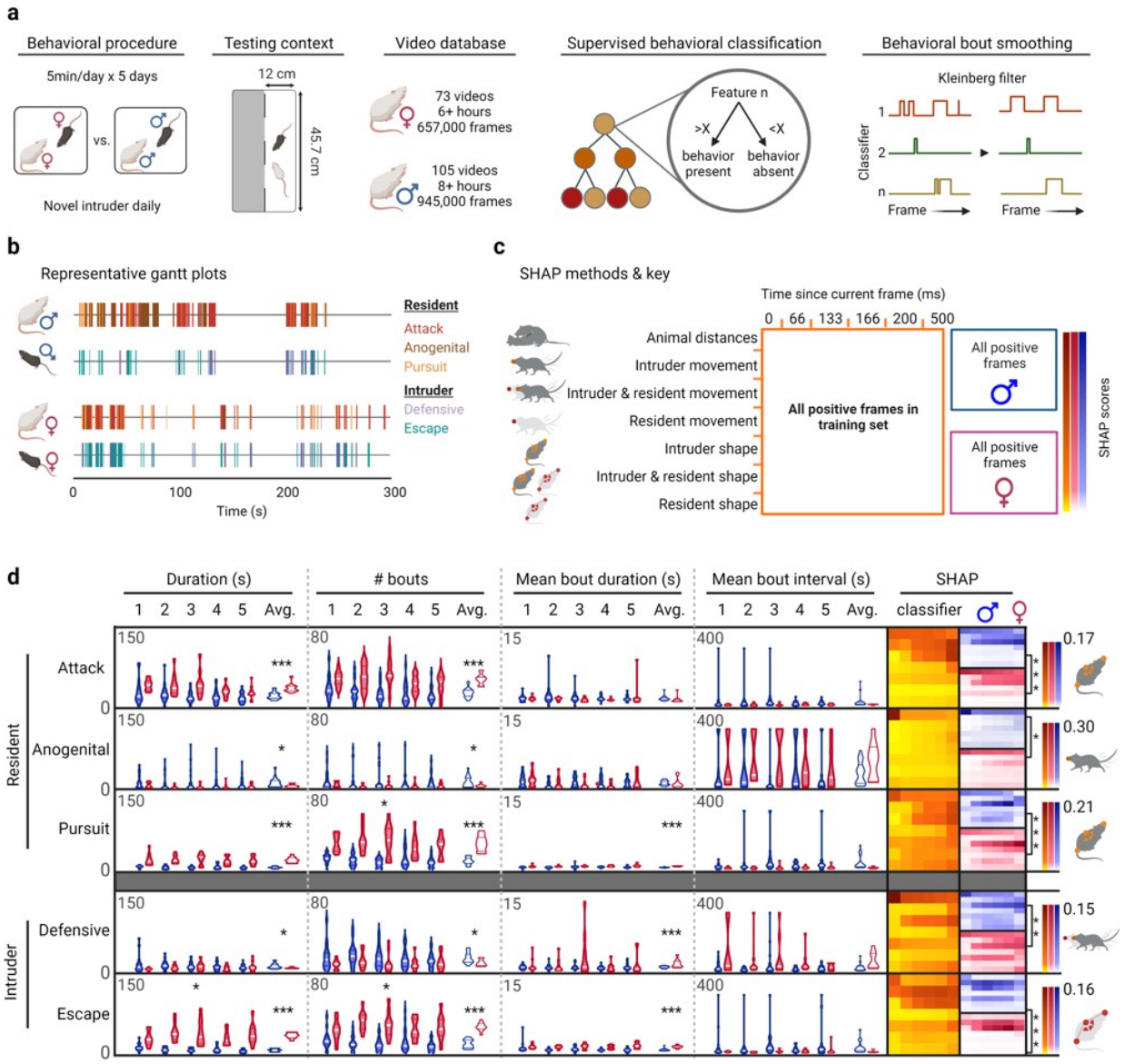


Figure 4.5. Chronic social defeat stress behaviors differ between sexes. ***(a)*** *Schematic representation of the mouse chronic social defeat (CSDS) behavioral protocol and the analysis pipeline for supervised machine learning behavioral classification.* ***(b)*** *Representative Gantt charts of classified male (top) and female (bottom) resident and intruder behaviors.* ***(c)*** *Key for the SHAP analysis and feature bin comparisons.* ***(d)*** *Supervised behavioral data and SHAP comparisons for five behavioral classifiers. Male data are represented in blue, and female in pink. For each classifier, SimBA provided the total duration (s), number of classified bouts, mean bout duration (s), and mean bout interval (s) across individual testing days (n = 21 males for all classifiers, 11 female residents for attack, anogenital sniffing, pursuit and classifiers, and 10 female intruders for defensive and escape classifiers). Males and females showed significant differences in all five assayed behaviors, with females showing higher average total durations and number of bouts in attack, pursuit, and escape behaviors ($p < 0.001$ for all), while males had higher levels of anogenital sniffing and defensive behaviors (duration: $p = 0.0461, 0.0374$; bouts: $p = 0.0298, 0.0146$). Only three metrics were significantly affected by day: escape duration and bouts (duration: interaction $p = 0.0122$, day $p = 0.3700$, sex $p < 0.001$; bouts: interaction $p = 0.0415$, day $p = 0.3947$, sex $p < 0.001$) and number of pursuit bouts (interaction $p = 0.0404$, day $p = 0.1724$, sex $p < 0.001$). Average SHAP values are reported in Figure S2. The color intensity for all three SHAP datasets per*

*classifier are on the same scale, as indicated by the scales on the right. For each behavior, comparisons with the lowest p-value per category are highlighted via a comparison bracket and the feature bin mouse icon. Asterisks denote significance levels * < 0.05, ** < 0.01, *** < 0.001. See Appendix A for full statistical analysis.*

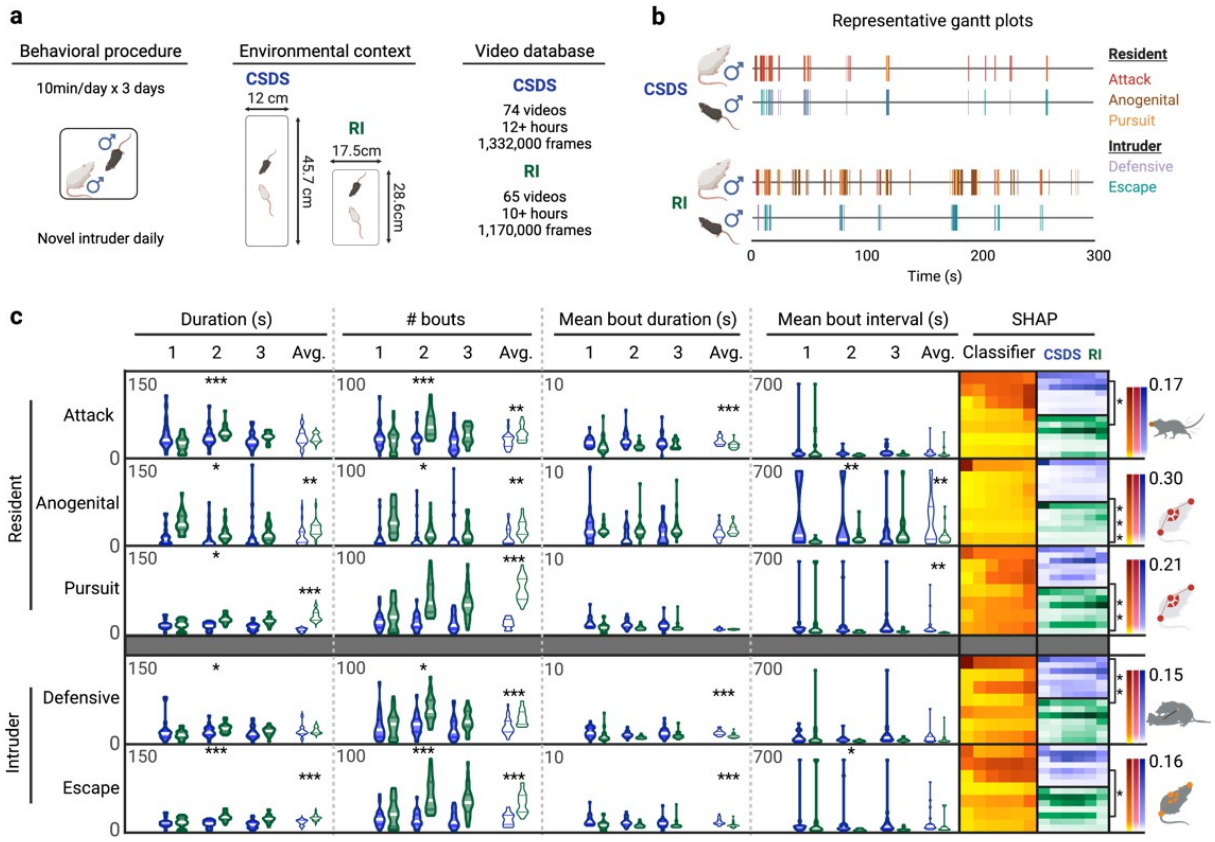


Figure 4.6. Environment and experience influence male aggression and coping behaviors. ***(a)*** *Schematic representation of the mouse chronic social defeat (CSDS) and resident intruder (RI) behavioral design.* ***(b)*** *Representative Gantt charts of classified CSDS (top) and RI (bottom) resident and intruder behaviors.* ***(c)*** *Supervised behavioral data and SHAP comparisons for five behavioral classifiers. CSDS data are represented in blue, while RI data are shown in green. For each classifier, SimBA provided the total duration (s), number of bouts, mean bout duration (s), and mean bout interval (s) across individual testing days (n = 21 CSDS, 24 RI). RI males showed a marked decrease in anogenital sniffing duration across days (interaction $p = 0.0123$, day $p = 0.0478$, environment $p < 0.0071$), with concomitant increases in attack (interaction $p < 0.001$, day $p = 0.0204$, environment $p = 0.9408$) and pursuit behaviors (interaction $p < 0.0258$, day $p = 0.0295$, environment $p < 0.001$). The color intensity for all three SHAP datasets per classifier are on the same scale, as indicated by the scales on the right. For each behavior, comparisons with the lowest p-value per category are highlighted via a comparison bracket and the feature bin mouse icon. Asterisks denote significance levels * < 0.05, ** < 0.01, *** < 0.001. See Appendix A for full statistical analysis.*

Supplemental methods

Animals

Mice. Male CD-1 (strain #022, Charles River Labs (CRL)) or female CFW (strain #024, CRL) white coat colored 12-week-old mice were used as residents. Female CFW residents were pair housed with 10-week-old castrated CFW male mice for the duration of the study to elicit aggression(Newman et al. 2019; Aubry et al. 2022), while male CD-1 residents were singly housed. Sex-matched, black coat colored C57BL6/J (C57; strain #000664, Jackson Labs) mice were used as intruders for all aggression assays. C57 females were > 10 weeks old, and males were > 8 weeks old. We chose to use differently-coated social partners to facilitate subject identification, which is improved by using social pairs with different coat colors. We gave all mice free access to standard food chow and water in all experiments. We housed all mice with enrichment (cotton padding) in standard Allentown clear polycarbonate cages covered with stainless-steel wire lids at least one week prior to experiments, and we maintained them on a reverse 12-h light/dark cycle (light off at 0900 am).

Rats. Male Long-Evans rats (Simonsen Laboratories, 120-140 days old) were used as residents. Residents were housed in isolation in standard rat cages for the duration of the study. Male Sprague Dawley rats (bred in house, 60-80 days old) were used as intruders. Intruders were pair housed in standard rat cages. We maintained all animals on a reverse 12-h light/dark cycle (light off at 7am).

All experiments were performed in accordance with the Guide for the Care and Use of Laboratory Animals under protocols approved by the local Animal Care and Use Committee at each institution (University of Washington, Tufts University, Stanford University, Columbia University, Washington State University, and the Icahn School of Medicine at Mount Sinai).

Code base

SimBA is maintained on GitHub (https://github.com/sgoldenlab/simba), pip (https://pypi.org/project/Simba-UW-tf-dev/), Read The Docs (https://simba-uw-tf-dev.readthedocs.io/), and Gitter (https://app.gitter.im/#/room/#SimBA-Resource_community). SimBA was originally conceptualized by SRON, NG, and SAG. Version 1 was released in 2020 with code written by SN and JJC, with minor contributions from NG, AI, and SH. Since then, the SimBA version 2 codebase, documentation, and support forums are written, maintained, and

developed independently by SRON with significant non-coding support from NG. The authors sincerely thank the collaborative and friendly open-source community for their continued expert feedback and suggestions.

Behavioral protocols

Chronic social defeat stress (CSDS). We used male(Golden et al. 2011) and female(Newman et al. 2019; Aubry et al. 2022) CSDS procedures as previously published. Briefly, we recorded dyadic encounters between male or female mice in clear polycarbonate cages (cage size: 28x19x12cm) divided in half by a clear acrylic barrier. Aggressive CD1 or CFW resident animals were housed on one side of the barrier where they encountered an unfamiliar sex-matched C57 intruder for 5 min (n = 11 female residents, n = 10 intruders) or 10 min (n = 21 male residents, 21 intruders). Female CFW mice were rendered aggressive through cohabitation with castrated males; castrated males were temporarily single-housed during daily 5-minute female defeat sessions. We analyzed five days of data per aggressive mouse.

Resident intruder (RI). We recorded dyadic encounters between male mice (n = 24 residents, 24 group-housed intruders) in standard shoebox home cages (17.8x28.6cm). C57 intruder animals were introduced into the center of the CD1 resident home-cage for 10 min and allowed to freely interact. Animals were tested once a day for three consecutive days. Rat resident-intruder assays were recorded between same-sex conspecifics in clear polycarbonate cages with fresh bedding (cage size: 33x46x19cm). Residents were tested three times, four to eight weeks apart with novel intruders. Assays lasted for up to ten minutes.

Video recordings

All recordings were made overhead. Male mice were recorded at 30-80 fps with USB 3.0 cameras (acA2040-120uc – Basler ace, Basler) using fixed-focal length lenses (Edmund Optics, NJ, 16mm/F1.4) at variable resolutions (W:1000-1200px, H:1255-2056px) using the pylon camera software (Basler). Male rats and female mice were recorded at 30-60 fps and 1280x720 resolution using a Logitech C922 camera. All recorded videos used to build behavioral classifiers were re-sampled in SimBA to 30 fps before being used to create machine learning classifiers. The Caltech Resident-Intruder Mouse (CRIM13)43 dataset was recorded at 25 fps and 640x480 resolution.

Video processing

Video recordings were pre-processed using tools available in SimBA. We shortened, cropped, and saved videos and frames in RGB, greyscale and CLAHE (Contrast-limited adaptive histogram equalization) enhanced formats at variable frame rates. CLAHE video-conversion can improve image quality and pose-estimation in non-optimal recording conditions. The CRIM13 dataset was recorded in SEQ file format and converted to MP4 format using in-built functions in SimBA (provided by Xiaoyu Tong, Lin Lab, NYU). We used SimBA to extract frames within specific time-periods, concatenate videos, down sample video resolutions, generate gifs and videos. An exhaustive list of SimBA video pre-processing tools and tutorials is available on the SimBA GitHub repository and Appendix B.

Pose estimation

We created pose estimation models in DeepLabCut (version 1.0)10 for rats, the CRIM dataset, and our experimental mouse videos. SimBA currently supports pose-estimation import from regular and maDeepLabCut(Mathis et al. 2018c; Lauer et al. 2021), SLEAP(Pereira et al. 2022), BENTO(Sun et al. 2021b), DANNCE 3D(Karashchuk et al. 2021b), DeepPoseKit(Graving et al. 2019b) and Animal Part Tracker(Branson). In rat and CRIM datasets, we tracked 8 points per animal consisting of nose, left and right ears, left and right sides, back, tail base, and tail end. For our mouse model, we tracked 7 points per animal (same points excluding tail end). We labeled frames from a diverse set of videos to create training sets for ResNet-50 based neural networks for all models. The pose estimation data were imported to SimBA for further analysis. All models are available on the SimBA Open Science Framework (OSF) repository. Further, SimBA includes methods for cross-video animal track swapping and pose scheme body part dropping to facilitate merging pose estimation datasets (Appendix B).

Classifier creation

Our original classifier set(Nilsson et al. 2020) consisted of 10 mouse classifiers (attack, pursuit, lateral threat, anogenital sniffing, allogrooming normal, allogrooming vigorous, mounting, scramble, flee, and upright submissive), seven rat classifiers (attack, anogenital sniffing, lateral threat, approach, boxing, avoidance, and submission), and 11 mouse classifiers from the open-source CRIM13(Xavier P. Burgos-Artizzu et al. 2021) dataset (approach, attack, chase, circle, copulation, drink, eat, sniff, up, clean, and walk away. All rat and CRIM13 data

presented in this manuscript use these original classifiers. The five University of Washington mouse classifiers (attack, anogenital sniffing, pursuit, escape, defensive) are optimized iterations of a subset of the original 10 mouse classifiers. All datasets are available, with detailed information on classifier creation, on the SimBA OSF repository. Importantly, all frames included in the random forest training sets need definite labels as positive or negative for the behavior of interest. Incorrectly labeled events negatively impact model performance. As such, we created strict operational definitions for our behaviors (Fig. S1).

We used SimBA to create random forest classifiers using scikit-learn50 with one classifier per behavior of interest (Appendix B). Random forest classifiers accept a range of hyperparameters that specify how decision trees are generated. Hyperparameters for our provided classifiers consist of: criterion = entropy, max features = square root, minimum sample leaf = 1, and number of estimators = 2k. Random under-sampling was used in cases of major class imbalances (Fig. S1, Fig. 2). All classifiers were iteratively trained and optimized (Fig. 2C). SimBA accepts a range of different random forest hyperparameter settings and sampling methods, but users unfamiliar with the available parameters can import recommended or previously successful settings (Appendix B) based on classifiers of behaviors with similar frequency and salience.

Classifier performance

We used SimBA to generate classifier learning curves (Fig. 2A, left). In these learning curves we evaluated F1-scores after performing 5-fold cross-validations using 1, 25, 50, 75 and 100% of the shuffled data sets to predict the classified behaviors on 20% of the datasets. Learning curves indicate how inclusion of further logged behavioral events affect classifier performance. Furthermore, we used SimBA to generate precision-recall curves (Fig. 2A, right), which informs the balance between sensitivity and specificity at varying classifier discrimination thresholds. F1 scores are a harmonic mean of precision and recall: F1 = 2 * (precision * recall) / (precision + recall), where precision quantifies the number of positive class predictions that actually belong to the positive class, and recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

SHapley Additive exPlanations (SHAP)

The Shapley value contribution(Covert et al. 2021) of a specific feature ϕ_i towards the final prediction generated by model f for a data instance with features x is computed as follows. For a given data instance, x, for all possible subsets $S \subseteq F$, where F represents the full set of features used in the model, the model is trained both with the feature of interest present, and without it present, denoted $f_{S\cup i}$ and f_S respectively. Predictions generated from the model excluding the feature are subtracted from the predictions generated from the model including the feature $f(x_{S\cup i}) - f(x_S)$. The difference in predicted values is multiplied by the number of possible permutations of the set, multiplied by the number of permutations of the remaining features, divided by the total possible number of permutations of features, where $|F|$ is the cardinality of the feature set F, ultimately generating a weighted average. The weighted averages of each set contribution are summed to generate the total contribution of feature i. Thus, the contribution of feature i to the prediction is:

$$\phi_i(f,x) = \sum_{S\subseteq F\setminus i} \frac{|S|!(|F|-|S|-1)!}{|F|!}(f_{S\cup i}(x_{S\cup i}) - f_S(x_S))$$

Explainability and classifier comparisons

We used Shapley Additive Explanations (SHAP)(Lundberg and Lee 2017) and TreeSHAP(Lundberg et al. 2019) to evaluate how feature values impact classification probabilities (Fig. 3-7). Due to the Shapley additivity axiom, we can collapse Shapley value contributions of features that measure similar, general, and often colinear characteristics of the behavior of interest into interpretable physical categories while maintaining consistency, accuracy and biological relevance. Hence, to aid interpretability(Goodwin et al. 2022), we collapsed the features into seven behaviorally-defined feature categories that measure general characteristics of social interactions (i.e., animal distances, resident and intruder movement, intruder movement, resident movement, resident shape, intruder shape, and resident and intruder shape). Each of the seven feature categories contained six temporal windows that represent different frame sampling frequencies (single frame to 500ms). The list of features encompassed by each collapsed feature category is modifiable for custom use cases in SimBA (Appendix B,

'Mixins'). We created the general structure of the feature categories to ensure compatibility with alternative and novel feature sets targeting similar behaviors.

The use of these categories ensures that classifiers targeting the same behavior, but using different feature sets, can be directly compared. A common reason that classifiers use different feature sets is a consequence of the initial pose-estimation scheme selected by experimenters, which influences the selection of features. For example, two research groups interested in the same behavior may use pose-estimation schemes that identify 5 versus 8 body parts due to their experimental needs; regardless, by binning their classification features within the same biologically determined classes and sampling frequencies, the use of Shapley values allows for direct comparisons between their classifiers.

SimBA was used for evaluating feature SHAP values in behavioral classifiers (Appendix B). The classifiers for cross-site SHAP comparisons (Fig. 3) were generated based on annotations from four different laboratories and showed similar performance for classifying attack behavior within their respective datasets (precision > 0.937, recall > 80.0, F1 > 86.3, attack present: > 4,437). For comparisons between rat and mouse attack classifiers (Fig. 4), a random sample of 4k attack and 4k non-attack frames from each site were analyzed using SHAP (total: 32k frames). Each of the five updated mouse classifiers (Fig. 2) were similarly evaluated on 10k (or greatest amount if < 10k available) positive frames per behavior (Fig. 6-7). From the supervised analysis of the biological datasets, we extracted 1k positive frames prior to Markov model(Lee et al. 2019) smoothing in SimBA, and evaluated these for SHAP values (Fig. 6-7).

Behavioral analysis

For the CSDS and RI datasets, we imported DeepLabCut pose information into SimBA. We used the SimBA GUI to calibrate video scales (pixels/mm). Outlier correction was performed using hedonic rules as documented in the SimBA API (movement criterion: 0.7, location criterion: 1.5). We extracted 498 features and analyzed videos with five classifiers (Resident: attack, anogenital sniffing, pursuit. Intruder: escape, defensive) using parameters outlined in Fig. 2. Multiple behaviors could be present in a single frame (i.e. attack and defensive behavior). We used a Kleinberg Filter[52] (sigma: 0.3, kappa: 2, hierarchy: 2 for anogenital and attack, 3 for defensive and escape, and 4 for pursuit) with the hierarchical search function for all classifiers as documented in the SimBA API. The Kleinberg algorithm using an infinite Markov chain[53] to delineate hierarchical temporal sequences, or 'bursts', where classified events are more or less

likely. For this we modified a version of the pyburst package available in SimBA (Appendix B). We analyzed the final machine learning results by calculating the total event duration, number of bouts, mean bout duration, and mean bout interval per video.

Statistical analysis

We visualized and calculated descriptive statistics of attack, anogenital sniffing, pursuit, escape, and defensive behaviors for three groups of mice using SimBA in-built functions (Fig. 5-6). For each classified behavior, we used SimBA to calculate total behavior duration, bouts per session, mean bout duration, and mean bout interval. For sex comparisons, the first five minutes of the male and female CSDS assays were used across five days. For environmental context comparisons (CSDS vs RI), the full ten minutes of RI and CSDS assays were used from the first three days of testing. Separate attack classifiers were built and used to analyze behavior for rats and each of the mouse labs participating in the SHAP consortium dataset. Data was analyzed in GraphPad Prism (v8.0.1) via t-tests, two-way ANOVAs, or linear mixed models as appropriate, and Bonferroni's correction was used for multiple comparisons. ($\alpha = 0.05$). * = $p < 0.05$, ** = < 0.01, *** = < 0.001 in all figures. Statistical comparisons are described in Appendix A.

Appendix 1.

> 0.05 < 0.05 < 0.001 < 0.0001

Attack present: CD1 movement features

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.1197±0.001
UW	Stanford	0.06429±0.001
UW	CRIM	0.02995±0.001
NYU	Stanford	0.184±0.001
NYU	CRIM	0.1496±0.001
Stanford	CRIM	-0.03434±0.001

Attack present: C57 movement features

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.09386±0.001
UW	Stanford	0.07733±0.001
UW	CRIM	0.1073±0.001
NYU	Stanford	0.1712±0.001
NYU	CRIM	0.2012±0.001
Stanford	CRIM	0.02999 ±0.001

Attack present: Animal movement

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.02477 ±0.0007
UW	Stanford	0.03076±0.0007
UW	CRIM	-0.09159 ±0.0007
NYU	Stanford	-0.05553 ±0.0007
NYU	CRIM	-0.06681±0.0007
Stanford	CRIM	-0.1223 ±0.0007

Attack present: Animal distances

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	0.0418 ±0.001
UW	Stanford	0.03645±0.001
UW	CRIM	-0.06837 ±0.001

NYU	Stanford	-0.005346 ±0.001
NYU	CRIM	-0.1102±0.001
Stanford	CRIM	-0.1048 ±0.001

Attack present: CD1 and C57 shape

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.002272 ±0.0003
UW	Stanford	-0.02139±0.0003
UW	CRIM	-0.005384 ±0.0003
NYU	Stanford	-0.01911 ±0.0003
NYU	CRIM	-0.003112±0.0003
Stanford	CRIM	0.016 ±0.0003

Attack present: C57 shape

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	0.03018 ±0.0007
UW	Stanford	0.02785±0.0007
UW	CRIM	-0.01878 ±0.0007
NYU	Stanford	-0.002322 ±0.0007
NYU	CRIM	-0.04896±0.0007
Stanford	CRIM	-0.04664 ±0.0007

Attack present: CD1 Shape

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.00615 ±0.001
UW	Stanford	-0.278±0.001
UW	CRIM	-0.03844 ±0.001
NYU	Stanford	-0.2718 ±0.001
NYU	CRIM	-0.03229±0.001
Stanford	CRIM	0.2395 ±0.001

Attack not present: CD1 and C57 shape

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.0004095 ±0.0001
UW	Stanford	0.004368±0.0001
UW	CRIM	-0.0005817 ±0.0001
NYU	Stanford	0.004777 ±0.0001
NYU	CRIM	-0.0001722±0.0001
Stanford	CRIM	-0.00495 ±0.0001

Attack not present: C57 movement

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.01766 ±0.0007
UW	Stanford	-0.01123±0.0007
UW	CRIM	-0.02797 ±0.0007
NYU	Stanford	0.006432 ±0.0007
NYU	CRIM	-0.01031 ±0.0007865
Stanford	CRIM	-0.01674 ±0.0007865

Attack not present: C57 shape

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.0118 ±0.000321
UW	Stanford	-0.008808±0.000321
UW	CRIM	-0.008811 ±0.000321
NYU	Stanford	0.002996 ±0.000321
NYU	CRIM	0.002992 ±0.000321
Stanford	CRIM	-0.000003544±0.000321

Attack not present: CD 1 movement

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.003512 ±0.0003705
UW	Stanford	-0.0102±0.0003705
UW	CRIM	0.0003151 ±0.0003705
NYU	Stanford	-0.006684 ±0.0003705
NYU	CRIM	-0.003827 ±0.0003705
Stanford	CRIM	0.01051 ±0.0003705

Attack not present: Animal distance

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.03349 ±0.001352
UW	Stanford	-0.008239±0.001352
UW	CRIM	-0.0339 ±0.001352
NYU	Stanford	0.02525 ±0.001352
NYU	CRIM	-0.0004122 ±0.001352
Stanford	CRIM	-0.02566 ±0.001352

Attack not present: CD1 and C57 movement

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.01814 ±0.000708
UW	Stanford	-0.00253±0.000708
UW	CRIM	-0.01875 ±0.000708
NYU	Stanford	0.01561 ±0.000708
NYU	CRIM	-0.0006114 ±0.000708
Stanford	CRIM	-0.01622 ±0.000708

Attack not present: CD1 shape

Variable 1	Variable 2	Mean difference (± SEM)
UW	NYU	-0.003723 ±0.000857
UW	Stanford	0.04717±0.000857
UW	CRIM	-0.00451 ±0.000857
NYU	Stanford	0.05089 ±0.000857
NYU	CRIM	-0.0007869 ±0.000857
Stanford	CRIM	-0.05168 ±0.000857

Consortium attack SHAP directionality statistics

SHAP bin	UW (R^2)	Stanford (R^2)	CRIM (R^2)	NYU (R^2)
Animal distances	0.4473	0.4890	0.4167	0.3176
Resident + intruder movement	0.4287	0.2859	0.3322	0.3493
Intruder movement	0.6611	0.5853	0.6253	0.6674
Resident shape	0.0221	0.6828	0.0002	0.1701
Resident movement	0.3611	0.3751	0.5934	0.4521
Intruder shape	0.3279	0.0396	0.2838	0.3395
Resident + intruder shape	0.0021	0.5308	0.1609	0.2280

Comparisons of rat and mouse SHAP values (Fig. 4)

Rat attack SHAP statistical analysis

Test	**Factor name**	**F-value**	***p*-value**
Two-way ANOVA main effects	Feature bin	F (6, 30) = 4.791	0.0016**
	Rolling window	F (5, 30) = 1.973	0.1116

Rat attack SHAP directionality analysis

SHAP bin	Rat (R^2)	Mouse (R^2)
Animal distances	0.2645	0.3445
Resident + intruder movement	0.4852	0.2568
Intruder movement	0.4940	0.5390
Resident shape	0.0915	0.0255
Resident movement	0.0265	0.2703
Intruder shape	0.5690	0.1863
Resident + intruder shape	0.1171	0.0799

Table 4.1. Male CSDS versus female CSDS statistical comparisons *(Fig. 5)*

Figure number	**Test**	**Factor name**	**F-value**	***p*-value**	**Multiple comparisons**
Figure 5D. Attack duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 119) = 1.059	0.3800	
Figure 5D. Attack duration averages	Unpaired t-test	Sex	df = 30	<0.001*	
Figure 5D. Anogen duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 119) = 0.1321	0.9704	
Figure 5D. Anogen duration averages	Unpaired t-test	Sex	df = 30	0.0461*	
Figure 5D. Pursuit duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Sex	F (4, 119) = 1.400	0.2381	
Figure 5D. Pursuit duration averages	Unpaired t-test	Sex	df = 30	<0.001*	
Figure 5D. Defensive duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Sex	F (4, 115) = 1.302	0.2735	
Figure 5D. Defensive duration averages	Unpaired t-test	Sex	df = 29	0.0374*	
Figure 5D. Escape duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Sex	F (4, 144) = 3.328 F (3.018, 108.6) = 1.059 F (1, 144) = 202.9	0.0122* 0.3700 <0.0001*	0.0174* Day 1 0.0008* Day 2 0.0118* Day 3 0.0072* Day 4 0.0026* Day 5
Figure 5D. Escape duration averages	Unpaired t-test	Sex	df = 29	<0.001*	
Figure 5D. Attack bouts by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 119) = 2.187	0.0745	

Figure 5D. Attack bouts averages	Unpaired t-test	Sex	df = 30	<0.001*	
Figure 5D. Anogen bouts by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 119) = 0.1134	0.9776	
Figure 5D. Anogen bouts averages	Unpaired t-test	Sex	df = 30	0.0298*	
Figure 5D. Pursuit bouts by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Sex	F (4, 119) = 2.587 F (3.693, 109.9) = 1.646 F (1, 30) = 77.10	0.0404* 0.1724 <0.0001*	0.0006* Day 1 0.0018* Day 2 0.0044* Day 3 0.0810 Day 4 0.0031* Day 5
Figure 5D. Pursuit bouts averages	Unpaired t-test	Sex	df = 30	<0.001*	
Figure 5D. Defensive bouts by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 115) = 0.4315	0.7857	
Figure 5D. Defensive bouts averages	Unpaired t-test	Sex	df = 29	0.0146*	
Figure 5D. Escape bouts by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Sex	F (4, 144) = 2.554 F (3.531, 127.1) = 1.019 F (1, 144) = 112.5	0.0415* 0.3947 <0.0001*	0.1349 Day 1 0.0001* Day 2 0.0039* Day 3 0.0633 Day 4 0.0051* Day 5
Figure 5D. Escape bouts averages	Unpaired t-test	Sex	df = 29	<0.001*	
Figure 5D. Attack bout duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 119) = 1.282	0.2811	

Figure 5D. Attack bout duration averages	Unpaired t-test	Sex	df = 30	0.3276	
Figure 5D. Anogen bout duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 119) = 1.165	0.3299	
Figure 5D. Anogen bout duration averages	Unpaired t-test	Sex	df = 30	0.8559	
Figure 5D. Pursuit bout duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 149) = 0.6693	0.6142	
Figure 5D. Pursuit bout duration averages	Unpaired t-test	Sex	df = 30	<0.001*	
Figure 5D. Defensive bout duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 144) = 2.357	0.0564	
Figure 5D. Defensive bout duration averages	Unpaired t-test	Sex	df = 29	<0.001*	
Figure 5D. Escape bout duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 115) = 1.397	0.2394	
Figure 5D. Escape bout duration averages	Unpaired t-test	Sex	df = 29	<0.001*	
Figure 5D. Attack bout interval by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 116) = 0.4298	0.7869	
Figure 5D. Attack bout duration averages	Unpaired t-test	Sex	df = 30	0.0880	
Figure 5D. Anogen bout interval by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 91) = 0.8302	0.5094	
Figure 5D. Anogen bout interval averages	Unpaired t-test	Sex	df = 30	0.0303*	

Figure 5D. Pursuit bout interval by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 116) = 0.8102	0.5211	
Figure 5D. Pursuit bout interval averages	Unpaired t-test	Sex	df = 30	0.0157*	
Figure 5D. Defensive bout interval by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 109) = 1.481	0.2130	
Figure 5D. Defensive bout interval averages	Unpaired t-test	Sex	df = 29	0.0013*	
Figure 5D. Escape bout interval by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (4, 111) = 0.2752	0.8934	
Figure 5D. Escape bout interval averages	Unpaired t-test	Sex	df = 29	0.0140*	
Figure 5D. Attack SHAP animal distances	Two-way ANOVA main effects	Time Sex	F (5, 5) = 10.65 F (1, 5) = 34.92	0.0107* 0.0020*	
Figure 5D. Attack SHAP intruder movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 142.1 F (1, 5) = 11.35	<0.0001* 0.0199*	
Figure 5D. Attack SHAP intruder & resident movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 7.667 F (1, 5) = 26.35	0.0216* 0.0037*	
Figure 5D. Attack SHAP resident movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 3.985 F (1, 5) = 3.740	0.0777 0.1109	
Figure 5D. Attack SHAP intruder shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 10.23 F (1, 5) = 418.0	0.0116* <0.0001*	
Figure 5D. Attack SHAP intruder & resident shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 2.800 F (1, 5) = 21.14	0.1414 0.0059*	
Figure 5D. Attack SHAP resident shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 1.823 F (1, 5) = 19.84	0.2630 0.0067*	
Figure 5D. Anogen SHAP animal distances	Two-way ANOVA main effects	Time Sex	F (5, 5) = 3.107 F (1, 5) = 0.006468	0.1195 0.9390	
Figure 5D. Anogen SHAP intruder movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 40.78 F (1, 5) = 14.13	0.0005* 0.0132*	

Figure 5D. Anogen SHAP intruder & resident movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 4.676 F (1, 5) = 9.741	0.0579 0.0262*	
Figure 5D. Anogen SHAP resident movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 7.733 F (1, 5) = 9.747	0.0212* 0.0262*	
Figure 5D. Anogen SHAP intruder shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 0.3145 F (1, 5) = 2.948	0.8850 0.1466	
Figure 5D. Anogen SHAP intruder & resident shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 3.893 F (1, 5) = 6.621	0.0811 0.0498*	
Figure 5D. Anogen SHAP resident shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 2.082 F (1, 5) = 0.5613	0.2200 0.4875	
Figure 5D. Pursuit SHAP animal distances	Two-way ANOVA main effects	Time Sex	F (5, 5) = 2.935 F (1, 5) = 2.818	0.1312 0.1541	
Figure 5D. Pursuit SHAP intruder movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 1.697 F (1, 5) = 1.971	0.2879 0.2193	
Figure 5D. Pursuit SHAP intruder & resident movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 2.380 F (1, 5) = 0.01491	0.1816 0.9076	
Figure 5D. Pursuit SHAP resident movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 0.3026 F (1, 5) = 2.480	0.8922 0.1761	
Figure 5D. Pursuit SHAP intruder shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 5.090 F (1, 5) = 20.64	0.0492* 0.0062*	
Figure 5D. Pursuit SHAP intruder & resident shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 13.30 F (1, 5) = 6.828	0.0065* 0.0475*	
Figure 5D. Pursuit SHAP resident shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 1.034 F (1, 5) = 2.018	0.4858 0.2147	
Figure 5D. Defensive SHAP animal distances	Two-way ANOVA main effects	Time Sex	F (5, 5) = 7.263 F (1, 5) = 0.1293	0.0242* 0.7338	
Figure 5D. Defensive SHAP intruder movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 7.323 F (1, 5) = 1.603	0.0238* 0.2612	
Figure 5D. Defensive SHAP intruder & resident movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 64.31 F (1, 5) = 34.39	0.0002* 0.0020*	

Figure 5D. Defensive SHAP resident movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 15.00 F (1, 5) = 4.945	0.0050* 0.0768	
Figure 5D. Attack SHAP intruder shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 5.488 F (1, 5) = 31.97	0.0425* 0.0024*	
Figure 5D. Defensive SHAP intruder & resident shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 1.203 F (1, 5) = 6.984	0.4221 0.0458*	
Figure 5D. Defensive SHAP resident shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 1.811 F (1, 5) = 1.837	0.2651 0.2333	
Figure 5D. Escape SHAP animal distances	Two-way ANOVA main effects	Time Sex	F (5, 5) = 1.418 F (1, 5) = 0.03812	0.3555 0.8529	
Figure 5D. Escape SHAP intruder movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 3.592 F (1, 5) = 3.060	0.0934 0.1407	
Figure 5D. Escape SHAP intruder & resident movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 1.604 F (1, 5) = 3.182	0.3084 0.1345	
Figure 5D. Escape SHAP resident movement	Two-way ANOVA main effects	Time Sex	F (5, 5) = 2.644 F (1, 5) = 9.386	0.1548 0.0280*	
Figure 5D. Escape SHAP intruder shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 1.040 F (1, 5) = 35.28	0.4835 0.0019*	
Figure 5D. Escape SHAP intruder & resident shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 1.260 F (1, 5) = 35.11	0.4029 0.0020*	
Figure 5D. Escape SHAP resident shape	Two-way ANOVA main effects	Time Sex	F (5, 5) = 2.071 F (1, 5) = 76.45	0.2217 0.0003*	

Male CSDS versus RI statistical comparisons (Fig. 6C)

Figure number	Test	Factor name	F-value	*p*-value	Multiple comparisons
Figure 6D. Attack duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Environment	F (2, 85) = 9.375 F (1.827, 77.64) = 4.261 F (1, 43) = 0.005583	0.0002* 0.0204* 0.9408	0.0508 Day 1 0.1796 Day 2 0.2135 Day 3
Figure 6D. Attack duration averages	Unpaired t-test	Environment	Df = 43	0.9804	
Figure 6D. Anogen duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Environment	F (2, 85) = 4.637 F (1.621, 68.89) = 3.422 F (1, 43) = 7.983	0.0123* 0.0478* 0.0071*	<0.0001* Day 1 0.4987 Day 2 >0.9999 Day 3
Figure 6D. Anogen duration averages	Unpaired t-test	Environment	Df = 43	0.0065*	
Figure 6D. Pursuit duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Environment	F (2, 85) = 3.820 F (1.797, 76.38) = 3.852 F (1, 43) = 141.4	0.0258* 0.0295* <0.0001*	<0.0001* Day 1 <0.0001* Day 2 <0.0001* Day 3
Figure 6D. Pursuit duration averages	Unpaired t-test	Environment	Df = 43	<0.0001*	
Figure 6D. Defensive duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Environment	F (2, 85) = 4.659 F (1.761, 74.82) = 2.575 F (1, 43) = 0.01712	0.0120* 0.0896 0.8965	0.4580 Day 1 0.1077 Day 2 >0.9999 Day 3
Figure 6D. Defensive duration averages	Unpaired t-test	Environment	Df = 43	0.8840	
Figure 6D. Escape duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Environment	F (2, 85) = 10.33 F (1.880, 79.88) = 2.948 F (1, 43) = 27.59	<0.0001* 0.0613 <0.0001*	>0.9999 Day 1 <0.0001* Day 2 <0.0001* Day 3

Figure 6D. Escape duration averages	Unpaired t-test	Environment	Df = 43	<0.0001*	
Figure 6D. Attack bouts by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Environment	F (2, 85) = 10.50 F (1.921, 81.66) = 5.734 F (1, 43) = 8.806	<0.0001* 0.0052* 0.0049*	>0.9999 Day 1 <0.0001* Day 2 0.0639 Day 3
Figure 6D. Attack bouts averages	Unpaired t-test	Environment	Df = 43	0.0042*	
Figure 6D. Anogen bouts by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Environment	F (2, 85) = 4.429 F (1.878, 79.82) = 2.928 F (1, 43) = 8.653	0.0148* 0.0625 0.0052*	0.0004* Day 1 0.6400 Day 2 >0.9999 Day 3
Figure 6D. Anogen bouts averages	Unpaired t-test	Environment	Df = 43	0.0047*	
Figure 6D. Pursuit bouts by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 85) = 2.721	0.0716	
Figure 6D. Pursuit bouts averages	Unpaired t-test	Environment	Df = 43	<0.0001*	
Figure 6D. Defensive bouts by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 85) = 4.774 F (1.800, 76.51) = 9.197 F (1, 43) = 12.50	0.0108* 0.0004* 0.0010*	>0.9999 Day 1 0.0001* Day 2 0.0585 Day 3
Figure 6D. Defensive bouts averages	Unpaired t-test	Environment	Df = 43	0.0009*	
Figure 6D. Escape bouts by day	Mixed-effects with Geisser-Greenhouse correction	Interaction Day Environment	F (2, 85) = 8.141 F (1.914, 81.36) = 6.085 F (1, 43) = 39.69	0.0006* 0.0039* <0.0001*	0.4343 Day 1 <0.0001* Day 2 <0.0001* Day 3
Figure 6D. Escape bouts averages	Unpaired t-test	Environment	Df = 43	<0.0001*	

Figure 6D. Attack bout duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 128) = 1.366	0.2588	
Figure 6D. Attack bout duration averages	Unpaired t-test	Environment	Df = 43	0.0003*	
Figure 6D. Anogen bout duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 128) = 1.758	0.1765	
Figure 6D. Anogen bout duration averages	Unpaired t-test	Environment	Df = 43	0.0810	
Figure 6D. Pursuit bout duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 128) = 0.0002332	0.9998	
Figure 6D. Pursuit bout duration averages	Unpaired t-test	Environment	Df = 43	0.5592	
Figure 6D. Defensive bout duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 128) = 0.05698	0.9446	
Figure 6D. Defensive bout duration averages	Unpaired t-test	Environment	Df = 43	<0.0001*	
Figure 6D. Escape bout duration by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 85) = 1.791	0.1730	
Figure 6D. Escape bout duration averages	Unpaired t-test	Environment	Df = 43	0.0008*	
Figure 6D. Attack bout interval by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 83) = 0.6215	0.5396	
Figure 6D. Attack bout duration averages	Unpaired t-test	Environment	Df = 43	0.4851	
Figure 6D. Anogen bout interval by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 111) = 4.950 F (1.974, 109.6) = 0.2576 F (1, 111) = 10.04	0.0087* 0.7705 0.0020*	0.0176* Day 1 0.3590 Day 2

					>0.9999 Day 3
Figure 6D. Anogen bout interval averages	Unpaired t-test	Environment	Df = 43	0.0037*	
Figure 6D. Pursuit bout interval by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 84) = 1.980	0.1444	
Figure 6D. Pursuit bout interval averages	Unpaired t-test	Environment	Df = 43	0.0067*	
Figure 6D. Defensive bout interval by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 85) = 2.159	0.1217	
Figure 6D. Defensive bout interval averages	Unpaired t-test	Environment	Df = 43	0.2169	
Figure 6D. Escape bout interval by day	Mixed-effects with Geisser-Greenhouse correction	Interaction	F (2, 84) = 3.476 F (1.132, 47.55) = 2.068 F (1, 43) = 0.7233	0.0355* 0.1553 0.3998	0.8226 Day 1 0.1818 Day 2 0.1838 Day 3
Figure 6D. Escape bout interval averages	Unpaired t-test	Environment	Df = 43	0.3897	
Figure 6D. Attack SHAP animal distances	Two-way ANOVA main effects	Time Environment	F (5, 5) = 3.821 F (1, 5) = 10.91	0.0838 0.0214*	
Figure 6D. Attack SHAP intruder movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 6.001 F (1, 5) = 15.76	0.0357* 0.0106*	
Figure 6D. Attack SHAP intruder & resident movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 7.954 F (1, 5) = 2.783	0.0200* 0.1561	
Figure 6D. Attack SHAP resident movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 6.242 F (1, 5) = 6.360	0.0329* 0.0530	
Figure 6D. Attack SHAP intruder shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 0.7648 F (1, 5) = 9.110	0.6121 0.0295*	
Figure 6D. Attack SHAP intruder & resident shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 22.23 F (1, 5) = 5.030	0.0020* 0.0750	
Figure 6D. Attack SHAP resident shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 3.203 F (1, 5) = 14.28	0.1136 0.0129*	

Figure 6D. Anogen SHAP animal distances	Two-way ANOVA main effects	Time Environment	F (5, 5) = 411.0 F (1, 5) = 4.713	<0.0001* 0.0820	
Figure 6D. Anogen SHAP intruder movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 125.5 F (1, 5) = 0.7301	<0.0001* 0.4319	
Figure 6D. Anogen SHAP intruder & resident movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 6.959 F (1, 5) = 8.071	0.0264* 0.0362*	
Figure 6D. Anogen SHAP resident movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 25.82 F (1, 5) = 0.3338	0.0014* 0.5885	
Figure 6D. Anogen SHAP intruder shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 1.002 F (1, 5) = 100.4	0.4993 0.0002*	
Figure 6D. Anogen SHAP intruder & resident shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 8.977 F (1, 5) = 51.01	0.0155* 0.0008*	
Figure 6D. Anogen SHAP resident shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 1.266 F (1, 5) = 117.6	0.4010 0.0001*	
Figure 6D. Pursuit SHAP animal distances	Two-way ANOVA main effects	Time Environment	F (5, 5) = 22.94 F (1, 5) = 7.569	0.0019* 0.0402*	
Figure 6D. Pursuit SHAP intruder movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 317.6 F (1, 5) = 2.665	<0.0001* 0.1635	
Figure 6D. Pursuit SHAP intruder & resident movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 1.672 F (1, 5) = 0.0003089	0.2932 0.9867	
Figure 6D. Pursuit SHAP resident movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 0.1357 F (1, 5) = 5.334	0.9766 0.0690	
Figure 6D. Pursuit SHAP intruder shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 6.646 F (1, 5) = 2.101	0.0290* 0.2069	
Figure 6D. Pursuit SHAP intruder & resident shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 19.10 F (1, 5) = 25.06	0.0028* 0.0041*	
Figure 6D. Pursuit SHAP resident shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 3.307 F (1, 5) = 27.06	0.1077 0.0035*	
Figure 6D. Defensive SHAP animal distances	Two-way ANOVA main effects	Time Environment	F (5, 5) = 39.43 F (1, 5) = 40.16	0.0005* 0.0014*	
Figure 6D. Defensive SHAP intruder movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 7.719 F (1, 5) = 1.913	0.0213* 0.2252	

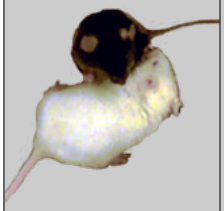
Figure 6D. Defensive SHAP intruder & resident movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 6.690 F (1, 5) = 8.981	0.0286* 0.0302*	
Figure 6D. Defensive SHAP resident movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 105.3 F (1, 5) = 7.207	<0.0001* 0.0436*	
Figure 6D. Attack SHAP intruder shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 0.6243 F (1, 5) = 11.53	0.6911 0.0194*	
Figure 6D. Defensive SHAP intruder & resident shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 17.61 F (1, 5) = 21.75	0.0034* 0.0055*	
Figure 6D. Defensive SHAP resident shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 12.68 F (1, 5) = 14.30	0.0072* 0.0129*	
Figure 6D. Escape SHAP animal distances	Two-way ANOVA main effects	Time Environment	F (5, 5) = 83.91 F (1, 5) = 1.790	<0.0001* 0.2385	
Figure 6D. Escape SHAP intruder movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 14.46 F (1, 5) = 1.640	0.0054* 0.2565	
Figure 6D. Escape SHAP intruder & resident movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 13.47 F (1, 5) = 6.711	0.0063* 0.0488*	
Figure 6D. Escape SHAP resident movement	Two-way ANOVA main effects	Time Environment	F (5, 5) = 4.411 F (1, 5) = 4.893	0.0646 0.0779	
Figure 6D. Escape SHAP intruder shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 3.241 F (1, 5) = 7.510	0.1113 0.0408*	
Figure 6D. Escape SHAP intruder & resident shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 14.79 F (1, 5) = 1.515	0.0051* 0.2731	
Figure 6D. Escape SHAP resident shape	Two-way ANOVA main effects	Time Environment	F (5, 5) = 190.8 F (1, 5) = 3.248e-030	<0.0001* >0.9999	

Table 4.2. Currently available open-source platforms for automated behavioral detection

Year	Name	Citation	Validated species	Behavioral Classifiers validated in the original publication	Software website
Supervised					
2009	CADABRA	(Dankert et al. 2009)	Fly	Lunging, tussling, wing threat/extension, circle, chase, copulation	http://www.vision.caltech.edu/cadabra/
2012	MiceProfiler	(de Chaumont et al. 2012)	Mouse	Head-head contact, anogenital contact, side-by-side, approach, leave, follow, chase	http://icy.bioimageanalysis.org/plugin/mice-profiler-tracker/
2013	JAABA	(Kabra et al. 2013)	Fly and mouse	Fly: walk, stop, crabwalk, backup, touch, chase, jump, copulation, wing flick/groom/extension, right, center pivot, tail pivot. Mouse: follow, walk.	https://sourceforge.net/projects/jaaba/files/
2013	Unamed*	Giancardo et al., 2013(Hong et al. 2015)	Mouse	Nose to body, nose to nose, nose to genitals, above, follow, stand together	
2015	Unamed*	(Hong et al. 2015)	Mouse	Attack, close investigation, mounting	
2020	SimBA	Nilsson et al. 2020	Mouse and rat	Attack, pursuit, lateral threat, anogenital sniff, mounting, upright submissive, allogroom, flee, scramble, box, approach, avoidance, drink, clean, eat, rear, walk away, circle	https://github.com/sgoldenlab/simba
2021	MARS	Segalin et al. 2021	Mouse	Attack, mount, close investigation, face-directed, genital-directed,	https://github.com/neuroethology/MARS
2022	DeepEthogram	Bohnslav et al. 2022	Fly and mouse	Nose-to-nose, nose-to-body, body-to-body, chase, anogenital	https://github.com/jbohnslav/deepethogram
2022	DeepCaT-z	Gerós et al. 2022	Rodent	Standstill, rear, walk, groom	https://github.com/AnaGeros/Deep-CaT-z-Software
2022	DeepAction	Harris et al. 2022	Mouse	Walk, rest, rear, mover, hang, groom, eat, drink, attack, approach, sniff, and copulation	https://github.com/carlwharris/DeepAction
2023	LabGym	Hu et al. 2023	Fly, larva, mouse and rat	Nest build, curl, uncoil, abdomen bend, wing extension, walk, rear, facial groom, sniff, ear groom, sit	https://github.com/umyelab/LabGym
Semi-supervised					
2022	SIPEC	Marks et al. 2022	Mouse and primate	Social groom, search, object interaction	https://github.com/SIPEC-Animal-Data-Analysis/SIPEC
Self-supervised					
2009	Ctrax	(Branson et al. 2009)	Fly	Walk, stop, sharp turn, crabwalk, backup, touch, chase	http://ctrax.sourceforge.net/
2021	TREBA	Sun et al. 2021	Fly and mouse	Sniff, mount, attack, lunge, tussle, wing extension	https://github.com/neuroethology/TREBA
Unsupervised					
2014	MotionMapper	Berman et al. 2014	Fly	Groom: wing, leg, abdomen, head. Wing waggle, run	https://github.com/gordonberman/MotionMapper

2017	DuoMouse	(Arakawa et al. 2017)	Mouse	Sniff, follow, indifferent	http://www.mgrl-lab.jp/DuoMouse.html
2019	LiveMouseTracker	de Chaumont et al. 2019	Mouse	Rear, look up, look down, contact	https://livemousetracker.org
2020	AlphaTracker	Chen et al. 2020	Mouse	Unnamed individual and social behavior clusters	https://github.com/ZexinChen/AlphaTracker
2021	Behavior Atlas	Huang et al, 2021	Mouse	Clusters identified: Walk, run, rear, trot, step, sniff, up stretch, left turn, right turn, rise, fall, dive	https://behavioratlas.tech/
2022	VAME	Luxem et al. 2022	Mouse	Motifs identified: exploration, turn, stationary, walk to rear, walk, rear, unsupported rear, groom, backward	https://github.com/LINCellularNeuroscience/VAME
2022	DBscorer	Nandi et al. 2022	Mouse and rat	Immobility	https://github.com/swanandlab/DBscorer
Heuristics					
2022	BehaviorDEPOT	Gabriel et al. 2022	Mouse	Freeze, jump, rear, escape, locomotion, novel object exploration	https://github.com/DeNardoLab/BehaviorDEPOT

Table 4.3. (Figure S1). Operational behavioral definitions for UW mouse classifiers

Classifier	Description	Start frame	Duration of behavior	End frame
Attack 	Clear physical antagonistic interaction initiated by the Resident.	Resident makes physical antagonistic contact with Intruder. Typically characterized by outstretched Resident forepaw(s) contacting the Intruder, while the Resident has an open mouth to initiate a bite. Can also be characterized by the first frame of a slap or quick bite without the forepaws being outstretched.	Resident remains oriented toward Intruder. Attacks can include tussling, biting, boxing, and corralling as part of the bout.	Resident orients away from Intruder. Typically, this is a slight turning of the head to look in a different direction, followed by a relaxation of the body and moving away from the Intruder.
Anogenital sniffing 	Resident is sniffing the anogenital region of the Intruder. Resident must be sniffing at base of tail, not further up on back or on legs.	Resident mouse is clearly sniffing anogenital region of Intruder, rather than side, back, or leg.	Uninterrupted sniffing of anogenital region.	Resident moves head away from anogenital region, either to move away from Intruder or to sniff non-anogenital region.
Lateral threat 	Resident is in close proximity (< one body length away from the Intruder) to face of Intruder with back arched and side displayed toward Intruder. Ears are often pinned with shoulder and side of face nearest to Intruder tilted toward ceiling.	Resident mouse orients side to Intruder and tilts front of body toward Intruder.	Resident will often circle Intruder or move front half side to side in front of Intruder, feigning attacks prior to actual attack.	Lateral threat posture is dropped. Animal will shift head away to look at other target or will begin an attack. If animal does not attack, the back posture will relax.
Pursuit 	Resident is following in the Intruder's path as the Intruder moves away from the Resident.	Resident is moving toward Intruder as Intruder moves away. Typically, this is characterized by the Intruder running away after an attack ends and the Resident follows directly after, or when the Intruder walks past the Resident and the Resident markedly changes directions to pursue the Intruder.	The Intruder is quickly moving away from the Resident, typically in a straight line until it reaches the edge of the area, at which point it will turn to face and watch the Resident. Resident is moving along the same path as the Intruder without distraction.	Either mouse stops moving, or the Resident deviates from path of the Intruder.
Mounting 	Can occur between same sex animals. Resident mounts intruder and paw movement and pelvic thrusts may be present.	Resident covers Intruder's back and tucks pelvis.	Resident continues to cover Intruder, can show pelvic thrusts but is not biting or otherwise attacking Intruder. Resident can continue trying to move away but Resident does not let go.	Resident untucks pelvis, starts to attack Intruder, or moves to the side.

Allogroom 	Resident mouse is grooming the Intruder.	Resident begins to groom the Intruder's fur.	Resident is often in a sitting position and head can be seen moving slightly side to side as it grooms the Intruder's fur.	Resident looks away, stops grooming, or starts biting the Intruder.
Escape 	Intruder mouse is running away from or attempting to get away from Resident mouse.	Intruder turns head and upper body away from Resident, attempting to quickly move away from Resident.	Intruder can be freely running away, can be biting or dragging Resident behind it, or can be "popcorning" trying to jump over the Resident when cornered.	Intruder stops trying to move away from Resident.
 **Defensive**	Intruder mouse is fighting back against Resident by pushing or biting. Intruder is not instigating aggression.	Intruder pushes, bites, or otherwise makes defensive physical contact with Resident.	Intruder continues to fight back against Resident. Intruder is most often facing and looking at Resident. This behavior can occur standing or when Intruder has been pushed to the ground.	Intruder stops fighting back, often transitioning into submissive or escape behaviors.

Table 4.4. Figure S2. SHAP values for positive frames of UW mouse classifiers used in Fig. 5-7

Attack SHAP (Positive frames)	0ms	66ms	133ms	166ms	200ms	500ms
Animal distances	0.0405	0.0377	0.0343	0.0370	0.0331	0.0194
Intruder movement	0.0173	0.0148	0.0326	0.0378	0.0482	0.0939
Resident + intruder movement	0.0011	0.0074	0.0179	0.0291	0.0365	0.0968
Resident movement	0.0030	0.0039	0.0059	0.0066	0.0076	0.0184
Intruder shape	0.0061	0.0063	0.0081	0.0082	0.0088	0.0104
Resident + intruder shape	0.0006	0.0011	0.0013	0.0015	0.0016	0.0025
Resident shape	0.0026	0.0036	0.0042	0.0044	0.0049	0.0066

Pursuit SHAP (Positive frames)	0ms	66ms	133ms	166ms	200ms	500ms
Animal distances	0.0704	0.0187	0.0192	0.0184	0.0187	0.0195
Intruder movement	0.0042	0.0048	0.0136	0.0238	0.0207	0.0721
Resident + intruder movement	0.0011	0.0074	0.0197	0.0274	0.0390	0.0781
Resident movement	0.0056	0.0051	0.0072	0.0078	0.0106	0.0200
Intruder shape	0.0125	0.0116	0.0144	0.0151	0.0165	0.0215
Resident + intruder shape	0.0059	0.0123	0.0150	0.0147	0.0154	0.0154
Resident shape	0.0109	0.0128	0.0147	0.0169	0.0173	0.0263

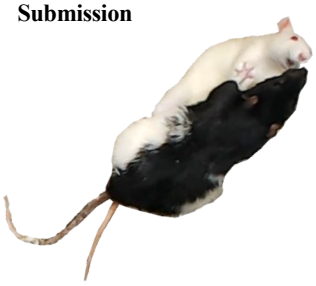
Anogen SHAP (Positive frames)	0ms	66ms	133ms	166ms	200ms	500ms
Animal distances	0.2393	0.0340	0.0336	0.0338	0.0335	0.0292
Intruder movement	0.0023	0.0050	0.0086	0.0106	0.0132	0.0216
Resident + intruder movement	0.0010	0.0054	0.0094	0.0121	0.0169	0.0314
Resident movement	0.0034	0.0053	0.0104	0.0123	0.0151	0.0234
Intruder shape	0.0109	0.0104	0.0111	0.0122	0.0128	0.0184
Resident + intruder shape	0.0037	0.0075	0.0081	0.0081	0.0086	0.0114
Resident shape	0.0201	0.0151	0.0168	0.0172	0.0178	0.0217

Defensive SHAP (Positive frames)	0ms	66ms	133ms	166ms	200ms	500ms
Animal distances	0.1306	0.0618	0.0581	0.0532	0.0461	0.0417
Intruder movement	0.0039	0.0042	0.0052	0.0060	0.0063	0.0174
Resident + intruder movement	0.0043	0.0168	0.0234	0.0253	0.0298	0.0556
Resident movement	0.0045	0.0031	0.0041	0.0047	0.0061	0.0149
Intruder shape	0.0164	0.0133	0.0130	0.0127	0.0125	0.0128
Resident + intruder shape	0.0027	0.0057	0.0067	0.0074	0.0075	0.0081
Resident shape	0.0059	0.0101	0.0135	0.0146	0.0154	0.0203

Escape SHAP (Positive frames)	0ms	66ms	133ms	166ms	200ms	500ms
Animal distances	0.0449	0.0104	0.0121	0.0125	0.0139	0.0253

Intruder movement	0.0237	0.0405	0.0585	0.0615	0.0669	0.0582
Resident + intruder movement	0.0030	0.0158	0.0327	0.0458	0.0526	0.0516
Resident movement	0.0021	0.0021	0.0025	0.0030	0.0037	0.0068
Intruder shape	0.0107	0.0109	0.0131	0.0137	0.0134	0.0210
Resident + intruder shape	0.0038	0.0088	0.0099	0.0097	0.0091	0.0094
Resident shape	0.0036	0.0057	0.0067	0.0071	0.0074	0.0109

Table 4.5. Figure S3. Operational behavioral definitions for WSU rat classifiers

Classifier	Description	Start Frame	Duration of behavior	End Frame
Attack 	Clear physical antagonistic interaction initiated by the Resident rat. Pauses in movement can occur if the Resident rat remains oriented toward the Intruder.	First frame when the Resident rat makes movement towards the Intruder that ends in a physical attack. A physical attack is typically characterized by outstretched Resident forepaw(s) contacting the Intruder, while the Resident has an open mouth to initiate a bite. Can also be characterized by a quick threatening movement resulting in the rearing up of both rats.	Attacks can include tussling, biting, pinning, boxing, and corralling as part of the attack bout. Breaks in movement may be present in attack behavior. This typically occurs after boxing where both rats are reared up and Resident is oriented towards the Intruder. Breaks in movement can also occur when the Resident has the Intruder pinned in a submissive posture.	First frame when Resident rat orients away from Intruder. Typically, this is a slight turning of the head to look in a different direction, followed by a relaxation of the body and moving away from the Intruder.
Submission 	The Intruder rat is in a supine posture and is pinned down by the Resident rat. The Intruder can be pinned in a corner or against a wall, but the back of the Intruder is close to horizontal with the bottom of the cage.	First frame when the Intruder rat assumes a supine posture and continues to be pinned down for at least 1 second.	Struggling usually occurs at the beginning of the submissive posture and is typically followed by breaks in movement with the Resident rat still on top of the Intruder. More bouts of struggling typically happen, and in some cases, may result in the moving of the interaction to a different location during which a submissive posture is still assumed by the Intruder.	First frame when the Intruder rat is no longer in a supine position. Typically characterized by moving upright onto its hind paws or on its side.
Lateral threat 	Resident rat is in proximity (typically < one body length away from the Intruder) to face of Intruder with back arched and side displayed toward Intruder. Ears are often pinned with shoulder and side of face nearest to Intruder rat tilted slightly toward ceiling.	First frame when Resident rat initiates the movement to orient towards the side of the Intruder, arches back, and tilts front of body toward Intruder.	Resident will often move toward the Intruder or move front half side to side in front of Intruder, feigning attacks prior to actual attack. Can also occur without initiation of an attack upon approach of the Intruder rat as a defensive behavior.	First frame when lateral threat posture is dropped. Animal will shift head away to look at other target or will begin an attack. If animal does not attack, the back posture will relax. Can occur by the Intruder rat increasing proximity to less than one body length from Resident.
Anogenital sniffing 	Either rat is sniffing the anogenital region of the other. The rat must be sniffing around the base of tail rather than the back or legs.	First frame when either rat is clearly sniffing the other's anogenital region near the base of the tail. In addition to what is pictured, anogenital sniffing can occur when one rat approaches from the front or side and crawls under the other to	Uninterrupted sniffing of anogenital region.	First frame when the rat moves head away from anogenital region, usually either to move away from Intruder or sniff non-anogenital region.

investigate.

Table 4.6. Figure S4. SHAP values for rat attack classifier (See Fig. 4)

Attack SHAP (Positive frames)	0ms	66ms	133ms	166ms	200ms	500ms
Animal distances	0.001397	0.001544	0.002811	0.003851	0.004815	0.015507
Intruder movement	0.008883	0.001855	0.003253	0.004434	0.00524	0.01376
Resident + intruder movement	0.002091	0.006047	0.006217	0.00688	0.007653	0.010801
Resident movement	0.000943	0.004572	0.008186	0.010696	0.013132	0.03583
Intruder shape	0.08918	0.012769	0.012295	0.012219	0.013057	0.015059
Resident + intruder shape	0.051291	0.018518	0.018442	0.020436	0.020672	0.028871
Resident shape	0.032857	0.025453	0.028869	0.033255	0.033684	0.051612

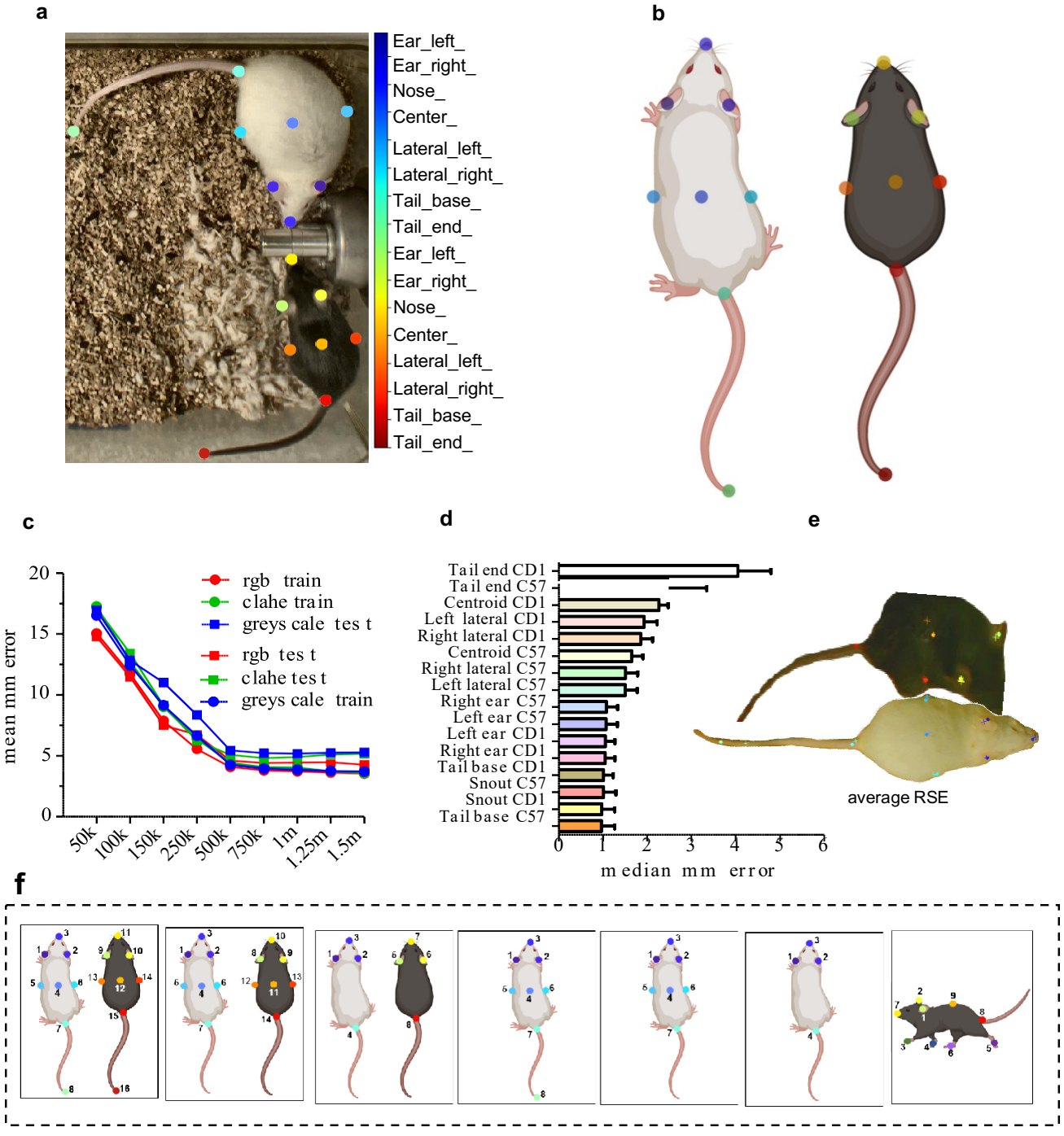
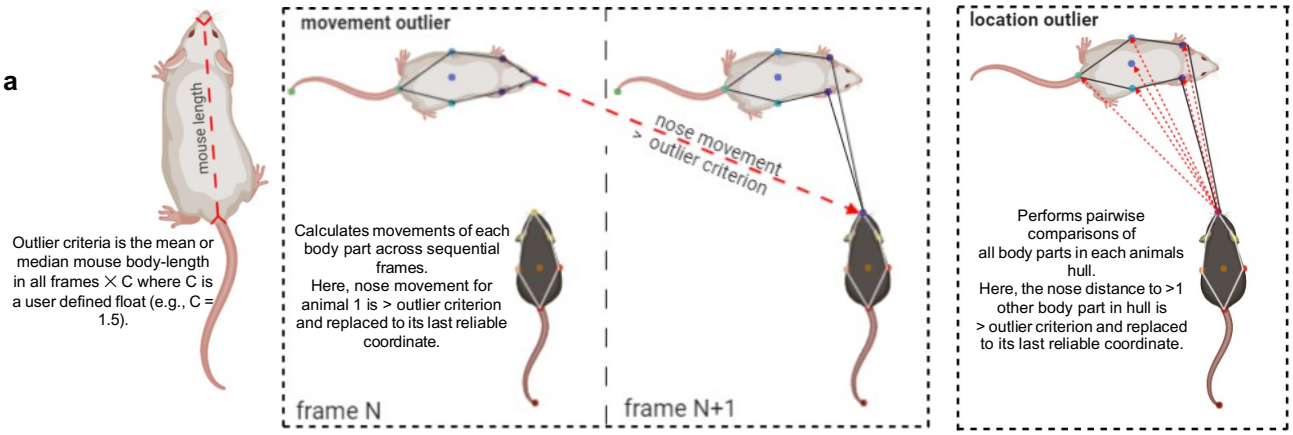


Figure 4.7. Figure S5. Example of DeepLabCut pose-estimation model for mouse resident-intruder behavior. *Example of DeepLabCut pose-estimation model for mouse resident-intruder behavior. (a) The 16 body-parts labeled. (b) Schematic depiction of the location of each of the 16 body part labels. (c) Evaluations of three models (rgb, clahe, greyscale) using the DeepLabCut evaluation tool. Pixel distances were converted to millimeter by using the lowest resolution images in the dataset (1000x1544px; 4.6px/millimeter). (d) Median millimeter error per body part. (e) Image representing the relative standard error (RSE) of the median millimeter error across all test images. The labelled images and DeepLabCut generated weights are available to download on the Open Science Framework, osf.io/mutws. (f) SimBA supports a range of alternative body-part settings for single animals and dyadic protocols through the File-> Create Project menu. Note: tail end tracking performance was insufficient for a tail rattle classifier, and the tail end body parts were dropped for all analysis in the main figures.*

a



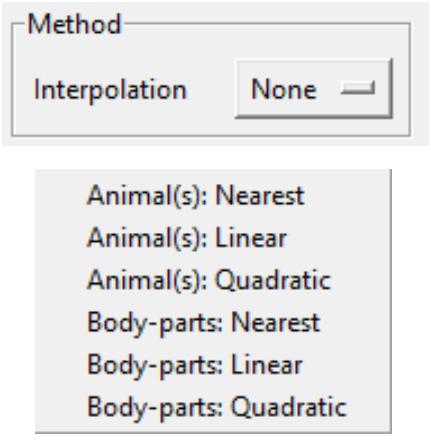
b

Body part	# frames	% of total
C57 tail end	24988	11.19
CD1 tail end	12261	5.49
CD1 right ear	2802	1.25
C57 tail base	1993	0.89
CD1 tail base	1423	0.64
C57 right ear	1141	0.51
CD1 centroid	268	0.12
C57 centroid	219	0.1
C57 left ear	155	0.07
CD1 left ear	141	0.06
CD1 lateral left	79	0.03
CD1 lateral right	40	0.02
C57 lateral left	53	0.02
C57 lateral right	51	0.02
C57 nose	35	0.02
CD1 nose	22	0.01

c

Body part	# frames	% of total
C57 tail base	3251	1.45
CD1 nose	2876	1.29
CD1 tail base	1624	0.73
CD1 left ear	625	0.28
CD1 right ear	565	0.25
CD1 lateral left	394	0.18
CD1 lateral right	375	0.17
CD1 centroid	350	0.16
C57 nose	360	0.16
C57 left ear	115	0.05
C57 right ear	88	0.04
C57 centroid	12	0.01
C57 lateral left	23	0.01
C57 lateral right	32	0.01

d



e

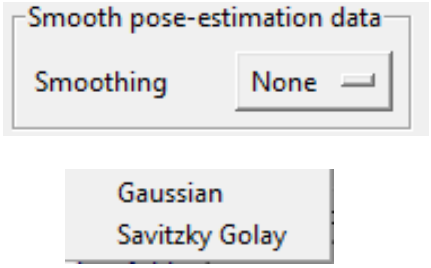


Figure 4.8. Figure S6. SimBA outlier correction options. *(a) SimBA calculates the mean or median distance between two user-defined body-parts across the frames of each video. We set the user-defined body-parts to be the nose and the tail-base of each animal. The user also defines a movement criterion value, and a location criterion value. We set the movement criterion to 0.7, and location criterion to 1.5. Two different outlier criteria are then calculated by SimBA. These criteria are the mean length between the two user-defined body parts in all frames of the video, multiplied by the either user-defined movement criterion value or location criterion value. SimBA corrects movement outliers prior to correcting location outliers. (b) Schematic representations of a pose-estimation body-part 'movement outlier' (top) and a 'location outlier' (bottom). A body-part violates the movement criterion when the movement of the body-part across sequential frames is greater than the movement outlier criterion. A body-part violates the location criteria when its distance to more than one other body-part in the animals' hull (except the tail-end) is greater than the location outlier criterion. Any body part that violates either the movement or location criterion is corrected by placing the body-part at its last reliable coordinate. (c) The ratio of body-part movements (top) and body-part locations (bottom) detected as outliers and corrected by SimBA in the RGB-format mouse resident-intruder data-set. For the outlier corrected in rat and the CRIM13 datasets, see the SimBA GitHub repository. We also offer (d) interpolation options for frames with missing body parts and (3) smoothing options to reduce frame-to-frame jitter.*

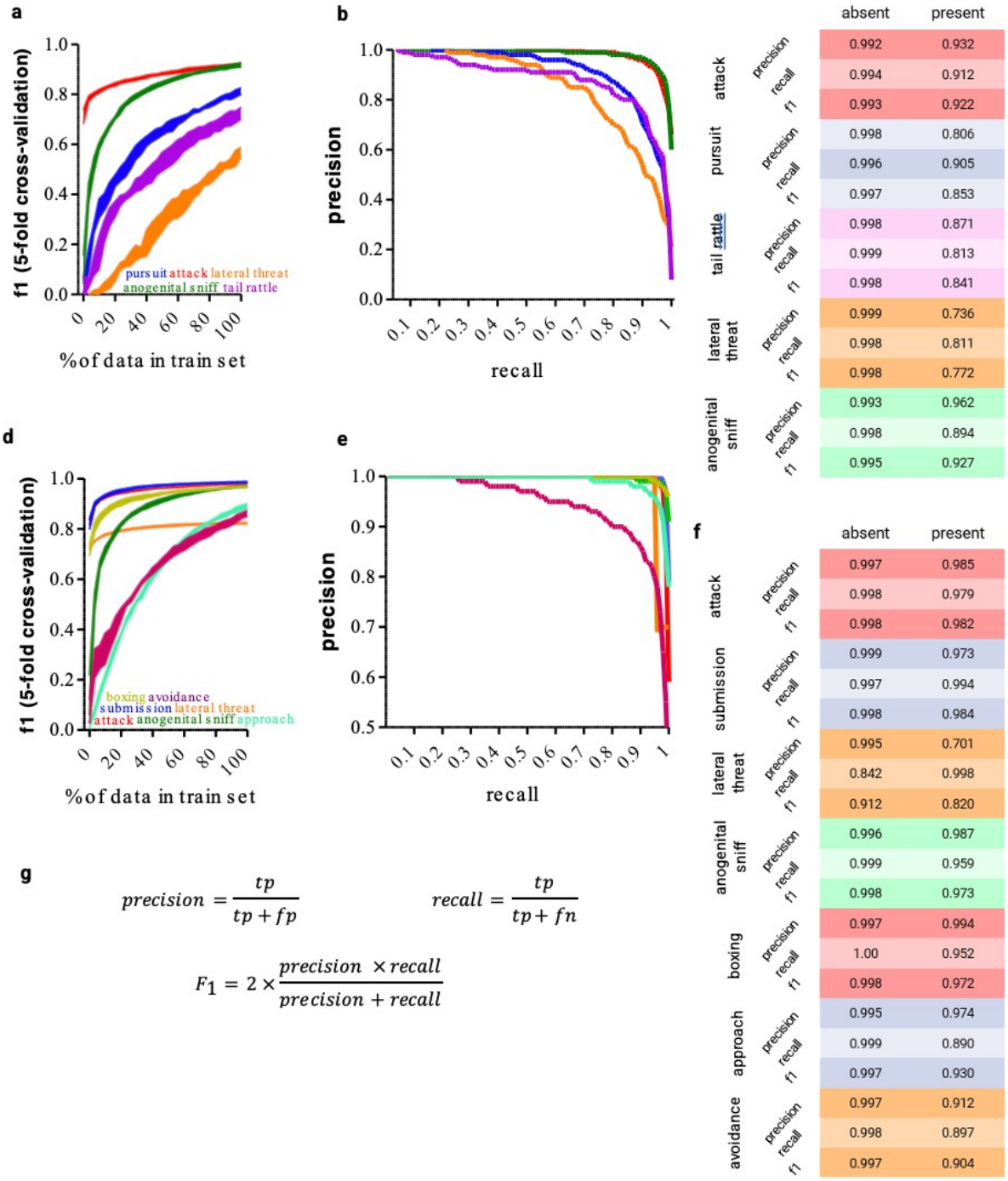


Figure 4.9. Figure S7. Evaluations of the mouse (a-c) and rat resident-intruder (d-f) models included in the original SimBA preprint. *All evaluations are presented for frame-by-frame predictions on 30fps videos. (a) Mouse resident intruder f1 learning curves with 5-fold cross-validation based 30 on train-set sizes performed prior to data balancing. (b) Precision-recall curves for the mouse resident-intruder dataset. (c) Precision, recall, and f1 for framewise predicted presence and absence of behaviors in the mouse resident-intruder dataset. (d) Rat resident-intruder f1 learning curves with 5-fold cross-validation based on 30 train-set sizes performed prior to data balancing. (e) Precision-recall curves for the rat resident-intruder dataset. (f) Precision, recall, and f1 for framewise predicted presence and absence of behaviors in the rat resident-intruder dataset. (g) Performance metrics equations, tp = true positive, fp = false positive, fn = false negative. Classification thresholds are 0.5 throughout.*

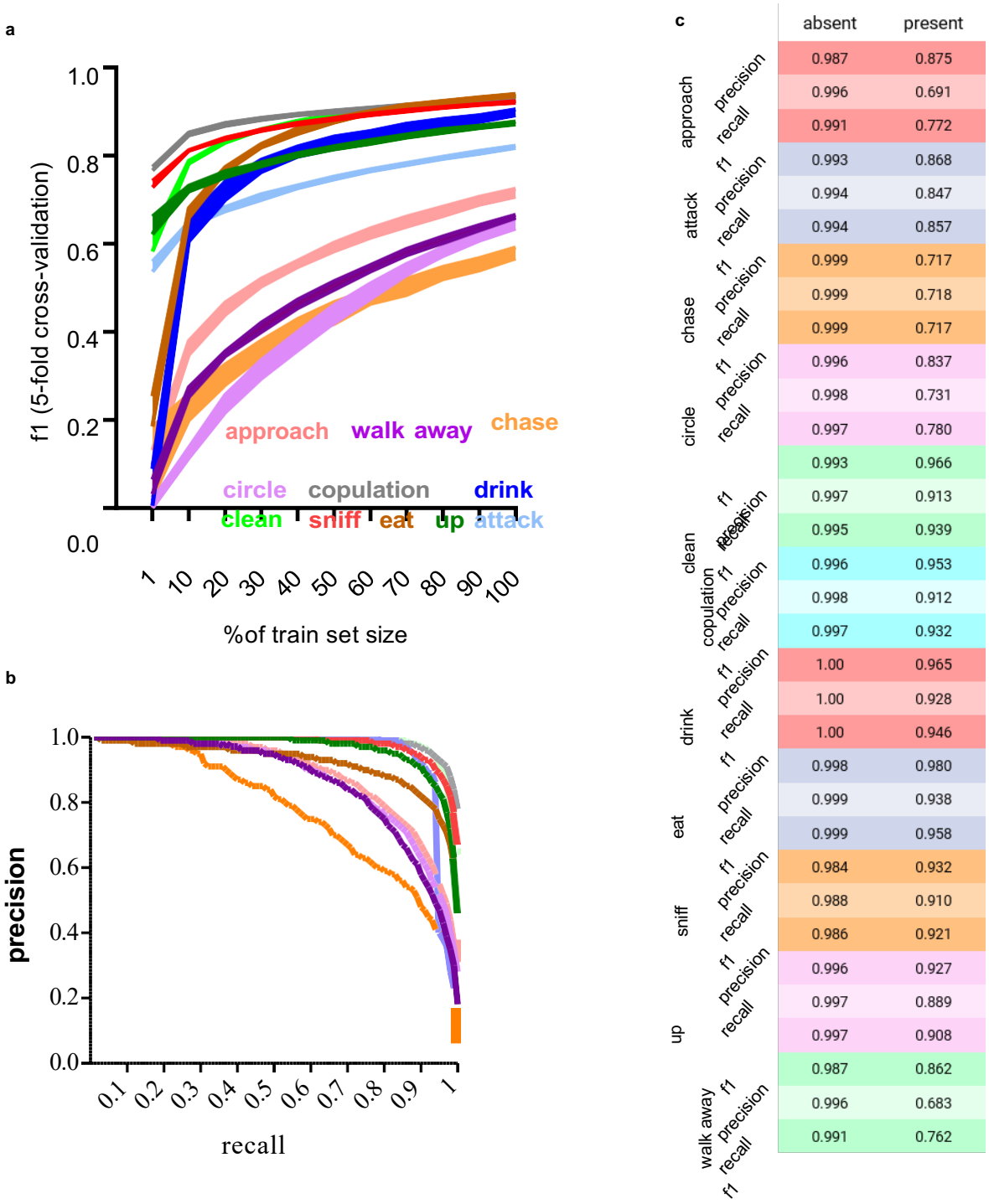


Figure 4.10. Figure S8. CRIM13 classifier performance. *Evaluations of the CRIM13 resident-intruder models. We built the models using 65 videos containing non-anesthetized black and white coat-colored animals. All evaluations are presented for frame-by-frame predictions on 25fps videos. (a) f1 learning curves with 5-fold cross-validation based on 10 training set sizes performed prior to data balancing. (b) Precision-recall curves for the CRIM13 dataset. (c) Precision, recall, and f1 for framewise predicted presence and absence of behaviors in the CRIM13 dataset. Classification thresholds are 0.5 throughout.*

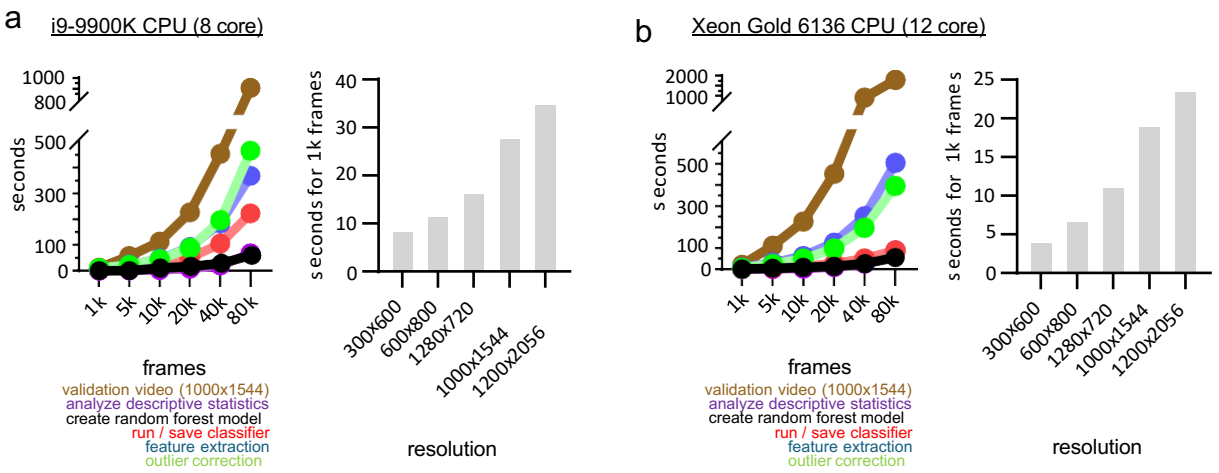
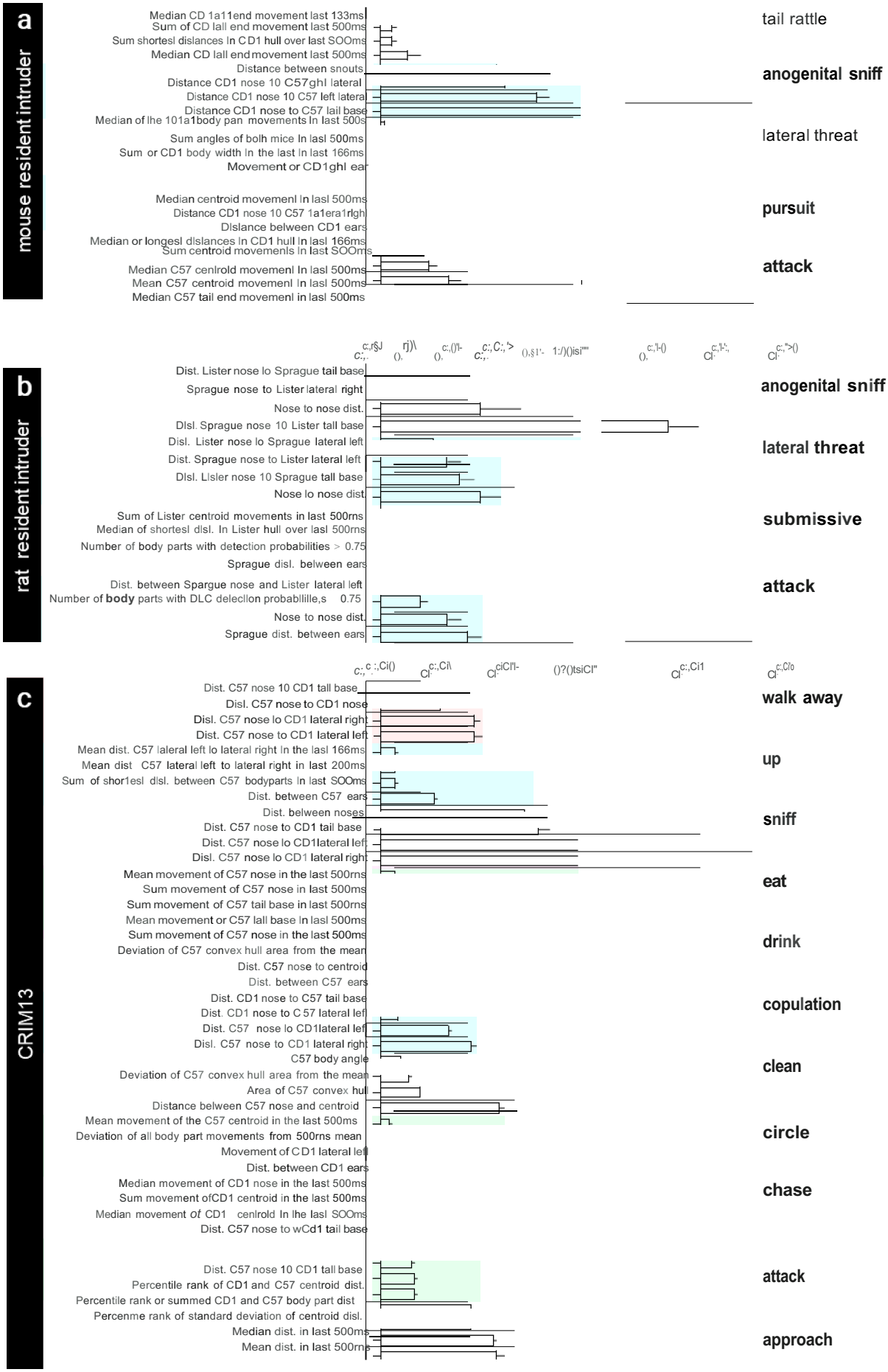


Figure 4.11. Figure S9. Approximate procedural runtimes for processing diOerent sized data-sets in SimBA *using (a) an 8-core Intel i9 CPU and (b) a 12-core Xeon Gold CPU. Time in seconds to perform outlier corrections, feature extraction using 16 body-parts, generating a random forest with 2k trees, performing / saving machine learning classifications, calculating descriptive statistics of machine classifications, generating a validation video, and extracting individual image frames from a video recorded at six diXerent resolutions (videos recorded at 975kbps bitrate). Note: only runtimes for creating a validation video and an extracting frames depend on the resolution of the videos. See the SimBA GitHub repository for more information.*


a
mouse resident intruder
Median CD 1a11end movement last 133ms
Sum of CD lall end movement last 500ms
Sum shortesl dislances In CD1 hull over last SOOms
Median CD lall end movement last 500ms
Distance between snouts
Distance CD1 nose 10 C57ghI lateral
Distance CD1 nose 10 C57 left latera
Distance CD1 nose to C57 lail base
Median of lhe 101a1body pan movements In last 500s
Sum angles of bolh mice In lasl 500ms
Sum or CD1 body width In the last In last 166ms
Movement or CD1ghI ear
Median centroid movemenI In lasl 500ms
Distance CD1 nose 10 C57 1a1era1rIgh
DIslance belween CD1 ears
Median or longesl dIslances In CD1 hull In lasl 166ms
Sum centroid movemenIs In last SOOms
Median C57 cenIrold movemenI In lasl 500ms
Mean C57 centroid movemenI In lasl 500ms
Median C57 tail end movemenI in lasl 500ms
tail rattle
anogenital sniff
lateral threat
pursuit
attack
b
rat resident intruder
Dist. Lister nose lo Sprague tail base
Sprague nose to Lister lateral right
Nose to nose dist.
DIsl. Sprague nose 10 Lister tall base
Disl. Lister nose lo Sprague lateral left
Dist. Sprague nose to Lister lateral left
DIsl. Lisler nose 10 Sprague tall base
Nose lo nose dist.
Sum of Lister centroid movements in last 500rns
Median of shortesl dIsl. In Lister hull over lasl 500rns
Number of body parts with detection probabilities > 0.75
Sprague disl. belween ears
Dist. between Spargue nose and Lister lateral left
Number of body parts with DLC deleclIon probablIlle,s 0.75
Nose to nose dist.
Sprague dist. between ears
anogenital sniff
lateral threat
submissive
attack
c
CRIM13
Dist. C57 nose 10 CD1 tall base
DisI. C57 nose to CD1 nose
Disl. C57 nose lo CD1 lateral right
Dist. C57 nose to CD1 lateral left
Mean dist. C57 laleral left lo lateral right In the lasl 166ms
Mean dist C57 lateral left to lateral right in last 200ms
Sum of shor1esl dIsl. between C57 bodyparts In last SOOms
Dist. between C57 ears
Dist. belween noses
Dist. C57 nose to CD1 tail base
Dist. C57 nose lo CD1lateral left
Disl. C57 nose lo CD1 lateral right
Mean movement of C57 nose in the last 500rns
Sum movement of C57 nose in last 500ms
Sum movement of C57 tail base in last 500rns
Mean movement or C57 lail base In lasl 500ms
Sum movement of C57 nose in the last 500ms
Deviation of C57 convex hull area from the mean
Dist. C57 nose to centroid
Dist. between C57 ears
Dist. CD1 nose to C57 tail base
Dist. CD1 nose to C 57 lateral lef
Dist. C57 nose lo CD1lateral lef
Disl. C57 nose to CD1 lateral right
C57 body angle
Deviation of C57 convex hull area from the mean
Area of C57 convex hull
Distance belween C57 nose and centroid
Mean movement of the C57 centroid in the last 500ms
Deviation of all body part movements from 500rns mean
Movement of CD1 lateral lef
Dist. between CD1 ears
Median movement of CD1 nose in the last 500ms
Sum movement ofCD1 centroid in the last 500ms
Median movement ot CD1 cenIrold In lhe lasl SOOms
Dist. C57 nose to wCd1 tail base
Dist. C57 nose 10 CD1 tall base
Percentile rank of CD1 and C57 centroid dist.
Percentile rank or summed CD1 and C57 body part dist
Percenme rank of standard deviation of centroid disl.
Median dist. in last 500ms
Mean dist. in last 500rns
walk away
up
sniff
eat
drink
copulation
clean
circle
chase
attack
approach

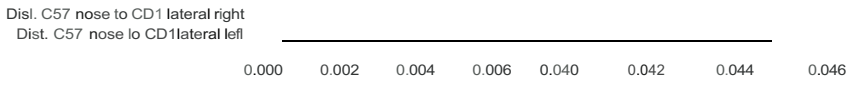

Disl. C57 nose to CD1 lateral right
Dist. C57 nose lo CD1lateral lefl
0.000
0.002
0.004
0.006
0.040
0.042
0.044
0.046

Figure 4.12. Figure S10. Feature contributions to classifiers calculated by permutation importance scores (Breiman, 2001). (*a) mouse resident intruder, (b) rat resident-intruder, and (c) CRIM13 data-sets. Feature permutation importance represents the performance classification degradation when the specific feature, and no other feature, is scrambled. A complete list of feature permutation and gini importance's are available through the SimBA OSF repository.*

	classifier	# annotated frames	# annotated videos	annotations: % present	annotations: present (hh:mm:ss)	annotations: absent (hh:mm:ss)	under sample ratio	test set: frames present	test set: frames absent	optimal threshold
mouse resident-intruder	allogroom normal	334680	26	1.49	00:02:46	03:03:10	19	983	65953	0.48
	allogroom vigorous	334680	26	0.73	00:01:21	03:04:35	0	494	66442	0.41
	attack	203841	37	4.93	00:05:35	01:47:40	9	1920	38849	0.56
	lateral threat	168738	21	1.49	00:01:24	01:32:21	18	503	33245	0.52
	mounting	470294	31	2.88	00:07:31	04:13:45	4	2639	91420	0.72
	pursuit	103538	33	2.02	00:01:10	00:56:22	20	419	20289	0.43
	upright submissive	272304	20	1.13	00:01:43	02:29:34	16	577	53884	0.54
	flee	395852	22	1.08	00:02:23	03:37:33	28	877	78294	0.60
	scramble	323852	18	2.16	00:03:53	02:56:02	3.9	1347	63424	0.79
	anogenital sniffing	103538	33	6.57	00:03:47	00:53:45	8	1268	19440	0.55
rat resident-intruder	attack	136830	12	16.40	00:12:28	01:03:33	0	4437	32119	0.55
	lateral threat	136830	12	9.09	00:06:55	01:09:06	0	9943	26613	0.55
	anogenital sniffing	136830	12	7.50	00:05:42	01:10:19	7.5	2095	22713	0.45
	submission	136830	12	11.27	00:08:34	01:07:27	5	3074	33482	0.59
	boxing	136830	12	4.50	00:03:25	01:12:36	0	1306	23502	0.47
	approach	136830	12	4.23	00:03:13	01:12:48	8.5	1053	23755	0.50
	avoidance	136830	12	2.48	00:01:53	01:14:08	10	686	24122	0.48
CRIM13 resident-intruder	approach	757350	64	4.13	00:20:51	18:01:39	12	6190	145280	0.44
	attack	757350	64	4.10	00:20:42	18:01:48	16	6266	145204	0.40
	chase	757350	64	0.49	00:02:28	18:20:02	46	769	150701	0.58
	circle	757350	64	1.32	00:06:40	18:15:50	22	1982	149488	0.54
	clean	757350	64	7.99	00:40:20	17:42:10	8	12095	139375	0.53
	copulation	757350	64	4.62	00:23:20	17:59:10	12	7080	144390	0.50
	drink	757350	64	0.36	00:01:49	18:20:41	10	570	150900	0.54
	eat	757350	64	2.54	00:12:49	18:09:41	12	3824	147646	0.43
	sniff	757350	64	14.94	01:15:26	17:07:04	0	22757	128713	0.54
	up	757350	64	3.71	00:18:44	18:03:46	6	5575	145895	0.50
	walk away	757350	64	3.89	00:19:38	18:02:52	8	5897	145573	0.50

Figure 4.13. Figure S11. Training set information for mouse, rat, and CRIM13 classifiers.

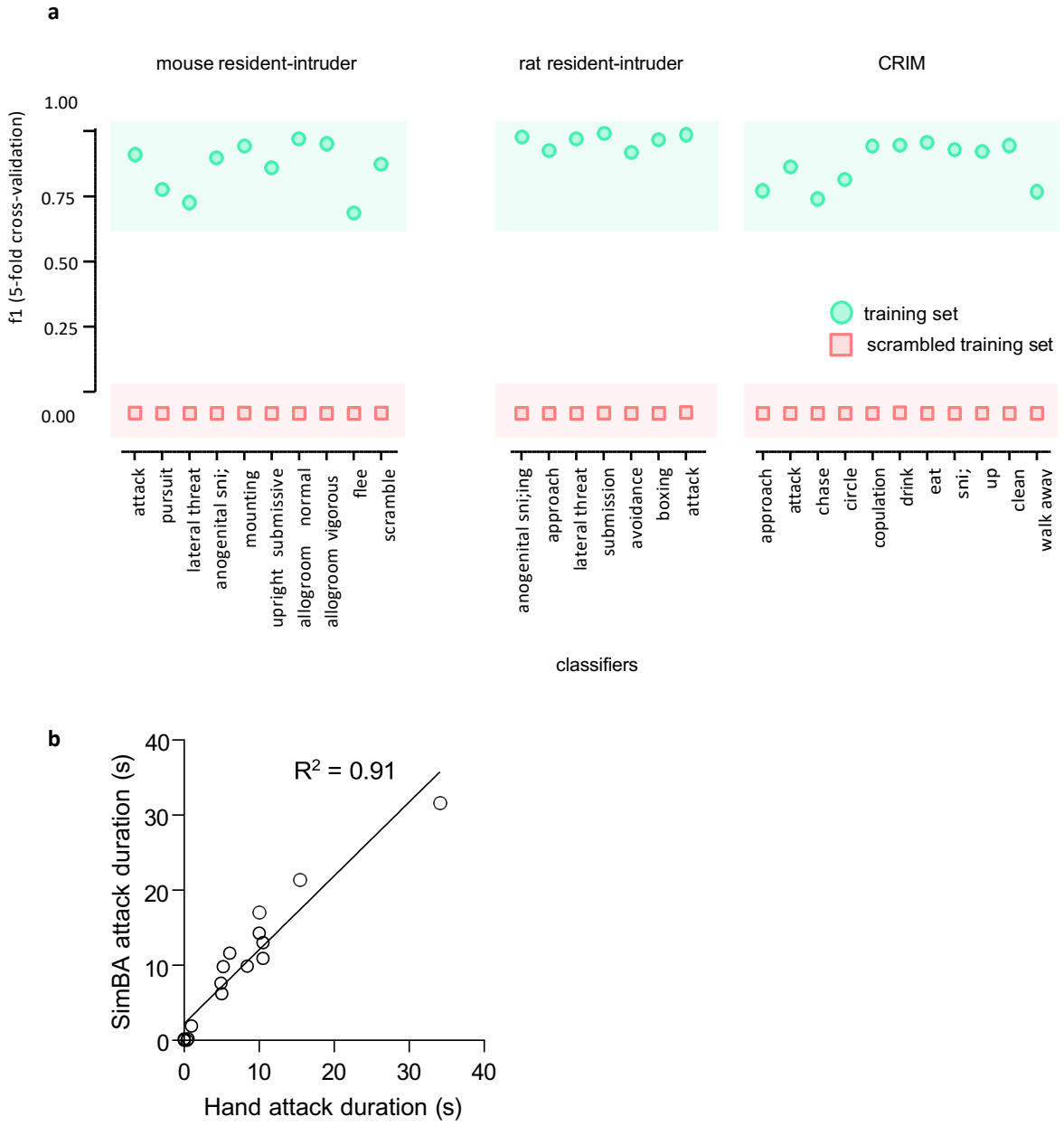


Figure 4.14. Figure S12. Additional classifier performance metrics*. (a) Classifier performance after randomly scrambling the human annotations in the training set. Performance was evaluated as f1 score for the presence of the target behavior, measured by shuffled 5-fold cross-fold validation after randomly scrambling the human annotations in the training set. The classifiers were tested using un-scrambled, correctly annotated, test sets. The green circles represents the performance of the classifiers when trained using un-scrambled annotations. Errors represent ± SEM (Note: errors are not discernible in the graph). (b) Hand annotation versus SimBA detections for attack behavior in 16 videos not included in the machine learning training set.*

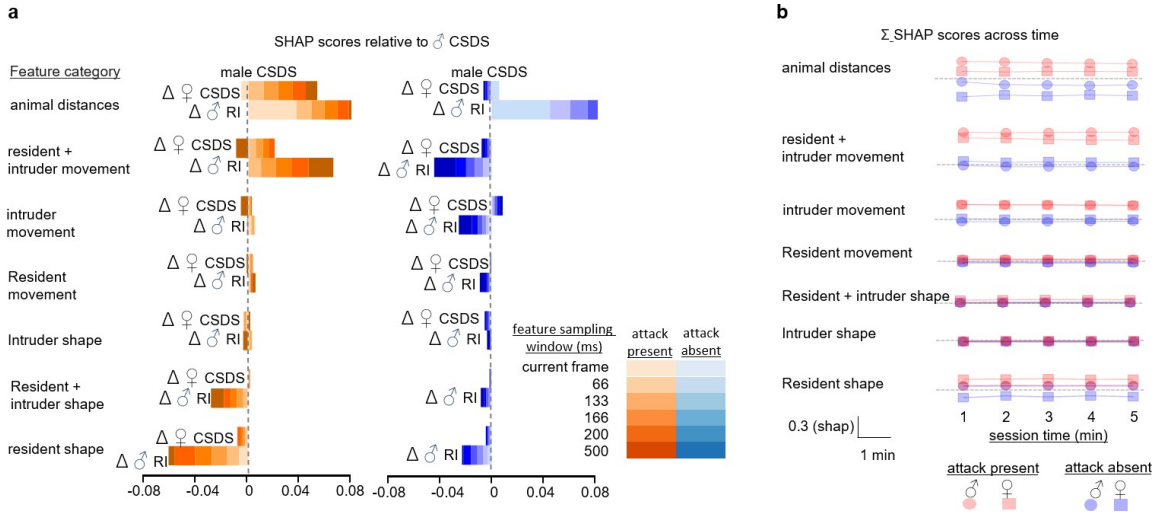


Figure 4.15. Figure S13. Attack SHAP values across groups and throughout testing sessions. *We calculated SHAP values for 1250 attack frames and 1250 non-attack frames within each experimental protocol. (a) We used these values to calculate delta shap values, where we evaluated the female CSDS and male RI SHAP values against male CSDS SHAP value baseline. The SHAP analyses revealed large similarities in how feature values aXected attack classification probabilities in the three experiments (all feature sub-category delta shap < 0.044). The most notable experiment diXerence was the importance of animal distance features within the current frame, which was associated with higher attack classification probabilities in the RI experiment than in the male CSDS experiment. Attack classification probabilities in the RI experiments were also less aXected by features of the resident shape than in the males CSDS experiment. These diXerences may relate to the diXerent attack strategies and experimental setup used in the experimental protocols. (b) Next, we analyzed SHAP vales for classifying attack and non-attack events in the male and female CSDS experiments within 1min bins and showed that SHAP values are not aXected by time of session.*

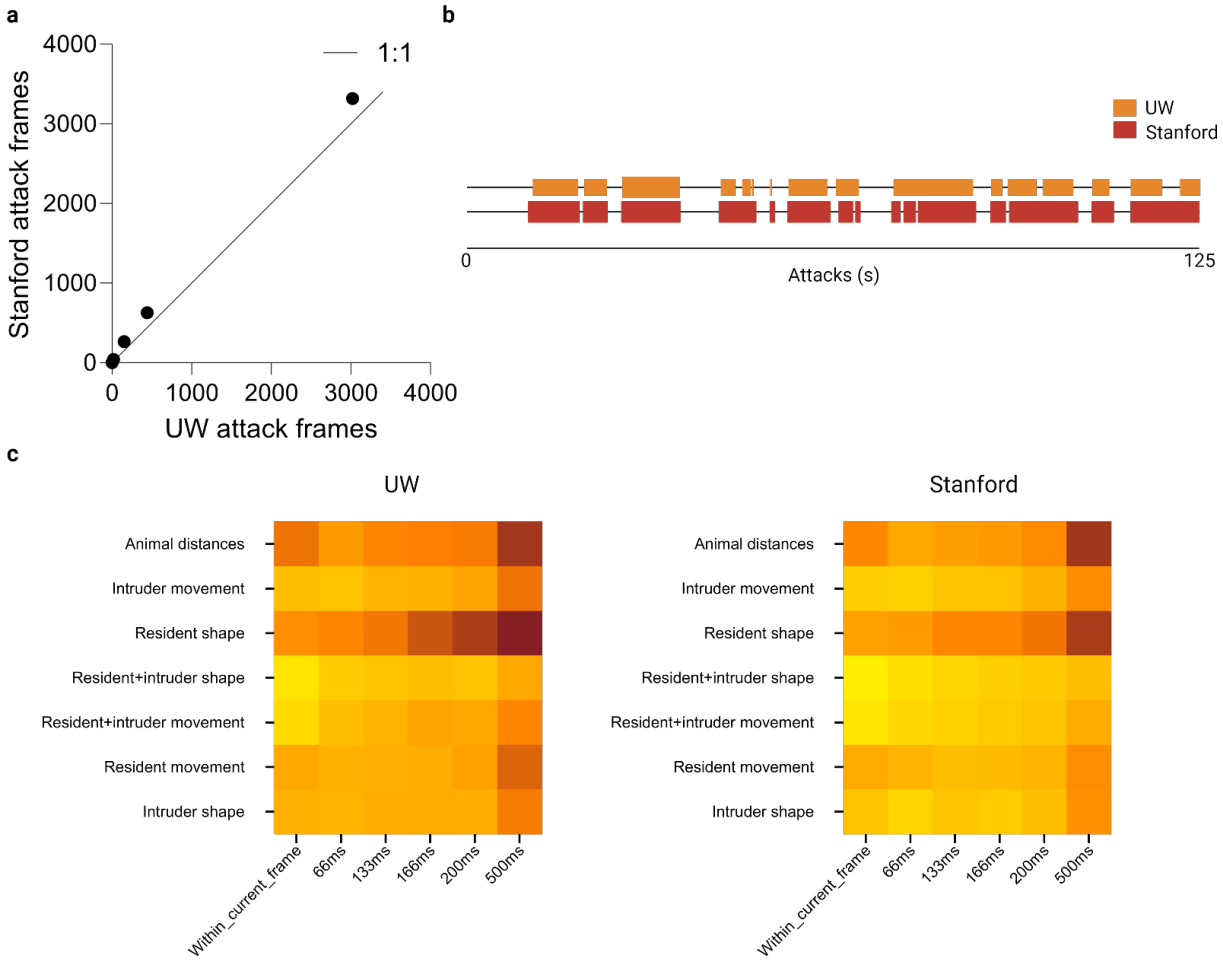


Figure 4.16. Figure S15. UW versus Stanford scoring and SHAP scores. *UW and Stanford manual scoring of the same dataset for attack behavior. (a) Manual annotations (n=9 videos) were highly correlated (R2 = 0.998). (b) Gantt plot of UW versus Stanford scores for a high-attack video. (c) SHAP scores for UW positive or Stanford positive attack frames. UW scores rely more on longer rolling windows of behavior than Stanford does.*

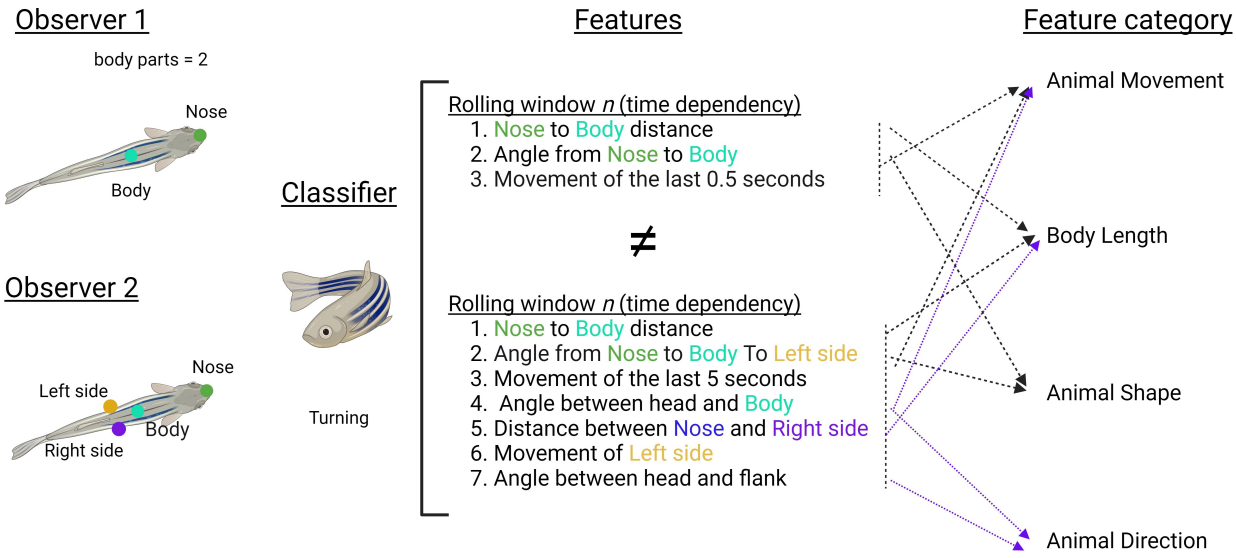


Figure 4.17. Figure S16. Feature binning for SHAP calculations. *Classifiers for the same behavior using diXerent pose estimation schemes will have diXerent feature lists, but can be directly compared via feature binning through the SHAP additivity axiom.*

Chapter 5 : Sex Differences in Appetitive and Reactive Aggression

Antonio V. Aubry[†1], C. Joseph Burnett[†1], Nastacia L. Goodwin[†2,3], Long Li,[1] Jovana Navarrete[2,3], Yizhe Zhang[2], Valerie Tsai[2], Romain Durand-de Cuttoli[1], Sam A. Golden[*2,3,4], Scott J. Russo[*1,]

[1] Icahn School of Medicine, Nash Family Department of Neuroscience and Friedman Brain Institute, New York, New York, USA

[2] University of Washington, Department of Biological Structure, Seattle, Washington, USA

[3] University of Washington, Graduate Program in Neuroscience, Seattle, Washington, USA

[4] University of Washington, Center of Excellence in Neurobiology of Addiction, Pain, and Emotion (NAPE), Seattle, Washington, USA

[†] Denotes equivalent contribution, [*] denotes co-corresponding authors

Introduction

Aggression exists along a spectrum from adaptive to maladaptive and is governed by both reactive (defensive) and appetitive (rewarding) drives. The transition away from an adaptive state can be associated with neuropsychiatric conditions and presents a challenge to patients and caregivers. Modeling and understanding the behavioral etiology of aggressive behavior is therefore a health priority with the potential to guide therapeutic interventions across a number of neuropsychiatric diseases. In mice, aggressive behavior serves as an evolutionary adaptation for survival (Kravitz & Huber, 2003) and engages highly conserved neural mechanisms (Lischinsky & Lin, 2020). However, while aggression is often a focus of both popular and scientific inquiry and highly evolutionarily conserved, very little is known about the neural and behavioral mechanisms controlling aggression-related sex differences.

Preclinical behavioral models developed in males have allowed for direct comparison of reactive and appetitive aggression (Aleyasin et al., 2018; Flanigan & Russo, 2019; Golden et al., 2019). Typically, reactive aggression is investigated using the resident intruder (RI) test in which a male intruder is introduced to the home cage of a male resident and they are allowed to freely interact (Golden et al., 2011). To assess aggression reward, aggression conditioned place preference (CPP) can be used where male mice will display a preference for contexts previously associated with opportunities to attack a naïve conspecific (Flanigan et al., 2020; Golden et al., 2017; Golden et al., 2016). However, like RI testing, this procedure uses forced involuntary social interactions and therefore cannot fully dissociate reactive from appetitive components. To overcome this, several groups have developed social operant tasks that measure voluntary appetitive aggression seeking in male mice (Bannai et al., 2007; Falkner et al., 2016; Fish et al., 2002; Fish et al., 2005), and established that appetitive aggression can transition to compulsive addiction-like behavior in some mice (Golden et al., 2017).

While these procedures have proven effective for studying aggressive behavior and the underlying neurobiology in animal models, studies have focused nearly entirely on males. Male aggression is typically studied in the context of isolated housing to enhance aggression, yet under similar conditions, naïve female mice do not show comparable intruder-directed aggression. To overcome this, alternative model organisms can be used such as Syrian hamsters (Grieb et al., 2021) and California mice (Silva et al., 2010), or with mouse models of gestational aggression (Blanchard et al., 1984; Hashikawa et al., 2017b; Unger et al., 2015). However, there are

limitations to these alternate models: (*i*) only mouse models can presently exploit the broad transgenic toolbox available to understand circuit and molecular mechanisms and (*ii*) gestational aggression models are not ideal for evaluating sex differences in aggressive behavior since these behaviors are linked to hormonal changes specifically associated with pregnancy, parturition and lactation.

Two recent studies have revisited female aggression during RI tests using outbred mouse strains, as opposed to the more typically used inbred strains, and have found that naïve outbred CFW mice display similar levels of aggression as males. Outbred mouse strains such as CD1 and CFW have gained popularity in aggression-related studies due to their high individual variability in innate aggressive behavior (Chia et al., 2005, Golden et al., 2016). Similarly, isolated sex-naive female outbred CFW, but not inbred C57BL6/J, mice will attack juvenile male or adult female C57BL6/J intruders in the home cage (Hashikawa et al., 2017), and female CFW mice pair-housed with a castrated male partner will attack adult female C57BL6/J mice, with similar but non-identical behavioral strategies to males (Newman et al. 2019). The establishment of a female RI procedures opens the door for sex comparisons of the neurobiological substrates of reactive aggression, but currently there are no direct comparisons of appetitive aggression in males versus females.

To further develop preclinical models of aggression, we directly compared adult male and female CFW mice in reactive and appetitive aggression procedures to evaluate sex as a biological variable. The evaluation of sex differences in aggression are compounded by the need for detailed annotation of behavioral actions and sequences. Typically, only gross measures of aggression, are reported. This lack of in-depth behavioral analysis obscures potential differences in specific behaviors and their sequences. Building upon a recent study in male mice (Kwiatkowski et al., 2021), we utilized a discrete state hidden Markov model (HMM) to define sex differences in aggression. HMMs analyze the ordering, clustering, and transitions between actions and are able to cluster time-series of behavior into distinct hidden states. These models have been used extensively in analyzing sequences of speech, gestures, animal behavior and the analysis of gene and protein sequences (Carola et al., 2011; Lee et al., 2019; Rabiner, 1989; Stanke & Waack, 2003). The discrete state HMM allowed us to examine the sequential composition of social behaviors and the hidden states which contribute to male versus female

aggression. We also tested whether there are differences in reactive or appetitive aggression between male and female mice.

We therefore set out in this study to accomplish two goals: (*i*) to elaborate upon a mouse model of aggressive behavior encompassing a broad suite of ethologically relevant and reward-related behavioral metrics, (*ii*) and to compare reactive and appetitive aggression between male and female mice. Our findings demonstrate clear sex differences in the behavioral sequences that make up bouts of aggressive and investigative behavior. Further, using both aggression CPP and operant social self-administration (SA) procedures, we show that female CFW mice in these contexts do not exhibit context-dependent aggression reward nor do they exhibit appetitive aggression seeking behavior. To our knowledge, this is the first report of sex differences in appetitive vs. reactive aggression in *mus musculus*. Together these data support distinct patterns of aggressive behavior between males and female outbred mice and underscore the importance of future research to identify detailed mechanisms by which different sexes express aggressive behavior.

Methods

Mice: 10-week-old CFW mice (Charles River Laboratories) were used for all studies. Females were pair-housed with a castrated male for the study duration (Newman et al., 2019) and males were pair-housed for 48-hr with stimulus females prior to isolate housing. Subject males were separated from group-housed cage mates and paired with a female for 2-d, then singly housed for an additional 10-d before all protocols. This procedure was used in order to acquire roughly equal amounts of aggressors (AGG) and non-aggressors (NON) without affecting the amount of aggression observed in AGGs. Females were housed with surgically castrated male CFW mice (see supplemental methods) for at least 14-d before all protocols. 12-week-old C57BL/6J mice were used as intruders for all social interactions. All studies were conducted during the light cycle. Procedures were performed in accordance with the National Institutes of Health Guide for Care and Use of Laboratory Mice and approved by the Icahn School of Medicine at Mount Sinai Institutional Animal Care and Use Committee. For studies conducted at the University of Washington, additional groups of two-week isolated males and females were also used as subject mice to assess the effects of housing conditions. All studies were conducted during the dark cycle. Procedures were performed in accordance with the National Institutes of Health Guide for

Care Use of Laboratory Mice and approved by the University of Washington Institutional Animal Care and Use Committee.

Aggressor Screening and Resident-Intruder (RI) Test: Mice were screened using protocols adapted from previous studies (Golden et al., 2011; Golden et al., 2016). Briefly, cage tops were removed and replaced with Plexiglas covers to monitor trials. Before initiating trials with paired female mice, the cohabiting male mouse was removed to a holding cage until completion of test. A novel C57BL/6J mouse matching the sex of the resident was introduced into each cage and mice were allowed to freely interact for 5-min. After 5-min elapsed, intruder mice were returned to their home cages and, in the case of female resident-intruder trials, cohabiting male mice were returned to their home cages. All videos were recorded for later analysis. Resident behaviors from the Mount Sinai videos were manually annotated using Observer XT 11.5 (Noldus Interactive Technologies).

Aggression Conditioned Place Preference (CPP): CPP testing was conducted in three phases as previously reported (Flanigan et al., 2020; Golden et al., 2016): pre-test, acquisition, and post-test. Mice were habituated to test rooms 1 hour before acquisition or test trials. All phases were conducted under red light and in sound-attenuated conditions. The CPP apparatus (Med Associates) has a neutral middle zone that allowed for unbiased entry and two conditioning chambers with different walls and floors.

Appetitive aggression SA: Following RI training, each group was further separated into mice who were either AGG or NON during RI testing. Only one male was NON, and as such only AGG males were tested in operant SA. While all CFW mice in this group underwent RI testing, previous work demonstrates that male CD1 mice that have not undergone previous RI testing acquire aggression self-administration at rates ranging from 57-81% (Golden et. al., 2017). These results suggest that prior resident intruder experience is not necessary to elicit appetitive aggression self-administration.

3-d following RI testing, mice underwent 1-d of magazine training, in which they were exposed to operant cues (house light and a two second tone) in addition to a same sex intruder mouse entering the operant chamber (3 times each). On the following day, mice underwent SA training every other day for 9-d as previously described (Golden et al., 2017). Researchers were present throughout all aggression testing to ensure that no mice were injured. Mice with an average of 3 presses or less across days 4-8 of training were considered non-acquirers.

See supplemental material for a full description of the methods.

Gross characterization of social behavior in male and female mice.

On the third day of RI, both male and female AGGs engaged in more aggressive behavior than NONs (Phenotype $F_{(1, 51)} = 36.20$, $p < 0.0001$, Male AGG vs. Male NON, p =0.0002, Female AGG vs Female NON, p = 0.001), and there were no differences in duration or latency to attack between male and female AGGs ($p > 0.05$) (Figures 1B & 1C). There were no differences between groups in investigative behavior (Sex: $F_{(1, 51)} = 2.947$, p = 0.0920. Phenotype: $F_{(1, 51)} =$ 0.6242 Figure 1D). Male NONs displayed a shorter latency to investigate intruders (Phenotype x Sex interaction $F_{(1, 51)} = 5.003$, p = 0.0297, Male NON vs Male AGG, p = 0.0139, Figure 1E).

Results

Male and female mice display distinct suites of social behavior.

We next investigated whether male and female AGGs/NONs displayed distinct aggressive and/or investigative behaviors. We did not find any differences in anogenital (Sex: $F_{(1, 51)} = 0.4110$, p = 0.5243 Phenotype: $F_{(1, 51)} = 1.141$, p = 0.2905, Interaction: $_{(1, 51)} = 0.2701$, p = 0.605) or flank investigation (Sex: $F_{(1, 51)} = 0.4182$ p = 0.4551, Phenotype: $F_{(1, 51)} = 0.0577$, p = 0.8904, Interaction: $F_{(1, 51)} = 0.3814$, p = 0.490) between any of the four groups (Figure 2A & 2C). Interestingly, we found that females engaged in significantly more facial investigation than males, regardless of their phenotype ($F_{(1, 51)} = 9.54$, p = 0.0032, Interaction: $F_{(1, 51)} = 0.03007$, p = 0.8630, Figure 2D). We also observed that AGGs, regardless of their sex engaged in more allogrooming than NONs (Phenotype: $F_{(1, 51)} = 4.574$, p = 0.0373), although only female AGGs were significantly different from female NONs (p = 0.0412, Figure 2B) We also found a main effect of both sex ($F_{(1, 51)} = 31.46$, $p < 0.0001$) and phenotype ($F_{(1, 51)} = 13.40$, p = 0.006) with no interaction ($F_{(1, 51)} = 3.690$, p = 0.06) on the number of withdrawals observed (Figure 3E. Female AGGs displayed a higher number of withdrawals than male AGGs ($p < 0.0001$) and Female NONs(p = 0.0022).

When examining aggression, we observed that male and female mice engage in qualitatively distinct behaviors on day 3. Male AGGs engaged in wrestling behavior, in which the resident male lunges at the intruder and tumbles around the home cage, while female AGGs did not engage in this behavior at all ($t(15) = 3.571$, p = 0.0034, Figure 2F). Although some females did engage in lunging behavior, it was to a lesser extent than male AGGs ($t(15) = 2.070$, p = 0.054,

Figure 2G). Females delivered more bites than males ($t(15) = 2.104$, $p = 0.046$, Figure 2I) with a and there was a trend for females to exhibit more kicks than males ($t(15) = 2.005$, $p = 0.0632$, Figure 2J). These single kicks were usually delivered following a single bite. In contrast, males were more likely to pin the intruder ($t(15) = 2.151$, $p = 0.0442$, Figure 2H) prior to delivering a bite.

Given that male and female mice display distinct sets of investigative and aggressive behavior, we used a random forest classifier to determine whether trials involving a male or a female as the resident were distinguishable based on the metrics quantified in Figure 2. Trials from day 3 were included in the model. We tested models in which 20, 40, 60, or 80 percent of the data was used for training the model (see Methods for details). We found that when 80% of the data was used to train the model, an F score of 1 was achieved, indicating a perfect classification of the remaining 20% of the trials (Figure 2K &L). We extracted the gini impurity metric to determine which variables were important for classifying males vs. females. The analysis indicated that withdrawals, facial investigation, and wrestling were important in classifying male vs. females.

Male and female mice display distinct sequences of social behavior.

For the HMM, we found that a 4-state model best fit the sequences of observations (see Methods for details). Inspection of the emission probabilities (Supplementary Table 1) suggests that states 1 and 2 (Persistent Attack & Intermittent Attack listed below as A1-A2 or I1-I2) were predominantly associated with aggressive actions, with bite being the most likely behavior to occur when the animal was in these states. Interestingly, state A2 was also characterized by a relatively high probability of investigation occurring, while state A1 was associated with relatively low probabilities of investigation (39% for state 2, 19% for state 1). Conversely, States 3 and 4 (Full Body Investigation & Anogenital Investigation) were predominantly associated with investigative behaviors, with aggressive behaviors being highly unlikely to occur (6% and .06% respectively). These investigative states were differentiated by the probability of specific investigative behaviors occurring. While in state I1, there was a roughly equal probability of anogenital (32%), facial (23%), and flank investigation (22%) (Supplementary Table 1). However, while in the anogenital investigation state, the mice were much more likely to engage in anogenital investigation (34%) rather than facial (14%) and or flank investigation (15%) (Supplementary Table 1). To determine whether certain groups were more likely to be in a particular state, we calculated the percentage of behavioral observations that occurred in each

state for each mouse. We found that male AGGs had a significantly higher percentage of their observations in the persistent attack state than female AGGs (Sex x Phenotype interaction $F_{(1, 51)} = 4.556$, $p = 0.0376$, Male AGG vs. Female AGG, $p = 0.0111$, Figure 3B).
Conversely, female AGGs had a significantly higher percentage of their observations in the intermittent attack state compared to male AGGs (Sex x Phenotype interaction $F_{(1, 51)} = 4.451$, $p = 0.0398$, Male AGG vs. Female AGG, $p = 0.0206$,Figure 3C). The difference between male and female AGGs is likely due to the fact that females are more likely to investigate the intruder before or after delivering a bite (36%) compared to males (14%) (Supplementary Table 3A & B)
With regard to the full body investigation state, there was a striking sex difference, with none of the males showing any observations in this state ($F_{(1, 51)} = 30.77$ $p < 0.0001$, Male AGG vs. Female AGG, $p = 0.0010$. Female NON vs Male NON $p = 0.0023$, Figure 3E). This phenomenon is due to the fact that females were more likely to string together multiple investigatory actions than males (Figure 3F, Supplementary Table 3A & B). Lastly, NON mice were more likely than AGGs to be in state I2, regardless of sex (phenotype $F_{(1, 51)} = 25.85$, $p < 0.0001$, Figure 3E).

Males, but not females display appetitive aggression

In the CPP assay (Figure 4A), there was a significant effect of time ($F_{(1,38)} = 8.269$, $p = 0.006$), sex ($F(1, 38) = 9.952$, $p = 0.003$), and a time x sex interaction ($F(1, 38) = 3.899$, $p = 0.055$). Post hoc analysis revealed that only male AGGs spent more time in the paired chambered in the post-test relative to the pre-test ($p < 0.05$, Figure 4B).

RI screening of mice used for appetitive aggression test:

All 29 males were aggressive during at least one RI trial, while 18 of 29 females were aggressive. The three resulting groups (male AGG, female AGG, female NON) differed significantly in latency to attack, with NON animals showing significantly longer latency to attack when compared to AGG male or female mice ($p < 0.0001$, $F(2, 56) = 14.47$) (Figure 4D).

Males and females learn to self-administer intruders similarly, but vary in attack behavior:

Between male and females AGGs, there was a significant sex x day interaction in reward and attack behavior (interaction $F_{24,306}=3.327$, $p < 0.001$, day $F_{8,306} = 6.787$, $p < 0.001$, sex $F_{3,306} = 108.2$, $p < 0.001$) with females showing significantly fewer attacks than males ($p < 0.001$, df =

306 Tukey's). Latency to press for an intruder significantly decreased over days in both males and females ($p = 0.0151$, $F (8, 148) = 2.475$), and there was no difference in exploratory head entry activity across days or sex. ($p = 0.9963$, $F (8, 153) = 0.1520$) (Figure 4E).

Female NONs showed significantly more rewards over time, but near zero attacks across training days (Interaction $F_{8,108}=9.277$, $p < 0.001$, Day $F_{8,108}=13.02$, $p < 0.001$, Attack v Reward $F_{1,108}=490.9$, $p < 0.001$). There were no significant differences across days in latency to press ($F_{(1.942,11.65)} = 2.658$, $p = 0.11$) or exploratory head entries ($F_{(3.977,23.86)} = 2.46$, $p = 0.07$). (Figure 4F).

When compared directly, female NONs and AGGS both showed increasing rewards over time, with stable but very low attack frequencies (Interaction $F_{24,270}=6.225$, $p < 0.001$). NON females showed slightly higher rewards than AGG females ($p = 0.0248$, means 7.857 and 6.833 respectively), with no differences in attack behavior ($p = 0.9309$). All mice that did not acquire self-administration were excluded from analysis.

RI aggression phenotype does not predict operant self-administration acquisition:
There was no significant difference in the proportion of mice per group that acquired operant SA, as evidenced by an average of > 3 presses per day for the last five days of training (Male AGG = 9/16, Female AGG = 10/15, Female NON = 7/11, Chi-square, df = 0.3752, 2, $p = 0.829$ Figure 4G).

Housing condition does not appear to impact aggression trends
Male and female AGGs that were housed in isolation for two weeks prior to the start of testing show similar trends to pair housed AGGs. Isolated males showed a trend toward increasing rewards over time ($p = 0.057$), as well as stable attack frequency ($p = 0.701$), latency to press ($p = 0.514$), and exploratory head entries ($p = 0.197$) over time (Figure 5B). Isolate females showed increasing rewards ($p < 0.001$), stable but low attack frequency ($p = 0.0763$), decreased latency to press ($p = 0.009$), and steady exploratory head entries ($p = 0.251$) over time (Figure 5C).

Food training data:

A subset of male and females that did not acquire aggression self-administration were tested for learning capability via food self-administration testing. There were no sex or housing differences in food self-administration performance ($p = 0.6169$, $F(6, 49) = 0.7440$), and we therefore collapse housing conditions across sexes for analysis. There were no differences between sexes in the amount of food (g/kg) self-administered ($p= 0.8402$, $F(6, 96) = 0.4544$), though rewards per day similarly increased over time in both sexes (Figure 5D).

Discussion

We sought to characterize differences in aggressive and investigative social behavior in outbred male and female CFW mice. To this end, we adopted the protocol of Newman et al. (2019) to quantify aggressive social behavior in females. Until now, most studies female aggression studies in laboratory mice have resorted to using lactating females during the postpartum period (Blanchard et al., 1984; Hashikawa et al., 2017a; Parmigiani et al., 1988; Unger et al., 2015). This is not ideal for evaluating sex differences in aggressive behavior since these behaviors are linked to hormonal changes specifically associated with pregnancy, parturition and lactation. Utilizing this protocol, we found that when grossly measured as "aggressive" or "investigative" males and females are largely similar. Although females tended to engage in more investigation than males, this effect was only significant on day 1 and waned with successive bouts of the RI test.

When rodents approach and contact a conspecific they engage in sniffing behavior of distinct body parts such as the face, anogenital, and flank regions (Arakawa et al., 2011). We observed that females, regardless of their RI phenotype, engaged in facial investigation for longer durations than males. The facial area contains different excretory glands that give off distinct signals to the investigating animal. The Harderian glands are located near the eyes and excrete a lipid containing porphyrins (Chen et al., 1997) and have been shown to provide information about the sex and reproductive status of the individual, which can influence social behavior in males (Cavaliere et al., 2020; Hattori et al., 2016). Whether certain facial cues more significantly impact female-female social interaction is unknown and requires further study.

Males and females also displayed distinct attacking behaviors. When male residents attacked the intruder, they displayed full-body lunges and wrestling behaviors that involved the two mice

tumbling around the cage at very high speeds. This is in contrast to females, who were more likely to deliver a series of bites followed by a single kick with their hindlimbs. Male aggression thus seems much more explosive and offensive whereas female aggression seems tamer and possibly defensive in nature. This is in line with a previous study (Blanchard 1984) which found that male bouts were more contact oriented with the male intruder having a higher chance of getting wounded from the bout relative to female intruders. In contrast, females were more likely to attack with a single bite or "jump-attack" followed by the resident withdrawing from the encounter.

As mentioned above, we employed a discrete state HMM. Although Markov chains have been used to examine aggressive behavior in males in the past (Haccou et al., 1988; Natarajan et al., 2009) this is the first instance of a hidden state model being used to compare aggressive behavior in male and female mice. We found that a 4-state model best fit our behavioral observations. Of these 4 states, states 1 and 2 were dominated by aggressive behaviors while states 3 and 4 were dominated by investigative behaviors. Interestingly, the "aggressive states" could be further differentiated by the probabilities of particular behaviors occurring. Although both states 1 and 2 were associated with a high level of aggression occurring, only state 2 was associated with a high level of investigation also occurring. This suggests that state 1 is characterized by persistent attacking for prolonged periods of time, while state 2 is characterized by a mix of both investigative and aggressive actions. Given the above discussion regarding the qualitative differences in attack behavior in males and females, it is not surprising that male AGGs had a greater proportion of their behaviors in state 1 whereas female AGGs had a greater proportion of their behaviors in state 2.

As with states 1 and 2, states 3 and 4 can also be further differentiated based on which particular behaviors were more likely to occur. State 3 was characterized by a roughly equal probability of any of the three main investigatory behaviors occurring, while state 4 was also characterized by a relatively high probability of AG investigation occurring relative to other modes of investigation. Strikingly, none of the behavioral sequences demonstrated by males were characterized as being in state 3. This is likely due to the fact that females were more likely to string together multiple investigative behaviors in succession, while males predominantly engaged in AG investigation or ended the interaction and then re-engaged in AG investigation during a separate bout. In contrast

males tend to engage in interaction bouts that consist solely of one of the two types of social behavior (aggressive or investigative), terminate the bout, and then re-engage in a separate bout. Although male and female AGGs displayed robust levels of reactive aggression, they differed with regard to aggression reward and the acquisition of appetitive aggression. The CPP experiment revealed that only male AGGs developed a preference for the side paired with aggressive experience, suggesting they find it to be rewarding or reinforcing. In line with these findings, while both males and females acquired SA behavior, only males attacked during the subsequent social interaction bout with the intruders. We can speculate that the robust female social self-administration may be affiliative, rather than aggressive, when social interactions are volitional rather than forced. These data agree with recently published work using outbred CD1 female mice, where female mice readily lever press for sensory contact to female partner mice (Ramsey et al., 2021). However, our data also caution against the use of purely barrier-based social self-administration procedures in males and females due to the potential incongruence in aggressive behavior between RI and SA testing. Use of barrier and purely sensory contact may mask, whether aggressive or affiliative, the ultimate motivation of the resident mouse.

These results indicate that female CFW mice are a valid model for studying reactive aggression, which is a departure from the historical narrative that female mice are only maternally aggressive and can therefore be excluded under the NIH sex as a biological variable initiative. In striking opposition to our reactive aggression observations, we find that female CFW mice that exhibited strong reactive aggression do not exhibit appetitive aggression seeking behavior under these housing and testing conditions. This mimics the sex difference observed using aggression CPP, and suggests a significant behavioral sex difference between male and female CFW mice regarding the reinforcing effects of aggression and aggression seeking behavior.

This work highlights the limitations of developing preclinical models entirely in males, and highlights the need for a more parametric exploration of female aggression. While our study demonstrates that female CFW mice do not demonstrate aggression reward under currently established male models, there may be additional manipulations which could induce aggression reward in female mice. Additionally, species in which females more typically show non-maternal aggression, including California mice, Syrian hamsters, and prairie voles may all provide opportunities to directly compare aggression reward across sexes.

Of note, we saw differences in the percentage of males that were NON versus AGG in RI testing between Mount Sinai and the University of Washington. Outbred lines, while helpful in studying individual differences in aggression, can exhibit batch differences due to the nature of their genetic variability, as has been seen in CD1 outbred mouse aggression testing (Kwiatkowski et al., 2021). As such, variation between sites in percentage of aggressive CFW mice during RI testing is not unexpected.

In summary, we show that despite similar levels of aggression and investigation, the actions displayed by male and female residents—which make up the gross measures of social behavior—are both qualitatively and quantitatively distinct. Our HMM revealed that females are more likely to switch between aggressive and investigative behaviors within a given interaction bout, while males typically engage in only one of these behaviors per bout. Furthermore, while female outbred CFW mice exhibit reactive aggression, only male outbred CFW mice displayed robust levels of appetitive aggression in CPP and SA experiments. Thus, future studies to disentangle the underlying biology driving these sex differences are critical.

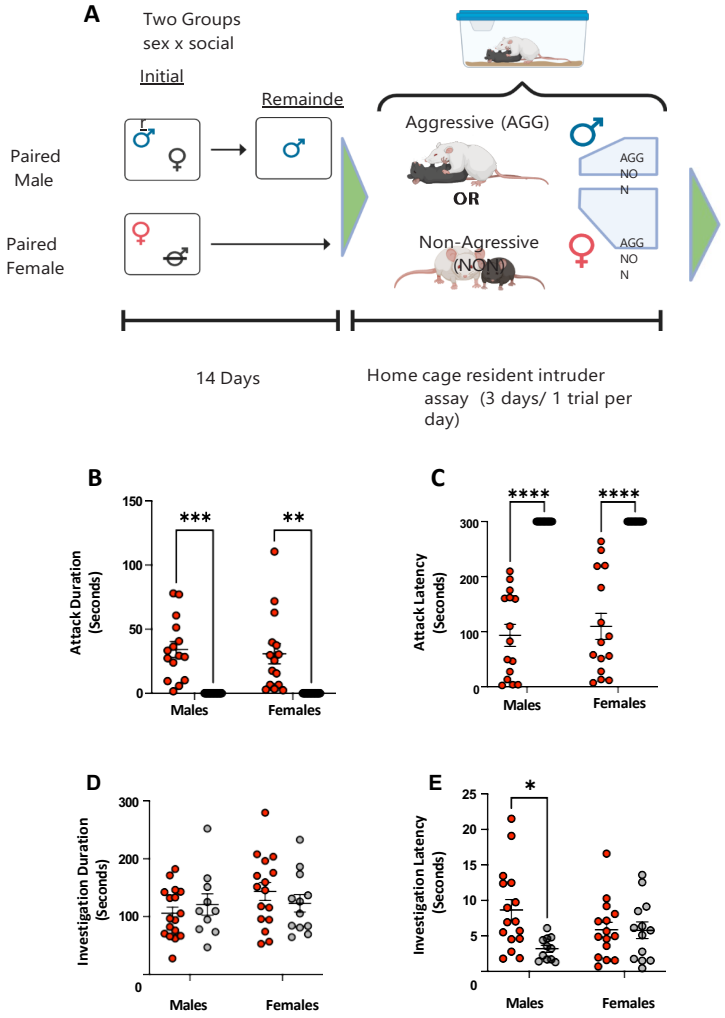


Figure 5.1 Male and Female SW mice engage in similar amounts of aggressive and investigative social behavior. *(A). Schematic illustrating the housing conditions prior to the resident intruder test. (B). Total attack duration (B) and latency (C) did not significantly differ in male and female AGGs. (D). Total investigation. All groups show similar levels of social investigation E. Investigation latency. Male NONs had a significantly shorter latency to investigate the intruder than male AGGs.*

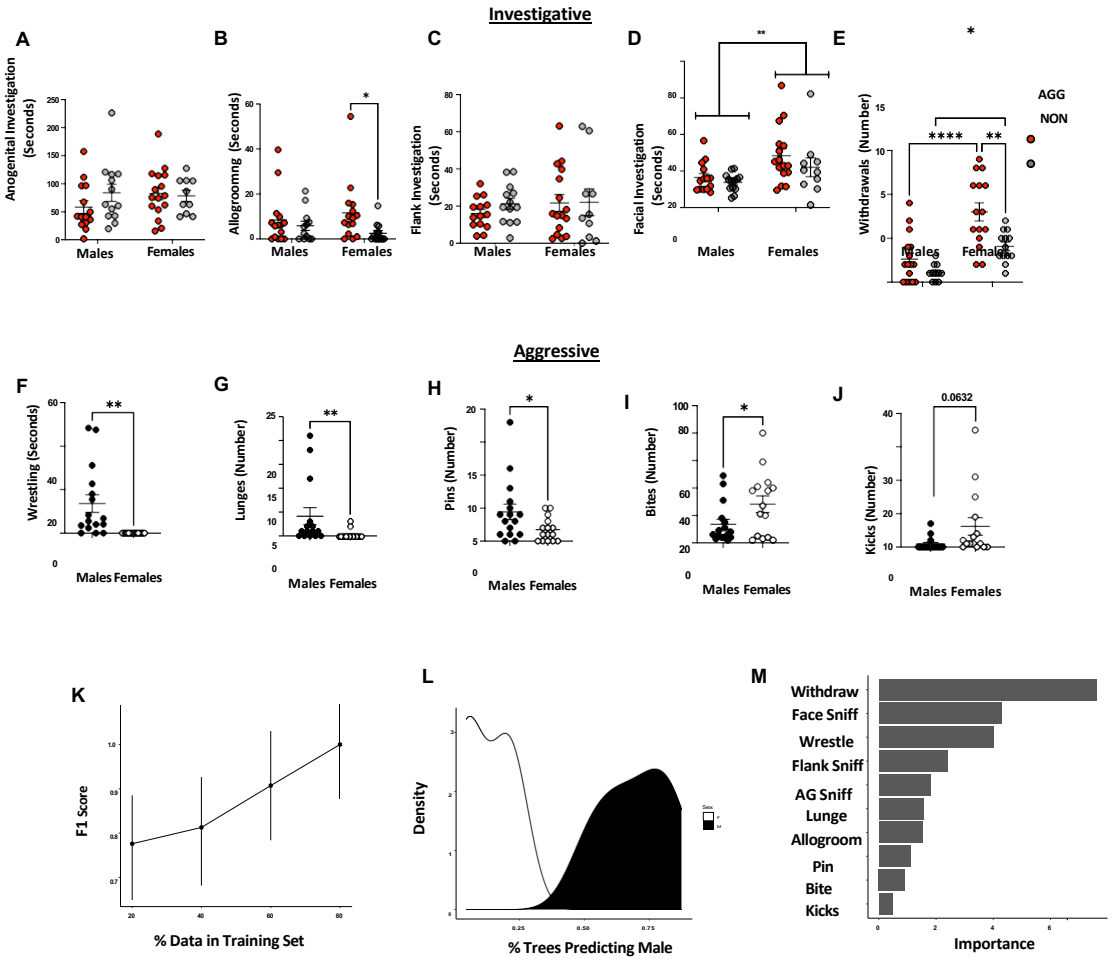


Figure 5.2 Male and female mice display distinct investigative and aggressive behaviors. *For investigative behaviors, there were no group differences in anogenital investigation (A) or flank Investigation (C) AGGs regardless of sex spent more time allogrooming (B). Females regardless of phenotype engaged in more facial investigation (C) and withdrew from interactions more frequently (E). For aggressive behaviors, males engaged in more wrestling (F), lunges (G), and pinned (H) the intruder more than females. Females delivered more bites (I) and kicks (J) K. Learning curves from Random Forrest classifier. Curves were created using 1K trees, 4 data splits (20-80%), and with shuffled 10-fold cross-validation at each data split. Errors represent ± SEM. (L). Density plot demonstrating probability of being classified as M or F as a function of the number of trees predicting male. (M). Variable importance plot for the random forest classifier.*

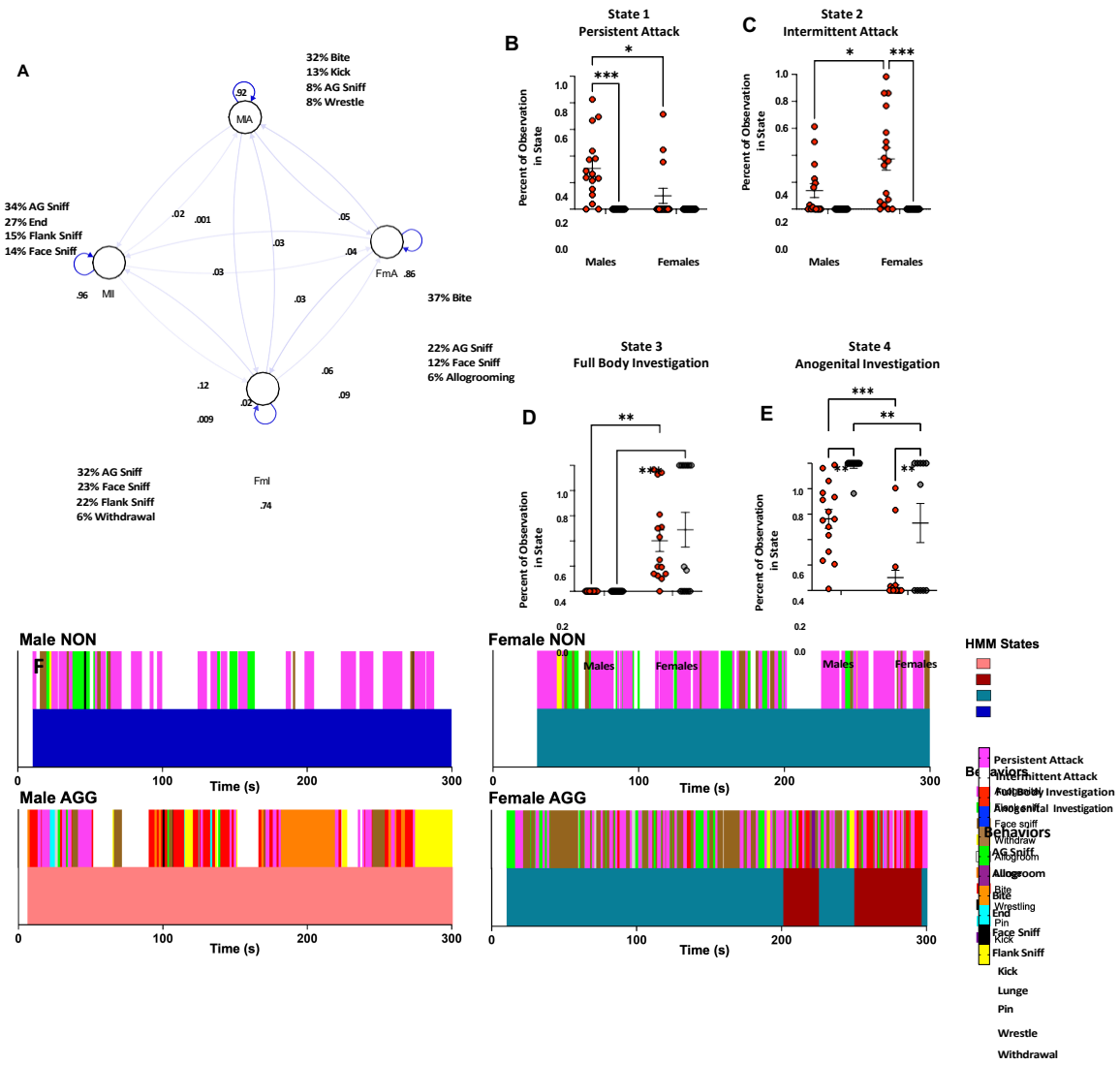


Figure 5.3 Hidden Markov Model of Social Behavior in the Resident Intruder Paradigm. *(A). Schematic of HMM. Each node represents a hidden state. Numbers along the arrows indicate the probabilities of transitioning between states. Listed behaviors indicate the probability of occurrence during each state. Male AGGs were more likely to be in a state of persistent aggression (B) while female AGGs were more likely to be in a state of intermittent aggression (C). Females regardless of phenotype were more likely to be in the full-body investigation state than males(D). NON's regardless of sex and males regardless of phenotype were more likely to be in the anogenital investigation state (E). (F) Representative examples of behavioral sequences (top) and predicted state (bottom) for all four groups.*

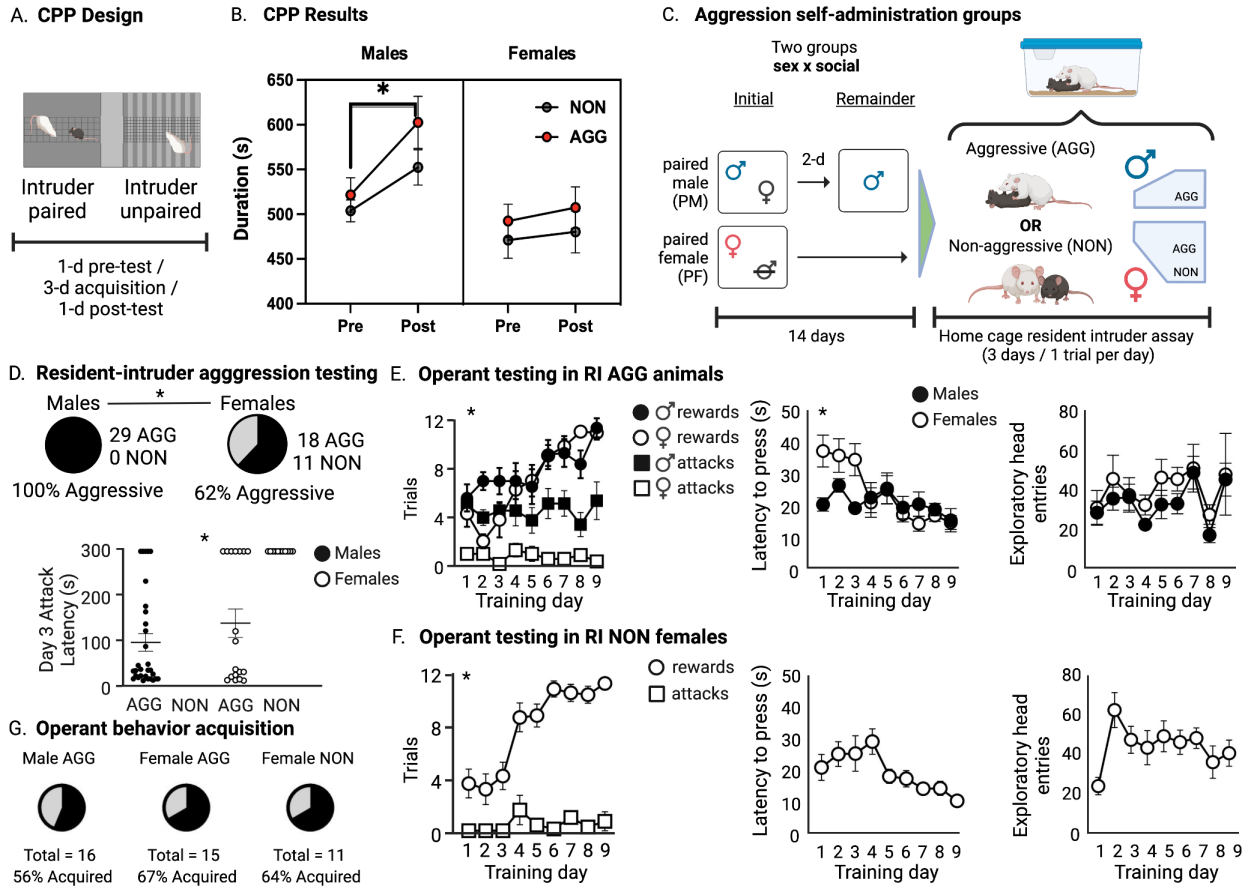


Figure 5.4. Males and females are similar in reactive but not appetitive aggression. *A) Schematic of CPP paradigm. B) Male AGGs but not NONs develop a CPP to the paired chamber. Neither female AGGs or NONs developed a CPP to the paired chamber. C) Schematic of social housing paradigm for self-administration animals. All males tested were aggressive during at least one trial of resident intruder screening, while the females separated into aggressive (AGG) and non-aggressive (NON) phenotypes. D) Latency to attack in the resident intruder assay differed significantly between groups, with female NONs having significantly higher latency to attack than the male or female AGGs. E) Females show slightly slower learning curves than males in acquiring the aggression self-administration task. Additionally, females show almost no attacks once they have self-administered a same-sex conspecific, while male aggression was steady across days. Females are initially slower than males to lever press, but both groups decrease latency over time. There were no differences in exploratory head entries across days or sex. F) Females who were not aggressive in the resident intruder screening show increasing rewards over time with steady attacks and decreasing latency to lever press. They show an increase in exploratory head entries initially which is steady thereafter. G) Similar percentages acquired operant self-administration across groups.*

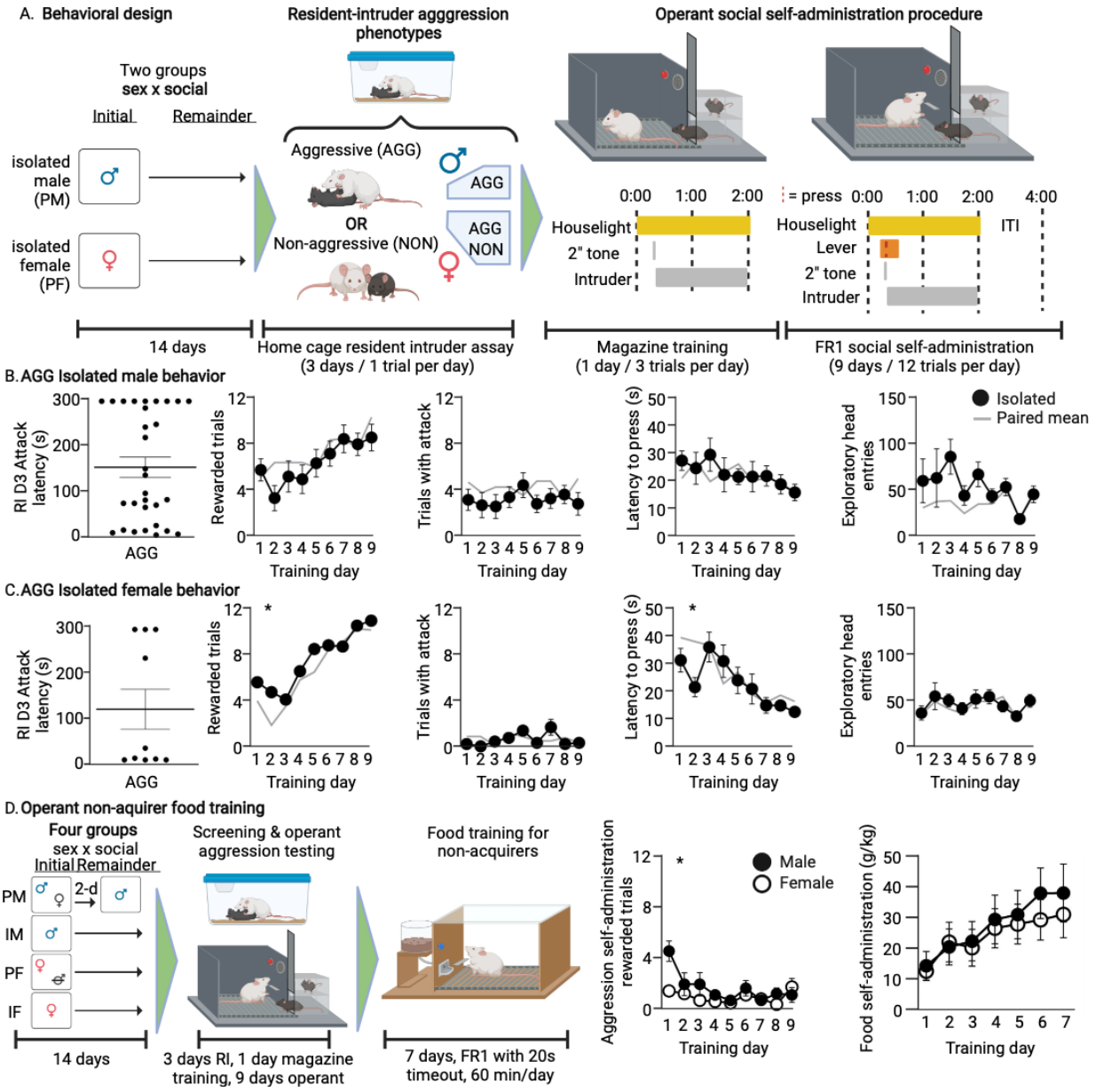


Figure 5.5. Isolate housing does not shift aggression patterns, and aggression self-administration non-acquirers rapidly learned food self-administration. *A) Schematic of isolate housing and behavioral paradigm. B-C) Isolated males and females (black circles) show similar trends as socially housed mice (full data in Figure 4, means showed here in gray). D) Abbreviated behavioral schematic showing housing conditions, resident intruder and aggression self-administration tasks, followed by seven days of sucrose pellet self-administration training for aggression non-acquirers. Males and females showed low rewards in aggression self-administration, with males initially slightly higher than females. Both males and females rapidly acquired sucrose pellet self-administration.*

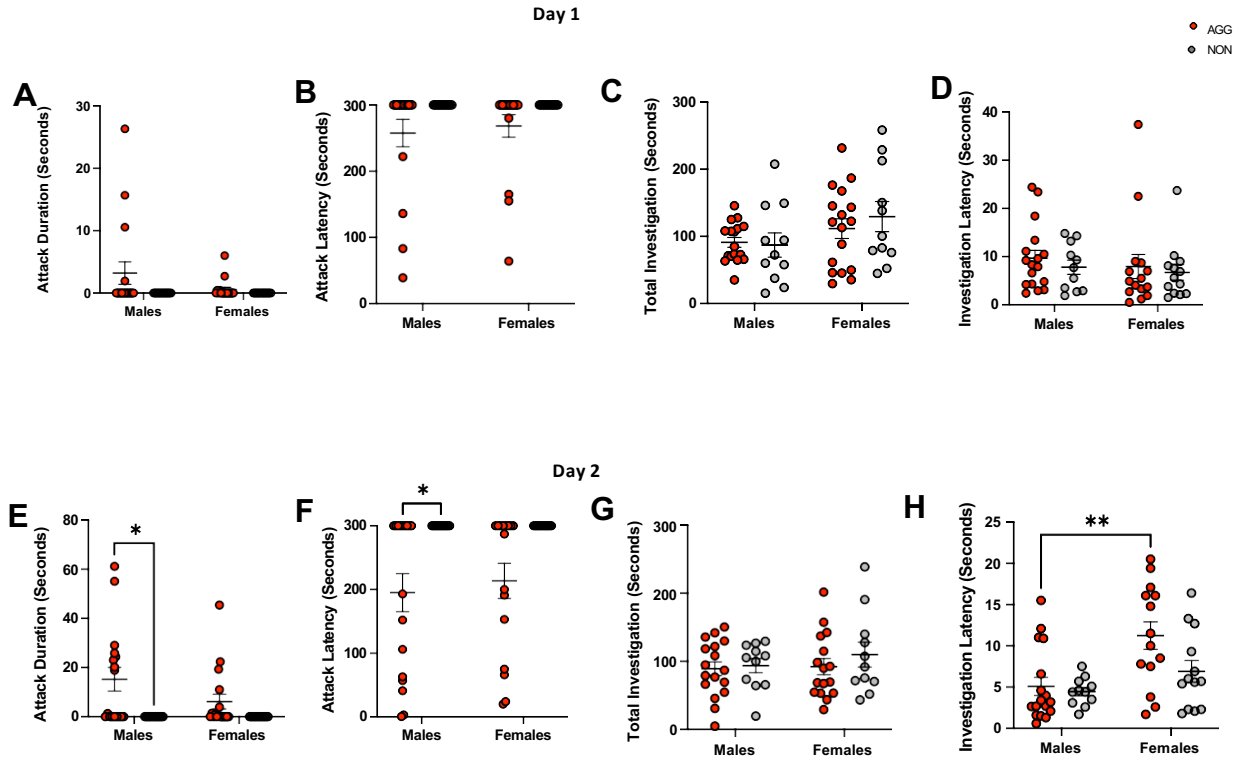


Figure 5.6. Figure S1. Gross measures of social behavior on Days 1 and 2. *A. Attack duration. Two-way ANOVA interaction F (1, 52) = 1.320, p = 0.258. B. Attack Latency. Two–way ANOVA, main eXect of phenotype F (1, 52) = 4.832, p = 0.0352. C. Total investigation. Two-Way ANOVA, main eXect of sex, F (1, 52) = 4.178, p = 0.046. D. Investigation latency. Two-way ANOVA interaction F (1, 52) = 0.02575, p = 0.8731. E. Attack duration (Day 2). Two-way ANOVA, main eXect of phenotype, F (1, 52) = 9.167, p = 0.0038. Tukey's post-hoc Male AGG vs. Male NON, p = 0.0183. F. Attack latency (Day 2). Two-way ANOVA, main eXect of phenotype F (1, 51) = 14.42, p = 0.004. G. Total investigation (Day 2). Two-way ANOVA, F (1, 51) = 0.2639, p = 0.6097. H. Investigation latency (Day 2). Two-way ANOVA, main eXect of sex F (1, 51) = 11.34, p = 0.0015. Tukey's post-hoc female AGG vs male AGG, p = 0.0035.*

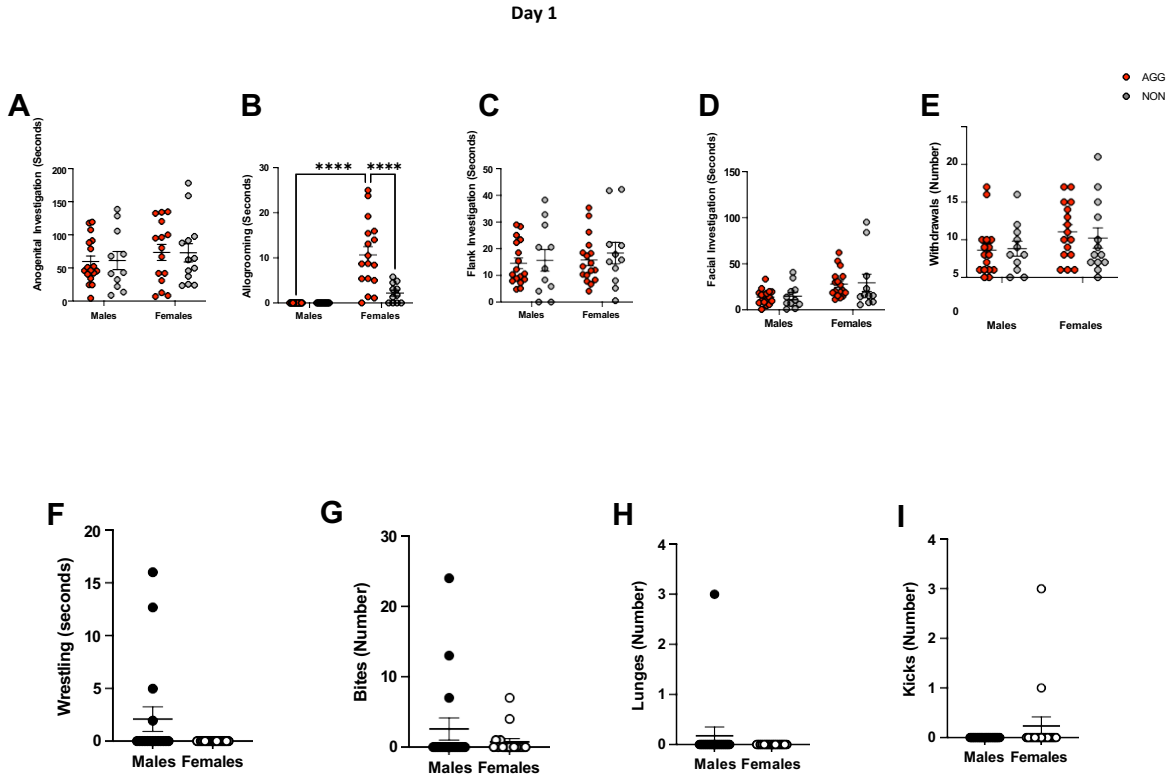


Figure 5.7. Figure S2. Quantification of distinct social behaviors on day 1. *A. Anogenital investigation. Two-way ANOVA, no eXect of sex or phenotype. B. Allogrooming. Two-way ANOVA, sex x phenotype interaction, F (1, 52) = 9.518, p = 0.0033. Tukey's post-hoc test, female AGG vs male AGG, p < 0.0001. Female AGG vs female NON, p = 0.003. C. Flank investigation. Two-way ANOVA, no eXect of sex or phenotype. D. Facial investigation. Two-way ANOVA, main eXect of sex F (1, 52) = 8.751, p = 0.0046. E. Withdrawals, no eXect of sex or phenotype F. Wrestling, Welch's t-test t (16) = 1.793, p = 0.09. G. Bites, Welch's t-test t (18.67) = 1.108, p = 0.281. H. Lunges, Welch's t-test t (16) = 1.00, p = 0.3322. I. Kicks, Welch's t-test t (16) = 1.289, p = 0.215.*

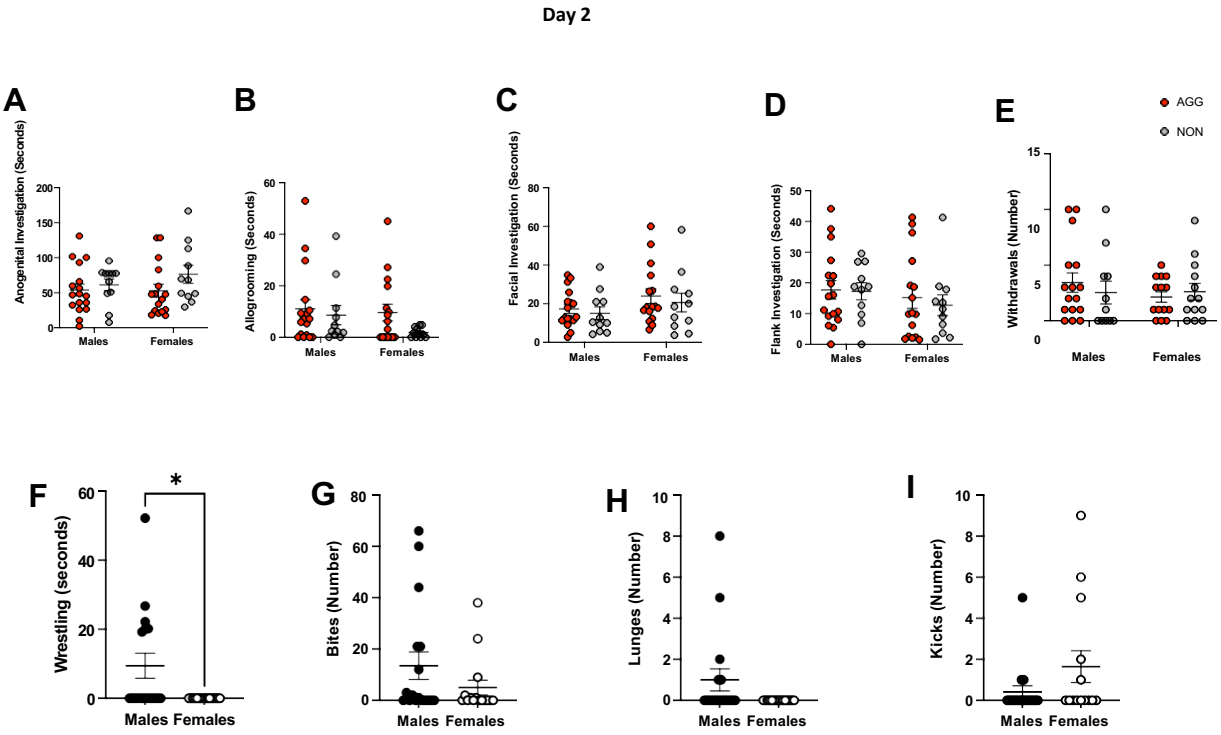


Figure 5.8. Figure S3. Quantification of distinct social behaviors on day 2*. A. Anogenital investigation Two-Way ANOVA, no eXect of sex or phenotype B. Allogrooming. Two-way ANOVA, no eXect of sex or phenotype. C. Flank investigation. Two-way ANOVA, no eXect of sex or phenotype. D. Facial investigation. Two-way ANOVA, main eXect of sex F (1, 51) = 3.142, p = 0.0823. E. Withdrawals. Two-Way ANOVA no eXect of sex or phenotype. F. Wrestling, Welch's t-test t (16) = 2.59, p = 0.0194. G. Bites, Welch's t-test t (24.22) = 1.402, p = 0.1735. H. Lunges, Welch's t-test t(16) = 1.867, p = 0.0803. I. Kicks, Welch's t-test t (16.83) = 1.483, p = 0.1565*

Table 5.1. Supplementary Table 1. Emission probability matrix. *Each cell indicates the probability of observing a given behavior in each of the four states.*

	S1	S2	S3	S4
AG sniff	0.08605625	0.22792224	0.32384628	0.34003168
Allogrooming	1.39E-26	0.06597771	0.05617866	0.03629472
Bite	0.32261289	0.3721225	0.01189169	5.79E-06
End	0.11603446	0.02531174	0.06355925	0.27885629
Face sniff	0.08426891	0.12043818	0.2316754	0.14977904
Flank sniff	0.02677525	0.04650898	0.23597293	0.15385922
Kick	0.13225733	0.03338049	0.00245818	5.73E-62
Lunge	0.08537152	6.75E-18	1.31E-68	6.39E-14
Pin	0.02424978	0.05413276	0.00619664	0.00841112
Withdraw	0.04750985	0.05280663	0.06822098	0.03190471
Wrestle	0.07486376	0.00139876	9.10E-57	0.00085743

***Table 5.2. Supplementary Table 2. Transition probability matrix**. Each cell indicates the probability of transition from one state to another.*

	S1	S2	S3	S4
S1	0.92266881	0.04333721	0.02364991	0.01034407
S2	0.05223531	0.86446045	0.05780613	0.02549812
S3	0.03172461	0.09440697	0.74857678	0.12529164
S4	0.00236107	0.00951385	0.02598266	0.96214243

	M	F
FA->AG	0.32	0.46
FA->FK	0.2	0.14
FK>FA	0.08	0.25
FK>AG	0.34	0.51
AG>FA	0.06	0.18
AG>FK	0.19	0.3

	M	F
B->AG	0.09	0.26
B->FA	0.01	0.10
B->FK	0.04	0.04
AG->B	0.05	0.21
FA->B	0.05	0.08
FK->B	0.04	0.03

Table 5.3. Supplementary Tables 3A and 3B. Transition probabilities *between investigate behaviors (A) and between bites and investigative behaviors in males and females (B). Key: AG: anogenital, B: Bite, FA: Face, FK: flank.*

Table 5.4. Supplementary Table 4. Data analysis from self-administration experiments

Figure number	**Test**	**Factor name**	**F-value**	***p*-value**
Figure 4B. CPP testing				
Figure 4D. RI Screening	Chi-square	Sex	Chi-square, df = 13.57, 1	<0.001*
Figure 4D. Day 3 RI attack latency	One-way ANOVA	Session (between)	$F_{2,56}$=14.47	<0.001*
Figure 4E. Attack & reward trials	Two-way ANOVA	Interaction Day Sex	$F_{24,306}$=3.327 $F_{8,306}$ = 6.787 $F_{3,306}$ = 108.2	<0.001* <0.001* <0.001*
Figure 4E. Latency to press	Two-way ANOVA	Interaction Day Sex	$F_{8,148}$=2.475 $F_{8,148}$=4.657 $F_{1,148}$=3.224	0.015* <0.001* 0.075
Figure 4E. Exploratory head entries	Two-way ANOVA	Interaction	$F_{8,153}$=0.1520	0.996
Figure 4F. Attack & reward trials	Two-way ANOVA	Interaction Day Attack	$F_{8,108}$=9.277 $F_{8,108}$=13.02 $F_{1,108}$=490.9	<0.001* <0.001* <0.001*
Figure 4F. Latency to press	RM one-way ANOVA	Day	$F_{1.942,11.65}$=2.658	0.113
Figure 4F. Exploratory head entries	RM one-way ANOVA	Day	$F_{3.977,23.86}$=2.46	0.073
Figure 4G. Operant acquisition	Chi-square	Sex, RI phenotype	Chi-square, df = 0.3752, 2	0.829
Figure 5B. Rewarded trials	RM one-way ANOVA	Day	$F_{2.487,17.41}$=3.194	0.057
Figure 5B. Attack trials	RM one-way ANOVA	Day	$F_{2.557,17.9}$=0.4350	0.701
Figure 5B. Latency to press	Mixed-effects with Geisser-Greenhouse correction	Day	$F_{3.892,26.27}$=0.833	0.514
Figure 5B. Exploratory head entries	RM one-way ANOVA	Day	$F_{2.224,15.57}$=1.794	0.197
Figure 5C. Rewarded trials	Mixed-effects with Geisser-Greenhouse correction	Day	$F_{3.111,24.5}$=7.995	<0.001*

Figure 5C. Attack trials	Mixed-effects with Geisser-Greenhouse correction	Day	$F_{2.571,22.82}$=2.707	0.0763
Figure 5C. Latency to press	Mixed-effects with Geisser-Greenhouse correction	Day	$F_{2.746,\ 21.28}$=5.199	0.009*
Figure 5C. Exploratory head entries	Mixed-effects with Geisser-Greenhouse correction	Day	$F_{3.134,\ 27.42}$=1.444	0.251
Figure 5D. Rewarded trials	Two-way ANOVA	Interaction Day Sex	$F_{8,142}$=2.105 $F_{8,142}$=3.945 $F_{1,142}$=8.584	0.039* <0.001* 0.004*
Figure 5E. Food self-administration	Two-way ANOVA	Interaction	$F_{6,96}$=0.4544	0.84

Chapter 6 : Dissociable neural mechanisms of reactive and appetitive aggression

Nastacia L. Goodwin[1,2], Michelle Jin, Jovana Navarrate[1,2], Pranav Anumolu, Yi Y. Zhang[2], Valerie Tsai, Drew-Elizabeth Barger, Sam Husarik, Simon R.O. Nilsson[1], Sam A. Golden[1,2,3]

[1] University of Washington, Department of Biological Structure, Seattle, Washington, USA
[2] University of Washington, Graduate Program in Neuroscience, Seattle, Washington, USA
[3] University of Washington, Center of Excellence in Neurobiology of Addiction, Pain, and Emotion (NAPE), Seattle, Washington, USA

* Corresponding Author: Sam A. Golden (sagolden@uw.edu)

Abstract

Reactive (fight or flight) and appetitive (sought after) aggression have long been recognized as separable phenomena in the clinical realm, but appetitive aggression remains preclinically understudied. Furthermore, females have been largely excluded from non-maternal aggression research due to the historical exclusion of female subjects in the biomedical sciences and the low propensity of inbred female mice to show aggression. We recently published a study indicating that while males who are reactively aggressive tend to go on to show appetitive aggression seeking, reactively aggressive females do not. Using whole-brain c-fos activity mapping in conjunction with network analysis, we here identify unique whole-brain activity patterns associated with the suppression of appetitive aggression in female mice using iDISCO+ clearing protocols. We further demonstrate that reactive and appetitive aggression phenotypes in males are controlled by distinct whole-brain networks, and we highlight the role of the lateral septum in the differential control of these aggression phenotypes. This work is the first direct comparison of both appetitive aggression seeking across the sexes, and of reactive versus appetitive aggression phenotypes in males. Establishing these whole-brain datasets and further identifying the lateral septum as a locus of differential aggression control is a key step in further establishing appetitive aggression seeking preclinical models.

Key words: Aggression, mice, sex, c-fos

Introduction

Maladaptive aggression seeking behavior and increased irritability are a component of many neuropsychiatric disorders, including intermittent explosive disorder, autism spectrum disorders, and post traumatic stress disorder. Inappropriate aggression seeking is detrimental to both individuals and society, and current treatment options are largely ineffective, or associated with significant unwanted side-effects. Clinically, reactive 'fight or flight' aggression and appetitive 'sought after' aggression are typically viewed and treated as distinct phenomena. Preclinically, however, the majority of studies consider reactive aggression phenotypes in males due to the relative ease of testing. Appetitive aggression testing in males has begun to expand with the advent of user-friendly operant aggression self-administration protocols, but there has been no direct comparison between the neural basis of reactive versus appetitive aggression phenotypes, and little work in either type of aggression in females.

Direct preclinical measures of appetitive (rewarding) aggression have existed for decades but their widespread adoption was stymied by the complexity of running the tasks. As different labs have begun to use operant boards and boxes for semi-automated procedures, however, operant aggression tasks are becoming more common in the literature (Fish et al. 2002, 2005; Bannai et al. 2007; Golden et al. 2017; Minakuchi et al. 2022). Studies on appetitive aggression in male mice have shown that a subset of outbred males show addiction-like aggression seeking and consummatory behaviors (Golden et al. 2017), and the motivational aspect of this seeking is driven in part by GABAergic basal forebrain projections to the lateral habenula (Golden et al. 2016) and DRD1 positive neurons in the nucleus accumbens (Golden et al. 2019). Furthermore, aggression motivation appears to be gated by a medial preoptic area to ventrolateral ventromedial hypothalamus GABAergic pathway (Minakuchi et al. 2022). While this growing body of literature is beginning to untangle the role of specific brain regions in aggression motivation in males, both reactive and appetitive aggression phenotypes in females are understudied.

Non-maternal female aggression has been excluded from study due to many factors, including the lack of historical inclusion of females in biological studies (Beery and Zucker 2011; Upchurch 2020; Merone et al. 2022), and the lack of a proclivity for aggression in mice of inbred strains (Jones and Brain 1987). What female mouse work has been conducted typically examines reactive aggression in outbred animals. Under isolate housing, female CFWs will

readily attack same sex juveniles (White et al. 1969; Hashikawa et al. 2017; Aubry et al. 2022), or adult males (Ferrari et al. 1996). When paired with a castrated male partner, female CFWs are highly aggressive toward intruders (Newman et al. 2019; Aubry et al. 2022). Outside of CFWs, isolate-housed and short photoperiod-exposed female California mice are aggressive toward same sex conspecifics and show increases in pERK activity in the bed nucleus of the stria terminalus and medial amygdala following reactive aggression opportunities (Silva et al. 2010). Furthermore, female house mice and prairie mice also show aggression toward same sex juvenile conspecifics (Gray 1979; Ayer and Whitsett 1980).

Using outbred CFW mice, we recently published a study showing that while they are reactively aggressive, very few females are subsequently aggressive in appetitive aggression tasks. In this study, we examine whole-brain fos activity from a subset of these animals to better understand the neural correlates driving appetitive aggression behavior in males versus those gating appetitive aggression in females. We also perform an unsupervised analysis of male and female reactive attack behavior, identifying separable latent motifs between the sexes. Whole-brain c-fos mapping serves as a proxy for strongly activated neurons and can identify regions and networks of regions that are activated by appetitive self-administration experiences. Briefly, we screened CFW males and females for reactive aggression in RI tests prior to ten days of self-administration training (test subjects) or operant box exposure (controls). On the tenth day of testing, we collected their brains 70 minutes after the start of testing, which we then immunostained, cleared, and analyzed for whole-brain c-fos activity using the optimized iDISCO+ protocol described in (Madangopal et al. 2022). We used the ClearMap pipeline (Renier et al. 2014, 2016; Kirst et al. 2020) for unbiased mapping of self-administration associated neural activation patterns throughout the mouse brain, and the SMART-R package for correlational and network analyses. Through this analysis, we uncovered dissociable networks of activity between males and females, indicating that the medial amygdala may gate female appetitive aggression following appetitive social seeking.

Methods

Animals

Mice for the whole-brain experiment were a subset of those used in Aubry et al., 2022. Briefly, 10-week old male CFW mice (Charles River Laboratories) were either housed in isolation or

with a stimulus female for 48-hr prior to two weeks of isolate housing. Females were either isolate housed or housed with a castrated male for two weeks. Same sex 12-week-old C57BL/6J mice were used as intruders, and all assays were performed during the dark cycle.
Mice used for the unsupervised reactive aggression analysis had supervised analysis data published in (Goodwin et al., 2024). Male CD-1 (strain #022, Charles River Labs (CRL)) or female CFW (strain #024, CRL) white coat colored 12-week-old mice were used as residents. Female CFW residents were pair housed with 10-week-old castrated CFW male mice for the duration of the study to elicit aggression, while male CD-1 residents were singly housed. Sex-matched, black coat colored C57BL6/J (C57; strain #000664, Jackson Labs) mice were used as intruders for all aggression assays. C57 females were > 10 weeks old, and males were > 8 weeks old. We gave all mice free access to standard food chow and water in all experiments. We housed all mice with enrichment (cotton padding) in standard Allentown clear polycarbonate cages covered with stainless-steel wire lids at least one week prior to experiments, and we maintained them on a reverse 12-h light/dark cycle (light off at 0900 am).
All procedures were performed in accordance with the National Institutes of Health (NIH) Guide for Care Use of Laboratory Mice and approved by the University of Washington Institutional Animal Care and Use Committee.

Behavior
For the whole-brain dataset, behavioral assays are described in Aubry et al., 2022. Briefly, all animals were pre-screened for reactive aggression toward a same-sex intruder in resident intruder (RI) assays for 5 minutes per day for three days. 3 days following RI screening, mice underwent one day of magazine training, followed by 10 days of self-administration training. Animals were trained in a specialized operant box with a guillotine door and side cannister for intruder mice as described in (Golden, 2017). Animals were able to press for access to an intruder on an FR1 schedule for 12 2-minute trials, with 2-minute inter-trial intervals per day. Animals had access to an empty food port with beam break readings and an inactive lever at all times. During trials, a house light went on in the chamber, followed 10 seconds later by an active lever extending. Following lever press, a two-second tone sounded, and the guillotine door opened five-seconds after lever press. An intruder mouse was ushered into the chamber with the resident, and the guillotine door closed 17 seconds after lever press. Five female mice showed at least one bout of aggression during the test day session, and were held out for separate analysis.

For the unsupervised reactive aggression analysis we used male and female repeated resident—intruder procedures as previously published. Briefly, we recorded dyadic encounters between male or female mice in clear polycarbonate cages (cage size: 28x19x12cm) divided in half by a clear acrylic barrier. Aggressive CD1 or CFW resident animals were housed on one side of the barrier where they encountered an unfamiliar sex-matched C57 intruder for 5 min (n = 11 female residents, n = 10 intruders) or 10 min (n = 21 male residents, 21 intruders). We analyzed five days of data per aggressive mouse.

All assays were filmed from above at 30FPS by WhiteMatter or Basler cameras at resolutions ranging from 320 x 300 to 1550 x 1050. Performance was assessed on videos of varying resolution.

Behavioral scoring

During operant training, observers marked rewarded trials and trials with attacks. Videos were preprocessed in SimBA prior to undergoing pose estimation via DeepLabCut (Version 2.2). In DeepLabCut, we tracked 7 points per animal (nose, ears, sides, body center, tail base) per animal and labeled frames from a diverse set of videos to create training sets for ResNet 50 based neural networks for all models. Pose estimation was loaded into SimBA, followed by outlier correction (location criteria: 2.0, movement criteria: 1.0) and video calibration. Classifiers for attack, anogenital sniffing, body sniffing, face sniffing, defensive behavior, and escape behavior were previously created using diverse sets of videos and subject animals and assessed for hand versus machine performance. Operational definitions are as follows:

Attack– Clear physical antagonistic interaction initiated by the resident, characterized by tussling, biting, boxing, and/or corralling.

Anogenital sniffing– Resident is sniffing the anogenital region of the intruder, behind the back legs.

Body sniffing– Resident is sniffing between the back legs and ears of the intruder.

Face sniffing– Resident is sniffing the face in front of the ears of the intruder.

Defensive– Intruder mouse is fighting back against resident by pushing or biting. Intruder is not instigating aggression.

Escape– Intruder mouse is running away or attempting to run away from resident.

Thresholds for the CFW operant videos were adjusted to best balance false positive and false negative detections, and were set as follows: Attack 0.45 & 0.3 (males and females respectively),

Escape 0.4, Defensive 0.2, Anogenital 0.4, Body sniffing 0.5, Face sniffing 0.35. All videos were analyzed with all classifiers in SimBA.

Unsupervised learning

Behavior-positive frames detected by SimBA supervised learning were further processed using unsupervised dimensionality reduction and clustering techniques. Importantly, unsupervised analyses of individual behavior frames may interfere with interpretability by introducing biases for clustering based on latent temporal identifiers of individual frames within attack bout sequences rather than differences across separate attack bouts. To overcome this, we first smoothed the time series of individually classified frames into bouts with the Kleinberg burst detection method using previously employed parameters (sigma: 0.3, kappa: 2, hierarchy: 2-4). We next calculated the mean feature values within each behavior burst event, collapsing each event into a single observation. Sex-dependent features are measurements of animal width, length and area; and their inclusion may bias unsupervised algorithms towards clustering observation by apparent size overt behavioral attributes. To account for this, preceding unsupervised analysis, we discarded features measuring animal size and used a remaining battery of 289 features for dimensionality reduction and clustering. Dimensionality reduction was performed using UMAP (spread: 1.0, distance: 0.1, neighbors: 4-50, dimensions: 2). Hierarchical clustering was performed using HDBSCAN (min cluster: 100, min samples: 1, metric: Euclidean).

We used SimBA to calculate feature permutation importances for decision trees targeting each unsupervised cluster. These scores gauge the importance of individual features for correctly classifying individual behaviors by evaluating the loss of predictive power when the feature, and no other feature, is scrambled. Hence, we first created supervised random forest classifiers (estimators: 2k, max features: sqrt, criterion: entropy, min sample leaf: 1) where the observations were the aggregated mean feature values during the attack bouts, and the target was the HDBSCAN-assigned cluster assignment. We further calculated Shapley values (SHAP) by creating binary random forest classifiers with cluster 1 as positive for each group, and additional clusters as negative in turn. We calculated SHAP values for 1,000 frames or all frames present if less were available.

Brain collection & clearing

On day 10 of SA training, animals were perfused with ice cold PBS and formalin 70 minutes after the start of their testing. Brains were cleared and stained for c-fos using a modified iDisco+ protocol (see Mandangopal and Szelenyi et al PNAS 2022). We imaged, stained, and cleared intact mouse brains using LSFM. We analyzed the data using ClearMap (1) and SMART (2) analysis pipelines as described below. Light-sheet fluorescent microscopy imaging (LSFM) We used a light-sheet microscope (UltraMicroscope II with Infinity Corrected Objective Lenses, Miltenyi Biotec) with an attached camera (Andor Zyla sCMOS), and a 1.1x/0.1NA objective (MI PLAN; LaVision BioTec). Imaging parameters and acquisition order were controlled through ImspectorPro software (v 7.1.4). We mounted cleared Fos-stained brains in horizontal orientation using a custom sample platform and imaged at 2.2x effective magnification (1.1x objective x 2x magnification slider) in DBE. We acquired images for autofluorescence (Excitation: 488 nm laser, Emission: 535/43 bandpass filter) and Fos-IHC (Excitation: 647 nm laser, Emission: 690/50 bandpass filter) in separate 2 x 1 tiled scans (scan order: z-x-y). We used the following fixed parameters for acquisition: exposure = ~100 ms for 488, ~135ms for 647; sheet NA = 0.16; sheet thickness = 3.89 um; sheet width = 100%; zoom = 2x; dynamic horizontal focus = 7 (Fos channel only); dynamic horizontal focus processing = blend; merge light-sheet = blend; 488 nm laser power = 20%; 647 nm laser power = 30%. Final image pixel resolution was 2.956 um X x 2.956 um Y x 3 um Z. Resulting tiles were stitched into full size coronal planes using Arivis Vision 4D (3.0.0) and exported as TIFFs. One test male brain was excluded due to damage, and one test female was excluded due to a total cell count > 2SD above average.

ClearMap analysis: We used the open-source program ClearMap 1.0 (1) for whole-brain volumetric analysis on a dedicated machine (Intel Xeon® CPU E5-2650 v4 @ 2.20GHz x 48; 4 x GeForce GTX 1080 Ti/PCIe/SSE2; 256GB RAM). We downsampled autofluorescence image stacks for each sample 7 and registered them to a common 25 μm isotropic serial two-photon (STP) tomography reference template (3). We manually validated registration for each sample by post-hoc inspection of overlaid reference template and post-transformation image stacks in ImageJ.

Fos detection: Next we created a random forest machine learning based classifier for fos detection using the open source ilastik program. To validate the classifier, an expert scorer hand counted fos positive neurons in nineteen field of views across four brains with a mean F1 score

of 0.69. We calculated precision (ratio of correctly predicted Fos+ cells to all predicted cells), recall (ratio of correctly predicted Fos+ cells to expert annotated Fos+ cells), and F-score (harmonic mean of precision and recall) in Microsoft Excel. We warped all ClearMap detected Fos+ cells into the reference space by applying transformation coordinates from the registration step and obtained counts for individual brain regions based on the Allen Brain Institute atlas ontology provided with the ClearMap installation package. We extracted Fos+ cell counts for all annotations and used custom python scripts to generate summed counts within regions of interest (ROIs) for analysis of activity changes between groups.

Data analysis

Statistical analysis was conducted using Prism 10 (Graphpad Software). Behavioral data was compared using two-way or repeated measures ANOVAs as appropriate. Whole-brain data was compared using either two-way ANOVAs or unpaired t-tests with FDR correction (Q = 2%) as appropriate.

Results

Males show higher appetitive aggression behavior than females

There were no significant differences in rewards between housing groups and sexes ($p = 0.5767$, $p = 0.5551$, females and males respectively), and we therefore collapsed housing groups. Males showed significantly lower latency to attack on day 3 of resident intruder screening than females ($p < 0.0001$, $t=4.520$, $df=28$, Fig. 1C). Rewards per day increased significantly by day, with no sex effect ($p = 0.0403$, $F (9, 251) = 1.995$ main, $p < 0.0001$ by day, $p = 0.3870$ by sex, Fig. 1D), while attacks were stable by day but lower in females than males ($p = 0.9970$, $F (9, 251) = 0.1666$ main; $p < 0.0001$, $F (1, 28) = 35.06$ sex; $p = 0.2582$, $F (3.557, 99.21) = 1.355$ day Fig. 1D). Latency to press decreases by day ($p < 0.0001$, $F (4.251, 115.7) = 7.818$), but not by sex ($p = 0.8884$, $F (1, 28) = 0.02004$, Fig. 1D). Similarly, exploratory head entries decrease by day ($p = 0.0015$, $F (4.433, 113.3) = 4.480$ Fig. 1D), but are not different between sexes ($p = 0.2956$, $F (1, 28) = 1.136$, Fig. 1D).

Males showed more attack behavior as measured by total duration, number of bouts, and mean bout duration than females, but only number of attack bouts differed significantly by both day

and sex ($p = 0.0413$, $F(3, 82) = 2.871$ main; $p = 0.0595$, $F(2.674, 73.09) = 2.679$ day, $p < 0.0001$, $F(1, 28) = 41.42$ sex; Fig. 2B). No other behaviors significantly differed ($p > 0.05$).

SHAP values reveal sex differences between feature bins

With the exception of attack intruder movement, all SHAP feature bins for all behaviors were significantly different between sexes and/or across time (Fig. 3B). Attack, defensive and escape behavior between the sexes most differed by animal distances, while anogenital sniffing was most significantly differentiated by intruder and resident shape. Face sniffing was most differentiated by resident shape, while body sniffing was most differentiated by intruder shape (Supplementary Note 1 - Statistical Appendix.).

Targeted unsupervised analysis of supervised attack bouts reveals latent aggression motifs

As a proof of principle, we first performed dimensionality reduction and clustering using an unlabeled dataset comprising five classified resident behaviors (male lateral threat, male anogenital sniffing, female and male allogrooming, pursuit, and attack events), and two intruder behaviors (female and male defensive and escape events). When clustering the five resident behaviors, we identified five behavioral clusters (Fig. 2b). The five clusters comprised of i) male pursuit, ii) male attack, iii) male anogenital sniffing, iv) male and female allogrooming, v) and female attack and pursuit behavior. Although a gradient could be observed with the female attack with pursuit cluster, the apparent delineation between the behaviors observed in the males was not present in the females. This could be explained by pursuit and attack behavior often co-occurring in the females but not the males. Similarly, when analyzing the two intruder behaviors, we identified three behavioral clusters (Fig. 2b). The three clusters comprised of i) male escape, ii) male defensive behavior, and iii) female defensive and escape behavior. For subsequent unsupervised analyses, we focused exclusively on attack behavior.

Female and male reactive testing attacks separated into three (Fig. 2d) and two clusters (Fig 2e) respectively. Feature permutation and SHAP scores indicated that the clusters differed most by animal movement. Contingent attacks separated into two clusters for both females (Fig. 2f) and

males (Fig. 2g), and also differed most by animal movement. Visual inspection of the cluster videos confirmed these differences, with clusters largely separating into longer attacks involving chasing, and shorter attacks involving lateral threats with defensive or upright submissive postures from the intruder mouse.

Test males show increased c-fos density in several regions

Our broadest level of analysis examined cell densities in the isocortex, olfactory areas, hippocampal region, cortical subplate, striatum, pallidum, thalamus, hypothalamus, midbrain, and hindbrain. Male test animals showed significantly higher c-fos densities in the striatum ($p = 0.0176$) and cortical subplate ($p=0.0455$), with no other significant differences.

Our next level of analysis examined c-fos densities in 836 sub-regions as defined by the Unified Anatomical Atlas. We excluded cerebellar regions from analysis due to damage in several brains. For an exhaustive list of significantly different sub-regions per group, please see Supplementary Note 1 - Statistical Appendix. At large, male test animals showed significantly increased c-fos in subregions of the medial amygdala (MeA), bed nucleus of the stria terminalus (BNST), amygdalohippocampal area, striohypothalamic nucleus, nucleus of the stria medullaris, zona incerta, caudoputamen, and the paraventricular hypothalamic nucleus (Extended Data Fig. 1, $q < 0.02$, df 25.00).

Test females show significantly higher c-fos density in the medial amygdala and suprachiasmatic nucleus

Females showed no significant difference in c-fos density between groups in the isocortex, olfactory areas, hippocampal region, cortical subplate, striatum, pallidum, thalamus, hypothalamus, midbrain, or hindbrain.

Examining subregions of these areas, female test animals showed significantly increased c-fos activity in subregions of the MeA and suprachiasmatic nucleus (Extended Data Fig. X, $q < 0.002$, df = 23.00).

Males and females show distinct neuronal response patterns to appetitive aggression opportunities

Females and male test animals did not differ in c-fos density in the isocortex, olfactory areas, hippocampal region, cortical subplate, striatum, pallidum, thalamus, hypothalamus, midbrain, or hindbrain.

Female test animals showed higher c-fos activity than test males in the agranular insular cortex and cingulate cortex A24b, while males showed higher activity in several subregions of the insular cortex, paraventricular hypothalamic nucleus, BNST, trigeminal nuclei, primary somatosensory cortex, motor cortex, salivary nucleus, red nucleus, striohypothalamic nucleus, lateral lemniscus, subcoeruleus nucleus, caudoputamen, cingulate cortex A25, nucleus of the stria medullaris (Fig. 3d).

Non-contingent intruder administration leads to increased aggression levels in CFW males

Contingent and non-contingent males showed a similar number and percentage of attack positive trials across training days, though the percent of attack trials decreased across days for both groups (number of attack positive trials: $F(9, 126) = 0.6290$, $p = 0.7705$, % attack trials: $F(9, 122) = 3.316$, $p = 0.0012$, Fig. 4c).

Despite this lack of trial differences, non-contingent males showed significantly higher total attack duration, number of attack bouts, and mean bout duration than contingent males (Duration: $F(3, 41) = 3.176$, 0.0340, Bouts: $F(3, 41) = 2.876$, 0.0476, Mean bout duration: $F(3, 41) = 3.895$, 0.0154, Fig. 4d). Among the resident mice, the groups showed similar levels of anogenital sniffing, body sniffing, and face sniffing. Among the intruders, defensive behavior was similar across groups but intruders facing non-contingent males showed higher average escape bouts and mean bout duration than the intruders facing contingently administering males (Bouts: $t=2.560$ $df=14.00$, $p = 0.022683$, Mean bout duration: $t=3.426$ $df=14.00$, $p=0.004099$, Fig. 4d).

With the exception of the feature bins animal distances, intruder and resident movement, and intruder and resident shape for the face sniffing classifier, all SHAP feature categories differed significantly between groups ($p<0.05$, Fig. 4d, Supplementary Note 1 – Statistical Appendix).

Unsupervised clustering revealed three primary attack clusters for non-contingent males, differing most by features including intruder area, mouse movement, and mouse Euclidean distances (Fig. 4e, See Supplementary Note 1- Statistical Appendix). Contingent male attacks separated into two clusters, differing by movement, size changes, mouse angles in relation to each other, and mouse distances (Fig. 4f).

SHAP analysis of unsupervised clustering revealed significant differences in feature bin importance for both non-contingent males and contingent males, but only non-contingent males showed a different reliance on time bins (Fig. 4g, Non-cont. feature bin $p < 0.0001$, time bins $p = 0.0106$, Contingent feature bin $p = 0.0002$, time bins $= 0.0594$).

Reactive and appetitive aggression in males are associated with different neural activity patterns

Non-contingent CFW males showed significantly higher c-fos density in the lateral septum, subregions of the somatosensory cortex, the caudate putamen, the bed nucleus of the stria terminalus, and several nuclei of the thalamus (q value < 0.02, Fig. 5d, Supplementary Note 1 – Statistical Appendix).

Discussion

Male and female aggression motifs differ between contingent and non-contingent aggression

When analyzed on a frame-by-frame basis, male and female social sniffing behavior is not significantly different. Males, however, show more aggression following appetitive self-administration than females. Using attack data from the few females who showed appetitive aggression (216 total bouts versus 1,889 in males), we found that females demonstrate two attack motifs, defined by similar top features and SHAP values, including resident and intruder movement, resident movement, and to a slightly different extent intruder movement. Using

unsupervised clustering when examining reactive aggression, however, we found that males and females show a different number of aggression motifs, defined by different top features including intruder movement in females and resident shape in males. These findings indicate that female and male reactive aggression techniques may be more diverse than appetitive aggression, potentially due to arena differences as I found in (Goodwin et al. 2024).

One consideration of our experimental setup is that it closely aligns with the rat empathy literature in which a partner rat is trapped in a cannister, and the subject rat has to work to free it (Bartal et al. 2011), with or without access to the partner rat afterwards (Cox and Reichel 2020). Unpublished data from our lab indicate that female mice show slightly higher rates of lever pressing for full physical interaction with an intruder than for interaction through jail-bar barriers (Zhang et al., in progress). Males, however show no difference in willingness to press for physical versus barrier contact. While there are strong arguments within the field regarding whether or not freeing a fellow trapped animal constitutes empathy, it may be the case that our experimental setup is driving females' willingness to self-administer a partner. Their choice to fight or not fight afterwards is volitional, and the mechanisms suppressing their appetitive aggression can be directly compared to those permitting male aggression.

Anterior cingulate activity is a likely candidate for the suppression of female contingent aggression

Female CFW mice who are reactively aggressive do not go on to show appetitive aggression in self-administration tasks, though they show similar levels of self-administration as males (Aubry et al. 2022), and we found similar levels of social sniffing. In females, we found that this suppression of appetitive aggression is associated with significantly increased c-fos densities versus males only in the agranular insular cortex (AIC) and the anterior cingulate cortex (ACC). The AIC is involved in reward prediction in rats (Kesner and Gilbert 2007), and the ACC has been implicated in empathy-like behavior in rats, voles, and mice (Burkett et al. 2016; Keum et al. 2018; Yamagishi et al. 2020; Kim et al. 2023). Furthermore, Oxytocin antagonism in the ACC inhibits both the learning of helping behavior in rats, and partner directed responses in prairie voles (Burkett et al. 2016; Yamagishi et al. 2020). Based on these results, I posit that ACC activity in females is inhibiting their appetitive aggression expression, likely via oxytocin

signaling. My future studies to examine this claim would include fiber-photometry and optogenetic manipulations within the ACC in males and females in OT-expressing neurons. I would expect to see reduced activity in males during appetitive aggression seeking, and would expect that oxytocin antagonism would increase appetitive aggression in both sexes. I would further expect that exciting these neurons would lead to increased affiliative behavior and decreased aggression in both sexes.

In males, we show that appetitive aggression seeking is associated with increased c-fos densities versus females in several motor and sensory regions. Although females show similar levels of sniffing as males, males show a clear signature of aggression behavior in the activation of their trigeminal nuclei, primary somatosensory cortex, motor cortex and salivary nucleus. Beyond these regions, males also show significantly elevated c-fos densities in the paraventricular nucleus of the hypothalamus (PVN), the bed nucleus of the stria terminalus (BNST), and the caudate putamen. The PVN is one of the main sources of oxytocin and vasopressin within the brain (Douglas 2005), and vasopressin is known to facilitate offensive aggression in hamsters, prairie voles, and rats (Ferris and Potegal 1988; Potegal and Ferris 1989; Koolhaas et al. 1990; Winslow et al. 1993; Delville et al. 1996a, b). Oxytocin, on the other hand, can increase aggression toward 'out-group' individuals (Beery 2015). My next study would include in situ hybridization of c-fos, oxytocin, and vasopressin to better understand which cell population is active following appetitive aggression experiences. We are currently working with a collaborator to conduct graph theory based network analyses on these whole-brain datasets, which will help us understand networks of co-activated regions for each of our groups.

C-fos expression is not an on/off switch, and activates in a neuronal activity dependent manner, often requiring higher activity levels than other immediate early genes to express (Okuno 2011; Fukuchi et al. 2017). While expression is dependent on calcium activity within neurons, the pattern of calcium influx appears to be more influential than total calcium levels, and a consistent influx appears to coordinate the activation of the mitogen-activated protein kinase (MAPK) pathway and phosphorylation of calcium response element binding protein (CREB), resulting in high levels of c-fos expression (Fields et al. 1997). C-fos expression can also be induced by other pathways including RhoA-actin and phosphoinositide 3 kinase (PI3K) cascades. Furthermore,

phosphorylation of ribosomal protein S6 (another activity marker) shows overlapping but not identical expression patterns with c-fos activity, with each marker displaying tropisms for certain tissues (Knight et al. 2012). Within my studies, whole-brain c-fos expression mapping is an inexpensive first pass to identify putative regions and networks of interest based on differences in c-fos density between groups. While this technique is likely to miss some differences in activity, it is only a first step prior to more intensive examination.

Contingent and non-contingent aggression in males are behaviorally separable

Non-contingent males showed slightly higher levels of aggression than contingent males, in addition to one more attack cluster than contingent males. While unsupervised behavioral analysis - in which you provide unlabeled behavioral events to an algorithm which then clusters them via mathematical rules on similarities and differences – is gaining popularity in behavioral neuroscience, few studies have mapped these clusters back onto neural data. In my ongoing fiber photometry study in GABAergic neurons in the lateral septum, I am currently working to align my neural recordings with different attack motifs to understand whether or not they are associated with different neural signals.

Contingent and non-contingent aggression in males are neurally separable

Males undergoing contingent versus non-contingent aggression assays show striking differences in c-fos activity (Figs. 5a-e). Unlike the female-male dataset, this dataset is able to control for the physical experience of aggression. While there are still differences in the somatosensory cortex, potentially due to differential attack motifs or due to the slightly higher level of aggression in non-contingent males, the majority of regional differences are in thalamic regions, the lateral septum, and the caudate putamen. In our dataset, c-fos activity is significantly increased in non-contingent versus contingent animals. This is in line with prior work which indicates that local inhibitory interneuron networks activate in the caudal and rostral LS to inhibit GABAergic output from the ventral LS to the VMHvl in reactive aggression (Wong et al. 2016; Leroy et al. 2018; Wang et al. 2019). The significant difference in LS activity between the contingent and non-contingent groups indicates that the LS may be differentially influencing the two aggression phenotypes. As stated in the introduction of this dissertation, the LS is centrally placed in known reactive and appetitive nodes, and has significant potential as a key region for the differential

control of male aggression phenotypes. Outside of the scope of my dissertation, our finding of differential LS activity in this whole-brain dataset laid the groundwork from ongoing DREADD and fiber photometry experiments examining LS GABAergic control over different aggression phenotypes.

Ultimately, these findings show that the anterior cingulate cortex is a candidate for the suppression of female appetitive aggression, and that reactive and appetitive aggression phenotypes in males are neurally distinct.

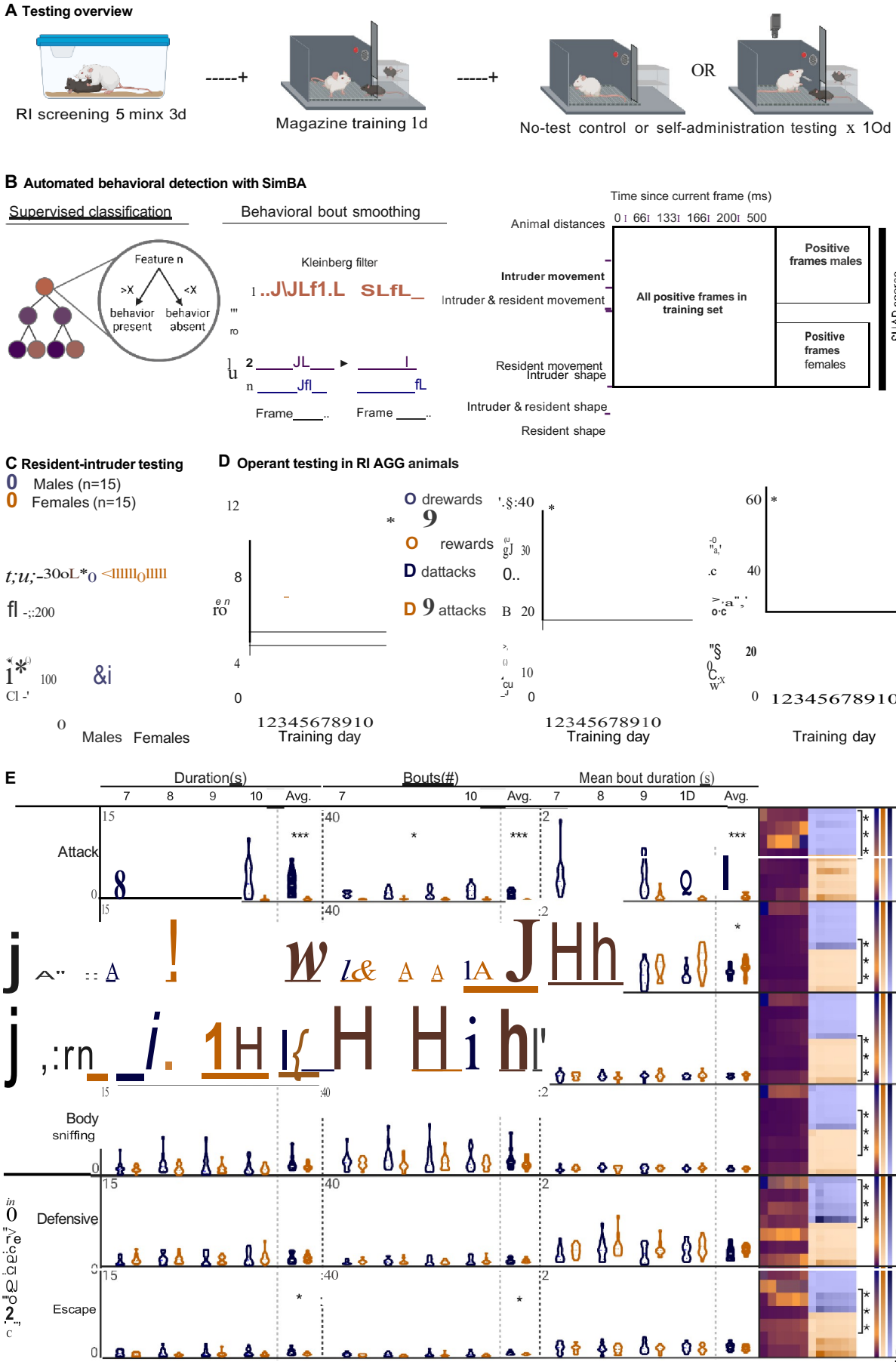
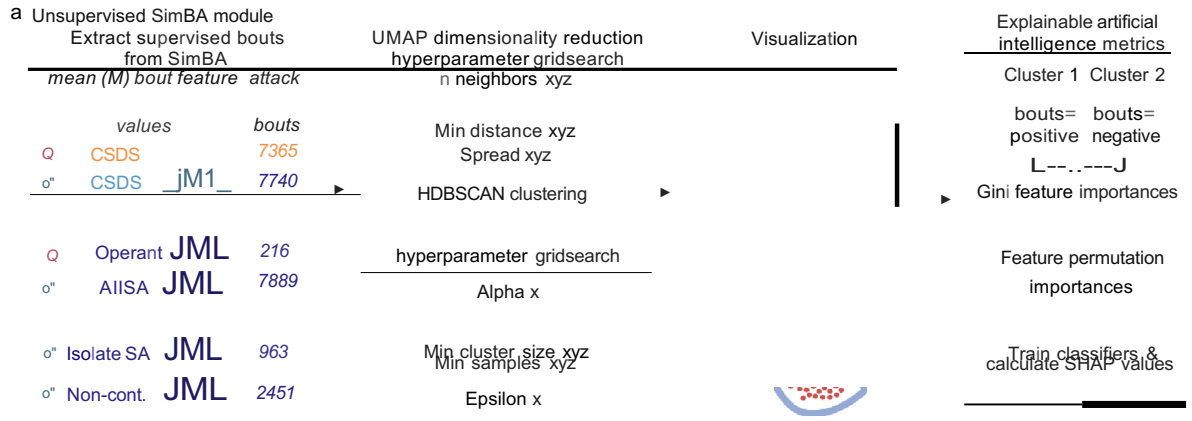
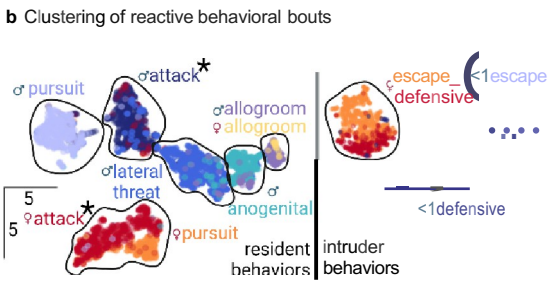
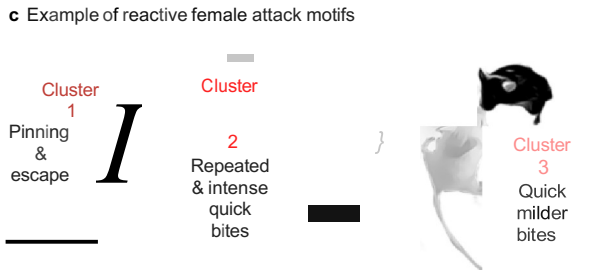
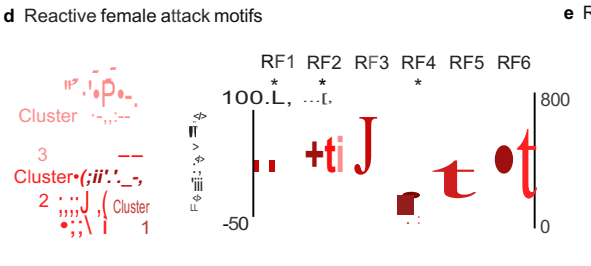

A Testing overview
RI screening 5 minx 3d
Magazine training 1d
No-test control or self-administration testing x 1Od
OR
B Automated behavioral detection with SimBA
Supervised classification
Behavioral bout smoothing
Kleinberg filter
Feature n
behavior present
behavior absent
Time since current frame (ms)
Animal distances
Intruder movement
Intruder & resident movement
Resident movement
Intruder shape
Intruder & resident shape
Resident shape
All positive frames in training set
Positive frames males
Positive frames females
SHAP scores
C Resident-intruder testing
Males (n=15)
Females (n=15)
Males
Females
D Operant testing in RI AGG animals
drewards
rewards
dattacks
attacks
Training day
Training day
Training day
E
Duration(s)
Bouts(#)
Mean bout duration (s)
Attack
Body sniffing
Defensive
Escape

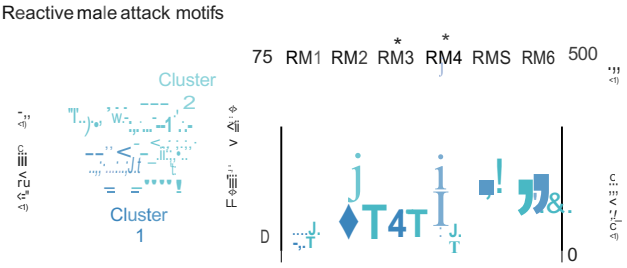
Figure 6.1. Males but not females show appetitive aggression in volitional social seeking tasks. *(a) Testing timeline overview. (b) Overview of automated behavioral detection and machine learning explainability metrics with Simple Behavioral Analysis (SimBA). (c) Females who showed reactive aggression on at least one of three days of testing show higher latency to attack than males on day three of testing. (d) These same animals who were reactively aggressive learn to lever press for social partners, showing increased learning over time, decreased latency to press, and no difference in exploratory head entries. Males but not females, however, show a steady level of attack positive trials over time. (e) Males show significantly higher attack duration, number of bouts, and mean bout duration than females. Sniffing behaviors and defensive behaviors are grossly similar between the sexes, but male intruders show significantly more escape behavior than females. SHAP values (right) indicate that there are significant differences in the ways in which males and females present these behaviors. Stars indicate the most significant feature bins between sexes. All SHAP values and statistical comparisons are listed in Statistical Summary Appendix A.*

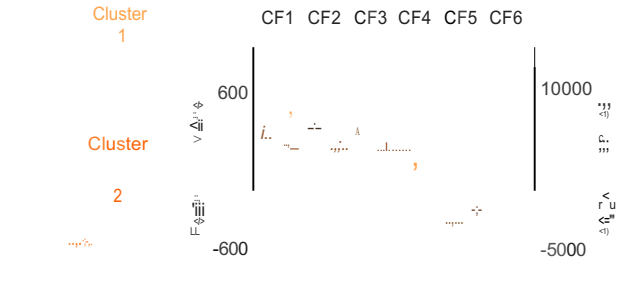

a Unsupervised SimBA module
Extract supervised bouts from SimBA
mean (M) bout feature attack
values bouts
Q CSDS 7365
o" CSDS _jM1_ 7740
Q Operant JML 216
o" AIISA JML 7889
o" Isolate SA JML 963
o" Non-cont. JML 2451
UMAP dimensionality reduction
hyperparameter gridsearch
n neighbors xyz
Min distance xyz
Spread xyz
HDBSCAN clustering
hyperparameter gridsearch
Alpha x
Min cluster size xyz
Min samples xyz
Epsilon x
Visualization
Explainable artificial intelligence metrics
Cluster 1 Cluster 2
bouts= positive bouts= negative
Gini feature importances
Feature permutation importances
Train classifiers & calculate SHAP values

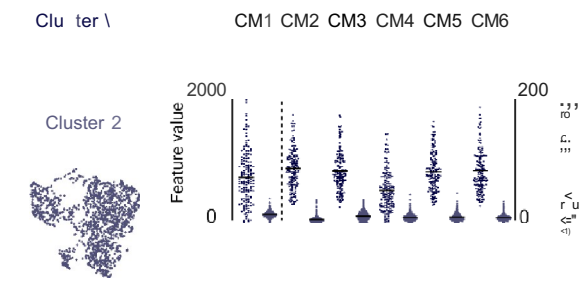

b Clustering of reactive behavioral bouts
♂pursuit
♂attack
♂allogroom
♀allogroom
♂lateral threat
♂ anogenital
♀attack
♀pursuit
resident behaviors
intruder behaviors
♀escape_defensive
<1escape
<1defensive

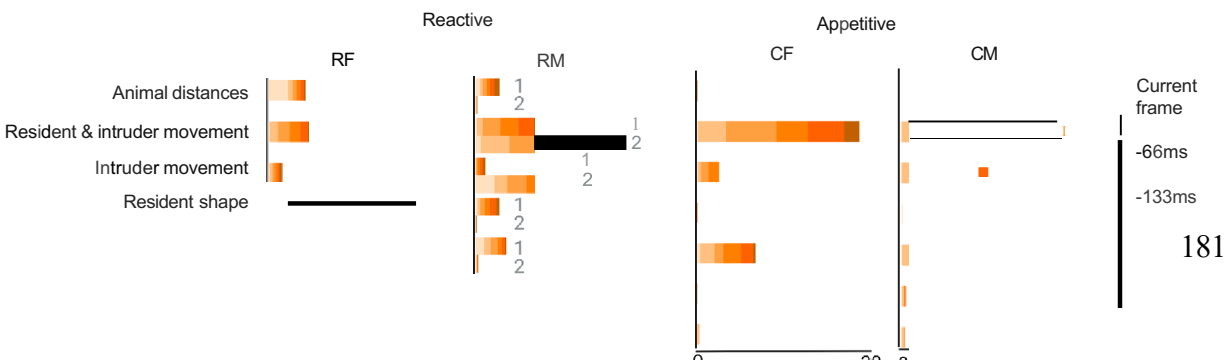

c Example of reactive female attack motifs
Cluster 1
Pinning & escape
Cluster 2
Repeated & intense quick bites
Cluster 3
Quick milder bites


d Reactive female attack motifs
Cluster 3
Cluster 2
Cluster 1
RF1 RF2 RF3 RF4 RF5 RF6
800
-50
0


e Reactive male attack motifs
Cluster 2
Cluster 1
75 RM1 RM2 RM3 RM4 RMS RM6 500
0


f Contingent CFW female attack motifs
Cluster 1
Cluster 2
CF1 CF2 CF3 CF4 CF5 CF6
600
-600
10000
-5000


g Contingent CFW male attack motifs
Cluster 2
CM1 CM2 CM3 CM4 CM5 CM6
Feature value
2000
0
200
0


h SHAP values for cluster 1 as positive versus cluster n as negative
Reactive
Appetitive
RF
RM
CF
CM
Animal distances
Resident & intruder movement
Intruder movement
Resident shape
Current frame
-66ms
-133ms


Resident movement
Intruder shape
Resident & intruder
1
2
1
2
Cluster n
-200ms
-500ms
30

Figure 6.2. Unsupervised analysis reveals distinct attack motifs across sexes*. (a) Schematic of supervised bout extraction, unsupervised analysis via UMAP dimensionality reduction and HDBSCAN clustering, and subsequent SHAP analysis of behavioral clusters. (b) Clustering of resident (attack, lateral threat, allogrooming, anogenital, pursuit) and intruder behaviors (defensive, escape) across sexes. Each dot represents one behavioral bout detected by SimBA classifiers. (c) Example of female attack motifs uncovered in (d). (d-g) Unsupervised analysis of SimBA-detected attack bouts per experimental condition (red = reactive female, light blue = reactive male, orange = contingent appetitive female, dark blue = contingent appetitive male). Each panel contains UMAP and HDBSCAN cluster output. Each panel also contains a feature permutation importances graph with the top 6 most important features plotted per cluster (acronyms listed above each plot and described in Fig. S1). Feature values for each bout in each cluster are plotted and compared via ANOVA or t-test as appropriate. (h) SHAP values for each feature bin across clusters, as in figures 2 and 3. In each group, cluster 1 is compared to cluster n, as denoted on the right. See Appendix A for full statistical analysis.*

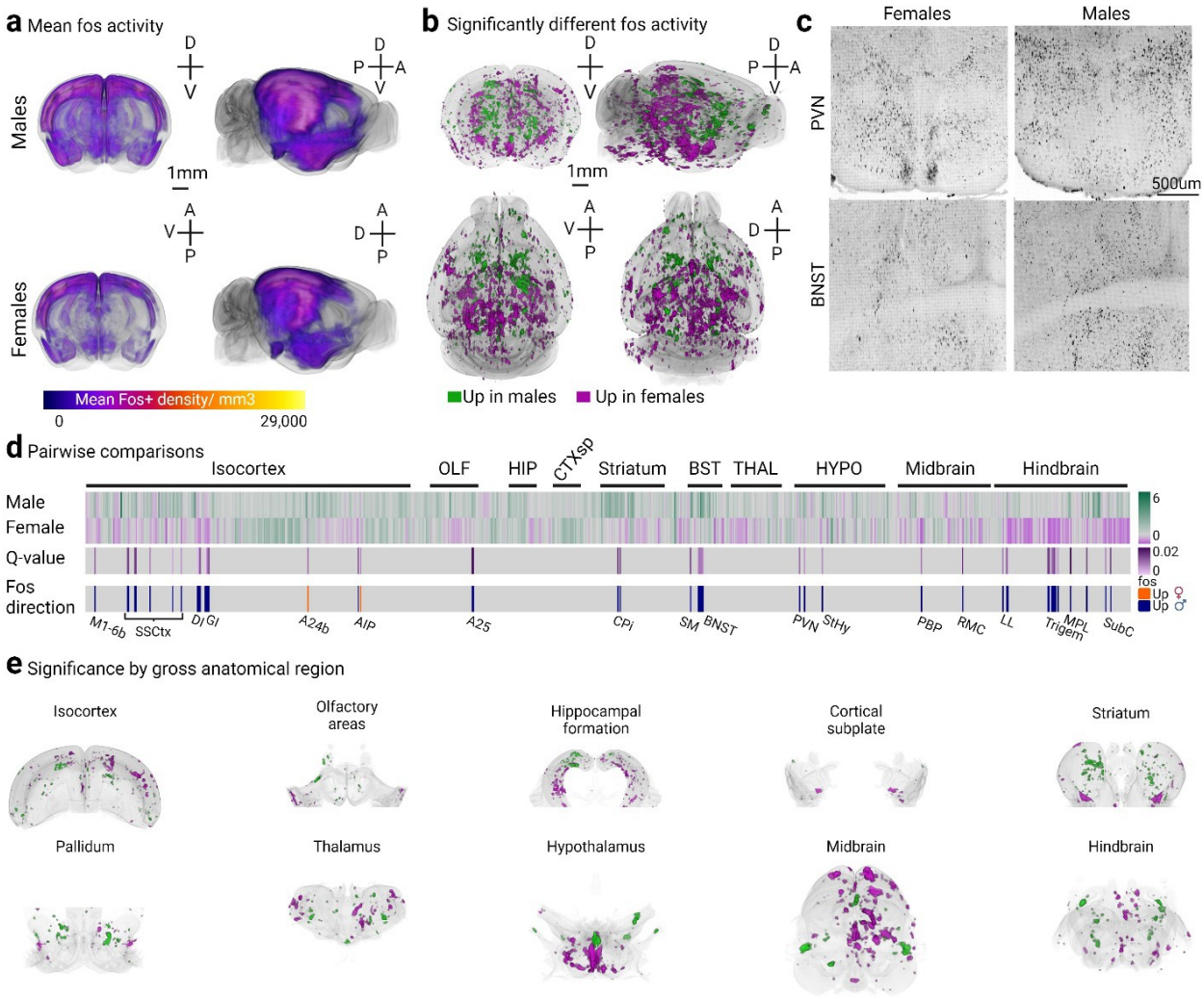
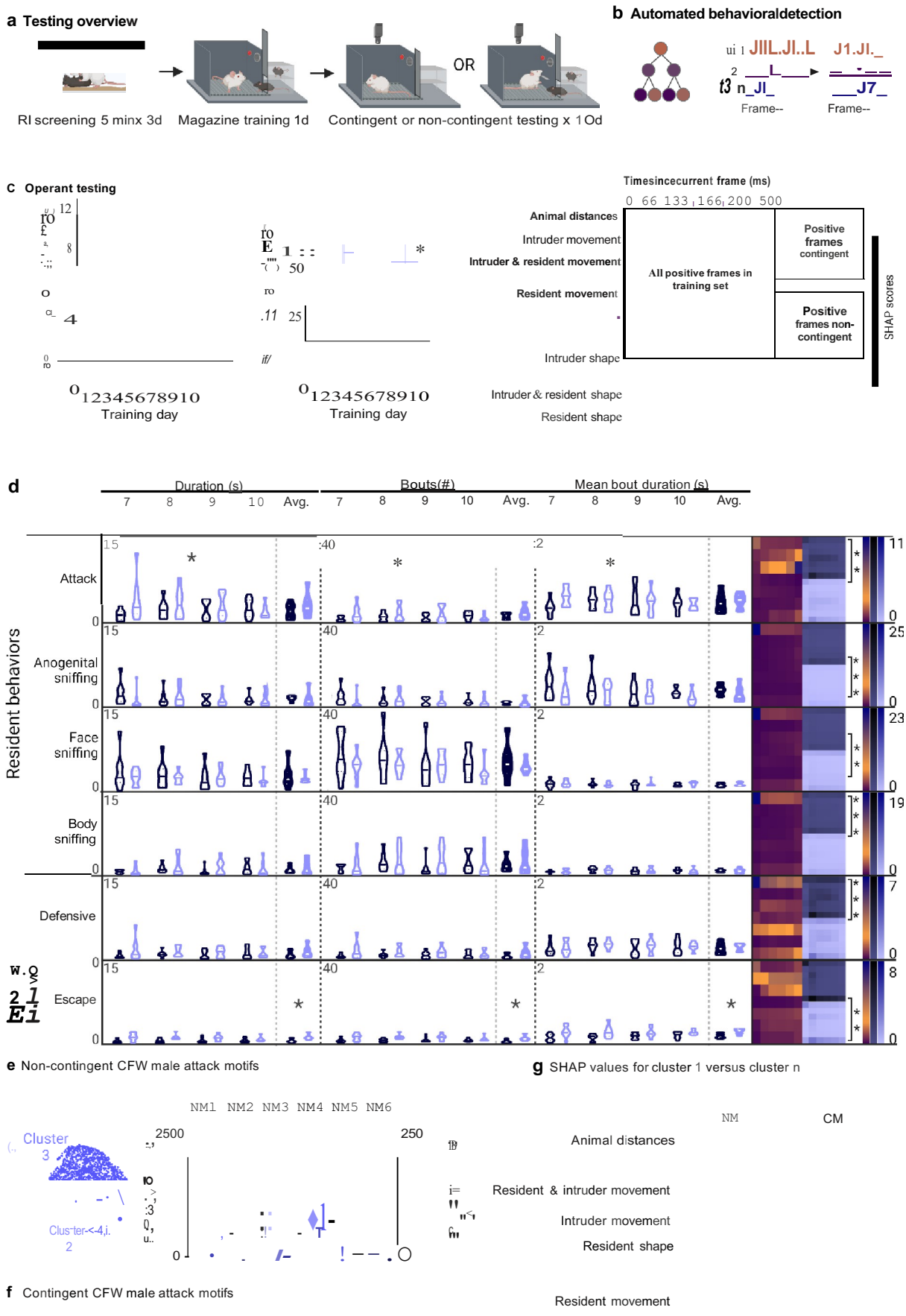
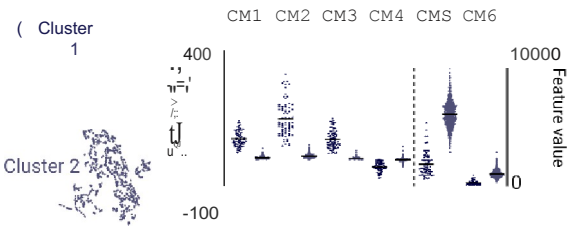


Figure 6.3. Females show a distinct whole-brain activity network inhibiting appetitive aggression behavior. *(a) Mean fos activity in whole brains of males (top) and females (bottom). (b) Significantly higher c-fos density in males (green) or females (pink). (c) Representative images of c-fos activity in the periventricular nucleus of the hypothalamus (top) and bed nucleus of the stria terminalus (bottom) in females (left) and males (right). (d) Pairwise comparisons of male and female c-fos density throughout the brain with a 2% false discovery rate correction. Mean c-fos density heatmaps by region for males and females (top), with significantly discoveries in purple (middle), and directionality indicated in orange (females) or blue (males). (e) p values within 10 anatomical regions throughout the brain.*


a Testing overview
RI screening 5 minx 3d
Magazine training 1d
OR
Contingent or non-contingent testing x 1Od
b Automated behavioraldetection
Frame--
Frame--
c Operant testing
Training day
Training day
Timesincecurrent frame (ms)
0 66 133 166 200 500
Animal distances
Intruder movement
Intruder & resident movement
Resident movement
Intruder shape
Intruder & resident shape
Resident shape
All positive frames in training set
Positive frames contingent
Positive frames non-contingent
SHAP scores
d
Duration (s)
Bouts(#)
Mean bout duration (s)
7 8 9 10 Avg.
Resident behaviors
Attack
Anogenital sniffing
Face sniffing
Body sniffing
Defensive
Escape
e Non-contingent CFW male attack motifs
NM1 NM2 NM3 NM4 NM5 NM6
Cluster 3
2500
250
f Contingent CFW male attack motifs
g SHAP values for cluster 1 versus cluster n
NM
CM
Animal distances
Resident & intruder movement
Intruder movement
Resident shape
Resident movement


(Cluster
1
CM1 CM2 CM3 CM4 CMS CM6
400
10000
Feature value
Cluster 2
-100
0

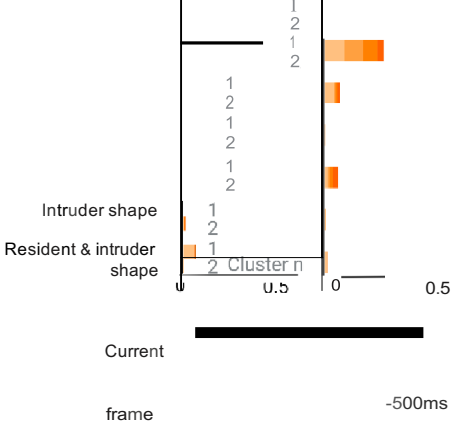

Intruder shape
Resident & intruder
shape
Cluster n
0.5
0
0.5
Current
frame
-500ms

Figure 6.4. Contingent and non-contingent operant administration behavior is grossly similar in males. *(a) Testing timeline overview. (b) Overview of automated behavioral detection and machine learning explainability metrics with Simple Behavioral Analysis (SimBA). (c) There was no significant difference between groups in the number of attack trials per day of training, but the proportion of attack trials per total rewarded trials decreased across training for both groups. (d) Non-contingently administered intruders led to higher attack duration, number of bouts, and mean bout duration than did contingently administered intruders. Sniffing and defensive behavior showed no significant differences between groups, but intruder escape behavior was slightly higher in the non-contingent groups. SHAP values (right) indicate that there are significant differences in the ways in contingently administering versus non-contingent animals present these behaviors. Stars indicate the most significant feature bins between sexes. All SHAP values and statistical comparisons are listed in Statistical Summary Appendix A. (e) Unsupervised analysis uncovered three attack clusters in non-contingent groups, (f) with only two attack clusters in contingent males. (g) SHAP analysis indicates that clusters rely heavily in resident and intruder movement, but that non-contingent male attacks are also defined by resident shape.*

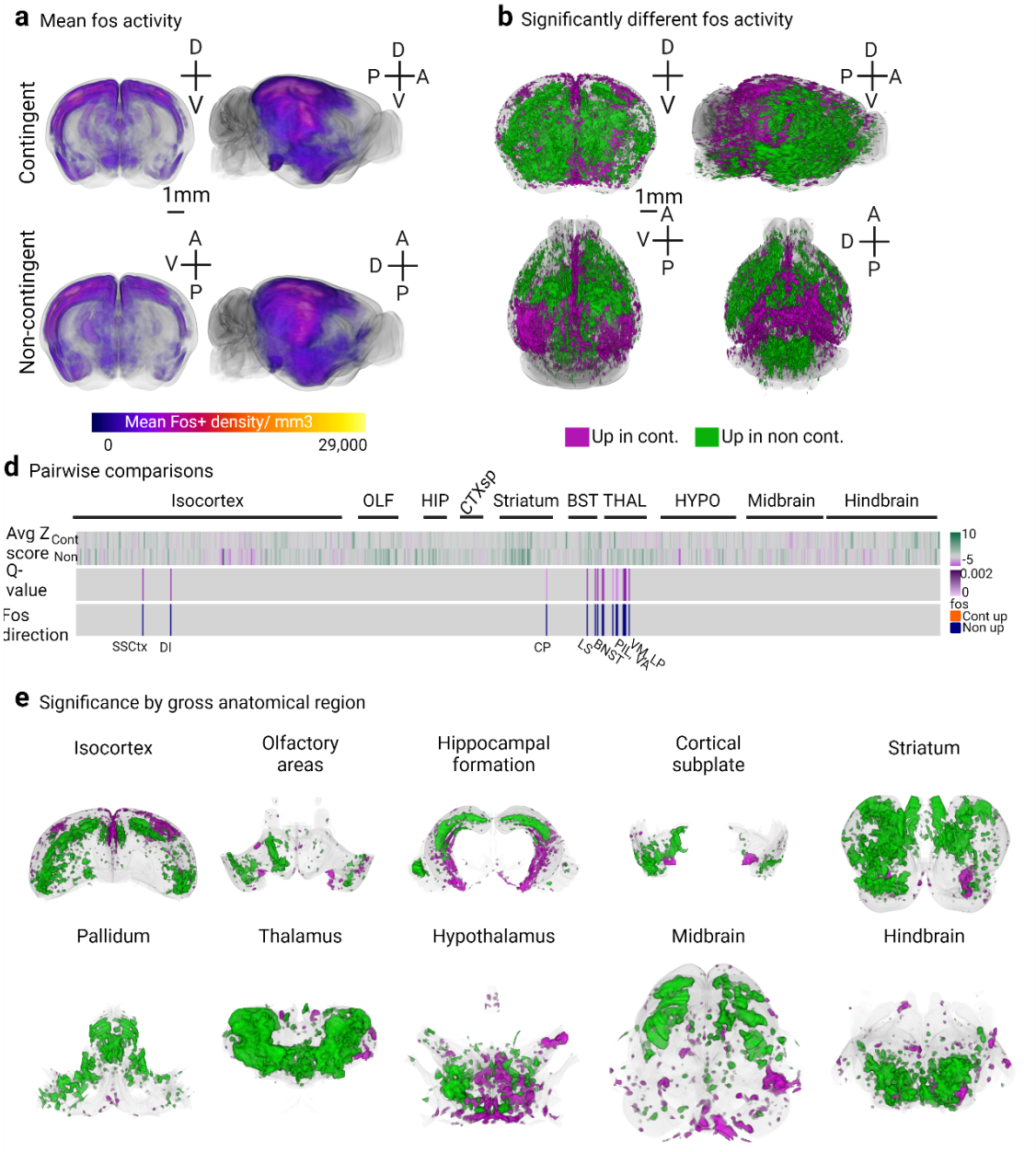


Figure 6.5. Contingent and non-contingent aggression are neurally separable*. (a) Mean fos activity in whole brains of contingent (top) and non-contingent males (bottom). (b) Significantly higher c-fos density in contingent (pink) or non-contingent males (green). (d) Pairwise comparisons of contingent and non-contingent c-fos density throughout the brain with a 2% false discovery rate correction. Mean c-fos density heatmaps by region (top), with significantly discoveries in purple (middle), and directionality indicated in orange (contingent) or blue (non-contingent). (e) p values within 10 anatomical regions throughout the brain.*

Chapter 7 : Discussion and future directions

Keeping it simple: a SimBA retrospective

I started my Ph.D. with zero machine learning experience, when DeepLabCut was still command line-based, and before SLEAP had been released. It has been incredible to be a part of the open-source community and to see how the field of computational neuroethology has grown and evolved over the last six years. A major early hurdle for these programs was increasing accessibility for individuals without computational or coding experience. As of 2020, despite a decade of calls for increased training, only ~15% of neuroscience graduate programs in the US required computational coding courses [33–37]. In my 2020 review (Chapter 2), I outlined several tenets for open-source programs including generalizability, ease of use, cost management, accuracy, easy expansion of training sets, and interpretability. Specific examples of these principles include providing graphical user interfaces rather than command line functionality to circumvent field-wide deficits in computational training and relying on non-specialized and inexpensive equipment. The majority of widely adopted packages today, including SimBA, adhere to these principles. In my next two publications (Chapters 2-3), interpretability and explainability of machine learning results come to the forefront.

Following our release of the SimBA preprint in 2020, more than a dozen additional open-source automated behavioral programs were released, all with relative strengths and weaknesses. As these programs continued to emerge, differences as fundamental as 2D versus 3D pose estimation and supervised versus unsupervised learning techniques arose. The use of different algorithms is not a negative - certain algorithms work better for different types of datasets and supervised versus unsupervised learning can answer different questions – but it does present challenges in working toward standardized behavioral definitions and methodology which can impact rigor and reproducibility. As one solution, our lab has championed the use of post-hoc SHAP explainability scores. Machine learning explainability techniques explain why an algorithm is giving a particular result, given the input data. While inherently explainable methods are always preferable, SHAP is useful as a post-hoc method specifically because it can be used across the many methods emerging within the field. Furthermore, SHAP scores are additive and

can be binned into biologically relevant categories which aids in interpretability, as discussed in Chapter 3. Most importantly, we posited that SHAP scores could be used to objectively quantify behavioral classifiers, moving them from subjective one-line descriptions in a methods section to a standardizable, shareable, and reportable resource.

Using SimBA supervised classification and SHAP explainability scores, we delve into two biological datasets in the peer-reviewed version of the SimBA publication (Chapter 4). Using these tools, I was able to analyze millions of frames of video for five behaviors across two mice within a matter of hours. We used SHAP scores to examine differences in attack classifiers created by expert labs around the country, between mice and rats, and to analyze differences in the ways that a) males and females present behaviors, and b) the ways in which testing environment influences behavioral displays in males. I further extended the use of SHAP to quantifying unsupervised clusters in Chapter 5. While unsupervised algorithms are posited as avoiding anthropomorphic biases and have the potential to identify previously unobserved behavioral clusters, researchers often describe their clustering results by watching several cluster videos and describing the differences they see, injecting subjectivity into the process. SHAP evaluation of unsupervised learning clusters allows for the objective quantification of features defining these clusters, and differences between them.

Challenges and future directions in machine learning

Of course, a major concern within the field is whether or not these classifiers perform to the level of humans, or at least to an acceptable speed/performance tradeoff. In Chapters 3-4 we speak fairly extensively about data cleanliness and the potential issues with standard machine learning performance metrics. As discussed in Chapters 3-4, most machine learning papers present F1 harmonic means as measures of supervised classifier performance, but may not cleanly separate test and training datasets, thereby inflating performance. In essence, a low F1 score means a classifier is trained poorly, but a classifier with a high F1 score may still not be generalizable to new data. This is a standing issue in the field, and both authors and reviewers of machine learning pipelines should work to understand test and training datasets to maximize data cleanliness and therefore the generalizability of models. Hand versus machine scoring

comparisons are still the gold standard, but are infrequently included in the publications for new packages.

As with SimBA, many of these projects begin as pre- or post-doctoral projects which often only lend enough time to develop a package and publish it before the authors need to move on to other projects, either for other stages of training or more lucrative opportunities. This often leaves PIs in the position of trying to hire on technicians or post-docs who are willing to work as developers on a package that has already been published, or to try to find computer scientists willing to work within academic pay scales to continue to upkeep the project. The field is reaching a turning point, and needs to begin to allotting appropriate funding to sustaining and further developing these packages.

One of our goals when we first released SimBA was to see repositories of classifiers that could be downloaded and shared across labs, and there have been parallel initiatives over the years providing open source annotated behavioral videos and classifiers including MABe (Sun et al. 2021a), the OpenBehavior project (2016b), the Jackson Laboratory Mouse Phenome Database, and the CRIM13 dataset (Xavier P. Burgos-Artizzu et al. 2021). There are some free storage solutions available, such as the Open Science Framework maintained by the non-profit Center for Open Science, but public projects are limited to 50GB of storage, and many of these datasets with associated videos can be multiple terabytes and are difficult to host online. As behavioral neuroscience catapults into the realm of big data, these storage issues are impacting both individual labs and the field at large, with few easy solutions.

I am most excited about real-time closed-loop applications of machine learning, which are rapidly becoming possible with current methods. DeepLabCut published DLC-Live in 2020 with less than 15ms delays (Kane et al. 2020), and SLEAP was published in 2022 with latencies less than 3.5ms (Pereira et al. 2022). This postural data can be used directly or can further be fed into behavioral prediction models with simple feature sets to rapidly classify ongoing behavior which in turn triggers optogenetic stimulation. Most current publications doing real-time closed loop experiments use postural data alone, but it is likely that behavioral data will be included in these studies within the next few years.

There is no excuse not to study female reactive aggression

A major theme in my dissertation work has been the use of outbred mice for aggression research. I touch on this directly in Chapter 2, section 2B, citing research indicating that outbred mice show more aggression than inbred lines. Most aggression research in female mice has used outbred females (discussed in Chapter 6), and we reproduced these results in Chapter 5, finding that a large proportion of female CFW mice are reactively aggressive in resident intruder tests. In Chapter 4, when analyzing attack classifiers from expert labs across the country with SHAP values, we find that inbred versus outbred male mice express aggression differently, with classifiers from inbred mice relying more on resident shape than any of the three classifiers from outbred mice. These results highlight the need to use outbred mice in order to study reactive aggression in females, and the need to carefully consider strain differences when comparing historical literature using only males. The common assertion that 'female mice aren't aggressive outside of maternal aggression' is nonsense, as indicated by the literature (White et al. 1969; Gray 1979; Ayer and Whitsett 1980; Ferrari et al. 1996; Silva et al. 2010; Hashikawa et al. 2017; Newman et al. 2019; Aubry et al. 2022).

While female mice are reactively aggressive, we have unfortunately not found a procedure or strain of mouse in which a reasonable number of females will display appetitive aggression. In Chapters 5 and 6, less than 10% of females showed any appetitive aggression, and did so at very low rates compared to males. This makes it unreasonable to study female appetitive aggression in this strain and context. Future directions should experiment with strains as well as housing and testing contexts to see if a lack of appetitive aggression is widely conserved in female mice, or if there are procedures that could be used to preclinically model female appetitive aggression.

Future experimental directions

Digging into my whole-brain c-fos datasets has been a delight. The sex finding of the anterior cingulate cortex involvement in females will be an excellent project for a future graduate student. For my final paper (Chapter 6), we will be including LS fiber photometry data (currently being

collected) as well as LS DREADD data looking at inhibition or excitation of GABAergic populations. I expect to see that inhibiting the LS will lead to increased reactive aggression due to the release of inhibition on the VMHvl, but we could see opposite or no effect in appetitive aggression if the LS is in fact involved in reactive but not appetitive aggression. Fiber photometry data will help us understand if different patterns of activity within the LS are driving these different aggression phenotypes as well.

Future directions at large include investigating female reactive aggression which has long been neglected, and working to further establish preclinical models of appetitive aggression. This is the first direct comparison of reactive and appetitive aggression, and they are clearly influenced by different regions (Fig. 6.5). Network analysis, currently being performed at Columbia by Michelle Jin, will help us understand these networks and identify co-activated regions to interrogate. Since there is little evidence that the VMHvl is involved in the elicitation of aggression in humans, and imaging studies have rather implicated upstream nuclei (Haller et al. 2006; Fanning et al. 2017), these datasets are a resource for the field in providing networks of other regions to begin to explore. I am excited to see what the next several years bring for this line of research.

References

1. Coccaro, E. F., Lee, R. J. & Kavoussi, R. J. A double-blind, randomized, placebo-controlled trial of fluoxetine in patients with intermittent explosive disorder. *J Clin Psychiatry* **70**, 653–662 (2009).
2. Carlson, G. A., Potegal, M., Margulies, D., Basile, J. & Gutkovich, Z. Liquid risperidone in the treatment of rages in psychiatrically hospitalized children with possible bipolar disorder. *Bipolar Disorders* **12**, 205–212 (2010).
3. Frogley, C., Taylor, D., Dickens, G. & Picchioni, M. A systematic review of the evidence of clozapine's anti-aggressive effects. *Int J Neuropsychopharmacol* **15**, 1351–1371 (2012).
4. Khushu, A. & Powney, M. J. Haloperidol for long-term aggression in psychosis. *Cochrane Database Syst Rev* **2016**, CD009830 (2016).
5. Pereira, T. D. *et al.* SLEAP: A deep learning system for multi-animal pose tracking. *Nat Methods* 1–10 (2022) doi:10.1038/s41592-022-01426-1.
6. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**, 1281–1289 (2018).
7. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1**, 206–215 (2019).
8. Zhu, Z. *et al.* A substantia innominata-midbrain circuit controls a general aggressive response. *Neuron* **109**, 1540-1553.e9 (2021).
9. Slotnick, B. M. & McMullen, M. F. Intraspecific fighting in albino mice with septal forebrain lesions. *Physiology & Behavior* **8**, 333–337 (1972).
10. Miley, W. M. & Baenninger, R. Inhibition and facilitation of interspecies aggression in septal lesioned rats. *Physiology & Behavior* **9**, 379–384 (1972).
11. Stark, P. & Henderson, J. K. Central cholinergic suppression of hyper-reactivity and aggression in septal-lesioned rats. *Neuropharmacology* **11**, 839–847 (1972).
12. Albert, D. J. & Richmond, S. E. Septal hyperreactivity: A comparison of lesions within and adjacent to the septum. *Physiology & Behavior* **15**, 339–347 (1975).
13. Lau, P. & Miczek, K. Differential effects of septal lesions on attack and defensive-submissive reactions during intraspecies aggression in rats☆. *Physiology & Behavior* **18**, 479–485 (1977).
14. Caroline Blanchard, D., Blanchard, R. J., Takahashi, L. K. & Takahashi, T. Septal lesions and aggressive behavior. *Behavioral Biology* **21**, 157–161 (1977).
15. Wallace, T. & Thorne, B. M. The effect of lesions in the septal region on muricide, irritability, and activity in the Long-Evans rat. *Psychobiology* **6**, 36–42 (1978).
16. Blanchard, D. C., Blanchard, R. J., Lee, E. M. & Nakamura, S. Defensive behaviors in rats following septal and septal-amygdala lesions. *Journal of Comparative and Physiological Psychology* **93**, 378–390 (1979).
17. Potegal, M., Blau, A. & Glusman, M. Effects of anteroventral septal lesions on intraspecific aggression in male hamsters. *Physiology & Behavior* **26**, 407–412 (1981).
18. Kishore, K. R. & Desiraju, T. INHIBITION OF POSITIVELY REWARDING BEHAVIOR BY THE HEIGHTENED AGGRESSIVE STATE EVOKED EITHER BY PAIN-INDUCING STIMULUS OR SEPTAL LESION. (1990).
19. Leroy, F. *et al.* A circuit from hippocampal CA2 to lateral septum disinhibits social aggression. *Nature* **564**, (2018).
20. Wang, L. *et al.* TMEM16B Calcium-Activated Chloride Channels Regulate Action

Potential Firing in Lateral Septum and Aggression in Male Mice. *J. Neurosci.* **39**, 7102–7117 (2019).
21. Wong, L. C. *et al.* Effective Modulation of Male Aggression through Lateral Septum to Medial Hypothalamus Projection. *Current Biology* **26**, 593–604 (2016).
22. Vega-Quiroga, I., Yarur, H. E. & Gysling, K. Lateral septum stimulation disinhibits dopaminergic neurons in the antero-ventral region of the ventral tegmental area: Role of GABA-A alpha 1 receptors. *Neuropharmacology* **128**, 76–85 (2018).
23. Soden, M. E. *et al.* Anatomic resolution of neurotransmitter-specific projections to the VTA reveals diversity of GABAergic inputs. *Nat Neurosci* **23**, 968–980 (2020).
24. Wang, D. *et al.* GABAergic Neurons in the Dorsal–Intermediate Lateral Septum Regulate Sleep–Wakefulness and Anesthesia in Mice. *Anesthesiology* **135**, 463–481 (2021).
25. Yu, Q. *et al.* Dopamine and serotonin signaling during two sensitive developmental periods differentially impact adult aggressive and affective behaviors in mice. *Mol Psychiatry* **19**, 688–698 (2014).
26. Golden, S. A. *et al.* Basal forebrain projections to the lateral habenula modulate aggression reward. *Nature* **534**, 688–692 (2016).
27. Putkonen, P. T. S. Attack elicited by forebrain and hypothalamic stimulation in the chicken. *Experientia* **22**, 405–407 (1966).
28. Chi, C. C. & Flynn, J. P. Neural Pathways Associated with Hypothalamically Elicited Attack Behavior in Cats. *Science* **171**, 703–706 (1971).
29. Lipp, H. P. & Hunsperger, R. W. Threat, Attack and Flight Elicited by Electrical Stimulation of the Ventromedial Hypothalamus of the Marmoset Monkey Callithrix jacchus; pp. 276–293. *Brain Behavior and Evolution* **15**, 276–293 (1978).
30. Kruk, M. R., Poel, A. M. V. D. & Vos-Frerichs, T. P. D. The Induction of Aggressive Behaviour By Electrical Stimulation in the Hypothalamus of Male Rats. *Behaviour* **70**, 292–322 (1979).
31. Kruk, M. R. *et al.* Discriminant analysis of the localization of aggression-inducing electrode placements in the hypothalamus of male rats. *Brain Research* **260**, 61–79 (1983).
32. Kruk, M. R., Laan, C., Poel, A., Van Erp, A. & Meelis, W. Strain differences in attack patterns elicited by electrical stimulation in the hypothalamus of male CPBWEzob and CPBWI rats. *Aggressive Behavior - AGGRESS BEHAV* **16**, 177–190 (1990).
33. Albert, D. J., Nanji, N., Brayley, K. N. & Madryga, F. J. Hyperreactivity as well as mouse killing is induced by electrical stimulation of the lateral hypothalamus in the rat. *Behav Neural Biol* **27**, 59–71 (1979).
34. Fuchs, S. A. G., Dalsass, M., Siegel, H. E. & Siegel, A. The neural pathways mediating quiet-biting attack behavior from the hypothalamus in the cat: A functional autoradiographic study. *Aggressive Behavior* **7**, 51–67 (1981).
35. Fuchs, S. A., Edinger, H. M. & Siegel, A. The organization of the hypothalamic pathways mediating affective defense behavior in the cat. *Brain Res* **330**, 77–92 (1985).
36. Siegel, A. & Pott, C. B. Neural substrates of aggression and flight in the cat. *Progress in Neurobiology* **31**, 261–283 (1988).
37. Lammers, J. H. C. M., Kruk, M. R., Meelis, W. & van der Poel, A. M. Hypothalamic substrates for brain stimulation-induced attack, teeth-chattering and social grooming in the rat. *Brain Research* **449**, 311–327 (1988).
38. Siegel, A., Roeling, T. A. P., Gregg, T. R. & Kruk, M. R. Neuropharmacology of brain-stimulation-evoked aggression. *Neuroscience & Biobehavioral Reviews* **23**, 359–389 (1999).

39. Lin, D. *et al.* Functional identification of an aggression locus in the mouse hypothalamus. *Nature* **470**, 221–226 (2011).
40. Hashikawa, K., Hashikawa, Y., Lischinsky, J. & Lin, D. The Neural Mechanisms of Sexually Dimorphic Aggressive Behaviors. *Trends Genet.* **34**, 755–776 (2018).
41. Yang, C. F. *et al.* Sexually Dimorphic Neurons in the Ventromedial Hypothalamus Govern Mating in Both Sexes and Aggression in Males. *Cell* **153**, 896–909 (2013).
42. Yang, T. *et al.* Social Control of Hypothalamus-Mediated Male Aggression. *Neuron* **95**, 955-970.e4 (2017).
43. Lee, H. *et al.* Scalable control of mounting and attack by Esr1+ neurons in the ventromedial hypothalamus. *Nature* **509**, 627–632 (2014).
44. Hashikawa, K. *et al.* Esr1+ cells in the ventromedial hypothalamus control female aggression. *Nat. Neurosci.* **20**, 1580–1590 (2017).
45. Yang, T. *et al.* Hypothalamic neurons that mirror aggression. *Cell* **186**, 1195-1211.e19 (2023).
46. Chang, C.-H. & Gean, P.-W. The Ventral Hippocampus Controls Stress-Provoked Impulsive Aggression through the Ventromedial Hypothalamus in Post-Weaning Social Isolation Mice. *Cell Reports* **28**, 1195-1205.e3 (2019).
47. Falkner, A. L. *et al.* Hierarchical Representations of Aggression in a Hypothalamic-Midbrain Circuit. *Neuron* **106**, 637-648.e6 (2020).
48. Haller, J., Tóth, M., Halasz, J. & De Boer, S. F. Patterns of violent aggression-induced brain c-fos expression in male mice selected for aggressiveness. *Physiology & Behavior* **88**, 173–182 (2006).
49. Fanning, J. R., Keedy, S., Berman, M. E., Lee, R. & Coccaro, E. F. Neural Correlates of Aggressive Behavior in Real Time: a Review of fMRI Studies of Laboratory Reactive Aggression. *Curr Behav Neurosci Rep* **4**, 138–150 (2017).
50. Numan, M. & Callahan, E. C. The connections of the medial preoptic region and maternal behavior in the rat. *Physiol Behav* **25**, 653–665 (1980).
51. Numan, M. & Smith, H. G. Maternal behavior in rats: Evidence for the involvement of preoptic projections to the ventral tegmental area. *Behavioral Neuroscience* **98**, 712–727 (1984).
52. Hammond, M. A. & Rowe, F. A. Medial preoptic and anterior hypothalamic lesions: Influences on aggressive behavior in female hamsters. *Physiology & Behavior* **17**, 507–513 (1976).
53. Bermond, B. Effects of medial preoptic hypothalamus anterior lesions on three kinds of behavior in the rat: Intermale aggressive, male-sexual, and mouse-killing behavior. *Aggressive Behavior* **8**, 335–354 (1982).
54. Albert, D. J., Walsh, M. L., Gorzalka, B. B., Mendelson, S. & Zalys, C. Intermale social aggression: Suppression by medial preoptic area lesions. *Physiology & Behavior* **38**, 169–173 (1986).
55. Harmon, A. C., Huhman, K. L., Moore, T. O. & Albers, H. E. Oxytocin Inhibits Aggression in Female Syrian Hamsters. *Journal of Neuroendocrinology* **14**, 963–969 (2002).
56. Gobrogge, K. L., Liu, Y., Jia, X. & Wang, Z. Anterior hypothalamic neural activation and neurochemical associations with aggression in pair-bonded male prairie voles. *Journal of Comparative Neurology* **502**, 1109–1122 (2007).
57. Wei, D. *et al.* A hypothalamic pathway that suppresses aggression toward superior opponents. *Nat Neurosci* **26**, 774–787 (2023).
58. McHenry, J. A. *et al.* The role of ΔfosB in the medial preoptic area: Differential effects

of mating and cocaine history. *Behavioral Neuroscience* **130**, 469–478 (2016).
59. Matsumoto, M. & Hikosaka, O. Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* **447**, 1111–1115 (2007).
60. Kawai, T., Yamada, H., Sato, N., Takada, M. & Matsumoto, M. Roles of the Lateral Habenula and Anterior Cingulate Cortex in Negative Outcome Monitoring and Behavioral Adjustment in Nonhuman Primates. *Neuron* **88**, 792–804 (2015).
61. Li, J. *et al.* The convergence of aversion and reward signals in individual neurons of the mice lateral habenula. *Experimental Neurology* **339**, 113637 (2021).
62. Flanigan, M. E. *et al.* Orexin signaling in GABAergic lateral habenula neurons modulates aggressive behavior in male mice. *Nat. Neurosci.* **23**, 638–650 (2020).
63. Takahashi, A. *et al.* Lateral habenula glutamatergic neurons projecting to the dorsal raphe nucleus promote aggressive arousal in mice. *Nat Commun* **13**, 4039 (2022).
64. Gan, G. *et al.* Habenula-prefrontal resting-state connectivity in reactive aggressive men – A pilot study. *Neuropharmacology* **156**, (2019).
65. Ginsburg, B. & Allee, W. C. Some Effects of Conditioning on Social Dominance and Subordination in Inbred Strains of Mice. *Physiological Zoology* **15**, 485–506 (1942).
66. Oyegbile, T. O. & Marler, C. A. Winning fights elevates testosterone levels in California mice and enhances future ability to win fights. *Horm Behav* **48**, 259–267 (2005).
67. Oyegbile, T. O. & Marler, C. A. Weak winner effect in a less aggressive mammal: correlations with corticosterone but not testosterone. *Physiol. Behav.* **89**, 171–179 (2006).
68. Kudryavtseva, N. N., Smagin, D. A. & Bondar, N. P. Modeling fighting deprivation effect in mouse repeated aggression paradigm. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **35**, 1472–1478 (2011).
69. Yu, Q. *et al.* Optogenetic stimulation of DAergic VTA neurons increases aggression. *Mol Psychiatry* **19**, 635–635 (2014).
70. Mahadevia, D. *et al.* Dopamine promotes aggression in mice via ventral tegmental area to lateral septum projections. *Nat Commun* **12**, 6796 (2021).
71. Gil, M., Nguyen, N.-T., McDonald, M. & Albers, H. E. Social reward: interactions with social status, social communication, aggression, and associated neural activation in the ventral tegmental area. *European Journal of Neuroscience* **38**, 2308–2318 (2013).
72. Filipenko, M. L., Alekseyenko, O. V., Beilina, A. G., Kamynina, T. P. & Kudryavtseva, N. N. Increase of tyrosine hydroxylase and dopamine transporter mRNA levels in ventral tegmental area of male mice under influence of repeated aggression experience. *Molecular Brain Research* **96**, 77–81 (2001).
73. Arnt, J. & Scheel-Krüger, J. GABA in the ventral tegmental area: Differential regional effects on locomotion, aggression and food intake after microinjection of GABA agonists and antagonists. *Life Sciences* **25**, 1351–1360 (1979).
74. Park, H. R. *et al.* Nucleus accumbens deep brain stimulation for a patient with self-injurious behavior and autism spectrum disorder: functional and structural changes of the brain: report of a case and review of literature. *Acta Neurochir* **159**, 137–143 (2017).
75. Harat, M., Kiec, M., Rudaś, M., Birski, M. & Furtak, J. Treating Aggression and Self-destructive Behaviors by Stimulating the Nucleus Accumbens: A Case Series. *Frontiers in Neurology* **12**, (2021).
76. Couppis, M. H. & Kennedy, C. H. The rewarding effect of aggression is reduced by nucleus accumbens dopamine receptor antagonism in mice. *Psychopharmacology* **197**, 449–456 (2008).

77. Golden, S. A. *et al.* Nucleus Accumbens Drd1-Expressing Neurons Control Aggression Self-Administration and Aggression Seeking in Mice. *J. Neurosci.* **39**, 2482–2496 (2019).
78. Aleyasin, H. *et al.* Cell-Type-Specific Role of ΔFosB in Nucleus Accumbens In Modulating Intermale Aggression. *J. Neurosci.* **38**, 5913–5924 (2018).
79. Staffend, N. A. & Meisel, R. L. Aggressive experience increases dendritic spine density within the nucleus accumbens core in female syrian hamsters. *Neuroscience* **227**, 163–169 (2012).
80. Borland, J. M. *et al.* Effect of Aggressive Experience in Female Syrian Hamsters on Glutamate Receptor Expression in the Nucleus Accumbens. *Front. Behav. Neurosci.* **14**, (2020).
81. Erp, A. M. M. van & Miczek, K. A. Aggressive Behavior, Increased Accumbal Dopamine, and Decreased Cortical Serotonin in Rats. *J. Neurosci.* **20**, 9320–9325 (2000).
82. Blair, R. J. R. The Neurobiology of Impulsive Aggression. *J Child Adolesc Psychopharmacol* **26**, 4–9 (2016).
83. Chester, D. S. & DeWall, C. N. The pleasure of revenge: retaliatory aggression arises from a neural imbalance toward reward. *Soc Cogn Affect Neurosci* **11**, 1173–1182 (2016).
84. Tyrer, P. *et al.* Violence and aggression: shortterm management in mental health, health and community settings. (2015).
85. Fazel, S. *et al.* Depression and violence: a Swedish population study. *The Lancet Psychiatry* **2**, 224–232 (2015).
86. Sinha, R. New Findings on Biological Factors Predicting Addiction Relapse Vulnerability. *Curr Psychiatry Rep* **13**, 398–405 (2011).
87. Ducrose, M. R., Alexia D. Cooper, & Howard N. Snyder. *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010*. (2014).
88. Miczek, K. A., Takahashi, A., Gobrogge, K. L., Hwa, L. S. & de Almeida, R. M. M. Escalated Aggression in Animal Models: Shedding New Light on Mesocorticolimbic Circuits. *Curr Opin Behav Sci* **3**, 90–95 (2015).
89. Miczek, K. A., DeBold, J. F., Gobrogge, K., Newman, E. L. & Almeida, R. M. M. de. The Role of Neurotransmitters in Violence and Aggression. in *The Wiley Handbook of Violence and Aggression* 1–13 (American Cancer Society, 2017). doi:10.1002/9781119057574.whbva019.
90. Golden, S. A., Jin, M. & Shaham, Y. Animal Models of (or for) Aggression Reward, Addiction, and Relapse: Behavior and Circuits. *J. Neurosci.* **39**, 3996–4008 (2019).
91. Covington, H. E., Newman, E. L., Leonard, M. Z. & Miczek, K. A. Translational models of adaptive and excessive fighting: an emerging role for neural circuits in pathological aggression. *F1000Res* **8**, (2019).
92. Flanigan, M. E. & Russo, S. J. Recent advances in the study of aggression. *Neuropsychopharmacol* **44**, 241–244 (2019).
93. Thompson, T. I. Visual Reinforcement in Siamese Fighting Fish. *Science* **141**, 55–57 (1963).
94. Anderson, D. J. & Perona, P. Toward a Science of Computational Ethology. *Neuron* **84**, 18–31 (2014).
95. Egnor, S. E. R. & Branson, K. Computational Analysis of Behavior. *Annu. Rev. Neurosci.* **39**, 217–236 (2016).
96. Robie, A. A., Seagraves, K. M., Egnor, S. E. R. & Branson, K. Machine vision methods for analyzing social interactions. *J Exp Biol* **220**, 25–34 (2017).
97. Gris, K. V., Coutu, J.-P. & Gris, D. Supervised and Unsupervised Learning Technology in the Study of Rodent Behavior. *Front Behav Neurosci* **11**, (2017).

98. Brown, A. E. X. & de Bivort, B. Ethology as a physical science. *Nature Physics* **14**, 653–657 (2018).
99. Akay, A. & Hess, H. Deep Learning: Current and Emerging Applications in Medicine and Technology. *IEEE J. Biomed. Health Inform.* **23**, 906–920 (2019).
100. Mathis, M. W. & Mathis, A. Deep learning tools for the measurement of animal behavior in neuroscience. *arXiv:1909.13868 [cs, q-bio]* (2019) doi:10.1016/j.conb.2019.10.008.
101. Datta, S. R., Anderson, D. J., Branson, K., Perona, P. & Leifer, A. Computational Neuroethology: A Call to Action. *Neuron* **104**, 11–24 (2019).
102. Vogelstein, J. T. *et al.* Discovery of Brainwide Neural-Behavioral Maps via Multiscale Unsupervised Structure Learning. *Science* **344**, 386–392 (2014).
103. Wiltschko, A. B. *et al.* Mapping Sub-Second Structure in Mouse Behavior. *Neuron* **88**, 1121–1135 (2015).
104. Rudolf, J., Dondorp, D., Canon, L., Tieo, S. & Chatzigeorgiou, M. Automated behavioural analysis reveals the basic behavioural repertoire of the urochordate Ciona intestinalis. *Sci Rep* **9**, 2416 (2019).
105. Nilsson, S. R. O. *et al.* Simple Behavioral Analysis (SimBA): an open source toolkit for computer classification of complex social behaviors in experimental animals. Preprint at https://doi.org/10.1101/2020.04.19.049452 (2020).
106. Golden, S. A. & Shaham, Y. Aggression Addiction and Relapse: A New Frontier in Psychiatry. *Neuropsychopharmacol.* **43**, 224–225 (2018).
107. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. *arXiv:1703.06870 [cs]* (2017).
108. Graving, J. M., Chae, D., Naik, H., Li, L. & Costelloe, B. R. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. 39 (2019).
109. Pereira, T. D. *et al.* Fast animal pose estimation using deep neural networks. *Nat Methods* **16**, 117–125 (2019).
110. Thompson, T. & Sturm, T. VISUAL-REINFORCER COLOR, AND OPERANT BEHAVIOR IN SIAMESE FIGHTING FISH. *J Exp Anal Behav* **8**, 341–344 (1965).
111. Craft, B. B., Velkey, A. J. & Szalda-Petree, A. Instrumental conditioning of choice behavior in male Siamese fighting fish (Betta splendens). *Behav. Processes* **63**, 171–175 (2003).
112. Craft, B. B., Szalda-Petree, A. D., Brinegar, J. L. & Haddad, N. F. Effect of various discriminative stimuli on choice behavior in male siamese fighting fish (Betta splendens). *Percept Mot Skills* **104**, 575–580 (2007).
113. Elcoro, M., Silva, S. P. & Lattal, K. A. Visual reinforcement in the female Siamese fighting fish, Betta splendens. *J Exp Anal Behav* **90**, 53–60 (2008).
114. Cole, J. M. & Parker, B. K. Schedule-induced aggression: Access to an attackable target bird as a positive reinforcer. *Psychon Sci* **22**, 33–35 (1971).
115. Thompson, T. I. Visual reinforcement in fighting cocks. *J Exp Anal Behav* **7**, 45–49 (1964).
116. Vlautin, C. & Ferkin, M. The Outcome of a Previous Social Interaction with a Same-sex Conspecific Affects the Behavior of Meadow Voles, Microtus pennsylvanicus. *ETHOLOGY* **119**, 212–220 (2013).
117. Momohara, Y., Minami, H., Kanai, A. & Nagayama, T. Role of cAMP signalling in winner and loser effects in crayfish agonistic encounters. *Eur. J. Neurosci.* **44**, 1886–1895 (2016).
118. Okada, K., Okada, Y., Dall, S. R. X. & Hosken, D. J. Loser-effect duration evolves

independently of fighting ability. *Proc. Biol. Sci.* **286**, 20190582 (2019).
119. Trannoy, S., Chowdhury, B. & Kravitz, E. A. Handling alters aggression and 'loser' effect formation in Drosophila melanogaster. *Learn. Mem.* **22**, 64–68 (2015).
120. Kim, Y.-K. *et al.* Repetitive aggressive encounters generate a long-lasting internal state in Drosophila melanogaster males. *PNAS* **115**, 1099–1104 (2018).
121. Kou, R., Hsu, C.-C., Chen, S.-C., Chang, P.-Y. & Fang, S. Winner and loser effects in lobster cockroach contests for social dominance. *Horm Behav* **107**, 49–60 (2019).
122. Schwartzer, J. J., Ricci, L. A. & Melloni, R. H. Prior fighting experience increases aggression in Syrian hamsters: implications for a role of dopamine in the winner effect. *Aggress Behav* **39**, 290–300 (2013).
123. Fuxjager, M. J., Montgomery, J. L. & Marler, C. A. Species differences in the winner effect disappear in response to post-victory testosterone manipulations. *Proc Biol Sci* **278**, 3497–3503 (2011).
124. Lagerspetz, K. Studies on the aggressive behaviour of mice. *Annales Academiae Scientiarum Fennicae* **Series B, 131**, 1–131 (1964).
125. Tellegen, A. & Horn, J. M. Primary aggressive motivation in three inbred strains of mice. *J Comp Physiol Psychol* **78**, 297–304 (1972).
126. Legrand, R. Reinforcing effect of aggressive behaviors preparatory to fighting in mice. *Bull. Psychon. Soc.* **11**, 359–362 (1978).
127. Legrand, R. Successful aggression as the reinforcer for runway behavior of mice. *Psychon Sci* **20**, 303–305 (1970).
128. Potegal, M. The reinforcing value of several types of aggressive behavior: A review. *Aggressive Behavior* **5**, 353–373 (1979).
129. Taylor, G. T. Reinforcement and intraspecific aggressive behavior. *Behavioral and Neural Biology* **27**, 1–24 (1979).
130. 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).
131. Brain, P. F., Benton, D., Childs, G. & Parmigiani, S. The effect of the type of opponent in tests of murine aggression. *Behavioural Processes* **6**, 319–327 (1981).
132. Kudryavtseva, N. N., Bakshtanovskaya, I. V. & Koryakina, L. A. Social model of depression in mice of C57BL/6J strain. *Pharmacol. Biochem. Behav.* **38**, 315–320 (1991).
133. Kudryavtseva, N. N., Smagin, D. A., Kovalenko, I. L. & Vishnivetskaya, G. B. Repeated positive fighting experience in male inbred mice. *Nat Protoc* **9**, 2705–2717 (2014).
134. Jones, S. E. & Brain, P. F. Performances of inbred and outbred laboratory mice in putative tests of aggression. *Behav Genet* **17**, 87–96 (1987).
135. Banerjee, U. An inquiry into the genesis of aggression in mice induced by isolation. *Behaviour* **40**, 86–99 (1971).
136. Chia, R., Achilli, F., Festing, M. F. W. & Fisher, E. M. C. The origins and uses of mouse outbred stocks. *Nat Genet* **37**, 1181–1186 (2005).
137. Tuttle, A. H., Philip, V. M., Chesler, E. J. & Mogil, J. S. Comparing phenotypic variation between inbred and outbred mice. *Nat Methods* **15**, 994–996 (2018).
138. Golden, S. A. *et al.* Compulsive Addiction-like Aggressive Behavior in Mice. *Biological Psychiatry* **82**, 239–248 (2017).
139. Golden, S. A. *et al.* Persistent conditioned place preference to aggression experience in adult male sexually-experienced CD-1 mice: Persistent aggression conditioned place preference

in CD-1 mice. *Genes, Brain and Behavior* **16**, 44–55 (2017).
140. Beach, H. D. Effect of morphine on the exploratory drive. *Can J Psychol* **11**, 237–244 (1957).
141. Mucha, R. F., van der Kooy, D., O'Shaughnessy, M. & Bucenieks, P. Drug reinforcement studied by the use of place conditioning in rat. *Brain Res.* **243**, 91–105 (1982).
142. Bardo, M. t. & Bevins, R. a. Conditioned place preference: what does it add to our preclinical understanding of drug reward? *Psychopharmacology* **153**, 31 (2000).
143. Panksepp, J. B. & Lahvis, G. P. Social reward among juvenile mice. *Genes, Brain and Behavior* **6**, 661–671 (2007).
144. Dölen, G., Darvishzadeh, A., Huang, K. W. & Malenka, R. C. Social reward requires coordinated activity of nucleus accumbens oxytocin and serotonin. *Nature* **501**, 179–184 (2013).
145. Goodwin, N. L., Lopez, S. A., Lee, N. S. & Beery, A. K. Comparative role of reward in long-term peer and mate relationships in voles. *Horm Behav* (2018) doi:10.1016/j.yhbeh.2018.10.012.
146. Meisel, R. L. & Joppa, M. A. Conditioned place preference in female hamsters following aggressive or sexual encounters. *Physiol. Behav.* **56**, 1115–1118 (1994).
147. Martínez, M., Guillén-Salazar, F., Salvador, A. & Simón, V. M. Successful intermale aggression and conditioned place preference in mice. *Physiol. Behav.* **58**, 323–328 (1995).
148. Miczek, K. A., Thompson, M. L. & Shuster, L. Opioid-like analgesia in defeated mice. *Science* **215**, 1520–1522 (1982).
149. Golden, S. A., Covington, H. E., Berton, O. & Russo, S. J. A standardized protocol for repeated social defeat stress in mice. *Nat Protoc* **6**, 1183–1191 (2011).
150. Fish, E. W., De Bold, J. F. & Miczek, K. A. Aggressive behavior as a reinforcer in mice: activation by allopregnanolone. *Psychopharmacology* **163**, 459–466 (2002).
151. Fish, E. W., DeBold, J. F. & Miczek, K. A. Escalated aggression as a reward: corticosterone and GABAA receptor positive modulators in mice. *Psychopharmacology* **182**, 116–127 (2005).
152. Bannai, M., Fish, E. W., Faccidomo, S. & Miczek, K. A. Anti-aggressive effects of agonists at 5-HT1B receptors in the dorsal raphe nucleus of mice. *Psychopharmacology (Berl.)* **193**, 295–304 (2007).
153. Covington, H. E. *et al.* The Urge to Fight: Persistent Escalation by Alcohol and Role of NMDA Receptors in Mice. *Front Behav Neurosci* **12**, 206 (2018).
154. May, M. E. & Kennedy, C. H. AGGRESSION AS POSITIVE REINFORCEMENT IN MICE UNDER VARIOUS RATIO- AND TIME-BASED REINFORCEMENT SCHEDULES. *J Exp Anal Behav* **91**, 185–196 (2009).
155. Falkner, A. L., Grosenick, L., Davidson, T. J., Deisseroth, K. & Lin, D. Hypothalamic control of male aggression-seeking behavior. *Nat Neurosci* **19**, 596–604 (2016).
156. Lacourse, E. *et al.* A longitudinal–experimental approach to testing theories of antisocial behavior development. *Dev Psychopathol* **14**, 909–924 (2002).
157. Provencal, N., Booij, L. & Tremblay, R. E. The developmental origins of chronic physical aggression: biological pathways triggered by early life adversity. *Journal of Experimental Biology* **218**, 123–133 (2015).
158. Anthony, J. C., Warner, L. A. & Kessler, R. C. Comparative Epidemiology of Dependence on Tobacco , Alcohol , Controlled Substances , and Inhalants : Basic Findings From the National Comorbidity Survey. in (1994). doi:10.1037//1064-1297.2.3.244.
159. Deroche-Gamonet, V. Evidence for Addiction-like Behavior in the Rat. *Science* **305**,

1014–1017 (2004).
160. Piazza, P. V. & Deroche-Gamonet, V. A multistep general theory of transition to addiction. *Psychopharmacology (Berl.)* **229**, 387–413 (2013).
161. Pickens, C. L. *et al.* Neurobiology of the incubation of drug craving. *Trends Neurosci.* **34**, 411–420 (2011).
162. Caprioli, D. *et al.* Effect of the Novel Positive Allosteric Modulator of Metabotropic Glutamate Receptor 2 AZD8529 on Incubation of Methamphetamine Craving After Prolonged Voluntary Abstinence in a Rat Model. *Biol. Psychiatry* **78**, 463–473 (2015).
163. Krasnova, I. N. *et al.* Incubation of methamphetamine and palatable food craving after punishment-induced abstinence. *Neuropsychopharmacology* **39**, 2008–2016 (2014).
164. Marchant, N. J., Campbell, E. J., Pelloux, Y., Bossert, J. M. & Shaham, Y. Context-induced relapse after extinction versus punishment: similarities and differences. *Psychopharmacology (Berl.)* **236**, 439–448 (2019).
165. O'kelly, L. I. & Steckle, L. C. A Note on Long Enduring Emotional Responses in the Rat. *The Journal of Psychology* **8**, 125–131 (1939).
166. Azrin, N. H., Hutchinson, R. R. & Hake, D. F. Attack, avoidance, and escape reactions to aversive shock. *J Exp Anal Behav* **10**, 131–148 (1967).
167. Vernon, W. & Ulrich, R. Classical conditioning of pain-elicited aggression. *Science* **152**, 668–669 (1966).
168. Ulrich, R., Wolfe, M. & Dulaney, S. Punishment of shock-induced aggression. *J Exp Anal Behav* **12**, 1009–1015 (1969).
169. Azrin, N. H. Punishment of elicited aggression. *J Exp Anal Behav* **14**, 7–10 (1970).
170. Baenninger, R. & Grossman, J. C. Some effects of punishment on pain-elicited aggression. *J Exp Anal Behav* **12**, 1017–1022 (1969).
171. Roberts, C. L. & Blase, K. Elicitation and punishment of intraspecies aggression by the same stimulus. *J Exp Anal Behav* **15**, 193–196 (1971).
172. Frischknecht, H.-R., Siegfried, B. & Waser, P. G. Postaggression footshock inhibits aggressive behavior in dominant but not in isolated mice. *Behavioral and Neural Biology* **44**, 132–138 (1985).
173. Coward, P. *et al.* Controlling signaling with a specifically designed Gi-coupled receptor. *Proc Natl Acad Sci U S A* **95**, 352–357 (1998).
174. Armbruster, B. N., Li, X., Pausch, M. H., Herlitze, S. & Roth, B. L. Evolving the lock to fit the key to create a family of G protein-coupled receptors potently activated by an inert ligand. *Proc Natl Acad Sci U S A* **104**, 5163–5168 (2007).
175. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience* **8**, 1263–1268 (2005).
176. Miczek, K. A., Maxson, S. C., Fish, E. W. & Faccidomo, S. Aggressive behavioral phenotypes in mice. *Behavioural Brain Research* **125**, 167–181 (2001).
177. Han, W. *et al.* Integrated Control of Predatory Hunting by the Central Nucleus of the Amygdala. *Cell* **168**, 311-324.e18 (2017).
178. Stagkourakis, S. *et al.* A neural network for intermale aggression to establish social hierarchy. *Nature Neuroscience* **21**, 834–842 (2018).
179. Takahashi, A. & Miczek, K. A. Neurogenetics of aggressive behavior: studies in rodents. *Curr Top Behav Neurosci* **17**, 3–44 (2014).
180. Newman, E. L. *et al.* A Role for Prefrontal Cortical NMDA Receptors in Murine

Alcohol-Heightened Aggression. *Neuropsychopharmacology* **43**, 1224–1234 (2018).
181. Schaefer, A. T. & Claridge-Chang, A. The surveillance state of behavioral automation. *Current Opinion in Neurobiology* **22**, 170–176 (2012).
182. Zhang, W. & Yartsev, M. M. Correlated Neural Activity across the Brains of Socially Interacting Bats. *Cell* **178**, 413-428.e22 (2019).
183. Klibaite, U., Berman, G. J., Cande, J., Stern, D. L. & Shaevitz, J. W. An unsupervised method for quantifying the behavior of paired animals. *Phys. Biol.* **14**, 015006 (2017).
184. Dolensek, N., Gehrlach, D. A., Klein, A. S. & Gogolla, N. Facial expressions of emotion states and their neuronal correlates in mice. *Science* **368**, 89–94 (2020).
185. Hsu, A. I. & Yttri, E. A. *B-SOiD: An Open Source Unsupervised Algorithm for Discovery of Spontaneous Behaviors*. http://biorxiv.org/lookup/doi/10.1101/770271 (2019) doi:10.1101/770271.
186. Todd, J. G., Kain, J. S. & Bivort, B. L. de. Systematic exploration of unsupervised methods for mapping behavior. *Phys. Biol.* **14**, 015002 (2017).
187. Branson, K. Distinguishing seemingly indistinguishable animals with computer vision. *Nat Methods* **11**, 721–722 (2014).
188. Berman, G. J., Choi, D. M., Bialek, W. & Shaevitz, J. W. Mapping the stereotyped behaviour of freely moving fruit flies. *J. R. Soc. Interface* **11**, 20140672 (2014).
189. Günel, S. *et al.* DeepFly3D: A deep learning-based approach for 3D limb and appendage tracking in tethered, adult Drosophila. *bioRxiv* 640375 (2019) doi:10.1101/640375.
190. Nath, T. *et al.* Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat Protoc* **14**, 2152–2176 (2019).
191. Hong, W. *et al.* Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proc Natl Acad Sci USA* **112**, E5351–E5360 (2015).
192. Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S. & Branson, K. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nat Methods* **10**, 64–67 (2013).
193. Jhuang, H. *et al.* Automated home-cage behavioural phenotyping of mice. *Nat Commun* **1**, 1–10 (2010).
194. Burgos-Artizzu, X. P., Dollar, P., Dayu Lin, Anderson, D. J. & Perona, P. Social behavior recognition in continuous video. in *2012 IEEE Conference on Computer Vision and Pattern Recognition* 1322–1329 (IEEE, Providence, RI, 2012). doi:10.1109/CVPR.2012.6247817.
195. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* 1135–1144 (ACM Press, San Francisco, California, USA, 2016). doi:10.1145/2939672.2939778.
196. Kuehlkamp, A., Becker, B. & Bowyer, K. Gender-From-Iris or Gender-From-Mascara? *arXiv:1702.01304 [cs]* (2017).
197. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data & Society* **3**, 205395171667967 (2016).
198. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
199. Hand, D. J. Classifier Technology and the Illusion of Progress. *Statist. Sci.* **21**, 1–14 (2006).
200. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

201. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R news* **2**, 6 (2002).
202. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *PNAS* **79**, 2554–2558 (1982).
203. LeCun, Y. *et al.* Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* **1**, 541–551 (1989).
204. Lietman, T., Eng, J., Katz, J. & Quigley, H. A. Neural networks for visual field analysis: how do they compare with other algorithms? *J Glaucoma* **8**, 77–80 (1999).
205. Nitze, I., Schulthess, U. & Asche, H. COMPARISON OF MACHINE LEARNING ALGORITHMS RANDOM FOREST, ARTIFICIAL NEURAL NETWORK AND SUPPORT VECTOR MACHINE TO MAXIMUM LIKELIHOOD FOR SUPERVISED CROP TYPE CLASSIFICATION. 7 (2012).
206. Liu, M., Wang, M., Wang, J. & Li, D. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. *Sensors and Actuators B: Chemical* **177**, 970–980 (2013).
207. Karpathy, A. *et al.* Large-Scale Video Classification with Convolutional Neural Networks. in *2014 IEEE Conference on Computer Vision and Pattern Recognition* 1725–1732 (IEEE, Columbus, OH, USA, 2014). doi:10.1109/CVPR.2014.223.
208. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
209. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002).
210. Batuwita, R. & Palade, V. Class Imbalance Learning Methods for Support Vector Machines. in *Imbalanced Learning* (eds. He, H. & Ma, Y.) 83–99 (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2013). doi:10.1002/9781118646106.ch5.
211. Freund, Y. & Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* **55**, 119–139 (1997).
212. Friedman, J., Hastie, T. & Tibshirani, R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Ann. Statist.* **28**, 337–407 (2000).
213. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **29**, 1189–1232 (2001).
214. Cortes, C. & Vapnik, V. Support-vector networks. *Mach Learn* **20**, 273–297 (1995).
215. Xu, L. Robust Support Vector Machine Training via Convex Outlier Ablation. 7 (2006).
216. Chih-Wei Hsu & Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* **13**, 415–425 (2002).
217. Caruana, R. & Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. in *Proceedings of the 23rd international conference on Machine learning - ICML '06* 161–168 (ACM Press, Pittsburgh, Pennsylvania, 2006). doi:10.1145/1143844.1143865.
218. Caruana, R., Karampatziakis, N. & Yessenalina, A. An empirical evaluation of supervised learning in high dimensions. in *Proceedings of the 25th international conference on Machine learning - ICML '08* 96–103 (ACM Press, Helsinki, Finland, 2008). doi:10.1145/1390156.1390169.
219. Rabiner, L. R. & Juang, B. H. An Introduction to Hidden Markov Models. 12 (1986).
220. Carola, V., Mirabeau, O. & Gross, C. T. Hidden Markov Model Analysis of Maternal

Behavior Patterns in Inbred and Reciprocal Hybrid Mice. *PLOS ONE* **6**, e14753 (2011).
221. Arakawa, T. *et al.* Automated Estimation of Mouse Social Behaviors Based on a Hidden Markov Model. in *Hidden Markov Models: Methods and Protocols* (eds. Westhead, D. R. & Vijayabaskar, M. S.) 185–197 (Springer New York, New York, NY, 2017). doi:10.1007/978-1-4939-6753-7_14.
222. Haccou, P. *et al.* Markov models for social interactions: analysis of electrical stimulation in the hypothalamic aggression area of rats. *Animal Behaviour* **36**, 1145–1163 (1988).
223. Natarajan, D., de Vries, H., Saaltink, D.-J., de Boer, S. F. & Koolhaas, J. M. Delineation of violence from functional aggression in mice: an ethological approach. *Behav. Genet.* **39**, 73–90 (2009).
224. Allahverdyan, A. & Galstyan, A. On Maximum a Posteriori Estimation of Hidden Markov Processes. *arXiv:0906.1980 [cond-mat, physics:physics, stat]* (2009).
225. Thanos, P. K., Restif, C., O'Rourke, J. R., Lam, C. Y. & Metaxas, D. Mouse Social Interaction Test (MoST): a quantitative computer automated analysis of behavior. *J Neural Transm* **124**, 3–11 (2017).
226. Rodriguez, A., Zhang, H., Klaminder, J., Brodin, T. & Andersson, M. ToxId: an efficient algorithm to solve occlusions when tracking multiple animals. *Sci Rep* **7**, 1–8 (2017).
227. Romero-Ferrero, F., Bergomi, M. G., Hinz, R. C., Heras, F. J. H. & Polavieja, G. G. de. idtracker.ai: tracking all individuals in small or large collectives of unmarked animals. *Nat Methods* **16**, 179–182 (2019).
228. Brox, T., Rosenhahn, B., Cremers, D. & Seidel, H.-P. High Accuracy Optical Flow Serves 3-D Pose Tracking: Exploiting Contour and Flow Based Constraints. in *Computer Vision – ECCV 2006* (eds. Leonardis, A., Bischof, H. & Pinz, A.) 98–111 (Springer, Berlin, Heidelberg, 2006). doi:10.1007/11744047_8.
229. Cao, Z., Simon, T., Wei, S.-E. & Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv:1611.08050 [cs]* (2017).
230. Pérez-Escudero, A., Vicente-Page, J., Hinz, R. C., Arganda, S. & de Polavieja, G. G. idTracker: tracking individuals in a group by automatic identification of unmarked animals. *Nat Methods* **11**, 743–748 (2014).
231. Weissbrod, A. *et al.* Automated long-term tracking and social behavioural phenotyping of animal colonies within a semi-natural environment. *Nat Commun* **4**, 2018 (2013).
232. de Chaumont, F. *et al.* Real-time analysis of the behaviour of groups of mice via a depth-sensing camera and machine learning. *Nat Biomed Eng* (2019) doi:10.1038/s41551-019-0396-1.
233. Peleh, T., Bai, X., Kas, M. J. H. & Hengerer, B. RFID-supported video tracking for automated analysis of social behaviour in groups of mice. *Journal of Neuroscience Methods* **325**, 108323 (2019).
234. Gomez-Marin, A., Partoune, N., Stephens, G. J. & Louis, M. Automated Tracking of Animal Posture and Movement during Exploration and Sensory Orientation Behaviors. *PLOS ONE* **7**, e41642 (2012).
235. Forys, B., Xiao, D., Gupta, P., Boyd, J. D. & Murphy, T. H. Real-time markerless video tracking of body parts in mice using deep neural networks. *bioRxiv* 482349 (2018) doi:10.1101/482349.
236. Stih, V., Petrucco, L., Kist, A. & Portugues, R. Stytra: An open-source, integrated system for stimulation, tracking and closed-loop behavioral experiments. 19 (2019).
237. Sumner, S. A. *et al.* Violence in the United States: Status, Challenges, and Opportunities. *JAMA* **314**, 478–488 (2015).

238. Martin, L. A., Neighbors, H. W. & Griffith, D. M. The Experience of Symptoms of Depression in Men vs Women: Analysis of the National Comorbidity Survey Replication. *JAMA Psychiatry* **70**, 1100–1106 (2013).
239. Ostinelli, E. G., Brooke-Powney, M. J., Li, X. & Adams, C. E. Haloperidol for psychosis-induced aggression or agitation (rapid tranquillisation). *Cochrane Database Syst Rev* **7**, CD009377–CD009377 (2017).
240. Calver, L., Drinkwater, V., Gupta, R., Page, C. B. & Isbister, G. K. Droperidol v. haloperidol for sedation of aggressive behaviour in acute mental health: Randomised controlled trial. *British Journal of Psychiatry* **206**, 223–228 (2015).
241. Branson, K., Robie, A. A., Bender, J., Perona, P. & Dickinson, M. H. High-throughput ethomics in large groups of Drosophila. *Nat Methods* **6**, 451–457 (2009).
242. Dankert, H., Wang, L., Hoopfer, E. D., Anderson, D. J. & Perona, P. Automated monitoring and analysis of social behavior in Drosophila. *Nat Methods* **6**, 297–303 (2009).
243. de Chaumont, F. *et al.* Computerized video analysis of social interactions in mice. *Nature Methods* **9**, 410–417 (2012).
244. Burgos-Artizzu, X. P., Dollar, P., Dayu Lin, Anderson, D. J. & Perona, P. Social behavior recognition in continuous video. in *2012 IEEE Conference on Computer Vision and Pattern Recognition* 1322–1329 (IEEE, Providence, RI, 2012). doi:10.1109/CVPR.2012.6247817.
245. Matsumoto, J. *et al.* A 3D-Video-Based Computerized Analysis of Social and Sexual Interactions in Rats. *PLOS ONE* **8**, e78460 (2013).
246. Unger, J. *et al.* An unsupervised learning approach for tracking mice in an enclosed area. *BMC Bioinformatics* **18**, (2017).
247. Rodriguez, A. *et al.* ToxTrac: A fast and robust software for tracking organisms. *Methods in Ecology and Evolution* **9**, 460–464 (2018).
248. Machado, A. S., Darmohray, D. M., Fayad, J., Marques, H. G. & Carey, M. R. A quantitative framework for whole-body coordination reveals specific deficits in freely walking ataxic mice. *eLife* **4**, e07892 (2015).
249. Landis, S. C. *et al.* A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490**, 187–191 (2012).
250. Reality check on reproducibility. *Nature* **533**, 437–437 (2016).
251. NIH. *NOT-OD-15-103: Enhancing Reproducibility through Rigor and Transparency*. https://grants.nih.gov/grants/guide/notice-files/not-od-15-103.html (2015).
252. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron* **93**, 480–490 (2017).
253. Gruene, T. M., Flick, K., Stefano, A., Shea, S. D. & Shansky, R. M. Sexually divergent expression of active and passive conditioned fear responses in rats. *Elife* **4**, (2015).
254. Greenberg, G. D. *et al.* Sex differences in stress-induced social withdrawal: role of brain derived neurotrophic factor in the bed nucleus of the stria terminalis. *Front. Behav. Neurosci.* **7**, (2014).
255. Meyer, A. F., O'Keefe, J. & Poort, J. Two Distinct Types of Eye-Head Coupling in Freely Moving Mice. *Current Biology* **30**, 2116-2130.e6 (2020).
256. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**, 1281–1289 (2018).
257. Dunn, T. W. *et al.* Geometric deep learning enables 3D kinematic profiling across species and environments. *Nat Methods* **18**, 564–573 (2021).

258. Karashchuk, P. *et al.* Anipose: A toolkit for robust markerless 3D pose estimation. *Cell Rep* **36**, 109730 (2021).
259. Bohnslav, J. P. *et al.* DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels. *eLife* **10**, e63377 (2021).
260. Mathis, M. W. & Mathis, A. Deep learning tools for the measurement of animal behavior in neuroscience. *arXiv:1909.13868 [cs, q-bio]* (2019).
261. Pereira, T. D., Shaevitz, J. W. & Murthy, M. Quantifying behavior to understand the brain. *Nature Neuroscience* 1–13 (2020) doi:10.1038/s41593-020-00734-z.
262. Goodwin, N. L., Nilsson, S. R. O. & Golden, S. A. Rage Against the Machine: Advancing the study of aggression ethology via machine learning. *Psychopharmacology* (2020) doi:10.1007/s00213-020-05577-x.
263. Das, A. & Rad, P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv:2006.11371 [cs]* (2020).
264. Shahroudnejad, A. A Survey on Understanding, Visualizations, and Explanation of Deep Neural Networks. *arXiv:2102.01792 [cs]* (2021).
265. Lundberg, S. M. *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* **2**, 56–67 (2020).
266. Doshi-Velez, F. & Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. (2017).
267. Vu, M.-A. T. *et al.* A Shared Vision for Machine Learning in Neuroscience. *J. Neurosci.* **38**, 1601–1607 (2018).
268. Markou, A., Chiamulera, C., Geyer, M. A., Tricklebank, M. & Steckler, T. Removing obstacles in neuroscience drug discovery: the future path for animal models. *Neuropsychopharmacology* **34**, 74–89 (2009).
269. Shapley, L.S. Stochastic Games*. **39**, 6 (1953).
270. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]* (2017).
271. Goldman, M. S. & Fee, M. S. Computational training for the next generation of neuroscientists. *Current Opinion in Neurobiology* **46**, 25–30 (2017).
272. Grisham, W., Lom, B., Lanyon, L. & L. Ramos, R. Proposed Training to Meet Challenges of Large-Scale Data in Neuroscience. *Frontiers in Neuroinformatics* **10**, (2016).
273. Pevzner, P. & Shamir, R. Computing Has Changed Biology—Biology Education Must Catch Up. *Science* **325**, 541–542 (2009).
274. Juavinett, A. L. The next generation of neuroscientists needs to learn how to code, and we need new ways to teach them. *Neuron* **110**, 576–578 (2022).
275. Society for Neuroscience. Surveys of Neuroscience Departments and Programs. *Surveys of Neuroscience Departments and Programs* https://www.sfn.org/careers/higher-education-and-training/neuroscience-training-program-survey (2016).
276. Covert, I. C., Lundberg, S. & Lee, S.-I. Explaining by Removing: A Unified Framework for Model Explanation. *Journal of Machine Learning Research* 90 (2021).
277. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **267**, 1–38 (2019).
278. Osborne, M. J. & Rubinstein, A. *A Course in Game Theory*. (MIT Press, Cambridge, Mass, 1994).
279. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802.03888 [cs, stat]* (2019).

280. Lam, S. K., Pitrou, A. & Seibert, S. Numba: a LLVM-based Python JIT compiler. in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15* 1–6 (ACM Press, Austin, Texas, 2015). doi:10.1145/2833157.2833162.
281. McKinney, W. pandas: a Foundational Python Library for Data Analysis and Statistics. 9 (2011).
282. RAPIDS: Collection of Libraries for End to End GPU Data Science. *RAPIDS* https://rapids.ai/index.html (2018).
283. Lundberg, S. slundberg/shap. (2022).
284. Tjandrasuwita, M., Sun, J. J., Kennedy, A., Chaudhuri, S. & Yue, Y. Interpreting Expert Annotation Differences in Animal Behavior. *arXiv:2106.06114 [cs]* (2021).
285. Sun, J. J. *et al.* Task Programming: Learning Data Efficient Behavior Representations. 10 (2021).
286. Pedregosa et al. scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation. https://scikit-learn.org/stable/.
287. Meng, X. *et al.* MLlib: Machine Learning in Apache Spark. 7.
288. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016) doi:10.1145/2939672.2939785.
289. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467 [cs]* (2016).
290. Verma, S., Dickerson, J. & Hines, K. Counterfactual Explanations for Machine Learning: A Review. *arXiv:2010.10596 [cs, stat]* (2020).
291. Lipton, Z. C. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]* (2017).
292. Graving, J. M. & Couzin, I. D. *VAE-SNE: A Deep Generative Model for Simultaneous Dimensionality Reduction and Clustering*. http://biorxiv.org/lookup/doi/10.1101/2020.07.17.207993 (2020) doi:10.1101/2020.07.17.207993.
293. Luxem, K. *et al. Identifying Behavioral Structure from Deep Variational Embeddings of Animal Motion*. http://biorxiv.org/lookup/doi/10.1101/2020.05.14.095430 (2020) doi:10.1101/2020.05.14.095430.
294. Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A. & Sander, J. Density-Based Clustering Validation. in *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)* 839–847 (Society for Industrial and Applied Mathematics, 2014). doi:10.1137/1.9781611973440.96.
295. Lab, S. DeepLabStream. (2022).
296. Kane, G. A., Lopes, G., Saunders, J. L., Mathis, A. & Mathis, M. W. Real-time, low-latency closed-loop feedback using markerless posture tracking. *eLife* **9**, e61909 (2020).
297. ML Kit. *Google Developers* https://developers.google.com/ml-kit.
298. TensorFlow Lite | ML for Mobile and Edge Devices. *TensorFlow* https://www.tensorflow.org/lite.
299. Winters, C. *et al.* Automated procedure to assess pup retrieval in laboratory mice. *Sci Rep* **12**, 1663 (2022).
300. Bandrowski, A. *et al.* The Resource Identification Initiative: A cultural shift in publishing. Preprint at https://doi.org/10.12688/f1000research.6555.2 (2015).
301. Chambers, K. *et al.* Towards minimum reporting standards for life scientists. Preprint at https://doi.org/10.31222/osf.io/9sm4x (2019).

302. White, S. R., Amarante, L. M., Kravitz, A. V. & Laubach, M. The Future Is Open: Open-Source Tools for Behavioral Neuroscience Research. *eNeuro* **6**, ENEURO.0223-19.2019 (2019).
303. Bogue, M. A. *et al.* Mouse Phenome Database: a data repository and analysis suite for curated primary mouse phenotype data. *Nucleic Acids Res* **48**, D716–D723 (2020).
304. Geuther, B. Q. *et al.* Action detection using a neural network elucidates the genetics of mouse grooming behavior. *eLife* **10**, e63207 (2021).
305. Falkner, A. L., Grosenick, L., Davidson, T. J., Deisseroth, K. & Lin, D. Hypothalamic control of male aggression-seeking behavior. *Nat Neurosci* **19**, 596–604 (2016).
306. Ferenczi, E. A. *et al.* Prefrontal cortical regulation of brainwide circuit dynamics and reward-related behavior. *Science* **351**, aac9698–aac9698 (2016).
307. Kim, Y. *et al.* Mapping Social Behavior-Induced Brain Activation at Cellular Resolution in the Mouse. *Cell Reports* **10**, 292–305 (2015).
308. Gunaydin, L. A. *et al.* Natural Neural Projection Dynamics Underlying Social Behavior. *Cell* **157**, 1535–1551 (2014).
309. Graving, J. M. *et al.* DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife* 8:e47994 (2019).
310. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**, 1281–1289 (2018).
311. Pereira, T. D. *et al.* Fast animal pose estimation using deep neural networks. *Nat Methods* **16**, 117–125 (2019).
312. Geuther, B. Q. *et al.* Robust mouse tracking in complex environments using neural networks. *Communications Biology* **2**, 1–11 (2019).
313. Gris, K. V., Coutu, J.-P. & Gris, D. Supervised and Unsupervised Learning Technology in the Study of Rodent Behavior. *Front Behav Neurosci* **11**, (2017).
314. Goodwin, N. L., Nilsson, S. R. O., Choong, J. J. & Golden, S. A. Toward the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience. *Current Opinion in Neurobiology* **73**, 102544 (2022).
315. Newton, K. C., Kacev, D., Nilsson, S. R. O., Golden, S. A. & Sheets, L. Lateral Line Ablation by Toxins Results in Distinct Rheotaxis Profiles in Fish. *bioRxiv* 2021.11.15.468723 (2021) doi:10.1101/2021.11.15.468723.
316. Jernigan, C. M., Stafstrom, J. A., Zaba, N. C., Vogt, C. C. & Sheehan, M. J. Color is necessary for face discrimination in the Northern paper wasp, Polistes fuscatus. *Anim Cogn* (2022) doi:10.1007/s10071-022-01691-9.
317. Dahake, A. *et al.* Floral humidity as a signal – not a cue – in a nocturnal pollination system. *bioRxiv* 2022.04.27.489805 (2022) doi:10.1101/2022.04.27.489805.
318. Dawson, M. *et al.* Sex-dependent role of hypocretin/orexin neurons in social behavior. *bioRxiv* 2022.08.19.504565 (2022) doi:10.1101/2022.08.19.504565.
319. Baleisyte, A., Schneggenburger, R. & Kochubey, O. Stimulation of medial amygdala GABA neurons with kinetically different channelrhodopsins yields opposite behavioral outcomes. *Cell Reports* **39**, 110850 (2022).
320. Cruz-Pereira, J. S. *et al.* Prebiotic supplementation modulates selective effects of stress on behavior and brain metabolome in aged mice. *Neurobiology of Stress* **21**, 100501 (2022).
321. Linders, L. E. *et al.* Stress-driven potentiation of lateral hypothalamic synapses onto ventral tegmental area dopamine neurons causes increased consumption of palatable food. *Nat Commun* **13**, 6898 (2022).
322. Slivicki, R. A. *et al.* Oral oxycodone self-administration leads to features of opioid

misuse in male and female mice. *Addiction Biology* **28**, e13253 (2023).
323. Miczek, K. A. *et al.* Excessive alcohol consumption after exposure to two types of chronic social stress: intermittent episodes vs. continuous exposure in C57BL/6J mice with a history of drinking. *Psychopharmacology* (2022) doi:10.1007/s00213-022-06211-8.
324. Cui, Q. *et al.* Striatal Direct Pathway Targets Npas1+ Pallidal Neurons. *J. Neurosci.* **41**, 3966–3987 (2021).
325. Chen, J. *et al.* A MYT1L syndrome mouse model recapitulates patient phenotypes and reveals altered brain development due to disrupted neuronal maturation. *Neuron* **109**, 3775-3792.e14 (2021).
326. Rigney, N., Zbib, A., de Vries, G. J. & Petrulis, A. Knockdown of sexually differentiated vasopressin expression in the bed nucleus of the stria terminalis reduces social and sexual behaviour in male, but not female, mice. *Journal of Neuroendocrinology* **n/a**, e13083 (2021).
327. Winters, C. *et al.* Automated procedure to assess pup retrieval in laboratory mice. *Sci Rep* **12**, 1663 (2022).
328. Neira, S. *et al.* Chronic Alcohol Consumption Alters Home-Cage Behaviors and Responses to Ethologically Relevant Predator Tasks in Mice. *bioRxiv* (2022) doi:10.1101/2022.02.04.479122.
329. Kwiatkowski, C. C. *et al.* Quantitative standardization of resident mouse behavior for studies of aggression and social defeat. *Neuropsychopharmacology* 1–10 (2021) doi:10.1038/s41386-021-01018-1.
330. Yamaguchi, T. *et al.* Posterior amygdala regulates sexual and aggressive behaviors in male mice. *Nat Neurosci* (2020) doi:10.1038/s41593-020-0675-x.
331. Nygaard, K. R. *et al.* Extensive characterization of a Williams Syndrome murine model shows *Gtf2ird1* -mediated rescue of select sensorimotor tasks, but no effect on enhanced social behavior. Preprint at https://doi.org/10.1101/2023.01.18.523029 (2023).
332. Ojanen, S. *et al.* Interneuronal GluK1 kainate receptors control maturation of GABAergic transmission and network synchrony in the hippocampus. *Molecular Brain* **16**, 43 (2023).
333. Hon, O. J. *et al.* Serotonin modulates an inhibitory input to the central amygdala from the ventral periaqueductal gray. *Neuropsychopharmacol.* **47**, 2194–2204 (2022).
334. Murphy, C. A. *et al. Modeling Features of Addiction with an Oral Oxycodone Self-Administration Paradigm*. http://biorxiv.org/lookup/doi/10.1101/2021.02.08.430180 (2021) doi:10.1101/2021.02.08.430180.
335. Neira, S. *et al.* Impact and role of hypothalamic corticotropin releasing hormone neurons in withdrawal from chronic alcohol consumption in female and male mice. 2023.05.30.542746 Preprint at https://doi.org/10.1101/2023.05.30.542746 (2023).
336. Lapp, H. E., Salazar, M. G. & Champagne, F. A. Automated Maternal Behavior during Early life in Rodents (AMBER) pipeline. 2023.09.15.557946 Preprint at https://doi.org/10.1101/2023.09.15.557946 (2023).
337. Barnard, I. L. *et al.* High-THC Cannabis smoke impairs working memory capacity in spontaneous tests of novelty preference for objects and odors in rats. 2023.04.06.535880 Preprint at https://doi.org/10.1101/2023.04.06.535880 (2023).
338. Ausra, J. *et al.* Wireless battery free fully implantable multimodal recording and neuromodulation tools for songbirds. *Nat Commun* **12**, 1968 (2021).
339. Friard, O. & Gamba, M. BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution* **7**, 1325–1330 (2016).

340. Spink, A. J., Tegelenbosch, R. A. J., Buma, M. O. S. & Noldus, L. P. J. J. The EthoVision video tracking system—A tool for behavioral phenotyping of transgenic mice. *Physiology & Behavior* **73**, 731–744 (2001).
341. Lauer, J. *et al. Multi-Animal Pose Estimation and Tracking with DeepLabCut*. http://biorxiv.org/lookup/doi/10.1101/2021.04.30.442096 (2021) doi:10.1101/2021.04.30.442096.
342. Segalin, C. *et al.* The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *eLife* **10**, e63720 (2021).
343. Goodwin, N. L., Nilsson, S. R. O. & Golden, S. A. Rage Against the Machine: Advancing the study of aggression ethology via machine learning. *Psychopharmacology* **237**, 2569–2588 (2020).
344. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. Preprint at https://doi.org/10.48550/arXiv.1602.04938 (2016).
345. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. in *Proceedings of the 34th International Conference on Machine Learning* 3319–3328 (PMLR, 2017).
346. Hatwell, J., Gaber, M. M. & Azad, R. M. A. CHIRPS: Explaining random forest classification. *Artif Intell Rev* **53**, 5747–5788 (2020).
347. Takahashi, A. *et al.* Establishment of a repeated social defeat stress model in female mice. *Sci Rep* **7**, 12838 (2017).
348. Newman, E. L. *et al.* Fighting Females: Neural and Behavioral Consequences of Social Defeat Stress in Female Mice. *Biological Psychiatry* **86**, 657–668 (2019).
349. Aubry, A. V. *et al.* Sex differences in appetitive and reactive aggression. *Neuropsychopharmacol.* (2022) doi:10.1038/s41386-022-01375-5.
350. Shemesh, Y. & Chen, A. A paradigm shift in translational psychiatry through rodent neuroethology. *Mol Psychiatry* **28**, 993–1003 (2023).
351. Bordes, J. *et al.* Automatically annotated motion tracking identifies a distinct social behavioral profile following chronic social defeat stress. 2022.06.23.497350 Preprint at https://doi.org/10.1101/2022.06.23.497350 (2022).
352. Winters, C., Gorssen, W., Wöhr, M. & D'Hooge, R. BAMBI: A new method for automated assessment of bidirectional early-life interaction between maternal behavior and pup vocalization in mouse dam-pup dyads. *Frontiers in Behavioral Neuroscience* **17**, (2023).
353. Lorbach, M., Poppe, R. & Veltkamp, R. C. Interactive rodent behavior annotation in video using active learning. *Multimed Tools Appl* **78**, 19787–19806 (2019).
354. Schweihoff, J. F., Hsu, A. I., Schwarz, M. K. & Yttri, E. A. A-SOiD, an active learning platform for expert-guided, data efficient discovery of behavior. 2022.11.04.515138 Preprint at https://doi.org/10.1101/2022.11.04.515138 (2022).
355. Whiteway, M. R. *et al.* Partitioning variability in animal behavioral videos using semi-supervised variational autoencoders. *PLOS Computational Biology* **17**, e1009439 (2021).
356. MABe 2022. https://sites.google.com/view/mabe22/home.
357. Sun, J. J. *et al.* The Multi-Agent Behavior Dataset: Mouse Dyadic Social Interactions. Preprint at http://arxiv.org/abs/2104.02710 (2021).
358. About OpenBehavior. *OpenBehavior* https://edspace.american.edu/openbehavior/ (2016).
359. MPD: About the Mouse Phenome Database. https://phenome.jax.org/about.
360. Kapoor, S. & Narayanan, A. Leakage and the Reproducibility Crisis in ML-based Science. Preprint at https://doi.org/10.48550/arXiv.2207.07048 (2022).

361. Karashchuk, P., Tuthill, J. C. & Brunton, B. W. The DANNCE of the rats: a new toolkit for 3D tracking of animal behavior. *Nat Methods* **18**, 460–462 (2021).
362. Branson, K. APT. *GitHub* https://github.com/kristinbranson/APT.
363. Xavier P. Burgos-Artizzu, Piotr Dollar, Dayu Lin, David J. Anderson & Pietro Perona. CRIM13 (Caltech Resident-Intruder Mouse 13). CaltechDATA https://doi.org/10.22002/D1.1892 (2021).
364. Lee, W., Fu, J., Bouwman, N., Farago, P. & Curley, J. P. Temporal microstructure of dyadic social behavior during relationship formation in mice. *PLoS ONE* **14**, e0220596 (2019).
365. Madangopal, R. *et al.* Incubation of palatable food craving is associated with brain-wide neuronal activation in mice. *Proceedings of the National Academy of Sciences* **119**, e2209382119 (2022).
366. Renier, N. *et al.* iDISCO: a simple, rapid method to immunolabel large tissue samples for volume imaging. *Cell* **159**, 896–910 (2014).
367. Renier, N. *et al.* Mapping of Brain Activity by Automated Volume Analysis of Immediate Early Genes. *Cell* **165**, 1789–1802 (2016).
368. Kirst, C. *et al.* Mapping the Fine-Scale Organization and Plasticity of the Brain Vasculature. *Cell* **180**, 780-795.e25 (2020).
369. Goodwin, N. L. *et al.* Simple behavioral analysis (SimBA) as a platform for explainable machine learning in behavioral neuroscience. *Nature Neuroscience* (2024).
370. Kesner, R. P. & Gilbert, P. E. The Role of the Agranular Insular Cortex in Anticipation of Reward Contrast. *Neurobiol Learn Mem* **88**, 82–86 (2007).
371. Burkett, J. P. *et al.* Oxytocin-dependent consolation behavior in rodents. *Science* **351**, 375–378 (2016).
372. Keum, S. *et al.* A Missense Variant at the Nrxn3 Locus Enhances Empathy Fear in the Mouse. *Neuron* **98**, 588-601.e5 (2018).
373. Yamagishi, A., Lee, J. & Sato, N. Oxytocin in the anterior cingulate cortex is involved in helping behaviour. *Behavioural Brain Research* **393**, 112790 (2020).
374. Kim, S.-W. *et al.* Hemispherically lateralized rhythmic oscillations in the cingulate-amygdala circuit drive affective empathy in mice. *Neuron* **111**, 418-429.e4 (2023).
375. Douglas, A. J. Chapter 2.5 - Vasopressin and oxytocin. in *Techniques in the Behavioral and Neural Sciences* (eds. Steckler, T., Kalin, N. H. & Reul, J. M. H. M.) vol. 15 205–229 (Elsevier, 2005).
376. Ferris, C. F. & Potegal, M. Vasopressin receptor blockade in the anterior hypothalamus suppresses aggression in hamsters. *Physiology & Behavior* **44**, 235–239 (1988).
377. Potegal, M. & Ferris, C. F. Intraspecific aggression in male hamsters is inhibited by intrahypothalamic vasopressin-receptor antagonist. *Aggressive Behavior* **15**, 311–320 (1989).
378. Koolhaas, J. M., Van Den Brink, T. H. C., Roozendaal, B. & Boorsma, F. Medial amygdala and aggressive behavior: Interaction between testosterone and vasopressin. *Aggressive Behavior* **16**, 223–229 (1990).
379. Winslow, J. T., Hastings, N., Carter, C. S., Harbaugh, C. R. & Insel, T. R. A role for central vasopressin in pair bonding in monogamous prairie voles. *Nature* **365**, 545–548 (1993).
380. Delville, Y., Mansour, K. M. & Ferris, C. F. Serotonin blocks vasopressin-facilitated offensive aggression: Interactions within the ventrolateral hypothalamus of golden hamsters. *Physiology & Behavior* **59**, 813–816 (1996).
381. Delville, Y., Mansour, K. M. & Ferris, C. F. Testosterone facilitates aggression by modulating vasopressin receptors in the hypothalamus. *Physiology & Behavior* **60**, 25–29

(1996).
382. Beery, A. K. Antisocial oxytocin: complex effects on social behavior. *Current Opinion in Behavioral Sciences* **6**, 174–182 (2015).
383. White, M., Mayo, S. & Edwards, D. A. Fighting in female mice as a function of the size of the opponent. *Psychon Sci* **16**, 14–15 (1969).
384. Hashikawa, Y., Hashikawa, K., Falkner, A. L. & Lin, D. Ventromedial Hypothalamus and the Generation of Aggression. *Front. Syst. Neurosci.* **11**, (2017).
385. Ferrari, P. F., Palanza, P. & Rodgers, R. J. Comparing Different Forms of Male and Female Aggression in Wild and Laboratory Mice: An Ethopharmacological Study. (1996).
386. Silva, A. L., Fry, W. H. D., Sweeney, C. & Trainor, B. C. Effects of photoperiod and experience on aggressive behavior in female California mice. *Behavioural Brain Research* **208**, 528–534 (2010).
387. Gray, L. E. The effects of the reproductive status and prior housing conditions on the aggressiveness of female mice. *Behavioral and Neural Biology* **26**, 508–513 (1979).
388. Ayer, M. L. & Whitsett, J. M. Aggressive behaviour of female prairie deer mice in laboratory populations. *Animal Behaviour* **28**, 763–771 (1980).e