bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A-SOiD, an active learning platform for expert-guided, data efficient discovery of behavior.

Jens F. Schweihoff [1,†], Alexander I. Hsu [2,†], Martin K. Schwarz[1,‡*], and Eric A. Yttri[2,3,‡*]

November 4, 2022

[1]Functional Neuroconnectomics Group, Institute of Experimental Epileptology and Cognition Research, Medical Faculty, University of Bonn, Bonn, Germany
[2]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA USA
[3]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA USA

Running title: Targeted pattern discovery of behaviors.

[†] These authors contributed equally
[‡] These authors contributed equally
[*] Email: eyttri@andrew.cmu.edu, Martin.Schwarz@ukbonn.de

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Behavior identification and quantification techniques have undergone rapid development. To this end, supervised or unsupervised methods are chosen based upon their intrinsic strengths and weaknesses (e.g. user bias, training cost, complexity, action discovery). Here, a new active learning platform, A-SOiD, blends these strengths and in doing so, overcomes several of their inherent drawbacks. A-SOiD iteratively learns user-defined groups with a fraction of the usual training data while attaining expansive classification through directed unsupervised classification. In socially-interacting mice, A-SOiD outperformed standard methods despite requiring 85% less training data. Additionally, it isolated two additional ethologically-distinct mouse interactions via unsupervised classification. Similar performance and efficiency was observed using non-human primate 3D pose data. In both cases, the transparency in A-SOiD's cluster definitions revealed the defining features of the supervised classification through a game-theoretic approach. To facilitate use, A-SOiD comes as an intuitive, open-source interface for efficient segmentation of user-defined behaviors and discovered subactions.

Keywords - computational neuroethology, naturalistic behavior, machine learning, social interactions, mouse behavior, primate behavior

Introduction

Naturalistic behaviors, particularly social interactions, provide a rich substrate to understand the brain and the decisions it makes. Cutting edge machine learning algorithms now enable researchers to capture the movement of individual body parts with markerless pose estimation [1–7]. These pose estimation algorithms can then be readily used to extract behavioral expressions in a previously unmatched level of detail and temporal resolution [5, 6, 8–15].

One approach that utilizes pose estimation data to extract behaviors is to reproduce expert human annotation in an automated fashion. A major advantage of this supervised approach is the direct control over the initial definition of the behavioral expression, incorporating the expertise of researchers into the classification process. However, for these supervised methods, a sizeable, manually-annotated data set is required to learn and reproduce human rate annotations [8, 16–19]. A potentially more serious issue, the reproducibility between and within research groups is known to suffer as the annotation process is prone to inherent biases and rater fatigue [6, 20]. Furthermore, as investigators aim to untangle a more complete and detailed behavioral repertoire, supervised algorithms are unable to generate new insights that build upon the what they have already found. Related to this, classification models often fail to capture a concise account of the learned reasoning, or decision boundaries of the algorithm - instead relying upon qualitative descriptions of the annotator's intuitive reference frame [21]. Consequently, reproduction

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and comparison of the classification process can become a matter of subjective re-interpretation and escalating inter-rater variability.

The alternative approach is agnostic to experimenter definitions, focusing instead on uncovering the conserved spatiotemporal structures within pose dynamics [6]. These unsupervised models can find known and hidden behavioral expressions without the need or influence of human annotation. Specifically, unsupervised pattern discovery can be directly applied to provide behavioral expressions with temporal resolution beyond human ability, into sub-second components and distinct sub-actions with high sensitivity [9–11, 22–24]. This major benefit is also its key drawback: the algorithm can only identify patterns that are statistically obvious given the provided input features - i.e. rare events or more subjective distinctions between behaviors will not be identified as unique clusters. Thus, behavioral expressions that are evident to the experimenter - but are not readily statistically discerned - often cannot be reconstructed. Additionally, identified behavioral patterns are often assigned semantic names that may obscure more complex underlying feature statistics if used without proper validation. Nevertheless, behavioral patterns may be discovered that do not conveniently fit traditional nomenclature, particularly at temporal scales beyond the typical spectrum. In these cases, researchers are often constrained to use associative descriptive names (e.g., grooming sub-type A) or token names (e.g., motif 1) to report their findings.

Supervised and unsupervised methods focus on different feature-integration approaches to segment behavior. As a result, each fails to combine both informed analysis and efficient pattern discovery within the same workflow. An ideal solution would be able to reproduce an expert researcher's informed annotations and translate them into a transparent, reproducible format. In addition, the researcher would be able to engage the power of agnostic discovery of conserved movement patters within the same framework to facilitate deeper behavioral understanding without occluding the functional rationale behind newly found components. Notably, in the field of machine learning there are already developments striving to replace black-box approaches for better *post hoc* explainability [25, 26]. With these, information about the feature composition of learned classes can be inferred which enables researchers to translate intuitive descriptions into transparent, operationalized definitions for the comparison between data sets and methods [21, 27].

Here, we present A-SOiD, an active learning platform that incorporates unsupervised discovery of spatiotemporal movement patterns. A-SOiD outperforms traditional supervised methods and does so with significantly less data than traditional supervised models. By automatically balancing annotation sets, A-SOiD reduced the amount of annotations required ($\approx 85\%$ reduction). Because A-SOiD requires only very small input data to reproduce annotations, we were able to expand an initial set of behavior categories to include newly discovered behaviors autonomously with high predictive performance. It also provides an

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

entry point for discovery to expand the classification annotation set through unsupervised segmentation of a selected behavior. To test its performance, we applied A-SOiD to two benchmark data sets. First, social interactions in rodents are complex and nuanced. While supervised approaches can be trained to identify select interactions, there are still substantial limitations on the specificity of behavioral segmentation - even with large training data sets. We investigated the sizeable human-annotated data set of social behavior in mice (CalMS21) [28] with A-SOiD and then extended the range of detected canonical social behaviors. Next, we note that there is currently a lack of behavioral segmentation methods for both 3D pose estimation and for non-human primates, despite rapid growth in these fields [7, 29]. To demonstrate the flexibility of A-SOiD, we analyzed a three-dimensional non-human primate pose data set. Finally, to facilitate A-SOiD's use in the rapid segmentation of a behavioral repertoire, we packaged the platform into an intuitive open-source graphical user interface that can be used without prior coding experience (Supp. Fig. 1).

Results

User definitions of complex behavior are not readily identified using unsupervised approaches

Experimenters often focus their initial behavioral quantification on a few selected behaviors but would like to expand their analysis by exploring the inherent data structure (unsupervised). The assumption being, that human definitions will be self-evident from the given data representation, so that further exploration can be directly aligned to previous findings. We therefore first investigated whether an unsupervised classification approach would be able to reproduce human annotations given a large benchmark data set of socially-interacting rodents.

The CalMS21 data set is a large social behavior benchmark data set consisting of annotations from four expert human raters for three distinct behavioral categories (attack, investigate and mount; see Methods) were annotated (Fig. 1a-b, see Methods and [6, 28]). The data set also provides pose estimation data of the two socially-interacting mice (Fig. 1a). To extract a single, homogenized set of annotations, we divided the behavioral classifications into non-overlapping segments. For this, we examined the distribution of bout lengths across annotations and found that a period of 400 ms (12 frames 10%tile, Fig. 1c) would be sufficient to resolve behavioral changes across annotations. Given that 12 frames could contain more than 1 type of annotation, we examined the annotation pattern in depth. We found that $\approx 92\%$ of the 400 ms non-overlapped segments contained only a single behavioral annotation. The remaining $\approx 8\%$ had at least 2 types of annotation within the same segments. Further examination of the remaining $\approx 8\%$ cases revealed a predominantly annotation dominated the segments, thus a tie-breaker was rarely required.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Our results suggest that, compared to frame-wise annotation, there is minimal loss of information upon downsampling of annotations to 400 ms.

To capture the spatiotemporal pose relationships of the annotated social interactions, we extracted features (see Methods) from both the individual animals (intra-animal; Fig 1d left) and the multi-animal interactions (inter-animal; Fig. 1d right). Notably, we observed that the distribution of single features was already indicative of the human annotations, e.g. mounting is characterized by a small inter-animal snout-snout distance while attacking typically possesses a greater speed (Fig. 1e top-left). However, significant overlap can occur (compare Fig. 1e bottom-left), which necessitates the evaluation of the composite feature distributions to fully represent a behavioral class (Fig. 1f).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

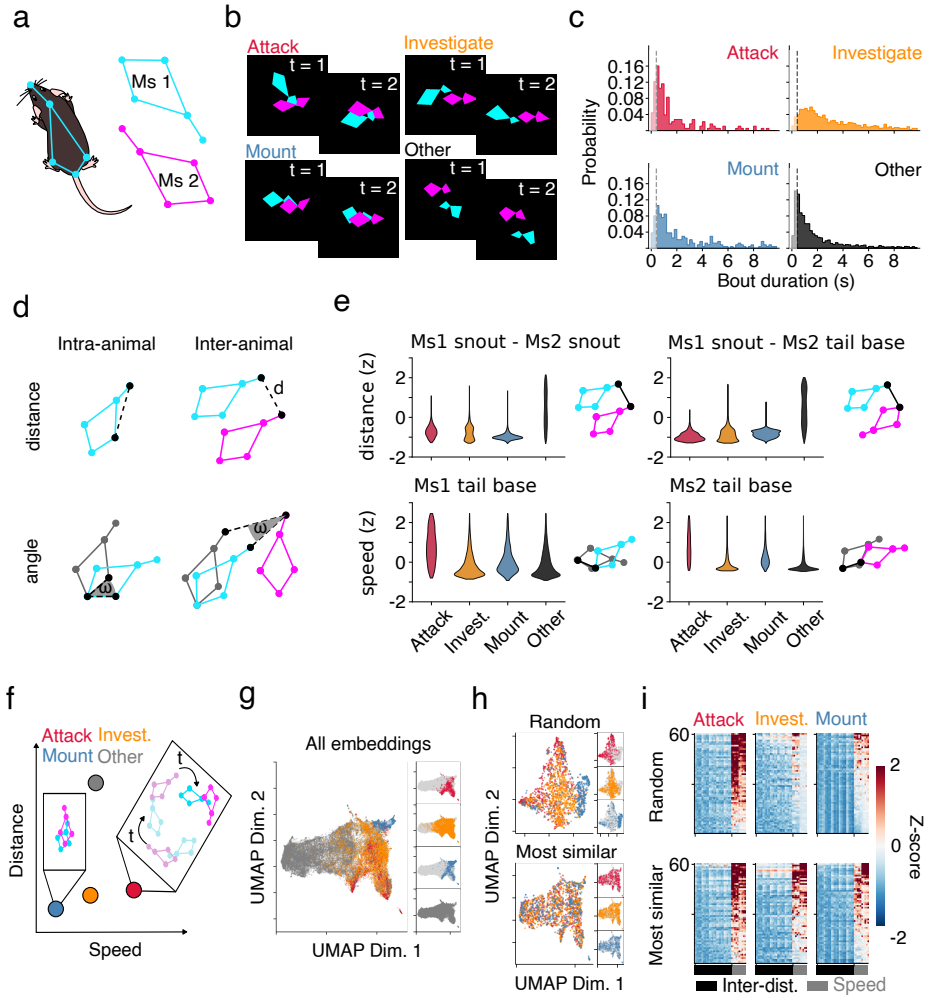

Figure 1: Human annotations of social interactions cannot be easily represented in unsupervised embeddings. a) Schematic representation of body positions on two mice (Ms1: cyan; Ms2: magenta) taken from the CalMS21 data set. b) Annotated example frames from the CalMS21 training data set (Red: attack; Orange: investigate; Blue: mount; Black: other). Behavioral color-code to be maintained throughout manuscript. c) Histograms showing the distribution of annotated frames before a transition into a different behavior occurs. Dashed lines indicate 400 ms (12 frames) that was used to integrate our features over, temporally. d) Intra-animal and inter-animal distances (top), and respective angular changes (bottom) are calculated for all combinations. e) Example feature distributions across annotated behaviors (top: snout-to-snout and tailbase-to-tailbase inter-animal distance, bottom: sample-to-sample tailbase speed of Mouse1 and Mouse2). f) Schematic representation of unique composite feature distribution for each behavioral expression. g) UMAP embeddings of composite features, colored by annotations. h) UMAP embeddings using a random subset (top) and using the same count, but more similar subset. i) 2D-histogram showing normalized feature values (z-scored) across features (columns) and selected samples (rows) of all three classes (attack, investigation, mount). Here we show a subset (n = 60) representing (h) for ease of visualization.

With this realization, we next explored whether an unsupervised approach [9] (see Methods) could resolve the human annotation space given the ex-

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tracted features (Fig. 1g). In the feature distribution (Fig. 1e), individual behaviors could be largely separated by a composite of several features. If the human annotations can be readily described by the selected features, a high-dimensional representation of all points in the feature space or its low dimensional embedding (UMAP) should provide distinct clusters that readily map onto each group (e.g., all examples of attack in the same area (high speed, low distance) whereas mount is in a different region (low speed, low distance); Fig. 1f). While an unsupervised approach has been proven successful for extracting common, single-animal behaviors [9–11, 22, 24, 30], when applied to the social interactions of the CalMS21 data set, we found that human annotation cannot be reliably represented as clusters of spatiotemporal pose relationships (Fig. 1g). Given the characteristics of unsupervised approaches, a potential cause for failure is the highly unbalanced distribution of data in its raw state. The majority of annotations are designated as "other" (n = 26409, see Methods) while all remaining behaviors are only represented by a substantially smaller amount of examples (attack: n = 1188; investigation: n = 12300, mount: n = 2378). The clustering algorithm is inclined to emphasize the differences between data points in "other", and consequently will overlook the differences in smaller classes. As such, we see that the "other" group spans the entire embedded space (grey, Fig. 1g), encompassing all other classes within. It is important to note that this bottom-up approach is agnostic to human classification and considers all points as equally important. The top-down human annotations were only added to the embeddings in order to visualize the lack of separation (Fig. 1g).

We further investigated whether a data set [28] comprised of a more balanced sample of all classes that contained clear definitions (attack, investigation and mount, see Methods) would yield better results in unsupervised embedding. Notably, this would not be possible without supervised annotations as an unlabeled data set would not be differentiable without prior human curation. First, we selected a random sample (n = 580, 872, 581 for attack, investigation, and mount, respectively) and embedded them together to determine how an unsupervised approach would perform. While these embeddings yielded better reproduction of the desired annotations, there appears to be subgroups within each annotation. Consequently, this approach does not resolve the entirety of the human annotation spectrum (Fig. 1h,i top).

Next, while maintaining the same number of samples from the three classes, we selected those that appear most similar amongst the three annotation types. We discovered that there are a select group of samples that rarely differ in features (Fig. 1h,i bottom). One possible explanation for these is that these samples are transitions between two behaviors and could be labeled as either or - i.e., where does a rater decide when a behavior stops and a new one starts. We first examined the preceding behavior for each behavior that the algorithm deemed uncertain and found that when examining the subset of samples that are most similar in feature space (hardest to predict), the consistency of annotations

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between prior frame to current frame (transition) is significantly lower than a random subset of samples (Supp. Fig. 2b). One key finding is that the rater's decision at which frame the animal goes from investigation to mount is inconsistent, as there is a higher rate of investigation to mount transitions and vice versa (Supp. Fig. 2a). As expected, this is in line with the type of predictions errors by baseline models from Sun et al. [28].
Moreover, when investigating whether our classifications show a bias towards transitions "out of" a behavior or transitions "into" one, we found that when considering the samples that are the most similar, that, overall, the consistency of annotations between current frame and the next frame is similar to transitioning "into" a behavior (Supp. Fig. 2d). More specifically , we found that, for the annotations in the used CalMS21 data, the particular inconsistency varies between behaviors for "into" vs "out of" transitions of the investigation class. For example, while the rater was less consistent with labeling a transition from mount into investigation, the inconsistency was higher for transitions out of investigation into other (Supp. Fig. 2a, c). Again, in line with prediction error types made by baseline models from Sun et al. [28].

Thus, a purely data-driven approach such as unsupervised classification is unlikely to fully reproduce human annotations in this social interaction data set. However, this interrogation of the data structure revealed key inroads to improved segmentation. Anticipating the challenges of an unbalanced data set and problematic edge scenarios, we focused on developing a solution that would automatically balance the training data and integrate human annotations in a data-efficient manner. Finally, we embraced the possibility that a data-driven approach, such as unsupervised clustering, may perform better when applied to the restricted, segmented classes - and allow further discovery of conserved patterns.

Active learning automatically balances training data and integrate human expertise with high performance

We developed A-SOiD - a GUI pipeline (Fig. 2a; Supp. Fig. 1) that includes both active learning and directed unsupervised behavioral segmentation. A-SOiD makes use of an iterative, active learning paradigm that selectively trains on low-confidence examples to improve classification robustness. Rather than blindly feeding in additional human annotations, A-SOiD queues a subset of low confidence predictions from the rest of the training using each previous classifier iteration. By focusing on the potentially problematic edge cases (Fig. 1h, i bottom, Supp. Fig. 2), the algorithm greatly reduces the overall number of annotated frames required (Fig. 2a, step 1-3). Beyond initial annotations, A-SOiD provides users an unsupervised clustering algorithm [9] to explore and further subdivide existing annotations (Fig. 2a, step 4).

To directly overcome the bottleneck of supervised approaches requiring massive training data sets while reducing the possibility of de-emphasizing the smaller

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

annotations like "attack", we first initialized A-SOiD by providing mere 1% of samples that were randomly selected from each of the three annotations. We maintained the relative proportions when initializing to maintain a relatively similar feature distribution to the entire pool. Not surprisingly, the resulting predictions underperform compared to using the entire data set (Fig. 2b, iteration 1). Next, a selected subsampling of low confidence predictions on the remaining training data (i.e., most similar across classes, Fig. 1h,i bottom) by the initial classifier (iteration 1) is refined by extracting annotations from the remaining training labels or by prompting human refinement. In subsequent iterations, the latest classifier is used to identify new low-confidence predictions and to refine them for the training of the following iteration. This establishes an iterative active learning scheme that focuses on clarifying annotation preference only for examples that lie at the decision boundary between classes (Supp. Fig. 2). With this approach to refinement, performance on a completely unseen, held out test data set improves beyond using the entire available training data with just 12% of the labels necessary, or fewer than 1000 samples per class (Fig. 2b, dashed line represents performance using all training data, black indicates average across all three classes). A-SOiD performance reached 0.874 ± 0.002 macro average f1 score and 0.918 ± 0.001 MAP with 20 cross-validation runs - on par with *Top-1* performance in the MABe 2021 Task1 Challenge [28], and without the use of additional unlabeled data set. This increase in performance is paralleled by a drop in additional samples per iteration, as the quantity of low-confidence samples sharply drops. Additionally, we observed that the number of training samples per behavioral expression became more balanced out over active learning iterations, even when the total available counts, and subsequently our initialized classifier inputs, were largely biased towards one behavior expression (investigation; Fig 2b bottom, orange of "Full" dataset). Therefore, these benchmark data suggest that even if A-SOiD starts from scratch with a small, random sample of frames, performance quickly exceeds that from classifiers trained on the full training data set at once by nearly an order of magnitude.

To understand the impact of individual features in the decision process of our iterative-learning classifier, we performed SHAP analysis (SHapley Additive exPlanations , see Methods) [25, 26] on representative iterations (1, 6, 11, and 20). SHAP is a game theoretic approach to assign credit to the individual feature underlying the performance of an algorithm. Each example (dot) from a feature provides both the explainability (X position) and the normalized feature value - e.g. red indicates greater speed or distance - with consistent local coloring indicating a conserved feature-explanability relationship.

In examining "attack", the class that was learned the slowest, we found that the critical spatiotemporal features quickly standardizes across subsequent iterations (Fig. 2c). While initial features such as the Resident mouse's nose speed ("RNose", grey; also Resident nose-Intruder neck distance "Rnose-Ineck", black) have a consistently high impact on the model's decision, other features

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

become more important and maintain their importance rank across iterations. As expected, the impact of speed features (grey) is mostly negative (SHAP value < 0.0) and skewed towards low values (blue dots) - i.e., low speed values are a good predictor against the attack class, as would be predicted from the full training data set (Fig. 1e). In this way, SHAP analysis provides a direct, transparent explanation for the performance of the autonomous active learning through an observer's point-of-view. It also provides potential insight into label discrepancy/misalignment amongst various human raters, a major area for improvement [6]. While investigation and mount classes improve less, we still see shuffling of the top five feature ranks (Supp. Fig. 3).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

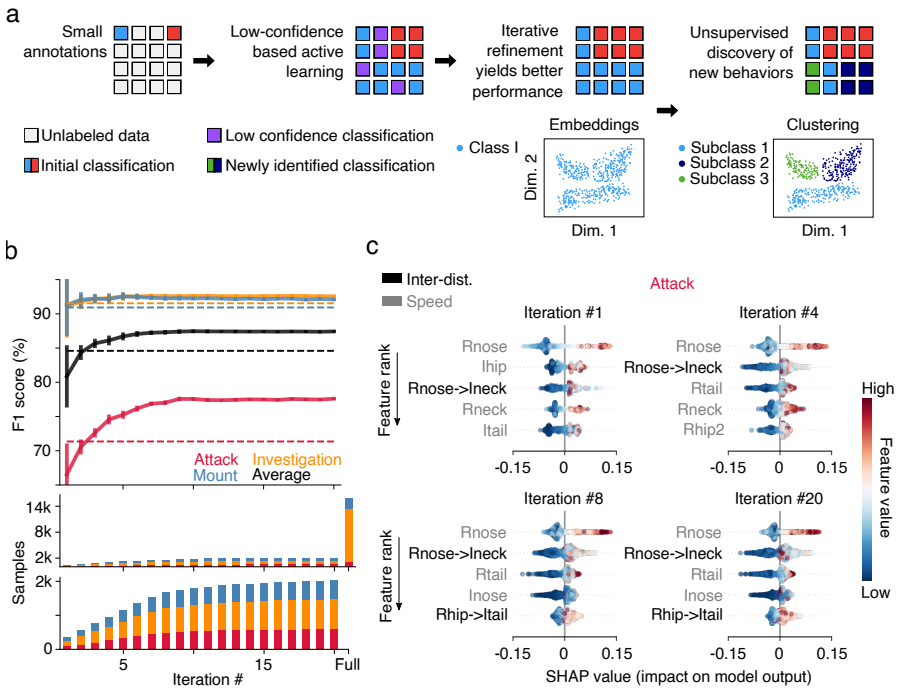

Figure 2: Active learning improves data efficiency and overall performance. a) The A-SOiD pipeline initializes a classifier with minimal training annotations (step 1). Next, low confidence predictions on the remaining training data are subsampled for researchers to refine manually (step 2) or automatically taken from the available ground truth. Step 2 reiterates until there are no remaining low confidence training samples (step 3). Lastly, A-SOiD can expand the annotation set by deploying directed, unsupervised pattern discovery (step 4). b) A-SOiD performance (F1 score, 20 cross-validations, top) on held out data outperforms a classifier trained on the full data at once (dashed lines). This is achieved through 12% of data (middle), with a more balanced representation across sampled annotations (bottom). Red: Attack; Orange: Investigate; Blue: Mount; Black: Average across classes. Full indicates complete annotated data set. c) Ranked order of the top five features (descending order) across iterations according to SHAP analysis (see Methods) for the "attack" class, including individual feature impact (x-axis) separated by relative feature value (High: red, Low: blue). The differences in top-5 features allow insights into the learning process (compare iteration 1, 4 and 8). In contrast, the top-5 features do not change after the plateau of test performance in (e, compare iteration 8 and 20). Features (inter-animal distance: black, speed: gray) are denoted by their corresponding animal (R: resident, I: intruder) and body part (e.g., nose).

Fast and efficient pattern discovery

While A-SOiD's active-learning component allows the reproduction of human annotations in a highly efficient manner, it lacks the ability to discover conserved patterns of behavior and thereby limits researchers to their initial set of behaviors. This limited set may be due to a lack of comprehension of the behavioral repertoire during annotation or difficulties in robustly defining

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

rare or nuanced sub-actions. For example, the CalMS21 data set does not inherently differentiate between the several known investigation types. As such, investigative behavior sub-types are hidden and dispersed within the "investigation" class (Fig. 3). However, it is often desirable to be able to split high-level behaviors into component, conserved sub-actions. Specifically, in social interactions encompassing both male and female mice, anogenital investigation is a primary behavioral expression when mice engage in olfactory investigation. Quantification of such behavior can therefore serve as a key identifier for social recognition, including habituation and discrimination [31, 32] (`https://mousebehavior.org/investigate-anogenital/`) which in turn indicates a branching point in behavioral strategies depending on the specific outcome.

Therefore, for the directed discovery of hidden sub-behaviors, we included an unsupervised classification step and discovered conserved sub-types of investigative behavior. Unsupervised embedding and clustering of the "investigation" class separately from the full data set revealed several sub-classes (Fig. 3a, see Methods). Of these, two clusters appear to be investigation specifically at the anogenital areas. More specifically, one of the sub-classes consists of the resident mouse approaching and directing its investigation at the anogenital area of the intruder mouse from behind (sub-class 2, "anogenital approach", n = 3234) - and the other of the resident mouse investigating the anogenital area of the other mouse while already in close proximity (sub-class 5, "anogenital investigation", n = 706; `http://mousebehavior.org/ethogram-index/`; Fig. 3c). To allow a direct comparison between unsupervised clusters and the heuristic criterion, that the snout of one or both mice must be close to the tail base of the other, we annotated all data points within the investigation class in which the distance between the resident's snout and the intruder's tail base was lower than a manually set threshold (15 pixels; Supp. Fig. 5). A comparison revealed an extensive overlap between the unsupervised clusters of sub-class 2 and sub-class 5 with the top-down, manually selected feature space.

We then computed the motion energy (see Methods) for each cluster. Motion energy analysis results in single images that are an average of the motion spectrum relative to an individual, aligned animal (Fig. 3b). Consequently, conserved behaviors with repeated distinct movements result in a bright, clearly defined image, while behaviors that include divergent movement patterns appear darker and blurry (Fig. 3b). In both example clusters the average resident snout's motion energy is concentrated close to the anogenital region of its con-specific, unlike the widely distributed motion energy found in the investigate class (Fig. 3c). Specifically, sub-class appears restricted to motion behind the centered animal, i.e., a targeted approach to the anogenital area, "anogential approach", while sub-class 5 consists of paralleled resident/intruder anogenital investigations, "anogenital investigation" (Fig. 3b). We confirmed the motion energy inspection by extracting example episodes (Supp. Video 1) of the found behavioral sub-types.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To test A-SOiD's ability to integrate discovered sub-types into our classifier, we provided additional labels for sub-class 2 ("anogenital approach") and sub-class 5 ("anogenital investigation") by splitting them from investigation , while all remaining clusters remain in the original class "investigation" (Fig. 3a). We then retrained a classifier using the expanded annotation set. Since we only clustered the provided training set from the CalMS21 data set, we decided to split the training set into a test and train set, which resulted in a reduced overall performance in this particular section compared to active learning using the full training set (compare with Fig. 2b). Similar to the active learning performance described in (Fig. 2b), we found that after 30 iterations, we reached higher predictive performance across all categories, while attack still improved the most (Fig. 3d). Taken together, these results demonstrate the possibility of iterating between supervised and unsupervised classification.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

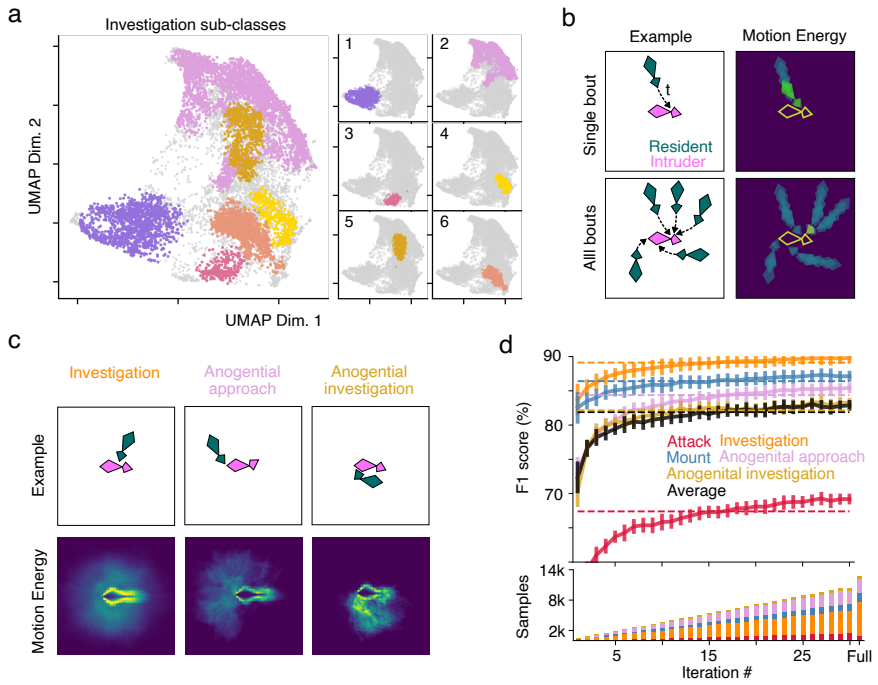

Figure 3: Unsupervised Clustering can be used to discover and integrate novel behavioral expressions in previously unspecific data. a) UMAP embeddings followed by HDBSCAN clustering (see Methods) of only the "investigation" class revealed 6 behavioral sub-classes. b) Schematic representation of motion energy in single bout (top) and average across all bouts (bottom). c) Representative examples (top row) and motion energy plot (bottom row) of the investigation class including two sub-types that closely align with heuristic identification of anogenital investigation (Suppl. Fig. 5) d) With new sub-classes redefined within the existing training data set, A-SOiD (F1 score, 20 cross-validations, top) outperforms a classifier trained on the full data at once (dashed lines). Note that in this case, we are only considering a subset of the CalMS21 training data (80%/20% split within the CalMS21 training data described throughout the manuscript) to allow testing on the remaining data. This performance increase is achieved through less data (bottom). Red: attack; Orange: investigate; Blue: mount; Pink: anogentital approach (sub-class 2); Gold: anogenital investigation (sub-class 5); Black: average across classes

A-SOiD demonstrates improved performance regardless of species or spatial dimensions

To demonstrate the flexibility of our approach, we applied A-SOiD to position information and human annotations of singly-housed non-human primates. Notably, pose information was computed using OpenMonkeyStudio [7], which provides 3D pose-estimations. The video was manually annotated, separating the animal's behavior into 5 categories: walk, rear, jump, ceiling climb (Climb C), and side-wall climb (Climb S), (Fig. 4a, or see Methods for detailed description of our annotations). Next, we trained A-SOiD to reproduce the human rater

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

annotations (n = 64, 177, 50, 214, 676 for ceiling climb, sidewall climb, jump, rear and walk respectively). In this case, 15 active learning iterations were sufficient to reach plateau performance (Fig. 4b). Similar to the performance in the social mouse benchmarking, we found that A-SOiD automatically balanced the training set to include examples of classes that were initially underrepresented across iterations (jump, Fig. 4 b). The overall predictive performance improved beyond the performance of a classifier trained on the full data all at once (compare Fig. 4b, dashed lines). SHAP analysis was again used to reveal the refinement process during active learning iterations that enabled the classifier to considerably increase predictive performance in this class. For example, the emerging feature importance (compare Fig. 4c iteration 1 and 15) of the speeds of hip and tail, followed by the distances between hip and right foot, as well as the distance between neck and tail, reveal how the jump was better defined with these composite features. While the classification of other monkey behaviors improved less, we still see shuffling of top five feature ranks (Supp. Fig. 4). These results demonstrated that A-SOiD performed just as well in a 3D, single-housed, non-human primate data set as it had with 2D, resident-intruder, mice data set (CalMS21).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

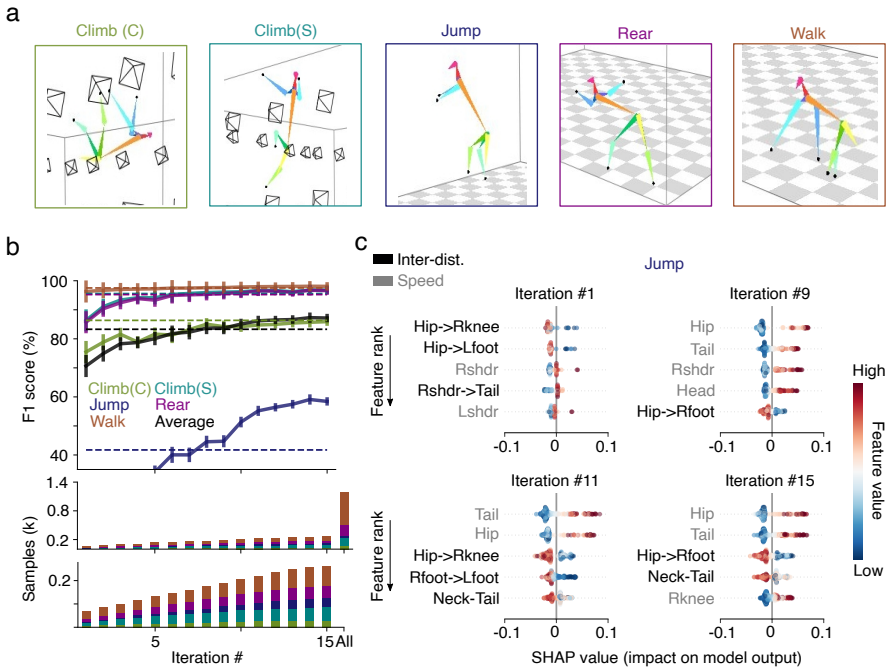

Figure 4: Efficient segmentation of monkey behavioral repertoire a) Representative frame examples for each annotated behavior. Images are reconstructed by OpenMonkeyStudio (3D pose). b) A-SOiD performance (F1 score, 20 cross-validations, top) on held out data outperforms classifier trained on all data at once (dashed lines). This is achieved through 12% of data (middle), with a more balanced representation across the annotations (bottom). Light Green: Ceiling Climb (Climb C); Cyan: Sidewall Climb (Climb S); Blue: Jump; Violet: Rear; Dark orange: Walk ; Black: Average across classes. c) Ranked order of the top five features (descending order) across iterations according to SHAP analysis for the "Jump" class, including individual feature impact (x-axis) separated by relative feature value (High: red, Low: blue). The differences in top-5 features allow insights into the learning process (compare iteration 1, 9, 11 and 15). The learning stopped prematurely due to lack of samples in a 5 minute video. Features (distance: black, speed: gray) are denoted by body part (e.g., Tail).

Discussion

As computational ethology explores the power of Big Data, the tools used to extract meaningful information must enable researchers to report transparent findings and allow generalizability across experiments. This is especially true in neuroethology, where variations in behavioral detection can fundamentally alter neural correlations or preclude their detection entirely [6, 9, 10, 20, 33]. To this extent, public benchmark data sets such as the social rodent data set CalMS21 [28] serve as a meaningful resource for developing and benchmarking new tools. We used this data set to develop a framework for the

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data-efficient integration of human-expert annotations and directed pattern discovery called A-SOiD. We then demonstrated the power of the approach, applying it to a completely unrelated data set - primate 3D pose estimation data.

Thus, A-SOiD facilitates building user-defined behavior analysis pipelines with high-data efficiency. A-SOiD's active learning framework resulted in an 8-fold decrease in required samples (total available labels = 15866, used labels at performance peak = 1866; see Supp. Table 2) to reproduce human annotations of socially active rodents with high performance (Fig. 2b). Even in the small non-human primate data set (n = 1181 total labels), A-SOid robustly increased the performance of the low-frequency behavior (jump, n = 50) by selectively training on low confidence examples (Fig. 4b, see Supp. Table 3). Moreover, the resulting analysis of the model's output can reach different levels of complexity depending on the research question. The unsupervised clustering of the embedded space of an annotation-based class was able to discover well-known behavioral phenotypes that were not the focus of the initial annotations and likely would not have been possible with a purely top-down approach(see Fig. 3, Supp. Fig. 5).

When annotating behaviors, experimenters are predisposed to mirror the biased distribution of behaviors in the data set, thus producing an unbalanced training data set. To improve the performance of supervised algorithms, experimenters add additional frames, but this rarely overcomes the initial imbalance. However, the root of the problem is that classifier objective functions strive to improve overall performance. Although there are good solutions for binary imbalanced classifications [34–37], multi-class imbalanced classification is not as well developed [38, 39]. The main issue lies on the varying relationships between classes, as there could be two out of three classes that are balanced (both large), while the third one is small, or vice versa [40]. Our solution is to implement intelligent selection of samples (active learning) within the already given annotations to reduce the bias towards larger classes. By starting with an absolutely minimal number of annotations, we let A-SOiD determine which samples are to be annotated. This approach effectively sparsifies representation, while focusing on the outliers.

Intuitive A-SOiD GUI for improved research integration

To improve the efficacy of this approach, A-SOiD comes as an app that can be installed on local computers without additional coding (Supp. Fig. 1). The app guides researchers through active learning and facilitates training high-performing classifiers based on an initially labeled data set. Researchers are then enabled to further refine their classifiers on unlabeled data using an intuitive interface in which low-confidence examples are presented for replay and evaluation. While the app is capable of reproducing the results from the CalMS21 data set reported in this study, we expanded its data importation capabilities to include the two most common open-source pose estimation

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

solutions (DeepLabCut [1] and SLEAP [4]). Further, once a classifier reaches satisfactory performance levels, the app can be used to predict behaviors on novel data directly. The classifier can also be exported and deployed in custom analysis pipelines, including closed-loop experiments [12].

A-SOiD targets an unmet need in behavior analysis tools and provide researchers with an accessible, conjoined solution of supervised and unsupervised classification.

Quantifying classifier performance

An important unresolved issue in behavioral classification is the quantification of classification performance. While there is a general consensus concerning the relative features of several behaviors, there is no ground truth. Thus, there is a need for , comparative methods that can be used to quantify algorithmic performance. In this manuscript we utilize two approaches that focus on the internal consistency of behavior classes themselves, rather than an external, top-down rule.

Motion energy can quickly asses the differences between movement patterns (e.g., behavioral sub-types). Motion energy is an intuitive and informative way to generate visual summary of the action within a found cluster [41]. In our hands, we utilized motion energy images to quickly differentiate sub-types of anogenital investigation (Fig. 3). Note, that comparative analysis can be done by analyzing the energetic variability within and across groups, providing a valuable statistic for cluster quality [9].

Another approach is using algorithmic explainability metrics, such as SHAP-based reporting [25, 26] of the underlying feature importance which can help to share and compare conserved patterns across studies (for a review see [21]). The feature value impact and ranking not only describe the refinement process but also serve as a looking glass into the underlying intuitive human reference frames by translating reproductions of human annotations into transparent, operationalized definitions. In this study, we employed SHAP-analysis to investigate the learning process across multiple iterations during active learning and could identify that specific sets of features accounted for the increased performance of our classifiers. Once the algorithm is trained, these same values can be used to compare the classification reference frame across various models.

Acknowledgements

We would like to acknowledge COVID, for pushing the field towards more computational and open formats. We thank Ann Kennedy, Jennifer Sun, David Anderson and the team that created the CalMS21 data set for providing a comprehensive data set that can be used to benchmark current and future ap-

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

proaches to the classification of social behavior in mice. We also thank Drs. David Barack, Felipe Parodi, and Michael Platt for the monkey pose data.

Author Contributions

JSF, AIH, MKS, and EAY wrote and reviewed the manuscript. JSF proposed the idea and AIH designed the core functionality and main analysis scripts. Both JSF and AIH created the app and worked on analysis and figures. JFS annotated the primate data. EAY and MKS provided support and funds.

Code Availability

The app, further documentation and the open-source code written in Python can be found under (https://github.com/YttriLab/A-SOID). The code, including the code to generate these figures is open-source and available through a GitHub repository.

Data Availability

The data set (CalMS21) used in this study is available online (https://data.caltech.edu/records/1991) [28].

Competing Interests

The authors declare no competing interests.

Methods

Data processing feature extraction

With increasing sampling frequency, the intra-frame differences that are critical to determining the spatiotemporal features (e.g. speed) diminish. For instance, 30 fps sampling provides an inter-frame interval of only 33.3 ms - relegating the changes in position to a similar magnitude to the jitter in the position signal itself. To improve the signal-to-noise ratio in CalMS21 social mice data set, we analyzed the duration distribution of these annotated behaviors, and established a non-overlapping 400 ms windows to integrate signal over, and then either sums (displacement, anglular change) or averages (distance) over all 10 fps samples. The window was defined as 200 ms for the single-housed non-human primate data set. Thus, for the 10 points used in the CalMS21 resident-intruder assay, the per-frame spatiotemporal features consisted of 45 distances ($\mathbf{D}$) and angular change (Θ) measures, and 10 total displacements ($\mathbf{L}$). As described, the social behavior described in this data set can be reliably described based on a set of

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

intra- and inter-animal features. These features are are based on the distance, angular change, and speed of the animal's body parts in relation to itself (speed, angular change) or to another body part, including its counterpart on the conspecific. For a detailed description refer to Supp. Table 1. This process is described in Algorithm 1 and the process pipeline diagram (Fig. 2a). As for annotation selection, we identified the most common annotation in that time window (mode). In the rarest case of tie-breakers, we used the smaller values and did not observe any difference by using a different method.

Algorithm 1: Feature extraction for N pose estimates

Initialize, for $m = 1$ to $\binom{N}{2}$:
$\mathbf{L}_m \leftarrow 0$
$\Theta_m \leftarrow 0$
for $m = 1, M$ **do**
  $m \leftarrow$ any pair of pose n and $\neq n$
  Store $\left\|(n_{m1}, n_{m2})\right\|^2$ in $\mathbf{L}_m$
  for $t = 1, T-1$ **do**
    Store $\arccos\left[(\mathbf{L}_{m,t+1} \times \mathbf{L}_{m,t})/(\|/\mathbf{L}_{m,t+1}\|\|\mathbf{L}_{m,t}\|)\right]$ in Θ_m
  end
  Discard the first index of $\mathbf{L}_m$
end
Initialize, for $n = 1$ to N:
$\mathbf{D}_n \leftarrow 0$
for $n = 1, N$ **do**
  $n \leftarrow$ 2D pose estimate
  for $t = 1, T-1$ **do**
    Store $\left\|(n_{t+1}, n_t)\right\|^2$ in $\mathbf{D}_n$
  end
end
return $\boldsymbol{L}$, Θ, $\boldsymbol{D}$

Behavioral and annotation downsampling to mimic annotation scale

Adjusted mutual information score calculates the similarity in two sets of label sequences. We employed this as the metric for A-SOiD to learn the granularity in human annotation. If AMI = 1, all annotations are identical within that binned segment of 400 ms. On the other hand, if AMI = 0, there are more than 1 annotation that deviate from the rest. We have shown that employing 10*percentile*, or 400 ms, as the minimum duration threshold for CalMS21 data set yields 92% in complete target consistency throughout the bin, even without any bin overlap.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Random forest classifier for accurate and fast prediction

To create a reproducible mapping between extracted features and aligned downsampled annotations, Random forest classifier design was chosen for high-dimensional pose relationships mapping to discrete multi-class behaviors. Random forest was iteratively implemented in the 'Classify' step in A-SOiD UI. Embedded in this step is a python implementation of ensemble.RandomForestClassifier() from scikit-learn v.1.1.2 (https://github.com/scikit-learn/scikit-learn). In addition to active learning automatically balance training class sizes, the random forest initialized weights were dependent on the remaining diversity in class sizes (class weight = balanced subsample). In the first iteration, we subsampled 1% per annotation class to mimic the time budget. We then followed an autonomous active learning schedule to curate a selection of refined samples to incorporate into our training set.

Autonomous iterative active learning

Upon initializing the classifier with 1% of each annotated class, we predict the probability for all training samples. In theory, similar training samples to the initial 1% would have a high predict probability for one class over the rest. However, if there exist a sample that does not have a predict probability > 0.5 for any of the classes, we defined it as a low confidence sample. These samples appear to be very similar in high feature dimensional space, as well as the reduced dimensional space (Fig. 1 h,i bottom). In an iterative manner, we incorporate the supposed annotation aligned with these low confidence samples to create a meaningful training data set. To test the model's performance generalizability, we used the same 20% held-out test data.

Frameshift prediction paradigm

Many end users may wish to apply the algorithm to higher frame-rate video. Because A-SOiD applies a temporal constraint depending on the temporal scale of user annotations, we designed A-SOiD to predict along a sliding window. This is mathematically implemented using offsets, pseudocode in Algorithm 2.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Algorithm 2: Frameshift implementation for F times higher sampling rate than 10fps

```
Initialize behavioral array:
G ← 0
Initialize downsampled behavioral array, for f = 1 to F:
g_f ← 0
for f = 1, F do
    Start at f, sample pose-relationships at 10fps, S frames
    for s = 1, S do
        Store the prediction (g | s) in g_f
    end
    Insert g_f at every F^th position in G starting at f
end
return G
```

Identify group assignments with UMAP and HDBSCAN

A-SOiD then projects the computed pose relationships (D, , and L) into a low-dimensional space, which facilitates behavioral identification without simplifying the data complexity. In simpler terms, similar mouse multi-joint trajectory will retain its similarity visualized in the low-dimensional space. A-SOiD achieves this through UMAP, a state-of-the-art algorithm that utilizes Riemannian geometry to represent real-world data with the underlying assumptions of the algebraic topology [42]. UMAP, as previously described in Hsu et al. [9], is chosen over the popular t-SNE for its advantage in computational complexity, outlier distinction, and most importantly, preservation of longer-range pairwise distance relationships [42–45]. Embedded in this step is a python implementation of umap-learn v.0.5.3 (https://github.com/lmcinnes/umap). Since our goal is to use UMAP space for clustering, we enforced the following UMAP parameters: (n_neighbours=60, min_dist=0.0, euclidean distance metric). In terms of n_components, we call python implementation of decomposition.PCA() from scikit-learn v.1.1.2 (https://github.com/scikit-learn/scikit-learn) and set n_components to explain 0.7 of total pose-estimation variance. UMAP embeddings were then clustered through HDBSCAN algorithm [9, 46]. It is particularly useful for UMAP outlier detections as it recognizes subthreshold densities. Embedded in this step is a python implementation of hdbscan v.0.8.28 (https://github.com/scikit-learn-contrib/hdbscan). We enforced the following HDBSCAN parameters: (min_cluster_size=a range of $2-2.5\%$ of the size of the data, whichever yields the most groupings).

Motion Energy

The term "motion energy" as previously described was first introduced by Stringer et al. [41] and refers to the absolute value of the difference of con-

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

secutive frames. Since the animals are freely moving in the environment, an initial pose alignment is necessary. For this, the intruder's neck and tail base coordinates are used. Following image registration using estimated outline of both animals at the start of each identified behavior, we compute the motion energy (ME, absolute value of the difference of all consecutive frames) within a bout using the numpy [47] functions *np.diff*, *np.absolute*, and *np.mean* (for more information see). We then performed averaging for each bout to reconstruct a single ME image per annotated class. In other words, each pixel in such reconstructed ME image represents the average absolute difference between consecutive frames at a given pixel location.

CalMS21 data set

Data set

While the data set consists of three parts [28]. For our purpose the first set (Task 1, Classical Classification) is the most relevant as it contains a complete training set of pose estimation sequences (70 sessions; total of 507,738 frames) including complete annotations of all frames. A separate test set of pose estimation sequences (19 sessions; total of 262,107 frames) is being used to benchmark against the challenge winner [6].

Annotation Descriptions

The behavior annotation as described in Sun et al. 2021 was not altered ([28]). Please refer to the original publication for detailed information. Notably, the majority of annotations did not include one of the three behaviors. These widely-divergent samples were collectively annotated as "other" and are therefore not considered in evaluations regarding this benchmark data set.

Pose estimation

The provided pose estimation of the data set were extracted using the MARS [6]. MARS identifies seven user-defined body parts (snout, ears, neck, hips, and tail base). For more information refer to Sun et al. 2021 [6, 28]. During feature engineering, we discarded the left and right ear key points as they did not provide additional information about the underlying behavior.

Non-human primate data set

A single housed monkey exploring the environment for 5 minutes. Monkey's pose is generated by OpenMonkeyStudio [7].

Annotation Descriptions

Behavior annotation was done using BORIS [48] by an expert annotator based on unpublished video displaying the 3D pose data. The animal's behavior was

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

separated into distinct categories that were exclusive to one another - i.e., only a single behavior could be shown at the same time.

1. **Walk**: The monkey moved across the arena using its feet or feet and hands touching the ground (floor or platform).
2. **Jump**: The monkey jumped from the ground onto a platform, or from a platform to another, leaving the ground and remaining for a certain duration in the air. This included the moment preparing the jump and immediately after landing.
3. **Climb sidewall**: The monkey left the ground completely and moved on a sidewall of the arena using its hands and feet. This does not include moments where the monkey transfers from the ground to the sidewall to separate the behavior from rearing.
4. **Climb ceiling**: The monkey left the ground or sidewall completely and climbed on the ceiling of the arena using its hands, or hands and feet. This includes moments, when the monkey transfers from the sidewall to the ceiling and vice versa.
5. **Rearing**: The monkey touches the sidewall at any point while remaining on the ground or a platform on its feet. This includes initiating and finishing the rearing until the next behavior is identified.

References

1. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience 2018 21:9* **21,** 1281–1289. ISSN: 1546-1726. `https://www.nature.com/articles/s41593-018-0209-y` (Aug. 2018).
2. Lauer, J. *et al.* Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods 2022 19:4* **19,** 496–504. ISSN: 1548-7105. `https://www.nature.com/articles/s41592-022-01443-0` (Apr. 2022).
3. Pereira, T. D. *et al.* SLEAP: Multi-animal pose tracking. *bioRxiv,* 2020.08.31.276246. `https://doi.org/10.1101/2020.08.31.276246` (Sept. 2020).
4. Pereira, T. D. *et al.* SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods 2022 19:4* **19,** 486–495. ISSN: 1548-7105. `https://www.nature.com/articles/s41592-022-01426-1` (Apr. 2022).
5. Graving, J. M. *et al.* DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife* **8.** ISSN: 2050084X. `/pmc/articles/PMC6897514/%20/pmc/articles/PMC6897514/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6897514/` (Oct. 2019).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6. Segalin, C. *et al.* The mouse action recognition system (MARS) software pipeline for automated analysis of social behaviors in mice. *eLife* **10.** ISSN: 2050084X (Nov. 2021).

7. Bala, P. C. *et al.* Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio. *Nature Communications 2020 11:1* **11,** 1–12. ISSN: 2041-1723. https://www.nature.com/articles/s41467-020-18441-5 (Sept. 2020).

8. Ro, S. *et al.* Simple Behavioral Analysis (SimBA) – an open source toolkit for computer classification of complex social behaviors in experimental animals. *bioRxiv,* 2020.04.19.049452. https://www.biorxiv.org/content/10.1101/2020.04.19.049452v2%20https://www.biorxiv.org/content/10.1101/2020.04.19.049452v2.abstract (Apr. 2020).

9. Hsu, A. I. & Yttri, E. A. B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature Communications 2021 12:1* **12,** 1–13. ISSN: 2041-1723. https://www.nature.com/articles/s41467-021-25420-x (Aug. 2021).

10. Luxem, K. *et al.* Identifying Behavioral Structure from Deep Variational Embeddings of Animal Motion. *bioRxiv,* 2020.05.14.095430. https://www.biorxiv.org/content/10.1101/2020.05.14.095430v3%20https://www.biorxiv.org/content/10.1101/2020.05.14.095430v3.abstract (Jan. 2022).

11. Wiltschko, A. B. *et al.* Mapping Sub-Second Structure in Mouse Behavior. *Neuron* **88,** 1121–1135. ISSN: 1097-4199. https://pubmed.ncbi.nlm.nih.gov/26687221/ (2015).

12. Schweihoff, J. F. *et al.* DeepLabStream enables closed-loop behavioral experiments using deep learning-based markerless, real-time posture detection. *Communications Biology 2021 4:1* **4,** 1–11. ISSN: 2399-3642. https://www.nature.com/articles/s42003-021-01654-9 (Jan. 2021).

13. Kane, G. A., Lopes, G., Saunders, J. L., Mathis, A. & Mathis, M. W. Real-time, low-latency closed-loop feedback using markerless posture tracking. *eLife* **9,** 1–29. ISSN: 2050-084X. https://pubmed.ncbi.nlm.nih.gov/33289631/ (Dec. 2020).

14. Nourizonoz, A. *et al.* EthoLoop: automated closed-loop neuroethology in naturalistic environments. *Nature Methods 2020 17:10* **17,** 1052–1059. ISSN: 1548-7105. https://www.nature.com/articles/s41592-020-0961-2 (Sept. 2020).

15. Klibaite, U. *et al.* Deep phenotyping reveals movement phenotypes in mouse neurodevelopmental models. *Molecular Autism* **13,** 1–18. ISSN: 20402392. https://molecularautism.biomedcentral.com/articles/10.1186/s13229-022-00492-8 (Dec. 2022).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

16. Giancardo, L. *et al.* Automatic Visual Tracking and Social Behaviour Analysis with Multiple Mice. *PLOS ONE* **8,** e74557. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0074557 (Sept. 2013).

17. De Chaumont, F. *et al.* Computerized video analysis of social interactions in mice. *Nature Methods 2012 9:4* **9,** 410–417. ISSN: 1548-7105. https://www.nature.com/articles/nmeth.1924 (Mar. 2012).

18. Hong, W. *et al.* Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proceedings of the National Academy of Sciences of the United States of America* **112,** E5351–E5360. ISSN: 10916490. https://www.pnas.org/doi/abs/10.1073/pnas.1515982112 (Sept. 2015).

19. Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S. & Branson, K. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature Methods 2012 10:1* **10,** 64–67. ISSN: 1548-7105. https://www.nature.com/articles/nmeth.2281 (Jan. 2013).

20. Von Ziegler, L., Sturman, O. & Bohacek, J. Big behavior: challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacology 2020 46:1* **46,** 33–44. ISSN: 1740-634X. https://www.nature.com/articles/s41386-020-0751-7 (June 2020).

21. Goodwin, N. L., Nilsson, S. R., Choong, J. J. & Golden, S. A. Toward the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience. *Current Opinion in Neurobiology* **73,** 102544. ISSN: 0959-4388 (Apr. 2022).

22. Berman, G. J., Choi, D. M., Bialek, W. & Shaevitz, J. W. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of the Royal Society, Interface* **11.** ISSN: 1742-5662. https://pubmed.ncbi.nlm.nih.gov/25142523/ (Oct. 2014).

23. Berman, G. J. *Measuring behavior across scales* 2018.

24. Marshall, J. D. *et al.* Continuous Whole-Body 3D Kinematic Recordings across the Rodent Behavioral Repertoire. *Neuron* **109,** 420–437. ISSN: 0896-6273 (Feb. 2021).

25. Lundberg, S. M., Allen, P. G. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* **30.** https://github.com/slundberg/shap (2017).

26. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence 2020 2:1* **2,** 56–67. ISSN: 2522-5839. https://www.nature.com/articles/s42256-019-0138-9 (Jan. 2020).

27. Mathis, M. W. & Mathis, A. Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology* **60,** 1–11. ISSN: 0959-4388 (Feb. 2020).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Caltech, J. J. S. *et al.* The Multi-Agent Behavior Dataset: Mouse Dyadic Social Interactions. https://arxiv.org/abs/2104.02710v4 (Apr. 2021).

29. Karashchuk, P. *et al.* Anipose: A toolkit for robust markerless 3D pose estimation. *Cell Reports* **36,** 109730. ISSN: 2211-1247 (Sept. 2021).

30. Todd, J. G., Kain, J. S. & de Bivort, B. L. Systematic exploration of unsupervised methods for mapping behavior. *Physical Biology* **14,** 015002. ISSN: 1478-3975. https://iopscience.iop.org/article/10.1088/1478-3975/14/1/015002 (Feb. 2017).

31. Winslow, J. T. Mouse Social Recognition and Preference. *Current Protocols in Neuroscience* **22,** 1–8. ISSN: 1934-8576. https://onlinelibrary.wiley.com/doi/full/10.1002/0471142301.ns0816s22%20https://onlinelibrary.wiley.com/doi/abs/10.1002/0471142301.ns0816s22%20https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/0471142301.ns0816s22 (Jan. 2003).

32. Yang, M., Loureiro, D., Kalikhman, D. & Crawley, J. N. Male mice emit distinct ultrasonic vocalizations when the female leaves the social interaction arena. *Frontiers in Behavioral Neuroscience* **0,** 159. ISSN: 16625153 (Nov. 2013).

33. Sturman, O. *et al.* Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. *Neuropsychopharmacology 2020 45:11* **45,** 1942–1952. ISSN: 1740-634X. https://www.nature.com/articles/s41386-020-0776-y (July 2020).

34. Sáez, J. A., Luengo, J., Stefanowski, J. & Herrera, F. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences* **291,** 184–203. ISSN: 0020-0255 (Jan. 2015).

35. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. & Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* **42,** 463–484. ISSN: 10946977 (July 2012).

36. He, H., Bai, Y., Garcia, E. A. & Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks,* 1322–1328 (2008).

37. Błaszczyński, J. & Stefanowski, J. Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing* **150,** 529–542. ISSN: 0925-2312 (Feb. 2015).

38. Cieslak, D. A., Hoens, T. R., Chawla, N. V. & Kegelmeyer, W. P. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery 2011 24:1* **24,** 136–158. ISSN: 1573-756X. https://link.springer.com/article/10.1007/s10618-011-0222-1 (June 2011).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.04.515138; this version posted November 4, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

39. Yu, H. *et al.* ODOC-ELM: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data. *Knowledge-Based Systems* **92,** 55–70. ISSN: 0950-7051 (Jan. 2016).

40. Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* **5,** 221–232. ISSN: 21926360. https://link.springer.com/article/10.1007/s13748-016-0094-0 (Nov. 2016).

41. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brain-wide activity. *Science* **364.** ISSN: 10959203. https://www.science.org/doi/10.1126/science.aav7893 (Apr. 2019).

42. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://arxiv.org/abs/1802.03426v3 (Feb. 2018).

43. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology 2018 37:1* **37,** 38–44. ISSN: 1546-1696. https://www.nature.com/articles/nbt.4314 (Dec. 2018).

44. Packer, J. S. *et al.* A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science (New York, N.Y.)* **365.** ISSN: 1095-9203. https://pubmed.ncbi.nlm.nih.gov/31488706/ (Sept. 2019).

45. Van Unen, V. *et al.* Mass Cytometry of the Human Mucosal Immune System Identifies Tissue- and Disease-Associated Immune Subsets. *Immunity* **44,** 1227–1239. ISSN: 1097-4180. https://pubmed.ncbi.nlm.nih.gov/27178470/ (May 2016).

46. Campello, R. J., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **7819 LNAI,** 160–172. ISSN: 03029743. https://link.springer.com/chapter/10.1007/978-3-642-37456-2_14 (2013).

47. Harris, C. R. *et al.* Array programming with NumPy. *Nature 2020 585:7825* **585,** 357–362. ISSN: 1476-4687. https://www.nature.com/articles/s41586-020-2649-2 (Sept. 2020).

48. Friard, O. & Gamba, M. BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution* **7,** 1325–1330. ISSN: 2041-210X. https://onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12584%20https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12584%20https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.12584 (Nov. 2016).