nature neuroscience

RESOURCE

https://doi.org/10.1038/s41593-020-00706-3

Revealing the structure of pharmacobehavioral space through motion sequencing

Alexander B. Wiltschko[1,2], Tatsuya Tsukahara[1], Ayman Zeine[1], Rockwell Anyoha[1], Winthrop F. Gillis[1], Jeffrey E. Markowitz[1], Ralph E. Peterson[1], Jesse Katon[1], Matthew J. Johnson[1,3] and Sandeep Robert Datta[1] ✉

Understanding how genes, drugs and neural circuits influence behavior requires the ability to effectively organize information about similarities and differences within complex behavioral datasets. Motion Sequencing (MoSeq) is an ethologically inspired behavioral analysis method that identifies modular components of three-dimensional mouse body language called 'syllables'. Here, we show that MoSeq effectively parses behavioral differences and captures similarities elicited by a panel of neuroactive and psychoactive drugs administered to a cohort of nearly 700 mice. MoSeq identifies syllables that are characteristic of individual drugs, a finding we leverage to reveal specific on- and off-target effects of both established and candidate therapeutics in a mouse model of autism spectrum disorder. These results demonstrate that MoSeq can meaningfully organize large-scale behavioral data, illustrate the power of a fundamentally modular description of behavior and suggest that behavioral syllables represent a new class of druggable target.

Animals interact with the world through freely expressed behaviors whose content reflects sensory information, prior experience and internal state. The brain composes these complex patterns of action by concatenating stereotyped motifs of movement into meaningful sequences[1,2]. Characterizing how naturalistic behaviors unfold over time—and how the content of behavior is altered by experimental manipulations or disease—offers a powerful lens to better understand how genes, receptors and neural circuits collaborate to enable brain function.

However, two practical challenges have hindered the effective use of naturalistic behaviors in the laboratory to understand the brain[3,4]. The first relates to measuring behavior, which in unrestrained animals often includes complex changes in pose and position. Recent technical advances are beginning to address this challenge, including the development of deep-learning-based platforms (such as LEAP, DeepLabCut and DeepPoseKit) that accurately track user-specified points in behavioral videos, of depth cameras that visualize mice in three dimensions (3D) as they freely behave and of miniaturized accelerometers that capture multi-axis head- or body-motion data[5–11].

The second challenge relates to understanding behavioral data. Traditionally, behavioral neuroscience has relied on summary statistics that are thought to reflect underlying neural or psychological processes of interest. Researchers studying anxiety, for example, often place mice in the open field and then take the number of center entries as a surrogate for their anxiety states[12,13]. Similarly, the total time struggling in a vat of water is taken to reflect the level of helplessness of a mouse[12,13]. Even under highly controlled conditions, however, these metrics tend to be unreliable (across mice, days and laboratories), and their narrow dynamic range obscures drug-specific behavioral effects, which prevents, for example, different drugs belonging to the same pharmacological class from being distinguished[14,15].

These limitations have prompted interest in developing unsupervised, data-driven methods that can discover the underlying structure of behavior and can characterize how that structure is altered by experimental interventions such as gene mutations or drug treatments[4,16,17]. We have recently developed one such method, referred to as Motion Sequencing (MoSeq), whose underlying model was inspired by the ethological insight that behavior is comprised of components that are organized into probabilistic sequences[2,9,18,19]. MoSeq combines 3D imaging and unsupervised machine learning to identify a set of reused and stereotyped subsecond 3D behavioral motifs out of which behavior is composed within a given experiment (for example, rears, turns, head-bobs, among others, referred to herein as behavioral "syllables"), as well as the statistics that govern how syllables transition from one to another over time (that is, behavioral "grammar"). Importantly, MoSeq recognizes syllables and grammar on the basis of the latent structure present in the behavioral data; it automatically learns the number and identity of behavioral syllables within any dataset, thereby enabling it to flexibly characterize new or unexpected patterns of behavior without human supervision.

While MoSeq was designed to identify repeated patterns in behavioral data, nothing in the MoSeq algorithm is explicitly optimized to distinguish different patterns of behavior or to identify behavioral relationships. To assess whether MoSeq can usefully organize large-scale behavioral data, here, we generate behavioral diversity in hundreds of individual mice using pharmacology and then quantify the ability of MoSeq (and, as a comparator, traditional behavioral metrics) to predict information about drug identity, dose and class. These experiments reveal that MoSeq can accurately predict (and therefore distinguish) which of the 30 drug–dose pairs any one of ~700 mice received while simultaneously maintaining key information about behavioral relationships. We then leverage these characteristics to identify the specific on- and off-target effects of both established and candidate therapeutics in the *CNTNAP2* mouse model of autism spectrum disorder (ASD)[20]. Taken together, this work demonstrates that MoSeq can effectively encapsulate

[1]Department of Neurobiology, Harvard Medical School, Boston, MA, USA. [2]Present address: Google Brain, Cambridge, MA, USA. [3]Present address: Google Brain, San Francisco, CA, USA. ✉e-mail: srdatta@hms.harvard.edu

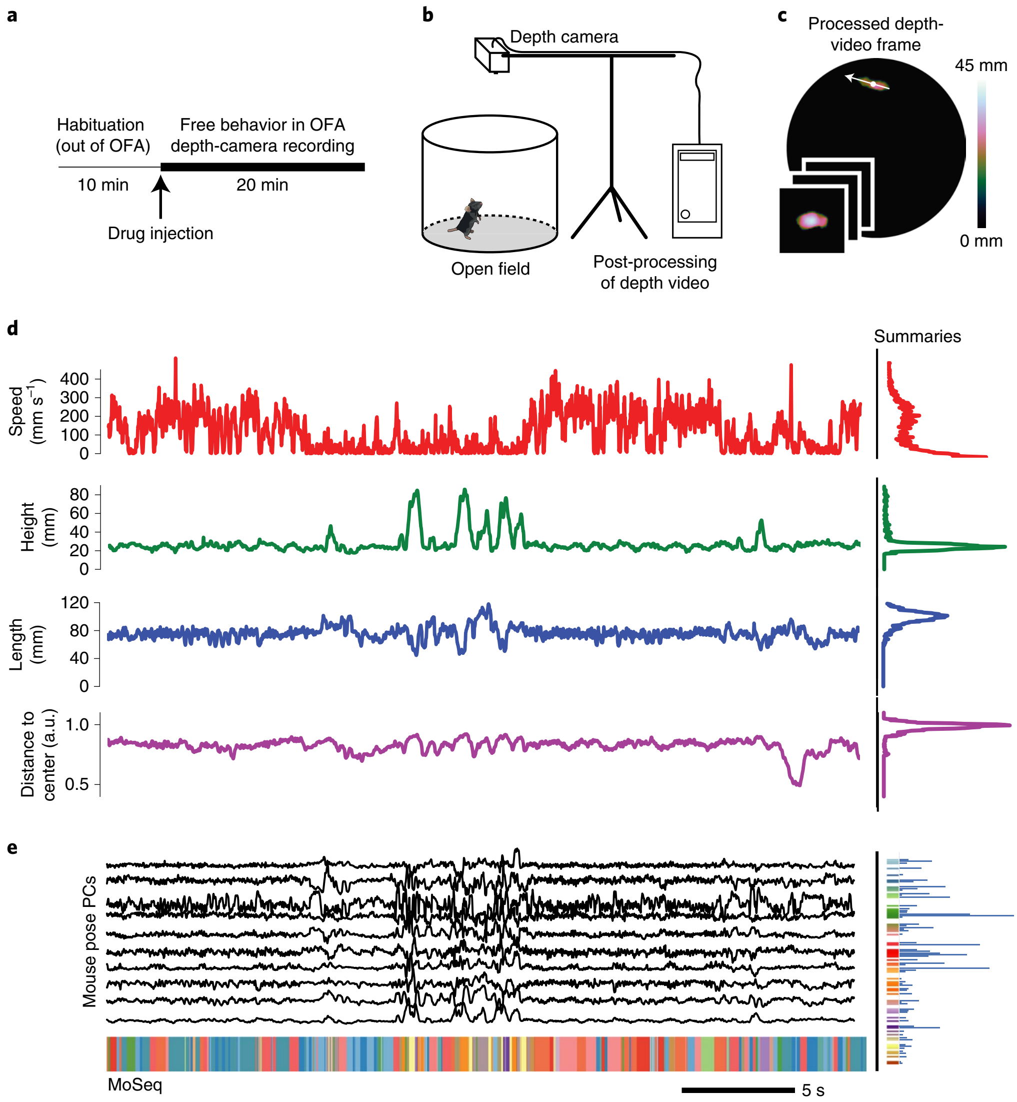


Fig. 1 | MoSeq captures 3D mouse pose dynamics after drug treatment. **a**, Schematic of the trial structure used for mouse OFA-based behavioral imaging. **b**, Mouse 3D pose dynamics were recorded using depth cameras placed above the arena, with raw frames stored locally and then processed in a cloud computing environment (Methods). **c**, A pre-processing pipeline identifies the mouse within the depth image, which enables analyses of 3D pose dynamics and quantification of scalar behavioral metrics (Methods). **d**, Imaging-based distributions of the speed, height, length and distance to arena center (given as arbitrary units (a.u.) for an example mouse during a 30-s example snippet. **e**, The first ten PCs of the pre-processed 3D imaging data (top) were fed to the MoSeq algorithm to assign each frame to a particular behavioral syllable (bottom) (Extended Data Fig. 1). The number of times each syllable is expressed during this 30-s example snippet is represented as a histogram (right). For each mouse, a MoSeq-based behavioral summary was generated using 20 min of data.

complex behavioral phenotypes in large-scale behavioral data and suggests that behavioral syllables represent a new category of therapeutic target for future drug development.

Results

To address whether the modular time-series description of behavior afforded by MoSeq can capture and organize behavioral variation in large-scale data, we acutely exposed mice to a panel of psychoactive or neuroactive pharmacological agents at multiple doses known to influence behavior. This drug-based strategy was designed to modulate activity across many neural circuits and neuromodulator systems, thereby eliciting diverse patterns of action in a neutral environment, the circular open field (Fig. 1a–c ($n=673$ mice in total) and Supplementary Table 1).

Two distinct behavioral summaries were computed for each imaged mouse: a “scalar” summary and a “MoSeq” summary. The scalar summary comprised parameters that are typically measured using point tracking over standard two-dimensional (2D) video, including distributions of length, speed and position, whereas the MoSeq summary consisted of how often each behavioral syllable was used (Figs. 1d,e and 2a–c, and see Methods and Extended Data Figs. 1 and 2 for details regarding construction of the behavioral summaries). Because imaging was performed using 3D cameras, the scalar summary was bolstered by the inclusion of the centroid height distribution, information that is typically not available with 2D cameras or beam breaks.

Visual inspection of the scalar behavioral summaries for each mouse offered intuitive insight into drug-induced behavioral states. For example, high-dose haloperidol caused low average speeds and frequent long-term pausing (apparent as a speckled pattern in the mouse position data), which is consistent with its known cataleptic effects[21] (Fig. 2a,b,e). In contrast, methylphenidate drove mice to the

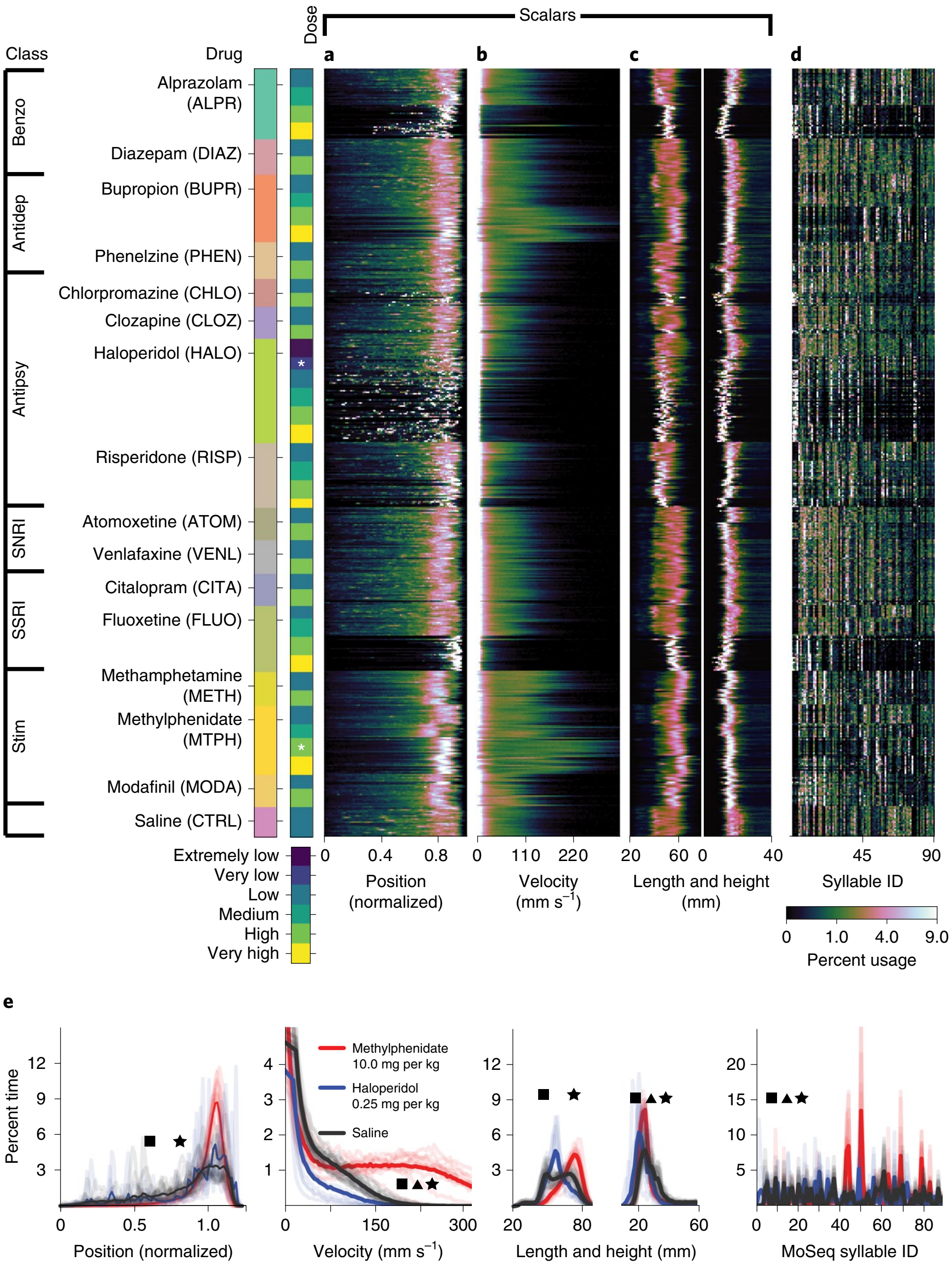


Fig. 2 | Generating behavioral diversity though pharmacology. a, Each mouse (rows) was treated with the indicated drug, and the distribution of mouse positions normalized to the arena center position was computed. The drug class is indicated on the left; here and in other figures, the abbreviations used are as follows: Benzo, benzodiazepine; Antidep, antidepressant; Antipsy, antipsychotic; SNRI, serotonin nonspecific reuptake inhibitor, SSRI, serotonin-selective reuptake inhibitor; Stim, stimulant. See Supplementary Table 1 for the number of mice used per treatment. **b**, Same as **a**, but for velocity. **c**, Same as **a**, but for length and height. **d**, Same as **a**, but the behavioral summary is composed of how often each MoSeq-identified (ID) syllable (arrayed on the *x* axis) was used. **e**, Comparisons of behavioral summaries for methylphenidate, haloperidol and saline at the doses indicated by the asterisks in the dose column in **a**. $P < 0.05$; the squares indicate significant differences between methylphenidate and haloperidol, the triangles between haloperidol and saline, and stars between methylphenidate and saline. Two-sided Mann–Whitney *U*-test was used on mean values for scalars, whereas two-factor MANOVA was used for MoSeq syllable differences; faint lines represent distributions of individual mice.

edge of the arena and substantially increased their velocity, which is consistent with its known stimulating properties[22]. MoSeq-based behavioral summaries captured a variety of subsecond stereotyped 3D actions (for example, darts, rears, pauses and turns) that differentiated most drugs and doses from control (Fig. 2d,e, mean duration ± s.d. = 425 ± 726 ms, and see Extended Data Fig. 3 for descriptions of behavioral syllables).

MoSeq enables effective behavioral classification. Scalar and MoSeq behavioral summaries for each mouse were submitted to a linear classifier to quantify the ability of each behavioral summary to distinguish each drug. As shown in Fig. 3, MoSeq outperformed traditional summaries at identifying individual drugs based on behavior (MoSeq $F_1 = 0.62 \pm 0.04$ versus scalar $F_1 = 0.40 \pm 0.05$; F_1 values represent the harmonic mean between precision and recall, and summarize the ability of a given method to capture true positives while rejecting both false positives and negatives). MoSeq was better at discriminating 14 out of the 16 drugs tested, including the saline controls (Fig. 3b and Supplementary Table 2, and see Methods for the use of randomized cross-validation to assess model reliability and statistical significance). Although absolute performance was reduced, MoSeq was also more effective at predicting the specific drug–dose combination that each mouse was administered (Extended Data Fig. 4 and Supplementary Table 3). Consistent with these classifier-based findings, the effective dimensionality of MoSeq, which measures its intrinsic capacity to describe behavioral variability, was higher than that for scalar metrics (Extended Data Fig. 4c). These experiments demonstrate that each drug elicits a specific pattern of behavior in treated mice and that—across nearly all drugs tested—MoSeq is more effective at capturing drug-specific behavioral effects than traditional metrics.

The data that make up each behavioral summary constrain its ability to convey information about behavioral variability, which raises the possibility that the specific composition of each summary limits its performance. To address this possibility, we modified both the scalar and MoSeq summaries to include additional measurements that were excluded in our initial analysis (such as acceleration, body angle, area, ellipticity and width in the case of scalar summaries, and syllable transition information for MoSeq). In neither case did performance exceed that of syllable-usage-based MoSeq alone (Extended Data Fig. 5a). In addition, MoSeq outperformed scalar metrics regardless of whether the scalar data were subject to dimensionality reduction, whether the scalar data were lumped into more or fewer bins or whether alternative classifier types were used to assess performance (Extended Data Fig. 5b,c).

These observations suggest that the time-series modeling approach used by MoSeq captures more relevant behavioral variance than simply aggregating behavioral data into histograms (as done by the scalar behavioral summary). To assess the importance of time-series modeling per se, we fed the frame-by-frame values of the parameters that make up the scalar behavioral summary to MoSeq, thereby identifying syllables on the basis of the scalar measurements instead of the 3D imaging data. This hybrid scalar/MoSeq summary exhibited improved performance relative to the scalar summary; however, it was still worse than classification performed using 3D imaging data (Fig. 3c). We also subjected the 3D pixel data to KMeans clustering, thereby generating a summary in which behavior is characterized by how often mice adopt one of the many possible 3D poses; this KMeans summary, in which behavior was clustered without regard to time, also significantly underperformed MoSeq (Fig. 3d). These findings demonstrate that time-series modeling can substantially improve the performance of even simple scalar metrics and that the 3D pixel data describing the full pose dynamics of the mouse contribute information important to behavioral classification that is absent from scalar metrics alone.

MoSeq separates treatment groups while capturing individual variation. Why is the behavioral summary generated by MoSeq effective at discriminating between closely related patterns of behavior? In principle, there are two (non-mutually exclusive) possible reasons. First, MoSeq might primarily act to separate treatment classes (here, mice treated with a given drug or drug–dose combination). If this is the case, the separation among the mean MoSeq behavioral summaries for each class should be large and greater than that observed when using scalar behavioral summaries. Alternatively, the mean class separation could be similar among summary types, but MoSeq might generate summaries with relatively low mouse-by-mouse variability, thereby reducing the confusion between drugs when assessed by the classifier.

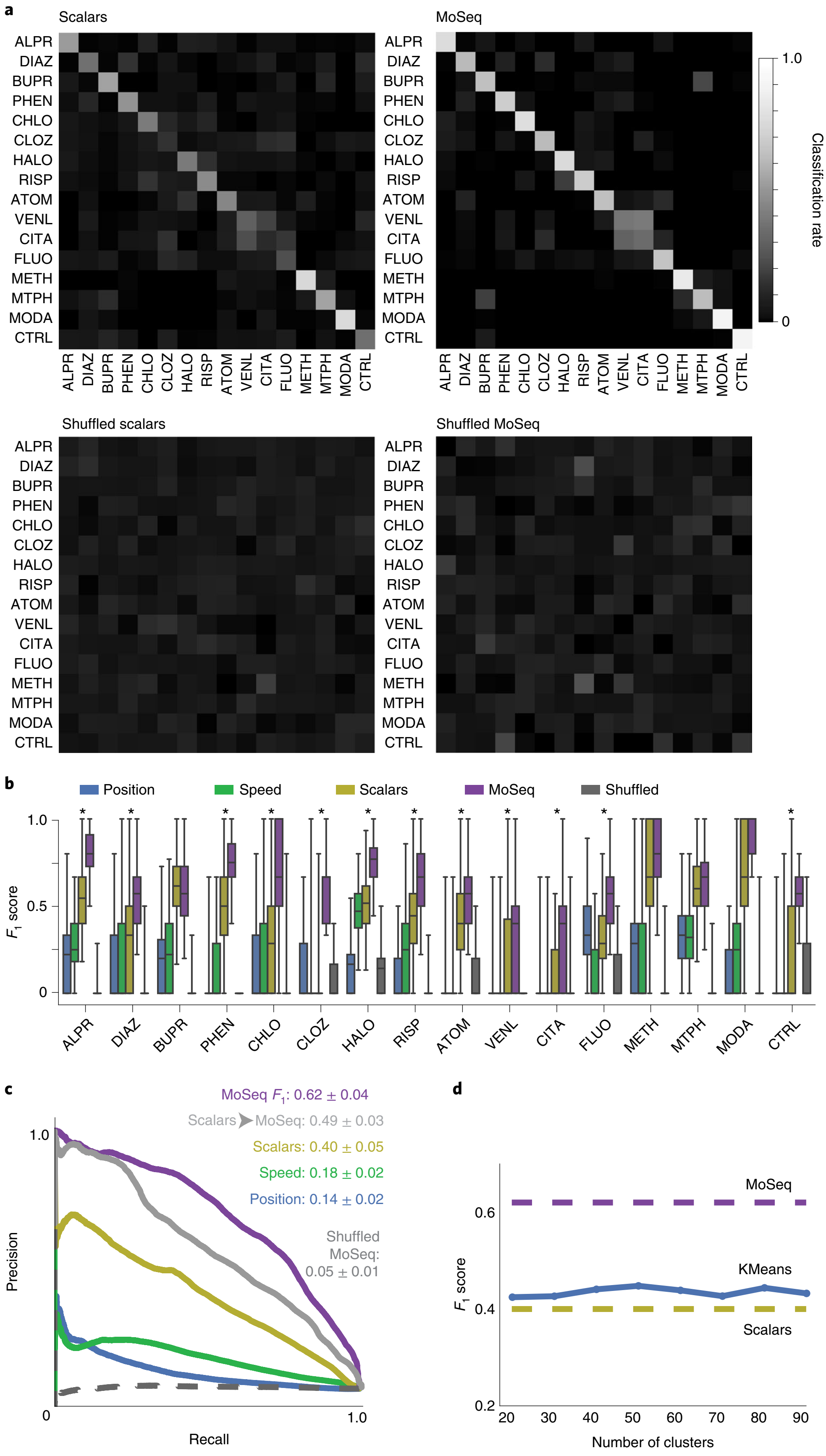
To explore these possibilities, we quantified the cosine distance that separated MoSeq summaries and compared these distances to those observed using scalar summaries. This analysis revealed that the mean separation between mice treated with different drugs, or drug–dose combinations, was greater when using MoSeq (Fig. 4a,b and Extended Data Fig. 6). Surprisingly, the cosine distances that separated individual mice within a given treatment class—that is, which all received the same drug—were also greater when using MoSeq than when using scalar summaries (Fig. 4a,b). Bootstrapping analysis demonstrated that these greater distances were not due to noise, but rather to bona fide behavioral differences between individual mice belonging to the same treatment class (Extended Data Fig. 7). Together, these results demonstrate that MoSeq supports behavioral classification by increasing the separation (relative to other metrics) between different treatment groups while at the same time maintaining information about the behavioral variability of individual mice within each treatment group.

MoSeq reveals behavioral relationships in large-scale datasets. These findings indicate that MoSeq effectively distinguishes patterns of behavior caused by specific drugs. However, it is not clear whether MoSeq also captures information about drug-related behaviors that are shared across drugs, which could be diminished if MoSeq simply decorrelates representations of behavior for each mouse. Indeed, the greater overlaps between the representations of individual mice observed in the scalar summaries (Fig. 4c) could enable those summaries to better represent behavioral relationships. However, classifier analysis revealed that MoSeq was uniformly more effective than traditional metrics at identifying the pharmacological class to which a given drug belongs (MoSeq $F_1 = 0.65 \pm 0.04$, scalar $F_1 = 0.42 \pm 0.06$, chance $F_1 = 0.12 \pm 02$, Fig. 5a,b).

Given that the notion of pharmacological class is not rigorous—as many drugs used in neurological and psychiatric practice are deployed for indications that cross diagnostic boundaries[23]—we asked whether MoSeq or scalar behavioral representations could identify behavioral relationships independent of constructed categories. Indeed, the pairwise correlation matrices describing

Fig. 3 | MoSeq discriminates drug-induced patterns of behavior. a, Normalized classification matrices (across rows and columns, plots represent classifier means after 500 cross-validation folds; see Methods for details and Supplementary Table 1 for the number of mice used per treatment) summarizing the performance of a linear classifier at distinguishing different drugs based on the indicated behavioral summary. Perfect classifier performance (in which each mouse is correctly assigned to its drug label) corresponds to white along the diagonal and black on the off-diagonal (that is, a classification rate of 1). For the shuffled control (bottom row), drug labels were shuffled on a per-mouse basis to compute a baseline of expected random performance. The heatmap indicates classification successes and errors (see Methods for summary definitions). Drug abbreviations here and in other figures are as indicated in Fig. 2. **b**, F_1 values, reflecting classification accuracy, for all behavioral summaries, including a label-shuffled random baseline. Box plots represent the distribution across 500 cross-validation folds, with whiskers representing 1.5-times the inter-quartile range. Shuffle controls are as in **a** ($P < 0.01$, paired two-sided t-test, Holm–Bonferroni step-down correction; asterisks indicate significant differences between MoSeq and scalars). **c**, Mean PR curves and F_1 values for all summary types across all drug treatments. Shuffle controls are as in **a**. "Scalars➤MoSeq" indicates the performance observed when modeling scalar values rather than 3D imaging data using MoSeq. **d**, Mean F_1 score of an alternative behavioral summary constructed by performing KMeans clustering (with the cluster number indicated) on the 3D image PCs (Methods). Note that the MoSeq summaries are composed of 90 syllables, which correspond to the maximum number of clusters chosen for analysis here. For comparison, the mean F_1 predictive performance scores are indicated for MoSeq and scalars.

behavioral similarities and differences revealed behavioral relationships between drugs across distinct pharmacological classes (Fig. 4c). To explore drug relationships from a classification perspective, we removed a single drug from our dataset and then built a linear classifier based on the MoSeq or scalar summaries of the remaining drugs to identify those agents that were most behaviorally similar



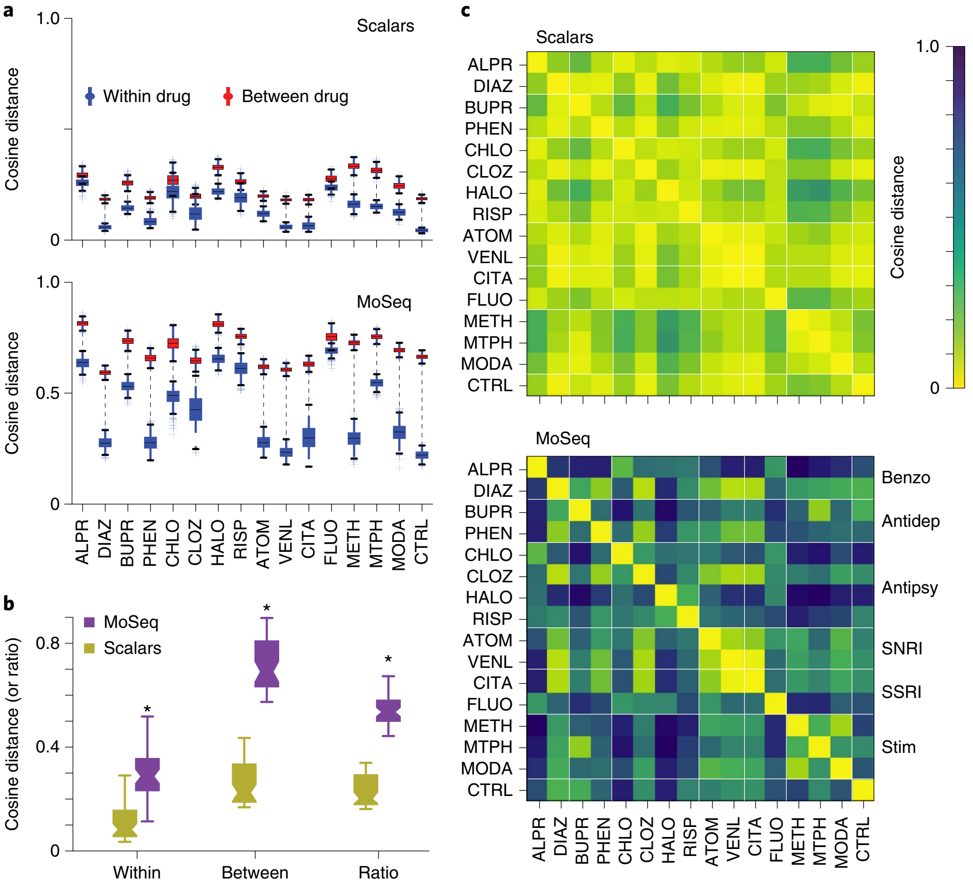


Fig. 4 | MoSeq enhances the separation between treatment classes relative to scalars. a, Average cosine distances for scalars (top) and MoSeq (bottom) of individual mice given the same drug (blue) compared with mice given different drugs (red) (±1 s.d. values are indicated; see Supplementary Table 1 for the number of mice used per treatment). **b**, Mean within- and between-treatment cosine distances, and their ratio, for scalar summaries and MoSeq ($P < 0.05$, asterisks indicate significant differences between MoSeq and scalars; paired two-sided *t*-test). Whiskers represent 1.5× the inter-quartile range, which is the range of values covered by the solid boxes. **c**, Average pairwise cosine distances between mice given the indicated drug treatments (the distance is indicated by the color bar; the white lines separate the drug classes, which are indicated to the right of the lower panel).

to the 'held-out' drug. By iteratively holding out each drug in the set, we could identify overlaps in the patterns of behavior evoked by all drugs in our dataset and then compare the overlaps identified by MoSeq and scalars (Fig. 5c).

When applied to MoSeq summaries, this approach identified relationships among drugs that belong to the same class (for example, modafinil–methamphetamine, haloperidol–risperidone), and three prominent inter-class relationships (for example, methylphenidate (stimulant)–bupropion (antidepressant), venlafaxine (selective serotonin reuptake inhibitor)–citalopram (serotonin nonselective reuptake inhibitor) and chlorpromazine (antipsychotic)–alprazolam (anxiolytic)). These same drug relationships were observed when embedding the MoSeq behavioral summaries into a 2D space using linear discriminant analysis (LDA) for visualization purposes (Fig. 5d), but were weaker or absent when held-out confusion matrices were computed using scalar summaries. Interestingly, data mining revealed that two of the inter-class pairs share clinical indications, while the third pair (chlorpromazine–alprazolam) shares sedation as a side effect[23] (Supplementary Table 4 and Methods). Thus, MoSeq can identify relationships among drugs that both include and transcend traditionally defined pharmacological classes; these behavioral relationships may in part reflect the observed effects of drugs in the clinic.

To pressure test the notion that MoSeq can simultaneously capture useful information about behavioral similarities and differences, we generated dose–response curves for three antipsychotic drugs—haloperidol, clozapine and risperidone—that all elicit a reduction in movement, albeit through different mechanisms. Haloperidol and risperidone both antagonize the dopamine D2 receptor (D2R) and trigger catalepsy, while clozapine and risperidone inhibit the serotonin 5-HT$_{2A}$ receptor (5-HT$_{2A}$R), which is thought to lead to sedation[24,25]. Clozapine is also a high-affinity histamine H$_1$ receptor antagonist, which contributes to its sedative effects[24,25]. Consistent with each of these agents antagonizing different receptors with distinct affinities[24,25], the classifier analysis demonstrated that MoSeq effectively distinguished nearly all drug–dose combinations (Extended Data Fig. 8a). Each drug altered a specific complement of behavioral syllables, many of which were unrelated to locomotion—for example, grooming and rearing (Extended Data Fig. 8b). Consistent with this observation, MoSeq also effectively classified the three drugs independent of their differential effects on velocity (Fig. 5e). Embedding the dose–response data using LDA revealed that at high doses, risperidone and haloperidol converged on a similar pattern of behavior distinct from that evoked by clozapine (Fig. 5f, compare the darkest blue square, green triangle and red star). These results demonstrate that MoSeq can differentiate between catalepsy (that is, haloperidol-typical behaviors) and sedation (that is, clozapine-typical behaviors), which both reduce movement and are often confused in traditional behavioral assays[26]. The fact that at high doses, risperidone acts predominantly as a cataleptic rather than a sedative suggests that its primary behavioral effects at high doses are caused by antagonism of the D2R rather than the 5-HT$_{2A}$R (despite the higher affinity of risperidone for the 5-HT$_{2A}$R relative to the D2R). Importantly, this inference (drawn on the MoSeq analysis alone) is consistent with a previous finding that locomotion is persistently reduced by risperidone in mice with the 5-HT$_{2A}$R knocked out[26].

MoSeq identifies subsets of behavioral syllables that encapsulate phenotypes. The ability of MoSeq to effectively distinguish drug effects while maintaining information about related patterns of behavior raises the question of how drug treatments alter the expression of behavioral syllables. Each drug appeared to significantly change the usage of a large subset of syllables when considered relative to control (Fig. 6a). However, LASSO regression analysis revealed that most of the information required to tell individual drugs apart from each other resides in a small subset of syllables (typically 5, nearly always fewer than 15; Extended Data Fig. 9). These small groups of drug-characteristic syllables reflected the similarities and differences between drugs as identified via the held-out classifier, including within drug-class relationships (for example, modafinil–methampehtamine) as well as across-class relationships (for example, citalopram–venlafaxine) (Fig. 6b and see Supplementary Figs. 1 and 2 for a description of the similarities and differences among drug-regulated and discrimination-relevant syllables).

In accord with its known role as a stimulant, all of the five most discriminant methamphetamine-related syllables encoded different forms of forward movement; three of these syllables overlap with the five most discriminant syllables for modafinil, with the two modafinil-specific syllables encoding exploratory behaviors, including a partial rear and a pause-and-head-flick motif (Supplementary Fig. 3). These observations demonstrate that modafinil shares at least some stimulant-related activity with methamphetamine, which is consistent with modafinil and methamphetamine acting through an overlapping set of molecular targets[27,28]. However, modafinil also recruited additional investigatory behaviors, which is consistent with modafinil engaging receptors distinct from those recruited by methamphetamine. Similarly, citalopram-related syllables encode forward movement and grooming behaviors; a subset of these syllables are shared with venlafaxine, which also recruited pausing and rearing behaviors that are not differentially upregulated by citalopram (Supplementary Fig. 4).

Behavioral syllables enable objective assessment of interactions between genes and candidate therapeutics. Given its ability to identify specific drug-related behavioral effects, we asked whether MoSeq could characterize the ability of a drug to revert behavioral phenotypes in a disease model. To explore this possibility, we used MoSeq to phenotype mice harboring a mutation in the *CNTNAP2* gene, which is associated with human ASD[20,29,30]. Consistent with prior results, velocity measurements revealed that the *CNTNAP2* mice are hyperactive[20,31] (Supplementary Fig. 5). MoSeq identified 16 behavioral syllables whose expression was statistically altered with respect to wild-type (WT) mice (Fig. 7a). Visual inspection revealed that many of these syllables would be predicted to be associated with a hyperactive phenotype (for example, downregulated pauses and upregulated micromovements and running); however, many high-velocity syllables were not affected by the *CNTNAP2* mutation (data not shown), which demonstrates that *CNTNAP2*-related hyperactivity does not reflect generalized arousal, but instead is composed of a specific array of syllabic changes (Fig. 7a).

Previous experiments have shown that the *CNTNAP2* hyperactivity phenotype can be reverted by treatment with risperidone, which is clinically used to treat hyperactivity and aggression in patients with ASD[20]. Of the 16 behavioral syllables that define the *CNTNAP2* mutant phenotype, seven were statistically normalized by risperidone treatment, seven were partially reverted and two remained uncorrected (Fig. 7a–c). Despite not fully reverting the observed mutant phenotype, risperidone also altered a large number of additional behavioral syllables, several of which represent high- velocity behaviors such as running. These results quantitatively demonstrate that risperidone has a specific (albeit partial) effect on the phenotype induced by mutation of the *CNTNAP2* gene and a much broader set of side effects on normal behavioral syllables.

We also wished to test the utility of MoSeq for characterizing the on- and off-target effects of novel or previously untested therapeutics in the *CNTNAP2* model. To identify candidates, we took advantage of a repurposing dataset in which possible therapies for ASD were nominated based on the intersection of genome-wide association data and drug-induced changes in gene expression[32]. From this

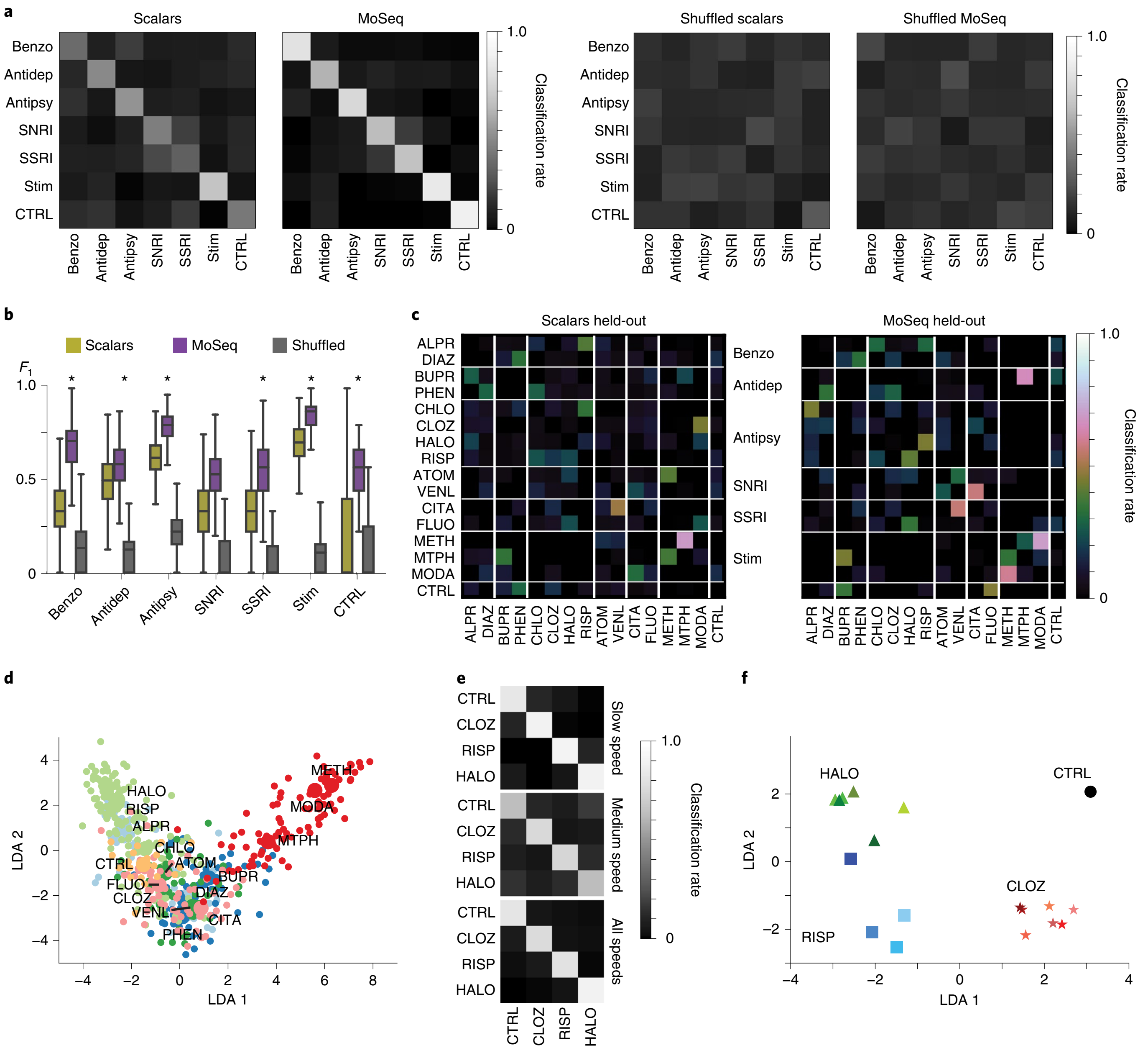
Fig. 5 | MoSeq reveals behavioral relationships between drug classes and can distinguish catalepsy from sedation. a, Normalized classification matrices (across rows and columns, plots represent means after 500 cross-validation folds; Methods) summarizing the classification performance of linear classifiers trained to predict drug class on a mouse-for-mouse basis (left). The heatmap indicates classification successes and errors; perfect classifier performance (in which each mouse is correctly assigned to its class label) corresponds to white along the diagonal and black on the off-diagonal (that is, a classification accuracy of 1). For the shuffled control (right), class labels were shuffled on a per-mouse basis to compute a baseline of expected random performance. See Supplementary Table 1 for the number of mice used per treatment. **b**, F_1 scores for linear classifiers designed to predict the pharmacological drug class on a mouse-for-mouse basis. Box plots represent the distribution across 500 cross-validation folds, with whiskers representing 1.5× the inter-quartile range ($P < 0.01$, asterisks indicate significant differences between MoSeq and scalars, paired two-sided t-test corrected with Holm–Bonferroni step-down procedure; Methods). The shuffle control was performed as in **a**. **c**, Held-out confusion matrices (across rows and columns) indicating the classification of a given drug when that drug was excluded from the drug classifier (and thus these matrices represent confusions made over 16 separate classifiers). This procedure identifies the drugs most confused with the query drug (given that, by design, the held-out classifier must identify a non-query drug as the correct label for each mouse). As correct within-drug classification is impossible in this representation, the diagonal is dark (plots depict means after 500 cross-validation folds, see Methods for details of held-out classification); drug classes are indicated. **d**, LDA plot indicating the similarity between the mean behavioral summaries of mice across drug treatments. Opaque circles indicate mean summary embeddings, and semi-transparent circles show the embedding location of each mouse. Colors indicate drugs from the same pharmacological class. **e**, Normalized classification matrices for different drugs, whereby the specific doses chosen for each drug were grouped on the basis of mouse speed (the mean centroid speed of the saline-treated control mouse = 74 mm s^{-1}; medium speed = 54 mm s^{-1}; slow speed = 24 mm s^{-1}; see Methods for a description of the Gaussian-mixture-model-based method for grouping doses based on speed). Perfect classification is indicated by white along the diagonal and black off-diagonal; the high degree of predictability when stratifying different below-normal speeds demonstrates that MoSeq can distinguish these drugs independent of their effects on gross movement. **f**, LDA plot indicating the observed mean MoSeq-characterized pattern of syllable usages for the three indicated drugs at doses tiling very low (light) to very high (dark; Methods). In general, all doses of each drug cluster together in the LDA space, and separate from a control saline treatment, although at the highest doses, risperidone and haloperidol elicit similar patterns of behavior (see the darkest blue square and the darkest green triangle, respectively).

list, we identified two drugs, loxapine and sulpiride, that have not been previously tested in *CNTNAP2* mutant mice and whose mechanisms of action overlap with—but are distinct from—risperidone (loxapine also antagonizes both the D2R and 5-HT$_{2A}$R, but with a lower relative inhibition ratio than risperidone[33], while sulpiride is a pure D2R antagonist).

Similar to risperidone, both sulpiride and loxapine reverted the gross hyperactivity of *CNTNAP2* mutant mice, as assessed by velocity measurements (Supplementary Fig. 5). However, MoSeq revealed that loxapine was less efficacious than risperidone at correcting *CNTNAP2*-specific syllables, and further recruited more side-effect syllables. In contrast, sulpiride exhibited nearly identical on-target effects with risperidone, but altered fewer off-target syllables (Fig. 7a–c). Importantly, with one exception, the off-target syllables induced by sulpiride—which specifically antagonizes the D2R—overlapped with the broader set induced by risperidone. These data suggest that D2R antagonism enables both risperidone and sulpiride to partially revert the *CNTNAP2* phenotype and that the risperidone-specific off-target effects (relative to sulpiride) are likely due to antagonism of other receptors, such as the 5-HT$_{2A}$R (Supplementary Figs. 5 and 6). These experiments reveal that MoSeq can identify a syllabic fingerprint that characterizes complex behavioral changes in a disease model. This fingerprint can be used to quantitatively assess the intended and inadvertent effects of candidate therapeutic agents and to deconvolve relationships between drugs, receptors and behavior.

Discussion

Before these experiments, it was not apparent whether MoSeq is more like a northern blot—a bespoke approach for understanding the relative expression levels of a small number of target RNAs from a limited set of samples—or RNA sequencing, which creates a broad and general representation of the transcriptome that can be effectively used to infer relationships among many different cell types and experimental interventions. This work reveals that MoSeq can experimentally parse induced behavioral variability



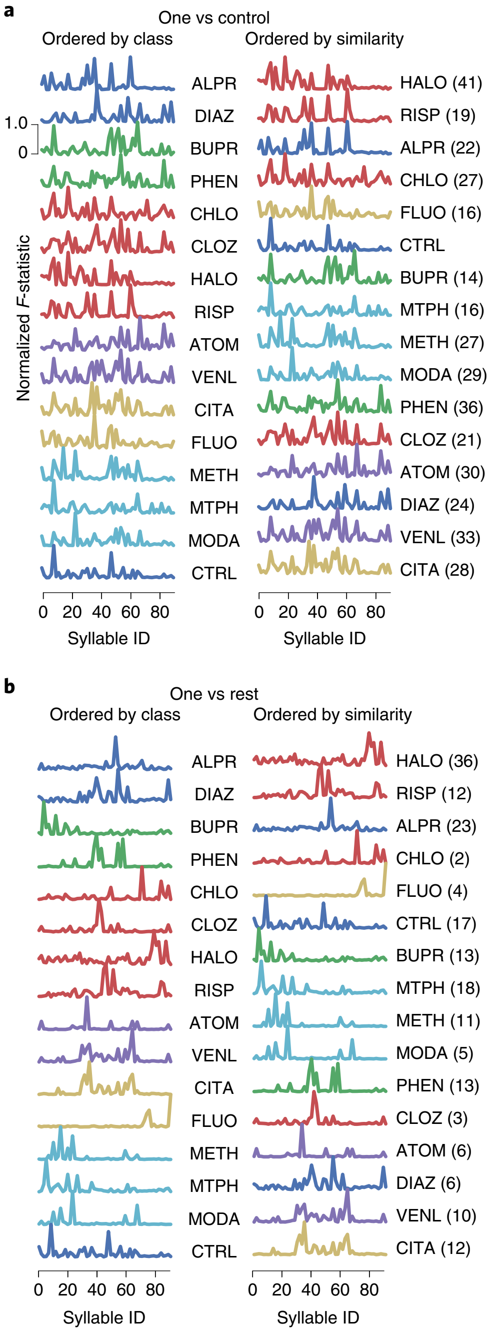


Fig. 6 | Subsets of syllables fingerprint each drug. a, A normalized *F* statistic identifies the quantitative relevance of each indicated syllable for discriminating a given drug treatment from a control saline treatment (one versus control). Ordering on left is based on the pharmacological class, ordering on the right is based on similarities in the *F*-statistic-identified syllables. The number of significant syllables is indicated in parentheses next to the drug treatment name on the right (Holm–Bonferroni-corrected $P < 0.01$ from the two-sided *F*-test). The control treatment *F* statistic is computed by comparing against all other treatments. **b**, Same as **a**, but computing the *F* statistic between a given drug treatment and all other treatments (one versus rest). The all-versus-all comparison revealed many fewer statistically significant syllables than when comparing to control alone. Note that those syllables that distinguish a given drug from control can be distinct from those that maximally distinguish a particular drug from all other tested drugs.

within large-scale and diverse datasets. Despite the fact that MoSeq is highly discriminative—and can therefore identify the specific behavioral effects of closely related drugs and doses—it retains information about behavioral relationships, allowing drug categorization independent of a presumed mechanism of action. These features also enable MoSeq to unveil the intersecting effects of gene and drug manipulations, even when the mechanistic consequences of those interventions are incompletely understood.

Drugs act at specific complements of receptors that selectively modulate the activity of neural circuits, which in turn cause changes in behavior. However, efforts to link drug effects to molecular mechanisms and behaviorally relevant circuits have been significantly complicated by the low dimensionality, the poor signal-to-noise and the lack of specificity of traditional behavioral metrics. The discriminative capacity of MoSeq suggests that it may ultimately enable receptor modulation to be causally mapped onto patterns of neural circuit activity and behavior, thereby allowing inferences to be drawn about the role of drug receptors in composing and shaping behavioral space. Our proof-of-concept experiments provisionally linking the differential expression of particular syllables to the modulation of specific receptors (made possible by phenotyping different drugs with distinct but overlapping receptor specificities) suggest that this sort of mapping could also enable accurate predictions of the mechanism of action of a drug from behavior alone.

We speculate that MoSeq outperforms traditional behavioral representations for four reasons. First, MoSeq organizes information about 3D pose dynamics based on the inherent structure of the behavioral data and in a manner that respects the observation that mouse behavior is both continuous and discrete. Second, MoSeq does not prespecify the number and identity of behavioral syllables, but instead learns these features on an experiment-by-experiment basis. Thus, the richness of the behavioral representation scales with the amount of observed behavioral variability, which enables MoSeq to summarize behavior in a manner that is simultaneously compact and expressive[9,34]. Third, MoSeq partly defines individual syllables on the basis of the order in which they occur, thereby leveraging the sequential nature of naturalistic behavior[3,35–37]. Finally, recent work suggests that the dorsolateral striatum encodes syllable identity and is required to assemble syllables into coherent sequences[11]. Thus, MoSeq may be particularly effective because it describes behavior, at least in part, in modular terms similar to those used by the brain to create it.

We explicitly chose to measure behavior in experiments in which mice explore featureless environments after acute drug exposure, reasoning that this represented a ground state in which behavioral differences should be difficult to quantify, thereby putting MoSeq to a rigorous test. It is clear that different patterns of behavior would be observed if mice were given drugs chronically rather than acutely or were placed in richer contexts that demand goal-oriented behaviors. For example, one might expect chronic methamphetamine (which is highly addictive) and chronic modafinil (which is not) to be more distinguishable than was observed here with acute treatment alone[23]. Similarly, drugs that influence frontal circuits (such as antipsychotics) might elicit greater behavioral differences in the context of social or stress assays. Furthermore, the relatively brief experiments carried out here almost certainly fail to capture the ability of many drugs (and associated neural circuits) to reshape behavior over long timescales. Future work will be required to assess the utility of MoSeq in long-term behavioral assays or in assays designed to elicit specific psychological reactions, such as the forced swim test or the three-chamber social assay.

Many of the chemical templates for currently used psychotherapeutics were discovered in the 1950s and 1960s on the basis of their behavioral effects[38]. This led to the widespread use of behavioral phenotypes (ranging from open-field entries to spider-web geometry) to screen for candidate therapeutics[39,40]; however, limited by low

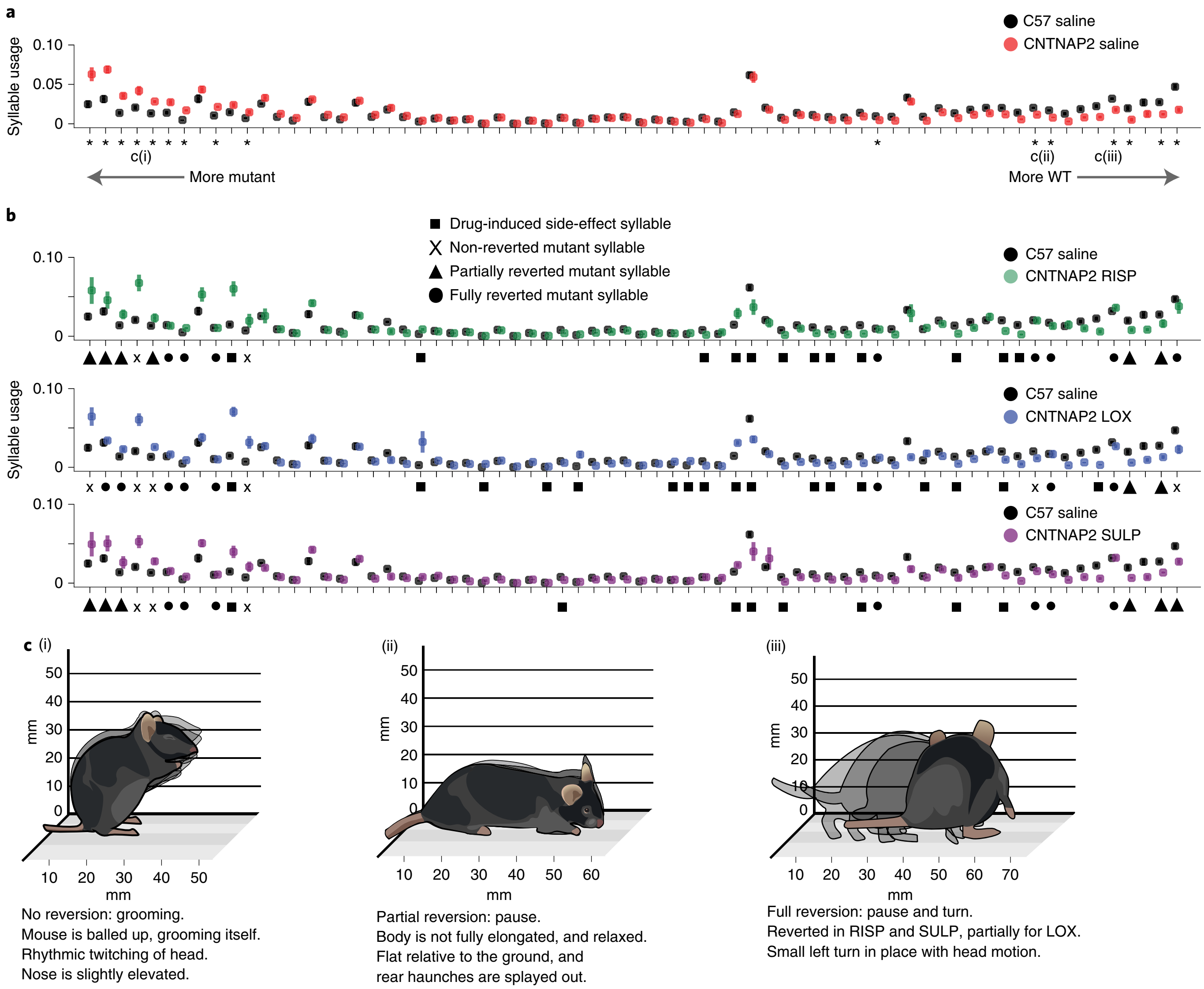


Fig. 7 | MoSeq-based phenotypic fingerprinting reveals on- and off-target drug effects in a mouse model of ASD. **a**, Usage plots for WT (black) and *CNTNAP2*$^{-/-}$ (red) mice injected with saline control (bootstrapped 95% confidence intervals indicated). Syllables were sorted by the degree to which they are overused in the mutant (Methods), with differentially used syllables marked by asterisks (for all statistical tests in this figure, Kruskal–Wallis and post hoc Dunn's two-sided test with permutation were used, with Benjamini–Hochberg FDR with $\alpha = 0.05$). Example syllables illustrated in **c** are indicated as (i), (ii) and (iii). See Methods for the number of mice used per treatment group. **b**, Usage plots for WT (black) and *CNTNAP2*$^{-/-}$ mice injected with risperidone (RISP; top), loxapine (LOX; middle) or sulpiride (SULP; bottom). Symbols indicate differentially used syllables (see Methods for definitions of reversions and side effects). **c**, Schematic illustrations of syllables that were not reverted (i), partially reverted (ii) or fully reverted (iii) by drug treatments. Note that syllable (iii) was fully reverted with RISP and SULP, but only partially reverted with LOX.

resolution and high variability, these behavior-based approaches have generally failed to yield novel pharmacology. More recent drug development efforts have focused on identifying risk genes and using medicinal chemistry to actuate or inhibit those specific targets. This alternative strategy has also not been entirely successful, perhaps in part because most clinically approved therapeutics exhibit mixed selectivity for multiple targets[23,25,38]. The observation that MoSeq summarizes complex behavioral phenotypes induced by drug and genetic manipulations—which almost certainly exert their effects through many receptors and neural circuit mechanisms in parallel—as discrete changes in subsets of behavioral syllables suggests that syllables themselves could serve as druggable targets. The ability of MoSeq to reveal on- and off-target effects of risperidone, sulpiride and loxapine in *CNTNAP2* mutant mice is consistent with this possibility. Given its low cost, scalability and interpretability, MoSeq may be useful as a discovery platform for characterizing the specific disease-relevant effects of candidate therapeutics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-020-00706-3.

Received: 1 July 2019; Accepted: 10 August 2020;
Published online: 21 September 2020

References

1. Tinbergen, N. *The Study of Instinct* (Clarendon Press, 1951).
2. Dawkins, R. in *Growing Points in Ethology* (eds Bateson, P. P. G. & Hinde, R. A.) 7–54 (Cambridge Univ. Press, 1976).
3. Datta, S. R., Anderson, D. J., Branson, K., Perona, P. & Leifer, A. Computational neuroethology: a call to action. *Neuron* **104**, 11–24 (2019).
4. Anderson, D. J. & Perona, P. Toward a science of computational ethology. *Neuron* **84**, 18–31 (2014).
5. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
6. Meyer, A. F., Poort, J., O'Keefe, J., Sahani, M. & Linden, J. F. A head-mounted camera system integrates detailed behavioral monitoring with multichannel electrophysiology in freely moving mice. *Neuron* **100**, 46–60.e7 (2018).
7. Klaus, A. et al. The spatiotemporal organization of the striatum encodes action space. *Neuron* **95**, 1171–1180.e7 (2017).
8. Pereira, T. D. et al. Fast animal pose estimation using deep neural networks. *Nat. Methods* **16**, 117–125 (2019).
9. Wiltschko, A. B. et al. Mapping sub-second structure in mouse behavior. *Neuron* **88**, 1121–1135 (2015).
10. Graving, J. M. et al. Fast and robust animal pose estimation. *eLife* **8**, e47994 (2019).
11. Markowitz, J. E. et al. The striatum organizes 3D behavior via moment-to-moment action selection. *Cell* **174**, 44–58.e17 (2018).
12. Crawley, J. N. Behavioral phenotyping of rodents. *Comp. Med.* **53**, 140–146 (2003).
13. Crawley, J. N. Behavioral phenotyping strategies for mutant mice. *Neuron* **57**, 809–818 (2008).
14. Crabbe, J. C. Genetics of mouse behavior: interactions with laboratory environment. *Science* **284**, 1670–1672 (1999).
15. Wahlsten, D. et al. Different data from different labs: lessons from studies of gene–environment interaction. *J. Neurobiol.* **54**, 283–311 (2002).
16. Egnor, S. E. R. & Branson, K. Computational analysis of behavior. *Annu. Rev. Neurosci.* **39**, 217–236 (2016).
17. Berman, G. J., Choi, D. M., Bialek, W. & Shaevitz, J. W. Mapping the stereotyped behaviour of freely moving fruit flies. *J. R. Soc. Interface* https://doi.org/10.1098/rsif.2014.0672 (2014).
18. Fentress, J. C. & Stilwell, F. P. Grammar of a movement sequence in inbred mice. *Nature* **244**, 52–53 (1973).
19. Berridge, K. C., Fentress, J. C. & Parr, H. Natural syntax rules control action sequence of rats. *Behav. Brain Res.* **23**, 59–68 (1987).
20. Peñagarikano, O. et al. Absence of CNTNAP2 leads to epilepsy, neuronal migration abnormalities, and core autism-related deficits. *Cell* **147**, 235–246 (2011).
21. Zetler, G. Haloperidol catalepsy in grouped and isolated mice. *Pharmacology* **13**, 526–532 (1975).
22. Millichap, J. G. & Boldrey, E. E. Studies in hyperkinetic behavior. II. Laboratory and clinical evaluations of drug treatments. *Neurology* **17**, 467–471 (1967).
23. Ebenezer, I. S. *Neuropsychopharmacology and Therapeutics* (Wiley, 2015).
24. Duncan, G. E., Zorn, S. & Lieberman, J. A. Mechanisms of typical and atypical antipsychotic drug action in relation to dopamine and NMDA receptor hypofunction hypotheses of schizophrenia. *Mol. Psychiatry* **4**, 418–428 (1999).
25. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3**, 353–359 (2004).
26. McOmish, C. E., Lira, A., Hanks, J. B. & Gingrich, J. A. Clozapine-induced locomotor suppression is mediated by 5-HT$_{2A}$ receptors in the forebrain. *Neuropsychopharmacolgy* **37**, 2747–2755 (2012).
27. Volkow, N. D. et al. Effects of modafinil on dopamine and dopamine transporters in the male human brain: clinical implications. *JAMA* **301**, 1148–1154 (2009).
28. Zolkowska, D. et al. Evidence for the involvement of dopamine transporters in behavioral stimulant effects of modafinil. *J. Pharmacol. Exp. Ther.* **329**, 738–746 (2009).
29. Alarcón, M. et al. Linkage, association, and gene-expression analyses identify *CNTNAP2* as an autism-susceptibility gene. *Am. J. Hum. Genet.* **82**, 150–159 (2008).
30. Rodenas-Cuadrado, P., Ho, J. & Vernes, S. C. Shining a light on CNTNAP2: complex functions to complex disorders. *Eur. J. Hum. Genet.* **22**, 171–178 (2014).
31. Brunner, D. et al. Comprehensive analysis of the 16p11.2 deletion and null *Cntnap2* mouse models of autism spectrum disorder. *PLoS ONE* **10**, e0134572 (2015).
32. So, H.-C. et al. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. *Nat. Neurosci.* **20**, 1342–1349 (2017).
33. Ferreri, F. et al. The in vitro actions of loxapine on dopaminergic and serotonergic receptors. Time to consider atypical classification of this antipsychotic drug? *Int. J. Neuropsychopharmacol.* **21**, 355–360 (2018).
34. Datta, S. R. Q&A: understanding the composition of behavior. *BMC Biol.* **17**, 44 (2019).
35. Brown, A. E. X., Yemini, E. I., Grundy, L. J., Jucikas, T. & Schafer, W. R. A dictionary of behavioral motifs reveals clusters of genes affecting *Caenorhabditis elegans* locomotion. *Proc. Natl Acad. Sci. USA* **110**, 791–796 (2013).
36. Berman, G. J., Choi, D. M., Bialek, W. & Shaevitz, J. W. Mapping the stereotyped behaviour of freely moving fruit flies. *J. R. Soc. Interface* **11**, 20140672 (2014).
37. Vogelstein, J. T. et al. Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. *Science* **344**, 386–392 (2014).
38. Swinney, D. C. & Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discov.* **10**, 507–519 (2011).
39. Hendriksen, H. & Groenink, L. Back to the future of psychopharmacology: a perspective on animal models in drug discovery. *Eur. J. Pharmacol.* **759**, 30–41 (2015).
40. Witt, P. N. Drugs alter web-building of spiders: a review and evaluation. *Behav. Sci.* **16**, 98–113 (1971).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Ethical compliance. All experimental procedures were approved by the Harvard Medical School Institutional Animal Care and Use Committee (protocol number 04930) and were performed in compliance with the ethical regulations of Harvard University as well as the Guide for Animal Care and Use of Laboratory Animals.

Data acquisition. Drugs were tested on $n=673$ 6–8-week-old C57/BL6 males (Jackson Laboratories). Mice were housed in standard animal facility conditions at a temperature of 71 ± 3 °F and at a relative humidity of 50 ± 15%. Mice were introduced into the colony at 5 weeks of age and group-housed for 1 week in a reverse 12-h light/12-h dark cycle. On the day of testing, mice were brought into the laboratory in a light-tight container, where they were habituated to the experiment room under red light for 10 min in disposable cages (Innovive) containing fresh bedding, with food and water available ad libitum. After the habituation period and subsequent drug injection, mice were placed in the middle of a circular 18″ diameter open-field assay (OFA) enclosure with 15″-high opaque walls (US Plastics), immediately after which video recording began. All experiments were performed under red light. Mice were allowed to freely explore the enclosure for the 20-min experimental period. At the end of the experiment, the enclosure was cleaned with 70% ethanol before reuse.

Drug treatments. Each mouse was treated with a single drug–dose combination, and used only once. Drug names, their concentration, the method used for dilution, the number of mice treated with each drug–dose combination and the supporting citations for the choice of dose are described in Supplementary Table 1. Drug doses were selected on the basis of the published literature to maximize the likelihood of observing a behavioral effect within the dose–response window. All drugs were delivered via intraperitoneal injection. All drug dilutions were prepared fresh on the day of experimentation, dissolved in accordance with previously published work and delivered intraperitoneally in a final volume of 200 µl. Drugs were generally diluted in lactated ringers solution (LRS), except for fluoxetine (at doses higher than 10 mg per kg), haloperidol (at doses higher than 0.25 mg per kg) and methylphenidate, which were diluted in ddH$_2$O. In instances where a drug was not soluble in LRS or ddH$_2$O, the drug was first diluted in dimethyl sulfoxide, then further diluted in LRS. Drug–dose pairs were tested in a pseudorandomized order, with control mice interspersed with drug treatments throughout the data acquisition phase of the experiment. The data acquisition phase of the experiment lasted for a period of 12 weeks (excluding the *CNTNAP2* experiments). Data collection and analyses were not performed blind to the conditions of the experiments. No statistical methods were used to predetermine sample sizes, but our sample sizes were similar to those reported in previous publications[1].

***CNTNAP2* mutant mouse experiments.** Male WT or mutant littermates from breeding pairs of heterozygous *CNTNAP2* mutants (Jackson Laboratories stock number 017482) were subjected to acute drug or saline injections as described above. Data from mice included in the analyses (for WT mice, $n=39$ with saline, $n=20$ with 0.1 mg per kg risperidone, $n=4$ with 0.5 mg per kg loxapine, $n=8$ with 20 mg per kg sulpiride; for *CNTNAP2* mice, $n=9$ with saline, $n=4$ with risperidone, $n=6$ with loxapine, $n=5$ with sulpiride) were separately modeled from the remainder of the drug data (see below).

Behavioral recording. Data acquisition was performed identically as previously described[9] using three parallel set-ups to maximize throughput. Mice were tracked in 3D using a Kinect for Windows v.1 (Microsoft). This camera projects structured infrared light onto the imaging field, and the 3D position of objects in the imaging field are computed based on parallax. A boom tripod (Manfrotto) was used to suspend the camera above the recording arena, affording a stable top-down view of the mouse. The Kinect v.1 has a minimum working distance (in near mode) of 0.5 m; by quantitating the number of missing depth pixels within an imaged field, we found that the optimal sensor position data are between 0.6 and 0.75 m depending on ambient light conditions and assay material.

Data from the Kinect were sent to an acquisition computer (hand-assembled, 16-GB RAM, Intel i7 CPU, 512-GB SSD) via a USB. A custom Matlab script was used to interface the Kinect via the official Microsoft.NET API that retrieves depth frames at a rate of 30 frames per second and saves the frames in raw binary format (16-bit unsigned integers) to disk. Relevant experimental metadata (mouse ID, drug ID and dose) were captured and saved in the same folder name into which the raw binary depth data recorded to disk. Because USB 3.0 has sufficient bandwidth to allow streaming of the data to an external hard drive in real-time, hot-swappable external hard drives were used for all data storage. After completion of the experiment, a region of interest was specified to delineate the area where the mouse could feasibly explore. This polygon was saved alongside the depth data and used to simplify the data extraction process by eliminating pixels outside the arena.

Data preprocessing and extraction. Raw frames recorded to external hard drives were immediately copied to the network-attached storage (NAS) associated with the Harvard Medical School Orchestra cluster. Custom mouse-tracking software was then run to extract the position, orientation and body morphometry of the mouse from the raw depth data. All extraction software was implemented in the Python programming language, using the MPI4Py, H5Py, joblib, pandas, OpenCV, Scikit-Learn, Scikit-Image, MoviePy, NumPy and SciPy libraries.

To extract and align the 3D image of the mouse from the video data, raw frame depth frames were first read in as rectilinear blocks of unsigned 16-bit integers, and then these bits were shifted right by three places, yielding distance measurements in millimeters. A background image, used for background subtraction, was then calculated by taking the median value of the first 1,000 frames of the recording. Noise in the depth image is highly correlated in both space and time due to the structured-illumination technique used to acquire depth information. Missing data were imputed by replacing missing depth pixel values with the spatially nearest valid pixel in both space and time. The raw depth images were resampled so that every pixel covered 2 square millimeters, using the published properties of the field of view of the camera. The resampled images were re-centered by subtracting them from the background image, yielding values indicating how high a given pixel is above the baseline background image. All negative values (portions of the image below the background, usually occurring because of spurious noise) were set to zero. All values above a maximum height (200 mm) were set to zero. Objects above the background that were smaller than a mouse were removed with morphological image operations using the Scikit-Image "remove_small_objects" and "binary_opening" functions. After these cleaning operations, the largest contiguous group of non-zero values in each frame is the body of the mouse, which was identified with the OpenCV "findContours" function. From this contour polygon, the area, the center-of-mass, the orientation and, using the "fitEllipse" function from OpenCV, the best-fit ellipse for each mouse were calculated. A square view measuring 120 mm × 120 mm centered on the mouse was then extracted in every frame using the center-of-mass and orientation of the mouse contour; the major axis of the ellipse defining the mouse was oriented along the horizontal axis of the square view.

Although, in an ideal case, this procedure would yield a square field of view in which a mouse was aligned horizontally along the virtual axis of its spine, in reality, the best-fit ellipse is not necessarily oriented in the direction of the head of the mouse. To correctly identify the head of the mouse, a random forest classifier was generated using Scikit-Learn and trained on a corpus of several thousand hand-oriented extracted mouse images. After acquiring a properly oriented extracted mouse image, and associated contour and positional data, the resultant aligned mouse movie was written to a HDF5 file. To accelerate the extraction process, the extraction over overlapping time-chunks of the experiment was parallelized using MPI. A recording from a single mouse was extracted into a single HDF5 file, and, for convenience, all mice were concatenated together into one central HDF5 file containing the entirety of the recorded data used in this study.

Data modeling. Once extraction of all experiments completed, the extracted data contained in a single HDF5 file were moved to a customized Starcluster on-demand high-performance compute cluster, hosted on Amazon Web Services Elastic Compute Cluster (EC2). Many of the processing steps either benefit from many CPU cores or require a very high memory budget, so much of the analysis was performed on a ×1.32xlarge EC2 machine, with 128 virtual CPU cores and 2 terabytes of onboard RAM. All cluster configuration and required code were saved on attached Elastic Block Store drives, and all imported data, and any further results of analyses, were saved on an attached Elastic File System drive, which was chosen because it did not require manual reformatting when additional storage was required. Local scratch drives were used for intermediate results that did not need persistence.

The extracted mouse images form a time-series that is 3,600 (60 pixels × 60 pixels) dimensional, sampled at 30 frames per second. These data were first dimensionally reduced using principal component analysis (PCA). All extracted mouse images were loaded into memory, and the RandomizedPCA model from Scikit-Learn was used to learn a ten-dimensional linear embedding of the image time-series. The principal component (PC) time-series was then whitened across all mice to remove covariance between PC dimensions. The PCs were saved onto the Elastic File System to avoid recomputing this step.

An autoregressive hierarchical Dirichlet process hidden Markov model (AR-HMM), identical to the model specified in Wiltschko et al.[9], was fit to the whitened PCs. All of the data were fit in a single model, except for the *CNTNAP2* data, which was separately modeled. Hyperparameters were validated via held-out likelihood assessment and qualitative inspection. Autoregressive observation distributions were initialized using empirical Bayes[41]. Kappa, the self-transition bias that controls the average duration of states, was set to produce states with duration distributions whose mode matches an independently specified changepoint detection model (Extended Data Fig. 3). The number of lags in the autoregressive distribution was selected with an automatic relevance detection prior and yielded the highest held-out likelihood (100 ms or 3 frames, see Wiltschko et al.[9]). As was observed in Wiltschko et al.[9], model output was insensitive to the hyperparameters of the hierarchical Dirichlet process prior. State sequences were randomly initialized. After initialization, the AR-HMM fit was burned-in with 1,000 iterations of Gibbs sampling, and then a maximum likelihood estimate was found using the Viterbi expectation-maximization algorithm. This model fitting procedure yielded 92 syllables capturing 95% of total frames in the main dataset (truncated to 90 syllables for convenience), and 67 syllables capturing 95% of total frames for the *CNTNAP2* experiment.

Data quality control. Data quality was assessed at several stages of the processing pipeline. First, each video recording was directly inspected to determine whether mouse tracking was successful. If there were persistent periods of the orientation of the mouse as being labeled as incorrectly flipped, these frames were added as new training data to the random forest flip classifier, described above, and the extraction procedure was run again. A heatmap of the location of the mouse body over the course of the entire experiment was next examined to identify any sharp boundaries or disproportionately bright areas that might indicate tracking of non-mouse objects. If a non-mouse object was tracked (typically the edge of the arena), the region of interest of the experiment was redefined, and the experiment was re-extracted. If, after applying all data quality correction methods listed above, the body of the mouse was not tracked and properly extracted, or more than 5% of total frames were dropped or unavailable, the recording was not used in the dataset or any further analyses.

Generating behavioral summaries. Preprocessed behavioral recordings of mice in the OFA were further summarized into fixed-length descriptions of behavior. A variety of summaries were constructed, based on the parameters described below.

Position. The center of a hand-drawn circle demarcating the edge of the OFA was considered the center of the arena. The 2D position of a mouse in the arena was subtracted from the center position of the circle. A histogram of these values was constructed with 90 bins equally spaced between 0 and 120 pixels.

Speed. Mouse speed was calculated as the absolute magnitude of the first time derivative of the 2D position of the mouse in the arena. A histogram of these values was constructed with 90 bins equally spaced between 0 and 20 pixels per frame.

Length. An ellipse was fit using the Python bindings of OpenCV to the top-down body contour of the animal in each recorded video frame. The length of the mouse for each frame was determined to be the length of the major axis of this ellipse. A histogram of these values was constructed with 45 bins equally spaced between 20 and 100 pixels.

Height. The height of the animal was determined to be the maximum height of the extracted mouse image in each frame. A histogram of these values was constructed with 45 bins equally spaced between 0 and 60 mm.

Length and height. The histograms of length and height were concatenated into a behavioral summary with 90 dimensions.

Acceleration. Mouse acceleration was calculated as the absolute magnitude of the second time derivative of the 2D position in the arena. A histogram of these values was constructed with 90 bins equally spaced between 0 and 5 pixels per frame^2.

Angle. A histogram of mouse orientation was constructed, in degrees, with 90 equally spaced bins between 0° and 360°.

Area. A histogram of the area of the best-fit ellipse to the top-down contour of the mouse was constructed, with 90 equally spaced bins between 0 and 12,000 pixels^2.

Ellipticity. A histogram of the ratio of a given length of a mouse to its width was constructed, which was derived from the best-fit ellipse of the top-down contour of the mouse, with 90 equally spaced bins between 1 and 3.

Width. A histogram of mouse width was constructed, which was derived from the best-fit ellipse of the top-down contour of the mouse, with 90 equally spaced bins between 20 and 50 pixels.

Scalars. The length, height, speed and position summaries were concatenated together.

Scalars++. We concatenated all of the parameters measured in the scalar summary together with the summaries for acceleration, angle, area, ellipticity and width.

MoSeq. MoSeq summaries were composed of a histogram describing the frequency of use of each of the 90 most-used syllables.

KMeans. We fit a KMeans model (using sklearn.cluster.KMeans method with kmeans++ initialization) on the PCs of aligned mouse images (the input to the MoSeq method) with varying numbers of clusters. The fingerprint was composed of the number of frames assigned to each cluster.

MoSeq on scalars. We fit an AR-HMM model on scalar data (as opposed to the PCs of aligned mouse images) using a four-dimensional time-series composed of the distance-to-center, the speed, the height and the length of the animal. To match the dimensionality of MoSeq, 90 states were used. The best-fit state sequence of the time-series data was summarized as a histogram of state frequencies, identically to the MoSeq summary described above.

Summaries are displayed (but not analyzed) in the paper as the square-root of their values to increase the visual dynamic range.

MoSeq-based behavioral distance measurements. To measure similarity between syllables, we performed MoSeq-based behavioral distance measurements as previously described[11]. Briefly, we assessed the similarity between pose trajectories of different syllables. We simulated pose trajectories for each syllable over 10 time steps (corresponding to 300 ms) using the autoregressive coefficients described by the AR-HMM model fit. Then, we computed the pairwise correlation distance (1 − Pearson's *r*) between the top 90 most used syllables to generate a distance matrix, where low distances (near 0) represent similar syllables and high distances (near 2) represent dissimilar syllables.

The cladogram was generated from the distance matrix using the Voor Hees hierarchical clustering algorithm (scipy.cluster.hierarchy.linkage).

Linear classification of behavioral summaries. Classification based on behavioral summaries was performed using logistic regression as implemented in the Scikit-Learn Python package. The underlying implementation took advantage of the liblinear C/C++ library, using a 'one-vs-rest' formulation of multi-class classification. A L2 weight penalty with an inverse regularization strength was also used. We scanned the values 0.01, 0.1, 1.0, 10.0 and 100.0 for each feature type, and presented results for the optimal choice per feature. To guard against overfitting, 500-fold cross-validation was performed using randomly shuffled folds with 10% of the data held-out per fold, keeping the relative proportion of each label the same in both train and held-out sets. To predict drug identity alone, data from all doses of a given drug were merged, and individual mice were held out. To predict drug class, data from all doses of all drugs belonging to a class were merged. For classification of drug pharmacological class, we used an additional stratification strategy, whereby all mice given a particular drug were placed in either the training or held-out set. We observed no appreciable difference in absolute or relative performance (data not shown). The mean and standard error of performance metrics on these randomly generated held-out folds are reported.

To evaluate performance, confusion matrices, precision–recall (PR) curves and F_1 scores were computed.

Each confusion matrix was a square matrix with each side length equal to the number of possible target labels, and each square indexed by *i*,*j* is the proportion of time a data point with true label *i* was classified as having label *j*. When $i = j$, the classifier correctly predicted the label. Confusion matrices were produced with the confusion_matrix function in Scikit-Learn. Matrices were normalized such that every row and column summed to one to indicate a probability of classification or misclassification. Held-out confusion matrices were calculated by repeating the linear model training and evaluation process *N* times, where *N* is the number of treatment groups. For each iteration, one target class was removed from the training set, but added into the held-out set for each fold. This forced the classifier to never correctly classify the removed treatment class and allowed analysis of the treatments the classifier deemed most similar to the target treatment class. This process was repeated for all treatments to generate the complete held-out confusion matrix that is presented.

Precision and recall are quantities computed from the number of true positives, t_p, the number of false positives, f_p, and the number of true negatives, t_n. Precision and recall are defined as follows:

$$\text{Precision} = \frac{t_p}{t_p + f_p}, \quad \text{recall} = \frac{t_p}{t_p + f_n}$$

The PR curve is a plot of the precision and recall of the model, as a decision threshold is varied. The curve is calculated for binary prediction problems by varying the decision threshold for binary predictions (for example, classifying a mouse as having received a specific drug versus not having received any other drug) and measuring the false-positive and true-positive rates at that decision threshold for all data in the validation set.

The F_1 score is the harmonic mean of precision and recall, and is a measure of binary classification performance as follows:

$$F_1 = 2\times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

The per-label class F_1 values were calculated using the f1_score function in Scikit-Learn. Class-weighted averaging was used across the F_1 score of all classes to report a single mean F_1 score for a behavioral summary; standard errors were also calculated.

Behavioral summary distance comparisons. *Cosine distance matrix.* Distances between two summaries *u* and *v* were directly assessed using the cosine distance, which was computed (using the SciPy Python package) as follows:

$$c(u, v) = 1 - \frac{u \cdot v}{||u||_2 \, ||v||_2}$$

The cosine distance was used because it is bounded between 0 and 2, thereby allowing comparisons between behavioral summaries with different units. Within- and between-treatment cosine distances were also computed. The between-treatment cosine distance was calculated as follows:

$$B(i,j) = \frac{1}{N_{\mathrm{p}}} \sum_{u \in G_i,\, v \in G_j,\, u \neq v} c(u,v)$$

where G_i is the set of behavioral summaries given treatment i, and N_{p} is the number (u,v) pairs in the sum. The within-treatment distance was calculated when $i=j$. The ratio of the within-treatment and between-treatment cosine distances was calculated as follows:

$$\frac{\frac{1}{N_{\mathrm{t}}}\sum_i B(i,i)}{1+\frac{1}{N_{\mathrm{t}}}\sum_{j \neq i} B(i,j)}$$

Where N_{t} is the number of treatments.

To visually highlight the relationships between behavioral summaries, we reordered a square matrix containing all pairwise cosine distances using hierarchical clustering (Ward's linkage) implemented by the SciPy Python scientific computing package.

Identifying syllables critical for classification. LASSO regression was used to identify how many syllables were on average needed to distinguish treatments in an all-to-all comparison. LASSO regression is a L1-regularized logistic regression; the regularization term was scaled from zero to a maximum value where no syllables were used, which results in random predictions. We densely sampled the L1 penalty so that we evenly sampled the number of used syllables. For each L1 value, we recorded the area under the receiver operating characteristics for each drug treatment, and the number of syllables with non-zero weight was used in the classifier.

To identify which syllables were most discriminative for a particular drug treatment (either relative to control or relative to all other drugs), a *F* univariate statistical test was used. We reasoned that syllables whose usage frequency in mice was statistically independent of the drug treatment given to mice would not be useful for linear classification. Conversely, syllables with high statistical dependence on the drug treatment would be useful for classification and therefore characteristic of a given treatment.

Visualizing behavioral summaries with low-dimensional embeddings. To visualize the relationship between drug treatments, as measured by behavioral summaries, we calculated low-dimensional 2D embeddings from MoSeq behavioral summaries. We used the LDA[42] algorithm to calculate a linear 2D projection of the MoSeq summaries that maximizes linear separability between all drug classes. We used the Scikit-Learn function call with the following defaults: discriminant_analysis.LinearDiscriminantAnalysis(solver='svd', n_components=2).

Calculating effective dimensionality of behavioral summaries. To quantify the effective dimensionality of both scalar and MoSeq behavioral summaries, we used both PCs and a method from Fukunaga and Olsen[43]. For the PCA method, we used Scikit-Learn's sklearn.decomposition.PCA method to calculate the number of components that were required to explain 95% of the variance in the behavioral summary data. Note that for this analysis, we apply PCA to the behavioral summaries output by MoSeq and the scalar analysis, not to the raw mouse depth images. For the Fukunaga and Olsen method, we calculated the eigenvalues of the behavioral summary array, normalized them so that their values fall between 0 and 1, and counted the number that fall above a threshold of 0.01.

Stratifying and classifying drug treatments by induced movement speed. Multiple doses of clozapine, haloperidol and risperidone were given to mice, each of which slowed overall mouse movement speed in reference to control treatment. We stratified the treatments by the mean movement speed of mice given the treatment to test whether a MoSeq fingerprint could disambiguate different drugs that each had an equal effect on overall locomotion. We bucketed each drug and dose into four movement speeds (very slow, slow, medium and fast) according to a four-component Gaussian mixture model fit on the full distribution of mean mouse movement speeds. The average movement speed in each group was 7 mm s^{-1}, 21 mm s^{-1}, 42 mm s^{-1} and 76 mm s^{-1}, respectively. The very slow and slow speeds were combined into a single slow movement speed bucket. The threshold movement speed dividing the slow and medium speed groups was 24 mm s^{-1}, and the threshold dividing the medium and fast groups was 53 mm s^{-1}. For each of the treatments that were placed in the slow and medium groups, we trained a linear classifier, as described above, to predict the drug identity given to each mouse using MoSeq fingerprints.

Querying clinical main effects and side effects. Food and Drug Administration (FDA)-approved and non-FDA approved indications, as well as main side effects, were manually scraped for each drug from the IBM Micromedex database (http://truvenhealth.com/Products/Micromedex).

Statistical tests. Error bars refer to the 95th percent confidence interval, standard error of the mean (s.e.m.) or standard deviation (s.d.) as indicated. For statistical tests that assumed normality, data distributions were assumed to be normal, but this was not formally tested.

Statistical differences in the mean scalar measurements of behavior between methylphenidate, haloperidol and saline treatments in Fig. 2f were established using the two-sided Mann–Whitney *U*-test. The mean, per mouse, for each of speed, length, height and distance from arena were first calculated. We then applied a two-sided Mann–Whitney *U*-test to assess whether treatments had either significantly greater or smaller values. The resultant *P* values for the four comparisons were then adjusted using the Holm–Bonferroni step-down procedure. For MoSeq summaries, which are not easily reduced into single scalar metrics per mouse, the significance between each of the three aforementioned treatments was assessed using a two-factor multivariate analysis of variance (MANOVA). The MANOVA calculation was performed using the R statistical language.

F_1 scores were tested for statistically significant differences using the two-sample *t*-test. F_1 scores were first calculated for each unique label (each drug identity irrespective of dose in Fig. 3, each pharmacological class in Fig. 5 and each unique drug–dose pair in Extended Data Fig. 6) on each held-out fold (of 500 total folds as described above). F_1 scores were compared between summary types using the two-sample *t*-test, with multiple comparison correction using the Holm–Bonferroni step-down procedure, with significance set at $P < 0.05$ after correction.

Differentially used behavioral syllables in the *CNTNAP2* experiment were identified using the Kruskal–Wallis and Dunn's post-hoc two-sided tests with permutation. In the Kruskal–Wallis test, for each syllable, we calculated the *H*-statistic from the actual data (*H*-data) and from the permuted data in which group labels were randomly shuffled for all four groups (*H*-permutation). Raw *P* values were then established by calculating the ratio of permutations where the *H*-permutation is larger than the *H*-data, and these *P* values were corrected using the Benjamini–Hochberg false discovery rate (FDR) across syllables. Syllables with FDR < 0.05 were identified as significant. For each of the syllables that passed the Kruskal–Wallis test, we then performed a Dunn's post-hoc test by calculating the *z*-statistic both from the actual data (*z*-data) and from the permuted data in which group labels of corresponding two groups were shuffled (*z*-permutation). We established the raw *P* values by calculating the ratio of permutations for which the *z*-permutation is larger than the *z*-data, and then corrected those *P* values using Benjamini–Hochberg FDR across all pairwise comparisons. Syllables with FDR < 0.05 were identified as significant.

For syllables differentially used between WT and *CNTNAP2*$^{-/-}$ mice treated with saline control, we considered the usage is fully reverted if a given syllable satisfied two criteria. First, a given syllable is within one standard deviation (of the overall differences in syllable usage observed between WT and *CNTNAP2*$^{-/-}$ mice) between WT mice treated with saline and *CNTNAP2*$^{-/-}$ mice treated with the drug. Second, that syllable is significantly different between *CNTNAP2*$^{-/-}$ mice treated with saline and *CNTNAP2*$^{-/-}$ mice treated with the drug. A given syllable is considered as "partially reverted" if it only satisfied one of these criteria, and considered "not reverted" if neither of these criteria was satisfied. Syllables were considered "side effects" if there was no statistical difference in their level of expression in WT and *CNTNAP2* mice, but treatment of the *CNTNAP2* mice with drug induced a significant change between the genotypes. Syllables in Fig. 7 and Supplementary Fig. 5 are sorted on the basis of how different their usage is in the *CNTNAP2*$^{-/-}$ and WT saline control mice (mutant − WT)/(mutant + WT).

We assessed whether the variability of syllable usage within each mouse met, exceeded or was less than the variability between mice given the same treatments or across different treatments. To quantify within-mouse variability, we randomly sampled the syllable labels for 1,000 frames with replacement, and constructed a MoSeq fingerprint using the labels associated with those frames (of the 36,000 total frames available per mouse) and measured the mean and standard deviation of all unique pairwise cosine distances after repeating that procedure 100 times. To measure between-mouse variability (either for mice given the same or different treatments), we computed the mean and standard deviation of all unique pairwise cosine distances.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All datasets generated and/or analyzed during the current study will be available from the corresponding author upon reasonable request. The raw per-frame data, MoSeq per-frame labels and per-mouse behavioral summary data organized as NumPy arrays are stored in a Python pickle file and are available for download on an open-access basis via GitHub (https://github.com/dattalab/moseq-drugs).

Code availability

All code used in this manuscript will be made available on GitHub at https://github.com/dattalab/moseq-drugs.

References

41. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
42. McLachlan, G. J. *Discriminant Analysis and Statistical Pattern Recognition* (Wiley, 2004).
43. Fukunaga, K. & Olsen, D. R. An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. Computers* **20**, 176–183 (1971).

Acknowledgements

We thank members of the Datta Lab for helpful comments on the manuscript. We thank O. Mazor and P. Gorelik from the Research Instrumentation Core Facility for engineering support, S. Knemeyer for mouse illustrations and C. Ashton for technical assistance. Core facility support is provided by NIH grant P30 HD18655. Pilot experiments for this paper were supported by Hoffman LaRoche. S.R.D. is supported by the National Institutes of Health (U24NS109520, RO11DC016222, U19NS113201 and RO1NS114020), by a SFARI grant from the Simons Foundation and by the Simons Collaboration on the Global Brain.

Author contributions

A.B.W., T.T., R.E.P. and S.R.D. conceived and designed experiments. A.B.W., T.T., A.Z., R.A., R.E.P. and J.K. acquired behavioral data and performed data pre-processing. A.B.W., T.T., A.Z., R.A., J.E.M., W.F.G. and M.J.J. carried out data analyses. A.B.W. and S.R.D. wrote the manuscript.

Competing interests

The authors declare the following competing interests: A.B.W., M.J.J. and S.R.D. are co-founders of Syllable Life Sciences, Inc. A.B.W. and S.R.D. are co-authors on awarded patents WO2013170129A1 and US10025973B2, which describe behavioral methods used herein.

Additional information

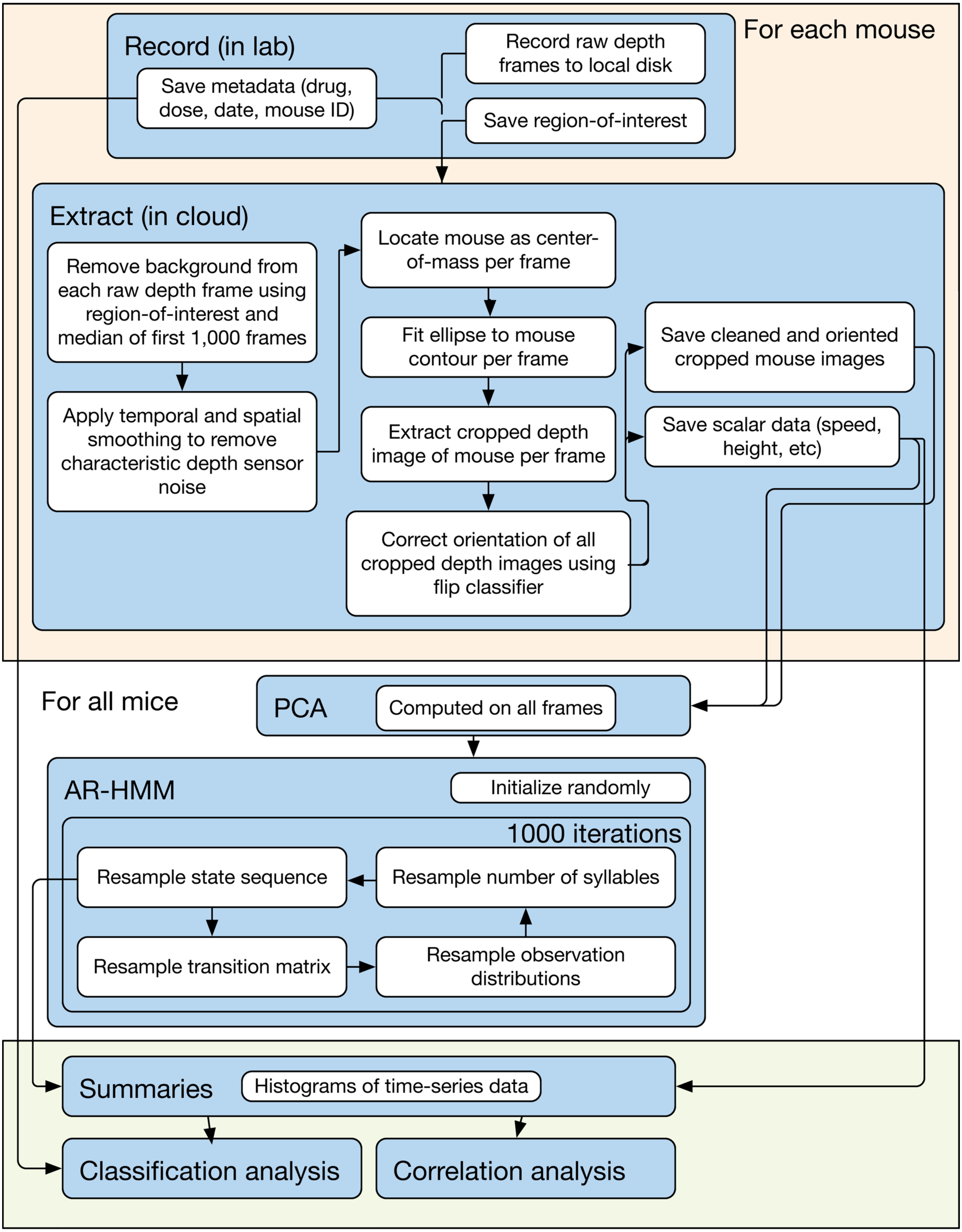
Extended data is available for this paper at https://doi.org/10.1038/s41593-020-00706-3.

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-020-00706-3.

Correspondence and requests for materials should be addressed to S.R.D.

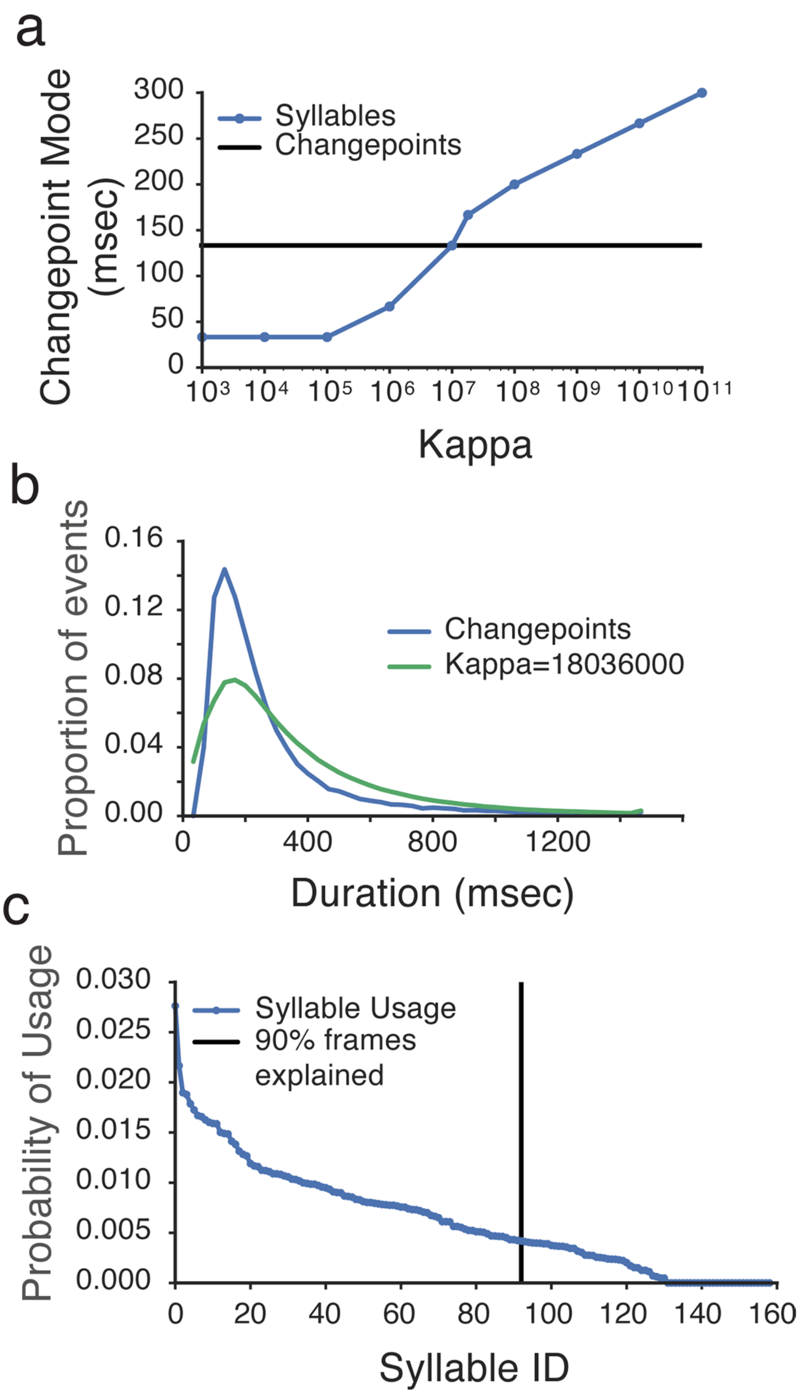
Peer review information *Nature Neuroscience* thanks Ann Kennedy, Paul Kenny, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

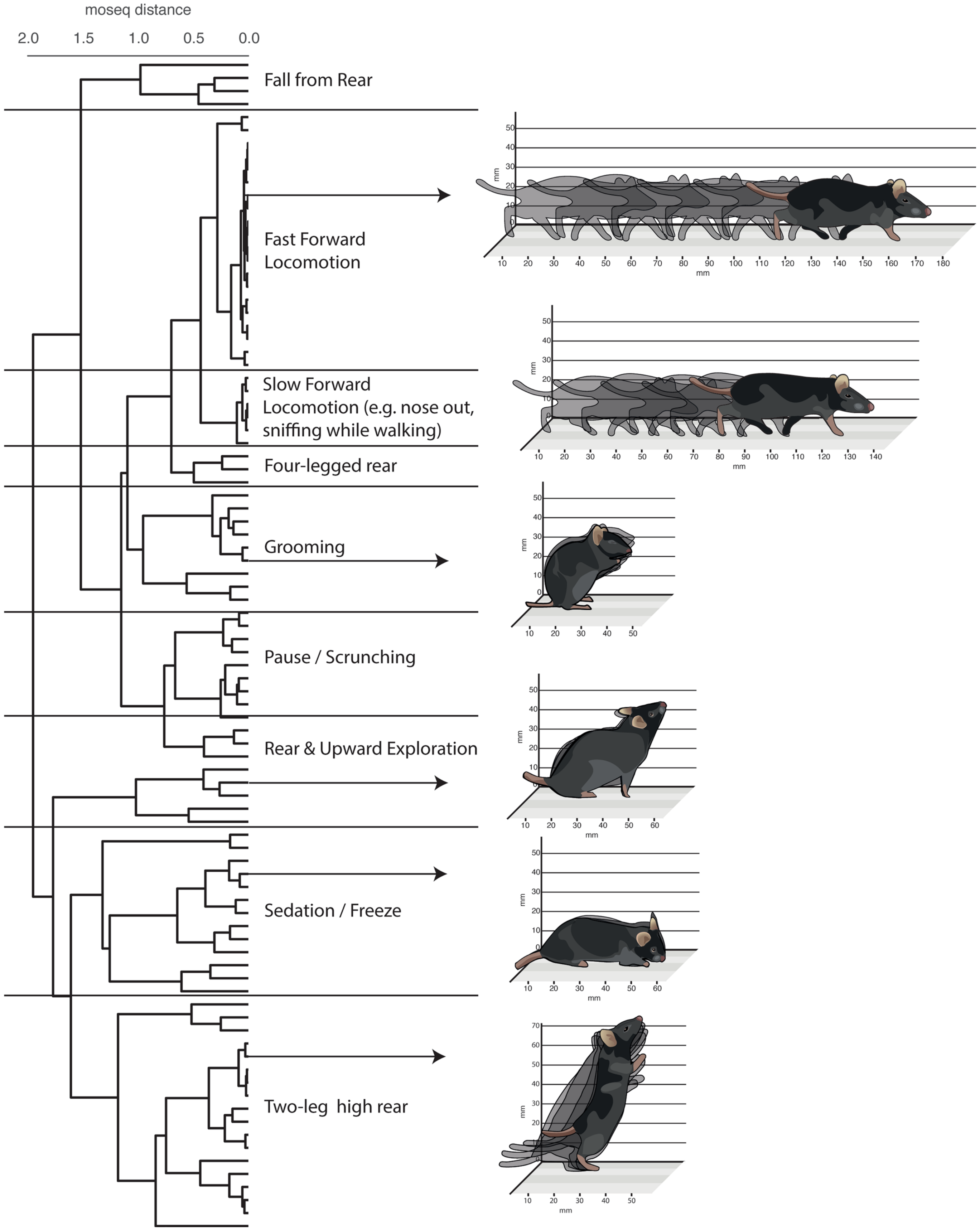

For each mouse
Record (in lab)
Save metadata (drug, dose, date, mouse ID)
Record raw depth frames to local disk
Save region-of-interest
Extract (in cloud)
Remove background from each raw depth frame using region-of-interest and median of first 1,000 frames
Apply temporal and spatial smoothing to remove characteristic depth sensor noise
Locate mouse as center-of-mass per frame
Fit ellipse to mouse contour per frame
Extract cropped depth image of mouse per frame
Correct orientation of all cropped depth images using flip classifier
Save cleaned and oriented cropped mouse images
Save scalar data (speed, height, etc)
For all mice
PCA
Computed on all frames
AR-HMM
Initialize randomly
1000 iterations
Resample state sequence
Resample number of syllables
Resample transition matrix
Resample observation distributions
Summaries
Histograms of time-series data
Classification analysis
Correlation analysis

Extended Data Fig. 1 | See next page for caption.

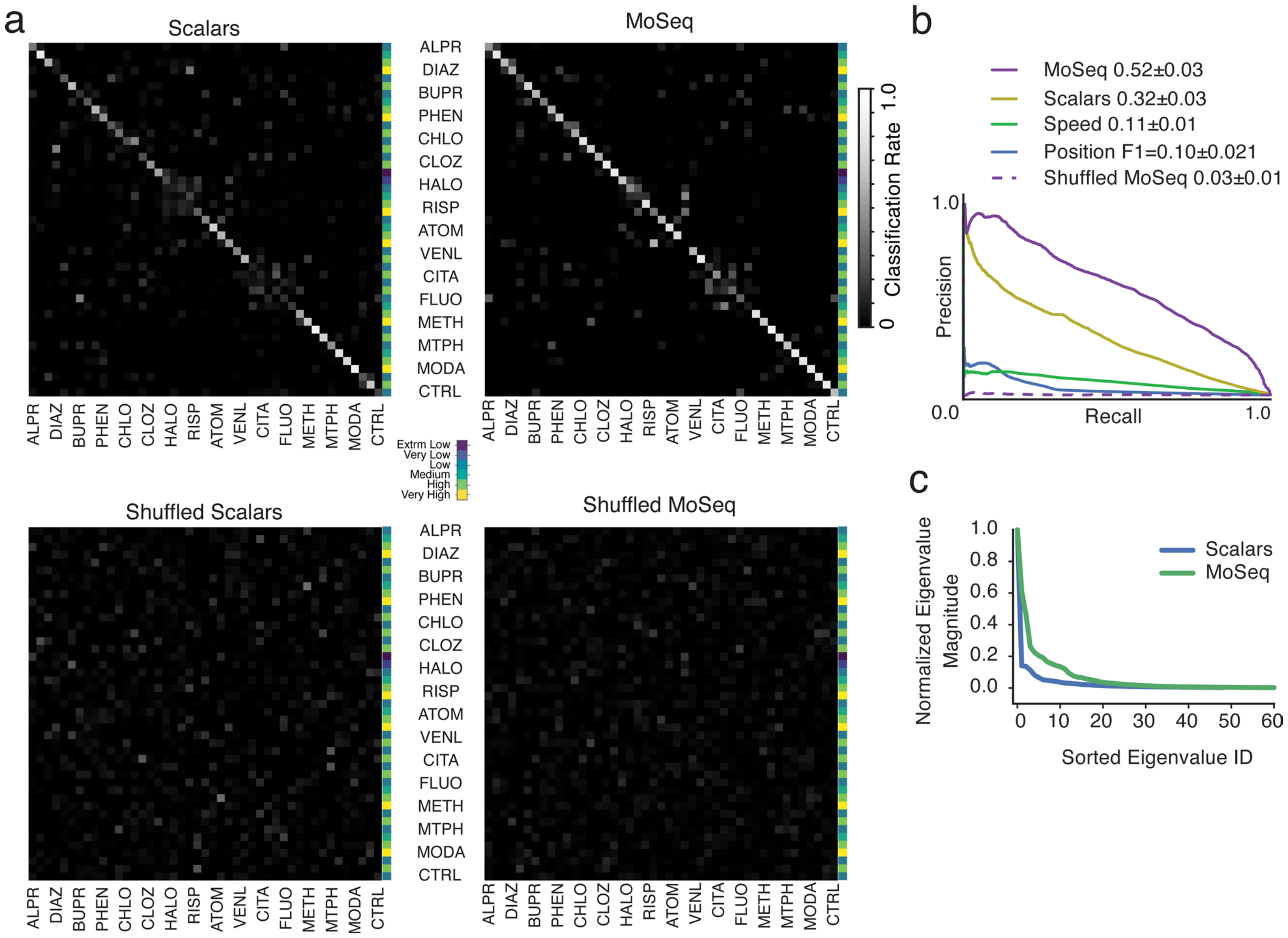
Extended Data Fig. 1 | Workflow for classifying drug effects in mice using scalars and MoSeq. Depth cameras are used to capture 3D video data encapsulating mouse postural dynamics in the open field. These data are saved locally before being uploaded to the cloud, where the videos are denoised and aligned. The image of the mouse is then extracted from the larger image; at this step, scalar behavioral metrics (like the position of the mouse within the arena, or its velocity) are computed. After extraction, aligned 3D mouse images are analyzed either locally or in the cloud, depending upon resource demands. 3D mouse images are compressed by PCA (for ease of computation), then these data are used to train an AR-HMM (as in Wiltschko et al[9]). The output of this training procedure is the optimal set of behavioral syllables that describe the 3D pose dynamics observed within the experiment (each of which is described as an autoregressive process through pose space). Every frame of the imaging data is then labeled with behavioral syllable MoSeq considers most likely, thereby revealing the behavioral grammar that governs the transitions from any given syllable to any other syllable. Herein, each mouse is characterized by a MoSeq behavioral summary that includes only information about how often each behavioral syllable is expressed during the experiment (without consideration of the syllable transition matrix), whereas the scalar summary includes a wide variety of data describing the mouse's behavioral comportment (including height, length, speed, position). These MoSeq and scalar behavioral summaries are then submitted to linear classifiers to predict the identity of the drug, drug and dose, or drug class to which each mouse was exposed.



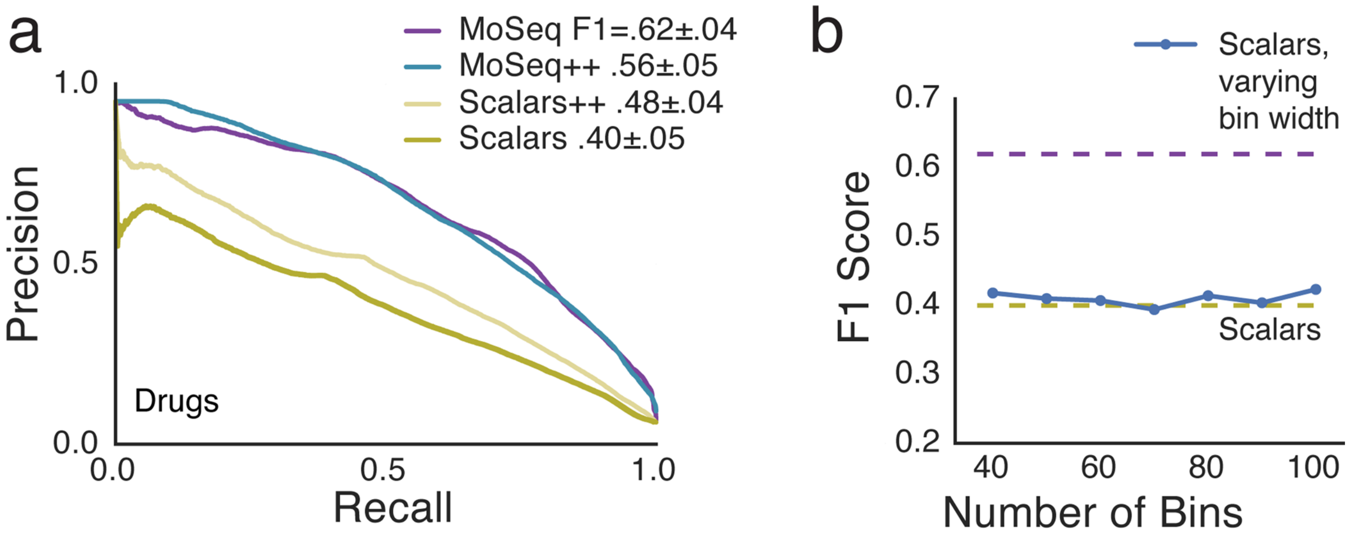
Extended Data Fig. 2 | Tuning MoSeq parameters. **a**, Scanning the MoSeq kappa parameter (which sets the timescale at which syllables are identified) reveals a value at which the modal syllable length matches the model-free block length identified by changepoints analysis (see Methods). **b**, The mode of the syllable duration distribution established by MoSeq, given the kappa established in **a**, matches that for the model-free changepoint distribution. **c**, Ninety percent of the total frames are explained by 92 behavioral syllables; for the sake of simplicity herein we analyze the top 90 syllables.



Extended Data Fig. 3 | Cladogram of syllables with representative illustrations. A cladogram describing behavioral relationships among syllables was computed using hierarchical clustering performed on the autoregressive matrices describing all syllables (see Methods). Nine general behavioral categories were identified after visual inspection and given natural language names. Illustrations are representative of syllables in each category.



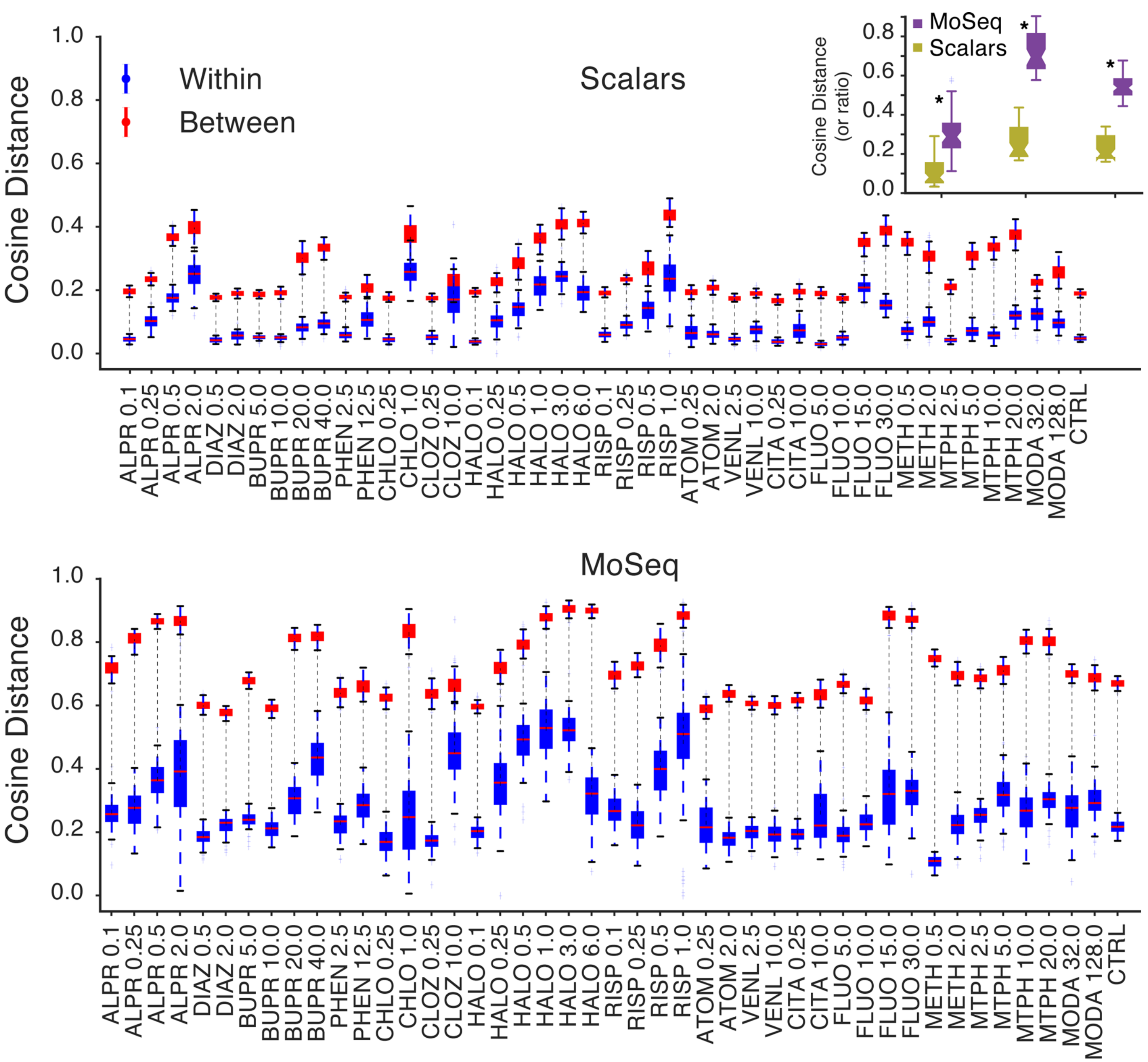
Extended Data Fig. 4 | MoSeq outperforms scalar metrics at all-vs-all drug discrimination. a, Normalized confusion matrices as in Fig. 3a, but computed for all drug/dose combinations. For the shuffled control (bottom row), syllable labels were shuffled on a per-mouse basis to compute a baseline of expected random performance. Heat map indicates classification successes and errors (see Methods for summary definitions). **b**, Mean precision-recall curves for all drugs and doses, computed for each behavioral summary type. **c**, The Fukunaga and Olsen method[43] was used to estimate the effective dimensionality of both scalar and MoSeq summaries; this analysis demonstrated that that MoSeq has a higher effective dimensionality than scalars (34 versus 26 dimensions), using a threshold value of 0.01 (see Methods).



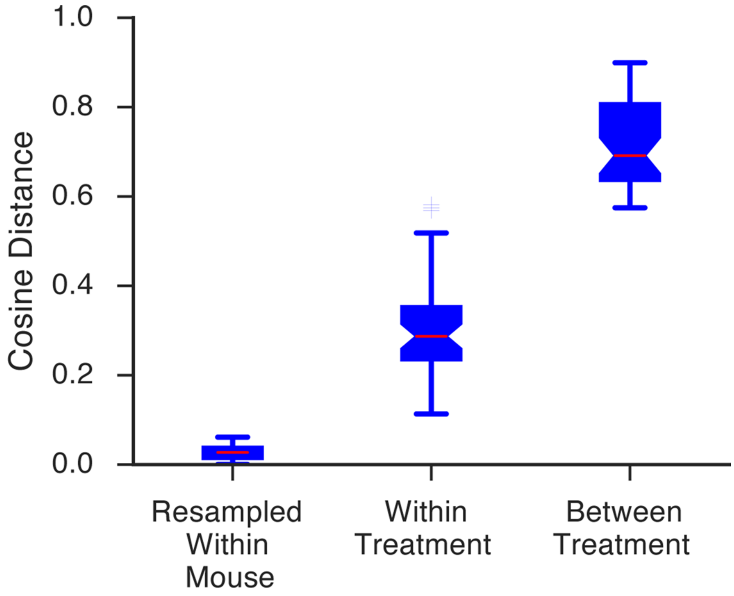
c

	Model	
Summary Type	Linear SVC (F1)	Logistic Regression (F1)
MoSeq	0.67 ± 0.04	0.62 ± 0.04
Scalars	0.36 ± 0.05	0.40 ± 0.05
Scalars: PCA 95%	0.45 ± 0.05	0.46 ± 0.05

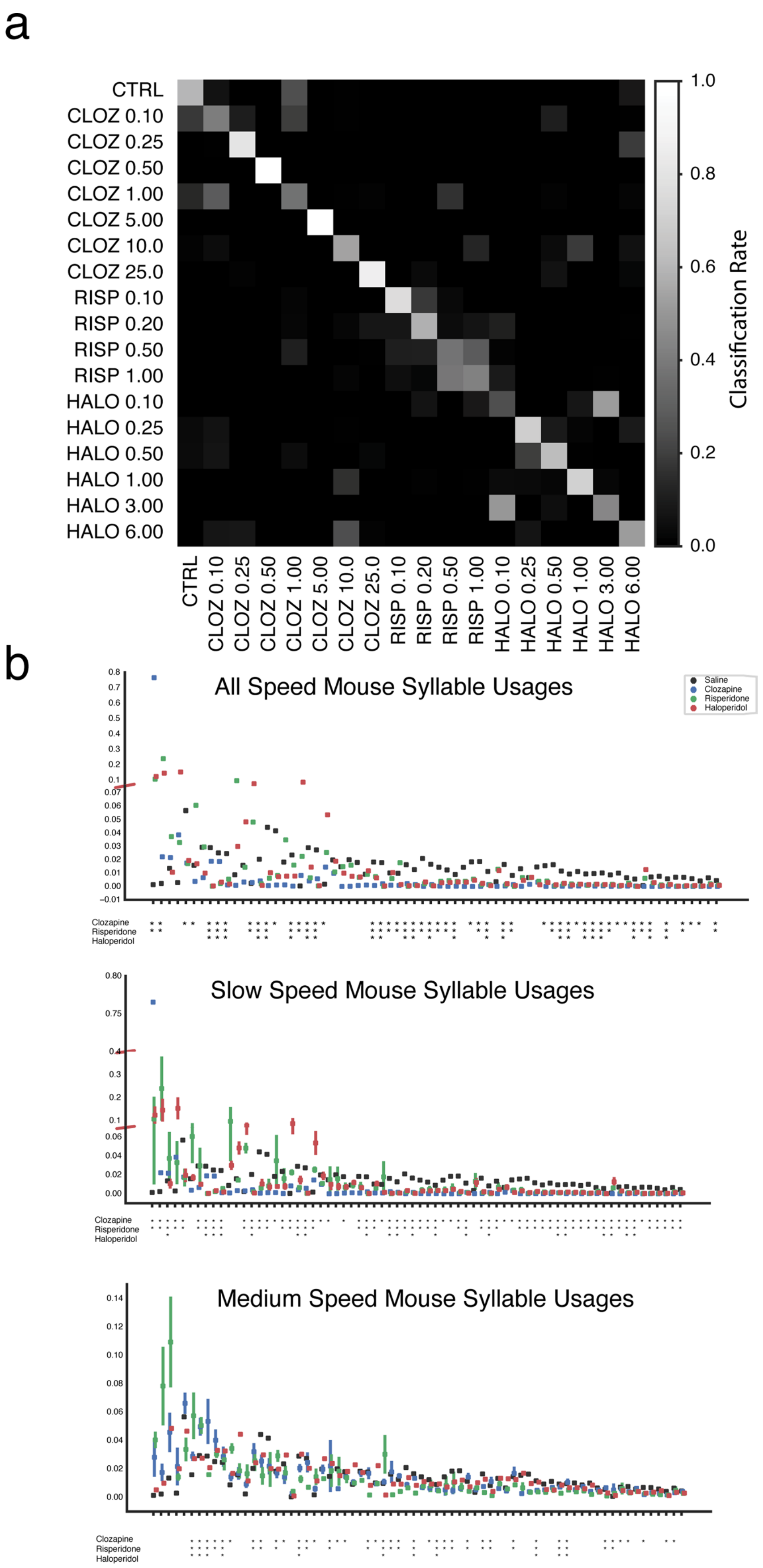
Extended Data Fig. 5 | Adding additional information to behavioral summaries or altering summary dimensionality does not improve performance. a, Additional information was added to the MoSeq and scalar behavioral summaries used to predict drug identity. For "MoSeq+ +," the empirical transition matrix derived from the syllable label sequence was calculated, flattened, and concatenated to the syllable usage frequency information. For "Scalars+ +," histograms of mouse acceleration, the mouse's heading, the area contained by the mouse's body contour, the ellipticity of the best-fit ellipse around the mouse's contour, and the mouse's width were added to the initial scalar behavioral summary. **b**, The granularity of the bins used to generate scalar behavioral summaries was systematically varied; bin size did not affect classification performance. **c**, To ensure that the higher dimensionality of the scalar summaries did not adversely affect performance, behavioral summaries containing scalars were also subjected to PCA to assess the consequences of dimensionality reduction (keeping the number of dimensions required to capture 95 percent of the variance; for scalars this is 33 dimensions); although performance was modestly improved, performance did not equal that observed for MoSeq.



Extended Data Fig. 6 | Exploring behavioral similarities elicited by specific drug/dose pairs. Average cosine distance ±1 standard deviation of mice given the same drug/dose pair (blue) and mice given different drug/dose pairs (red) using either scalar- (top) or MoSeq-based behavioral summaries (bottom). The difference observed between mice given the same drug/dose pair and different drug/dose pairs is uniformly larger when behavior is summarized using MoSeq when compared to scalars. Inset: summary of mean within- and between- class differences and their ratio for either scalar- and MoSeq-based analysis. MoSeq shows larger differences (two-sided paired t-test, $p < 0.05$, stars indicate statistically significant differences between MoSeq and scalars).

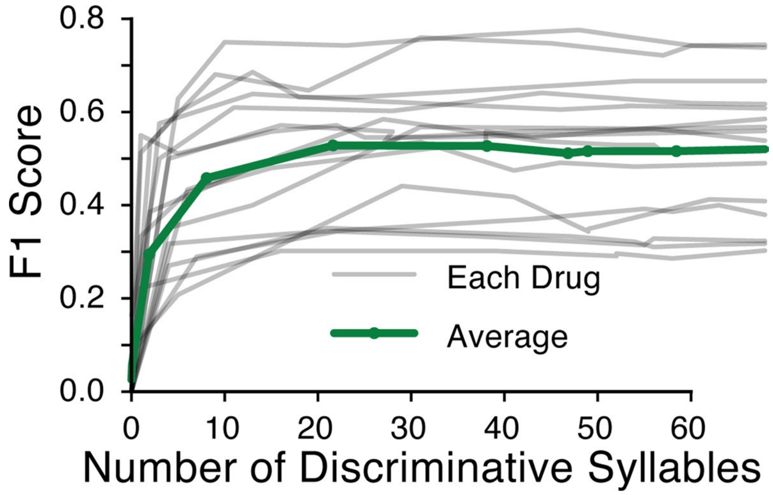


Extended Data Fig. 7 | MoSeq captures the behavioral variability of individual mice. To test whether the cosine distances that separate individual mice within a treatment class reflect individual variability or technical noise, we subsampled the data from each individual mouse and then asked how these sub-samples of each individual mouse compared to each other; observing low variability in these sub-samples would be consistent with each individual mouse expressing a stable set of behavioral syllables within an experiment, and with the within-condition variability observed across mice reflecting differences in individual mouse responses to a given drug and dose. In specific, within-mouse variability of MoSeq was assessed by randomly picking 1000 frames (with replacement) of the 3D imaging data (which for each mouse was constituted of approximately 36,000 frames), identifying the syllable associated by MoSeq with that frame, and then using those syllable labels to compute overall syllable usages; this procedure is roughly equivalent to randomly choosing less than one third of the syllables to quantify the pattern of syllable usage within a mouse. We repeated this procedure 100 times, and by computing cosine distances between each sub-sample within-mouse variability could be assessed. The bootstrapped estimate of individual variability (Resampled Within Mouse) was lower than the treatment-induced variability (Within tTreatment), as measured by the cosine distance between all pairs of mice given the same treatment, and was also lower than the cosine distance between pairs of mice given different treatments (Between Treatment). Thus the observed within-treatment variability reflects stable differences in behavior expressed by individual mice.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | MoSeq distinguishes the behavioral effects of drugs independent of effects on mouse movement speed. a, Similar as Fig. 3a, but classifying drug/dose identity instead of drug identity, across the entire risperidone, haloperidol, clozapine dose-response experiment. Many significant syllables that differentiated drug-treated mice from controls were, by inspection, behaviors like grooming or rearing that do not include significant two-dimensional velocity components (data not shown). **b**, Syllable usages for all mice and all drug/dose combinations (top), doses which resulted in slow mouse movement speed (middle) or moderate movement speed (bottom). Slow and medium speeds (relative to normal) were identified via a Gaussian Mixture Model (mean centroid speed of saline control mouse = 74 mm/sec; "medium speed" = 54 mm/sec; "slow speed" = 24 mm/sec; see Methods). Significant differential syllable usage for each drug versus control indicated with an asterisk (Kruskal-Wallis and post-hoc Dunn's two-sided test with permutation, with Benjamini/Hochberg FDR with alpha = 0.05).



Extended Data Fig. 9 | High classification performance by MoSeq is supported by a limited set of syllables. Sparsification reveals the number of syllables required to correctly distinguish each drug, as assessed by F1 scores emerging from linear classifiers trained on subsets of syllables (see Methods).

nature research

Corresponding author(s): Sandeep Robert Datta

Last updated by author(s): July 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Data was collected using the Microsoft Kinect SDK (v1.0) with custom LabVIEW software (v 2014) interfacing with the camera's .NET drivers.

Data analysis

Data was extracted and analyzed using open-source Python tools. The versions of the libraries are as follows.
All algorithms utilized in the text, excluding the core MoSeq algorithm, were implemented in the libraries below.

```
accelerate=2.3.1=np111py27_0
accelerate_cudalib=2.0=0
agate=0.7.0=pypi_0
altair=2.2.2=pypi_0
asn1crypto=0.24.0=py27_0
autograd=1.1.3=pypi_0
autoregressive=0.1.0=pypi_0 # https://github.com/mattjj/pyhsmm-autoregressive
awscli=1.10.1=pypi_0
backports=1.0=py27_0
backports_abc=0.4=py27_0
baker=1.3=pypi_0
blas=1.0=mkl
boto=2.39.0=pypi_0
botocore=1.3.23=pypi_0
ca-certificates=2019.1.23=0
cairo=1.12.18=6
cairocffi=0.6=pypi_0
certifi=2019.3.9=py27_0
cffi=1.12.1=py27h2e261b9_0
```

```
chardet=3.0.4=py27_1
click=4.0=pypi_0
colorama=0.3.3=pypi_0
conda=4.6.9=py27_0
conda-env=2.6.0=1
configparser=3.5.0=pypi_0
crl=0.0.0=dev_0
cryptography=2.5=py27h1ba5d50_0
cudatoolkit=7.0=1
cycler=0.10.0=py27_0
cython=0.24=py27_0
dask=0.10.0=pypi_0
decorator=4.0.10=py27_0
dill=0.2.4=pypi_0
django=1.8=pypi_0
docutils=0.12=pypi_0
dpca=0.1=pypi_0
ecdsa=0.13=pypi_0
entrypoints=0.2.3=pypi_0
enum34=1.1.6=py27_0
ez-setup=0.9=pypi_0
fabric=1.10.2=pypi_0
fastcluster=1.1.20=pypi_0
fontconfig=2.11.1=5
freetype=2.5.5=1
funcsigs=1.0.2=py27_0
functools32=3.2.3.2=py27_0
future=0.15.2=pypi_0
futures=3.0.5=pypi_0
geos=3.4.2=0
get_terminal_size=1.0.0=py27_0
gitdb=0.6.4=pypi_0
gitpython=1.0.0=pypi_0
gizeh=0.1.10=pypi_0
glances=2.5.1=pypi_0
glib=2.43.0=0
graphviz=2.38.0=1
h5py=2.8.0=py27h989c5e5_3
harfbuzz=0.9.39=0
hdf5=1.10.2=hba1933b_1
idna=2.8=py27_0
imageio=1.2=pypi_0
intel-openmp=2019.3=199
ipaddress=1.0.18=py27_0
ipdb=0.8.1=pypi_0
ipykernel=4.3.1=py27_0
ipyparallel=5.0.1=pypi_0
ipython=4.0.0=pypi_0
ipython-cluster-helper=0.5.1=pypi_0
ipython_genutils=0.1.0=py27_0
ipywidgets=4.1.1=py27_0
jbig=2.1=0
jdcal=1.0=pypi_0
jinja2=2.8=py27_1
jmespath=0.9.0=pypi_0
joblib=0.8.4=pypi_0
jpeg=8d=0
jsonschema=2.5.1=py27_0
jupyter=1.0.0=py27_3
jupyter_client=4.2.2=py27_0
jupyter_console=4.1.1=py27_0
jupyter_core=4.1.0=py27_0
lcms=1.19=0
libedit=3.1.20181209=hc058e9b_0
libffi=3.2.1=0
libgcc=5.2.0=0
libgcc-ng=8.2.0=hdf63c60_1
libgfortran=3.0.0=1
libgfortran-ng=7.3.0=hdf63c60_0
libopenblas=0.3.3=h5a2b251_3
libpng=1.6.17=0
libsodium=1.0.10=0
libstdcxx-ng=8.2.0=hdf63c60_1
libtiff=4.0.6=2
libxml2=2.9.2=0
linecache2=1.0.0=py27_0
```

```
llvmlite=0.15.0=py27_0
markupsafe=0.23=py27_2
matplotlib=1.4.3=pypi_0
mistune=0.7.2=py27_0
mkl=2018.0.3=1
mkl-rt=11.1=p0
mkl-service=1.1.2=py27_3
mock=1.0.1=pypi_0
moseq=0.0.1=dev_0 # freely available via MTA from the dattalab
moviepy=0.2.2.11=pypi_0
mpi4py=2.0.0=py27_1
mpich2=1.4.1p1=0
mpld3=0.3git=pypi_0
nbconvert=4.2.0=py27_0
nbformat=4.0.1=py27_0
ncurses=6.1=he6710b0_1
netifaces=0.10.4=pypi_0
networkx=1.9=pypi_0
nose=1.3.7=py27_1
notebook=4.2.1=py27_0
numba=0.30.1=np111py27_0
numbapro_cudalib=0.2=0
numexpr=2.6.8=py27hd89afb7_0
numpy=1.11.3=py27h3dfced4_4
numpy-base=1.15.0=py27h1793315_0
onedrivesdk=1.1.8=pypi_0
openblas=0.2.14=4
opencv=2.4.11=nppy27_0
openssl=1.1.1b=h7b6447c_1
pandas=0.23.4=py27h04863e7_0
pango=1.39.0=0
paramiko=1.16.0=pypi_0
path.py=8.2.1=py27_0
pathlib2=2.1.0=py27_0
patsy=0.4.1=py27_0
pexpect=4.0.1=py27_0
pickleshare=0.7.2=py27_0
pil=1.1.7=py27_2
pillow=2.8.1=pypi_0
pip=8.1.2=py27_0
pixman=0.32.6=0
prettytable=0.7.2=pypi_0
psutil=4.0.0=pypi_0
ptyprocess=0.5.1=py27_0
py2cairo=1.10.0=py27_2
pyasn1=0.1.9=pypi_0
pybasicbayes=0.2.1=pypi_0
pycairo=1.10.0=py27_0
pycosat=0.6.3=py27h470a237_1
pycparser=2.14=py27_1
pycrypto=2.6.1=pypi_0
pydot=1.0.28=py27_0
pygments=2.1.3=py27_0
pygraphviz=1.2=pypi_0
pyhsmm=0.1.6=pypi_0 # github.com/mattjj/pyhsmm
pyopenssl=16.2.0=py27_0
pyparsing=1.5.7=pypi_0
pyqt=4.11.4=py27_3
pyslds=0.0.1=pypi_0 # github.com/mattjj/pyslds
pysocks=1.6.8=py27_0
python=2.7.15=h9bab390_6
python-dateutil=2.5.3=py27_0
python-igraph=0.7.1.post6=pypi_0
python-louvain=0.9=pypi_0
pytimeparse=1.1.5=pypi_0
pytz=2016.4=py27_0
pywavelets=0.2.2=pypi_0
pyyaml=3.11=py27_4
pyzmq=15.2.0=py27_1
qt=4.8.7=2
qtconsole=4.2.1=py27_0
readline=7.0=h7b6447c_5
requests=2.21.0=py27_0
rsa=3.3=pypi_0
ruamel_yaml=0.11.14=py27_1
scikit-image=0.13.0=np111py27_0
```

```
scikit-learn=0.18.dev0=pypi_0
scipy=1.1.0=py27hd20e5f9_0
seaborn=0.8.1=pypi_0
setuptools=40.8.0=py27_0
shapely=1.5.13=py27_0
simplegeneric=0.8.1=py27_1
singledispatch=3.4.0.3=py27_0
sip=4.16.9=py27_0
six=1.6.1=pypi_0
smmap=0.9.0=pypi_0
snakeviz=1.0.0=py27_0
sqlite=3.26.0=h7b6447c_0
ssl_match_hostname=3.4.0.2=py27_1
statsmodels=0.9.0=py27h035aef0_0
syllables=0.1.0=pypi_0 # freely available from dattalab via MTA
system=5.8=2
terminado=0.6=py27_0
tk=8.6.8=hbc83047_0
toolz=0.8.0=pypi_0
tornado=4.3=py27_1
tqdm=1.0=pypi_0
traceback2=1.4.0=py27_0
traitlets=4.2.1=py27_0
typing=3.6.6=pypi_0
unittest2=1.1.0=py27_0
urllib3=1.24.1=py27_0
vega=1.4.0=py27_1
wheel=0.29.0=py27_0
xlsxwriter=0.7.2=pypi_0
xlwt=1.0.0=pypi_0
xz=5.2.2=0
yaml=0.1.6=0
zeromq=4.1.4=0
zlib=1.2.11=h7b6447c_3
```

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All datasets generated and/or analyzed during the current study will be available from the corresponding author on reasonable request. The raw per-frame data, MoSeq per-frame labels, and per-mouse behavioral summary data organized as NumPy arrays are stored in a Python pickle file, and available for download on an open-access basis via github.com/dattalab/moseq-drugs. Data derived from Micromedex is accessible at micromedex.com. Correspondence and requests for materials should be addressed to srdatta@hms.harvard.edu.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were chosen according to standards in the field of behavioral neuroscience, and follow the guidelines in Wiltschko et al., 2015.
Data exclusions	Data quality was assessed at several stages of the processing pipeline. First, each video recording was directly inspected to determine whether mouse tracking was successful. If there were persistent periods of the mouse's orientation being labeled as incorrectly flipped, these frames were added as new training data to the random forest flip classifier, described above, and the extraction procedure was run again. A heatmap of the mouse's body location over the course of the entire experiment was next examined to identify any sharp boundaries or disproportionately bright areas that might indicate tracking of non-mouse objects. If a non-mouse object was tracked (typically the edge of the arena), the ROI of the experiment was redefined, and the experiment was re-extracted. If, after applying all data quality correction

	methods listed above, the mouse's body was not tracked and extracted properly, or more than 5% of total frames were dropped or unavailable, the recording was not used in the dataset or any further analyses.
Replication	An average of n=10 biological replicates were performed for each drug & dose pair. Mice were only used once per drug treatment.
Randomization	Mice were randomly placed into both a drug and dose group, and which drugs and doses were delivered on a given experiment day were randomly chosen.
Blinding	Experimenters were not blinded, as they both prepared the drug samples and performed injections. However, modeling was carried out independent of any information about drug treatments per se.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	C57/BL6 male mice, aged 6-8 weeks old were used. Male wild-type or mutant littermates from breeding pairs of heterozygous CNTNAP2 mutants (JAX stock No. 017482) were used, aged 6-8 weeks.
Wild animals	No wild animals were used
Field-collected samples	No field-samples were used
Ethics oversight	All experiments were completed according to approved Harvard Medical School IRB guidelines.

Note that full information on the approval of the study protocol must also be provided in the manuscript.