ARBUR: a Machine Learning-based Analysis System for Relating Behaviors and Ultrasonic Vocalizations of Rats

Guanglu Jia[1,2†], Zhe Chen[1,3†], Qijie Zhou[1,2], Yulai Zhang[1,3], Zhenzhen Quan[4], Xuechao Chen[1,2], Toshio Fukuda[1,2], Qiang Huang [1,2*], Qing Shi [1,2*]

[1]Key Laboratory of Biomimetic Robots and Systems, Beijing Institute of Technology, Ministry of Education, Beijing, China.
[2]Intelligent Robotics Institute, School of Mechatronical Engineering, Beijing Institute of Technology, Beijing, China.
[3]School of Medical Technology, Beijing Institute of Technology, Beijing, China.
[4]Key Laboratory of Molecular Medicine and Biotherapy, School of Life Science, Beijing Institute of Technology, Beijing, China.

*Corresponding author(s). E-mail(s): qhuang@bit.edu.cn; shiqing@bit.edu.cn;
†These authors contributed equally to this work.

Abstract

Deciphering how different types of behavior and ultrasonic vocalizations (USVs) of rats interact can yield insights into the neural basis of social interaction. However, the behavior-vocalization interplay of rats remains elusive because of the challenges of relating the two communication media in complex social contexts. Here, we propose a machine learning-based analysis system called ARBUR. ARBUR can cluster without bias both non-step (continuous) and step USVs via three steps, hierarchically detect eight types of behavior of two freely behaving rats with high accuracy, and locate the vocal rat in 3-D space. By simultaneously recording the video and ultrasonic streams of two freely behaving rats, we show that ARBUR can not only automatically reveal the well-understood behavior-associated vocalizations that were carefully concluded by other behavioral researchers, but also hold the promise to indicate novel findings that can be hardly found by manual analysis, especially regarding step USVs and the active/passive rat-associated vocalizations during aggressive/intimate social behaviors. Using ARBUR, we further confirm that rats communicate via distinct USVs when engaging in different types of social behavior. This work could help mechanistically understand the interactive influence between the behavior and USVs of rats.

Keywords: ultrasonic vocalization, rat social behavior, biomimetic robots, animal-robot interaction

1 Introduction

Nonverbal communication and vocalization are two vital means of natural communication during social interaction across the animal kingdom and human society [1, 2]. Nonverbal communication such as facial expression [3], body posture, waggle dance [4], and social play [5] can be received by nearby conspecifics and often analyzed generally by human observers as social behavior [6], whereas vocalization can be communicated remotely to convey information in dark and tortuous environments. Both social behavior and vocalization can be readily recorded experimentally as outer observable states that reflect the inner emotional states [3] or even neural circuit activities [7–9], and therefore are actively investigated in animal behavior research [10]. During social engagement, both the social behavior and vocalization of individuals influence each other in a continuous and interactive manner, contributing to the variable group social dynamics. Deciphering the behavior-vocalization interplay [11, 12] would thus reveal insights into the neural basis of social interaction. For such deciphering efforts, rats are widely used as social animal models because of their innate sociability [13] and proven emotion-related ultrasonic vocalizations (USVs) [14]. For example, the alarm sounds (22-kHz USVs) of trapped rats may induce overwhelming distress in freely moving rats via emotional contagion and further evoke pro-social behavior in the free ones [15]. However, a dedicated system for analyzing social

communications of freely behaving rats is still missing to promote the mechanistic understanding of the interplay between social behavior and USVs.

The barrier is caused by the difficulty to relate the different types of behavior and USVs of rats in complex social contexts (to reveal the underlying behavior-vocalization interplay) because of the following challenges. First of all, a method for the unbiased and automatic clustering of USVs that not only covers both non-step (continuous) and step (non-continuous) signals in spectrograms [12] but also comprehensively incorporates the structure of frequency and duration [16], is still missing. Moreover, the automatic behavior detection of rats under social interaction poses a major challenge in constructing the behavior-specific features [17, 18]; therefore, it is difficult to discriminate different types of easy-to-confuse social behavior [19–22]. In addition, allocating the recorded USVs to the vocal rat also brings an obstacle because of the close proximity and top-view partial overlapping of socially engaged rats [23].

In this Article, we propose a machine learning-based Analysis system for Relating Behaviors and USVs of Rats (named ARBUR). ARBUR automatically clusters ultrasonic syllables into user-defined or automatically calculated subgroups from a comprehensive perspective, detects eight types of behavior (Table 1) based on the lateral camera view, and locates the vocal rat in two-rat scenarios with free social interaction. Using ARBUR, we reveal that the distinct behaviors of rats are associated with different USVs. Therefore, based on the simultaneously recorded raw video and audio streams (see Extended Data Fig. 1), ARBUR can relate the behavior and USVs of rats in freely interacting social contexts for downstream qualitative and quantitative analyses. For example, the pinned rat produces significantly more 22-kHz aversive USVs compared with during other social behavior or solitary state.

Table 1 Behavioral definitions

Behavior	Definition
SO (solitary)	Both rats engage in separate activities.
ST (still)	Both rats keep still.
MA (moving away)	A rat moves away from another rat.
AP (approaching)	A rat moves towards another rat until contact.
FO (following)	A rat follows another rat without contact.
PIN (pinning)	A rat holds down another rat lying on its back.
POU (pouncing)	A rat lies on the back of another rat.
SNC (social nose contact)	A rat touches another rat with the tip of its nose.

2 Results

2.1 Unsupervised clustering of the vocal repertoire

We recorded 43,357 ultrasonic vocalizations (USVs) to create the vocal repertoire. To cluster the USVs in an unbiased and comprehensive way, we present the unsupervised three-step clustering algorithm ARBUR:USV. In the first step, ARBUR:USV divides the USVs into 2 clusters according to their mean frequency-indicated emotional states (aversive 22-kHz USVs (ranging from 15 kHz to 32 kHz) or appetitive 50-kHz USVs (above 32 kHz)). This is a special treatment for rat USVs analysis since rats produce meaningful USVs with frequency near 22 kHz while mice don't. Another special consideration in ARBUR:USV is the inclusion of step signals, which are usually overlooked in other research. Moreover, note that 50-kHz USVs are classified into four groups according to their mean frequency and duration: high-peak-frequency (HPF) 50-kHz USVs and low-peak-frequency (LPF) 50-kHz USVs, because two aggregated areas were observed in the density map (Fig.2b), indicating two kinds of 50-kHz USVs may be present despite of their similar frequency contours (shape in the spectrograms). Therefore, in the second step, the 50-kHz USVs are classified into four groups (groups A-D) according to their continuity (non-step or step) and distribution in frequency-duration space (HPF or LPF) (Fig.2a). In the third step, ARBUR:USV separates the four 50-kHz groups (groups A-D) into subgroups within each group according to their frequency contours, thus allowing users to investigate the correlation between USV contours with other biological factors (e.g. behavior types) in relevant research. In the mean time, the 22-kHz group (group E) is divided into five subgroups based on mean frequency and duration because the frequency contours of 22-kHz USVs do not vary much.

While the first two steps divides the USVs according to their biological relevance (mean frequency, natural distribution in frequency-duration space), the third clustering step separates the USVs within each group according to their contour features, which raises the question of determining the optimal number of clusters to balance over-clustering with under-clustering. Here, ARBUR:USV provides three solutions for users to choose. First, the number of clusters for groups A-E can be arbitrarily chosen according to their experiences. This may be suitable for experienced animal behaviorist to investigate USV-related rat behaviors with fixed cluster size. Second, ARBUR:USV can calculate the optimal number of clusters for each group according to the elbow method (see Methods) [29]. It basically relies on the relationship (curve) of total within-cluster error (TWCE) versus increasing number of clusters. With increasing cluster number, TWCE tends to decrease exponentially. The elbow method finds the elbow point of the

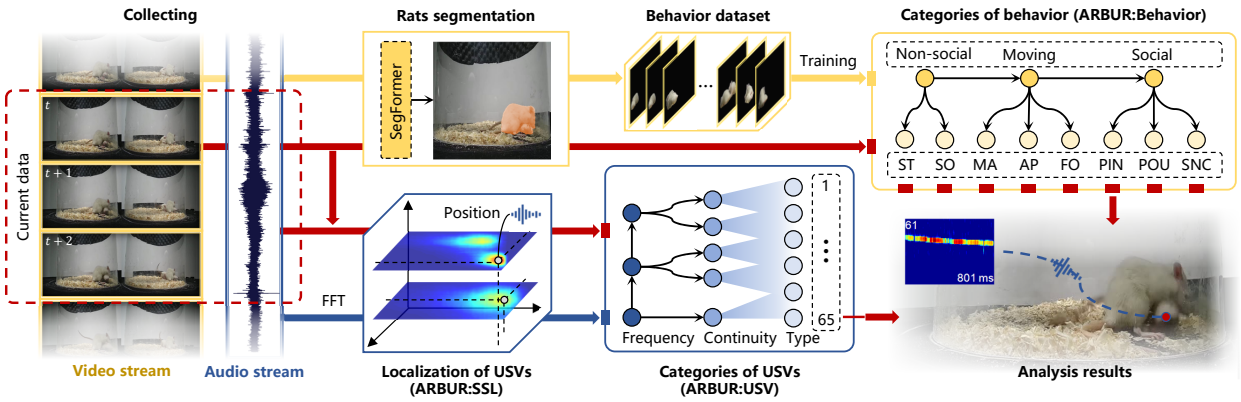


Fig. 1 Illustration of the working flow of ARBUR. ARBUR takes simultaneously recorded video and audio streams as input and outputs the current behavior type, the ultrasonic vocalization (USV) in spectrogram with cluster type and duration indicated, and the labeled position of the vocal rat for each frame of the binocular images in the video frame.

Table 2 ARBUR:USV achieves comprehensive unsupervised clustering of rodent's USVs compared with current research.

	Manner of clustering	Mean frequency	Duration	Non-step contour	Step contour	Choice of optimal clusters (number)	Rodents
Wright et al. [24]	Manual	✓	-	✓	✓	Arbitrary (15)	Rats
Riede et al. [25]	Manual	✓	-	✓	✓	Arbitrary (6)	Rats
Burgdorf et al. [26]	Manual	✓	-	✓	✓	Arbitrary (3)	Rats
Fonseca et al. [27]	Supervised	✓	-	✓	✓	Arbitrary (11)	Mice
Sangiamo et al. [12]	Unsupervised	-	-	✓	-	Progressive (22)	Mice
Takahashi et al. [16]	Unsupervised	✓	✓	-	-	Arbitrary (3)	Rats
MUPET [28]	Unsupervised	-	✓	✓	-	Arbitrary (20-200)	Mice
DeepSqueak [29]	Unsupervised	-	✓	✓	-	Automatic (20)	Mice
AVA [30]	Unsupervised	-	✓	✓	-	Arbitrary (-)	Mice
ARBUR:USV	**Unsupervised**	✓	✓	✓	✓	**Multiple choices (65)**	**Rats**

curve to be the optimal number. For the collected dataset, the optimal numbers for groups A-E are 13, 28, 13, 14 and 8, respectively (see Extended Data Fig. 3). Third, ARBUR:USV also provides the progressive method, which considers the clustering degree for each cluster (see Methods). In particular, the cluster number, starts with two, increases by one if the average innerpoint percent of clustering results does not reach a user-defined threshold. For example, if the threshold is set to be 0.98, the optimal cluster number for group E is 8 (Fig.S1). In short, ARBUR:USV provides three methods for choosing the number of clusters for groups A-E. Here, to balance simpler visualization and the elbow method result, we arbitrarily set the numbers for groups A-E to be 10, 25, 10, 15 and 5, respectively.

Fig.2a shows the clustering examples. Note that the USV clusters were re-sorted in descending order according to contour slope (for clusters 1-60) or duration (for clusters 61-65). It is intuitive that the USVs within each cluster share similar contours, and the mean contours of each cluster differ from those of the others (Fig.S2). We then quantified the clustering results. We show that ARBUR:USV achieves much higher inter-cluster distance compared with intra-cluster variance across the 65 clusters (Fig.2c; Extended Data Fig.2a, b, c). Moreover, we quantified the 65 clusters of the USVs, and show that they vary substantially in terms of duration, frequency range, mean frequency, and signal counts across clusters (Fig.S3).

Rats produce diverse USVs in terms of duration, mean frequency, and frequency contours [24]. Existing studies have only clustered the USVs in a way that considers some of these factors in an unsupervised [12, 16] or manual [14, 25] manner (Table 2). Among them, a recent study leveraged unsupervised learning to cluster mouse non-step USVs into 22 categories based on their contours, which showed potential for unbiased clustering of a broader spectrum of USVs [12]. However, an unsupervised clustering algorithm that comprehensively takes all these three factors and step signals under consideration remains missing. To our knowledge, ARBUR:USV is the first unsupervised clustering method that considers the step USVs of rodents. Although it is designed and optimized to cluster rat USVs, it can be readily generalized for mouse USVs applications. Compared with algorithms from existing studies, the proposed unsupervised clustering algorithm features: 1) the comprehensive consideration of frequency, duration, and contour information; 2) the inclusion of step

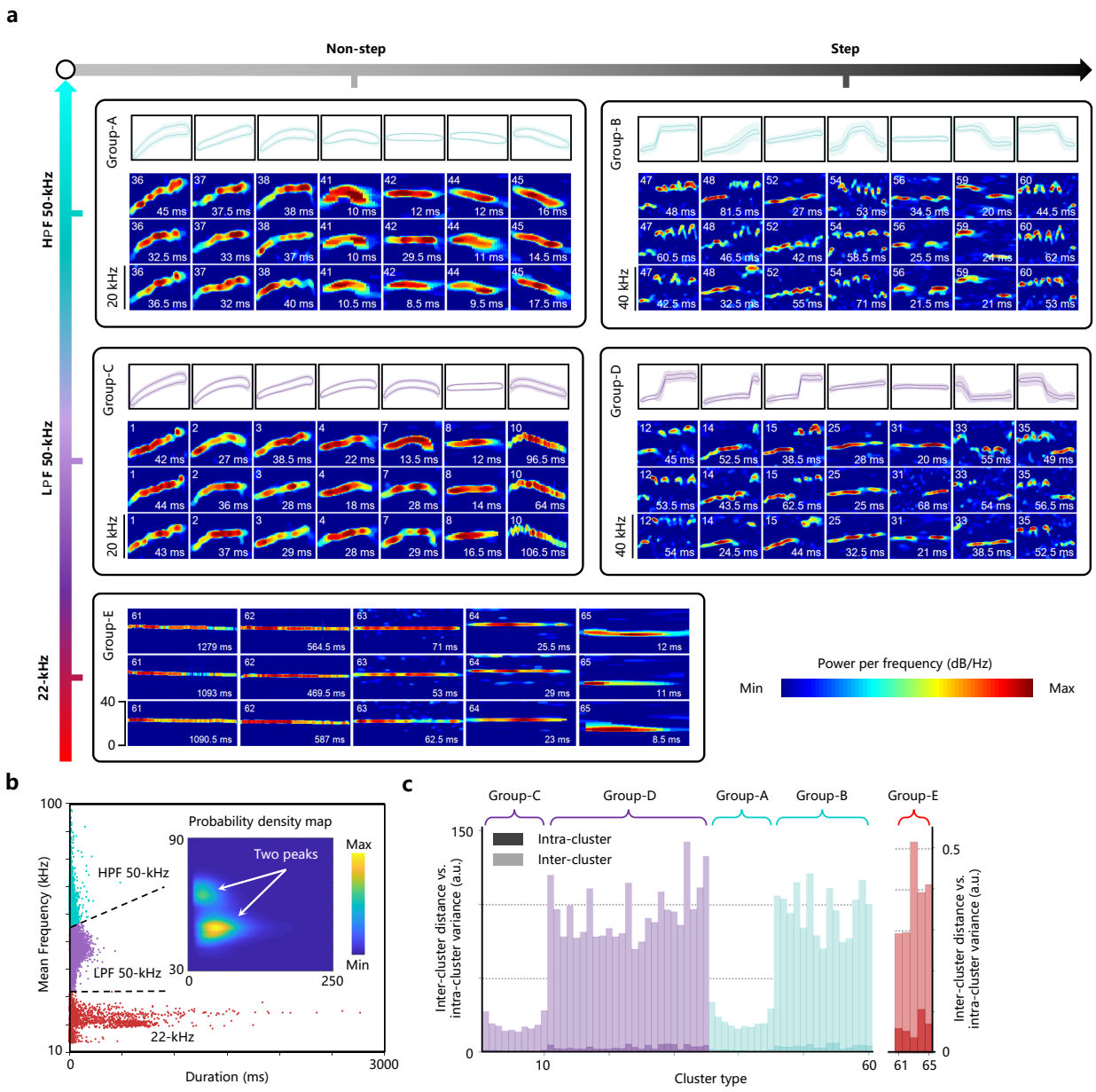


Fig. 2 ARBUR:USV clusters the vocal repertoire in a unbiased and comprehensive manner. **a**, Representative examples show that USVs within the same clusters share similar contours. Groups A-D, top panels, mean value of contour ± s.e.m. Bottom three panels, three selected USVs spectrograms with cluster type (top left) and duration (bottom right) indicated. Each column represents one cluster. Intensity is indicated by color (red, maximum; blue, minimum). Groups A-E represent high-frequency non-step 50-kHz USVs (clusters 36-45), high-frequency step 50-kHz USVs (clusters 46-60), low-frequency non-step 50-kHz USVs (clusters 1-10), low-frequency step 50-kHz USVs (clusters 11-35) and 22-kHz USVs (clusters 61-65), respectively. Frequency range of spectrograms: 20 kHz for group A and C; 40 kHz for group B and D; 0-40 kHz for group E. **b**, Intuitive demonstration of the classification within the 50-kHz USVs, made because two independent peaks were observed in the probability density map (inset). **c**, Quantification of the clustering results, which show high inter-cluster distances and small intra-cluster variances.

USVs. Though the fine-grained clustering process of 50-kHz USVs produces 60 clusters in this work, the proposed algorithm can be run in a progressive manner to produce a higher or lower number of clusters if stricter or looser thresholds are used for particular purposes.

2.2 Hierarchical behavior classification of two freely behaving rats

ARBUR:Behavior is based on a hierarchical classification architecture, which we optimizeded for analyzing the species-specific behaviors of two freely behaving rats (see Methods). We constructed different classification features based on behavioral states (non-social, moving, and social) and designed three classifiers to distinguish behavioral categories further (see Extended Data Fig.3). ARBUR:Behavior uses the binocular side-view video stream (segmented

Table 3 ARBUR:Bheavior achieves comprehensive behavior classification

	Input modality	Algorithm characteristics	Background subtraction	Social behavior	Behavior categories	F1-score/ Accuracy
DeepEthogram [19]	RGB (top view)	End-to-end (supervised)	-	✓	6	0.638/-
SIPEC:BehaveNet [31]	RGB (top view)	End-to-end (supervised)	✓	-	3	0.720/-
DeepAction [20]	RGB (top view)	Hand-crafted (supervised)	-	✓	12	-/0.739
VSAMBR [32]	RGB (side view)	Hand-crafted (supervised)	✓	-	8	-/0.771
DeepCaT-z [33]	RGB-D (top view)	End-to-end (supervised)	✓	-	4	0.822/-
OpenLabCluster [34]	Pose and tracking data	Hand-crafted (semi-supervised)	-	-	8	-/0.842
MARS [21]	RGB (top view)	Hand-crafted (supervised)	-	✓	3	0.878/-
B-SOiD [35]	Pose and tracking data	Hand-crafted (unsupervised)	-	-	11	-/0.915
ARBUR:Behavior	**RGB (side view)**	**Hand-crafted (supervised)**	✓	✓	**8**	**0.868/0.870**

Table 4 Comparison of ARBUR:SSL with mainstream sound source localization methods for rodents' USVs.

	Dimen. of SSL	Number of mics	Principle of SSL	Median error (cm)	Mic-to-source distance (cm)	Estimation area (cm^2)	Rodents
Matsumoto et al. [36]	2-D	4	Direction of arrival	1.58	47	20×20	Mice
Heckman et al. [37]	1-D	2	Hyperbolic	0.85	46	7×7.5	Mice
Warren et al. [38]	2-D	8	Hyperbolic	1.80	38	66×66	Mice
Neunuebel et al. [23]	2-D	4	Hyperbolic	3.87	33	66×66	Mice
Oliveira-Stahl et al. [39]	2-D	4	Hyperbolic	1.31	27.8	30×40	Mice
Sterling et al. [40]	2-D	4+1(64-channel)	Hyperbolic	0.48	27.8	30×40	Mice
ARBUR:SSL	**3-D**	**4**	**Hyperbolic**	**5.01**	**55**	**D50**	**Rats**

by SegFormer [41] to remove the background) as input to discriminate eight behaviors of rats (Fig.1b). ARBUR:Behavior can also run in a single-shot mode if moving behaviors are not considered.

We annotated 1,265 randomly extracted segments from all recorded videos (including 43,357 segments) to test the performance of ARBUR:Behavior. As demonstrated in Fig.3a, the module can classify the behaviors of two freely behaving rats well, especially the moving behaviors with large displacement and the social behaviors with complex postures. We show that the desired detection precision, recall, and F1-score of each behavior are achieved (average precision: 0.874, average recall: 0.862, average F1-score: 0.868, Fig.3b). The confusion matrix shows that ARBUR:Behavior can discriminate the easy-to-confuse social behaviors with high F1-scores, such as PIN (pinning) and POU (pouncing) (Fig.3c). In contrast to studies based on end-to-end architectures [19, 31, 33], ARBUR:Behavior is capable of discriminating specific social behaviors (PIN, POU, and SNC (social nose contact) and can classify a higher number of behavior categories while achieving comparable accuracy.

In addition, we tested the performance of current mainstream deep learning classification algorithms [42–45] for the direct classification of the eight behaviors on our dataset. ResNet achieves the best performance with an accuracy of only 0.333 using eight-class classifier (Extended Data Fig.4a). For the two- and three-class classification of three social behaviors, the highest accuracies of these algorithms are 0.767 and 0.578 (Extended Data Fig.4a). We argue that features with behavioral specificity can not be directly extracted in the images using the end-to-end approach, resulting in poor classification performance. Therefore, some studies extracted the pose and tracking data of rats to classify behaviors in an unsupervised way and achieved high accuracies [34, 35]. However, such methods are unable to discriminate specific social behaviors. In contrast, ARBUR:Behavior detects non-social behaviors using the location feature of rats in the image and moving behaviors by estimating the optical flow of the rat as

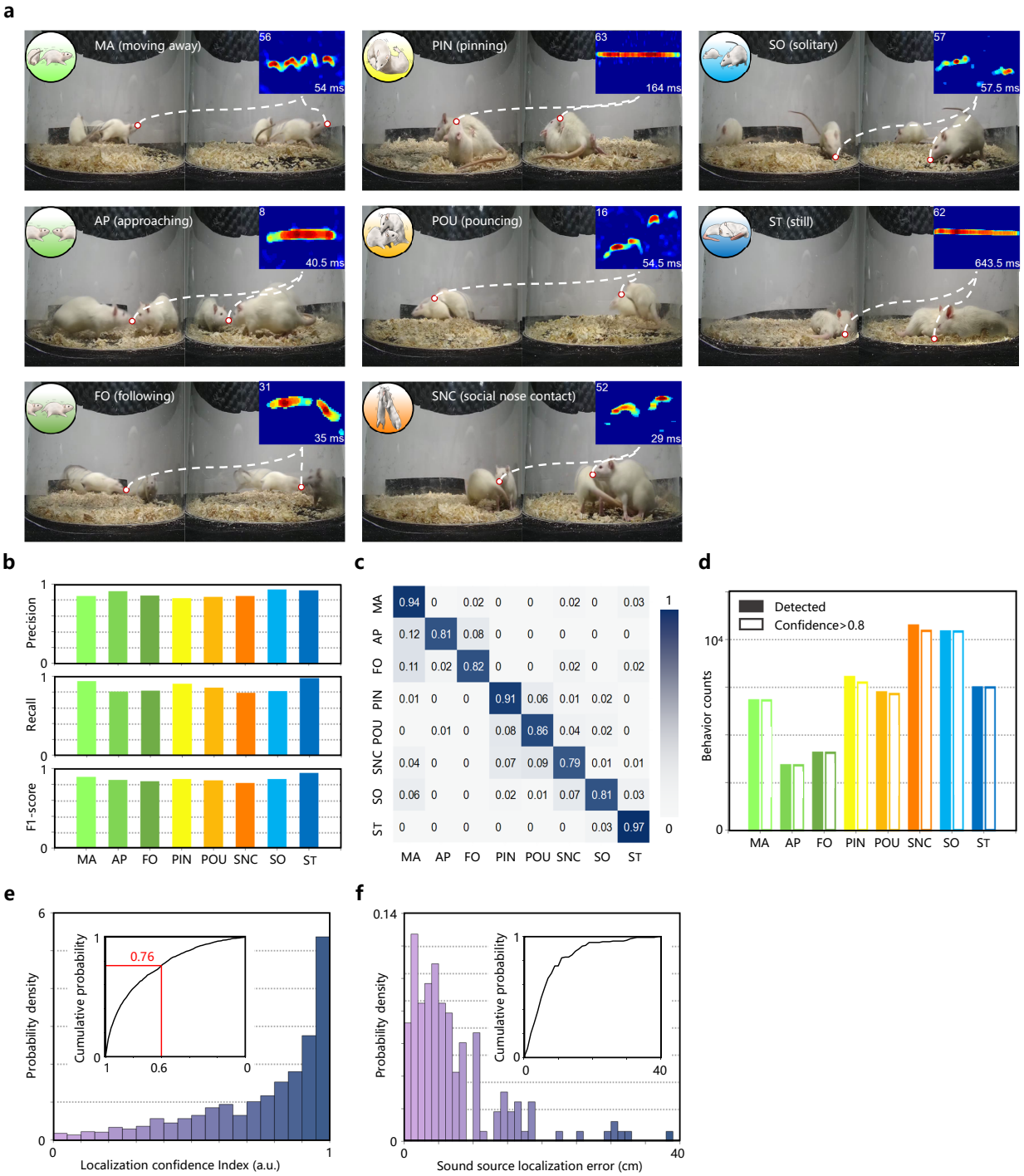


Fig. 3 ARBUR simultaneously achieves accurate behavior detection and sound source localization in three-dimensions. **a**, Examples of the ARBUR outputs for eight behaviors: left and right camera view, behavior type (top left of each panel), spectrograms of USVs with cluster (inset, top left) and duration (inset, bottom right) and location of the vocal rat in the two camera views (red circle). Frequency range of spectrograms: 20 kHz for cluster 8; 0-40 kHz for clusters 62 and 63; 40 kHz for all other clusters. **b**, Quantification of the performance of ARBUR:Behavior. **c**, Confusion matrix showing the identification rates of ARBUR:Behavior. **d**, Counts of different types of behavior detected (logarithmic) across behavior types. **e**, Distribution of the localization confidence index (LCI) obtained by ARBUR:SSL; inset, cumulative probability vs. LCI. **f**, Distribution of the sound source localization error by ARBUR:SSL; inset, cumulative probability vs. error.

well as its relative orientation. To discriminate social behaviors, it extracts histogram of oriented gradients (HOG) [46] descriptors to represent the statistics of social posture features in a standardized way.

We also tested the performance of mainstream machine learning-based classifiers for two- and three-class classifiers on three social behaviors (Extended Data Fig.4b). The support vector machine (SVM) achieved the highest classification accuracy (0.843) in binary classification, and the Receiver Operating Characteristic (ROC) curve shows that it has the best overall classification performance (Extended Data Fig.4c). Therefore, we designed an SVM-based decision tree classifier for discriminating social behaviors

(see Methods). Compared with hand-crafted features used in existing studies [20, 21, 32], ARBUR:Behavior has a greater advantage in terms of F1-score/accuracy and the number of behavior categories. On balance, ARBUR:Behavior outperforms existing studies with respect to comprehensive detection performance (Table 3). ARBUR:Behavior successfully detected 33,428 behaviors (with a confidence of over 0.8) out of 43,357 segments of video frames corresponding to recorded USVs (Fig.3d).

2.3 Locating the vocal rat in 3-D space

To locate the vocal rat during free behaving scenarios, we developed a novel algorithm (ARBUR:SSL) that incorporates the lateral binocular view (for rat nose reconstruction in the Cartesian space), behavior classification results, height-sensitive distributed microphone configuration, and 3-D probabilistic triangulation of sound source to locate the vocal rat in 3-D space (see Extended Data Fig. 7). ARBUR:SSL takes the 3D reconstructed noses position, the current behavior type, and four-channel audio streams as input. It can comprehensively leverage these information and output the pixel position of the vocal rat in both binocular images, which is useful for follow-up expert quantitative and qualitative analysis.

The rat-rat social interaction scenario makes the microphone array far above the rats (50 mm above ground) compared with the experimental sets for mouse interaction, which makes the recorded USVs signal intensity (therefore signal-to-noise ratio) much weaker than others. This poses a new challenge to us because the mainstream sound source localization (SSL) suitable for mouse interaction experiments works much worse in our dataset, therefore undesirable for our purposes. To address this challenge, we optimized a classic method [23] and proposed a new index (localization confidence index, LCI), which measures the consistency between 3-mics SSL and 4-mics SSL results, and hence can rule out those low-SNR USVs. Using the proposed method, we achieved a median SSL error of 5.01 cm in the test dataset, which is suitable for rat-rat social scenarios. Since the large size of adult rats (>18 cm), snout-to-snout distance in rat-rat interaction scenarios are much bigger than mouse-mouse interactions (see Fig.S4). Such an SSL error can be usable for safely assigning the USVs to the vocal rat in rat-rat interaction scenario.

Another concern that haunted the experts working on SSL is the lack of accurate test dataset. Mainstream studies reportedly induced mouse USVs using heterosexual urine [39, 40], but the visual localization of rodent snout would not only bring unnecessary errors, but lack the ability of snout height detection. In this paper, we constructed a rat USV test dataset for evaluating SSL errors using USVs collected from still rats, which brings two advantages: 1) all the USVs are produced at the same spatial source position because the rat kept still throughout the process; 2) 3D position of the vocal rat nose/snout is provided, therefore allowing evaluating 3D SSL algorithms.

We further show that a higher LCI is associated with smaller sound source localization (SSL) error (Extended Data Fig.6b). Therefore, only USV signals with an LCI over 0.6 are accepted, which consist of 76% of the valid vocal repertoire (29772 USVs, intersection of the valid outputs of ARBUR:USV and ARBUR:Behavior, Extended Data Fig.7). ARBUR:SSL shows decent horizontal-plane localization precision, with less than 10 cm error for over 78% of the data, which is considerable, considering the much weaker USVs in our dataset, compared to existing algorithms (see [12, 23]) (Fig.3f, Extended Data Fig.6a), and more importantly, suitable for our purposes. Moreover, by evaluating the LCI at different heights (reconstructed from the binocular view), ARBUR:SSL can pinpoint the vocal rat even if the two rats overlap in the top-down projection plane (which is more frequent in, for example, pinning and pouncing). ARBUR:SSL successfully assigned 17,235 USVs out of 29,772 USVs to the vocal rat. Fig.3a shows examples. In contrast to other sound source localization methods for rodents' USVs, ARBUR:SSL features the ability to locate the vocal rat in 3-D space while preserving the localization precision with weaker USV signals (Table 4).

2.4 Revealing the behavior-vocalization interplay

By seamlessly incorporating these three modules, ARBUR can build connections between different types of rat behavior and vocalizations. Fig.4a shows examples. It is intuitive that when the rats are engaged in non-aggressive play behavior, such as social nose contact and pouncing, more appetitive 50-kHz USVs are communicated. On the contrary, the submissive one emits more aversive 22-kHz USVs when engaged in aggressive play behavior, i.e., pinning. We investigated whether these two hypotheses would hold in the whole repertoire. Therefore, we quantified the USV proportions of rats in different types of behavior (Fig.4b). Results show that the submissive rat produces significantly more aversive 22-kHz USVs during the pinning process. Moreover, in social nose contact and pouncing, both rats produce more 50-kHz USVs.

The finding that during an aggressive interaction (i.e., pinning), the submissive one produces significantly more aversive 22-kHz USVs, is consistent with former research [26, 47, 48]. We also show that even though pouncing and pinning are similar types of behavior, they actually have distinct USV distributions. In particular, during pinning, the submissive rat rarely emits appetitive 50-kHz USVs, whereas during pouncing, the submissive one produces mainly 50-kHz USVs. This may indicate that the submissive rat has distinct feelings and emotional states when engaging in the two types of behavior. ARBUR

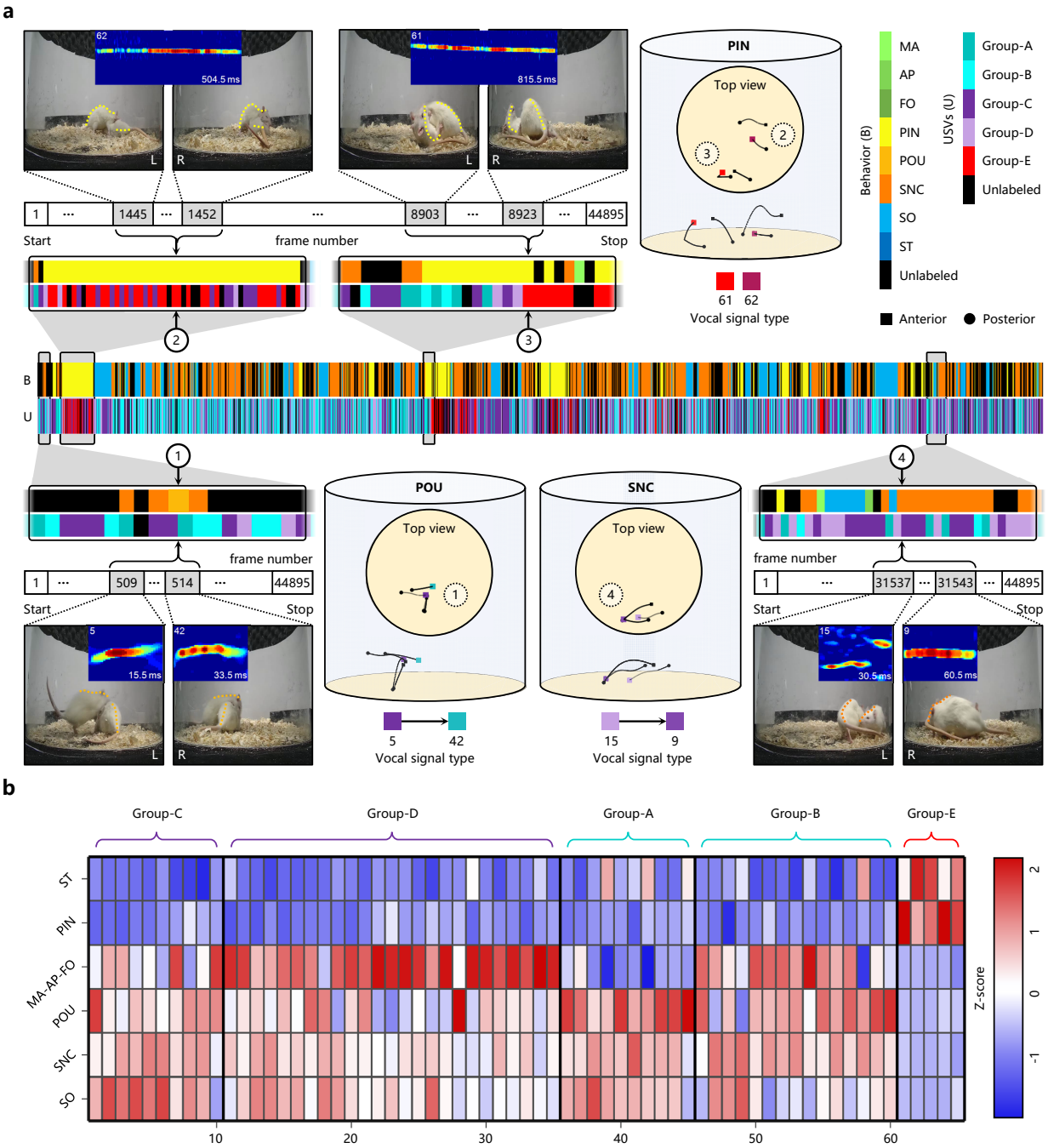


Fig. 4 Revealing the behavior-vocalization interplay using ARBUR. **a**, Examples showing how 22-kHz USVs are associated with pinning (top), and appetitive 50-kHz USVs are associated with pouncing (bottom left) and social nose contact (bottom right). Frequency range of spectrograms: 20 kHz for clusters 5, 9, and 42; 0-40 kHz for clusters 61 and 62; and 40 kHz for cluster 15. **b**, Quantification of behavior-associated USV distributions, which indicates whether a certain cluster of USVs is emitted during a specific type of behavior above chance (red), at chance (white) or below chance (blue). Groups A-E, the same as in Fig.2.

can reveal such a distinction because it can not only accurately discriminate these behaviors, but locate the submissive rat spatially. These findings are in line with former studies and therefore validate the effectiveness of ARBUR.

In addition, ARBUR reveals that several USVs, mostly 22-kHz USVs, are recorded when both rats are still (Fig.4b), indicating that some USVs may be emitted unconsciously (possibly during sleeping), and these 22-kHz USVs may neither convey emotional information nor trigger behavioral changes of the other rat. However, the still-associated (or sleeping-associated) USVs have rarely been focused on because of their independence from social interaction. Furthermore, ARBUR shows that rats produce more step 50-kHz USVs than non-step ones during motion (Fig 4b), indicating a part of step 50-kHz USVs may be caused by rats' locomotion. Recent research has demonstrated the tight correlation between respiratory activity and frequency features of USVs [25], which indicated that locomotion may induce

more step 50-kHz USVs by affecting rat respiratory activity.

3 Discussion

Understanding the behavior-vocalization interplay of rats is inhibited by the difficulty of relating the behaviors and the USVs of freely behaving rats in complex social contexts to objectively reveal the underlying behavior-vocalization interplay.

In this study, we initially proposed a machine learning-based analysis system (named ARBUR) to relate rat vocalizations to their free behavior. ARBUR contains three modules: 1) an unsupervised learning algorithm (ARBUR:USV) that clusters rat USVs in an unbiased manner by comprehensively considering their mean frequency, duration, and both non-step (continuous) and step (non-continuous) contours in the spectrograms; 2) a machine learning-based hierarchical framework (ARBUR:Behavior) that detects the type of social behavior from a laterally binocular view; and 3) a sound source localization algorithm (ARBUR:SSL) that spatially allocates the USVs to the vocal rats.

ARBUR combines the advantages of comprehensive and unbiased USV clustering, hierarchical high-accuracy rat behavior detection, and 3-D sound source localization to reveal the latent behavior-vocalization interplay of rats. We also note that although ARBUR is designed to investigate the behavior-vocalization interplay of rats, it can be readily generalized to other rodents that also communicate through both non-verbal signals and USVs (for example, mice and mole rats). However, ARBUR is still limited in the following aspects. First, ARBUR:USV rejects the concurrent USVs (simultaneously recorded USVs at different frequencies), which may be important during social communication and could possibly be separated by intelligent segmentation algorithms combined with high-precision SSL, further enriching the vocal repertoire. Moreover, ARBUR:Behavior lacks the ability to detect potential behavioral changes in freely behaving rats, which is essential for discovering the underlying mechanisms of novel behavior-vocalization and could possibly be enhanced by incorporating unsupervised learning to uncover latent structures and categories in behavioral space [49, 50]. In addition, ARBUR:SSL could be improved in terms of the limited number of visual reconstructions of rat noses by incorporating more camera views in the future, fulfilling the need for a more reliable three-dimensional sound source localization.

Using ARBUR, we revealed that rat USV distribution is biased by rat behavior, indicating that the USVs of rats are associated with their behaviors. For example, a submissive rat significantly up-regulates the aversive 22-kHz USVs during pinning. Moreover, during pouncing and social nose contact, both rats produce substantially more appetitive 50-kHz USVs than they do during pinning.

In summary, we proposed a machine learning-based analysis system combined with an autonomous rat-robot interaction framework, which can reveal the underlying behavior-vocalization interplay of rats. This work could help mechanistically understand the interactive influences between the behavior and USVs of rats.

4 Methods

4.1 Animals and Experiment preparation

Animals. Adult (8-10 weeks) male (n = 4) and female (n = 4) SD rats (stock number 198499, 212822, SPF (Beijing) biotechnology co., LTD.) were used. Rats were kept on a 12/12 h light/dark cycle at a consistent ambient temperature (26 1° C) and humidity (50 5%), and all experiments were performed during the light cycle. Food and water were accessed ad libitum. All experimental procedures were approved by the Institutional Animal Care and Use Committee of the Beijing Institute of Technology, Beijing, China. Prior to all recording experiments and rat-robot interaction experiments, rats were singly housed for at least 3 days to minimize group housing effects on the social behavior of the rats.

4.2 Vocal signals clustering

Extraction of frequency contours. The segmented spectrograms (time-frequency matrices) were filtered sequentially by the median, Gaussian, and threshold filters to enhance the signal-to-noise ratio (for audio segmentation, see Supplementary Notes). The filtered spectrograms were then binarized for detecting the boundaries of the frequency contours (edges in the spectrograms) by the modified Moore's neighbor tracking algorithm with the Jacobian's termination condition. Those contours with less than 22 points and a maximum intensity inside the contour of less than 15 were detected as isolated islands and eliminated. For the discontinuous USVs with multiple contours, only the longest four contours were selected as the final contours to recapitulate the original complex signals. Harmonics were considered as those frequency contours with 90% overlap in time, and only the frequency contour of the lowest mean frequency was preserved whereas others were eliminated. Other than harmonics, those frequency contours that share an overlap of duration over 3 ms were considered overlapping vocalizations emitted by two rats and also abolished.

Coarse clustering. Coarse clustering consists of two steps. In the first step, the USVs were first clustered coarsely into 2 groups according to their mean frequency: aversive 22-kHz and appetitive 50-kHz USVs. In the second step, the 2 groups were further divided into 5 groups (A-D) according to

their distribution of duration and mean frequency of frequency contours: so-called 22-kHz, low-peak-frequency/high-peak-frequency step/non-step 50-kHz USVs (Extended Data Fig.1a). Specifically, those USVs with a mean frequency of less than 32 kHz were classified as 22-kHz signals. The rest of the USVs (50-kHz) were clustered into 2 groups (low-peak-frequency 50-kHz and high-peak-frequency 50-kHz) by k-means clustering according to two features: duration and mean frequency. These USVs were further classified into 4 groups according to their continuity of frequency contours: non-step (continuous) USVs containing only one contour and step USVs containing more than one contour.

Determining the optimal number of clusters. In the third step, groups A-E were further divided using unsupervised learning within their groups. For the appropriate setting of the cluster number, ARBUR:USV provides users with three choices: arbitrary, automatic and progressive methods. The arbitrary method allows users to set integer number for each cluster. For example, in Fig. 2, cluster number for groups A-E are set as 10, 25, 10, 15, and 5, respectively. The automatic method relies on the well-known elbow method to avoid over-clustering, which works as follows. For each cluster number, the total within-cluster error (TWCE) is calculated. By sliding from 2 clusters to 100 clusters (200 for group B for its complexity), forming a TWCE-clusters curve. Then, for each point on this curve, its left and right side are used for linear fitting, and the total fitting error is stored, which forms a second curve: fitting error versus cluster number (see 3 as an example). The global bottom of the second curve is called the inflection point, which indicates the optimal number of clusters. The progressive method takes a user defined intra-cluster variance parameter to find the appropriate number. It calculates the innerpoint percentage (percent of points falling within the mean±2.5SD range of each cluster). Theoretically, with increasing clusters, the intra-cluster variance decreases. This method starts with 2 clusters and calculate the average innerpoint percent and stops when the user-defined threshold is first reached. For example, if the threshold is set as 98%, that is, average 98% points should be innerpoints for all clusters and all features. The optimal cluster should be 8 for group E (see Fig.S1).

Refined clustering of non-step USVs. The non-step (continuous) signals were clustered further according to their contour shapes. The detected boundary points were centered by subtracting the mean frequency. To normalize these centered contours with different duration, we constructed a fixed-length feature vector containing 100 features (Extended Data Fig.1b). The first 50 and last 50 features represent the upper and the bottom boundaries of the centered contour, respectively. Each contour was linearly mapped to a 100-feature vector. By doing so, frequency contours with different duration or mean frequencies can be clustered appropriately according to their shapes. These USVs were categorized using k-means clustering based on the 100-feature vectors. The number of clusters for low-frequency and high-frequency USVs were both arbitrarily set as 10. Alternatively, the number of clusters can be determined progressively as in [12].

Refined clustering of step signals. For the step (discontinuous) signals with multiple contours, a similar 100-feature vector was constructed based on these contours (Extended Data Fig.1b). All contours of each USV were centered by subtracting their mean frequencies. Then, the upper boundaries of each contour were sorted in ascending order in time and mapped linearly to a 50-feature vector. By doing so, those inter-contour intervals were deleted, and contours were placed compactly. Similarly, the bottom boundaries were mapped to the second 50-feature vector. These step USVs were then categorized using k-means clustering based on the 100-feature vectors concatenating these two vectors together. The number of clusters for low-frequency and high-frequency USVs were arbitrarily set as 15 and 25, respectively.

Refined clustering of 22-kHz USVs. 22-kHz USVs vary a lot in duration, ranging from several microseconds to several seconds, such a feature should not be ignored in clustering. In contrast, the contour shapes varied little across USVs. Therefore, we constructed a 2-feature vector for each 22-kHz signal containing the mean frequency and duration. These 22-kHz signals were then categorized using k-means clustering based on the 2-feature vectors. The number of clusters was arbitrarily set to 5.

Validating clustering. To validate clustering results, dissimilarities within each cluster (intra-cluster variance) and between different clusters (inter-cluster distance) were evaluated, respectively. They were calculated based on the constructed 100-feature vectors for both step and non-step USVs (Extended Data Fig.2a). The inter-cluster distance of the ith cluster D_i was defined as the average difference between each pair of frequency contours from two different clusters within each group (one of 5 groups by coarse clustering):

$$D_i = \sum_{j|\neq i}^{N_x} \left\|\overline{\mathbf{V}}_i - \overline{\mathbf{V}}_j\right\|_2 / (n_x - 1), \tag{1}$$

where $\|\cdot\|_2$ denotes l_2-norm, N_x is the index integer set of one of five particular groups, $\overline{\mathbf{V}}_i \in R^{100}$ is the average feature vector of the ith cluster in that group and n_x is the number of individuals in N_x. The intra-cluster variance of ith cluster VAR_i was calculated as:

$$\mathrm{VAR}_i = \sum_{m}^{N_i} \left\|\mathbf{V}_m - \overline{\mathbf{V}}_i\right\|_2 / n_i, \tag{2}$$

where N_i is the index integer set of the ith cluster, $\mathbf{V}_m$ is the feature vector of the mth USV in this cluster, and n_i is the number of individuals in N_i.

4.3 Behavior classification

The workflow of ARBUR: Behavior is shown in Extended Data Fig.3. ARBUR:Behavior uses the binocular side-view video stream (segmented by SegFormer to remove the background) as input to discriminate eight behaviors of rats. First, the rat regions within each input image were labeled by edge detection. Detected regions were recognized as outlier regions if their areas are below the maximum error range:

$$L_a^e = L_f L_a^{max}, \tag{3}$$

where L_a^{max} is the size of the maximum area with labels in the current image, and L_f is the filter threshold. The segmented image and labels were then input into a hierarchical algorithm to classify the non-social, moving, and social state behaviors of rats as follows.

Non-social state behaviors. The movement variation (in pixels) M_p of the center position of each label was calculated to determine whether rats have moved:

$$M_p = L_{center}^N - L_{center}^1, \tag{4}$$

where L_{center}^N and L_{center}^1 are the center pixel positions of the labels in the last and the first image. The value of M_p for sleeping rats hardly changes. For individually moving rats, we consider the total number of labels in each image to reflect whether the rats come into contact with each other (solitary) within that sequence of images.

Moving state behaviors. Based on the behavioral definition in Table 1, we used the number of labels in the first and last images of the image sequence to initially determine the moving state behaviors. Then, we estimated the motion optical flow of the rat in the images and define its centroid motion vector as:

$$\overrightarrow{V_c^{label}} = \sum\sum \overrightarrow{V_{u,v}^{label}}, \tag{5}$$

where u and v are the pixel coordinates of the image. As shown in the moving state behaviors section of Extended Data Fig.3, we determined the specific behavior by judging the vector $\overrightarrow{V_c^{label}}$ relative direction of the two labels.

Social state behaviors. We designed a decision tree algorithm based on multiple binary SVM classifiers to detect social state behaviors. A total of 860 images were labeled (PIN: 280, POU: 300, SNC: 280), and the HOG features were extracted as the training dataset. The SVM models were trained using a third-degree polynomial kernel function, and parameter optimization was performed for each model. Specifically, for each image, we deployed three SVM models simultaneously (PIN-POU, PIN-SNC, POU-SNC) to determine the category of the image with maximum probability. We defined the total confidence of the behavioral category for all images:

$$P_{label} = \frac{1}{2N}\sum_{n=1}^{N}\left(S_{label}\left(I_{Rc}^n\right) + S_{label}\left(I_{Lc}^n\right)\right), \tag{6}$$

where

$$S_{label}\left(I\right) = \begin{cases} 1, P\left(I = label\right) \geq I_l \\ 0, P\left(I = label\right) < I_l \end{cases} \tag{7}$$

where $P\left(I = label\right)$ is the probability that the current image category is *label* (PIN, POU, SNC), and I_l is the likelihood of a specific image belonging to the category. Behavioral category I_c was mapped through the maximum probability in each P_{label}:

$$\max\left(P_{PIN}, P_{POU}, P_{SNC}\right) \rightarrow I_c. \tag{8}$$

Image sequences that do not meet the aforementioned criteria were labeled as undetected.

4.4 Sound source localization

3D keypoints detection of rats. To obtain the spatial position of the two rats, we designed a deep learning-based keypoint detection network capable of detecting seven keypoints distributed across the rats' heads, spines, and tails. We then calculated the 3D coordinates of the keypoints via triangulation by utilizing the calibrated camera parameters and the detected key points in two image spaces. It is worth noting that for rats engaged in social behaviors, there can be occlusions of certain body parts, resulting in only a part of keypoints being detectable in each frame. Using these 3D coordinates, a curve that represents the posture of the rat in the current frame could be fitted (Fig.4), which could be useful to determine the submissive rat and the dominant one during PIN (pinning) and POU (pouncing). For other cases, only the 3D coordinates of the rats' noses were used for the downstream SSL processes.

Estimating sound sources. The sources of USVs were estimated using an algorithm modified from Neunuebel et al. [23]. The workflow of this algorithm is illustrated in Extended Data Fig.5. For each extracted USV from four microphones, a data augmentation process was performed by slicing the recognized frequency contours (see Method 4.2) into m segments. Then, four groups with 3 channels (mics) and a group with 4 mics were input to a steered response power-based sound source localization method. That is, $m \times 5$ estimates were calculated. The estimates with 4 mics ($m \times 4$) estimates and those with 3 mics ($m \times 5$) were used to calculate two 2-D probability density functions ($f_3(x, y)$, $f_4(x, y)$, $x, y \in \mathbb{R}$), respectively. Then, the localization confidence index (LCI) P_{LCI} that represents the reliability

of estimation results was produced as follows.

$$P_{LCI} = \frac{f_3(x,y)|[(x,y), f_4(x,y) = \max(f_4(x,y))]}{\max(f_3(x,y))}. \tag{9}$$

If the LCI is above the pre-set threshold (0.6), the SSL result is accepted as the center of estimates with maximum probability densities in $f_3(x,y)$ and $f_4(x,y)$. The SSL results are rejected otherwise.

Combining visual and SLL estimations. Because rats have a body length as long as 25 cm, the height of their sound source cannot be ignored in some cases. For example, if a rat presents a high-rearing pose, the nose can be as high as over 20 cm. Under these circumstances, the height should be considered to avoid incorrect SLL prediction.

For those behaviors with no apparent height difference between the noses of two rats (besides POU and PIN), the SSL results were assigned to rat noses based on horizontal information only. That is, the USV was assigned to the nearer rat if two noses were detected. If only one rat nose was detected, the USV was assigned to the detected one if it lies within the SLL estimation area (10 cm from the estimated point), and the USV was assigned to the undetected one otherwise. If no rat noses were detected, the USV cannot be assigned.

For pouncing and pinning behaviors, the SSL results were assigned based on not only horizontal information (see last paragraph) but also 3-D information. Other If two rat noses were detected, SSL results were calculated again with the nose height added. The USV was allocated to the rat with a higher LCI. If only one rat nose was detected and an obvious height difference was observed and labeled, the USV can be allocated to the rat with a higher lCI by assuming the height of another rat. This 3D SSL process is useful if two rats are in close proximity but have obvious height differences, which is common in pinning and pouncing.

4.5 Real-time tracking of freely behaving rats

In order for the robot to track the rat in real-time, we first need to detect and locate the rats, and then control the robot to approach it.

Detection. The real-time image used for tracking was acquired from the built-in binocular camera of the robot. We deployed the YOLOv4 [51] algorithm to detect the head, body, and tail of rats. 6000 hand-labeled images were used to train the model. The detection speed of the model is 64.1 frames per second (fps), and the average precision at an Intersection over Union (IoU) threshold of 0.5 (AP50) is 0.95% (GPU: NVIDIA RTX3090).

Estimating rat location. The location of the rat with respect to the robot was obtained by a stereo-depth estimation model. First, to balance speed and accuracy, we optimized the stereo-matching strategy by incorporating the detection results of the rats. Here, only the ORB feature points within the regions containing the rats in the image were extracted and matched. The center coordinates u_a and v_a of the detection bounding box were utilized as a substitute for the centroid position of the rat on the image. The location of the rat in the camera coordinate system can be estimated as:

$$\begin{cases} X_a^{rat} = Z_a^{rat}\dfrac{u_a - c_x}{f_x} \\ Y_a^{rat} = Z_a^{rat}\dfrac{v_a - c_y}{f_y} \\ Z_a^{rat} = \dfrac{n f_x b}{\sum (u_l - u_r)} \end{cases} \tag{10}$$

where c_x, c_y, f_x, and f_y are the intrinsic parameters of the camera, which can be obtained by calibration; n is the number of valid feature point matches; b is the baseline length of the stereo camera. u_l and u_r are the horizontal coordinates of the feature point in the left and right images. The corresponding location can then be obtained by using YOLOv4 to detect different attention areas of the rat (T_a).

Visual servo control. The process of real-time tracking of rats using visual servo control is illustrated in Figure 1c. We define the state feedback variable as follows:

$$S = \left[X_a^{rat}, Y_a^{rat}, Z_a^{rat}\right]. \tag{11}$$

The robot updates the state variable S at a frequency of 25 Hz. Given T_s and T_a, the objective of visual servo control can be transformed to:

$$e\left(T_a, T_s, S_t\right) = S_t - S^*. \tag{12}$$

where, $S^* = [0, 0, T_s]$, S_t is the state at the current moment t. The robot is controlled to reach the target position by minimizing $e\left(T_a, T_s, S_t\right)$ and mapping it into the robot's movement variables $[J^s, W^s]$:

$$\min e\left(T_a, T_s, S_t\right) \rightarrow [J^s, W^s]. \tag{13}$$

where,

$$\begin{cases} J^S = [J_{Tp}^S, J_{Wp}^S, J_{Wy}^S, J_{Hy}^S, J_{Hp}^S] \\ W^S = [W_v^S, W_\omega^S] \end{cases} \tag{14}$$

J^S is the variable in the robot's joint space (Fig.5b), W_v^S and W_ω^S are the line velocity and angular velocity of the robot controlled by the differential wheel. To enhance tracking speed and accuracy, we compensate for the errors caused by the multi-joint structure of the robot in each variable of J^S and W^S. The compensation coefficients are determined by a tracking experiment with a stationary scene.

Based on the relative location of the rat to the robot in the interaction, we designed three types of robot motion strategy: 1) Searching for the rat during the initialization phase or retrieving the rat if it

has been lost. The robot searches for the rat by rotating the differential wheel. Once the rat is detected, it is tracked according to strategies 2 and 3. 2) The rat's location is within a range of $5T_s$. The multi-joint motion of the robot allows it to track the rat within a 360 range. When the robot reaches the joint limits and the rat continues to move towards the limit, the robot can utilize the differential wheel rotations in place to continue tracking the rat. 3) The rat's location is outside a range of $5T_s$. To improve the efficiency of the interaction, the strategy specifies that the robot preferentially uses the differential wheel to track the rat until the rat's location is within a range of $5T_s$.

4.6 Statistics

All statistical analyses were performed in MATLAB (version 2021b, MathWorks) using two-sample single-tailed Welch's t-tests with a significance level of $\alpha = 0.05$. Data distributions were assumed to be normal, but this was not formally tested.

Declarations

Data Availability

All data needed to evaluate the conclusions in the paper are present in the Article and/or the supplementary information.).

Code Availability

The open-source software ARBUR and its tutorials are freely available at Github (https://github.com/Guanglu-Jia/ARBUR) and Zenodo (https://doi.org/10.5281/zenodo.8081539) [52] .

Acknowledgments

This study was funded by the National Natural Science Foundation of China under Grant 62022014, the Science and Technology Innovation Program of Beijing Institute of Technology under Grant 2022CX01010, and the National Science and Technology Key major projects (STI2030-Major Projects 2022ZD02068000).

Author contributions

Q.H. and Q.S. conceived and supervised the project. G.J. and Z.C. implemented the methods, conducted the experiments, processed the data, analyzed the results, and wrote the manuscript. Q.Z. and Y.Z. assisted in conducting the experiments. All authors contributed to discussions. All authors edited and approved the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary information

The supplementary information contains the supplementary notes, Figs. S1-S4.

References

[1] Barker, A. J. *et al.* Cultural transmission of vocal dialect in the naked mole-rat. *Science* **371**, 503–507 (2021).

[2] Padilla-Coreano, N. *et al.* Cortical ensembles orchestrate social competition through hypothalamic outputs. *Nature* **603**, 667–671 (2022).

[3] 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).

[4] Dong, S., Lin, T., Nieh, J. C. & Tan, K. Social signal learning of the waggle dance in honey bees. *Science* **379**, 1015–1018 (2023).

[5] VanRyzin, J. W. *et al.* Microglial phagocytosis of newborn cells is induced by endocannabinoids and sculpts sex differences in juvenile rat social play. *Neuron* **102**, 435–449 (2019).

[6] Robinson, G. E., Fernald, R. D. & Clayton, D. F. Genes and social behavior. *Science* **322**, 896–900 (2008).

[7] Wei, D., Talwar, V. & Lin, D. Neural circuits of social behaviors: Innate yet flexible. *Neuron* **109**, 1600–1620 (2021).

[8] Li, S. W. *et al.* Frontal neurons driving competitive behaviour and ecology of social groups. *Nature* **603**, 661–666 (2022).

[9] Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioral and neural analysis. *Nature* **617**, 360–368 (2023).

[10] Murugan, M. *et al.* Combined social and spatial coding in a descending projection from the prefrontal cortex. *Cell* **171**, 1663–1677 (2017).

[11] Knutson, B., Burgdorf, J. & Panksepp, J. Anticipation of play elicits high-frequency ultrasonic vocalizations in young rats. *Journal of comparative psychology* **112**, 65 (1998).

[12] Sangiamo, D. T., Warren, M. R. & Neunuebel, J. P. Ultrasonic signals associated with different types of social behavior of mice. *Nature neuroscience* **23**, 411–422 (2020).

[13] Venniro, M. & Shaham, Y. An operant social self-administration and choice model in rats. *Nature protocols* **15**, 1542–1559 (2020).

[14] Brudzynski, S. M. Ethotransmission: communication of emotional states through ultrasonic vocalization in rats. *Current opinion in neurobiology* **23**, 310–317 (2013).

[15] Bartal, I. B.-A., Decety, J. & Mason, P. Empathy and pro-social behavior in rats. *Science* **334**, 1427–1430 (2011).

[16] Takahashi, N., Kashino, M. & Hironaka, N. Structure of rat ultrasonic vocalizations and its relevance to behavior. *PloS one* **5**, e14115 (2010).

[17] Lorbach, M. *et al.* Learning to recognize rat social behavior: Novel dataset and cross-dataset application. *Journal of neuroscience methods* **300**, 166–172 (2018).

[18] Vogt, N. Automated behavioral analysis. *Nature Methods* **18**, 29–29 (2021).

[19] Bohnslav, J. P. *et al.* Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels. *Elife* **10**, e63377 (2021).

[20] Harris, C., Finn, K. R., Kieseler, M.-L., Maechler, M. R. & Tse, P. U. Deepaction: a matlab toolbox for automated classification of animal behavior in video. *Scientific Reports* **13**, 2688 (2023).

[21] Segalin, C. *et al.* The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice. *Elife* **10**, e63720 (2021).

[22] Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S. & Branson, K. Jaaba: interactive machine learning for automatic annotation of animal behavior. *Nature methods* **10**, 64–67 (2013).

[23] Neunuebel, J. P., Taylor, A. L., Arthur, B. J. & Egnor, S. R. Female mice ultrasonically interact with males during courtship displays. *Elife* **4**, e06203 (2015).

[24] Wright, J. M., Gourdon, J. C. & Clarke, P. B. Identification of multiple call categories within the rich repertoire of adult rat 50-khz ultrasonic vocalizations: effects of amphetamine and social context. *Psychopharmacology* **211**, 1–13 (2010).

[25] Riede, T. Stereotypic laryngeal and respiratory motor patterns generate different call types in rat ultrasound vocalization. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology* **319**, 213–224 (2013).

[26] Burgdorf, J. *et al.* Ultrasonic vocalizations of rats (rattus norvegicus) during mating, play, and aggression: Behavioral concomitants, relationship to reward, and self-administration of playback.. *Journal of comparative psychology* **122**, 357 (2008).

[27] Fonseca, A. H., Santana, G. M., Bosque Ortiz, G. M., Bampi, S. & Dietrich, M. O. Analysis of ultrasonic vocalizations from mice using computer vision and machine learning. *Elife* **10**, e59161 (2021).

[28] Van Segbroeck, M., Knoll, A. T., Levitt, P. & Narayanan, S. Mupetmouse ultrasonic profile extraction: a signal processing tool for rapid and unsupervised analysis of ultrasonic vocalizations. *Neuron* **94**, 465–485 (2017).

[29] Coffey, K. R., Marx, R. E. & Neumaier, J. F. Deepsqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology* **44**, 859–868 (2019).

[30] Goffinet, J., Brudner, S., Mooney, R. & Pearson, J. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *Elife* **10**, e67855 (2021).

[31] Marks, M. *et al.* Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. *Nature machine intelligence* **4**, 331–340 (2022).

[32] Jhuang, H. *et al.* Automated home-cage behavioural phenotyping of mice. *Nature communications* **1**, 68 (2010).

[33] Gerós, A., Cruz, R., de Chaumont, F., Cardoso, J. S. & Aguiar, P. Deep learning-based system for real-time behavior recognition and closed-loop control of behavioral mazes using depth sensing. *bioRxiv* 2022–02 (2022).

[34] Li, J., Keselman, M. & Shlizerman, E. Openlabcluster: Active learning based clustering and classification of animal behaviors in videos based on automatically extracted kinematic body keypoints. *bioRxiv* 2022–10 (2022).

[35] Hsu, A. I. & Yttri, E. A. B-soid, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature communications* **12**, 5188 (2021).

[36] Matsumoto, J. *et al.* Acoustic camera system for measuring ultrasound communication in mice. *Iscience* **25**, 104812 (2022).

[37] Heckman, J. J. *et al.* High-precision spatial localization of mouse vocalizations during social interaction. *Scientific reports* **7**, 1–16 (2017).

[38] Warren, M. R., Sangiamo, D. T. & Neunuebel, J. P. High channel count microphone array accurately and precisely localizes ultrasonic signals from freely-moving mice. *Journal of neuroscience methods* **297**, 44–60 (2018).

[39] Oliveira-Stahl, G. *et al.* High-precision spatial analysis of mouse courtship vocalization behavior reveals sex and strain differences. *Scientific Reports* **13**, 5219 (2023).

[40] Sterling, M. L., Teunisse, R. & Englitz, B. Rodent ultrasonic vocal interaction resolved with millimeter precision using hybrid beamforming. *Elife* **12**, e86126 (2023).

[41] Xie, E. *et al.* Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems* **34**, 12077–12090 (2021).

[42] Touvron, H. *et al. Training data-efficient image transformers & distillation through attention*, 10347–10357 (PMLR, 2021).

[43] Liu, Z. *et al. Swin transformer: Hierarchical vision transformer using shifted windows*, 10012–10022 (2021).

[44] Liu, Z. *et al. A convnet for the 2020s*, 11976–11986 (2022).

[45] He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition*, 770–778 (2016).

[46] Dalal, N. & Triggs, B. *Histograms of oriented gradients for human detection*, Vol. 1, 886–893 (Ieee, 2005).

[47] Lukas, M. & Wöhr, M. Endogenous vasopressin, innate anxiety, and the emission of pro-social 50-khz ultrasonic vocalizations during social play behavior in juvenile rats. *Psychoneuroendocrinology* **56**, 35–44 (2015).

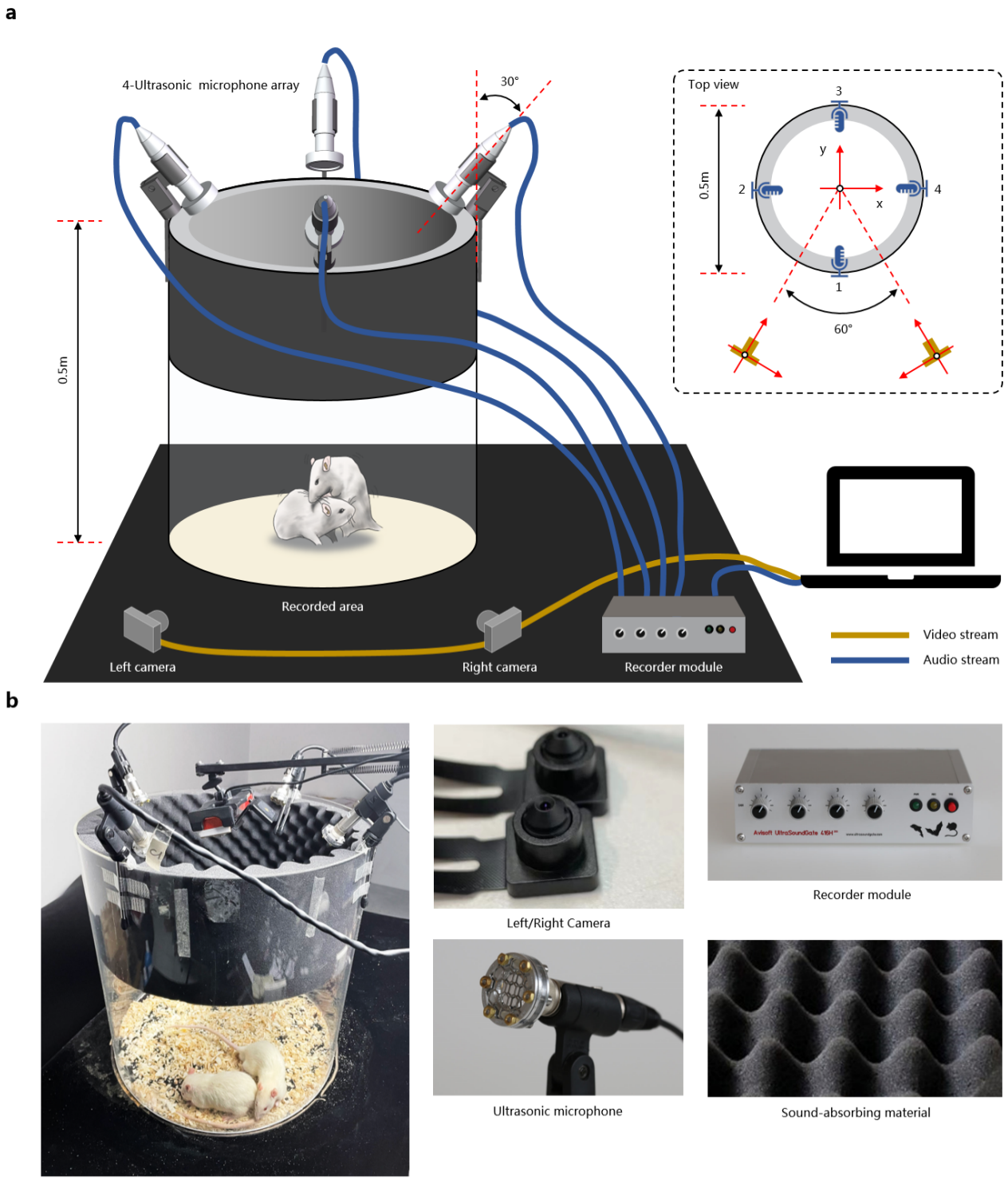
[48] Litvin, Y., Blanchard, D. C. & Blanchard, R. J. Rat 22 khz ultrasonic vocalizations as alarm cries. *Behavioural brain research* **182**, 166–172 (2007).

[49] Brattoli, B. *et al.* Unsupervised behaviour analysis and magnification (ubam) using deep learning. *Nature Machine Intelligence* **3**, 495–506 (2021).

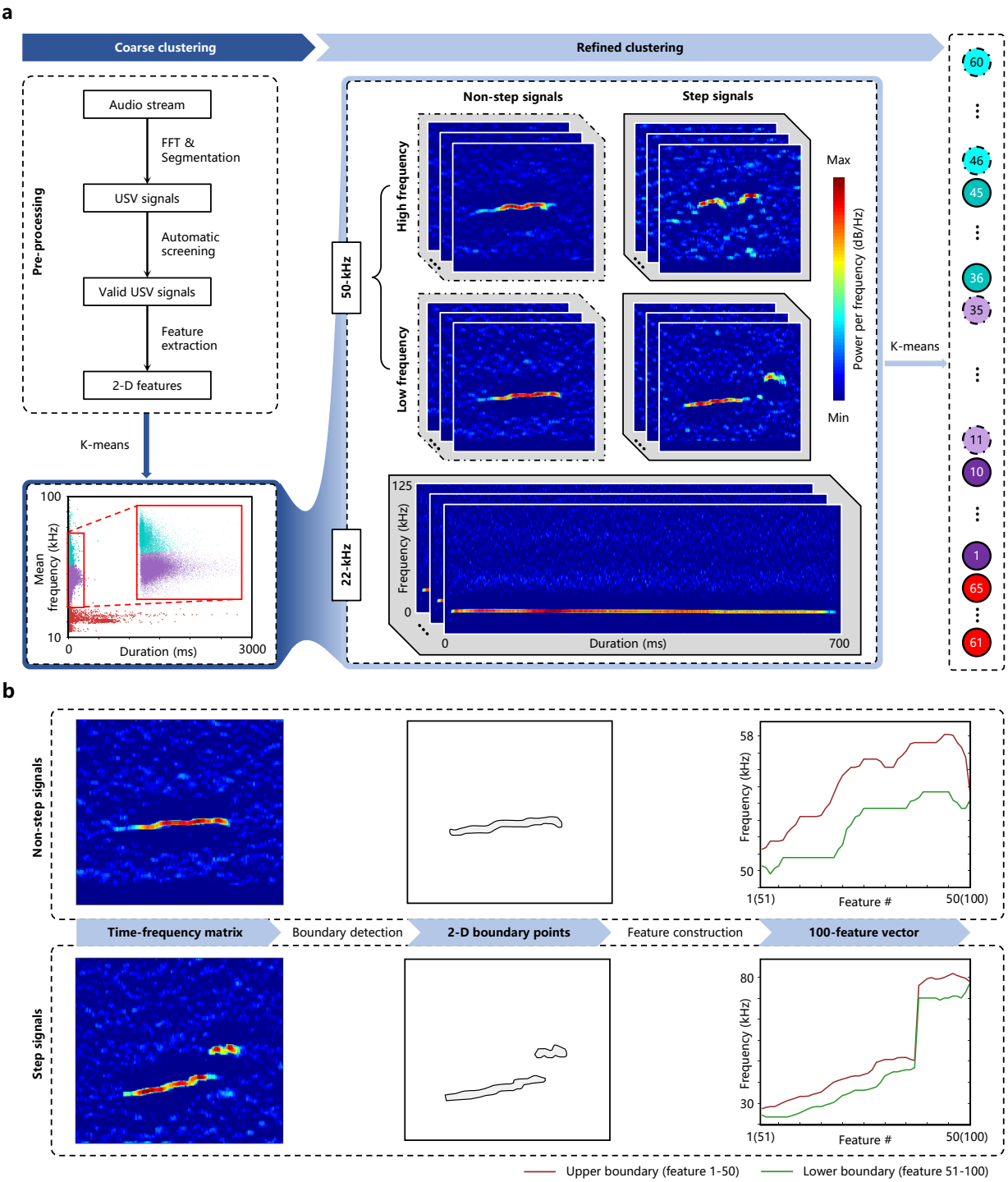
[50] Huang, K. *et al.* A hierarchical 3d-motion learning framework for animal spontaneous behavior mapping. *Nature communications* **12**, 2784 (2021).

[51] Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

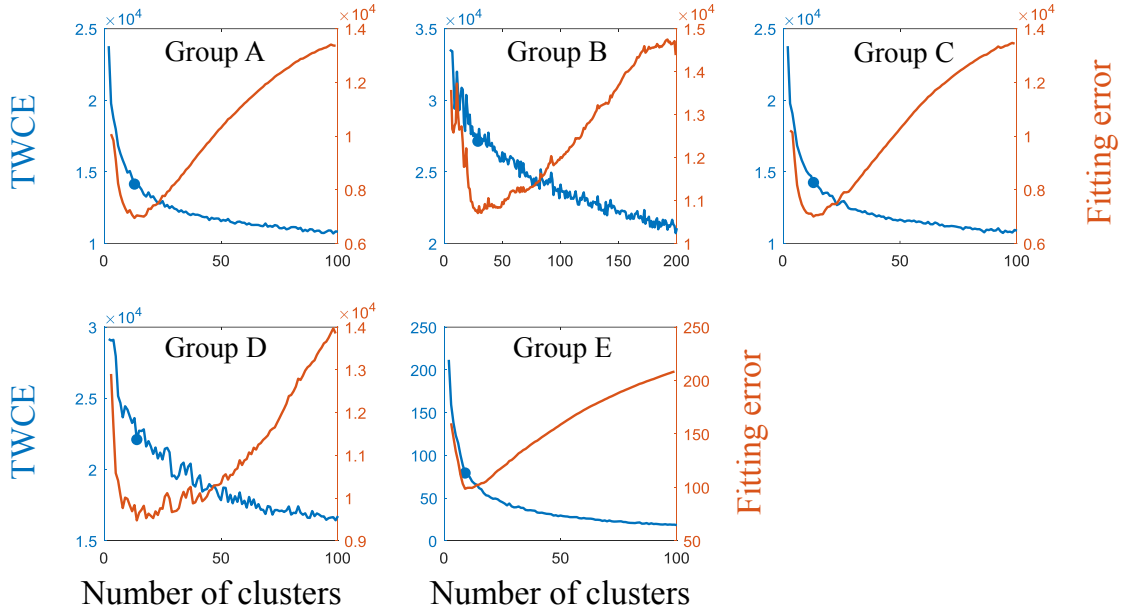
[52] Jia, G. *et al.* Arbur: A machine learning-based analysis system for relating the behavior and usvs of rats. *Zenodo* (2023).



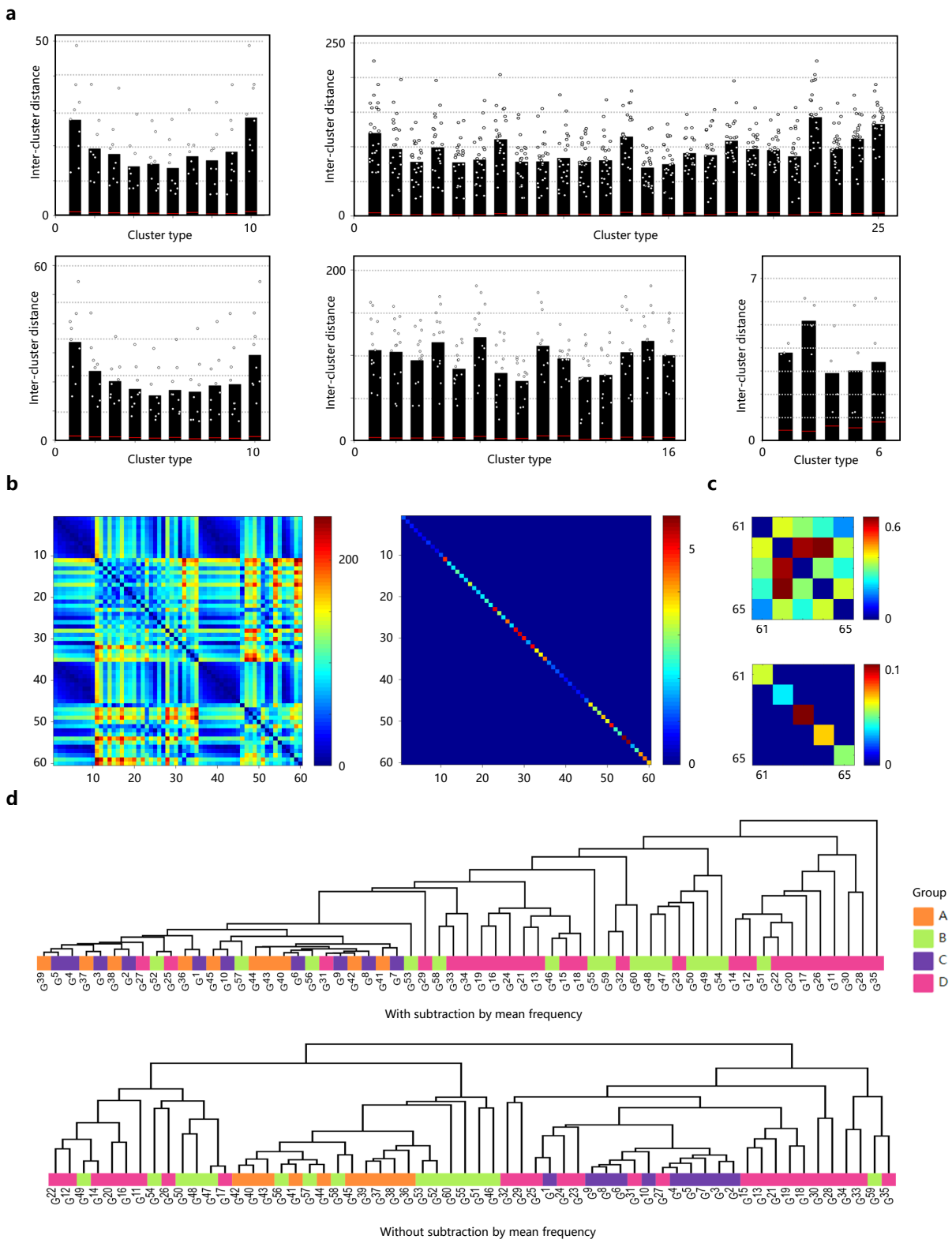
Extended Data Fig. 1 Experimental setup for recording video and audio streams of rats simultaneously. **a**, Schematic illustration of the setup. The video stream is recorded by two lateral cameras fixed on the ground, and the audio stream is recorded by a 4-mic array, which is fixed above the transparent acrylic arena. **b**, Images of the setup and devices.



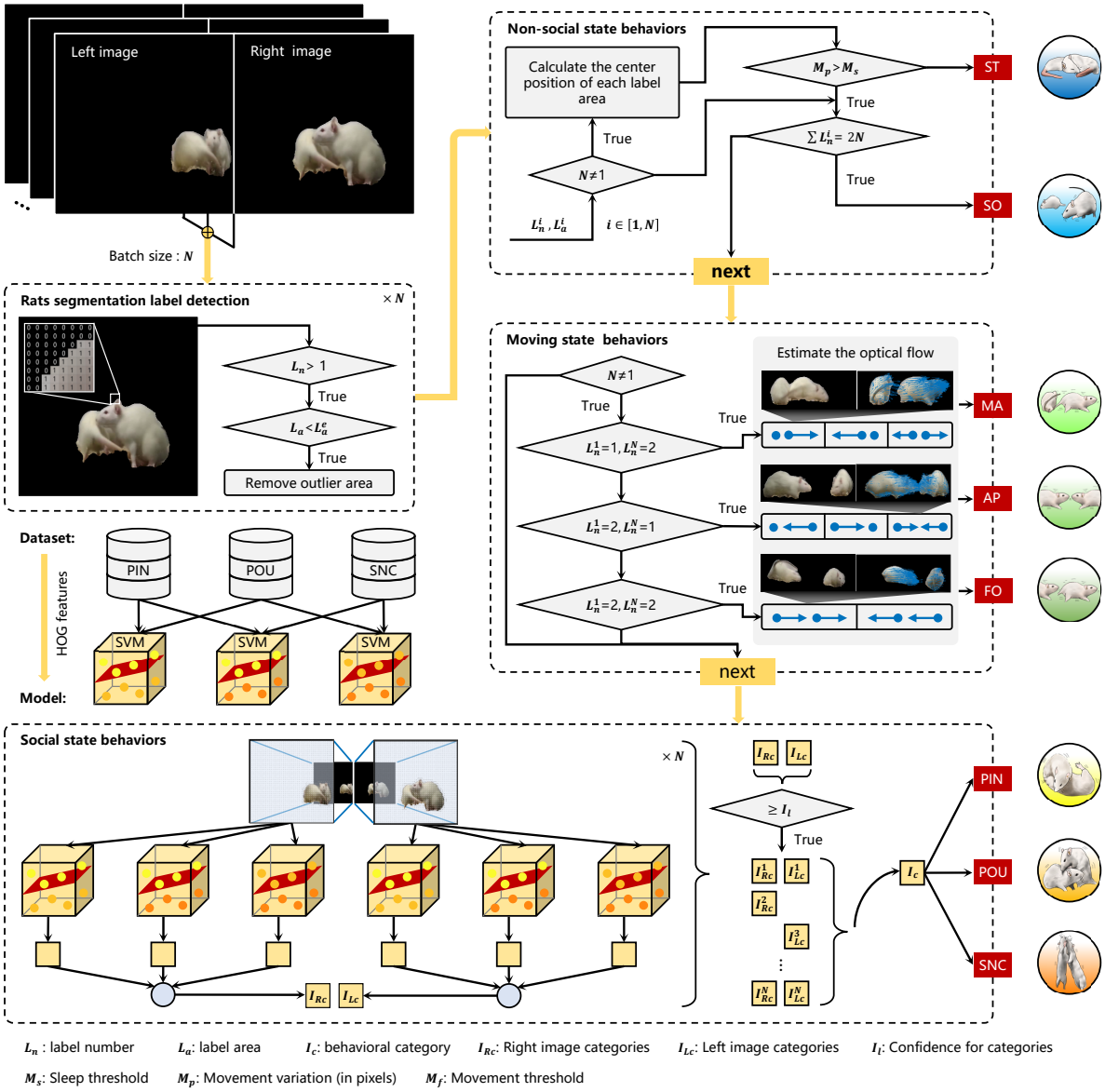
Extended Data Fig. 2 Workflow of the unsupervised clustering algorithm ARBUR:USV. **a**, In the first step (coarse clustering), the recorded audio stream was Fast Fourier Transformed, segmented, screened, and clustered into five groups sequentially; in the second step (refined clustering), USVs were clustered according to their frequency contours (for 50-kHz USVs) or mean frequency and duration (for 22-kHz USVs) within each group. **b**, Feature construction processes for non-step and step 50-kHz USVs.



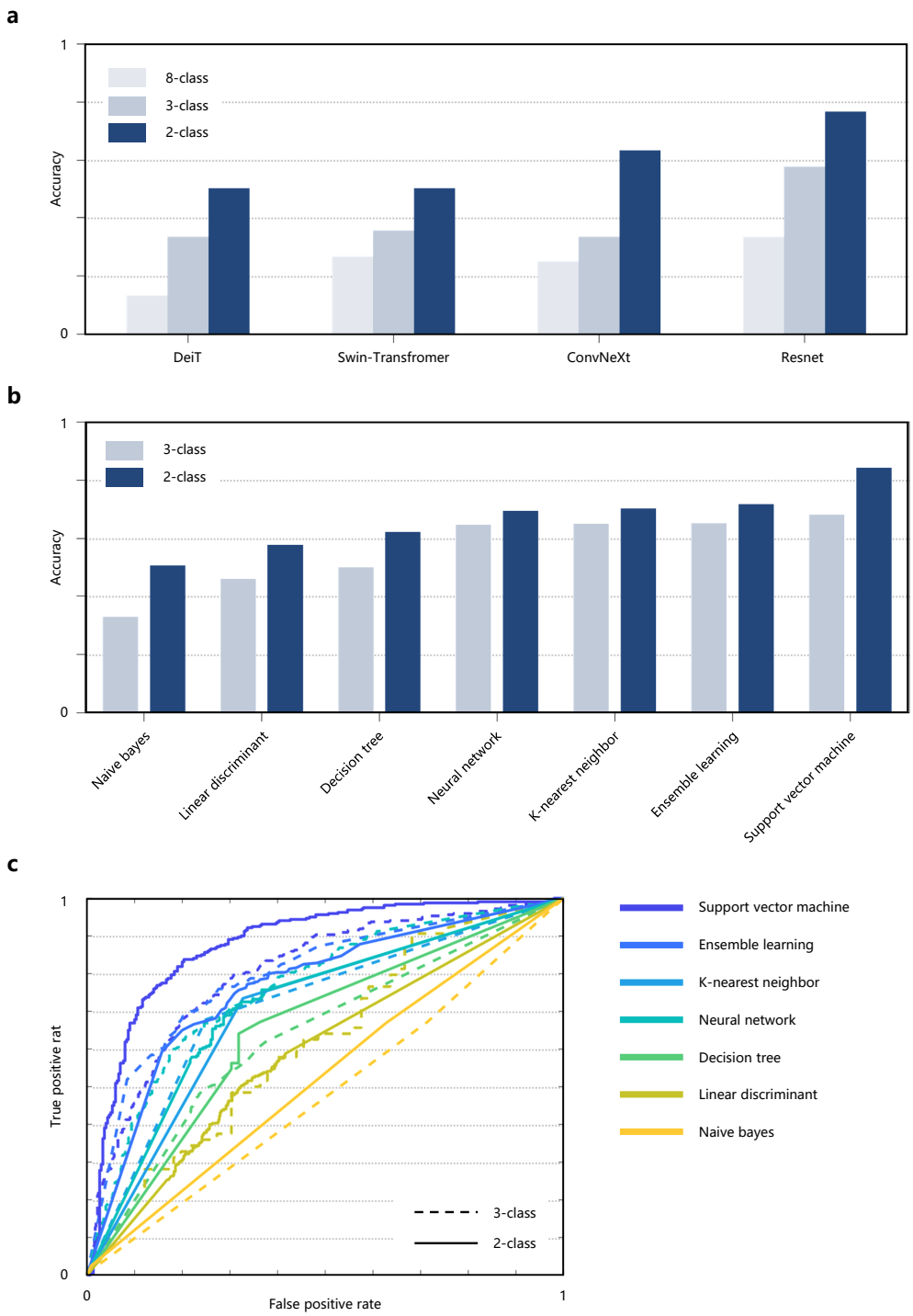
Extended Data Fig. 3 Determining the optimal number of clusters for groups A-E using the elbow method in our dataset. The optimal number for groups A-E are 13, 28, 13, 14 and 8, respectively. Left axis: total within-cluster error (TWCE); right axis: fitting error of two lines to the left and right side of certain number. The lowest fitting error reveals the inflection point of the TWCE-number curve, which is the optimal number of clusters.



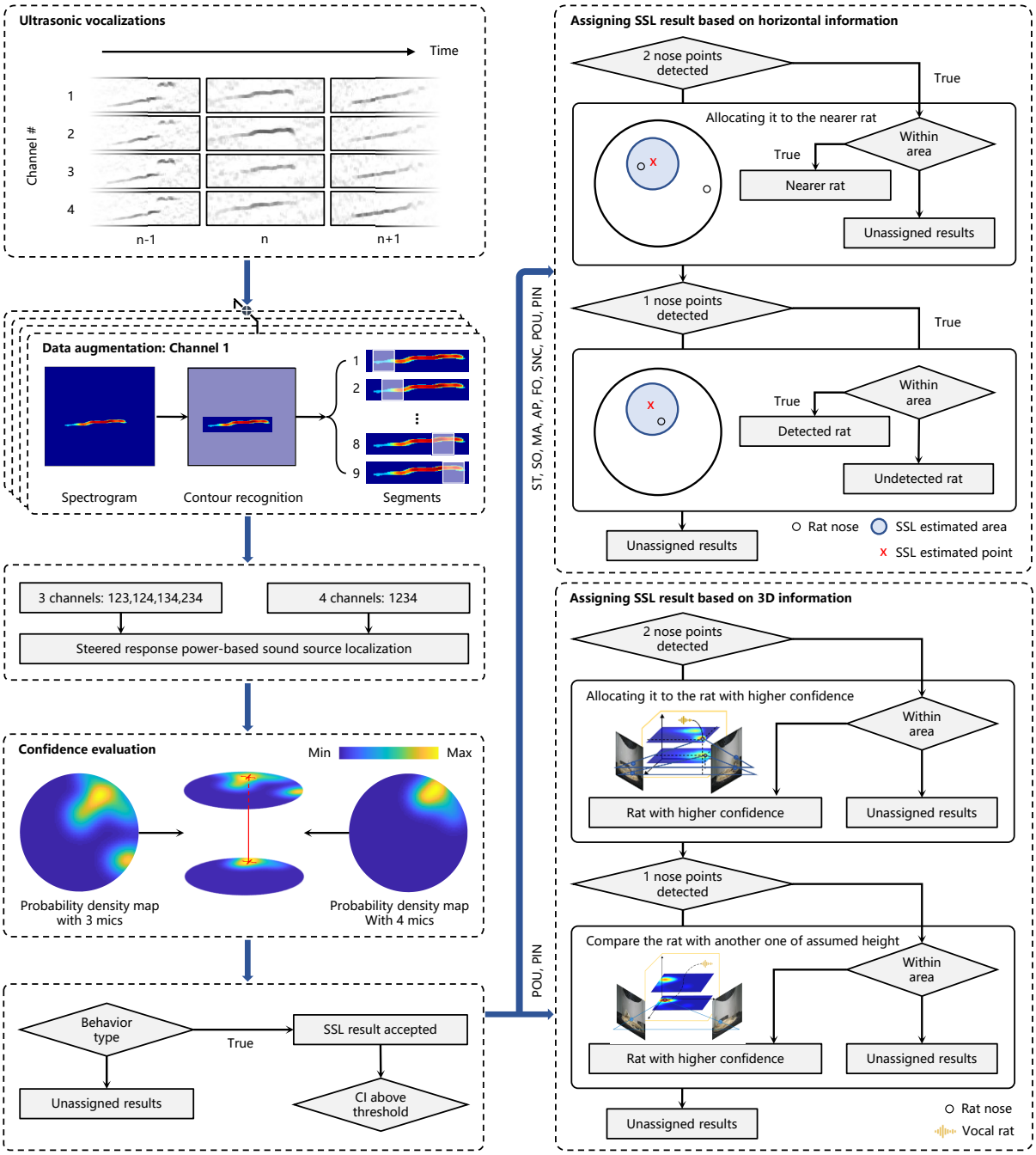
Extended Data Fig. 4 Validating ARBUR:USV. **a**, Inter-cluster distances between two clusters within each group, indicated by black circles. Black bars represent the average inter-cluster distance of this cluster with respect to the rest clusters within the same group. Intra-cluster variances of each cluster are indicated by red lines. **b**, Matrices of inter-cluster distance (left) and intra-cluster variance (right) for clusters 1-60. **c**, Matrices of inter-cluster distance (left) and intra-cluster variance (right) for clusters 61-65. **d**, Dendrograms of hierarchically clustered results for clusters 1-60 with (top) and without (bottom) feature vectors being subtracted by mean frequency.



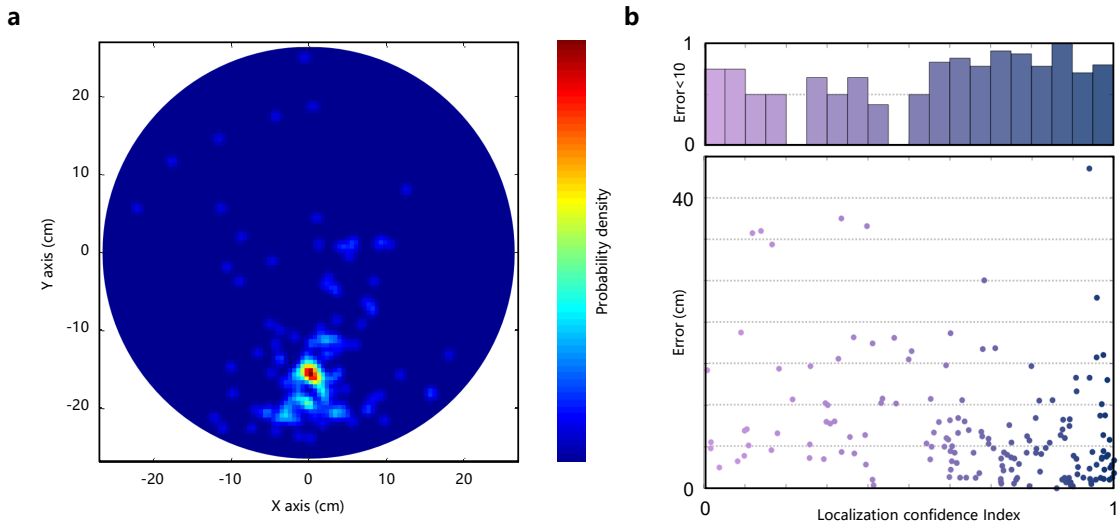
Extended Data Fig. 5 Workflow of ARBUR:Behavior. Rat regions (foreground) of the binocular image sequences (recognized by SegFormer) are labeled by edge detection. Both rat regions and labels are input to a hierarchical algorithm to classify rats' non-social, moving, and social state behaviors sequentially. Non-social state and moving state behavior are detected by displacement and motion optical flow of the rats. A decision tree algorithm based on multiple binary SVM classifiers is used to discriminate easy-to-confuse social state behaviors. ST: still; SO: solitary; MA: moving away; AP: approaching; FO: following; PIN: pinning; POU: pouncing; SNC: social nose contact.



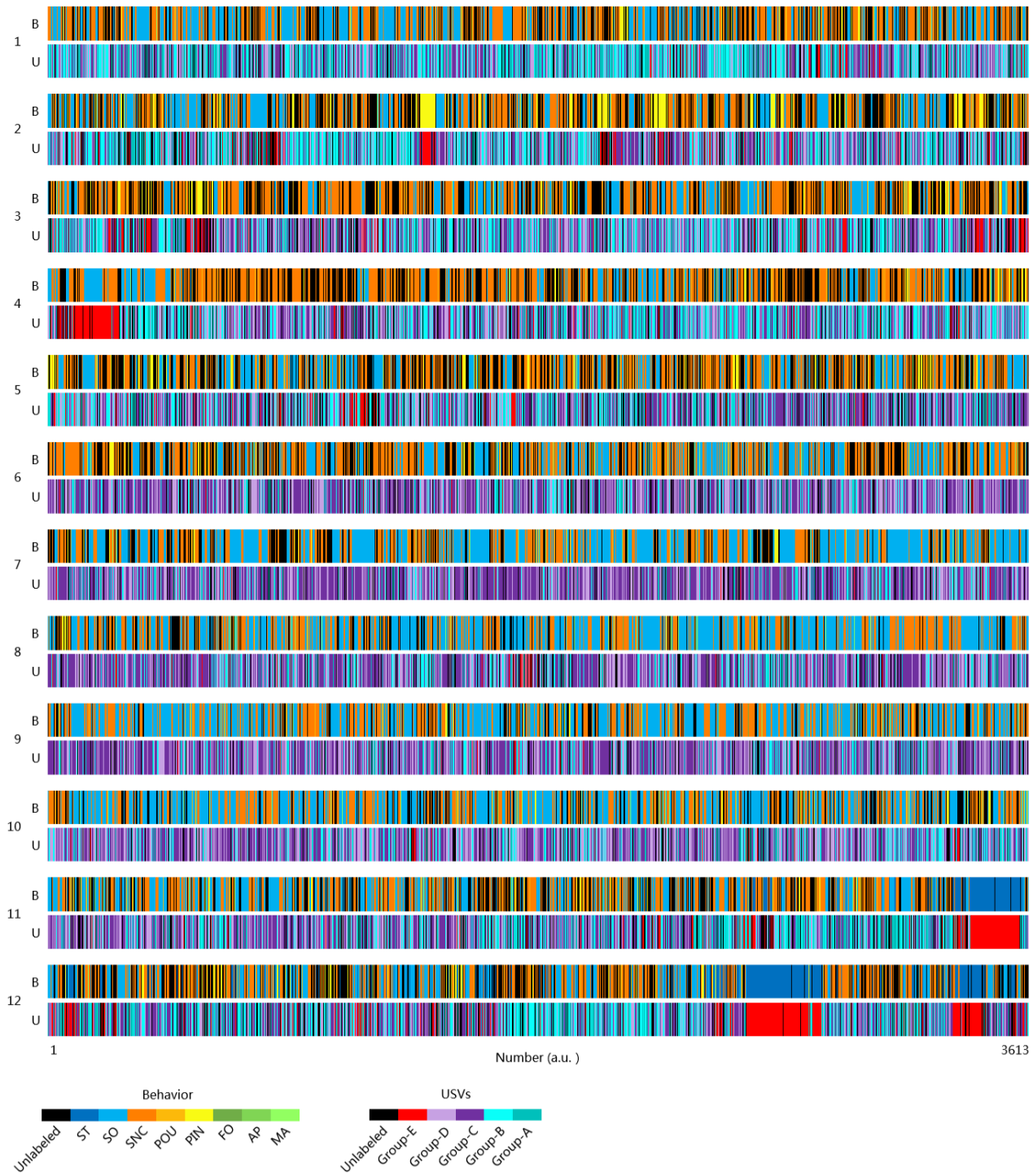
Extended Data Fig. 6 Comparison of algorithms for discriminating the easy-to-confuse social state behaviors. **a**, the average accuracies of four mainstream end-to-end classification networks for detecting 2-class, 3-class, and 8-class behaviors (include all behaviors), respectively. **b**, the average accuracies of seven classic machine learning-based algorithms for 2-class and 3-class behavior discrimination. **c**, The ROC (receiver operating characteristic) curves of the seven algorithms for 2-class and 3-class behavior discrimination.



Extended Data Fig. 7 Workflow of ARBUR:SSL. Segmented ultrasonic vocalizations (USVs) from four channels (mics) were augmented, assembled (with 3 or 4 channels), and input into a steered response power-based sound source localization algorithm. Then, the localization confidence indices (LCIs) were evaluated and the SSL results were calculated, which were combined with 3D coordinates of rats' noses to assign the USVs to the vocal rats by either horizontal or 3D information.



Extended Data Fig. 8 Validating ARBUR:SSL. **a**, The probability density map of the results estimated by ARBUR:SSL. USVs of the rat during still were used for evaluation. The ground truth of the horizontal coordinate of rat nose is (0.5, -14.5), near the maximum of the density map. b, the SSL error (bottom) and the percentage of the results with an error less than 10 cm (top) versus the localization confidence index (LCI).



Extended Data Fig. 9 The behavior (B, top panels) and USVs (U, bottom panels) ethograms of the simultaneously recorded vocal and behavioral repertoires. ST: still; SO: solitary; MA: moving away; AP: approaching; FO: following; PIN: pinning; POU: pouncing; SNC: social nose contact. Groups A and B: high-frequency non-step/step appetitive 50-kHz USVs; Groups C and D: low-frequency non-step/step appetitive 50-kHz USVs; Groups E: aversive 22-kHz USVs.