Gaining Explainability from a CNN for Stereotype Detection Based on Mice Stopping Behavior

Raul Alfredo de Sousa Silva[1], Yasmine Belaidouni[2], Rabah Iguernaissi[1], Djamal Merad[1], and Séverine Dubuisson[1]

[1] Laboratoire d'Informatique et des Systèmes, CNRS, Aix-Marseille University, Marseille, France
{raul-alfredo.de-sousa-silva, rabah.iguernaissi, djamal.merad, severine.dubuisson}@lis-lab.fr
https://im.lis-lab.fr/

[2] Aix-Marseille University, CNRS, IBDM, UMR7288, Marseille, France
yasmine.belaidouni@univ-amu.fr

Abstract. Understanding the behavior of laboratory animals is a key to find answers about diseases and neurodevelopmental disorders that also affects humans. One behavior of interest is the stopping, as it correlates with exploration, feeding and sleeping habits of individuals. To improve comprehension of animal's behavior, we focus on identifying trait revealing age/sex of mice through the series of stopping spots of each individual. We track 4 mice using LiveMouseTracker (LMT) system during 3 days. Then, we build a stack of 2D histograms of the stop positions. This stack of histograms passes through a shallow CNN architecture to classify mice in terms of age and sex. We observe that female mice show more recognizable behavioral patterns, reaching a classification accuracy of more than 90%, while males, which do not present as many distinguishable patterns, reach an accuracy of 62.5%. To gain explainability from the model, we look at the activation function of the convolutional layers and found that some regions of the cage are preferentially explored by females. Males, especially juveniles, present behavior patterns that oscillate between juvenile female and adult male.

Keywords: Automated Behavior Analysis · CNN · Animal Tracking

1 Introduction

Automated Animal Behavior Analysis has become an important active research field during the last decade. It has been furnished with increasingly powerful algorithms by computer vision and machine learning. With the possibility of recording experiments with laboratory animals, the manual analysis became obsolete and unfeasible for a human, once the amount of data is beyond the capacity of processing for an experimenter. In special, manual annotations are prone to experimenter bias and drift throughout the annotation process. Thus, researchers have proposed algorithmic solutions to tackle a number of tasks in the field. The animals studied varied from ants [6] to fruit flies [8], mice [9,11], zebrafish [14], and many others.

Understanding the behavior of laboratory animals is crucial to the discovery of disease mechanisms, neural correlations to behaviors, and efficient treatments, which can be later tested on humans. That is why it is important to have consistent analysis protocols and metrics.

In general, experiments are designed to have a minimal population of animals that ensures statistical significance for tests, but this population is most often not enough to train a deep learning model for detecting differences between individuals. Thus, automatic methods capable of reliably highlighting behavioral differences would represent a new step towards understanding stereotypes in neuroscience. Here, we borrow the definition of stereotype from Battle, which says “a stereotype is a held popular belief about specific social groups or types of individuals” [1]. In this case, sex and age are characteristics separating those “types of individuals”.

Although many proposed algorithmic solutions tackle different problems in fields such as tracking, identity recognition and behavior detection, the recognition of impaired and stereotyped behaviors is still not fully explored. In this work, we propose a pipeline to recognize mice age and sex, using their series of stopping points.

The stopping spots of the mice are related to exploration, food and water consumption and sleeping patterns. Our hypothesis is that distinct stereotypes can be classified based on the temporal sequence of stops, because of the different dominance strategies employed by individuals in each gender and age, which affects exploration, feeding, and sleeping behaviors of individuals.

Using the Live Mouse Tracker (LMT) tracking system [4], we recorded for multiple cages the position and identity, as well as a set of behaviors of four house-grouped mice during 3 days. Then, with a concatenation of single layers Convolutional Networks, we can show that stopping spot sequences actually store relevant information about animal characteristics and can be sufficient to predict the age and sex of individuals.

The following sections are structured as follows: examination of the related work (Section 2), description of the proposed model (Section 3), presentation of the experiments and results (Section 4) and a conclusion (Section 5).

2 Related Work

A variety of approaches were already proposed to distinguish animals of different sexes, ages or disease conditions, mostly based on social interactions. Mice are particularly studied due to the extensive genetic knowledge already discovered by researchers and their similarity with human [13]. Some works were presented with the goal of detecting behaviors in mice, without considering behaviors that could only appear in one stereotype or the change in the way of expressing those behaviors due to stereotypical variations. Burgos-Artizzu *et al.* [2] proposed a classifier to categorize behaviors on a dataset of mice in resident-intruder interactions. The approach is based on spatio-temporal energy and agent trajectories that are used to classify behavior into 13 classes using AdaBoost classifiers. York *et al.* [15] on their turn, presented a pipeline based on postural and displacement features to predict a sequence of behaviors and find neural correlations with movement patterns.

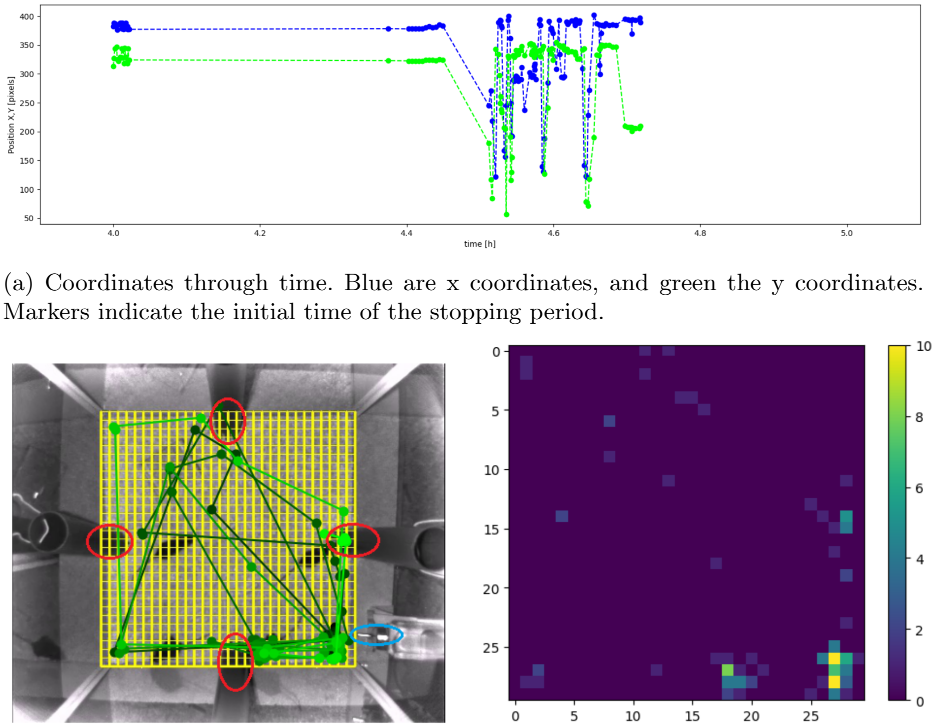
Other works relied on behaviors that could change between stereotypes, in particular sex, but only measured statistical differences at the population level without deepening the discussion about how to identify, at the individual level, the stereotype. Netset *et al.* [12], for example, presented a new tracking system for analysis of social preference dynamics, and highlight that a loss of social preference is observable in females. The time passed investigating the novel individual reduces quickly over time for females, while males keep investigating the new mouse up to 5 minutes after introduction. Zilkha *et al.* [16] studied the behavioral personalities in male and female mice cages to conclude that, although males and females both establish dominance, each sex has its own strategies. Clein *et al.* [3] investigated dominance hierarchies in mixed-sex groups of mice. Through the detection of sequences of behaviors, the authors showed that the submissive mouse has a mechanism to de-escalate social conflicts. We can also find in the literature multiple works where significant statistical difference was found between groups of tested animals, either due to specific gene variation [5,7], or exposition to drugs [10].

Despite these findings, none of these works tried to infer individuals' sex, strain or hierarchy, based on the extracted features. Neither of them intended to unveil distinctive features at the individual level that could identify patterns in the animals related to any of the stereotypes studied. In this work, we propose an explainable neural network model to identify stereotypes in individuals through the series of stopping spots of each animal.

3 Stack of Histograms Feature Space

We recorded 24 male and 24 female mice for 3 days. Each mouse was recorded at two different ages. The first recording happened around 40 days of life, and the second recording happened when mice were about 6 months. We have, then, 96 recordings in total. Animals were housed in groups of four. The recordings were done into a 50x50 cm cage with food and water *ad libitum*. The arrangement of objects in the cage was the same for all experiments. A system called Live Mouse Tracker (LMT) [4] was used to manage recordings and to retrieve positions (center of mass) of mice into the cage at 30 fps. Based on these stopping coordinates and on the definition of stop as the mice velocity being $\leq$ 5 cm/s, we retrieve all stopping spots in which the mouse remained for at least 1 s. This allowed us to have stop detections consistent with what is seen in video recordings.

The set of stops can be modeled as an irregular time series, because the event sequences have different lengths among mice, and they happen at irregular intervals. This is illustrated by Figure 1a. In order to have regularly sampled data and the same dimension for every recording, we compute a stack of 2D histograms of the stopping spots. Firstly, the cage surface is split into a regular grid of size $(N \times N)$, as shown in figure 1b. Also, the stopping coordinates sequences are split in T regular intervals. Then, we count the number of stop events $\mathbb{S}^t \in \mathbb{R}^{N \times N}$ happening in a tile $\{b_x, b_y\} = \{[1 \ldots N], [1 \ldots N]\}$ of the grid for each interval $t = [1 \ldots T]$. Thus, we have a stack of histograms $\mathbf{S} = [\mathbb{S}^1; \ldots; \mathbb{S}^T] \in \mathbb{R}^{T \times N \times N}$ for each recording. These histograms are the features used to distinguish animals and to detect stereotypes. They represent the evolution in time of the spots' preferences of each individual throughout the experiment. We process these sets of histograms as stacked 2D maps, characterizing the behavior of the animal. Each map in the time dimension is treated as a channel of an image. An example of such maps is shown in Figure 1c.



(a) Coordinates through time. Blue are x coordinates, and green the y coordinates. Markers indicate the initial time of the stopping period.

(b) Stop spots over time. The clearer the marker, the further in time. The yellow grid shown is the grid created over the cage coordinates. Feeders are in red, water spot in blue.

(c) Histogram map of the stop spots. The lighter, the higher is the number of stops in this spot. Here, the mouse spent most of its time next to the water spot.

Fig. 1: An example of a 1-hour time series of stops of one mouse. The most often visited spots are the nest and the feeders.

We build a simple CNN architecture. It is composed of 2 convolutional layers receiving as input the histogram maps. The channels of both layers are flattened and concatenated to pass through a single linear layer. Figure 2 illustrates the chosen architecture. The choice of such a network has two goals: avoiding overfitting due to the reduced size of the dataset, and extracting the explainability of the model choices through the convolution kernel outputs assessment. We train the model to classify data into 4 classes: `adult-males`(AM), `juvenile-males`(JM), `adult-females`(AF) and `juvenile-females`(JF).

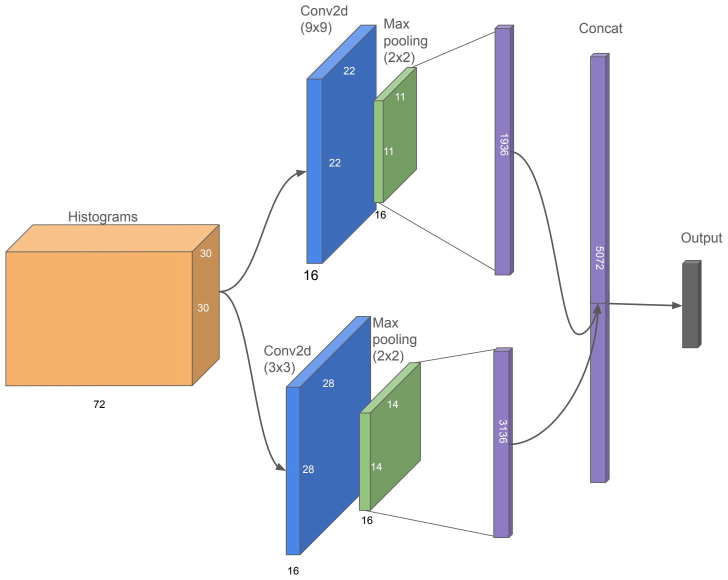


Fig. 2: Proposed CNN architecture to predict sex and age from histogram maps of mice.

4 Experiments

For the construction of the histogram maps, stopping spots series were split into $T = 72$ 1-hour intervals. The number of subdivisions in each dimension of the cage was $N = 30$. Our architecture was composed of two convolutional layers with ReLU as activation function. Each layer has different kernel sizes of (3×3) and (9×9), both with 16 channels. Output of both layers are subsampled with a Max-pooling layer and then, flattened and concatenated to be sent to a single linear layer for classification. Learning was done with a Cross Entropy Loss, using an SGD optimizer, and learning rate was set to $1e^{-3}$.

As we observed that maps from mice of the same cage had strong common traits, we adopted a more challenging learning strategy: dataset is separated per cages and trained using a leave-one(cage)-out cross validation strategy. This avoids the model to rely on common cage features to recognize gender or age, forcing it to learn from generalizing the dynamic features of other cages.

We see the classification results in Table 1. Looking at the table, we observe that `juvenile-males` are misclassified as `juvenile-females`, meaning that they share some patterns. We also realize that the model struggles to separate `adult-males` and `juvenile-males` mice. This seems an indicative that male mice take less time to transit from juvenile age to adulthood, which justifies the same displacement patterns at both ages. On the other hand, females had a distinguishable shift in behavior between the two ages.

		Prediction			
		AM	JM	AF	JF
Labels	AM	22	2	0	0
	JM	9	8	1	6
	AF	0	4	20	0
	JF	0	0	0	24

Table 1: Confusion matrix of sex-age classification. Average accuracy is 77.1%, but accuracy is better among females (91.7%) than males (62.5%).

In order to have a baseline for comparison, we tested k-Nearest Neighbors and SVM classifiers on PCA reduced data. We applied the dimensionality reduction (PCA) over the flattened stack of histograms (D=72*30*30=64800) and kept the first 3 dimensions. The 1-Nearest Neighbor resulted in an accuracy of 0.62, while SVM resulted in an accuracy of 0.72, all of them, considerably lower than the accuracy obtained with the CNN (0.77).

Given these results, we look at the activation maps of the convolution kernels in the network. In Figure 3, we see the average activation maps over all the individuals of each class for some of the 32 convolutional layers channels (16 each). We can see that there's actually no misclassification between `adult-males` and `adult-females` mice because kernels learned quickly how to distinguish them. In Figure 3a it is possible to observe the patterns related to the upper left corner of the cage, where the dome is placed, for male and female mice. Activation maps are exactly opposite one to another. While adult females have a strong predilection for the region where a dome was installed, the adult males seem to correlate much less with the same region. However, In Figure 3c an activation pattern change between `adult-females` and `juvenile-females` can be seen.

We also see that males correlate more with the lower part of the cage, especially between the lower feeder and the water spot. Maps of Figure 3b for both `adult-males` and `juvenile-males` are similar, and response is particularly high in the lower border of the cage, which is also visible in Figures 3c and 3d. In the same figure it is evident that `adult-females` also show a similar pattern but slightly displaced to the left. It's worth noting that no relevant features are observed in `juvenile-males` with relation to other groups, in general, it has either the features of `adult-males` or `juvenile-female` (see figures 3b and

3d), which justify the strong level of misclassification in this class. This suggests that `juvenile-males` mice may take less time to mature, and that their stopping behavior choices at juvenile age are still not fully modulated by sexual differences.

5 Conclusion

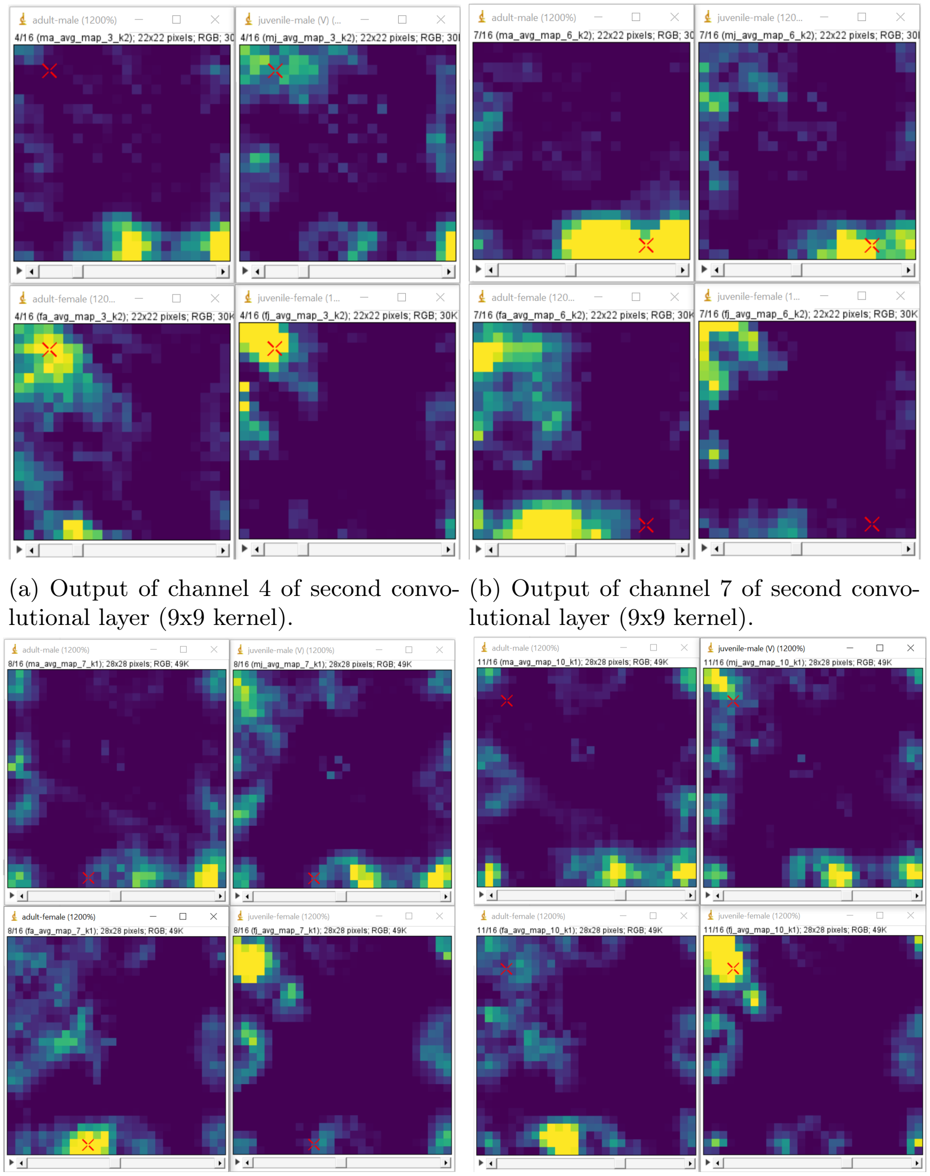
In this work, we proposed a concatenation of single-layer convolutional neural network to identify the sex and age of animals based on stopping spots over time. Our work shows that, while females have more straightforward, identifiable features, stopping spots did not show distinctive features in males over age.

Observation of activation maps helped to understand which regions are more decisive in classification and may help in focusing behavioral analysis on what mice actually do in these regions. The initial cues indicate that nidification and feeding patterns are the ones that distinguish sex, while the age patterns are still a challenge for further work.

As next steps, we propose to keep investigating the stopping spots time series in the cage level, to determine which traits correlate with the hierarchy and main stereotypes of dominant/submissive individuals, as well as the variability of these traits with age and sex.

References

1. Battle, D.E.: Chapter 1 - Communication disorders in a multicultural and global society. In: Battle, D.E. (ed.) Communication Disorders in Multicultural and International Populations (Fourth Edition), pp. 1–19. Mosby, Saint Louis, fourth edition edn. (2012). https://doi.org/10.1016/B978-0-323-06699-0.00010-8
2. Burgos-Artizzu, X.P., Dollár, P., Lin, D., Anderson, D.J., Perona, P.: Social behavior recognition in continuous video. In: CVPR. pp. 1322–1329. IEEE (2012)
3. Clein, R.S., Warren, M.R., Neunuebel, J.P.: Automated behavioral analysis reveals that mice employ a bait-and-switch escape mechanism to de-escalate social conflict (Jan 2024). https://doi.org/10.1101/2024.01.12.575321
4. De Chaumont, F., Ey, E., Torquet, N., Lagache, T., Dallongeville, S., Imbert, A., Legou, T., Le Sourd, A.M., Faure, P., Bourgeron, T., et al.: Real-time analysis of the behaviour of groups of mice via a depth-sensing camera and machine learning. Nature biomedical engineering **3**(11), 930–942 (2019)
5. Ferhat, A.T., Verpy, E., Biton, A., Forget, B., De Chaumont, F., Mueller, F., Le Sourd, A.M., Coqueran, S., Schmitt, J., Rochefort, C., Rondi-Reig, L., Leboucher, A., Boland, A., Fin, B., Deleuze, J.F., Boeckers, T.M., Ey, E., Bourgeron, T.: Excessive self-grooming, gene dysregulation and imbalance between the striosome and matrix compartments in the striatum of Shank3 mutant mice. Frontiers in Molecular Neuroscience **16**, 1139118 (Mar 2023). https://doi.org/10.3389/fnmol.2023.1139118
6. Gal, A., Saragosti, J., Kronauer, D.J.: anTraX, a software package for high-throughput video tracking of color-tagged insects. eLife **9**, e58145 (Nov 2020). https://doi.org/10.7554/eLife.58145


adult-male (1200%)
4/16 (ma_avg_map_3_k2); 22x22 pixels; RGB; 30
juvenile-male (V) (...
4/16 (mj_avg_map_3_k2); 22x22 pixels; RGB; 30
adult-female (120...
4/16 (fa_avg_map_3_k2); 22x22 pixels; RGB; 30K
juvenile-female (1...
4/16 (fj_avg_map_3_k2); 22x22 pixels; RGB; 30K
adult-male (1200%)
7/16 (ma_avg_map_6_k2); 22x22 pixels; RGB; 30
juvenile-male (120...
7/16 (mj_avg_map_6_k2); 22x22 pixels; RGB; 30
adult-female (120...
7/16 (fa_avg_map_6_k2); 22x22 pixels; RGB; 30K
juvenile-female (1...
7/16 (fj_avg_map_6_k2); 22x22 pixels; RGB; 30K
adult-male (1200%)
8/16 (ma_avg_map_7_k1); 28x28 pixels; RGB; 49K
juvenile-male (V) (1200%)
8/16 (mj_avg_map_7_k1); 28x28 pixels; RGB; 49K
adult-female (1200%)
8/16 (fa_avg_map_7_k1); 28x28 pixels; RGB; 49K
juvenile-female (1200%)
8/16 (fj_avg_map_7_k1); 28x28 pixels; RGB; 49K
adult-male (1200%)
11/16 (ma_avg_map_10_k1); 28x28 pixels; RGB; 49K
juvenile-male (V) (1200%)
11/16 (mj_avg_map_10_k1); 28x28 pixels; RGB; 49K
adult-female (1200%)
11/16 (fa_avg_map_10_k1); 28x28 pixels; RGB; 49K
juvenile-female (1200%)
11/16 (fj_avg_map_10_k1); 28x28 pixels; RGB; 49K

(a) Output of channel 4 of second convolutional layer (9x9 kernel).

(b) Output of channel 7 of second convolutional layer (9x9 kernel).

(c) Output of channel 8 of first convolutional layer (3x3 kernel).

(d) Output of channel 11 of first convolutional layer (3x3 kernel).

Fig. 3: Activation maps of a few channels of the convolution layers averaged over animals of the same class. Red crosses highlights the spots where we see considerable difference between activation maps. For each figure (a, b, c, d), AM is on top-left, JM on top-right, AF on bottom-left and JF on bottom-right.

7. Huzard, D., Martin, M., Maingret, F., Chemin, J., Jeanneteau, F., Mery, P.F., Fossat, P., Bourinet, E., François, A.: The impact of C-tactile low-threshold mechanoreceptors on affective touch and social interactions in mice. Science Advances **8**(26), eabo7566 (Jul 2022). https://doi.org/10.1126/sciadv.abo7566
8. Karashchuk, P., Rupp, K.L., Dickinson, E.S., Walling-Bell, S., Sanders, E., Azim, E., Brunton, B.W., Tuthill, J.C.: Anipose: A toolkit for robust markerless 3D pose estimation. Cell Reports **36**(13), 109730 (Sep 2021). https://doi.org/10.1016/j.celrep.2021.109730
9. Le, V.A., Sterley, T.L., Cheng, N., Bains, J.S., Murari, K.: Markerless Mouse Tracking for Social Experiments. eneuro **11**(2), ENEURO.0154–22.2023 (Feb 2024). https://doi.org/10.1523/ENEURO.0154-22.2023
10. Maisterrena, A., De Chaumont, F., Longueville, J.E., Balado, E., Ey, E., Jaber, M.: Female mice prenatally exposed to valproic acid exhibit complex and prolonged social behavior deficits. Progress in Neuro-Psychopharmacology and Biological Psychiatry **131**, 110948 (Apr 2024). https://doi.org/10.1016/j.pnpbp.2024.110948
11. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience **21**(9), 1281–1289 (Sep 2018). https://doi.org/10.1038/s41593-018-0209-y
12. Netser, S., Haskal, S., Magalnik, H., Wagner, S.: A novel system for tracking social preference dynamics in mice reveals sex- and strain-specific characteristics. Molecular Autism **8**(1), 53 (Dec 2017). https://doi.org/10.1186/s13229-017-0169-1
13. Perlman, R.L.: Mouse models of human disease: an evolutionary perspective. Evolution, medicine, and public health **2016**(1), 170–176 (2016)
14. Xu, C., Govindarajan, L.N., Zhang, Y., Cheng, L.: Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups. IJCV **123**, 454–478 (2017)
15. York, R.A., Carreira-Rosario, A., Giocomo, L.M., Clandinin, T.R.: Flexible analysis of animal behavior via time-resolved manifold embedding (Oct 2020). https://doi.org/10.1101/2020.09.30.321406
16. Zilkha, N., Chuartzman, S.G., Sofer, Y., Pen, Y., Cum, M., Mayo, A., Alon, U., Kimchi, T.: Sex-dependent control of pheromones on social organization within groups of wild house mice. Current Biology **33**(8), 1407–1420.e4 (Apr 2023). https://doi.org/10.1016/j.cub.2023.02.039