# Characterizing Behavioral Dynamics in Bipolar Disorder with Computational Ethology

**Authors:** Zhanqi Zhang[1], Chi K. Chou[2], Holden Rosberg[3], William Perry[3], Jared W Young[3], Arpi Minassian[3], Gal Mishne[4,5,#], Mikio Aoi[4,6,#]

1. Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA
2. Department of Mathematics, University of California San Diego, La Jolla, CA
3. Department of Psychiatry, University of California San Diego, La Jolla, CA
4. Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA
5. Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA
6. Department of Neurobiology, University of California San Diego, La Jolla, CA

# These authors contributed equally.
Correspondence can be addressed to gmishne@ucsd.edu and maoi@ucsd.edu.

## Abstract

New technologies for the quantification of behavior have revolutionized animal studies in social, cognitive, and pharmacological neurosciences. However, comparable studies in understanding human behavior, especially in psychiatry, are lacking. In this study, we utilized data-driven machine learning to analyze natural, spontaneous open-field human behaviors from people with euthymic bipolar disorder (BD) and non-BD participants. Our computational paradigm identified representations of distinct sets of actions (*motifs*) that capture the physical activities of both groups of participants. We propose novel measures for quantifying dynamics, variability, and stereotypy in BD behaviors. These fine-grained behavioral features reflect patterns of cognitive functions of BD and better predict BD compared with traditional ethological and psychiatric measures and action recognition approaches. This research represents a significant computational advancement in human ethology, enabling the quantification of complex behaviors in real-world conditions and opening new avenues for characterizing neuropsychiatric conditions from behavior.

## Main

Behavior, particularly in novel contexts, can be highly informative about neuropsychiatric conditions and illness states. For example, open field studies, which observe individuals in unstructured environments, can provide unique insights into how different conditions manifest in real-world settings. Bipolar disorder (BD), a chronic psychiatric illness that can have devastating functional consequences, is hallmarked by increased energy, which often manifests as more motor activity and engagement in goal-directed behaviors[1]. Quantifying such behavior is critical to identify symptoms, formulate diagnoses, and ultimately advance treatment approaches. Contemporary machine learning can automate this process to identify signature behavior patterns that potentially reflect underlying brain functions of conditions such as BD and other neuropsychiatric illnesses.

Currently, to assess the underlying psychiatric disorders, clinicians heavily rely upon observer-rating scales such as the Hamilton Depression Rating Scale (HAM-D)[2,3], Young Mania Rating Scale (YMRS)[4] and other self-reported rating scales[5]. However, self-reported rating scales have limitations in reliability. Rating scales can address broad classifications but may fail to accurately address fine motor skills and behaviors or effectively differentiate between conditions. For example, 'Increased Motor Activity-Energy' in YMRS may represent a group of symptoms that are present in conditions other than BD (such as ADHD). These scales aggregate multiple experiences over various timeframes and milieus — such as work, home, and leisure activities — which may not best represent real-time behavior. Additionally, these rating scales reduce complex, high-dimensional experiences into integer ranges from severe to mild, where the relative magnitude between ranges can vary inconsistently (e.g., the difference between 0 and 1 is not necessarily equivalent to the difference between from 1 and 2). Therefore, quantification of behavior on a continuous scale would be preferable for more accurate assessments.

An additional concern is that psychiatric conditions often manifest symptoms cyclically and extend over timescales[6], such that individuals with BD can exhibit distinctive patterns of behavior depending on their illness state[7]. While people with BD experiencing manic episodes have high motor activity, the activity of those in a euthymic state, defined by the absence of a manic, hypomanic, or depressed episode, may appear indistinguishable from that of a healthy person. Moreover, due to inter-individual differences in pathology, the idiosyncrasies of each individual’s life history, and the time-varying nature of mental health and psychiatric disorders, two patients even when experiencing the same BD episode may not present in precisely the same way. This difference means that population averages may not reflect the best possible assessment of a given individual[8,9]. Therefore, it remains a challenge to identify and quantify the subtle behavioral features among individuals with BD until they present with prominent manic or depressive symptoms, at which point the opportunity for preventative intervention has been missed.

There have been some recent inroads in the quantification of undirected human behavior in medical settings. The human Behavioral Pattern Monitor (hBPM), a human version of the classic rodent open-field activity assessment, was develorped to better quantify human exploratory behavior[10]. hBPM uses spatial information (for example, Spatial-D) and temporal statistics to identify signature patterns of behavior of human patients[10,11]. However, the hBPM still relies on observers to label behavior using *a priori* established criteria. This time-consuming process is susceptible to subjective biases in behavioral labels and can be undermined by insufficient inter-rater reliability. Moreover, manual observer-based methods face challenges in scaling to the extensive sizes of modern datasets. To overcome these limitations and discover relevant behavior repertoire in an exploratory manner, data-driven behavioral identification is needed.

Behavior as a reflection of cognition often displays repeated patterns, i.e., behavioral *motifs*. *Motifs* are recurring, identifiable sequences of actions, reactions, or responses, exhibited as a characteristic feature of a population. *Motifs* are often considered meaningful units of behavior that may provide insights into underlying psychological or physiological processes[12–14]. *Motifs* also appear in rating scales, described as specific actions. For example, the HAM-D describes “agitation” based on a collection of actions (i.e., *fidgetiness*; *playing with hands, hair, etc.*; *moving about*, *can’t sit still*, *hand wringing*, *nail-biting*, *hair-pulling*, *biting of lips*). These subtle *motifs* usually do not belong to a generic label and are ignored during manual behavior annotation. This raises a question: can we automatically identify *motifs* from free-moving spontaneous human behavior in a rich real-world context?

Progress towards this direction has been made in animal models, where automated behavioral segmentation methods (e.g., MoSeq-based models[15,16], VAME[17], MotionMapper[18–20], and B-SOiD[21]) have proven useful for identifying stereotyped behavioral *motifs* that can be related to neurological[19] and pharmacological manipulations[14] in animals. However, there is little research applying such methods to understanding human behavior, let alone in a psychiatric context. In recent years, computer vision-based supervised methods of animal- (e.g., DeepLabCut[22], DeepPoseKit[23], Deep Graph Pose[24], DeepOF[25], and SLEAP[26]) and human-pose estimation (e.g., MoveNet[27] and OpenMMLab[28]) can produce accurate key points tracking and skeleton estimates of animal or human participants and can even automatically label actions. Built on deep-learning-based architectures, these models have significantly increased the efficiency of behavioral quantification with little to no direct human supervision. However, these methods are limited by their training sets of gait movements, which are often constrained to not only a small subset of camera angles, lens distortions, and action labels, but also a narrow scope of human behaviors. Thus, pose estimation models alone cannot identify distinct behavioral *motifs*, making them relatively impoverished descriptions of behavior for clinical settings.

Our objective was to quantify spontaneous human behavior in real-world contexts among euthymic BD individuals and differentiate them from a healthy comparison (HC) population. We aimed to use an “unsupervised” machine learning model (meaning a model that is not explicitly told how to structure the relationships between data points) to objectively characterize patterns of behavior without relying on a predetermined catalog of behaviors. Here, we introduce a novel approach to address these challenges. Specifically, we identified recognized behavioral features of BD that aligned with previously known clinical observations and were uniquely expressed in our analysis. Our machine learning framework also consistently identified patterns and relationships that may not be immediately obvious to human observers. By exploring new behavioral features and providing psychiatric interpretations of these features, our approach shows the potential to lead discoveries in the field to better understand symptoms, formulate diagnoses of psychiatric disorders, and ultimately advance treatment approaches.

## Results

Study participants have been described previously in hBPM studies[29]. Briefly, 25 participants (12 men) were diagnosed with bipolar disorder (BD). Twenty-four were diagnosed with BD Type I or Type II, and one participant was diagnosed with the cyclothymic subtype of BD. All diagnoses were determined by the Structured Clinical Interview for DSM-IV[30]. All BD participants were in a euthymic state as defined by scores of HAM-D < 10 and YMRS < 12 (**Supplementary Table 1**). Healthy comparison (HC) volunteers (n = 25; 15 men) who had never met the DSM-IV criteria for neurological or psychiatric disorders participated in the study as the HC group. All participants gave written consent and were assessed by the YMRS (to assess symptoms of mania) and HAM-D (to assess symptoms of depression). Higher scores on the measures reflect more severe symptoms of mania or depression. Each participant was introduced to a previously unexplored room containing furniture and small objects along the periphery of the room (**Supplementary Fig. 1**) and remained there for 15 minutes. Videos were recorded from a commercial camera with a fisheye lens placed at the center of the ceiling (**Fig. 1a**). For full details, please refer to **Methods**.

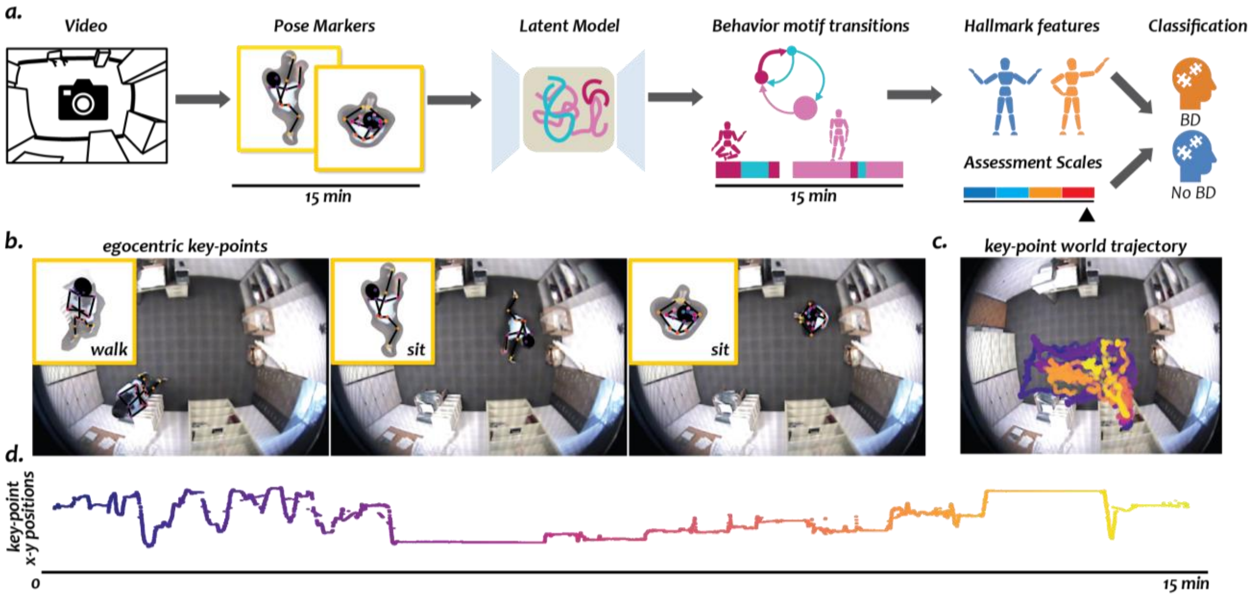

**Figure 1. Data and Methods.** a. Videos of free-moving human behavior from participants with bipolar disorder (BD) during euthymic episodes and healthy comparison (HC) participants for 15 minutes in an unexplored room with objects. We utilized DeepLabCut to label 20 markers placed on key-points of human participants (e.g., elbows). Pose markers were fed into a latent-variable model and the latent representations were used to segment the videos into *motif*s. We identified hallmark behavioral features that characterized BD in different time scales and these features were used to classify if a participant is from the BD or HC groups. Classification was benchmarked against assessment scales YMRS and HAMD and other action segmentation approaches. b. Three example frames from the videos of human behavior with key-points marking the skeleton. Inset: Egocentric view of the human skeleton with key-points are shown with action label from manual behavior annotations. c. Example of center-of-feet key-point x-position trajectory in the room. d. Trajectory of the center-of-feet key-point x-position over time.

### A Latent-variable model identified context-dependent behavioral motifs of human participants.

While the full repertoire of human behaviors is vast, we expect the distribution of behaviors a person expresses in a given context to be highly constrained and specific. We, therefore, sought to best characterize the distribution of behaviors relevant to the context of our experiment, rather than a predetermined catalog of behaviors that may not be as well matched. To characterize patterns of context-dependent, naturalistic human behaviors, we required an unbiased way of annotating our video data. We, therefore, developed a data-driven approach for discovering behavioral features of freely-moving humans with two key functional modules: (1) pose estimation (using DeepLabCut) for accurately labeling anatomical key points of the human participants in every frame[22] (**Fig. 1b-d**), and (2) a latent-variable model (VAME) for embedding these key points into a low-dimensional representation[17] (**Fig. 2a, b**). Clustering on the latent representation provided a set of behavioral *motifs* corresponding to distinct actions or sequences of actions (**Fig. 2c, d**). We compared our approach to manually annotated labels determined by clinically trained human experts; as well as pre-trained computer vision (CV) action detection models[28,31], which automatically generated a set of labels (**Supplementary Fig. 2a, b**). As an additional control, we applied k-means clustering to the key points themselves (rather than the latent coordinates) to obtain an alternative set of clusters.

We found the distribution of manually labeled behaviors was imbalanced — among 50 videos, the vast majority of time frames are labeled as “stand” or “walk” (median(IQR) BD: 65.2%(34.7%), 17.9%(23.1%); HC: 77.3%(55.3%) 7.9%(12.2%), **Fig. 2e**). For the CV models, while they have access to up to 400 available action labels[32], most labels were irrelevant to the clinical setting, such as “canoeing or kayaking,” “changing wheel”, and “playing musical instrument”. We therefore found that the majority of the identified actions among CV models were only distributed among a few labels. For example, MMAction[28] identified “stand,” “sit” and “lie/sleep” (median (IQR) BD: 55.56% (40.00%), 17.11% (20.89%), 7.11% (7.11%); HC: 42.44% (25.11%), 17.33% (13.55%), 11.33% (13.99%)). Most concerning was that the top three actions detected by S3D[31] were erroneously identified as “biking through snow,” “folding napkins,” and “folding clothes” (median(IQR): BD: 28.81% (22.27%), 17.17% (23.64%), 13.37% (42.74%); HC: 42.74% (37.16%), 24.55% (30.71%), 10.64% (15.06%)).

In contrast, the *motifs* obtained from the latent-variable model captured a broad array of interpretable behaviors in the clinical context. Clips from the same *motif* showed visually similar combinations of actions and activities. Interestingly, our *motifs* spanned multiple time scales, varying from a few seconds to a couple of minutes, indicating diverse scales of complexity in behavioral dynamics and underlying cognitive processes[33]. To accurately quantify these nuances observed in human behavior, each *motif* clip was described using natural language, instead of discrete labels employing single verbs (**Methods**). While some *motifs* represented intuitively simple activities (e.g., *standstill*), the majority of *motifs* captured higher-order behavioral sequences that reveal previously undefined actions, even behavioral intentions. For example, *motif 1* included a collection of clips related to the *stretch of one body part*, such as *upper body bend*, *arm swing,* and *wrist/ankle rotation*. *Motif 4* revealed *fidget*, meaning small movements in hands and feet, such as *nose picking*. In addition, *motif 9* showed an active exploratory behavior, in which participants *approached objects and then inspected them*, but did not necessarily directly interact with objects as in *motif 8*. Notably, *motif 9* is an intentional exploration, i.e. the subject typically had a targeted object or a destination in mind after scanning around the environment, as opposed to the *aimless wander* in *motif 6* and the *depart after exploration* in *motif 2*. **Table 1** includes the actions in all *motifs*.

**Table 1**
*Motif descriptions in natural language*

| Motif | Description | Examples |
|---|---|---|
| motif 0 | *torso rotation* | *turn walking direction,*<br>*lean left and right,*<br>*bend forward* |
| motif 1 | *stretch (one body part)* | *upper body bend,*<br>*wrist/ankle rotation,*<br>*arm swing* |
| motif 2 | *depart (from the previous action)* | *step away from the window,*<br>*walk away from the desk,*<br>*turn away from the bulletin board* |
| motif 3 | *arm and hand movement* | *touch clothes,*<br>*pull open drawers,*<br>*reach objects* |
| motif 4 | *static or fidget* | *pick nose,*<br>*remove the candy wrapper,*<br>*detangle and braid hair* |
| motif 5 | *standstill* | *standstill by the bulletin board,*<br>*standstill in the middle of the room,*<br>*standstill by window* |
| motif 6 | *wander and scan (aimlessly)* | *wander towards the bookcase,*<br>*scan across the room,*<br>*look at the cradle swing* |
| motif 7 | *turn from/to* | *turnaround from the window,*<br>*step back and turn,*<br>*turn head left and right* |
| motif 8 | *examine/interact with objects* | *look at the desk,*<br>*reach objects on the bookcase,*<br>*wear clothes placed on the bookcase* |
| motif 9 | *approach (with aim) and/or inspect* | *approach the bookcase and inspect it,*<br>*go to the door and peek,*<br>*read from the bulletin board* |

The timing and duration of motif occurrences were similar to those of manually annotated labels. For example, as we divided the video into three 5-minute epochs, both approaches showed many behavior occurrences in epoch 1, and few occurrences in epoch 3 (**Fig. 2c)**. Although there is not a one-to-one correspondence between manually annotated labels and learned *motifs*, 87.10% of the onset and offset of *motifs* align with those of manually annotated labels (**Methods**). *Motifs* displayed a more fine-grained and broader distribution of behavior compared with manually annotated labels. For periods where there is only one human annotated label like “stand,” the latent-variable model has revealed more fine-grained motifs such as *tucking shirts using hands while standing*. This demonstrates that the latent-variable model not only captured the actions that are explicitly perceivable by the eye but also identified finer categories of actions that are data-dependent.

## Motif dwell times suggest perseveration and impairment of attention in BD.

Our *motifs* produced relevant representations of the human pose for understanding the behavioral characteristics of the euthymic state of BD. People with BD are considered in a euthymic state when they do not meet the criteria for a manic, hypomanic, or depressed episode although they may still exhibit some symptoms. We were interested in whether we could identify distinct behavioral features of euthymic BD patients that distinguished them from HCs, even in the absence of a depressive or manic episode.

To this end, we measured the average *motif* usage dwell time, which is the time spent in each *motif*, for BD and HC during the entire recording period (**Fig. 2e**). Previous work on the hBPM has shown that manic BD patients displayed high motor activity in the first epoch, but quickly attenuated in the second and third epochs[11]. Consistent with this setting, we also calculated the mean dwell time of each *motif* in the three 5-minute epochs.

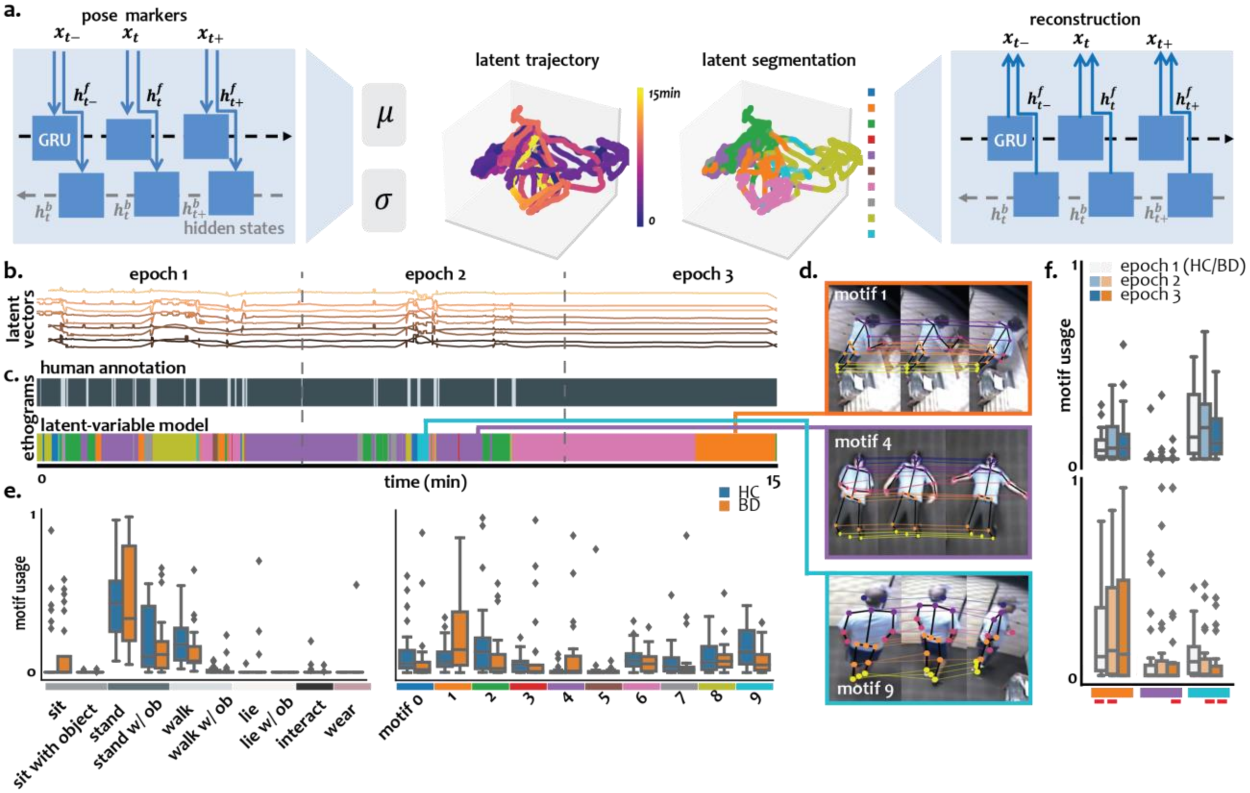

**Figure 2. Latent-variable Model and Dwell time.** a. Pose markers were fed into the VAME variational autoencoder and the latent representations were used to segment motifs. The input were the past $x_{t-}$, current $x_t$, and next $x_{t+}$ pose markers time series which were encoded as corresponding hidden states. The model would learn to reconstruct the input, and the learned latent representation was a 15-min vector that were segmented into *motifs*. b. Example of latent vectors for video in Fig. 1b. c. Top*:* Each video was manually annotated by experts into 10 behavior categories (e.g., sit, stand). Ethogram of manual annotation. Bottom*:* Ethograms of motif segmentation from latent segmentation*.* d. examples of *motif 1*, *motif 4* and *motif 9* in the dataset. e. Motif usage dwell time from human annotation (left) and latent variable model (right) in BD (orange) and HC (blue). f. Motif dwell time for *motif 1*, *motif 4* and *motif 9* in three epochs in BD (light to dark shades of orange), and HC (light to dark shades of blue). Red bars on the x-axis indicates significance.

We detected differences between BD and HC in overall dwell time for *motif 1* (*stretch of one body part*), *motif 4* (*static or fidget*), and *motif 9* (*approach objects then inspect them*) (two-sample t-test p-value: 0.010, 0.027, 0.015). Furthermore, dwell time in *motif 9* was positively correlated with HAM-D (Pearson Correlation r: 0.44, p-value: 0.03), and dwell time in *motif 2* (*depart*) was positively correlated with YMRS (Pearson Correlation: r: 0.53, p-value: 0.01) in the BD group.

For clusters obtained by k-means clustering of the key point trajectories, cluster 4 and cluster 6 displayed differences between the populations (two-sample t-test, p-value: 0.033, 0.007) but these were not correlated with assessment scales. Cluster 2 demonstrated no difference in dwell time but was correlated with higher YMRS scores in the BD group (Pearson Correlation: r: 0.44, p-value: 0.03). In contrast, for manually annotated and CV-identified actions, dwell times associated with their labels either did not distinguish between the populations or were different between populations but did not correlate with assessment scales (**Supplementary Table 2**).

The dwell time of *motifs* varied between epochs. We found the dwell time of *motif 1* was higher in the BD population in the first and second epochs (two-sample t-test, p-value: 0.04, 0.026), higher in *motif 4* in the third epoch (two-sample t-test, p-value: 0.047), lower in BD in *motif 9* in the second and third epochs (**Fig. 2f**, two-sample t-test, p-value: 0.026, 0.044). We found *motif 9* became more correlated with HAM-D (Pearson correlation r in epoch 1 to epoch 3: -0.02, 0.38, 0.61, p-value: 0.93, 0.06, 0.00) but not with YMRS. *Motif 2* was correlated with YMRS in the second epoch (Pearson Correlation: r: 0.52, p-value: 0.01).

For k-means clustering of the key points, cluster 2 showed a correlation with YMRS in the first two epochs (Pearson Correlation: r: 0.42, 0.45, p-value: 0.04, 0.02). Cluster 4 showed a difference in dwell time in epoch 2 (two-sample t-test, p-value: 0.015), and cluster 6 showed a difference in all epochs (two-sample t-test, p-value: 0.015,0.012, 0.016), but no correlation with either HAM-D or YMRS. For the manually annotated categories, no difference was found in dwell time, but “stand” time was negatively correlated with HAM-D in the first epoch (Pearson Correlation: r: -0.47, p-value: 0.02), and “sit” time was correlated with HAM-D in the last epoch in the BD population (Pearson Correlation: r: 0.42, p-value: 0.04).

To compare the describing power on the distribution of behaviors, we introduced a measure of motif entropy. Specifically, the entropy of the dwell time distributions is the highest for our method (**Supplementary Fig. 2c**). Lower entropy dwell time distributions suggest a model mismatch, as they indicate that most of the probability mass is allocated to a small number of motifs. An ideal fit, according to the principle of maximum entropy, should have a uniform dwell time distribution.

Overall, we found BD had increased time stretching, fidgeting, and less time in interaction with objects, indicating potential perseveration and impairment of attention[34,35]. In summary, *motifs* identified by our data-driven machine learning approach showed stronger and more consistent correlations with clinical assessments than either general-purpose annotation methods or more traditional manual annotations.

### Motif transitions displayed less activation and more stereotypy in BD.

The behavioral dynamics, as measured by the transition frequency between *motifs,* and the variety of the behavioral repertoire, changed as the participants spent more time in the environment. Specifically, visual inspection of ethograms highlighted periods during which participants frequently transitioned between *motifs*, indicating a richer and more diverse behavioral repertoire, in contrast to periods where participants remained consistently within a single *motif*, or a small subset of *motifs*. To quantify these fluctuations in behavioral transitions and their variety, we can view *motifs* as *states* within a Markov Chain and quantify the temporal relationships between them.

We computed the weighted adjacency matrices $A$, and transition probability matrices $P$ separately for each participant to capture the dynamics between *motifs* (**Fig. 3a, b**). Adjacency matrices $A$ tally how often every *motif* $S_i$ transitions to every other *motif* $S_j$, where $j \neq i$. The sum of all entries in the adjacency matrix, $\sum_{i,j} A_{ij}$, provides the transition frequency, and the overall number of transitions during the period of interest. Transition matrices $P$ assess the rate of transitions between *motifs* by calculating the probability of every motif $S_i$ transitioning into every other *motif* $S_j$. We computed $A_T$ and $P_T$ for the entire duration of the recording $T$, as well as $A_\tau$ and $P_\tau$ at each epoch $\tau$. These measurements enable us to quantify how frequently individuals shift between different *motifs* and the likelihood of such transitions occurring. As a control, we computed $A$ and $P$ for setting the latent variable model to identify either $n = 10$ or $30$ *motifs* to explore the impact of the number of *motifs* on transition dynamics.

While both BD and HC groups experience an overall decrease in transition frequency, the decline is more pronounced in BD over time (**Fig. 3c**, linear regression fitting over three epochs: BD: slope: -0.06, p-value: 9.80 x $10^{-4}$, SE: 0.02; HC: slope: -0.01, p-value: 0.57, SE: 0.02). This indicates that the behavioral repertoire within the BD group becomes narrower and more stereotyped over time. Note that there is a distinction between a narrower range in behavioral repertoire and true inactivity (i.e., no change in key point positions): a decrease in transition frequency does not necessarily indicate inactivity; instead, it signifies an increase in stereotypy of behavioral patterns. For example, the increase in stereotypy reflected as $P_\tau$ became sparser (more zeros) in BD, in comparison to *idiosyncrasy* which was reflected as a consistent number of zeros in $P_\tau$ of HC.

To quantify *stereotypy*, we introduced the *effective-number-of-accessible-states* (*ENAS*) of the transition matrix. *ENAS* is a measure of the number of accessible *motifs* (states) for each period (overall time, or epoch) by weighting the count of *motifs* by their relative accessibility (probability). Intuitively, given a *motif* that the participant occupied within the period, if every

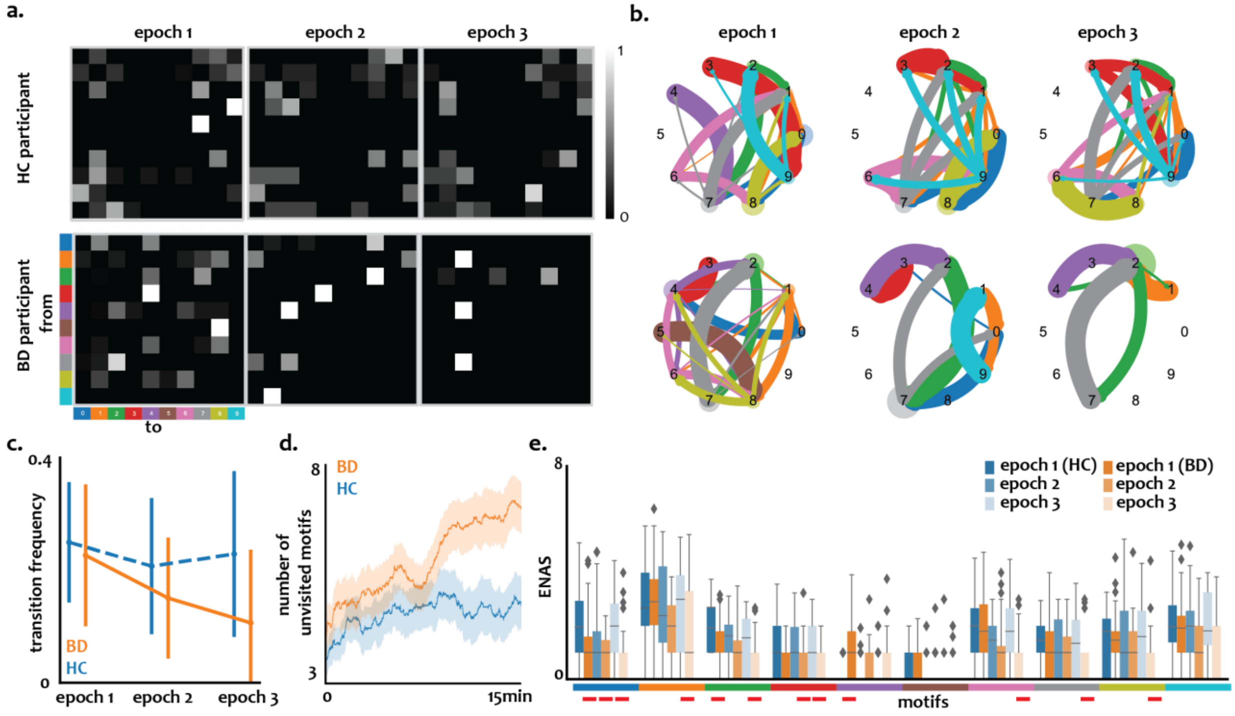

**Figure 3. Motif Transition.** a. Transition matrices in three epochs for an HC participant and a BD participant, where each pixel represents the transition probability from every *motif* into every other *motif*. b. Graphs representing the transition matrices in a. where nodes represent motifs and directed edges are colored by the 'from' *motif* color. The thicker the edges the higher transition probability. The larger the nodes the higher dwell time of the *motif*. c. Transition frequency of three epochs in HC (blue) and BD (orange). d. Number of unvisited *motifs* of the HC (blue) and BD (orange) population over time. e. Effective-number-of-accessible-states (ENAS) of three epochs of HC (blue) and BD (orange) of ten motifs. Epoch 1 – epoch 3 marked by dark to light shades in each population. Significance marked by red bars.

other *motif* $i$ is visited equally from this *motif*, *ENAS* of this *motif* is equal to $n$; if no other *motif* is visited, *ENAS* is equal to 1; if the *motif* was not occupied during the period, then the *ENAS* is 0.

We counted the number of unvisited motifs in the transition matrices to quantify sparsity, i.e. whether or not the behavior was dominated by only a few stereotypical transitions between *motifs*. We found the number of unvisited *motifs* became higher in BD than in HC (**Fig. 3d**). In addition, *ENAS* became smaller for BD over time in all *motifs* and often was smaller compared with HC, especially in epoch 3. This indicated that BD participants tended to not only display a smaller behavior repertoire, but also had fewer accessible *motifs* over time in this repertoire (**Fig. 3e**).

We experiment with denser motif segmentations $(n = 30)$ and observed BD to also have a decrease in motif transitions (**Supplementary Fig. 3**), suggesting that an increase in stereotypy over time are hallmark of BD, independent of the set of actions, or the complexity of actions chosen in the given environment. Moreover, our analysis of transition provides a quantification on the level of dynamic characteristics of *activation*, an important dimension of BD that is associated with many terms including *arousal*, *excitation*, *novelty seeking*, *agitation*[36]. Together, we provide quantifications on behavioral dynamics and these results suggest that the behavior of the BD population tends to become more stereotyped, and less in *activation* during the course of recording, even in euthymic episodes.

### Latent representations displayed behavioral variability in BD.

Transition analysis explored the temporal relationships between *motifs*, shedding light on their sequences but not on the diversity of actions occurring within specific *motifs*. For example, in *motif 1*, one participant may stretch by *rolling their arms*, while another may *kick their legs*. To examine within-motif variability, we measured *motif-volume*. Actions expressed similarly in physical space are represented by trajectories nearby in the latent space. Therefore, the variability observed in movements is reflected in the variability of the latent variables. *Motif-volume* $v_i(\tau)$ is computed as the total variance of the latent representation of motif $i$ at time $\tau$ (**Fig. 4a, b, Methods**). A larger *motif-volume* indicates greater variability of motif expression in the population, whereas a smaller *motif-volume* suggests a more uniform motif expression among the same groups of participants.

We observed BD *motif-volume* was consistently lower than HC *motif-volume* in *motifs 0* and *2* (two sample t-test p-value of epoch 1-3, *motif 0*:0.009, 0.006, 0.011, *motif 2*: 0.709, 0.094, 0.011), and consistently higher than HC in *motifs 4* and *5* (two sample t-test p-value of epoch 1-3, *motif 4*:0.004, 0.234, 0.061, *motif 5*: 0.080, 0.917, 0.356, **Supplementary Fig. 4a, b**). However, motif volume in BD was not significantly different from HC in the first epoch but was lower than HC in the second and third epochs in *motifs 2, 3, 6, 7, 8,* and *9* (two sample t-test p-value of epoch 3, 0.011, 0.031, 0.002, 0.042, 0.025, 0.001)*.* Notably, *motif-volume* is not necessarily correlated with dwell time (**Supplementary Table 4**), indicating that volume is not merely a consequence of more time spent in a given *motif*.

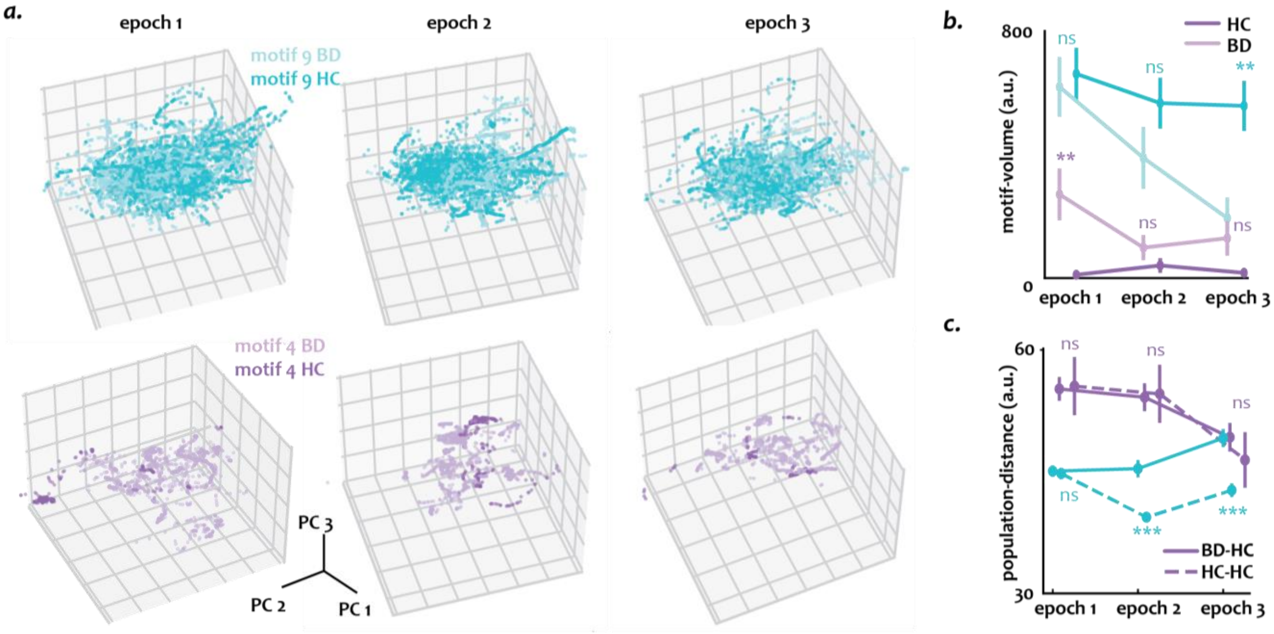

**Figure 4. Latent Shifting of Motif Representation.** a. *motif 9* and *motif 4* of BD (lighter shades) and HC (darker shades) latent vector in three epochs represented in the top three PC. Latent vectors were shuffled in index and subsampled for visualization. b. *Motif-volume* over time for *motif 9* and *motif 4* in BD (lighter shades) and HC (darker shades) population. c. Interpopulation-distance between BD and HC (solid lines) in epoch 1, epoch 2, and epoch 3. As control, intrapopulation-distance of HC (dashed lines) were shown. Significance were marked by asterisks.

To quantify within-motif variability between populations over time, we computed the *interpopulation distance* between BD and HC latent representations of each *motif $i$* in each epoch. As a control, we computed the *intrapopulation distance* within BD and within HC in each epoch (**Fig. 4c, Supplementary Fig. 4c**). If latent representations are getting more dissimilar between BD and HC, then the interpopulation distance would increase and the volumes representing the *motif* for both populations would overlap less. We found the *interpopulation distance* consistently increased in *motifs 1,6,9* from epoch 1 to epoch 3, decreased in *motif 4*, decreased and then increased in *motifs 0,3,5,7*, and increased then decreased in *motifs 2, and 8.* In addition, the *interpopulation distance* is higher than the *intrapopulation distance* in *motifs 1, and 9* in the last epoch, indicating the expressions of these motifs in terms of specific actions and movements for BD and HC become more distinct over time (2 sample t-test p-value epoch 1-3: *motif 1*: 0.49, 0.39 5.54 x $10^{-7}$; *motif 9*: 0.76, 3.36 x $10^{-7}$, 1.31 x $10^{-5}$). Together, these findings not only highlight the progressive divergence between BD and HC but also suggest that BD may be associated with the development of more stereotypical and more distinct behavior, which provides a potential avenue for monitoring disease progression.

## Behavioral features from the latent space better discriminate BDs from HCs than traditional measurements.

The behavioral features we derived from the segmented latent representations of actions are consistent with the phenotype of increased activity and energy, which is a hallmark feature of BD. These features arguably encompassed a less biased set of behavioral markers of BD compared to CV models, expert human annotation, and even established clinical assessment scales as they were discovered from spontaneous human behavior in real-world contexts, rather than pre-defined catalogs of behaviors. We thus hypothesized that the identified behavioral features would better distinguish euthymic BD participants from HCs, than alternative methods. To test this hypothesis, we first performed feature selection in our framework among assessment scales (HAM-D and YMRS) and our behavioral features. We found the most predictive features of BD are difference of behavioral features between epochs 3 and 1 (**Supplementary Table 3**). Since our framework, human annotation, and CV-based models all provide a way of segmenting the behaviors, we can compute behavioral features except for the latent representations (such as motif dwell time, *ENAS*, zeros in transition matrix, and latent volume) from all models. The selected features were used in a logistic regression model for classification. The dataset was randomly split among participants into training and validation sets. The average accuracy, recall, and precision were calculated with 3-fold cross-validation. As controls, we benchmarked the classifier on (1) assessment scales that encompassed a range of psychometric measures, (2) behavioral features identified by human annotations, or (3) CV models.

**Table 2**

*Classification accuracy of BD vs HC across approaches*

| model | pretrained | accuracy (mean ± std) |
|---|---|---|
| Assessment Scales | - | 0.53 (0.11) |
| Spatial-D | - | 0.53 (0.13) |
| K-means on DLC | - | 0.70 (0.12) |
| hBPM video ratings | - | 0.65 (0.12) |
| S3D | Kinetics-400 | 0.70 (0.13) |
| MMAction2 | Kinetics-400 | 0.65 (0.13) |
| **Ours** | - | **0.75* (0.11)** |

**Table 2** shows the cross-validated classification accuracy using selected input features. We found that the classification accuracy using our behavioral features outperformed human annotation, CV models, and clinical assessment scales (Tukey HSD p-value Ours vs other approaches in Table 2 order all $< 0.001$). Our results underscore the potential of data-driven identified behavioral motifs to effectively differentiate BD from HC.

## Discussion

Current data-driven machine learning techniques offer significant improvements over traditional observational methods across a wide range of domains, as the latter methods are prone to bias. Our study demonstrates that an “unsupervised” machine learning model, which does not rely on hundreds of person-hours of data annotation, can assist in clinical characterization. By integrating computer vision, deep learning, and probabilistic reasoning to study activation in BD, we present a novel approach to better understand subtle behavior patterns in individuals under clinical context.

Our model automatically identifies patterns in the data relevant to our participants and the specific conditions of our experiment, rather than adhering to traditional characterizations of mental disorders. We demonstrate several advantages of our approach. Firstly, human video annotation is time-consuming, as it not only requires extensive training and practice, but also assessment of the validity and reliability of the annotator. Our method surpasses human annotation by more accurately describing the dwell time distribution of behaviors, as measured by motif entropy.

Through an “end-to-end” design, we are able to validate our model by evaluating it in a BD vs non-BD classification task that was downstream from the learning of the latent states. Our approach exhibits superior performance when benchmarking against traditional approaches for diagnostics. This result not only suggests the behavioral features (*motif* quantification, transition dynamics, and latent representations) could be robust metrics for evaluating patient behavior in euthymic BD, but also implies that a more precise representation of the psychopathology of the participants has been learned by the model, and can be used in various downstream tasks that could offer valuable insights for clinical assessment and treatment planning. In addition, although a sample of people with BD was used here to develop and validate our methods, our general approach is agnostic to patient diagnosis and environmental setting and is modular by design.

Central to our methodology is analyzing various features downstream of the latent variable representations of motifs, including dwell time, motif transitions, and variability of latent representations. Our approach identifies clinically meaningful *motifs* that may reflect aspects of the condition that are not easily perceptible to human observers. For example, people with BD display shorter dwell times for motif *approached some objects and inspected them*, potentially reflecting impairment in attention span, set shifting, and task switching[34]. This observation aligned with previous studies where euthymic BD patients were observed to perform worse than controls on the digit subtest (Wechsler Adult Intelligence Scale) attention task[37,38], and may reflect impulsive reward-seeking behavior, a characteristic feature of BD[39]. As another example, the observation of fidgeting movements, such as *tapping feet* or *scratching hair*, in euthymic BD patients may signify deficits in inhibitory control, consistent with perseverative behavior[35] observed in manic and hypomanic BD patients[29]. However, these subtle behaviors are not included in established behavior rating criteria and were missed by both general-purpose action detection software and human annotators viewing our videos.

The motif identification process also enables us to establish parallels between human and animal behavior, enhancing our understanding of underlying mechanisms. For example, human *fidgeting* could be analogous to *grooming* behavior in rodents, reflecting similar responses to environmental stressors or internal states. Future studies on cross-species comparisons will broaden our perspective on behavior patterns to a more comprehensive understanding of the underlying brain and mind states.

*Motif 2* (*depart*) encompassed movements from the periphery (where objects are placed) to the center of the room (no object placed), as opposed to a seemingly more natural trajectory along the periphery. This observation could be consistent with the overactive goal-directed behavior observed in manic and hypomanic states in BD[11,40–42]. These relationships suggest that behavioral features characteristic of depressive or manic states of BD patients may persist during the euthymic state, albeit subtly, such that data analysis methods that are less sensitive may overlook this persistence. We also found that BD participants displayed sparser transition matrices, indicating more stereotyped modes of behavior, and altered variability in motif expression, as evidenced by variance of latent representations. The emergence of this collection of features as discriminators of BD from HC participants suggests that they are impacted by behavioral parameters such as attention, exploratory activity, novelty-seeking, and overall modulation of motor activity for people with BD euthymia.

While the focus of our study was on BD, our results highlight the potential of methods for automatic annotation of spontaneous behavior across species to assess individual responses to psychiatric treatments and uncover novel behavioral features across a range of neuropsychiatric disorders. Our approach can be straightforwardly applied across species, e.g., to animal models of psychiatric and cognitive conditions, critical to the understanding of biological mechanisms as well as drug discovery. Future endeavors aim to integrate our methodology with neural activity analyses to elucidate the neural mechanisms underlying behavioral abnormalities in humans and animals.

**Acknowledgements.** We thank Kristen Kraffert, Haili Song, and Nathan Wood for their help in data collection. Zhanqi Zhang is supported by the HDSI Ph.D. Fellowship at the University of California San Diego. Chi K. Chou was supported by the Shenoy Research Mentor Fellowship in Neuroscience (SURFiN). This study was conducted under IRB #180344 and was supported by NIH grants NIDA R01DA043535 (W.P./J.Y.) and NIDA R01DA051295 (A.M./J.Y.).

**Contributions**. A.M., J.Y., and W.P. designed the experiments and collected the data. Z.Z, G.M., and M.A. conceptualized the experiment analysis and analyzed the data with assistance from C.C., and H.R. Z.Z wrote the manuscript under G.M. and M.A.'s supervision. Z.Z., G.M., M.A., A.M., J.Y., and W.P. reviewed and edited the manuscript.

## Methods

### *Data and Procedure.*

All Patients (n = 25; 12 men) were between the ages of 18 to 55. Among the population, all but one patient was diagnosed with bipolar disorder (BD) Type I or Type II(defined by the Structured Clinical Interview for DSM-IV[30]). The remaining patient was diagnosed with the cyclothymic subtype of BD. All BD participants were in a current euthymic episode. Non-patient participants (n = 25; 15 men) of matching years of age who had never met the DSM-IV[30] standard for alcohol or substance abuse or dependence, tested positive on a urine toxicology screen, had a neurological ailment, or had a condition affecting their motor skills were recruited for the study as the healthy control group (HC). Participants from both BD and HC populations were evaluated with the Young Mania Rating Scale (YMRS)[4] and Hamilton Depression Rating Scale (HAM-D)[2], and all BD and HC participants had YMRS < 12 and HAM-D < 10. Most of the BD patients were treated with one or a combination of mood-stabilizing, antipsychotic, antidepressant, and sleep aid medication; other BD patients were not on medication during testing. See **Supplementary Table 1** for full information.

Participants consented to have their activities filmed during an unspecified segment of the research session. The video data was collected at the UCSD Medical Center in an unused office room that was designed to appear in transition. The room was 2.7 m × 4.3 m with a periphery lined with various pieces of furniture, such as a desk, both small and large filing cabinets, and two sets of bookshelves. No furniture that could directly lead to sedentary behavior was set in the room. Eleven small objects were placed evenly on items of furniture. These items were selected based on the condition that they are safe, vibrant, tactile, easily handled, and are likely to encourage exploration by humans[43].

Participants were directed to wait in the room with minimal instructions until the examiner returned. Participants were not allowed to leave the room or bring personal items into the room. The videos were recorded for $T = 15$ minutes continuously from a commercial camera with a fisheye lens hiddenly placed at the center of the ceiling. The recordings had a resolution of 640 x 480 pixels and a frame rate of 30 frames per second. Following the procedure in the previous studies on the dataset[10,11], the recorded session of 15 minutes was evenly divided into three 5-minute epochs for analysis in this study.

Human experts reviewed the video recordings afterward to count instances of 11 exploration action categories, including sitting with or without an object, standing with or without an object, walking with or without an object, lying with or without an object, wearing an object, exercising, and interacting with objects such as drawers and window blinds[11].

The spatial scaling exponent (Spatial-d) estimated the geometric structure of the path of the participants, first introduced in animal behavior studies[44] and used as a metric in previous human behavior studies on this dataset. It estimates the linear slope of $log(L_k)$ with respect to $log(k)$ where $L_k$ is the average length of the path and $k$ is the measuring resolution of the movements.

### *Human Pose Tracking and Estimation.*

Existing methodologies for human motion tracking were not developed for a single top-view camera with fish-eye distortion and thus performed poorly on this dataset. To characterize the participant's behavior, we used DeepLabCut[22]. Specifically, in DeepLabCut we first clustered the frames using k-means and selected frames from different clusters to obtain 20 - 50 frames from each video. This process ensures that the selected frames cover different poses of the person. We labeled these frames with markers at 20 anatomical landmarks (left eye, right eye, left ear, right ear, mouth, the center of the neck, left shoulder, right shoulder, left elbow, right elbow, left hand, right hand, the center of hip, left hip, right hip, left knee, right knee, left foot, right foot, the center of feet). The labeled frames were used for training a ResNet-50[45] model to learn and predict marker position in the remaining frames. In order to have accurate marker estimation, the training involved 3 iterations, with 1,030,000 epochs each. After each iteration, 10 outlier frames (DeepLabCut confidence score below 0.1) with inaccurate marker estimates from every video were relabeled and added to the training set for the next iteration. Training iterations were terminated when the training and testing errors of the DeepLabCut marker estimation were 2.03 pixels and 3.71 pixels, respectively. The x-y position estimates of the 20 body parts for each frame were used for subsequent analyses.

### *Key Point Marker Postprocessing.*

We aligned the skeleton markers of the human to egocentric coordinates. To accomplish this, we cropped the frame to the size of a bounding box (300 x 300 pixels) such that the whole person would fit in the bounding box. Then we aligned the skeleton using the key points of the center of the hip, and center of the feet markers as reference. As a result, the upper body markers were located at the top of the cropped frame, and the lower body markers at the bottom. Marker estimates with less than 90% confidence level determined by DeepLabCut were removed.

### *Encoding the Pose into Latent Space.*

To identify distinct behavioral motifs from times series of pose coordinates, we adapted the pipeline in the Variational Animal Motion Embedding (VAME) model[17], which has been used previously to identify open-field mouse behaviors using a bidirectional RNN variational autoencoder (VAE) and clustering. The VAME model was used to encode and reduce the dimensionality of the pose sequence of the human participants. Specifically, the latent dimensionality was set to $d = 10$, a value less than the input dimension of 40 (20 markers with x and y coordinates). The resulting latent representation $Z$ for each subject is thus a matrix of size $d \times T$.

The original VAME model used a hidden Markov model for extracting 50 motifs of the animal, used hierarchical clustering of *motifs* to obtain a tree-structured graph, and then grouped *motifs* into *communities* by cutting the tree at a certain level/depth of the branches. However, because human behavior may be more complex, the hierarchical representation of human behavior varied across *motifs* and was not visually similar in each *community*. We instead performed k-means clustering on the latent representation to obtain the behavioral *motifs*. As a direct comparison with 10 labels from human annotation, we included the results of 10 clusters in the main results of this study. We also reproduced our analysis using $k = 30$ clusters with results included in **Supplementary Fig. 3**.

### *Matching Annotation Labels with Motif Labels.*

For each video, we obtained a list of human annotations and a list of motif labels. Since the labels from human annotations and motifs obtained from the latent-variable model do not necessarily match one-to-one, we measured how many times the onset and offset of each label matched between the two labels. Both lists were filled with integers representing the action labels at each frame. For example, the first 8 frames from one video may be represented as `[a, a, a, b, b, b, c, c]`, with `as, bs, cs` standing for the labels of the action on that frame. We first divided the lists into chunks `[a, a, a], [b, b, b], [c, c]` so that each chunk represented an epoch with only one label, and a delimiter '0' was added between chunks. The output of the example frames would be `[[a, a, a], 0, [b, b, b], 0, [c, c]].` Since the objective was to find the onset/offset alignment, which was marked by the location of the `0`s only, the labels could be simplified as `[[1, 1, 1], 0, [1, 1], 0, [1, 1, 1]], with 1s` representing the chunks of labeled frames while `0` representing the chunk boundaries.

We computed the total number of chunks in human annotations, and the number of matching chunks between human annotation and motif labels in terms of onset/offset timestamp. Because human annotations of onset and offset of actions had inherent uncertainty, we defined a specified `offset` value allowing for a certain number of frames of mismatch.

For example, between

`list1 = [0, 1, 1, 1, 1, 1, 1, 0, 1, 1]` and

`list2 = [1, 0, 1, 1, 1, 0, 1, 1, 1, 0],`

with an offset of `2,` there are two matching labels chunks: `[1, 1, 1, 1, 1, 1] with [1, 1, 1] and [1, 1] with [1, 1, 1]`. We reported the ratio of matching labels to total human annotation labels. There are 33.19% of labels that were matched when the offset was 1 second, 76.90% when the offset was 5 seconds, and 87.10% when the offset was 10 seconds.

### *Computing effective-number-of-accessible-states (ENAS).*

Each $i \in n$ row of the transition matrix $P$ is composed of the transition probability, $P_{i,j}$ from *motif* $S_i$ into every other *motif* $S_j$. The intuition behind the *ENAS* is to measure how many *motifs* could be accessible based on the current observed transition matrix. If $\sum_{j=1}^{n} P_{i,j} = 0$, this indicates no other *motif* was visited from *motif* $i$, resulting in *ENAS* of *motif* $i$ to be 0 (self-transitions were excluded from computations). Otherwise, we compute *ENAS* of the motif $i$ in the following manner.

$$E_{S_i} = \left( \sum_{j \in [0, n]} p_{ij}^2 \right)^{-1}$$

The $E_{S_i}$ represents the number of accessible *motifs* from the current *motif* $i$, which is a number between 0 to $n$, where $n$ is the number of total *motifs*. If there is no motif accessible from the current motif, then $E_{S_i}$ will be 0.

The overall ENAS $E$ is the average of $E_{S_i}$overall *motif* $S_i$ for $i \in n$

$$E = \frac{1}{n}\sum_{i\in[0,\,n]} E_{S_i}$$

The pseudo-code for ENAS is the following:

```
ENAS(P):
for $row_i$ in $P$:
    if $\sum_{j=1}^{n} P_{i,j} = 0$:
        $E_{S_i} = 0$
    else:
        $E_{S_i} = \left( \sum_{j\in[0,\,n]} p_{ij}^2 \right)^{-1}$
```

### ***Computing Volume and Distance of Latent Representations.***

To compute the *latent-volume*, we first mean-centered the latent vectors of all *motifs* during the entire time $T$. The *latent-volume* $v_i(\tau_m, p)$ of the latent representation $Z_{i,\tau_m,p}$ of *motif* $i$ at the time $\tau_m$ of population, $p$ was quantified by the trace of the covariance of the latent vector $Z_i$

$$v_i(\tau, p) = Tr(Cov(Z_{i,\tau_m,p})).$$

To compute the *population-distance*, let's define the following:

At each motif $i \in [1, 2, \dots, k]$ and during each epoch $\tau_m$, the latent representation of a BD subject to be $X_i, \tau_m$ of $\mathbb{R}^d$, and the latent representation of an HC subject to be $Y_{i,\tau_m}$ of $\mathbb{R}^d$, where $d$ is the latent dimension.

Assume $X_{i,\tau_m} \sim N(m_1, \Sigma_1)$ and $Y_{i,\tau_m} \sim N(m_2, \Sigma_2)$, meaning each point in $X_{i,\tau_m}$ and $Y_{i,\tau_m}$ is an independent sample from its respective Gaussian distribution, with expected values and covariance.

We computed the 2-Wasserstein distance between $(X_{i,\tau_m}, Y_{i,\tau_m})$ at each motif $i \in [1, 2, \dots, k]$ and during each epoch $\tau_m$. Specifically,

$$d_{i,\tau_m}^2 = W_2\left(X_{i,\tau_m}, Y_{i,\tau_m}\right)^2 = \left|\left|m_1 - m_2\right|\right|_2^2 + Tr(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2})$$

where, $m_1$, $m_2$ and $\Sigma_1$, $\Sigma_2$ are sampled means and covariances. The 2-Wasserstein distance was computed with the Python function below.

*Interpopulation-distance* was the mean of pairwise 2-Wasserstein distance between every subject in BD and every subject in HC. For comparison, we computed *intrapopulation-distance*, as the mean pairwise 2-Wasserstein distance within the HC group and within the BD group.

```
def wasserstein_distance(m1, C1, m2, C2):
    """
    Calculate the 2-Wasserstein distance between two Gaussian distributions.

    Parameters:
    m1, m2: Mean vectors of the two Gaussian distributions (numpy arrays).
    C1, C2: Covariance matrices of the two Gaussian distributions (numpy arrays).

    Returns:
    W2: The 2-Wasserstein distance.
    """
    # Euclidean distance between the means
    mean_diff = np.linalg.norm(m1 - m2)

    # Principal square roots of the covariance matrices
    # Calculate the trace term
    term = sqrtm(sqrtm(C2) @ C1 @ sqrtm(C2))
    trace_term = np.trace(C1 + C2 - 2 * term)

    # Wasserstein distance squared
    W2_squared = mean_diff ** 2 + trace_term

    return np.sqrt(W2_squared).real
```

### *Visualization of the Latent Representation.*

Since the latent representation is in a dimension of $d \times T$, we transformed the latent space using PCA, and the first three principal components (PCs) were plotted for visualization purposes. The motif centroids and centroid distances defined above were also computed separately in PC space and plotted in the top three PCs for proper visualization. All latent representations were visualized in the PC space (computed from the entire latent representation).

### *Baseline Computer Vision Models***.**

We selected two state-of-the-art computer vision action recognition models, MMAction2[28] and S3D-CNN[31] since not many models would detect the person in the setting of the top view fisheye camera used in the study.

We adapted OpenMMLab's official repository for MMAction2 (https://github.com/open-mmlab/mmaction2). MMAction2 consists of two modules: a human detection using faster RCNN

ResNet50 with COCO dataset, and an action detection using SlowFast ResNet50 network pretrained on Kinetics-400 first for action classification and then fine-tuned on AVA v2.2 dataset for person detection. All pretrained weights and configuration files were downloaded from the repository. We used the following configuration and checkpoints for MMAction2:

```
--config
configs/detection/ava/slowfast_kinetics_pretrained_r50_8x8x1_cosine_10e_ava22_rgb.py
--checkpoint slowfast_kinetics_pretrained_r50_8x8x1_cosine_10e_ava22_rgb-b987b516.pth
--det-checkpoint faster_rcnn_r50_fpn_mstrain_3x_coco_20210524_110822-e10bd31c.pth
--det-score-thr 0
--action-score-thr 0
--label-map tools/data/ava/label_map.txt
```

For S3D-CNN[31], we used the unofficial PyTorch implementation (https://github.com/kylemin/S3D), which was pretrained on the Kinetics-400 dataset with pretrained weights downloaded from the same repository. S3D takes in the video dataset and outputs the labels from Kinetics-400 for each frame in the video.

### Selecting *Features for Classification.*

Our data is comprised of numerical input features and categorial output labels (BD and HC). We applied backward feature selection using `SequentialFeatureSelector(n_features_to_select=15, direction="backward",scoring='accuracy', cv=4)` from `sklearn.feature_selection`. This is a greedy sequential feature algorithm that sequentially removes features from all features based on a 4-fold cross-validated score of the accuracy of the logistic regression classifier. The feature selector stops removing features when the desired number of selected features is reached. Before feature selection, there are 67 input features of each human video, including each motif's dwell time at three epochs, ENAS of each motif at three epochs, ENAS of all motifs at three epochs, number of zeros in transition matrices, motif volume at three epochs, YMRS scale, and HAMD scale. After feature selection, 15 features were selected from each approach (**Supplementary Table 3**).

### *Classifying BD from Behavior Features.*

Selected features were fed into a binary logistic regression classifier. We utilized a logistic regression classifier from scikit-learn (`LogisticRegression`) with a maximum number of iterations set to 1000. Each feature of the dataset was min-max scaled using `MinMaxScaler` from `sklearn.preprocessing`. For each iteration, we split the data randomly into 75% training and 25% testing sets using stratified sampling, then trained a logistic regression classifier for each iteration, and computed accuracy, precision, and recall scores (using the `accuracy_score`, `precision_score`, and `recall_score` functions from scikit-learn) on the test set for each iteration. We conducted cross-validation with 3 folds to estimate model performance using `cross_validate` from scikit-learn. We reported mean and standard deviation of accuracy, precision, and recall scores across all iterations. We performed Tukey's range test between pairwise scores between our model and other models and reported the p-values.

# Reference

1. Martinowich, K., Schloesser, R. J. & Manji, H. K. Bipolar disorder: from genes to behavior pathways. *J. Clin. Invest.* **119**, 726–736 (2009).
2. Hamilton, M. Development of a rating scale for primary depressive illness. *Br. J. Soc. Clin. Psychol.* **6**, 278–296 (1967).
3. A RATING SCALE FOR DEPRESSION | Journal of Neurology, Neurosurgery & Psychiatry. https://jnnp.bmj.com/content/23/1/56.
4. A rating scale for mania: reliability, validity and sensitivity - PubMed. https://pubmed.ncbi.nlm.nih.gov/728692/.
5. Möller, H. J. Rating depressed patients: observer- vs self-assessment. *Eur. Psychiatry* **15**, 160–172 (2000).
6. Hitchcock, P. F., Fried, E. I. & Frank, M. J. Computational Psychiatry Needs Time and Context. *Annu. Rev. Psychol.* **73**, 243–270 (2022).
7. Mc Reynolds, P. Exploratory Behavior: A Theoretical Interpretation. *Psychol. Rep.* **11**, 311–318 (1962).
8. Meyer-Lindenberg, A. The non-ergodic nature of mental health and psychiatric disorders: implications for biomarker and diagnostic research. *World Psychiatry* **22**, 272–274 (2023).
9. Fisher, A. J., Medaglia, J. D. & Jeronimus, B. F. Lack of group-to-individual generalizability is a threat to human subjects research. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6106–E6115 (2018).
10. Young, J. W., Minassian, A., Paulus, M. P., Geyer, M. A. & Perry, W. A Reverse-Translational Approach to Bipolar Disorder: Rodent and human studies in the Behavioral Pattern Monitor. *Neurosci. Biobehav. Rev.* **31**, 882–896 (2007).
11. Perry, W. *et al.* A reverse-translational study of dysfunctional exploration in psychiatric disorders: from mice to men. *Arch. Gen. Psychiatry* **66**, 1072–1080 (2009).

12. Growing Points Ethology | Animal behaviour. *Cambridge University Press* https://www.cambridge.org/us/academic/subjects/life-sciences/animal-behaviour/growing-points-ethology, https://www.cambridge.org/us/academic/subjects/life-sciences/animal-behaviour.

13. Tinbergen, N. *The Study of Instinct*. xii, 237 (Clarendon Press/Oxford University Press, New York, NY, US, 1951).

14. Wiltschko, A. B. *et al.* Revealing the structure of pharmacobehavioral space through motion sequencing. *Nat. Neurosci.* **23**, 1433–1443 (2020).

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

16. Weinreb, C. *et al.* Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. 2023.03.16.532307 Preprint at https://doi.org/10.1101/2023.03.16.532307 (2023).

17. Luxem, K. *et al.* Identifying Behavioral Structure from Deep Variational Embeddings of Animal Motion. 2020.05.14.095430 Preprint at https://doi.org/10.1101/2020.05.14.095430 (2022).

18. Berman, G. J., Bialek, W. & Shaevitz, J. W. Predictability and hierarchy in Drosophila behavior. *Proc. Natl. Acad. Sci.* **113**, 11943–11948 (2016).

19. Cande, J. *et al.* Optogenetic dissection of descending behavioral control in Drosophila. *eLife* **7**, e34275 (2018).

20. Berman, G. J., Choi, D. M., Bialek, W. & Shaevitz, J. W. Mapping the stereotyped behaviour of freely moving fruit flies. *J. R. Soc. Interface* **11**, 20140672 (2014).

21. Hsu, A. I. & Yttri, E. A. B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nat. Commun.* **12**, 5188 (2021).

22. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).

23. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning | eLife. https://elifesciences.org/articles/47994.

24. Wu, A. *et al.* Deep Graph Pose: a semi-supervised deep graphical model for improved animal pose tracking. in *Advances in Neural Information Processing Systems* vol. 33 6040–6052 (Curran Associates, Inc., 2020).

25. Bordes, J. *et al.* Automatically annotated motion tracking identifies a distinct social behavioral profile following chronic social defeat stress. *Nat. Commun.* **14**, 4319 (2023).

26. Pereira, T. D. *et al.* SLEAP: A deep learning system for multi-animal pose tracking. *Nat. Methods* **19**, 486–495 (2022).

27. MoveNet: Ultra fast and accurate pose detection model. | TensorFlow Hub. *TensorFlow* https://www.tensorflow.org/hub/tutorials/movenet.

28. MMAction2 Contributors. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. (2020).

29. Henry, B. L. *et al.* Inhibitory deficits in euthymic bipolar disorder patients assessed in the Human Behavioral Pattern Monitor. *J. Affect. Disord.* **150**, 948–954 (2013).

30. Bell, C. C. DSM-IV: Diagnostic and Statistical Manual of Mental Disorders. *JAMA* **272**, 828–829 (1994).

31. Xiong, X. *et al.* S3D-CNN: skeleton-based 3D consecutive-low-pooling neural network for fall detection. *Appl. Intell.* **50**, 3521–3534 (2020).

32. Kay, W. *et al.* The Kinetics Human Action Video Dataset. Preprint at https://doi.org/10.48550/arXiv.1705.06950 (2017).

33. Monosov, I. E., Zimmermann, J., Frank, M. J., Mathis, M. W. & Baker, J. T. Ethological computational psychiatry: Challenges and opportunities. *Curr. Opin. Neurobiol.* **86**, 102881 (2024).

34. Ravizza, S. M. & Carter, C. S. Shifting set about task switching: Behavioral and neural evidence for distinct forms of cognitive flexibility. *Neuropsychologia* **46**, 2924–2935 (2008).

35. Oosterloo, M., Craufurd, D., Nijsten, H. & van Duijn, E. Obsessive-Compulsive and Perseverative Behaviors in Huntington's Disease. *J. Huntingt. Dis.* **8**, 1–7.

36. Activation in Bipolar Disorders: A Systematic Review | Bipolar and Related Disorders | JAMA Psychiatry | JAMA Network. https://jamanetwork.com/journals/jamapsychiatry/article-abstract/2592473.

37. Ozdel, O., Karadag, F., Atesci, F. C., Oguzhanoglu, N. K. & Cabuk, T. Cognitive functions in euthymic patients with bipolar disorder. *Ann. Saudi Med.* **27**, 273–278 (2007).

38. Henry, B. L. *et al.* Cross-species assessments of Motor and Exploratory Behavior related to Bipolar Disorder. *Neurosci. Biobehav. Rev.* **34**, 1296–1306 (2010).

39. Decision-making and trait impulsivity in bipolar disorder are associated with reduced prefrontal regulation of striatal reward valuation - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4107743/.

40. Benazzi, F. Testing new diagnostic criteria for hypomania. *Ann. Clin. Psychiatry Off. J. Am. Acad. Clin. Psychiatr.* **19**, 99–104 (2007).

41. Angst, J. *et al.* Toward a re-definition of subthreshold bipolarity: epidemiology and proposed criteria for bipolar-II, minor bipolar disorders and hypomania. *J. Affect. Disord.* **73**, 133–146 (2003).

42. Akiskal, H. S., Azorin, J. M. & Hantouche, E. G. Proposed multidimensional structure of mania: beyond the euphoric-dysphoric dichotomy. *J. Affect. Disord.* **73**, 7–18 (2003).

43. Pierce, K. & Courchesne, E. Evidence for a cerebellar role in reduced exploration and stereotyped behavior in autism. *Biol. Psychiatry* **49**, 655–664 (2001).

44. Paulus, M. & Geyer, M. Paulus MP, Geyer MA. A temporal and spatial scaling hypothesis for the behavioral effects of psychostimulants. Psychopharmacology (Berlin) 104: 6-16. *Psychopharmacology (Berl.)* **104**, 6–16 (1991).

45. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. Preprint at https://doi.org/10.48550/arXiv.1512.03385 (2015).