Analyse du comportement social des souris non biaisé par l'interprétation humaine utilisant un machine learning non-supervisé

Oscar Bauer[1,2,3], Anne-Marie Le Sourd[2], Giacomo Nardi[1], Thomas Bourgeron[2], Jean-Christophe Olivo-Marin[1], Elodie Ey[2], Fabrice de Chaumont[1]

[1] Unité d'Analyse d'Images Biologiques, CNRS UMR 3691, Institut Pasteur, France
[2]Génétique Humaine et Fonctions Cognitives, CNRS UMR 3571, Institut Pasteur, Université Paris-Diderot, Sorbonne Paris Cité, France
[3]Ecole Doctorale Frontières du Vivant (FdV), Programme Bettencourt, France

[1,2,3]oscar.bauer@pasteur.fr,
[1]fabrice.de-chaumont@pasteur.fr, [2]elodie.ey@pasteur.fr

Résumé - Les modèles murins sont largement utilisés dans le but d'étudier les mécanismes des troubles neuropsychiatriques et tester de potentielles stratégies thérapeutiques. L'automatisation du suivi des différences comportementales en interaction sociale est encore limitée pour ces modèles. Nous proposons, dans la présente étude, une nouvelle méthode originale de classification comportementale automatique basée sur du machine learning non-supervisé. Nous avons appliqué la méthode proposée à des souris muté pour *Shank2*, un gène associé à des troubles du spectre autistique. Nous avons validé nos résultats en les comparant à ceux automatiquement extraits par la précédente classification basée sur des règles géométriques. Nous avons découvert sept états comportementaux correspondants à hauteur de 80 à 95 % à ceux du répertoire précédemment défini. Nous avons également trouvé deux comportements inattendus. Nous avons enfin remarqué des différences liées au génotype dans les deux catégories comportementales, locomotion et « faire face à son congénère ».

Abstract - Mouse models are broadly used to study the mechanisms of neuropsychiatric disorders and to test potential treatments. In these models, automation to monitor behavioural differences during social interactions is currently limited. We propose in the present study a new method to conduct automatic behavioural classification, using an original unsupervised machine learning. We applied the proposed method to mice mutated in *Shank2*, a gene associated with autism spectrum disorders. We validated our results by comparing automatically extracted results to rule-based classifier labelling. We discovered seven behavioural states matching from 80 to 95% previous rule-based classification, and two unsuspected behaviours. Interestingly, we also highlighted genotype-related differences in two behavioural categories, namely locomotion and facing the conspecific.

1 Introduction

Les troubles neuropsychiatriques tel que les troubles du spectre autistique, la schizophrénie, l'addiction, ou la dépression affecte profondément la vie sociale des patients. Suivre leur comportement constitue la part principale du diagnostic car les marqueurs physiologiques sont rares. Les modèles murins sont utilisés pour étudier les mécanismes sous-jacents de ces troubles comme pour tester des traitements potentiels. Pour cette raison, les recherches se concentrent sur le comportement - et plus particulièrement les interactions sociales - de ces modèles.

A ce jour, la caractérisation du comportement social des modèles murins repose sur un petit nombre d'événements sociaux sélectionnés lors d'études précédentes. Examiner le comportement de manière plus exhaustive permettra de révéler des effets cachés des traitements pharmacologiques de façon à les raffiner.

Pour être capables de mener de nouvelles mesures quantitatives avec moins de biais expérimentaux, les études comportementales s'orientent actuellement vers l'analyse computationnelle [1]. Les techniques émergentes de traitement de vision par ordinateur permettent de suivre les animaux, transformant les données vidéo en trajectoires. De façon à interpréter le comportement, une description de ces composantes pertinentes - de manière objective et quantitative - est nécessaire. Nous avons donc pour but de développer une classification comportementale automatique.

2 Contraintes

Afin d'extraire automatiquement les comportements, plusieurs techniques non-supervisées ont été développées depuis 2001 [2-6]. Parmi celles-ci, Braun *et al* [3] ont analysé le comportement de la mouche à viande. Dans ce but, ils ont utilisé une classification globale, prenant en compte toutes les informations disponibles. Cela permet une définition des événements moins anthropomorphique qu'une classification humaine manuelle. Mais, en contrepartie, cette analyse se limite à décrire le comportement comme une succession d'événements. Elle part du postulat qu'un seul événement existe en un temps donné, ce qui n'est pas adapté à la complexité du comportement social. Dans nos expériences, les animaux peuvent produire différentes actions indépendantes simultanément. Par exemple ils peuvent être en mouvement ou bien à l'arrêt et dans le même temps flairer ou non leur congénère.

Les seules méthodes capables de décrire ces événements simultanés sont les classifications basées sur règles [7-9]. Celles-ci filtrent les données selon des indices géométriques extraits des trajectoires, tels que la position relative des animaux, leur distance et leur vitesse. Ces règles, aussi appelées répertoire d'événements [8], permettent de décrire des comportements simultanés, mais n'extraient que les comportements précédemment définis par des experts. Ainsi ces classifications dépendent aussi des seuils qu'ils ont déterminés.

Cette catégorisation est anthropomorphique et empêche l'observateur de découvrir des comportements inattendus qui pourraient être clefs dans la différenciation du phénotype social au sein de groupes d'animaux.

Dans cet article, nous proposons de combiner les avantages des deux approches : une classification non-supervisée capable de prendre en compte des événements simultanés, basée sur des informations simples extraites des trajectoires des animaux.

3 Classification comportementale non-supervisée

Nous utilisons une méthode de suivi qui renseigne la position de la tête et de la base de la queue des animaux. Nous utilisons ces données en héritage de notre précédent traqueur [8] comme il a été conçu pour suivre spécifiquement ces parties de l'animal qui sont connus en tant qu'éléments clefs de l'interaction des animaux (la base de la queue correspondant à la région ano-génitale). Nous avons cependant conçu cette méthode pour qu'elle puisse prendre en compte des informations plus fouillés.

Nous détaillons tout d'abord comment calculer ces informations. Celles-ci sont ensuite traités par une analyse en composante indépendantes (ACI) qui extrait des descripteurs indépendants. Chaque composante indépendante (CI) offre une combinaison linéaire de ces informations. Nous utilisons par la suite un modèle de mélange gaussien (MMG) unidimensionnel afin de générer pour chaque IC une classification de comportements exclusifs (tel qu'animaux en contact, ou à distance). Le fait de diviser les données en descripteurs indépendants donne lieu à une classification différente pour chacun d'entre eux. Ces classifications multiples permettent d'affecter chaque point de temps à plusieurs catégories d'évènements comportementaux qui sont donc, entre eux, non-exclusifs.

3.1 Calcule des informations

Les résultats du suivi [8] sont les coordonnées bidimensionnelles de la tête et de la base de la queue, pour deux souris, à chaque point de temps. Les CIs, extraites par l'ACI, sont orthogonales. Nous maximisons donc l'information extraite de ces sorties en choisissant toute information, sans considération des corrélations entre elles.

Ces variables consistent en des mesures de distances et angles, relevés sur chaque image, ou entre images successives pour les informations dynamiques tels que des vitesses. Le tableau 1 récapitule les informations extraites des coordonnées des têtes et queues des deux animaux. Ces informations sont individuelles (1 à 4, i.e. longueur de l'animal, vitesse), ou sociales (5 à 13, i.e. distances, angles entre animaux). Nous alimentons alors l'ACI avec le set complet d'informations mesurées.

3.2 Extraction des composantes indépendantes

Certaines des informations précédemment calculées sont fortement corrélées. Nous extrayons donc des descripteurs indépendants en appliquant une ICA globale, prenant en compte toutes les informations, sur l'ensemble de tous les points de temps de toutes les expériences réunies. Nous utilisons l'algorithme fastICA [10]. Ce dernier applique une rotation orthogonale des données pré-blanchies qui maximise une mesure de non-gaussianité.

Tab 1 : Jeu d'informations basé sur les trajectoires, individuelles et sociales.

Dans les formules, H et T se renvoient respectivement aux points des têtes et bases de queues. Les indices A et B indiquent si le point appartient à l'animal étudié A ou à son congénère B, t étant le numéro de l'image analysée. ||.|| est la norme euclidienne.

1 longueur du vecteur animal A $\left\| \overrightarrow{T_t^A H_t^A} \right\|$

2 vitesse de la tête de l'animal A $\left\| \overrightarrow{H_t^A H_{t+1}^A} \right\|$

3 vitesse de la queue de l'animal A $\left\| \overrightarrow{T_t^A T_{t+1}^A} \right\|$

4 changement de direction du vecteur de l'animal A $\overset{\frown}{T_t^A H_t^A T_{t+1}^A H_{t+1}^A}$

5 distance entre les têtes de A et B $\left\| \overrightarrow{H_t^A H_t^B} \right\|$

6 distance entre la tête de A et la queue de B $\left\| \overrightarrow{H_t^A T_t^B} \right\|$

7 distance entre la queue de A et la tête de B $\left\| \overrightarrow{T_t^A H_t^B} \right\|$

8 distance entre les queues de A et B $\left\| \overrightarrow{T_t^A T_t^B} \right\|$

9 vitesse de A vers la position actuelle de B $\left\| \overrightarrow{\frac{T_t^A + H_t^A}{2} \frac{T_t^B + H_t^B}{2}} \right\| - \left\| \overrightarrow{\frac{T_{t+1}^A + H_{t+1}^A}{2} \frac{T_t^B + H_t^B}{2}} \right\|$

10 vitesse de A vers la prochaine position de B $\left\| \overrightarrow{\frac{T_t^A + H_t^A}{2} \frac{T_{t+1}^B + H_{t+1}^B}{2}} \right\| - \left\| \overrightarrow{\frac{T_{t+1}^A + H_{t+1}^A}{2} \frac{T_{t+1}^B + H_{t+1}^B}{2}} \right\|$

11 angle entre le vecteur de A et le vecteur du milieu de A vers la tête de B $\overset{\frown}{\overrightarrow{T_t^A H_t^A} \overrightarrow{\frac{T_{t+1}^A + H_{t+1}^A}{2} H_{t+1}^B}}$

12 angle entre le vecteur de A et le vecteur du milieu de A vers la queue de B $\overset{\frown}{\overrightarrow{T_t^A H_t^A} \overrightarrow{\frac{T_{t+1}^A + H_{t+1}^A}{2} T_{t+1}^B}}$

13 angle entre les vecteurs de A et B $\overset{\frown}{\overrightarrow{T_t^A H_t^A} \overrightarrow{T_t^B H_t^B}}$

Dans cette analyse, le nombre de CIs extraites doit être ajusté. Les CIs doivent respecter les deux critères d'orthogonalité et de non-gaussianité pour être considérés comme descripteurs sources pertinents du comportement. Aussi, pour avoir les meilleurs CIs, nous devons restreindre le nombre d'axes produits par l'ACI en extrayant moins de CIs que le nombre originel de variables.

Pour cela nous devons trouver un nombre de CIs qui correspond à des sources réelles. Dans ce nombre de dimensions, l'extraction d'un CI supplémentaire ajoute une nouvelle source, affectant peu les sources extraites précédemment. Au-dessus d'un certain nombre de CIs, tout ou partie des composantes sont alors réarrangés pour en extraire une supplémentaire.

Nous recherchons donc l'extraction du nombre maximum de CIs qui préserve les composantes trouvées lorsqu'elles étaient extraites en plus petit nombre.

L'ACI extrait un set *CI(n)* de *n* CIs. CI_n^i Désigne l'élément *i* de *CI(n)*. Nous cherchons à estimer dans quelle proportion les éléments de *CI(n-1)* sont préservés dans leur forme d'origine dans *CI(n)*. Pour considérer qu'un CI de *CI(n-1)* est bien préservé, il doit être fortement corrélé avec un des *IC(n)* et le moins possible avec les autres. Dans ce but nous avons mis au point un critère pour évaluer la préservation des CIs. La préservation d'un CI_{n-1}^i $CI(n-1)$ en *CI(n)* est estimée par la fonction :

$$F(i,n) = \left[\left[2 \max_{j(1,n)} \left(Cor\left(IC_{n-1}^i, IC_n^j \right) \right) \right] - \sum_{j=1}^{n} Cor\left(IC_{n-1}^i, IC_n^j \right) \right]$$

Nous cherchons à sélectionner le nombre n de CIs préservant le plus *IC(n-1)*. Dans ce but, l'ACI est utilisé pour générer un nombre croissant de CIs, en commençant par 2. Le nombre choisi sera le premier maximum local de la moyenne : $C(n)=\frac{\sum_{i=1}^{n-1} F(i,n)}{n-1}$

3.3 Classification par modèle de mélange gaussien

Nous traitons dorénavant les CIs précédemment générés séparément. Afin de décrire des expressions comportementales qualitatives, tous les points de temps doivent être classifiés d'après ce descripteur quantitatif. La distribution d'un CI étant non-gaussienne, nous choisissons une classification par MMG unidimensionnelle [11], Ce qui nous assure de modéliser la CI par plus d'une distribution gaussienne (DG) et de générer plus d'une catégorie comportementale. Le MMG est ajusté par maximisation de la vraisemblance $L=\prod_{i=1}^{T}\sum_{j=1}^{N} p_j N\left(x_i,\mu_j,\sigma_j^2\right)$ où T et N sont respectivement le nombre de points de données et de DGs et $p_j N\left(x_i,\mu_j,\sigma_j^2\right)$ la densité de probabilité pondérée de x_i dans la *j-éme* DG. Pour mener ce calcule nous utilisons l'algorithme de maximisation de l'espérance.

De façon à éviter le sur-ajustement au jeu de données, nous restreignons le nombre de DGs mélangées dans le modèle. Le but étant de définir un modèle qui permette une classification pertinente, nous sélectionnons le nombre de distributions qui s'ajuste le mieux aux données tout en mélangeant le moins possible les classes. Pour cela, plusieurs MMG sont ajustés avec un nombre croissant de DGs en commençant par 2. Nous sélectionnons le modèle produit en maximisant la fonction $\prod_{i=1}^{t} p_{k(i)} N\left(x_i,\mu_{k(j)},\sigma_{k(j)}^2\right)$, où *k(i)* désigne le numéro de la DG qui donne la plus haute densité de probabilité pour x_i. Cela pénalise les modèles sur la base du chevauchement de leurs DGs et restreint leur nombre. Pour chaque CI, le MMG produisant le premier maximum local de ce critère est retenu.

Le point d'égalité des probabilités de deux DGs est défini comme seuil segmentant les comportements le long de la CI (Fig 1B). Pour chaque CI, tous les points de temps sont classés en un certain nombre d'événements comportementaux. Un animal exhibera donc autant d'états comportementaux à chaque point de temps que le nombre de CIs sélectionne automatiquement.

3.4 Montage automatique de vidéos représentant une compilation d'événements semblables.

Grâce aux résultats du MMG, nous retrouvons les événements exclusifs et retournons aux mesures pour produire la vidéo correspondant à cet axe de la classification. Ces événements pouvant être brefs (moins d'une seconde à quelques-unes), nous avons créé un programme dans le logiciel d'analyse d'image Icy [12], qui édite automatiquement les vidéos source. Cela permet d'obtenir le résultat final de la classification. Pour chaque seuil de chaque CI, nous procédons à deux éditions. La première regroupe toutes les séquences durant lesquelles la valeur sur la CI est inférieure au seuil. La seconde regroupe le reste, celles avec une valeur supérieure au seuil. Ces vidéos sont donc une succession de nombreuses séquences où l'animal présente à chaque fois le même comportement.

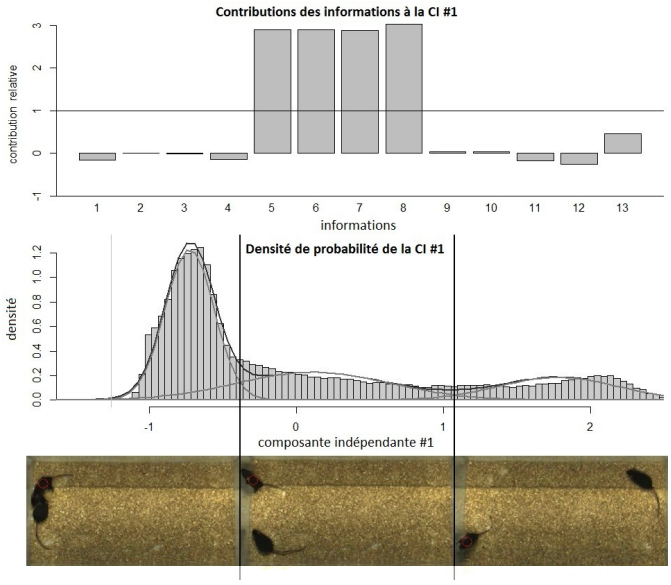

Fig 1 :
A. Contributions des informations relatives à la première CI.
B. Densité de la distribution au long de la première CI.
C. Exemple d'images montées automatiquement
(2 souris – contraste augmenté pour raisons d'impression).

Nous plaçons ces deux vidéos côte à côte (Fig 1C) pour aider un expert qui visionne plusieurs fois le même événement afin d'être capable d'affecter un nom « compréhensible par les humains » à chaque événement trouvé par le programme. Nous donnons aussi accès aux mélanges des informations impliquées dans les CIs (Fig 1A).

4 Résultats expérimentaux

4.1 Données biologiques

Nous avons étudié le comportement social de 13 souris *Shank2*$^{-/-}$ et 16 souris wild-type, toutes femelles. Les expériences comportementales ont été approuvées par le comité d'éthique CETEA Institut Pasteur n°89. La souris sujet était placé dans la cage de test 30 minutes avant qu'une femelle inconnue soit introduite [13-14]. Leur interaction de 4 minutes a été filmée en vue zénithale. Les souris ont été suivies avec Mice Profiler [8]. Nous avons obtenu les coordonnées (x,y) de la tête et de la base de la queue des deux souris à la cadence de 15 images par seconde.

4.2 Implémentation

L'ACI et le MMG ont été calculés avec R, en utilisant respectivement les fonctions icafast du package ica et Mclust du package mclust.

Sur la base des contraintes précédemment décrites, nous avons utilisé l'ACI pour extraire n=7 CIs, chacun d'eux étant segmenté par 1 à 2 seuils. Nous nous sommes concentrés sur les composantes comportementales fréquemment exprimées, en négligeant les seuils qui isolaient moins de 2,5% des images.

Nous avons aussi procédé à une cross-validation, sur 10 sous-échantillons du jeu de données d'origine. L'algorithme a calculé des sets de 7 CIs, similaires à ceux générés d'après le jeux complet, chacun segmenté par le même nombre de seuils de valeurs similaires.

4.3 Validation de la méthode : comparaison à Mice Profiler

Nous avons trouvé 7 comportements homologues aux événements élémentaires présents dans le répertoire de Mice Profiler [8]. (coefficients de corrélations entre parenthèses) Contact (95%), A derrière B (93.2%), A fait face à B (92.9%), A évite B (87.2%), A suit B (80.7%), côte à côte dans le même sens (85.9%) et côte à côte en sens opposés (83.3%).

Tab 2 : Set de comportements définis par la nouvelle méthode.

A renvoie à l'animal sujet et B à son congénère, pour chaque CI le seuil S ou les seuils S1 etS2 permettent de localiser les catégories qui sont en dessous ou au dessus de ceux-ci.

Nom du comportement d'après l'édition	CI	catégories
contact	1	<S1
distance intermédiaire	1	$[S1, S2]$
longue distance	1	>S2
A derrière B	2	<S
A fait face à B	3	>S
A s'éloigne de B	4	<S
A va vers B	5	<S
Axes têtes-queues parallèles dans la même sens	6	<S1
Axes têtes-queues parallèles en sens opposés	6	>S2
Trajectoires parallèles dans la même direction	7	<S1
Trajectoires parallèles en directions opposées	7	>S2

Deux comportements additionnels ont été détectés avec la 7^{éme} CI (Table 2), voire résultats biologiques. L'algorithme permet donc de discriminer des comportements qui n'étaient pas attendus précédemment.

4.4 Résultats biologiques

Des différences significatives liées au génotype ont été observées pour 4 événements comportementaux (table 2). Les femelles *Shank2*$^{-/-}$ passèrent significativement : plus de temps à une distance intermédiaire de leur congénère ; moins de temps à lui faire face ; plus de temps à venir vers lui ; et plus de temps à s'en éloigner en comparaison avec des femelles wild-types des mêmes portées (tests de Mann-Whithney, avec correction de Bonferroni pour 13 tests, P-values respectives de : 0.033 ; 0.016 ; 1.7×10^{-5} ; 6.2×10^{-3}). L'augmentation des comportements locomoteurs est cohérent avec l'hyperactivité diagnostiquée [13] et le fait de moins faire face au congénère correspond au manque d'intérêt social [14].

Nous avons aussi révélé un nouveau comportement : le fait que les animaux marchent dans des trajectoires parallèles, avec ou sans contact. Ce comportement relié aux trajectoires parallèles n'a aucun événement homologue décrit dans Mice Profiler. Ces comportements étaient donc inattendus avant d'utiliser l'algorithme non-supervisé pour les découvrir et ils seront étudiés plus en détails.

5 Conclusion

Dans cette étude nous avons détaillé un nouvel algorithme capable de classifier des comportements sociaux sans supervision humaine. Par rapport aux travaux précédents, nous avons ajouté la capacité, pour la classification, de définir différents comportements indépendants exprimés simultanément.

Des états comportementaux plus complexes peuvent dorénavant être décrits par la co-occurrence de ces comportements.

Nous avons aussi montré que cette classification révélait de nouveaux événements qui n'étaient pas définis auparavant, comme les événements de trajectoires parallèles. Ces nouveaux comportements peuvent dorénavant être utilisés comme nouveaux indicateurs de la restauration du comportement social chez les souris *Shank2*$^{-/-}$ femelles, et leur étude chez d'autres modèles de troubles neuropsychiatriques sera probablement informative.

Dans nos futurs travaux, nous augmenterons le nombre d'informations traitées pour révéler plus de comportements. Nous espérons aussi que cette méthode bénéficiera d'être appliquée à d'autres jeux de données composés d'autres types d'informations, en particulier les futurs logiciels de suivi automatique capables d'extraire des informations plus détaillées.

6 Références

[1] S.R. Egnor, and K. Branson, "Computational Analysis of Behaviour." *Annual review of neuroscience 0*, pp. 217-236, 2016.

[2] L. Zelnik-Manor, and M. Irani, "Event-based analysis of video." *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Vol. 2*, pp. II-123, 2001.

[3] E.Braun,B. Geurten, and M. Egelhaaf, "Identifying prototypical components in behaviour using clustering algorithms." *PloS one 5.2*, e. 9361, 2010.

[4] A.B. Wiltschko, M.J. Johnson, G. Iurilli, R.E. Peterson, J.M. Katon, S.L. Pashkovski, V.E. Abraira, R.P. Adams, and S.R. Datta, "Mapping sub-second structure in mouse behaviour." *Neuron 88.6*, pp. 1121-1135, 2015.

[5] G.J. Berman, D.M. Choi, W. Bialek, and J.W. Shaevitz, "Mapping the stereotyped behaviour of freely moving fruit flies." *Journal of The Royal Society Interface 11.99*, 2014.

[6] U. Kilbaite, G.J. Berman, J. Cande, D.L. Stern, and J.W. Shaevitz, "An unsupervised method for quantifying the behaviour of interacting individuals." *arXiv:1609.09345*, 2016.

[7] H. Dankert, L. Wang, E.D. Hoopfer, D.J. Anderson, and P. Perona, "Automated monitoring and analysis of social behaviour in Drosophila." *Nature methods 6.4*, pp. 297-303, 2009.

[8] F. de Chaumont, R.D.S. Coura, P. Serreau, A. Cressant, J. Chabout, S. Granon, and J.C. Olivo-Marin, "Computerized video analysis of social interactions in mice." *Nature methods 9.4*, pp. 410-417, 2012.

[9] A. Weissbrod, A. Shapiro, G. Vasserman, L. Edry, M. Dayan, A. Yitzhaky, L. Hertzberg, O. Feinerman, and T. Kimchi, "Automated long-term tracking and social behavioural phenotyping of animal colonies within a semi-natural environment." *Nature communications 4*, 2013.

[10] A. Hyvärinen, and O. Erkki, "A fast fixed-point algorithm for independent component analysis." *Neural computation 9.7*, pp. 1483-1492, 1997.

[11] D. Lipkind, A. Sakov, N. Kafkafi, G.I. Elmer, Y. Benjamini, and I. Golani, "New replicable anxiety-related measures of wall vs. center behaviour of mice in the open field." *Journal of Applied Physiology 97.1*, pp. 347-359, 2004.

[12] F. de Chaumont, S. Dallongeville, N. Chenouard, N. Hervé, S. Pop, T. Provoost, T. Lagache, ... and J.C. Olivo-Marin, "Icy: an open bioimage informatics platform for extended reproducible research." *Nature methods 9.7*, pp. 690-696, 2012.

[13] M.J. Schmeisser, E. Ey, S. Wegener, J. Bockmann, A.V. Stempel, A. Kuebler, … and D. Balschun, "Autistic-like behaviours and hyperactivity in mice lacking ProSAP1/Shank2." *Nature 486.7402*, pp. 256-260, 2012.

[14] A.T. Ferhat, A.M. Le Sourd, F. de Chaumont, J.C. Olivo-Marin, T. Bourgeron, & E. Ey, "Social communication in mice–Are there optimal cage conditions?" *PloS one 10.3*, e. 0121802, 2015.