B-SOiD: An Open Source Unsupervised Algorithm for Discovery of Spontaneous Behaviors

Alexander I. Hsu[1] and Eric A. Yttri[1,2*]

[1] *Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA USA*

[2] *Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA USA*

Capturing the performance of naturalistic behaviors remains a prohibitively difficult objective. Recent machine learning applications have enabled localization of limb position; however, position alone does not yield behavior. To provide a bridge from positions to actions and their kinematics, we developed Behavioral Segmentation of Open-field in DeepLabCut (B-SOiD). This unsupervised learning algorithm discovers natural patterns in position and extracts their inherent statistics. The cluster statistics are then used to train a machine learning algorithm to classify behaviors with greater speed and accuracy due to improved ability to generalize from subject to subject. Through the application of a novel frame-shift paradigm, B-SOiD provides the fast temporal resolution required for comparison with neural activity. B-SOiD also provides high-resolution behavioral measures such as stride and grooming kinematics, that are difficult but critical to obtain, particularly in the study of pain, compulsion, and other neurological disorders. This open-source platform surpasses the current state of the field in its improved analytical accessibility, objectivity, and ease of use.

***Keywords*—** naturalistic behavior, neuroethology, kinematics, open-source, action selection, po-

sition estimation, open-field, behavioral sequences

Main

The brain has evolved to support the generation of individual limb movements strung together to create action sequences. The selection, performance, and modification of these actions is key to an animal's continued survival[1]. The neural underpinnings of this behavioral repertoire is one of the foundations of neuroscience [2]; however, research largely focuses on stereotyped, reductionist, and over-trained behaviors due to their ease of study. Beyond the potential confounds associated with an artificial or over-trained task, this line of interrogation discards most of the behavioral repertoire and its intricate transition dynamics [3–5]. Comprehensive behavioral tracking permits behavioral quantification of action, but even with the few, expensive commercial options available, it has been an extremely time and labor intensive process, while yielding limited temporal resolution and behavior detail.

Recent advances in computer vision and machine learning has accelerated automatic tracking of geometric estimates of body parts[6–9]. Although establishing the location of a body part can be informative given the appropriate experimental configuration, the behavioral interpretability is quite low. For instance, the extracted position of where a paw is in an open-field may be able to be used to determine stride length, but it does not capture what the animal is doing. Moreover, the minimum thresholds for ambulatory bout or

turns are all subjective and may vary with animal size, video capture technique, or behavioral raters. The uncertainty inherent in these seemingly straightforward user-definitions is only compounded when top-down edicts are applied to delineate more complex behaviors. Thus, classification presents a severe impediment to the impactful lessons to be learned from dissecting the neural correlates of spontaneous behavior. These challenges motivated our investigation into how partnering dimensionality reduction algorithms with unsupervised pattern recognition could create a viable tool in automatic extraction of an animal's behavioral repertoire.

In a recent behavioral study, Markowitz and colleagues were able to automatically identify subgroups of animal's behaviors using depth sensors. Their algorithm, MoSeq [10], utilized the principal components of "spinograms" to identify action groups. The authors uncovered the striatal neural dynamics of action, consistent with the notion that population of medium spiny neurons (MSNs) encodes certain behavior[3], and the organization of[11–15]. Taking a different approach, Klaus et al. employed a unique dimensionality reduction method, t-Distributed Stochastic Neighbor Embedding (t-SNE), to cluster behaviors based on time-varying signals such as angle, speed and body acceleration. There, they discovered that different behavioral clusters were accompanied by distinct changes in striatal neural dynamics[15]. The powerful rationale behind selecting t-SNE over other dimensionality reduction methods is to compress variable high dimensional features onto low-dimensional space while preserving the contrast, or "local structure", within the original feature space [16]. Together, these studies suggest that data-driven algorithms can

define actions more accurately and objectively to study the neural substrates of behavior.

Building on this progression, we were inspired to investigate the organization of high-dimensional spatiotemporal features - such as body length, distance and angle between body parts, speed, and occlusion of a body part from view. In conjunction with recent advances in pose estimation, including the open-source platform DeepLabCut [7], we provide an open-source tool that integrates dimensionality reduction, pattern recognition, and machine learning to enable autonomous multi-dimensional behavioral classification. B-SOiD (Behavioral-Segmentation of Open-field in DeepLabCut) automatically extracts behavior categories with excellent reliability and accuracy (Supp Video 1). Due to the unsupervised nature of the algorithm, behavioral categories are discovered based upon their natural statistics. In this way, we greatly reduce tool and user bias. It is easily applied across subjects, arenas, and cameras. Moreover, B-SOiD can provide two essential features that are currently unattainable: temporal resolution on the order of milliseconds (required for pairing with electrophysiology data) and individual limb kinematics during those behaviors. For the study of behavior, establishing how an action is performed can be more important than whether it is performed. Finally, in benchmarking the tool, we characterize an action sequence behavioral structure - an organization of "what to do next", and how this changes over time and following a cell-type specific lesion.

Results

To reveal the natural patterns of body part positions that accompany different behaviors, we first established body part locations (snout, paws, and proximal tail). Any marking system can be used. We then computed spatiotemporal features of these points (speed, angle, distance between tracked points - Fig. 1). After clustering the features, we employed pattern recognition tools to extract the behavior classes based upon their natural statistics. Although this clustering is sufficient to achieve the desired behavioral identification (Fig. 1b,d), doing so creates a computationally expensive process. Moreover it yields datasets that are impossible to compare with each other owing to the stochastic cluster seeding. Therefore, we created a more robust tool by using the clusters to train a machine learning algorithm that will predict animal behaviors based on pose. The advantages of building a machine model to predict new data are the consistency across datasets, speed, and the agnostic nature of assignment. Our goal is to provide an openly available, easily generalizable "plug-and-play" tool that can achieve behavioral categorization at millisecond timescales - all with a single USB camera. We provide this to the research community (https://github.com/YttriLab/B-SOiD). B-SOiD resolves many known challenges in the field, namely the expense/applicability, lack of kinematics, and most importantly, the temporal resolution that is quintessential for neurobehavioral studies. While we describe here the utility of a machine learning classifier based on unsupervised grouping, the package also provides several options for users only interested in pose patterns defined *a priori*.

B-SOiD extracts behavioral clusters in high-dimensional space. An animal's behavior can be parsed into a sequence of changes in physical features. Based on the three independent categories clustered out by Hierarchical Clustering Analyses (HCA), we performed multi-dimensional embedding from seven down to three using t-distributed Stochastic Neighbor Embedding (t-SNE) [16]. t-SNE embeds high-dimensional features based on conditional probability, i.e. how probable is a given data point positioned in a dimensionally reduced space. This technique simplifies the output without simplifying the complexity of naturalistic behaviors. In addition, with the maintenance of original feature differences ($P(A|A,B)$ vs $P(B|A,B)$ mimics A vs B), it enables the artificial embedding of a signal occlusion marker to skew the data more compared to pose estimate jitters. This critical step transforms the lack of information itself into information that can be used. The pose information segregates into distinct nodes in the 3-dimensional t-SNE space. (Fig. 1d). When the clusters are mapped onto behavioral categories (e.g. given descriptive names), we find a motion axis, direction axis, and length axis, similar to the output via hierarchical clustering analysis (see Supp Video 1). Due to the stochastic nature of embedding data point as a function of a randomly selected joint distribution, it is not mandated that any t-SNE parameter set (perplexity, exaggeration, learning rate, number of dimensions, etc.) will output distinguishable axes that agree with the greedy algorithm HCA. The ability to capture these axes enables unsupervised grouping algorithms to be applied here.

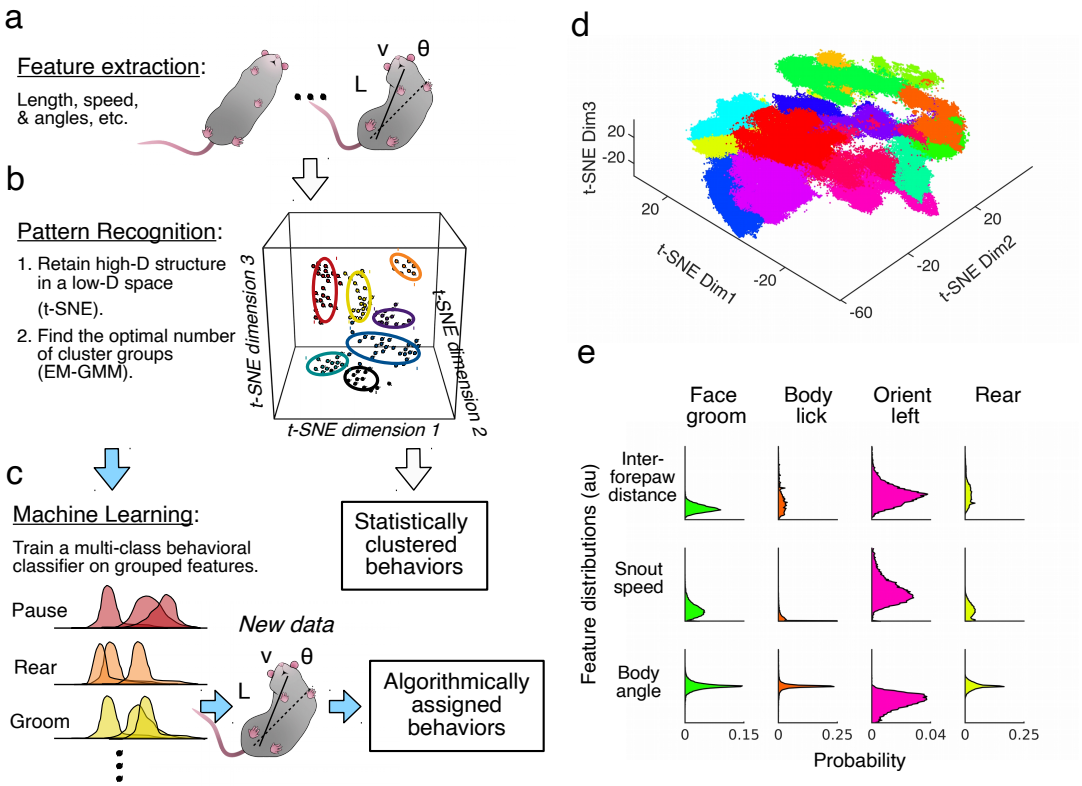

Figure 1: Flow-chart and implementation of B-SOiD algorithm. (a) After obtaining the six body positions that used to provide the three feature classes that define movement (speed, angle and inter-body part distance), we (b) cluster high-dimensional input features in a low-dimensional space utilizing t-distributed stochastic neighbor embedding (t-SNE) and identify clusters using Gaussian mixture models (GMM). The high-dimensional statistics of these clusters are fed as inputs to (c) train a support vector machine (SVM) classifier to recognize behavior based on pose. (d)Even when provided a maximum of 50 clusters, EM-GMM routinely converges to 16 assignments given these inputs.(e) All 16 assignments generated distributions that were easily distinguishable based upon their original features (See S1 for all distributions).

Due to the stochasticity of t-SNE, we cannot predetermine where behavioral cluster centers will be. Therefore, to assign data points to clusters, we employed iterative expectation maximization (EM). Specifically, we chose to fit the parameters of Gaussian mixture models (GMM) [17] because t-SNE utilizes Gaussian distribution for the probability of observing a pair of poses. B-SOiD iteratively initializes the model mean, covariance, and priors with random values to provide data-driven group determination. For our dataset, B-SOiD identified 16 GMM classes in the low-dimensional space (Fig. 1d). Based upon the conserved behavioral motifs, we assigned names to the clusters. For organizational purposes, we grouped these behaviors according to movement type (red=pause, pink=poke, black=rear, blue=groom, green=move; to be used throughout this manuscript). Although automated behavioral class definitions can suffer from either over- or under-splitting of behaviors, we found that this was not the case. When given the opportunity to initialize 50 classes, B-SOiD converged to 16. To verify that we are not errantly merging behaviors, we randomly isolated short videos based on behavioral class assignments (See https://github.com/YttriLab/B-SOID/tree/master/segmented_behaviors for details) and found that behavioral assignments were internally consistent. In addition, meta-analyses on physical features showed distinct multi-feature distributions (Fig. 1e). This includes the incorporation of signal occlusion, i.e. body lick and rear will have instances where the snout is marked with 0, or null signal. For some behaviors, B-SOiD discovered two distinct classes of the same behavior. Upon examining these classes, it appeared to be based upon the vigor of their performance, therefore the definitions + and - were given

(see S1 for feature distribution ranges).

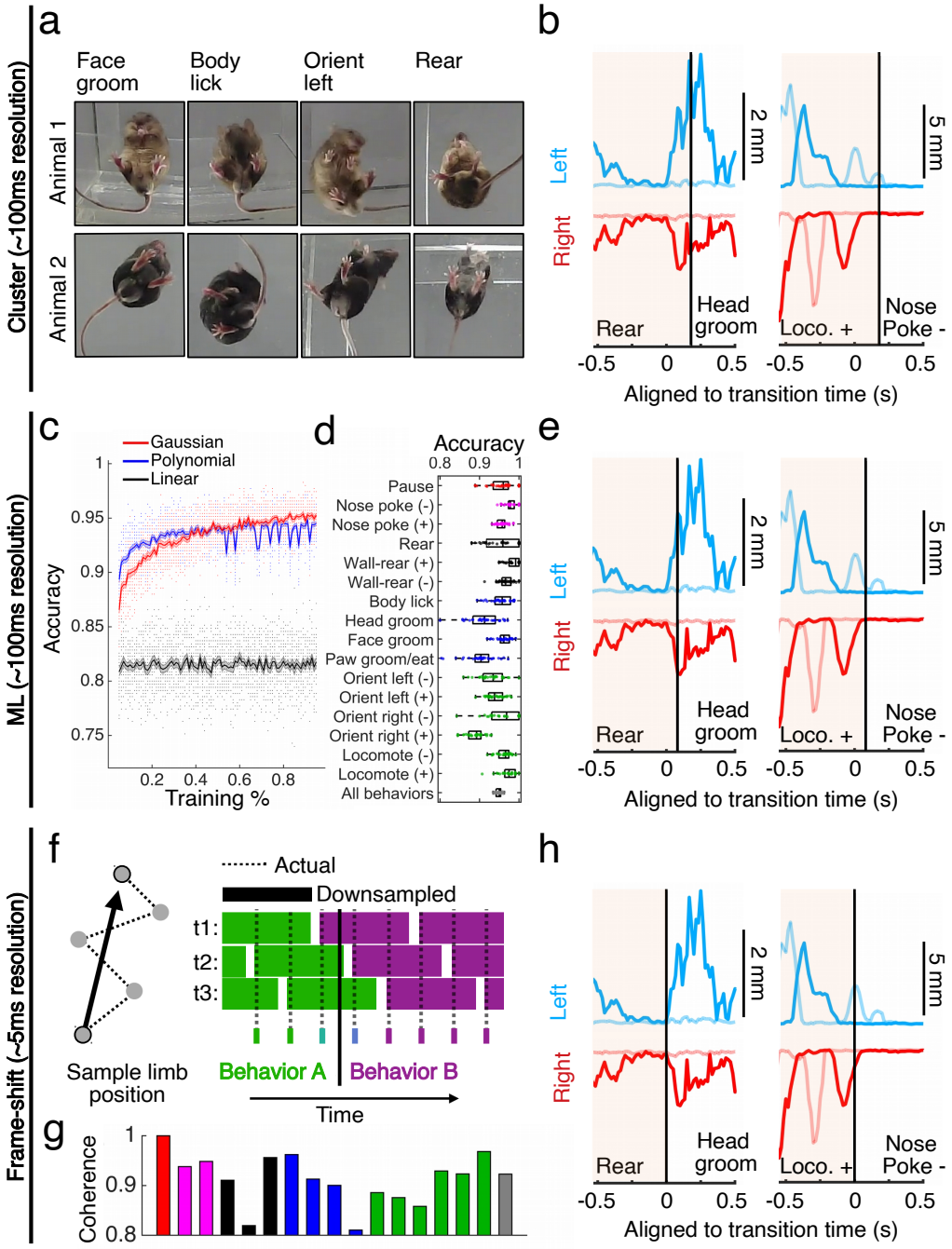

Figure 2: Performance quantification across multiple temporal resolutions with novel machine learning algorithms. (a) Snapshots of 300ms into example behaviors.

Figure 2: Note the differences in mouse color and size. (b) Trajectory plots of right (red) and left (blue) for the forepaw (darker) and hind paw (lighter) showcasing an example transition from rear -> head groom (left), and locomote + -> nose poke (left). Vertical lines denote identified transition time. (c) Cross-validated accuracy on a portion of held-out data (350 data points/test, 30 non-repeated randomly selected iterations), using 5 - 95% of the dataset as training. Linear (black), polynomial (blue), and Gaussian (red) kernel functions were tested using identical training and test data. (d) With 80% of the data used to train our support vector machine, accuracy for individual and all behaviors were decoupled. (e) Same trajectory plots of as in (b), but with transitions defined by with machine learning algorithm. (f) Cartoon example of the potential for noise jitter to override the movement signal at high sampling rates (left). To overcome this, we executed a frame-shift computation to derive high resolution timing from downsampled high signal data (right). (g) Percent coherence between downsampled data and high resolution, frame-shifted results for individual behaviors. Color code as in (d). (h) Same trajectory plots of as in (b), now incorporating novel frame-shift algorithms for detecting millisecond resolution transitions.

We can see that the clustered groups (here, we provided descriptive names for each) does not overfit an animal versus another (Fig. 2a). The behavioral segmentation gained from B-SOiD also was broad enough to incorporate similar sequences of multi-joint movements, i.e. grooming different places on the torso were both considered to be the same behavioral group (Fig. 2a). To our surprise, B-SOiD did not group all grooming behaviors together; rather it segmented out the canonical grooming types, described as the syntactic chain of self-grooming in rodents, paw groom, face groom, head groom, and

body lick [18–20]. The generalizability for different animals as well as the discriminability for different behaviors hinted at the potential for a support vector machine classifier in identifying behaviors based on pose.

Machine learning classifier improves B-SOiD's generalizability across animals and temporal resolution. To improve consistency, speed, and applicability in classifying behaviors, we further equipped B-SOiD with a multi-class SVM (other classifiers could also work, but SVM has proven sufficient). Traditional Generalized Linear Models are insufficient to compute this multi-class classification. Recent computational advances have enabled SVMs to significantly improve decoding accuracy using Error Correcting Output Codes (ECOC) [21]. Based on our utilization of normal distribution with t-SNE and GMM, we hypothesized that transforming the feature space with Gaussian kernels would best separate behavioral groups for classifier training, thereby supplying the most robust and accurate decoding results. To test our hypothesis, we trained our classifier with three types of kernels: Linear, Gaussian, or Polynomial. Our results showed that, with sufficient training data ($\geq 70\%$), the model predicts most reliably with Gaussian kernel functions when training a multi-class support vector machine classifier (Fig. 2c-d). The accuracy quantification is referenced against cluster assignments, not necessarily better or worse (our data in the following section would argue it is the former). The Gaussian utilization of physical features is a good starting point, and has proven effective when working with time-varying signals, i.e. Gaussian Process Factor Analysis (GPFA) [22], small dataset automatic speech recognition [23], etc. Accurate labeling of behaviors requires precision in

both classification and timing. We present two examples of action sequence transitions, displaying the constituent paw positions (forepaws are shaded darker, Fig. 2b). In these instances, paw trajectory appears to precede the start of the behavioral assignments. Upon sufficient training (80%) of all cluster data using spatiotemporal poses, we were able to improve cluster space mis-assignments at the intersection of two behaviors (Fig. 2e). The results suggest that the previously described prediction error were, in fact, an improvement from mere clustering.

Frame-shift paradigm enabled behavioral discrimination at temporal resolution sufficient for electrophysiology. To provide reliable alignment to behavioral dynamics with electrophysiological measures, it is ideal to approach the millisecond resolution of electrophysiology. This is unavailable given current technology. However, a particular challenge in defining behaviors at a high sampling rate is that pose-estimation jitter will dominate the signal (Fig. 2f (left)). Here, we propose a novel "frame-shift" manipulation, inspired by current state-of-the-art automatic speech recognition [24], to resolve behavioral transitions at the millisecond-scale (Fig. 2f (right)). Briefly, B-SOiD SVM predicts on a high frame-rate video downsampled to 10 frames per second (fps) so as to maintain a high signal to noise ratio in the spatiotemporal dynamics of the markers. This is then repeated many times, but offset by one frame each time (t1, t2, t3 in Fig. 2f (right)). For example, in this and all subsequent analyses in the manuscript we use 200 fps video downsampled to 10fps 20 times, each time offset by 5ms (one frame). We found that, after accounting for the change in temporal resolution (see Methods), the frame-shifted prediction provided

$\sim 90\%$ coherence with the single 10fps data to which the ML algorithm had also been applied (Fig. 2g). The frame-shift feature of B-SOiD provides additional improvement over non-frameshift applications due to weighted signal over noise. Given that the behavioral content remains largely identical, this frame-shift paradigm then allows B-SOiD to predict behaviors at a temporal resolution matching the sampling rate of the video camera, identifying behavioral initiation at the scale of a few milliseconds (Fig. 2e,h). This is a critical advancement in analyzing neural correlates of spontaneous behaviors [3]. In addition, this enables a much more comprehensive analysis of action kinematics over only clustered or non-frame-shift applications (Fig. 2b,e,h).

B-SOiD uncovers history-dependent effects in millisecond-resolution kinematics. The use of body position markers for high-resolution behavioral classification provides an additional advantage: movement kinematics [25]. For any identified action, we can extract the speed and trajectory of each limb, and do so at the millisecond-scale. To study these kinematics, we applied frame-shift B-SOiD to data from six mice exposed to a novel open-field, splitting the hour-long sessions into early versus late exploration blocks. In our dataset, animals traversed less distance as a function of time in open-field (Fig. 3 a). We first determined whether average bout duration changed. Interestingly, for both locomotion + and locomotion - (see S1 for distinction), the distribution of bout durations did not change over time (Fig. 3 b), nor did total time spent locomoting ($p > 0.3$ for both + and -). Thus, we predicted that the reduced kinematic vigor of each stride must be the cause of of the decreased distance covered. Indeed, the distribution of stride length

and peak speed per stride over both types of locomotor bouts were decreased (Fig. 3 c-d, see S2 for kinematics analysis technique). These findings are similar to what has been documented previously but using a very different methods (a split-belt treadmill, [25]. The unique ability of this tool to both discover conserved movement types and effortlessly quantify their kinematic properties stands to be impactful in the neurobehavioral study of disease. For instance, the ability to determine if a ischemic lesion-induced reduction in speed is the result of shorter or slower stride (as well as the interactions across the broader behavioral repertoire) is a powerful but rare experimental insight.

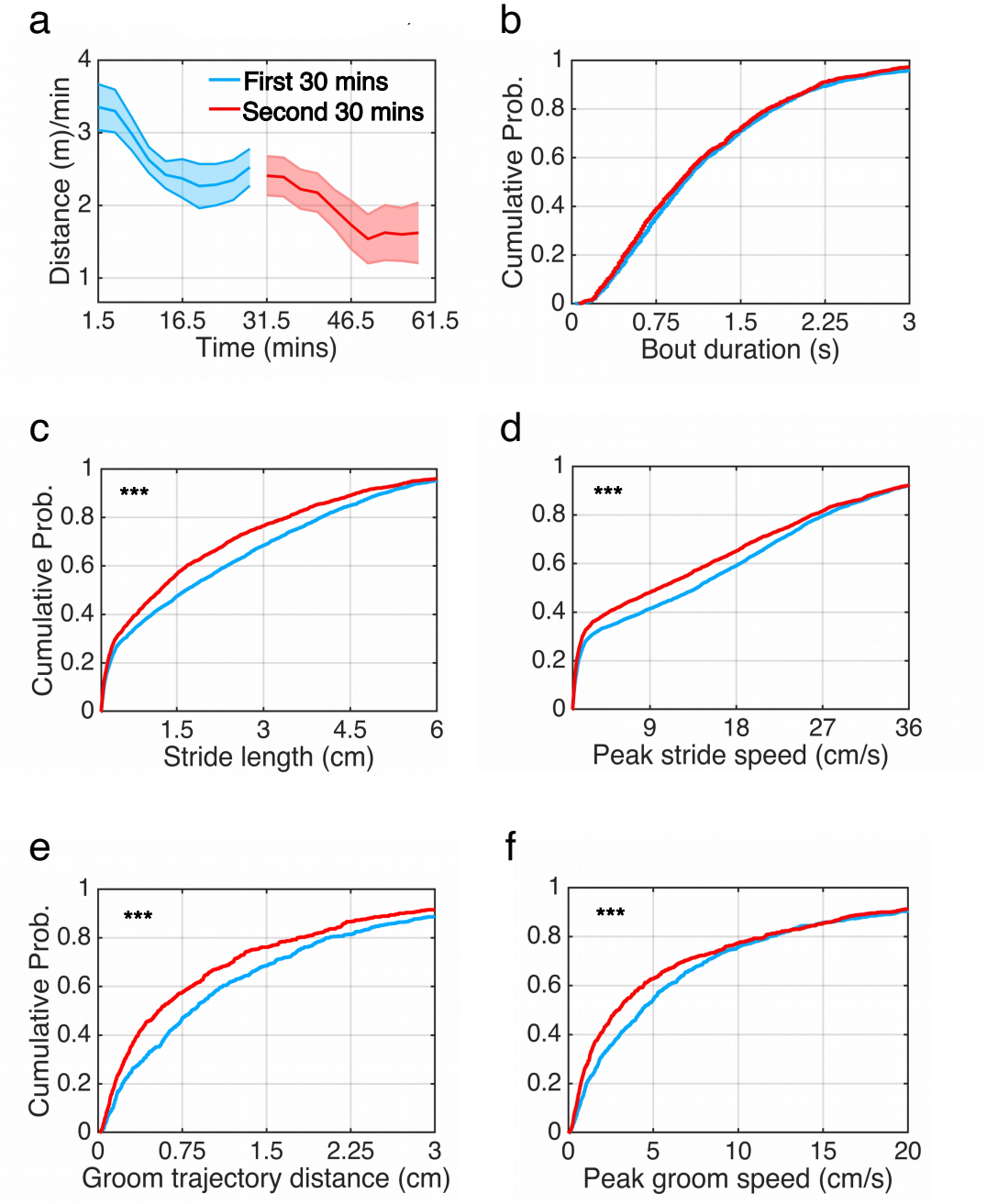

Figure 3: B-SOiD enables millisecond-resolution kinematic analyses. (a) Distance covered over the session, shaded = SEM, split into first (blue) versus the second (red) 30 minutes.

Figure 3: Cumulative histograms of (b) time locomoting, (c) stride length, and (d) peak stride speed in the first (blue) versus the next (red) 30 minutes. Locomotor + and - groups were combined in order to study the full range of stride kinematics. Cumulative of (e) groom stroke trajectory distance and (f) peak groom speed for head groom. ***$p < 0.001$.

As another example of this application, rodent research concerning obsessive compulsive disorder, anxiety, and pain use grooming behaviors to assess the disease model state [26]. Attempts to quantify these behaviors, particularly beyond merely the amount of time spent grooming, has proven exceedingly difficult and arduous. In wild-type animals, we again used time as an experimental manipulation. We found that head groom trajectory length and peak speed diminished with time (Fig. 3 e-f). This may be due to many causes, and there may be an interaction with the preponderance of time spent grooming, as we will look into later.

Cell-type specific perturbation biased animals towards larger and faster forepaw movements in face grooms, decoded by B-SOiD. We more thoroughly tested of B-SOiD's utility to quantify grooming-type behaviors, using video from mice with and without cell-type specific lesions of the indirect pathway of the basal ganglia (A2A-cre, with or without cre-dependent caspase virus injected into striatum, Fig. S3). The basal ganglia is thought to be involved in action selection and sequencing [3,18], the dysfunction of which may give rise to diseases like OCD and Huntington's, in which unwanted actions occur, or occur too quickly [27]. Additionally, activation of the indirect pathway has been suggested

to contribute to hypokinesis, or smaller and slower actions [28]. We first compared the two canonical groom types, head and face grooms across animals (Fig. 4 a). We found no difference in the temporal patterning of these two groom types ($p = 0.90$, chi-square test of ratio head before face, face before head, N= four animals control, one session each; Fig. 4 b-c for first half of two example sessions). However, consistent with a hypokinetic role, we found that animals lacking striatal indirect pathway neurons demonstrated a significant rightward shift in the speed and distance of face grooms, particularly pronounced for the smaller movements in the distribution (Fig. 4 d,f (top)). In addition, these effects was not observed in the similar, but generally larger head grooming behavior (See S1 for G6 versus G8) - head grooming (Fig. 3 d, f (bottom)). Importantly, there was no effect in either face groom or head groom on bout duration (Fig. 4 e). Currently, most all studies of grooming behavior focus on the duration of bouts, and due to difficulties in quantification, head and face grooms are typically combined. These findings, made possible because of B-SOiD's ability to both dissociate groom types and measure kinematics, further support the notion that indirect pathway contribute to the performance of less vigorous behavior. In demonstrating the functionality of B-SOiD, we have taken a vital step in extending hypotheses gained from a reductionist approach to a naturalistic setting.

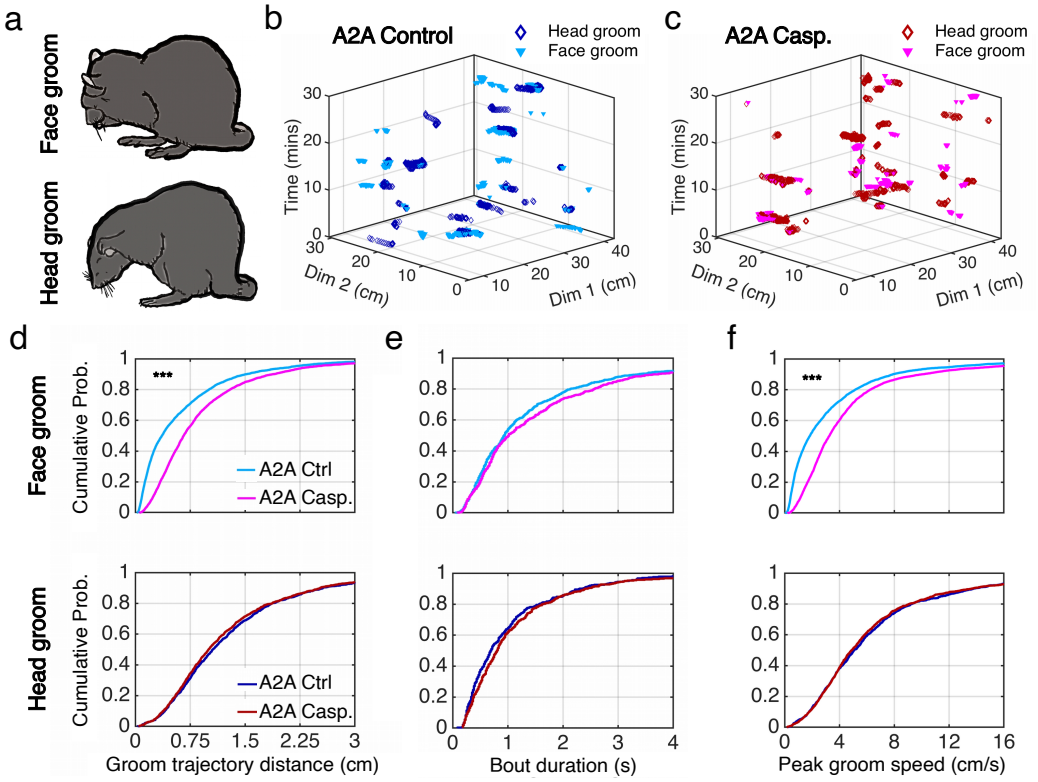

Figure 4: Detection of robust, hard to detect kinematic changes in grooming behavior following cell-type specific lesion. (a) Schematic of the two canonical groom types, (top) face and (bottom) head (image adapted from summary by Aldridge et al., 1990). (b)-(c) Time and location of B-SOiD defined sequential grooming from example A2A-Cre transgenic animals (c) with and (b) without caspase virus induced cell-type specific lesion. (d-f) Cumulative histograms of groom stroke trajectory distance (d), groom bout duration (e), and peak groom speed (f) of face and head grooming (control -blue; lesioned - red, N=4 animals, one session each). ***$p < 0.001$.

Observation of behavioral structure changes with decreased novelty. In addition to millisecond-scale kinematics, we can interrogate the greater behavioral structure that

exists in animals exploring an open-field. The ability for an animal to explore a novel environment is critical for survival. We documented the occurrence of our naturally discovered behaviors in the context of the exploratory strategies mice employ after being introduced to a new cage. We analyzed the transitional probabilities between each extracted behavior and the next. We found that for several behavioral types, the average transition probability to a similar type of behavior is greater. Similarly, locomotor type behaviors (green) show a stronger, more predictable transtion to poking behaviors (pink). When split into the first and second 30 minute periods exploring an open-field for the first time, we found that animals tend to be less predictable in their transitions as a function of time. More specifically, the polarized regions of transitions in the first 30 minutes (Fig. 5 a (top)) became more even across behaviors in the second 30 minutes (Fig. 5 b (top)). To better visualize these transitions, we plotted a directed graph of all transition probabilities exceeding 0.5 percent. We observe the same transition structure (e.g. similar behavior types are clustered together) in both the first and second thirty minute data. However, the strength of the transition structure (shown by the thickness of the transition arrows and tightness of points), is greatly decreased. Using a fix-point attractor framework in dynamical systems, this suggests a basin of attraction early on that reflects the animals' drive to explore the novel open-field. This drive appears to diminish with time. Thus, the larger action sequence framework is conserved, but the decreased novelty diminishes the attraction forces within nodes of transitions. Our results highlighted the fact that, our algorithm spans meso- to macro-level analyses for a richer account of naturalistic behaviors.

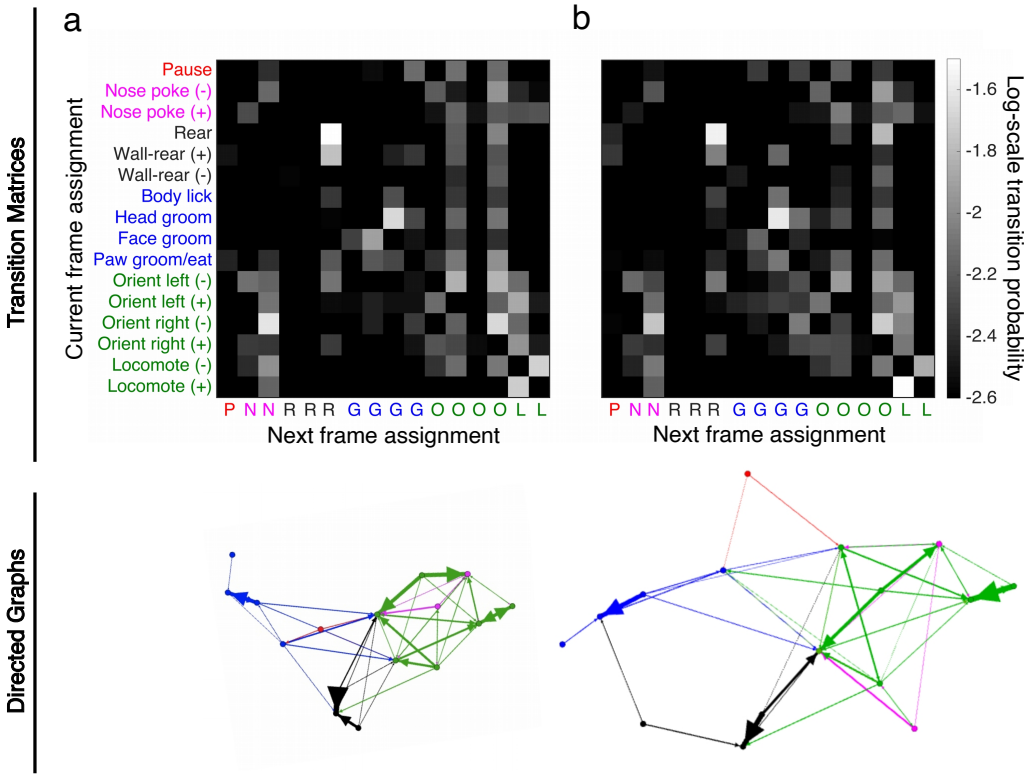

Figure 5: Frame-to-frame transition analyses demonstrates patterned transitions that dissipates with time in open-field. Transition matrix of the sixteen classified behaviors going from current action (Y axis) to next (X axis) in (a) the first and (b) next 30 minutes of an open-field session. (Bottom) Directed graph using data from a,b, but limited to transition probabilities $\geq 0.5\%$ for clarity. Thicker line indicates greater transition probability. Network attraction strength and scale are identical in both plots. N= 6 animals, one session each.

A change in latent recurrency upon exposure to novel open-field. In addition to analyzing "what to do next" in an mouse exploring an open-field, we could also ask "how long does it take to repeat" a behavior. Limited studies have attempted to investigate the

time between repetitions of a certain behavior. Recent neural population studies have suggested that the between trial neural dynamics may hold latent information about an animal's internal drive to perform an action [22]. To understand the animal's internal state from a purely behavioral perspective, we can analyze the recurrency, or "how long before an action is repeated". Over the hour session, we discovered that the interval between performing the same behavior converged to the mean of all behavioral recurrencies (Fig. 6 a). This was also evident if we consider the standard deviation from the mean as a function of time (Fig. 6 b). To provide a better visualization, we summarize the convergence of recurrence in an understandable plot, where the length of the arrow denotes the mean recurrence time, and the thickness of the arrow indicates the recurrence SEM across all bouts and animals (Fig. 6 c). Quantitatively, we found that there was a significant positive correlation between distance from the mean and time-dependent change in recurrence. Our findings suggest that the animals' internal state changes with the amount of exposure in the novel open-field. This also points at the potential utility mapping neural activity to these latent variables.

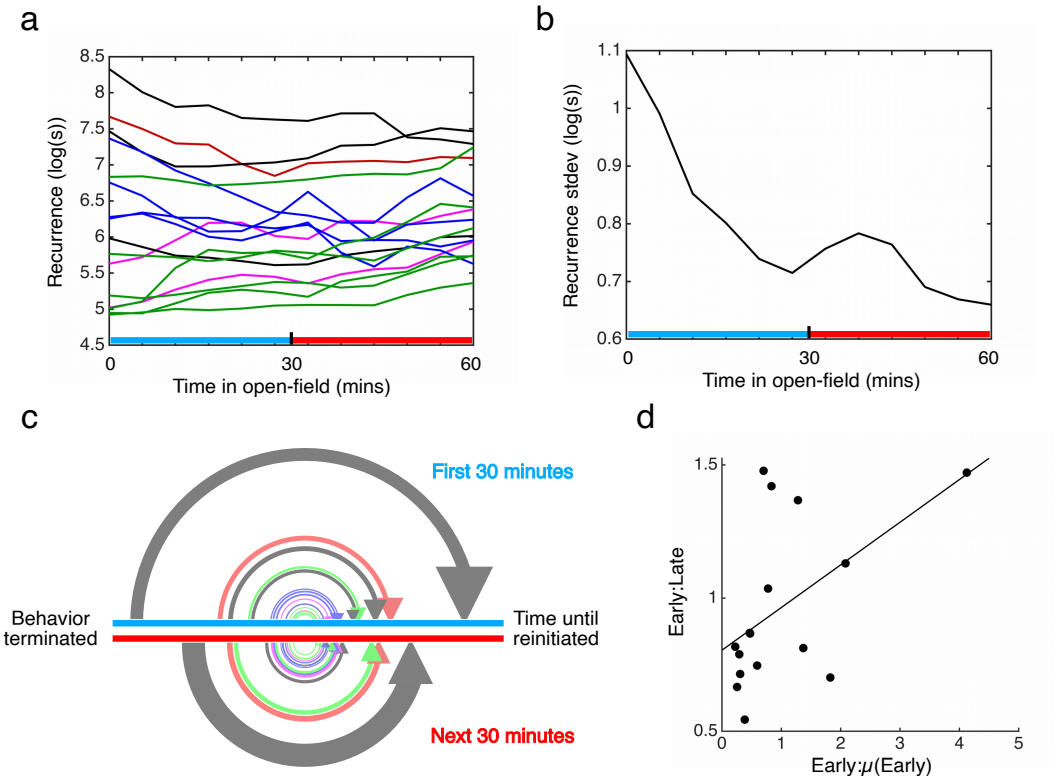

Figure 6: B-SOiD reveals change in behavioral recurrence rate over time. (a) Mean recurrence time (time until an action is repeated) each behavioral (color codes as in 2d, blue and red bars indicate first and second 30 minute blocks). (b) Standard deviation of recurrence time reduces throughout session. (c) Recurrency plot, based on data in (a), summarizes over first versus the next 30 minutes blocks. Length of arrow correlates to mean recurrence time. (d) Ratio of early to early mean recurrence positively correlated ratio of early to late recurrence. ($r = 0.52$, $p = 0.04$). N= 6 animals, one session each; same sessions as Figure 5.

Discussion

Naturalistic, open-field behavior provides a rich account of an animal's motor decisions and repertoire. Until recently, capturing these behaviors with precision and accuracy was prohibitive. Still, new methods require extensive investments in specialized knowledge and tools (spinogram scanner, image capture systems), while still lacking the temporal resolution to meaningfully couple behaviors and their kinematics with neural recordings. Our unsupervised algorithm, B-SOiD, captures the natural statistics of limb and action dynamics with off-the-shelf technology and a simple user interface. This tool serves as the vital bridge between recent breakthroughs in establishing the position of body part[7,8] and the behaviors those parts perform. It also demonstrates the utility and potential of artificial intelligence in behavioral assessment, specifically the integration of multi-dimensional embedding, iterative expectation maximization, and a multi-learner design coding matrix in classifying behavior. It is the natural statistics of the inherent behaviors that are discovered, extracted, and used to inform the ML classifier, thus tailoring the tool to the animal and eliminating top-down bias. With the additional insight that the presence of missing information is information itself, our tool even permits the extraction of three-dimensional movements from two-dimensional data.

To extract 3-D behaviors from a 2-D video, we need to dissociate visual obstruction from loss of signal. One of the key advantages of incorporating t-SNE in B-SOiD is its ability to seed pose-estimation jitters away from the "0" signal occlusion marker. For rear-

ing and certain grooming behaviors (i.e. body lick), providing a distinct visual obstruction marker pushes out the clusters in the dimensionally reduced space. Moreover, due to the probabilistic positioning of data in the low-dimensional space, t-SNE innately generates a wide distribution of seeding encompassing the dynamic range of any given behavior. Determining the assignment in arbitrary 3-D t-SNE space can be taxing manually. Fortunately, EM-GMM allows for unsupervised grouping based on low root-mean-squared error between EM iterations. Through iteratively determining the lowest root-mean-squared error that the EM algorithm converges, the algorithm escapes getting stuck in a suboptimal local optimum.

Recently, Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) has proven to be more computationally efficient and faster for clustering than t-SNE [29]. In addition to local structure preservation like t-SNE, UMAP preserves the "global structure", or relative cluster placement of the dimensionally reduced space: is 1 closer to 2 or 3? Certainly, these are technological advances that can be useful for future development and explainability. By developing a machine learning algorithm that learns on the high dimensional features, relative cluster placements in the low dimensional space becomes unnecessary. Also, a powerful advantage of B-SOiD is the consistent behavioral classification across experimental datasets - a robustness that actually would be lost with UMAP.

The ability to decompose behaviors into their constituent movements is a key fea-

ture of B-SOiD. By using limb position, we extract not only the action performed, but also its kinematics (stride speed, paw trajectory, etc). While recent work has benefited from access to such performance parameters [25,30], it stands to be an even more potent advantage in the study of disease models. Obsessive compulsive disorder research in particular has long sought improved identification and quantification of grooming behavior [19,31,32]. B-SOiD enables not only the automated detection of grooming - a key metric in the study of several neurological disorder models - but also its speed and relation to other behaviors. Therefore, understanding both the structure and substructure of actions increases the potential of research.

Lastly, it has been historically difficult to decode an animal's motivational drive purely from a behavioral perspective. It is unclear how actions, their selection, and their performance should change. Thus, in addition to the effects of a cell-type specific lesion, we examined B-SOiD's utility in analyzing the greater behavioral structure that may be reflective of the animal's internal states. This structure, which is considerably more difficult to access than more basic behavioral measures, governs action sequence transitions ("what to do next") and recurrency (which may indicate a behavioral urgency signal [33]). We understand that these types of analysis can be gained from other methods [10]. One key feature of B-SOiD, however, is the greater than six times improvement in temporal resolution. This improvement provides a decided advantage for alignment of behavior with electrophysiology, as well as a deeper comprehension of the composite kinematics forming those actions. The flexibility in our algorithm allows users to access the richness

of naturalistic behaviors with a single, off-the-shelf video camera. The open source code can be run in full with one command line input, but is easily adapted to increasing degrees of user-guided classification. In sum, B-SOiD delivers the quality of data and ease of use sought after by several neuro-behavioral communities - and does so by eliminating potential financial or intellectual barriers to research.

Methods

All analysis code, as well as the data used to create these figures, is open sourced and freely available from gitHub (https://github.com/YttriLab/B-SOiD).

Data processing feature extraction We applied an adaptive high-pass filter to attenuate all sub-threshold (p $<$ rising phase, from low to high, in smoothed bimodal distribution histogram for signals that are likely to be occluded, and p $<$ 0.2 for robust body parts such as tail base and hind limbs) estimated positions by replacing with the previous likeliest position, marking occlusion with change of 0. Because the pose estimation in a program like DeepLabCut and LEAP never outputs identical positions in consecutive frames, this particular workaround allows the 3-D behavior to be represented in a 2-D space. We isolated 7 spatiotemporal features based on the three independent categories determined by HCA, speed, length and angle. The body length, or distance from snout to base of tail, or d_{ST}, is formulated as,

$$||d_{ST}|| = \sqrt{\sum_{D=1}^{2}(S_D - T_D)^2} \tag{1}$$

where S_D and T_D represent the likeliest position of snout and base of tail, respectively, and D denoting x or y dimension. The front paws to base of tail distance relative to body length, or d_{SF}, is computed with the equation,

$$\begin{aligned} d_{SF} &= ||d_{ST}|| - ||d_{FT}|| \\ &= \sqrt{\sum_{D=1}^{2}(S_D - T_D)^2} - \sqrt{\sum_{D=1}^{2}(F_D - T_D)^2} \end{aligned} \tag{2}$$

where d_{FT} is the distance between front paws and base of tail, F_D is the mean x and y position of the two front paws. The back paws to base of tail distance relative to body length, or d_{SB}, is calculated using the formula,

$$\begin{aligned} d_{SB} &= ||d_{ST}|| - ||d_{BT}|| \\ &= \sqrt{\sum_{D=1}^{2}(S_D - T_D)^2} - \sqrt{\sum_{D=1}^{2}(B_D - T_D)^2} \end{aligned} \tag{3}$$

where d_{BT} is the distance between back paws and base of tail, B_D is the mean x and y position of the two back paws. The distance between two front paws, d_{FP}, is derived from,

$$||d_{FP}|| = \sqrt{\sum_{D=1}^{2}(FR_D - FL_D)^2} \tag{4}$$

where FR_D and FL_D are the likeliest positions of right and left front paws, respectively. The snout speed, v_S, or displacement over period of 16 ms, uses the following equation,

$$||v_S|| = \sqrt{\sum_{D=1}^{2}(S_D^{t+1} - S_D^t)^2} \tag{5}$$

where S_D^{t+1} and S_D^t refer to the current and past likeliest snout positions, respectively. The base of tail speed, v_T, or displacement over period of 16 ms, similar to the formula above,

as follows,

$$||v_T|| = \sqrt{\sum_{D=1}^{2}(T_D^{t+1} - T_D^t)^2} \tag{6}$$

where T_D^{t+1} and T_D^t refer to the current and past likeliest base of tail positions, respectively. The snout to base of tail change in angle, $\Delta\theta_A^{A'}$, is formulated as follows,

$$\begin{aligned}\Delta\theta_A^{A'} =& sign(A'_xA_y - A_xA'_y)(\frac{180}{\pi})\arctan(\frac{A'_xA_y - A_xA'_y}{A'_xA_x + A'_yA_y})(\frac{180}{\pi}) \\ & sign(A'_xA_y - A_xA'_y)(1 - sign(A_xA'_x + A_yA'_y))\end{aligned} \tag{7}$$

where A and A' represent body length vector at past (t) and current (t+1) time steps, respectively, $sign$ equals 1 for positive, -1 for negative, 0 for 0, and $x \bullet ||x||$ for complex numbers. Note that the Cartesian product and dot product are necessary for four-quadrant inverse tangent and that the sign is flipped to determine left versus right in terms of animal's perspective.

In addition, these features are then smoothed over, or averaged across, a sliding window of size equivalent to 60 ms (30 ms prior to and after the frame of interest). This is important for distinguishing the pose estimate jitter for finer movements that the animal makes, such as the different groom types.

Data clustering With sampling frequency at 60 Hz (1 frame every 16.7 ms) the data capture only fragments of movements. To improve the signal-to-noise ratio, we implemented an approach to either sum over all fragments for speed and angles (features in eqs. (5) to (7)), or the average across the length (features in eqs. (1) to (4)) every six frames. Due to our sliding window function at about double the resolution of the bins prior to this step,

we were not particularly concerned with washing out inter-bin behavioral signals. Upon multi-dimensional embedding of our features using t-SNE, our objective function finds the minimal divergence between the distribution of high-dimensional versus the compressed dimension. This algorithm has been selected due to preservation of local structures going from high to low dimensions, allowing simplification without reducing the complexity. The locations of the embedded points in the low dimensional space, y_{ij}, are determined by minimizing the Kullback-Leibler divergence between joint distributions P and Q, by first assuming normal distribution of distance between any two observations,

$$p_{ij} = \frac{\exp(\frac{-||x_i - x_j||^2}{2\sigma_i^2})}{\sum_{k \neq i} \exp(\frac{-||x_i - x_k||^2}{2\sigma_i^2})}, \qquad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \tag{8}$$

where σ for each sample will be determined by some constraints (perplexity). To reduce overcrowding in the process of embedding, we utilize Student's t-distribution for distances between pairs of points in the low dimensional space,

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}. \tag{9}$$

To match the two distributions, we compute the Kullback-Leiber divergence KL between P and Q,

$$C_\epsilon = KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log(\frac{p_{j|i}}{q_{j|i}}) \tag{10}$$

and minimize the KL divergence loss with respect to given high dimensional data point y_i.

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1} \tag{11}$$

In simpler terms, similar objects, or mouse multi-joint trajectory in this case (high values of p_{ij}) will retain its similarity visualized in the low-dimensional space (high values of q_{ij}), scaled with a normalization constant $\sum_{k \neq l}(1 + ||y_i - y_j||^2)^{-1}$. To accelerate the dimensionality reduction process, we opted to perform Barnes-Hut approximation [34].

Grouping Expectation Maximization based on Gaussian Mixture Models [17] was performed to guarantee convergence to local optimum. We opted to randomly initialize the Gaussian parameters μ_k, Σ_k and π_k, over number of times to escape a suboptimal local optimum.

First, we evaluate the responsibilities using the initialized parameter values, or E-step,

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{k} \pi_j N(x_n|\mu_j, \Sigma_j)} \tag{12}$$

Second, we re-estimate the parameters using current responsibilities $\gamma(z_{nk})$, or M-step,

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \tag{13}$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T \tag{14}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{15}$$

Finally, we evaluate the log likelihood,

$$\ln p(X|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k) \} \tag{16}$$

to determine if the parameters or log likelihood has converged. If the convergence criterion is not satisfied, then let $\mu_k, \Sigma_k, \pi_k \longleftarrow \mu_k^{new}, \Sigma_k^{new}, \pi_k^{new}$, and repeat the E and M-steps.

Classifier design Since we are dealing with more than two classes, we performed error correcting output codes (ECOC) [21,35] to reduce the problem from multi-class discrimination into a set of binary classification problems. To build this SVM classifier, we consider the exemplar table of multi-learner design coding matrix (Supp. Table.1).

Following constructing the coding design matrix M with elements m_{kl}, and s_l as the predicted classification score for positive class of learner l. ECOC algorithm assigns a new observation to the $\hat{k}^{th}$ class that minimizes the aggregate loss for L binary learners.

$$\hat{k} = argmin \frac{\sum_{l=1}^{L} |m_{kl}| g(m_k l, s_t)}{\sum_{l=1}^{L} |m_{kl}|} \tag{17}$$

where g is the loss of the decoding scheme.

Frame-shift prediction paradigm Many end users may wish to apply the algorithm to higher frame-rate video. Because B-SOiD has a temporal constraint of $\sim$10Hz to maintain an optimal signal-to-noise ratio, we designed B-SOiD to predict along a sliding window. This is mathematically implemented using offsets. Assume $\hat{k}$ now to be a vector of behavioral assignments over time, let j be the number of maximum offsets in 100ms to be repeated for frame-shift paradigm,

$$\hat{k}(i) = \hat{k}_i \quad for \quad 1 \leq i \leq j \tag{18}$$

and weave together the assignment arrays $\hat{k}$. The behavioral output will match the resolution of the video camera's sampling rate. In Fig. 2g, to accurately quantify the consistency between predicting $\hat{k}$ using frame-shift output at 5ms versus the non-frame-shift at 100ms,

we masked out the transitional differences in resolution - 100ms before or after the non-shifted transition time.

Extraction of kinematics Stride length is computed as the euclidean distance of left hind paw displacement averaged amongst all locomotor bouts across our six C57/BL6 animals, and only when the left hind paw moves. Maximum stride speed is computed using MATLAB function $findpeaks$ per stride. Similarly, average groom distance is computed as the euclidean distance of right forepaw displacement averaged amongst all head groom bouts across our six C57/BL6 animals. Maximum groom speed is computed using MATLAB function $findpeaks$ per scratch.

Behvioral subjects and experimental set-up Normal, non-lesion experiment subjects were six, adult C57/BL6 mice, (3 females, Jackson Laboratory). Individual animals were placed in a clear, 15 x 12 inch rectangular arena for one hour while a 1280x720p video-camera captured video at 60Hz (cluster and ML) or 200Hz (frame-shift). Following each data session, any feces were cleaned out and the arena was thoroughly sprayed and wiped down with Virkon S solution. The arena was then allowed to dry and air out for several minutes. This video was acquired from below, 19 inches under the center of the field. Offline analysis was performed on MATLAB (MathWorks). While this program and the GPU required for data processing are not open source, the software package described in this manuscript is openly available https://github.com/YttriLab/B-SOiD. All statistical measures on behavior were non-parametric, two-tailed Kolmogorov-Smirnov tests unless otherwise mentioned. All animals were handled in accordance with guidelines

approved by the Carneige Mellon Institutional Animal Care and Use Committee (IACUC).

Indirect pathway cell-type specific lesion experiment Eight adult Adora2a-cre (often called A2A, Jackson Laboratory stock 036158, four females) mice were used to study the effects of cell-type specific lesion of striatal indirect pathway neurons. Half of these animals (two females) were injected with AAV2-flex-taCasp3-TEVP [36] 4×10^{12} vg/mL from UNC Vector Core bilaterally into the dorsomedial striatum: AP +0.9, ML +/-1.5, DV -2.65. Animals were allowed to recover at least 14 days prior to open-field experiments. The virus is designed to kill only cells expressing cre recombinase. To help visualize virus spread, a non-Cre dependent GFP virus, AAV2-CAG-GFP 4×10^{12} was co-injected. 1μL in each hemisphere was injected with a virus ratio of 2:1, Casp:GFP and a rate of 200nL/min. The GFP virus was added because only a portion of neurons (the cre positive cells) die as a result of this injection, and thus gross cell loss is more difficult to quantify. Quantifying decreased cell density by eye also produced similar lesion maps (see S3).

References

1. Gallistel, C. R. Representations in animal cognition: An introduction. *Cognition* (1990).

2. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience Needs Behavior: Correcting a Reductionist Bias (2017).

3. Markowitz, J. E. *et al.* The Striatum Organizes 3D Behavior via Moment-to-Moment

Action Selection. *Cell* **174**, 44–49 (2018). URL https://doi.org/10.1016/j.cell.2018.04.019.

4. Tanaka, S., Young, J. W., Halberstadt, A. L., Masten, V. L. & Geyer, M. A. Four factors underlying mouse behavior in an open field. *Behavioural Brain Research* **233**, 55–61 (2012).

5. Van Lier, H., Coenen, A. M. & Drinkenburg, W. H. Behavioral transitions modulate hippocampal electroencephalogram correlates of open field behavior in the rat: Support for a sensorimotor function of hippocampal rhythmical synchronous activity. *Journal of Neuroscience* **23**, 2459–2465 (2003).

6. Nath, T. *et al.* Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols* (2019).

7. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience* (2018).

8. Pereira, T. D. *et al.* Fast animal pose estimation using deep neural networks. *Nature Methods* **16**, 117–125 (2019).

9. Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S. & Branson, K. JAABA: Interactive machine learning for automatic annotation of animal behavior. *Nature Methods* **10**, 64–67 (2013).

10. Wiltschko, A. B. *et al.* Mapping Sub-Second Structure in Mouse Behavior. *Neuron* **88**, 1121–1135 (2015).

11. Everitt, B. J. & Robbins, T. W. Neural systems of reinforcement for drug addiction: From actions to habits to compulsion (2005).

12. Jin, X. & Costa, R. M. Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature* **466**, 457–62 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20651684http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3477867.

13. Ding, L. & Gold, J. I. The basal ganglia's contributions to perceptual decision making (2013).

14. Jin, X., Tecuapetla, F. & Costa, R. M. Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nature Neuroscience* **17**, 423–430 (2014).

15. Klaus, A. *et al.* The Spatiotemporal Organization of the Striatum Encodes Action Space. *Neuron* (2017).

16. Maaten, L. v. d. & Hinton, G. Visualizing Data Using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008). URL http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

17. Jordan, M., Kleinberg, J. & Schölkopf, B. Pattern Recognition and Machine Learning. Tech. Rep.

18. Aldridge, J. W., Berridge, K. C. & Rosen, A. R. Basal ganglia neural mechanism of natural movement sequences. In *Canadian Journal of Physiology and Pharmacology*, vol. 82, 732–739 (2004).

19. Kalueff, A. V. *et al.* Neurobiology of rodent self-grooming and its value for translational neuroscience (2016).

20. Berridge, K. C., Fentress, J. C. & Parr, H. Natural syntax rules control action sequence of rats. *Behavioural brain research* **23**, 59–68 (1987). URL http://www.ncbi.nlm.nih.gov/pubmed/3828046.

21. Escalera, S., Pujol, O. & Radeva, P. On the decoding process in ternary error-correcting output codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**, 120–134 (2010).

22. Yu, B. M. *et al.* Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology* **102**, 614–635 (2009).

23. Fujimoto, M. & Ariki, Y. Robust speech recognition in additive and channel noise environments using GMM and EM algorithm. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1 (2004).

24. Bartkova, K. & Jouvet, D. Impact of frame rate on automatic speech-text alignment for corpus-based phonetic studies. In *ICPhS'2015-18th International Congress of Phonetic Sciences* (2015). URL https://hal.inria.fr/hal-01183637.

25. Darmohray, D. M., Jacobs, J. R., Marques, H. G. & Carey, M. R. Spatial and Temporal Locomotor Learning in Mouse Cerebellum. *Neuron* **102**, 217–231 (2019).

26. O'Brien, D. E., Brenner, D. S., Gutmann, D. H. & Gereau IV, R. W. Assessment of Pain and Itch Behavior in a Mouse Model of Neurofibromatosis Type 1. *Journal of Pain* **14**, 628–637 (2013).

27. Rapoport, J. L. & Wise, S. P. Obsessive-compulsive disorder: evidence for basal ganglia dysfunction. *Psychopharmacology bulletin* **24**, 380–4 (1988). URL http://www.ncbi.nlm.nih.gov/pubmed/3153497.

28. Albin, R. L., Young, A. B. & Penney, J. B. The functional anatomy of basal ganglia disorders. *Trends in Neurosciences* **12**, 366–375 (1989).

29. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2018). URL http://arxiv.org/abs/1802.03426.

30. Yttri, E. A. & Dudman, J. T. Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature* **533**, 402–6 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27135927http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4873380.

31. Graybiel, A. M. & Saka, E. A genetic basis for obsessive grooming. *Neuron* **33**, 1–2 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/11779470.

32. Berridge, K. C., Aldridge, J. W., Houchard, K. R. & Zhuang, X. Sequential super-stereotypy of an instinctive fixed action pattern in hyper-dopaminergic mutant mice: A model of obsessive compulsive disorder and Tourette's. *BMC Biology* **3** (2005).

33. Cisek, P., Puskas, G. A. & El-Murr, S. Decisions in changing conditions: The urgency-gating model. *Journal of Neuroscience* **29**, 11560–11571 (2009).

34. Van Der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. Tech. Rep. (2014). URL http://homepage.tudelft.nl/19j49/tsne;.

35. Dietterich, T. G. & Bakiri, G. Solving Multiclass Learning Problems via Error-Correcting Output Codes. Tech. Rep. (1995).

36. Yang, C. F. *et al.* Sexually dimorphic neurons in the ventromedial hypothalamus govern mating in both sexes and aggression in males. *Cell* **153**, 896–909 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23663785http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3767768.

Acknowledgements We would like to acknowledge to acknowledge Mary Cundiff in the Aryn Gittis lab for providing A2A lesion animals. Special thanks to members of the Yttri lab for comments on the manuscript. This work was supported by the Brain Research Foundation Fay/Frank Seed Grant.

Author contributions - AIH acquired and analyzed the data, conceived the algorithmic advancements, created the software and drafted the manuscript. EAY conceived the study, analyzed

and interpreted the data, and drafted the manuscript.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to Eric A. Yttri (email: eyttri@andrew.cmu.edu).