

AI-driven decoding of naturalistic behaviors enables tailored detection of depressive-like behavior in mice

Received: 17 October 2024

Accepted: 2 December 2025

Published online: 30 December 2025

 Check for updates

Hyeonsik Oh ^{1,8}, Sian Choi ^{1,8}, Jinsu Lee ^{2,3}, Heeyoung Lee ¹,
Jongpil Shin ¹, Seungkyu Son⁴, Bobae Hyeon ^{1,5} & Won Do Heo ^{1,6,7} 

Major depressive disorder (MDD) is an etiologically diverse psychiatric disease with heterogeneous manifestations, making it difficult to diagnose with conventional assessment standards. In addition, the obvious incompatibility of the standard survey-based tests for human MDD and the behavioral assays for depressive-like phenotypes in mice makes clear the requirement for a non-invasive method for quantifying the expression of depressive-like state in naturalistic contexts. Here, we introduce a self-supervised machine learning platform, CLOSER (Contrastive Learning-based Observer-free analysis of Spontaneous behavior for Ethogram Representation), to monitor the spontaneous behavior in a depressive disease model with enhanced precision, reliability, and efficiency. This framework incorporates 3D pose skeleton data and kinematic features in a unique data augmentation strategy to characterize semantic behavioral syllables with a high-quality feature space. Using CLOSER, we uncovered distinct motion profiles in chronically stressed mice across both sexes and different disease stages. Furthermore, we quantified the drug-specific recovery of psychomotor symptoms, highlighting CLOSER's discriminative power for identifying drug efficacy. In offering an artificial intelligence (AI)-driven decoding of exploratory behaviors, CLOSER proposes the standardization of depressive-like phenotype assessment in mouse models, thereby bridging preclinical and clinical diagnostics for psychiatric drug discovery.

Major depressive disorder (MDD) is a stress-induced disorder associated with persistent feelings of sadness and loss of interest. MDD has diverse causes, including stress, and is associated with an elevated risk of developing a host of other severe medical disorders. Despite its prevalence across the world, the heterogeneity of MDD symptoms and

their severity in patients of different ages and genders constrains the therapeutic effectiveness of pharmacological interventions^{1,2}. Recently, researchers have tried to discriminate genetic and pathophysiological depression subtypes, while little attempt has been made to classify depressive episodes according to the motion parameters,

¹Department of Biological Sciences, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea. ²The Gene Lay Institute of Immunology and Inflammation, Brigham and Women's Hospital, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. ³Ann Romney Center for Neurologic Diseases, Brigham and Women's Hospital, Boston, MA, USA. ⁴Center for Cognition and Sociality, Institute for Basic Science, Daejeon, Republic of Korea. ⁵Molecular Neurobiology Laboratory, McLean Hospital and Department of Psychiatry, Harvard Medical School, Belmont, MA, USA. ⁶KAIST Institute for the BioInnovation (KIBI), Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, Republic of Korea. ⁷Department of Cognitive Brain Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea. ⁸These authors contributed equally: Hyeonsik Oh, Sian Choi. ✉ e-mail: wondo@kaist.ac.kr

which are the symbolic symptoms of psychiatric diseases^{3–6}. We aimed, therefore, to develop a diagnostic module encompassing the heterogeneity of depression for use in symptom-based antidepressant drug discovery.

Traditional behavioral screening tests in rodents (e.g., the tail suspension test (TST), forced swim test (FST), and sucrose preference test) have been used to evaluate depressive-like symptoms that are affected by stress exposure and antidepressant treatment. Unfortunately, they show low construct validity⁷, making them incapable of reflecting the complex psychiatric symptoms and prognoses of MDD. The highly controlled environments and oversimplified conditions used to elicit non-spontaneous behaviors in these tests have limited the development of the most common antidepressants, resulting in high false positive rates for novel antidepressant discovery^{8,9}. Moreover, the over-reliance of the tests on manual scoring introduces subjectivity and inter-observer variability, further limiting the accuracy and reproducibility of depressive disorders models¹⁰.

In human mental diseases, the motion-associated symptoms, or psychomotor activity, have been assessed through behavioral observation of limb movements, trunk posture, voice, facial expression, and eye movements⁵, in parallel with identifications of circuitual and functional connectivity between the amygdala and motor cortex^{11–14}. For rodents, the machine learning approaches (both supervised and unsupervised learning) have started to distinguish the psychiatric state of mice with distinct motion features, decomposing behavioral structures induced by genetic manipulation or antipsychotics^{15–17}. In general, the supervised method trains the machine to provide a theory-driven classification for the reproduction of human annotation, while the performance of such models is strongly limited because the supervised classifier can never outperform human perception¹⁸. The alternative approach is a data-driven method that mathematically organizes behavior syllables to preserve the natural principle of a motion sequence¹⁹. The major advantage of the unsupervised approach is the ability to uncover hidden behavioral patterns without human intervention. A popular unsupervised algorithm, contrastive learning, is a self-supervised model that learns behavioral representations to capture relevant features in the data^{20–22}, by contrasting pairs of augmented views from the same sample against unrelated negative samples. During the training process of contrastive learning, data augmentation compels the machine to generate discriminative representations such that similar data pairs are located near one another in the embedding space. Importantly, contrastive learning semantically strengthens unsupervised approaches by modulating the combination of appropriate augmentations to capture invariance in behavioral data.

Therefore, we present Contrastive Learning-based Observer-free analysis of Spontaneous behavior for Ethogram Representation (CLOSER), a platform for scalable measurement of exploratory behaviors in depressive disorder model mice, employing 3D skeleton coordinates obtained from the precise motion recognition framework, AVATAR²³. Trained on both 3D action skeleton coordinates and kinematic features, CLOSER surpasses the quality of embedding structures produced by current unsupervised pose estimation algorithms and shows improved predictive performance for manually annotated behavior datasets. Using CLOSER, we identified motion profiles for spontaneous movements in depressive mouse models that are affected by sex and disease severity. This discovery permits the computational evaluation of depressive behaviors with consideration for symptom variance between distinct cohorts. Finally, our study demonstrates that CLOSER accomplished a pivotal improvement in the identification of distinctive signatures for monoaminergic and non-monoaminergic antidepressants in naturalistic behavior, facilitating the interpretation of behavioral data required in future preclinical screens for new psychiatric drugs.

Results

Self-supervised learning of mouse exploratory behaviors using 3D action skeleton representations

To assess mouse behavioral structure in a quantitative manner, we constructed a contrastive learning module leveraging a distillation that interchanges information learned from joint and motion modalities to acquire both allocentric (global) and egocentric (local) action skeletal knowledge from recorded video in AVATAR studio (Fig. 1a). The module was trained with a motion dataset established by shifting a 10-frame-long sliding window through the recorded videos one frame (or about 33 msec) at a time. Since our model learns the characteristic invariances of mouse behavior by capturing semantic similarities from two augmented views (positive pairs), we conducted an ablation study for eight different combinations of augmentation strategies to optimize the feature space of the behavior sequence (Supplementary Table 1). In the estimation of model prediction accuracy, the algorithm incorporating size scaling, joint angle distortion, motion crop, and x- or y-axis mask presented the best performance on all evaluation metrics for the classification of manually labeled mouse behavior segments. We observed that reported augmentation sets optimized for human 3D skeleton representations produced relatively low accuracy (Supplementary Table 1, column a), suggesting human and mouse behaviors exhibit distinct semantic invariance.

Next, to evaluate the geometric variance of latent spaces across multiple augmentation strategies, we measured Pearson correlation coefficients for representations of 13 manually labeled behaviors (Supplementary Table 2) in an unseen human dataset to compare within- (or intra-) and between (or inter-) cluster relationships (Supplementary Fig. 1). A coefficient of +1 or –1 denotes perfect positive or negative linear association, respectively, while 0 indicates no linear relationship. We found that augmentation schemes with superior classification accuracy also exhibited stronger intra-cluster correlation than inter-cluster correlation. Importantly, our mouse-optimized augmentation yielded enhanced embedding for grooming behaviors (clusters 6 and 7) compared with augmentation combinations lacking channel masking (strategy 3) or scaling (strategy 2), highlighting the necessity of a complete augmentation strategy for mouse behavior latent structures.

To determine and categorize behavioral syllables, which are minimal motion units involving ethological context, we leveraged K-means clustering in 128 dimensions to partition behavior segments into 50 clusters. Then, a cluster label for each frame was determined by majority voting on candidates assigned from respective sliding windows at the first learning step (Fig. 1a). Given the continuous nature of mouse behavior, a behavioral syllable was denoted as a consecutive frame with the same cluster label. For validation of temporal segmentation, we monitored whether the dissection of a postural sequence captured the temporal dynamics of behavior shifts from one motion to another by utilizing a keypoint change-score, which highlighted the robust shift within the 3D action skeleton. Observing the transition of behavior motifs longer than 10 frames (or about 330 msec) with z-scored keypoint change scores, we found that the start frame (0% frame fraction) recorded the maximal change-score compared to 50% frame fractions in either direction, which matched the change-point of keypoint-MoSeq, the existing unsupervised motion analysis frameworks, and manually annotated segments (Fig. 2a). The duration of the defined clusters (17.69 ± 24.58 frame, 590 ± 819 ms) was slightly shorter than both the manually defined behavior syllables (28.78 ± 25.83 frame, 959 ± 861 ms) and segments created by keypoint-MoSeq (27.93 ± 34.95 frame, 931 ± 1165 ms) but longer than those created by Variational Animal Motion Embedding (VAME) (7.75 ± 12.52 frame, 258 ± 417 ms) (Fig. 2b), while also satisfying the length requirement for general sub-second behavior motifs (<500 ms)¹⁸.

To estimate whether the models could follow the criteria of human annotation, we evaluated the prediction power of five

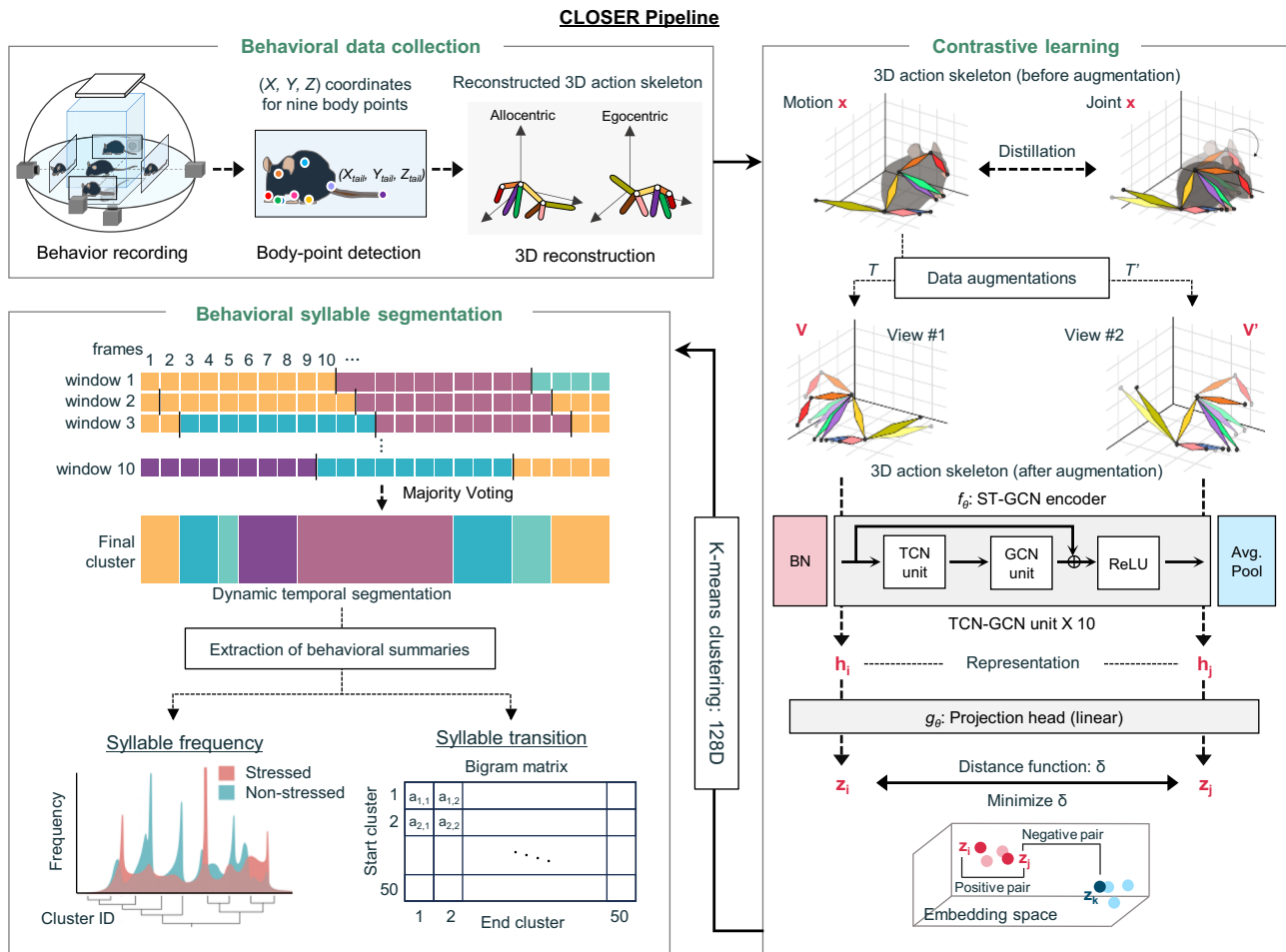


Fig. 1 | CLOSER workflow for self-supervised learning recognizing motion semantic units in the exploratory behavior of mice. Schematic representation of the CLOSER pipeline for 3D pose estimation of freely moving mice. (Left, Behavioral data collection) Five-camera recording system records 2D motion sequence of mice to obtain the x-, y-, and z-coordinates of nine body parts (the nose, head-top, body, anus, tail-tip, and four paws), reconstructed in 3D action skeletons. (Right) The contrastive learning-based machine learning platform workflow. The framework

was trained through the joint and motion modalities along with knowledge distillation. Augmented skeletons were embedded using an ST-GCN encoder and a projection head, where the InfoNCE loss is applied during embedding construction, followed by K-means clustering in the 128D embedding space. (Left, Behavioral syllable segmentation) Dynamic temporal segmentation employing a majority voting strategy. Consecutive frames with the same label were denoted “behavioral syllable”.

individual methods on the discrimination of labeled behaviors through k-nearest neighbor (k-NN) probing, a reliable indicator of downstream task performance²⁴ (Fig. 2c). In benchmarking, we compared VAME trained for 20 epochs (to match CLOSER) against the default 500-epoch setting and employed the default setting, achieving optimal performance, although the improvement over 20 epochs was marginal (Supplementary Fig. 2). Moreover, since keypoint-MoSeq employs an HMM that learns transition probabilities between states, behavioral syllables IDs, and emission probabilities from a given 13-dimensional Principal Component Analysis (PCA) projection of motion sequences, we leveraged 270-dimension representation with 13-dimension PCA and Uniform Manifold Approximation and Projection (UMAP) embedding for 10-frame length behavioral segments, instead of keypoint-MoSeq. In behavioral classification, CLOSER provided the highest balanced (66%) and top-1 accuracy (55%), accompanied by consistently strong precision, recall, and F1 score. For further intrinsic evaluation of embedding spaces, we computed the Pearson correlation, which quantifies the strength and direction of the linear relationship between two behavioral segments. Using this method, we evaluated intra- and inter-cluster correlation coefficients for manually annotated behaviors within the latent spaces of multiple embedding strategies (Supplementary Fig. 3). These results demonstrated that CLOSER consistently yielded higher intra-cluster correlations than

inter-cluster correlations, signifying superior embedding quality compared to all other algorithms. Among alternative methods, UMAP achieved the second-highest quality of embedding structure, whereas VAME produced relatively high correlations in both intra- and inter-cluster, indicating limitations in discriminative ability.

We then applied the dimensionality reduction algorithm, UMAP, to visualize the structure of feature space for CLOSER (Fig. 2d). This result implied that 128-dimensional behavior representations are well-organized even in two-dimensional (2D) space, with explicit compartmentalization of embedding space for individual behavioral clusters. Next, to determine if the resulting clusters comprised behavior syllables with uniform ethological contexts, we examined the embedding structures and kinematic features of the clusters with strong annotation as representative compositional behaviors, such as walking, turning, rearing up, and rearing down (Fig. 2e). Notably, behavioral clusters classified in the same groups closely distributed with unique cluster areas. When we extracted the moved distance, head-body angle, and nose height from the 3D pose skeleton, each cluster class showed motion hallmarks appropriate for its corresponding behavior label (Fig. 2f-i). These model qualifications with typical postural phenotypes demonstrate that CLOSER produces a fine characterization of sub-second motion segments containing meaningful behavioral sequences, which is a fundamental building block of the mouse ethogram.

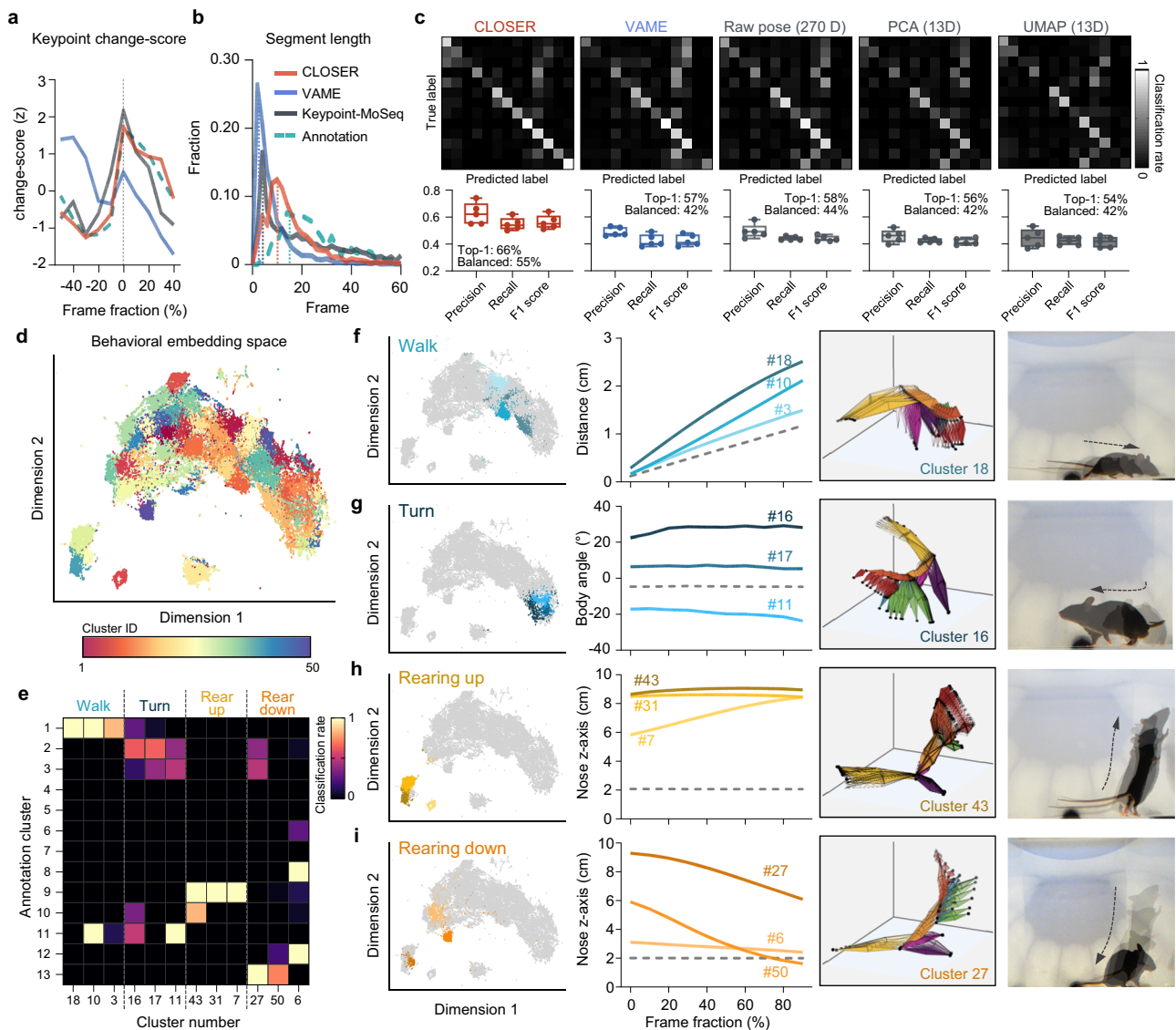


Fig. 2 | The CLOSER framework recognized temporal and behavioral characteristics for classifying behavior segments into semantic syllables. **a** Z-scored change-scores of the keypoint dynamics for motion segments longer than ten frames, followed by a normalization of the x-axis using the relative frame fraction for total segment length (100%). A higher change-score indicates a more robust shift in the 3D coordinates of the keypoints, improving inferences of the proper borderline for each behavioral motif. **b** The distribution of durations of the behavioral motifs recognized by CLOSER (red), VAME (blue), Keypoint-MoSeq (gray), and by manual annotation (green, dotted line). The data are presented as a fraction of the total syllable length for each frame length, where 30 frames equal 1 s. The dotted lines denote the mode of each distribution across analytic methods. **c** Normalized confusion matrix (top) and associated performance metrics (bottom) showing the accuracy of the k-nearest neighbor classifier based on the latent spaces of different embedding methods. Accumulated performance over the fivefold cross-validation is shown. The classification rate is represented in grayscale with

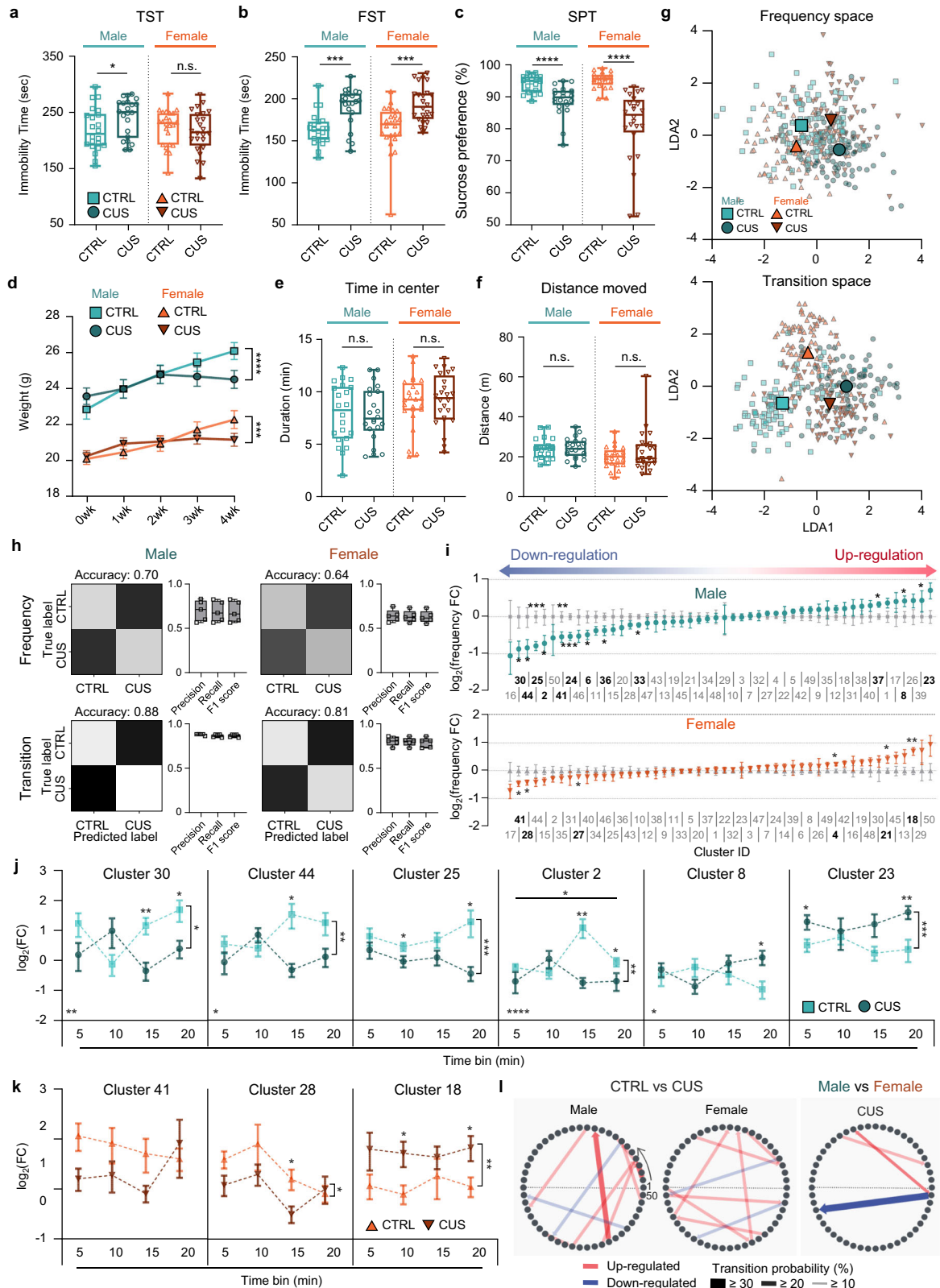
white for a classification rate of 1 and black for a classification rate of 0. Box plots show the 25th–75th percentiles of the precision, recall, and F1 score for methods with center lines for median; center lines indicate the median, and whiskers show the minimum and maximum values. **d** Behavioral feature space of CLOSER dimensionally reduced by UMAP. **e** Magma-scale heatmap of manual scoring results for representative ethograms, including “walk” (clusters 18, 10, and 3), “turn” (clusters 16, 17, and 11), “rear up” (clusters 43, 31, and 7), and “rear down” (clusters 27, 50, and 6). Annotation cluster numbers correspond to previous coarse and fine-grained labels in the Supplementary Table 2. **f–i** The location of ethograms in behavioral feature space, temporal dynamics of kinematic characteristics of ethograms against the relative time points, reconstructed 3D action skeletons, and corresponding motion sequences for walking (**f**), turning (**g**), rearing up (**h**), and rearing down (**i**). In each graph, the numbers next to the line denote the cluster IDs, and dashed lines (gray) demonstrate the average values for each locomotor metric. Source data are provided as a Source Data file.

To determine whether the latent space of CLOSER learned to rely on pose dynamics rather than confounding variables, we assessed mouse-size robustness in CLOSER and benchmarking pipelines, the former having received explicit scaling augmentation. Eighty male and female mice were binned into five weight groups, from 18 × g upward in 2-g steps, and we extracted embedding vectors in non-overlapping 100-frame strides for all mice. These embedding vectors were used as inputs to a Random Forest classifier trained with fivefold validation to predict weight group. CLOSER yielded the lowest accuracy, precision,

recall, and F1 scores among all algorithms (Supplementary Fig. 4), indicating that weight information is the hardest to extract from its latent space and confirming that the scaling augmentation effectively enforces size invariance in mouse behavioral segmentation.

Chronic unpredictable stress (CUS) induces typical depressive-like behaviors in mice in a sex-dependent manner

While chronic stress is known to modify the behavior of mice in a sex-dependent manner, most studies have included only a partial



examination of depressive-like phenotypes, requiring an array of behavior tests to compare depressive behaviors in males and females. To address this issue, we exposed male and female mice to 4 weeks of CUS to generate an etiologically relevant depressive disorder model, followed by an evaluation of depressive-like behaviors. In male CUS mice, we observed a significant increase in immobility at TST and FST (Fig. 3a, b, left) and decrease in sucrose preference (Fig. 3c, left), as well

as a decrease in body weight (Fig. 3d). Although female CUS mice showed reduced sucrose preference (Fig. 3c, right) and a notable loss of body weight (Fig. 3d), we detected immobility increase only with the FST, not the TST (Fig. 3a, b, right). The reduced appearance of stress-related behavior in females during TST sessions may be attributed to hormonal influences associated with estrous cycle fluctuations, which were not measured in this experiment. When we evaluated motor

Fig. 3 | Chronic unpredictable stress modifies the structure of exploratory behavior in a sex-dependent manner. **a** Chronic stress significantly increased immobility time in the TST for CUS males but not in CUS females. **b** CUS exposure induced a significant increase in immobility time in the FST for both males and females. **c** Stress exposure reduced sucrose preference in the SPT. **d** The body weight in both male and female mice was significantly decreased in weight in the fourth week by the CUS paradigm. **e, f** Total time spent in the center (**e**) and distance traveled during behavioral recordings (**f**). **g** 2D visualizations of the LDA projection for syllable frequency (top) and transition (bottom) information detected by the CLOSER framework. The large, bright points indicate the means of the LDA coordinates for each cohort, and the small, dim points depict the LDA results for individual mice. **h** Classification matrices summarizing the prediction accuracy of the random forest classifier discriminating stressed mice from normal mice using frequency (top) or transition datasets (bottom) in male (left) and female (right) mice. **i** Logistic graphs showing the fold change (FC) in cluster frequency,

altered by exposure to chronic stress, in male (top) and female (bottom) mice, ordered from stress-down-regulated to stress-upregulated clusters. Differentially expressed clusters are marked with significance levels and highlighted in bold in the cluster ID table. Values are represented as means \pm SEM. **j, k** Time-bin fold changes for the usage of differentially expressed syllables in male (**j**) and female (**k**) CUS mice. **l** Network graph depicting grammar dissimilarities between the control and CUS groups (left) or between different sexes (right). Colors indicate the mode of divergence (upregulation: red, downregulation: blue), and width represents the degree of change. In box plots, the box covers the 25th–75th percentiles of the dataset; center lines denote the median, where whiskers show the minimum and maximum values. All individual mice appear as data points. $N = 22$ for control (sky blue, square) and CUS (blue-green, circle) males and $N = 24$ for control (triangle, orange) and CUS (inverted triangle, brown) females. See Supplementary Data 1 for statistical details. Source data are provided as a Source Data file.

activity, we did not see any critical changes as a result of stress (Fig. 3e, f).

The CLOSER pipeline tracks the sex-specific behavior of CUS model mice in an open field

Rather than contribute to the current oversimplified model of depressive-like symptoms, we traced sophisticated chronic stress-induced behaviors of freely moving mice. To detect variations in the spontaneous behavior of individual mice, we used the AVATAR system to record mice at the beginning (0 weeks) and end (4 weeks) of the CUS paradigm, followed by normalization of behaviors at four weeks with the initial behavior frequency. Initially, we performed a Linear Discriminant Analysis (LDA) to transform the syllable frequency and transition summaries of five-minute intervals into a low-dimensional subspace (Fig. 3g), which enhances the expressiveness of stress-induced ethograms to address the sexual disparity. As LDA searches the axes showing the best classification performance, the larger distances between samples from different classes indicate the higher disparity in the original data²⁵. The result revealed a divergence in the summaries of syllable frequency between stressed male and female mice. In contrast, syllable transition information converged from distinct embeddings, implying a discrete pattern of alteration in syllable usage and bigram. Using these behavioral structures, we measured the diagnostic accuracy of a random forest classifier trained by syllable frequency or inter-syllable transitions (Fig. 3h). In both males and females, CLOSER-identified syllables and sequences enabled the classifier to predict depressive disorder, with the sequence-trained decoder performing greater and suggesting larger changes in the temporal dynamics than syllable usage with the CUS model. Thus, our behavior analysis framework can extract fundamental motion features from exploratory behaviors affected by stress exposure.

To further evaluate the efficiency of CLOSER in the detection of depressive-like behavior, we applied the same analysis on male and female groups using the syllable combinations identified by VAME and keypoint-MoSeq. Classification accuracy indicated that CLOSER outperformed VAME and keypoint-MoSeq in the male mice, achieving superior predictive performance for both classifiers trained using syllable frequency of transition (Supplementary Fig. 5a, b). In contrast, for non-stressed and stressed females, the random forest classifier leveraging syllable frequency from keypoint-MoSeq demonstrated greater performance compared to other behavior recognition pipelines (Supplementary Fig. 5c); nevertheless, the syllable transition calculated by CLOSER consistently yielded the highest accuracy (Supplementary Fig. 5d). Notably, behavioral deconstruction in males offered higher classification rates than in females for all segmentation methods, aligning with the observed sex-dimorphic alterations of syllable frequency and transition (Fig. 3i, l).

Before interpreting the stress-induced changes in behavioral structure, we annotated the clusters with one or two stereotyped

components while considering the discrete property of behavior. In the hierarchical structure of defined behavior syllables, clusters assigned analogous labels were located adjacent to one another in the hierarchy table (Supplementary Fig. 6), for example, the “wall rear up” clusters (clusters 49, 8, and 34), revealing the strong proximity of relevant behavior segments in the CLOSER embedding space. With this knowledge, we next explored the heterogenic alteration in behavior that emerged from chronic stress exposure. When evaluating syllable frequency in CUS mice, males and females showed statistical changes in the usage of twelve and six syllables, respectively (Fig. 3j), consistent with the greater predictive performance of the frequency-based classifier for males (Fig. 3h, frequency). Although there was a single overlap (cluster 41) in significant clusters, the trends in the major alterations were maintained across the sexes, with minor sex-dimorphic syllables (clusters 30, 24, and 6) showing sexual dimorphism in the depressive mouse model. In further kinematic profiling, chronic stress selectively reduced the occurrence of motifs with elevated body angles and sniff-related behaviors with large nose shifts relative to body shifts (Supplementary Fig. 7a), while higher nose heights were observed in stress-elevated syllables (Supplementary Fig. 7b). In female mice, stress-elevated syllables displayed greater body angles, in agreement with manual annotation, turning behaviors (Supplementary Fig. 7c, d).

Once we identified discriminant-relevant syllables, we wondered whether depressive phenotypes in naturalistic behavior had a direct correlation with conventional depressive-like behaviors. Hence, we used a linear regression model to predict the immobility time or sucrose preference using discriminant-relevant syllables and calculated a coefficient of determination (R^2 score) to evaluate the accuracy of predictive value compared to actual data (Supplementary Fig. 8). In contrast, the structure of spontaneous behavior showed only a weak linear correlation with TST, FST, and SPT in females, implying a sexually dimorphic relationship between naturalistic behaviors and depressive-like behaviors (Supplementary Fig. 8d, f). In addition, syllable frequency was more closely related to stress-related behaviors in TST and FST (Supplementary Fig. 8a, b, d, e) than anhedonia-like behavior in SPT (Supplementary Fig. 8c, f) in both males and females, as an analogous result has been observed in the previous study utilizing social defeat stress mice for a depressive disorder model²⁶.

We then dissected the total recording into five-minute time bins and looked for the temporal windows in which those syllables showed statistically significant changes (Fig. 3j, k). Interestingly, while only one significant stress effect (cluster 23) was found in the initial time bin, all other frequency alterations (clusters 25, 2, 23, and 18) were observed primarily in the third and fourth time bins. Given that initial cluster frequencies were similar in chronically stressed and non-stressed mice, we concluded that chronic stress disturbs the frequency of behavioral motifs in a semi-naturalistic environment only after sufficient exposure. Moreover, to investigate the “grammar” of the behavioral outcomes of stress exposure, we visualized the cluster transition matrix in

a circular diagram depicting significantly adjusted conversions probability from one cluster type to another (Fig. 3l). Even though the transition summaries revealed a robust transition shift in CUS-treated male and female mice compared to unstressed controls, there were minimal differences between stressed male and female mice, implying that chronic stress induces greater changes in behavioral grammar than does sex. Together, these results demonstrate that the CLOSER pipeline identifies sex-specific modifications in naturalistic behavior structure elicited by chronic stress.

Model-specific alterations in behavioral syntax and architecture in depressive-like behavior mouse models

To explore the characteristics of spontaneous behavior across depressive-like behavior models, we first established stress paradigms using distinct pathological triggers. We employed two depressive-like behavior models: an immune system-based model through a single intraperitoneal (i.p.) injection of lipopolysaccharide (LPS); and a steroid hormone-based model using corticosterone administration for three weeks. Given prior reports suggesting that 1-week corticosterone does not elicit depressive-like phenotypes²⁷, an additional group receiving a 1-week corticosterone injection was included to estimate the validity of the model. In the LPS-induced depressive-like phenotype model, mice exhibited hallmark stress-induced behaviors, as evidenced by increased immobility in the FST and reduced sucrose preference (Supplementary Fig. 9a). In the corticosterone paradigm, stress-related behaviors in TST and SPT emerged exclusively after 3 weeks of administration, with no significant alterations in the 1-week group (Supplementary Fig. 9b), in line with existing literature²⁷.

In the first step of behavioral analysis, we evaluated syllable frequency and transition dynamics employing LDA to interrogate the broader architecture of spontaneous behaviors across these models (Supplementary Fig. 9c). While non-stressed groups and saline-treated groups, regardless of models, clustered together, chronically stressed and LPS-treated groups diverged markedly from controls in both syllable usage and bigram transitions. Interestingly, while both stress- and inflammation-induced models displayed distinct separations, their behavioral trajectories diverged in opposing directions. In contrast, corticosterone-treated mice were closely positioned with controls, suggesting comparatively modest modifications in spontaneous behavior patterns. Then, we implemented a random forest classifier to quantify the discriminability among stress models (Supplementary Fig. 9d). Consistent with LDA findings, confusion matrices indicated that CUS and LPS models were readily separable, whereas corticosterone-induced models were less distinguishable from controls.

We next applied an analysis of individual syllable contributions to unveil model-specific alterations. LPS treatment elicited marked changes in the frequency of 16 syllables, whereas corticosterone treatment affected only three, all of which were also altered in the LPS-treated model (Supplementary Fig. 9e). A subset analysis using the Venn diagram illustrated the degree of overlap across the different mouse models (Supplementary Fig. 9f). Of note, while six syllables were commonly altered between stress- and inflammation-induced models, the directionality of syllable changes was predominantly opposite between two models, aside from one shared cluster (cluster 2). These data suggest that frequency alterations are limited to a subset of syllables and exhibit distinct patterns depending on the etiology of depressive-like behavior. These results may also account for the divergent trajectories of depressive-like behavior models observed in the LDA (Supplementary Fig. 9c). Beyond the frequency alterations, we observed pronounced disruptions in grammatical structure, particularly in the LPS model (Supplementary Fig. 9g), and few conserved transition shifts were identified across depressive-like behavior models (Supplementary Fig. 9h). These findings imply that the source of stress (stress, inflammation) can be inferred based on distinctive patterns of spontaneous behavior identified in the CLOSER platform, extending

beyond conventional dichotomous assessments of depressive-like behaviors.

Anomaly detection revealed no irregular motion in the stress mouse model

A key challenge in the psychomotor analysis of depressive disorder is the identification of abnormal behavioral syllables that only appear after exposure to chronic stress. To determine whether the CUS paradigm induces abnormal behaviors, we performed a regularity detection method with global and local pose representations to scan skeleton trajectories for anomalies. In this analysis, we utilized a recurrent neural network to calculate the frame-based anomaly scores to quantify the predictability of behavior sequences in an experimental group. We examined behavior datasets from male and female depressed mice, using the A53T Parkinson's disease model²⁸ as a positive control group with clear motor dysfunction. Consistent with our expectations, there was a clear segregation between the loss distributions of normal ($\mu = 3.844$, $\sigma = 0.789$) and A53T mice ($\mu = 5.412$, $\sigma = 0.635$) (Supplementary Fig. 10a), with discrete distribution patterns of behavior segments in embedding spaces for global and local components built with latent space-based anomaly detection²⁹ (Supplementary Fig. 10b, c). In contrast, the computational losses and embedding structures of chronically stressed mice were similar for both male (control: $\mu = 3.834$, $\sigma = 0.946$ /CUS: $\mu = 3.793$, $\sigma = 0.877$) and female groups (control: $\mu = 3.661$, $\sigma = 0.932$ /CUS: $\mu = 3.679$, $\sigma = 0.825$) (Supplementary Fig. 10d–i), indicating that most of the behavior patterns of CUS mice can be generalized as behavior sequences from normal mice. We, therefore, concluded that chronic exposure to stressors modifies the structure of free exploratory mouse behaviors rather than causing irregular behaviors.

CLOSER uncovers gradual changes in the structure of mouse exploratory behavior depending on the duration of exposure to chronic stress

Since the discrimination of stages in the appearance of depressive symptoms would provide an additional regulatory node in the clinic, we asked whether CLOSER can discriminate time-dependent behavioral phenotypes in exploratory movements generated by chronic stress. We obtained 20-minute behavioral sequences of stressed and non-stressed male mice from 3D motion videos sampled every week of the four-week CUS protocol, along with weekly assessments of depressive-like behaviors. This analysis excluded sucrose preference since its outcome is sensitive to adaptation with repetitive uptake (Supplementary Fig. 11a–d). In the fourth week, we observed clear stress-induced effects in the TST, FST, SPT, and body weight measurements of CUS mice compared to control mice. Additional tests for locomotion identified a critical change in time spent in the arena's center for the control group (Supplementary Fig. 11e, g), with a significant reduction in distance traveled regardless of the presence of chronic stimuli, suggesting a decline in exploratory instinct in familiar environments (Supplementary Fig. 11f, h). These results with conventional assays were sufficient to confirm the expression of depressive-like behaviors, but they failed to show a direct correlation with specific time points in the CUS paradigm. In other words, this type of behavioral analysis was insufficient for tracking the development of depressive-like behaviors.

To resolve this limitation, we devised a method to capture the characteristics of the behavioral repertoire of stressed mice following various durations of exposure to chronic stress. As previously described, we employed the random forest algorithm to estimate the predictive accuracy of the classifier for time points (1–4 weeks after), training it with syllable usage or with transition probability calculated from behavior sequences recorded in the AVATAR studio. The resulting confusion matrix indicated that the classifier discriminated between the experimental conditions (CTRL vs. CUS) and captured the

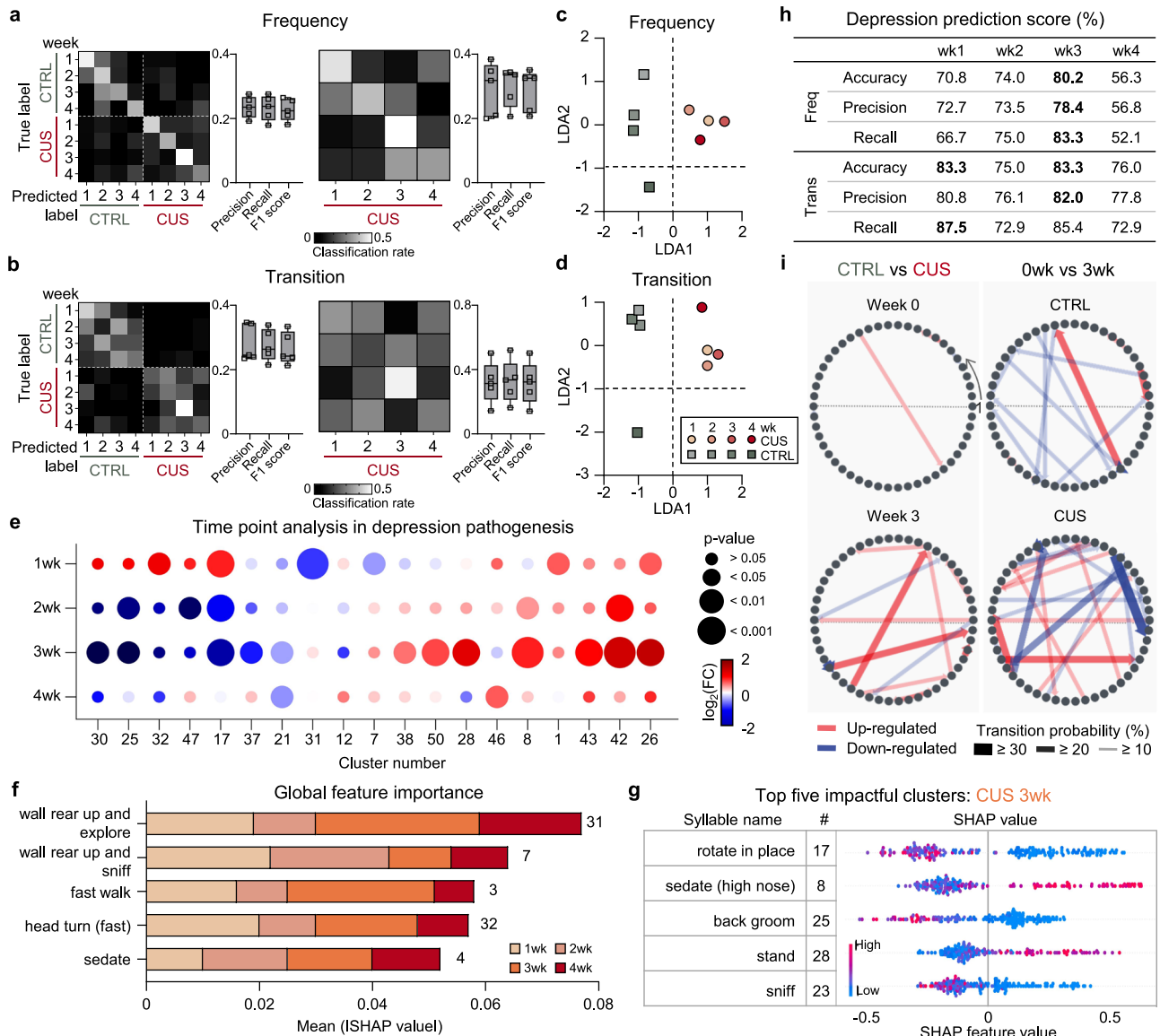


Fig. 4 | Heterogenic changes in spontaneous behaviors as depressive-like behavioral pathogenesis progresses. **a, b** Confusion matrices for the predictive performance of the random forest classifier in discriminating the timing of behavior recordings (left) or the duration of stress exposure (right) based on syllable usage (**a**) or bigram (**b**) in exploratory behavior of stressed and non-stressed mice ($N = 12$ for both groups). The classification rate represents how mice with true labels (row) were classified into predicted labels (column). For the box plots, the boxes span the interquartile range (25th–75th percentiles); center lines indicate the median, and whiskers indicate the minimum and maximum values. **c, d** The mean LDA points for syllable frequencies (**c**) and transitions (**d**) for the four experimental time points (1–4 weeks) in stressed (circle) or non-stressed (square) mice. **e** A color-coded dot plot showing chronic stress-induced changes in syllable usage across the stress paradigm, where the sizes of the circles indicate the significance level of

alteration estimated with Benjamini–Hochberg post hoc following two-way ANOVA. **f** Top five most impactful syllables in the classification of time points for stressed mice, sorted by overall absolute Shapley value. The ethogram is indicated along the y-axis, with each cluster ID presented next to its corresponding bar. **g** Beeswarm plot for syllables with the top five highest contributions in distinguishing third-week CUS mice. **h** Prediction performance of the classifier diagnosing the psychiatric state of individual mice at different time points (1–4 weeks) based on syllable frequencies (top) or transitions (bottom). The top prediction scores in accuracy, precision, or recall are indicated in bold to highlight the pathogenic features of depressive symptoms. **i** Circular network graph representing significant modifications in transition probability induced by stress (left) or the time point (right) (multiple two-tailed unpaired t -tests). Source data are provided as a Source Data file.

manifestation of behavioral change from the beginning of depressive disorder, as represented by the isolation of prediction errors within the same conditions (Fig. 4a, b, left). Likewise, we were able to diagnose experimental conditions and exposure periods through LDA1 and LDA2 axes, respectively, which were derived from information on syllable frequency (Fig. 4c) and transitions (Fig. 4d). Then, we performed an additional investigation of syllables differentially expressed throughout the various stages of stress paradigm (Fig. 4e), which indicated decisive modification in syllable usage over the course of psychiatric pathogenesis.

Notably, we observed a peak in classification performance and the largest fold change at three weeks of stress exposure (Fig. 4e), marking it as the point of the largest deviation from the normal exploration pattern. To overcome the black-box nature of the machine learning framework, we augmented the explainability of the classifier model by harnessing SHAP analysis to identify clusters that most strongly impact the model’s decision-making process (Fig. 4f, g). The SHAP summary plot characterized the behavior signature for 3-week CUS, showing increases in “stand” (cluster 28) and “sedate (high nose)” (cluster 8) and decreases in “rotate” (cluster 17), “back groom” (cluster 25), and

“sniff” (cluster 23) (Fig. 4g). Conversely, the minimal changes in cluster frequency we observed in at the final week of the CUS protocol indicated that context adaptation by means of repetitious (in our case, five) experiences reduced the behavioral disparity caused by chronic stress, considering the larger variance in syllable usage between depressed and control mice exposed to the AVATAR studio only twice (Fig. 3i). The examination of diagnosis performance for depressive disorder further supports this observation (Fig. 4h). Further analysis of the transition map revealed that three-week exposure to chronic stressors altered the transition matrix from its original structure (Fig. 4i, left). Familiarity with the open field environment strongly affected syllable bigrams in a single experimental group (Fig. 4i, right), suggesting a large contribution of context familiarity to the transition matrix of spontaneous behavior.

CLOSER can quantify the pharmacological effects of monoaminergic and non-monoaminergic antidepressants on psychomotor symptoms

To develop targeted therapeutic strategies for depressive disorders, we examined whether CLOSER captures antidepressant-induced reversal of chronic stress-related behavioral modifications, assessing distinct behavioral phenotypes in freely moving mice. We administered traditional monoaminergic antidepressants (fluoxetine and imipramine) or emerging non-monoaminergic candidates (agmatine, scopolamine, and rapastinel) to observe the distinct action of different drug groups (Fig. 5a). In chronically stressed male mice, most antidepressants significantly reduced immobility durations in FST, while imipramine alone enhanced sucrose preference rather than influencing immobility (Supplementary Fig. 13a, b). Additionally, rapastinel treatment significantly elevated sucrose preference, whereas other drugs did not induce significant improvements in either the TST or SPT. Thus, despite the high predictive validity of FST in male mice, traditional behavioral assays generally failed to differentiate individual or grouped antidepressant actions, with the exception of imipramine and rapastinel.

As the initial step of our ML-guided behavioral evaluation, we implemented a machine learning-guided behavioral evaluation to discern subtle alterations within exploratory behaviors. A random forest classifier, informed by syllable frequencies and transitions, yielded sensitivities of 40.35% and specificities of 88.00% (Fig. 5b). Incorporating transition information enhanced performance, raising sensitivity to 50.07% and specificity to 90.01%. We examined the classification errors, which revealed strong confusion within drug groups; however, inter-group and vehicle-associated misclassifications remained moderate (Supplementary Fig. 14).

Further analysis of syllable frequency subsets was performed to explore on- and off-target effects of antidepressants (Fig. 5c). Chronic stress significantly altered the usage patterns of four behavior clusters (31, 12, 49, 28). Notably, cluster 31 was consistently restored by all tested drugs, while others displayed selective restoration profiles (Fig. 5d). Imipramine and agmatine demonstrated complete recovery across stress-relevant clusters, whereas other treatments elicited partial effects. Off-target behavioral alterations were more pronounced with non-monoaminergic compounds, displaying considerable overlap within this drug class, unlike the monoaminergic group, which lacked common off-target syllables (Fig. 5c, column 12). Eight additional clusters appeared at the intersection of two non-monoaminergic drugs, emphasizing high group similarity. Utilizing SHAP analysis for deeper behavioral profiling, we identified syllables that critically influence antidepressant classification, indicative markers reflecting antidepressant heterogeneity (Fig. 5e). Different variants of “rear up” behaviors emerged prominently as discriminative features. Specifically, increased occurrences of clusters 48, 47, and 4 and decreased cluster 3 activity characterized positive predictions for non-monoaminergic antidepressants, while inverse patterns

distinguished monoaminergic treatments (Supplementary Fig. 16a). Within groups, fluctuations in cluster 19 contributed to precise drug identification. These findings underline that the CLOSER pipeline effectively leverages profound behavioral compositions of mice to distinguish psychiatric drug actions, offering potential for personalized therapeutic recommendations grounded in behavioral phenotypes.

In grammatical structure, we revealed partial normalization of stress-induced behavioral transitions following antidepressant treatment, though aberrant transitions persisted (Fig. 5f). Fluoxetine and imipramine demonstrated modest efficacy in restoring syllable bigram structures compared to the recovery in syllable frequency, accompanied by considerable off-target transition modifications (Fig. 5f, monoaminergic). Conversely, non-monoaminergic antidepressants presented comparatively lower off-target transition alterations (Fig. 5f, non-monoaminergic).

For the drug administration without stress exposure (Drug Only), the syllable frequency and transitions of spontaneous behavior were strongly discriminated from those of vehicle-treated male mice (Supplementary Fig. 17a). Consistent with observations in the stressed group, fluoxetine induced more changes in syllable frequency compared to imipramine (Supplementary Fig. 17b, c). In SHAP-based fingerprint identification, clusters 19 (“stand”) and 48 (“wall rear up”) overlapped with the top five impactful features of the stressed group. Instead, other syllables emerged as unique fingerprints of antidepressants, such as cluster 37 (“sniff and pause”) for fluoxetine (Supplementary Fig. 16b). This observation underscores the importance of pathological context in assessing the efficacy of psychiatric medications and their side effects via the monitoring of psychomotor behavior. Therefore, we have demonstrated that CLOSER can decode the hidden patterns of psychomotor symptoms into interpretable information for quantifying the pharmacological effects of antidepressants.

Sex-dependent behavioral fingerprints in naturalistic behavior for monoaminergic antidepressants

To elucidate sex-dependent antidepressant responsiveness and its impact on naturalistic behavior, we evaluated female mice with the same experimental paradigm previously applied to male mice using monoaminergic antidepressants. Unlike the male CUS model, female mice exhibited antidepressant efficacy exclusively in SPT and not in FST, highlighting substantial limitations in the predictive validity of traditional assays (Supplementary Fig. 13d, e).

To test if the behavioral structure identified by CLOSER could be applied to reliable discrimination of stress and antidepressant administration states within the female cohort (Supplementary Fig. 18a). Analysis based on syllable usage yielded a sensitivity of 47.22% and specificity of 82.41%, whereas classifiers leveraging transition profiles demonstrated enhanced performance, achieving sensitivity and specificity values of 61.11% and 87.04%, respectively, showing an analogous trend observed in males. Subsequent SHAP analyses elucidated antidepressant behavioral fingerprints, revealing significant sex-dependent differences (Supplementary Fig. 18b). In contrast to the rearing behaviors pivotal in males, female fingerprints predominantly featured head movements and sniffing behaviors, with cluster 22 exhibiting decreased frequencies following monoaminergic antidepressant treatment, while clusters 14 and 29 differentiated fluoxetine from imipramine (Supplementary Fig. 16c). Although cluster 4 displayed analogous antidepressant responsiveness across sexes, disparities in other syllable fingerprints underscored sex-specific differences in responses of naturalistic behavior to depressive disorders and their pharmacological interventions.

Next, we investigated drug-induced on- and off-target effects within naturalistic behavior syllables (Supplementary Fig. 18c). Chronic stress significantly altered five behavioral clusters; fluoxetine and

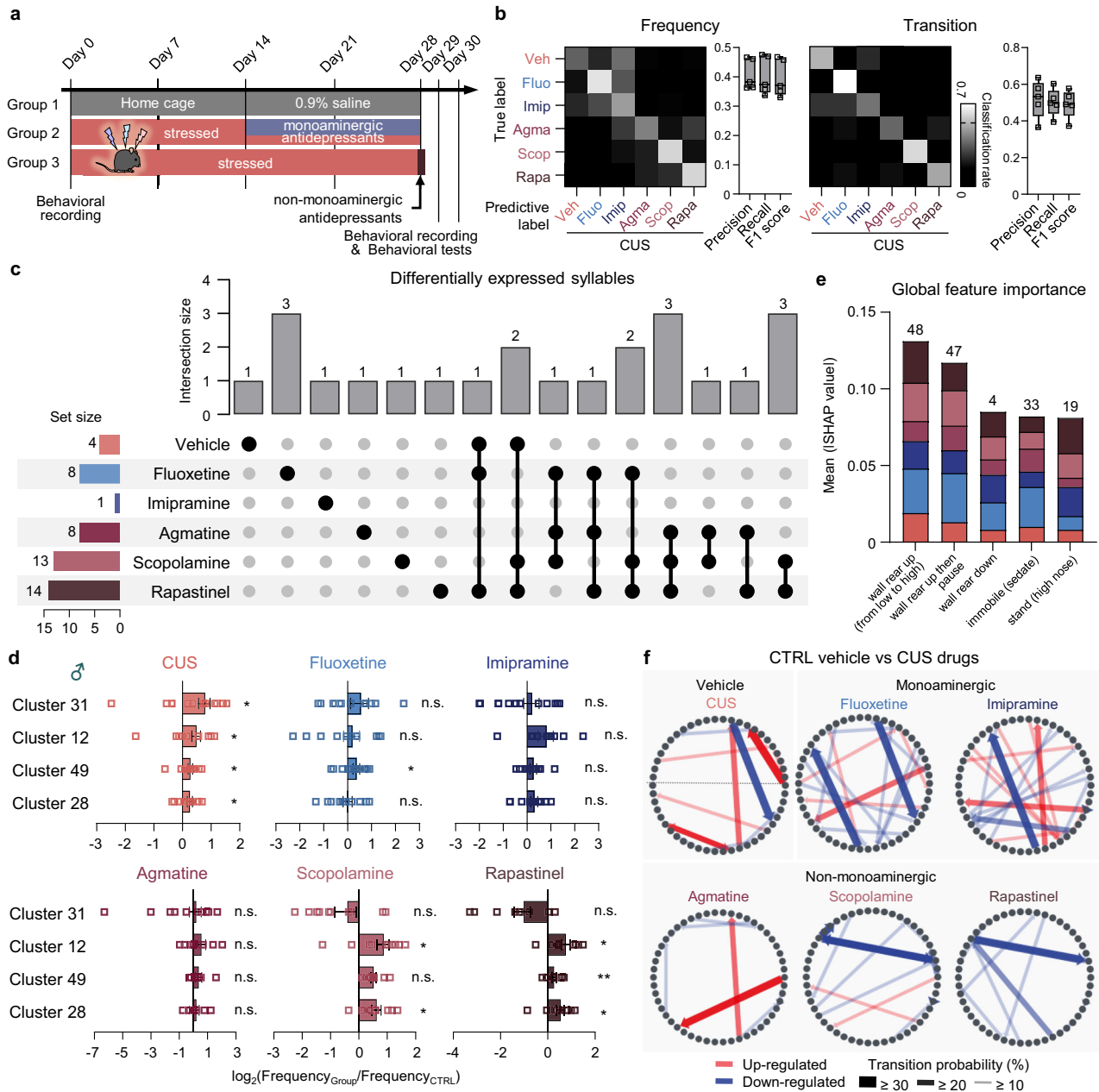


Fig. 5 | ML-guided discovery of antidepressant individuality expressed in mouse exploratory behavior. **a** Graphical scheme for the experimental timeline for vehicle- or antidepressant-treated mouse groups. **b** Confusion matrices describing stress condition and drug type classification accuracy for classifiers, utilizing syllable frequency (left) or transition probability (right). **c** An upset plot illustrating the intersections of differentially expressed syllables identified under each treatment condition. The total number of altered syllables for individual groups is shown as the set size (left). Linked black filled circles denote which groups are included in the intersection. **d** Drug-specific efficacy on stress-relevant change of syllable usage. The bar graph describes the logarithm of relative expression for stress-related syllables in mice treated with each drug; the centers indicate mean, and the error bars denote the s.e.m. (two-tailed unpaired *t*-tests between non-stressed saline-treated mice and stressed groups: CUS: cluster 31: **p* = 0.0268;

cluster 12: **p* = 0.0453; cluster 49: **p* = 0.0135; cluster 28: **p* = 0.0362, fluoxetine: cluster 49: **p* = 0.0466, scopolamine: cluster 12: **p* = 0.0423; cluster 28: **p* = 0.0194, raspastinel: cluster 12: **p* = 0.0317; cluster 49: ***p* = 0.0028; cluster 28: **p* = 0.0135, *N* = 12 for monoaminergic groups and *N* = 10 for non-monoaminergic groups). **e** Overall importance of syllables (x-axis) for the random forest classifier trained to predict administrated antidepressants. The cohort type in the bar is specified by the color of the square next to the group name in (c). **f** Modifications in syllable transition probability reveal that antidepressant treatment induced a moderate reversal of the transition shift associated with chronic stress. In box plots, the box shows the 25th–75th percentiles of the dataset; whiskers indicate the minimum and maximum values, and center lines indicate the median. Source data are provided as a Source Data file.

imipramine successfully reversed three to baseline states (Supplementary Fig. 18d). Notably, off-target effects were substantially pronounced in females, with fluoxetine altering eleven syllables and imipramine influencing nineteen. Cross-sex comparisons revealed minimal overlap between male and female responses to fluoxetine (two common syllables) and none for imipramine, underscoring sexual

dimorphism in behavioral responses due to divergent baseline behavioral structures post-stress induction. Moreover, fluoxetine consistently exhibited more off-target effects than imipramine across both sexes.

Finally, analysis of syllable bigram transitions revealed marked differences between female and male responses to antidepressants.

Fluoxetine robustly normalized stress-induced transition modifications toward non-stressed states, whereas imipramine treatment predominantly induced off-target transitions (Supplementary Fig. 18e). In conclusion, while off-target effects on syllable usage were more evident in females, off-target transitions within syllable sequences were dominant in males, further emphasizing substantial sex-dependent variability in antidepressant pharmacodynamics.

Discussion

Over several decades, researchers have measured depressive-like behavior using oversimplified tests in non-natural conditions. However, since these methods provide only all-or-none information regarding the onset of depressive disorder, they have suffered from low face validity, failing to represent the progressive nature of many mental disorders³⁰. The current study overcomes the limitation by quantitatively estimating complex, individual symptoms in exploratory behaviors recorded in a semi-natural environment. Importantly, the neurological and molecular linkages between emotion and motor activity allow it to produce high face validity for depressive symptoms appearing in spontaneous movement.

Many ML-based 3D pose estimation tools employ autoencoders, neural networks, or simple dimensionality reduction (PCA, t-SNE, or UMAP) to extract representations of animal behavior. However, considering the compositional nature of spontaneous behavior, these models are inefficient in learning the invariance of behavior elements and understanding similarity and dissimilarity in behavioral architecture. We addressed this problem by applying a self-supervised contrastive learning algorithm with augmentation modules that force the machine to embrace the underlying behavioral structure using invariance and contrast. This leads to the development of a transferable and generalizable ML toolkit. Indeed, the algorithm generated high-quality embedding for naturalistic behavior compared with recent unsupervised toolboxes, with clear recognition of latent motion patterns imperceptible on visual inspection. We also incorporated the sequential property of behavior through a majority voting strategy, an exploratory application to motion segmentation. The major advantage of this dynamic temporal segmentation is the linear time complexity, $O(n)$, of majority voting. This complexity cannot be achieved with the Hidden Markov Model (HMM), whose time complexity is $O(k^2n)$ for sequence length n and state number k . Using the method, we were able to successfully examine all possible sequences and reflect the true diversity of the behavior sequence, as evidenced by a high change point scores at the segment edges.

In this study, we demonstrated depressive-like phenotypes in mouse exploratory behaviors in an open field and subtype classification of a mood disorder. First, we found that chronic stress induces psychomotor symptoms in a sex-dependent manner, characterizing them as distinct behavioral representations and disparities in behavioral summaries. Using CLOSER, we were able to capture a stage-specific expression of motion features over the course of pathogenesis. This enabled the early detection of depressive disorder because the behavioral structure of stressed mice diverged from that of non-stressed mice beginning from the first week of the stress paradigm. Furthermore, a time bin analysis with weekly monitoring revealed the effect of spatial novelty in the diagnosis of depressive symptoms, suggesting that temporal and environmental conditions combine to evoke depressive-like phenotypes. Thus, we provide the comprehensive screening of conventional behavior tests for depressive-like behavior according to sex, etiology, stress duration, and antidepressant therapy, encompassing heterogeneity of depressive-like states (Fig. 6).

Understanding the mechanisms by which antidepressants modulate depressive-like behaviors has traditionally centered on an ideational domain^{31–34}. While this has elucidated key aspects of the neurobiology of psychiatric disorders, the interpretability of data

remains constrained by the divergence between rodent behavioral paradigms and self-reported human symptomatology. We resolved this limitation by benchmarking the pose estimation algorithm for human motion tracking to construct a generalizable diagnostic framework^{20–22}. Our framework identified the innate efficacy of each antidepressant in restoring distinct subsets of behavioral syntax and architecture and, thus, enables a subtle assessment of antidepressant effects, including quantitative evaluations of both on- and off-target actions. This insight allows for the automated prediction of drug indications tailored to the recognized symptom clusters within heterogeneous depressive disorders, offering refined therapeutic strategies with improved predictive power. A promising direction for future studies could leverage intrinsic behavioral structure under depressive states to infer treatment responsiveness, which substantially strengthens the importance of spontaneous behavior as a determinant in prescription and potential application to clinical frameworks.

Future research is necessary to validate the molecular and neurological linkages of depressive symptoms in exploratory behavior to expand the applicability of the CLOSER pipeline in neuroscience research. For instance, RNA-seq transcriptional analysis has been used to observe the sexually dimorphic modification in the expression of genes in human and mouse brains to discover the cell-type-specific molecular mechanisms underlying sex-specific stress responses³⁵. From a different perspective, neuronal studies about sexual dimorphisms in social behaviors generally focused on the intent behind behaviors, arguing that a specific neural code or type of neuron (mPOA *Esr1+* neuron, for example) governs sex-biased social behaviors³⁶. These approaches will provide an explanation for the behavioral results we observed in this study, ultimately enhancing the face validity of exploratory behavior monitoring in depression research.

For broader application of CLOSER in depressive state diagnosis, the action recognition efficiency and diagnostic accuracy must be validated in diverse recording environments and across different species and, ultimately, in human patients. It is important to note, however, that the psychomotor retardation and agitation are primarily characteristic of melancholic or catatonic depression^{5,6,37}, and the relationship between motion patterns and other subtypes, for example, atypical, psychotic, and seasonal depression, remains elusive³⁸. Thus, despite our results, which demonstrate that naturalistic behavioral profiles captured by CLOSER reveal distinct psychomotor signatures associated with stress, additional studies on the clinical relevance of psychomotor symptoms with the full spectrum of depressive subtypes are needed to prove the applicability of CLOSER and the translational potential of these behavioral markers in human depression diagnosis.

In conclusion, the CLOSER platform offers a semantically aware ML toolkit for representing naturalistic behavior and defining hidden psychomotor signatures of mood disorders in mice. We demonstrated that the improved representations of mouse 3D motion sequences our pipeline provides allow it to recognize subtypes of depressive-like symptoms from behavioral symptoms according to sex and pathogenic stage. Although the clinical applications of spontaneous behavior-based diagnosis of psychiatric diseases demand further behavioral tests in depressed human patients, we suggest our results form a strong foundation promoting the compatibility of diagnosis and drug indications between rodents and primates, with clear applicability to the field of drug discovery.

Methods

Animals

All experimental procedures were performed in accordance with the guidelines of and approved by the Institutional Animal Care and Use Committee at KAIST (approval number No. KA2024-007-v4). Wild-type C57BL/6J male and female mice were used in the behavioral

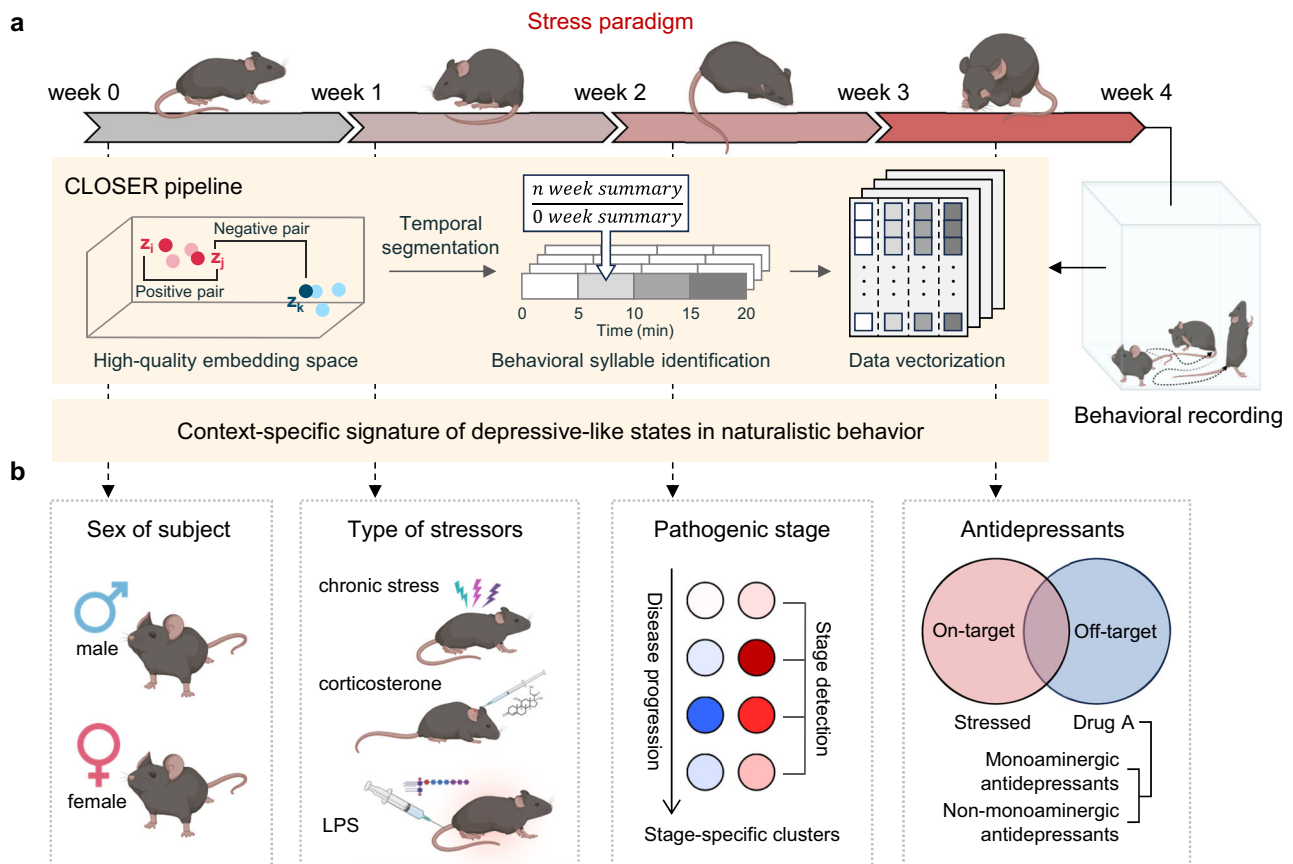


Fig. 6 | The workflow of the CLOSER framework for the discovery of latent depressive-like behavior in naturalistic movement. **a** Graphical summary of the exploratory behavior-based diagnosis of psychiatric disorders. Behavioral structures (i.e., syllable frequencies and inter-syllable transitions) were extracted from motion sequences segmented into five-minute bins to estimate the context-specific signature of depressive-like states in naturalistic behavior using various analytic methods. **b** The learned descriptors reveal sex-dependent psychomotor

phenotypes; etiology-specific effects across chronic stress, corticosterone, and LPS; and stage-specific clusters along disease progression, enabling week-by-week detection from the onset of the paradigm. The framework quantifies antidepressant actions, resolving on-target versus off-target restoration of behavioral syntax across monoaminergic and non-monoaminergic drugs. Illustrations were created in BioRender. Heo, W. (2025) <https://BioRender.com/2gfjr30>.

studies. The mice were group-housed for the first week at 7 weeks of age and then transferred at 8 weeks of age to separate cages under a 12-h light-dark cycle (lights on from 8:00 to 20:00) at room temperature (22 °C) and 40% humidity to allow them to acclimate to the housing facility. We used mice from at least three litters, and mice were randomly allocated into groups to balance potential genetic or developmental variance. Food and water were provided ad libitum.

Mouse models for depressive-like behavior

Chronic unpredictable stress (CUS) model. Wild-type C57BL/6 male or female mice (8-week-old) were subjected to nine different stressors: (i) 24 h of food and water deprivation, (ii) 8 h of cage tilting (45°), (iii) 24 h of light cycle reversal, (iv) 5 min of swimming in cold water (13 ± 1 °C), (v) 2 min of tail pinching 1 cm from the end of the tail, (vi) 8 h of moist bedding, (vii) 30 min of cold exposure (4 °C), (viii) 12 h of sleep deprivation with a cylindrical pedestal (2.5 cm diameter and 3 cm height) placed on the cage floor filled with 1 cm-deep tap water to prevent the sleep during dark phase, and (ix) 1 h of physical restraint. These stressors were randomly scheduled over four weeks, with one or two stressors performed daily in a random combination to magnify unpredictability. During the test week (week 4), we omitted the cold swimming and tail pinching to prevent repetitive exposure and habituation effects. In periodic assessments of depressive-like behaviors, behavioral tests (TST and FST) were conducted at the beginning of each week, while forced swimming and tail-pinch stressors were

excluded from the CUS paradigm to avoid confounding effects. Control mice were individually housed for four weeks with short daily handling in the housing room. We measured the mouse body weight each week at a fixed time during depressive-like behavior induction. Experimental batches and the number of mice per batch are listed in the Supplementary Table 4.

Lipopolysaccharide-induced mouse model. LPS (serotype O55:B5, Sigma-Aldrich, Cat#: L2880) was freshly prepared on the test day by dissolving in sterile 0.9% NaCl solution (normal saline). The solution was administered to 8-week-old male C57BL/6 mice with an intraperitoneal injection (i.p.) at a volume of 0.1 ml per mouse. The dose of LPS (0.83 mg/kg) was selected based on the meta-analysis for experimental studies on inflammation-associated depressive-like behavior, which is known to elicit the full spectrum of acute sickness behaviors³⁹, including physiological and behavioral changes relevant to depressive-like phenotypes⁴⁰. Control animals received an equivalent volume of sterile saline.

Corticosterone-induced mouse model. Eight-week-old male C57BL/6 mice were randomly assigned to three groups and received the following administration protocol: (i) corticosterone (20 mg/kg, Sigma-Aldrich, Cat#: D27840) for consecutive 21 days, (ii) vehicle (normal saline containing 0.1% dimethyl sulfoxide and 0.1% Tween-80) for consecutive 21 days, or (iii) vehicle for 14 days followed by

corticosterone for 7 days. All injections were delivered subcutaneously once daily at a volume of 5 ml kg⁻¹ body weight^{27,41}.

Behavioral studies

AVATAR studio recording. All mice were handled in the behavior room for 10 min per day for 3 consecutive days and habituated to the room for an hour before the behavioral experiments. AVATAR studio consists of an opaque hemispheric chamber with five recording cameras (Full-HD CMOS, 1200 × 1200 pixels, 30 frames per second) and an LED light device (20 lux) at the bottom²³. An open field apparatus, 5-mm-thick transparent acrylic panels (20 cm × 20 cm × 30 cm), is located at the center of the AVATAR studio. The mice were placed in the center of the chamber and allowed to freely explore the apparatus for 20 min. For the investigation of individual variance, we conducted recordings a day before the induction of a depressive-like state model. To estimate the effects of CUS on exploratory behaviors, the mice were recorded a day after exposure to the last stressor. From the video, the 3D coordination of nine key body points (nose, head, anus, tail tip, torso, left and right hindlimb, left and right forelimb) was calculated from multiple 2D images following a computer-vision 3D calibration and reconstruction algorithm, which calculates the intersection of the ray from the optical center of the camera through anatomical landmarks for a given camera arrangement⁴². The obtained 3D coordinate parameters were processed to generate joint connections for the formation of the action skeleton, a vectorial representation of body configuration in 3D space, on both allocentric and egocentric coordinate systems. The accuracy of AVATARnet for keypoint prediction was evaluated using three standard metrics:

$$\text{Intersection over Union (IoU)} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{\text{GT} \cap \text{Prediction}}{\text{GT} \cup \text{Prediction}} \quad (1)$$

where GT denotes the ground truth region manually labeled by human annotators,

$$\text{Average Precision (AP)} = \int_0^1 p(r) dr \quad (2)$$

where $p = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$, $r = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$ with true positives (IoU > 50%), false positives (IoU < 50%), false negatives (not detected), true negatives (no object), which estimates the degree of overlap between the predicted body part areas and actual body part areas, and Root Mean Squared Error (RMSE), which evaluates the square root for the mean square error distance for each keypoint class. AVATARnet achieved IoU consistently above 75% across nine anatomical keypoints and mAP values around 90% after sufficient network training. The positional error, measured as MSE, ranged from 7 to 15 pixels, equivalent to 1.4–4.5 mm in physical space, and the average RMSE recorded 10.19 pixels, corresponding to approximately 2–3 mm²³.

Sucrose preference test. For habituation, we provided mice with two bottles, one containing 1% sucrose (Sigma, Cat#S7903) solution and the other containing regular water alone, for 48 h before the sucrose preference test (SPT). We changed the left-right position of the bottles every 24 h in the adaptation phase to avoid any location preference. In the formal experiment, mice were allowed to freely access the 1% sucrose and water bottles for 24 h with food withholding. The positions of the bottles were switched after 12 h of testing to exclude the potential effects of side preference. After the test, we moved the mice to the home cage and measured the total intake of each fluid. Sucrose preference was calculated as follows: sucrose preference (%) = 100 × sucrose consumption (g)/(sucrose + water consumption)

(g). A minimal layer of sawdust was added to each cage to detect any bottle leakage.

Tail suspension test. Mice were suspended from a stick by their tails 50 cm above the ground for six minutes. The sticky tape was used to attach the stick to a spot 1 cm from the end of the tail⁴¹. We recorded video from the side and measured the immobility duration during 6 min of suspension using a video tracking system (Smart 3.0). Mice were considered immobile when hanging motionless, swinging passively, or remaining completely motionless.

Forced swimming test. The forced swimming test (FST) was conducted to assess the behavioral effect of stress exposure, at least four hours after the TST to minimize any potential effect of physical discomfort caused by tail suspension. A clear glass cylinder (20 cm height, 12 cm diameter) was filled to a depth of 15 cm with water (24 ± 1 °C) to prevent any physical contact of the tail or hind limbs with the bottom of the container. Mice were placed in a water-filled cylinder for six minutes. Video was recorded from the side with a web camera recording system and monitored using a video tracking system (Smart 3.0). We only quantified the duration of immobility during the final four minutes in an observer-blinded manner, as most mice move around at the beginning of the FST⁴³. We defined immobility as the status in which mice floated (motionless) or merely moved their paws to maintain balance in the water.

Time in the center and moving distance. We extracted the time spent in the arena center and the distance traveled from the AVATAR studio recording videos to measure the locomotive function of mice. In 20-min sessions, we measured the time mice stayed in the center square of the open field (10 × 10 cm) and the total distance traveled by tracking the x- and y-axis coordinates of the body center.

Pharmacological treatment of animals

Monoaminergic antidepressants. Male or female 10-week-old C57BL/6J mice were used to investigate the effects of monoaminergic antidepressants on free exploratory behaviors. All drugs were dissolved in sterile 0.9% NaCl solution (normal saline) and injected intraperitoneally for 14 consecutive days. For both fluoxetine hydrochloride (Abcam, Cat#: ab120077) and imipramine hydrochloride (Abcam, Cat#: ab141046), 15 mg kg⁻¹ was dissolved in 200 μl of vehicle. Control mice were only injected with the vehicle (normal saline). For all behavioral experiments, mice were tested 24 h after the final injection.

Non-monoaminergic antidepressants. Non-monoaminergic antidepressants were administered to 12-week-old male C57BL/6J mice exposed to chronic stress for four weeks. Agmatine sulfate salt (50 mg kg⁻¹, Sigma-Aldrich, Cat#: A7127)⁴⁴, scopolamine hydrobromide (0.5 mg kg⁻¹, Tocris Bioscience, Cat#: 1414)^{45,46}, and Rapastinel (or GLYX-13) (10 mg kg⁻¹, Tocris Bioscience, Cat#: 3406)⁴⁷ were dissolved in 0.9% saline and injected intraperitoneally at a volume of 5 ml kg⁻¹ body weight, 24 h before behavioral experiments.

Manual label assignment and cluster annotation

We established a mouse behavior database for the evaluation of our machine learning framework. Thirteen different behavior labels were assigned to each frame of a 20-min video of a freely moving mouse. Coarse behavior phenotypes were defined according to the Mouse Ethogram database taxonomy (www.mousebehavior.org). We further concretized these coarse labels with fine-grained labels to provide more detailed discrimination of the spontaneous motions of well-known behavior sequences (Supplementary Table 2). After this process, each annotated behavior segment contained information about the start/end frame number and a single fine-grained label. To annotate the unsupervised learning-driven clusters with behavior

components, experts explored twenty corresponding video clips for each cluster and determined if the behavior sequences in the video clip could be associated with fine-grained labels. The behavioral labels that appeared in the video clip were scored and averaged as part of a numerical evaluation of cluster characteristics. We then summarized the features of the behavioral phenotypes (a maximum of two) with more subtle descriptions.

CLOSER Framework

Data augmentation. Contrastive learning requires an appropriate data augmentation strategy to maximize the performance of the machine learning model. We introduced four augmentation methods: “Scale”, “ChangeStartPoint”, “ChangeAngle”, and “ChannelMask”.

“Scale” is accomplished by random scaling of the original mouse 3D skeleton vector size by a factor of 0.5 to 1.5. For the given motion sequence $f_t^i = (x_t^i, y_t^i, z_t^i)$, the augmented motion sequence was defined as

$$\text{Scale}(f_t^i) = \alpha * (x_t^i, y_t^i, z_t^i) \tag{3}$$

while $0.5 \leq \alpha \leq 1.5$. Scale enables the model to learn the irrelevance of mouse size in its behavior. “ChangeStartPoint” simulates the temporal heterogeneity of behavior sequences that originates in variations in the duration of the motion segments. To accomplish this, we randomly cropped 60% of the original motion sequence and then performed a linear interpolation to the original length according to this relationship:

$$\text{ChangeStartPoint}(f_{1..T}^i) = \text{Interpolate}(f_{s..e}^i), 1 \leq s, e \leq T, e - s + 1 = 0.6 * T \tag{4}$$

We employed “ChangeAngle” to distort the angle between the skeleton vectors of a chosen joint and its parent skeleton vectors with a maximum distortion of 1.75 radians. Three joints were selected randomly. The angles between the skeleton vectors v_1 involving the chosen joints and their parent skeleton vectors v_2 were calculated using the following equation:

$$\theta = \text{acos} \frac{\sum v_1 \cdot v_2}{|v_1| |v_2|}; \theta^{\text{new}} = \theta + \Delta, -30^\circ \leq \Delta \leq 30^\circ \tag{5}$$

For the investigation of new cartesian coordinates X , we built a linear system as follows:

$$\text{Concat}(n, v_1, v_2)X = \text{Concat} \left(d, \sum v_1 \cdot p_2 + |v_1| |v_2| \cos(\theta^{\text{new}}), \sum v_2 \cdot p_2 + |v_1| |v_2| \cos(\theta^{\text{new}} - \theta) \right) \tag{6}$$

with $n = v_1 \times v_2$ and $d = \sum n \cdot \text{joint}$, where joint denotes the cartesian coordinates of the chosen joints. p_1 and p_2 refer to the parent joints of the chosen joints. Similarly, we rotated the skeleton vectors with the chosen joints orthogonal to the plane containing v_1 and v_2 by at most 0.875 radians. Through this augmentation, the model learned the angular difference in key joints. “ChannelMask” is a random mask of the x- or y-axis with a 25% probability of implementing independence of behavior contents on a specific axis. For example, if the y-axis is masked in a motion sequence $f_t^i = (x_t^i, y_t^i, z_t^i)$,

$$\text{ChannelMask}(f_t^i) = (x_t^i, 0, z_t^i) \tag{7}$$

To validate augmentation combinations through ablation analysis, we introduced existing augmentation strategies widely used in human action recognition. For the first augmentation combination²⁰, we adopted two spatial skeleton augmentations (“Shear” and “JointJittering”) and a single temporal skeleton augmentation (“Temporal Crop-resize”). The “Shear” transformation gave variations in viewpoint

and distance to the camera with pose conservation. In more detail, for the segment $X \in \mathbb{R}^{\text{batch_size} \times \text{num_joints} \times \text{joint_size}}$, the modified action representation

$$\text{Shear}(X) = X \cdot \begin{bmatrix} 1 & r_{01} & r_{02} \\ r_{10} & 1 & r_{12} \\ r_{20} & r_{21} & 1 \end{bmatrix} \tag{8}$$

“JointJittering” inserts a perturbation in joint connections through the irregular movement of random joints to teach the model invariance to noise in joint positions. To achieve this, we multiplexed three randomly selected joints j , with a jitter matrix as

$$\text{JointJitter}(X) = X[:, j, :] \cdot \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix} \tag{9}$$

All entries in the shear and jitter matrices are randomly drawn from a uniform distribution $[-1, 1]$. Since “Temporal Crop-resize” adjusts a similar distortion to the “ChangeStartPoint” method, described above, we utilized our augmentation module without modification.

In a second augmentation strategy, we integrated augmentation methods proposed for human depression detection⁴⁸: “Shear”, “Rotate”, “GaussianNoise”, “JointMask”, and “ChannelMask”. In “Rotate”, we randomly rotated a 3D action skeleton with respect to the x-, y-, and z-axes with random variables $\alpha, \beta, \gamma \in \mathbb{R}^{\text{batch_size} \times 1 \times \text{joint_size} \times \text{joint_size}}$:

$$\text{Rotate}(X) = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix} \cdot X \tag{10}$$

“GaussianNoise” adds appropriate Gaussian noise to the action sequence. We applied the Gaussian noise matrix $\varepsilon \in \mathbb{R}^{\text{batch_size} \times 1 \times \text{num_joints} \times \text{joint_size}}$, whose entries are taken from a Gaussian distribution with mean of 0 and a standard deviation of 1, to the action segment $X \in \mathbb{R}^{\text{batch_size} \times \text{input_length} \times \text{num_joints} \times \text{joint_size}}$ as

$$\text{GaussianNoise}(X) = X + \varepsilon \tag{11}$$

“JointMask” generates positive pairs by concealing three random joints by

$$X[:, :, j, :] = 0 \tag{12}$$

Model training. For action representation learning, we employed a self-supervised contrastive learning framework based on Cross-modal Mutual Distillation (ST)⁴⁹, following the structure of Spatio-Temporal Graph Convolutional Network (ST-GCN) building encoders⁵⁰. A given full-length motion sequence obtained from AVATAR studio was divided into segments with a single motion segment unit of 10, introducing a 1-frame stride to provide generalization in input action sequences. The dataset consisted of 120 label-free videos of control and depressive-like behavior model mice for 4 weeks, and each video contains a whole 36,000 frame motion sequence of freely moving mice. In the CLOSER framework, training, validation, and test datasets are generated with a ratio of 60:20:20 while randomly sorting motion sequences. We used encoders composed of 9 layers of the spatio-temporal graph convolution operators; 3 layers of 32 output channels, 3 layers of 64 output channels, and 3 layers of 128 output channels stacked in order. This implies that the encoder is designed to map the input space of $R^{10 \times 27}$ to an embedding space of R^{128} .

In the CMD modality design, we implemented two approaches, joint movement and motion change. Joint modality represents Cartesian coordinates of keypoints from a “fixpose” motion sequences, and motion modality defines a difference in Cartesian coordinates of

keypoints between adjacent frames from “pose” motion sequences. For the projection head module, linear layers of input dimension $I \subset R^{128}$ and output dimension $O \subset R^{128}$ without any activation function. The model was trained to optimize InfoNCE (Info Noise-Contrastive Estimation) loss⁵¹ and Kullback–Leibler (KL) divergence. For InfoNCE loss L_{SCL} and CMD loss L_{CMD} , the total loss function.

$$L = L_{SCL}^{joint} + L_{SCL}^{motion} + L_{CMD} \tag{13}$$

was calculated. For the size of a queue-based memory bank N and a temperature hyper-parameter τ_c , InfoNCE loss for joint and motion modality is defined as

$$L_{SCL} = -\log \frac{\exp\left(\frac{z_q^T z_k}{\tau_c}\right)}{\exp\left(\frac{z_q^T z_k}{\tau_c}\right) + \sum_{i=1}^N \exp\left(\frac{z_q^T m_i}{\tau_c}\right)} \tag{14}$$

where z_q and z_k represents the query and key embedding, respectively. The knowledge distillation was achieved by minimizing Kullback–Leibler (KL) divergence of similarity distributions

$$L_{CMD} = KL\left(p(z_k^{joint}, \tau_t) \parallel p(z_k^{motion}, \tau_s)\right) + KL\left(p(z_k^{motion}, \tau_t) \parallel p(z_k^{joint}, \tau_s)\right) \tag{15}$$

$$p_i(z, \tau) = \frac{\exp\left(\frac{z^T n_i}{\tau}\right)}{\sum_{i=1}^K \exp\left(\frac{z^T n_i}{\tau}\right)}, i = 1, 2, \dots, K \tag{16}$$

where τ_t and τ_s are a teacher and a student asymmetric temperature, respectively. All hyperparameters used to optimize the CLOSER framework are listed in Supplementary Table 3.

Temporal segmentation. For consideration of temporal dynamics in behavior sequences, we modified the majority vote algorithm to split sequential frames into semantic motion segments. An action representation z_i was computed by

$$z_i = ST - GCN^{joint}(\text{segment}_i) \tag{17}$$

for i^{th} segment segment_i and the projection head-free query encoder $ST - GCN^{joint}$ for joint modality. In dynamic temporal segmentation, we define the set S_i as

$$S_i = \left\{ \text{segment}_j \mid |i - T + 1 \leq j \leq i + T - 1 \right\} \tag{18}$$

where T is the segment length, which is set to 10. We labeled each segment in S_i with k-means clustering in the embedding space, then assigned the most frequent frame-level label.

$$\text{Label}(\text{frame}_i) = \text{argmax}_{s \in S_i} \left\{ \text{Label}\left(ST - GCN^{joint}(s)\right) \right\} \tag{19}$$

The consecutive frames with identical labels were paired to represent a behavioral unit. In addition, we employed a keypoint change score function⁵² for further estimation of the ability of the framework to capture the robust shift in the action skeleton.

Linear regression

The linear correlation of conventional depressive-like behaviors and spontaneous behavior was driven by fitting a linear regression model that predicts the immobility time (s) in TST and FST or sucrose preference (%) in SPT using the fold change of behavioral syllables. For a given variable X_i and observations y , the linear regression model

calculates coefficients β_i and residual β_0 by minimization of the following equation.

$$L = \frac{1}{2} (y - \hat{y})^2 \tag{20}$$

where $y = \beta_0 + \sum_i \beta_i X_i$. Then, we estimated the coefficient of determination, the R^2 score, for the numerical expression of predictability for the dependent variable from discriminant-relevant syllables.

Anomaly detection

Model configuration. To estimate the behavioral abnormalities of the disease model mice, we employed a Message-Passing Encoder-Decoder Recurrent Neural Network (MPED-RNN) model⁵³. At each time point t , we decomposed mouse 3D pose representations $f_t^i = (x_t^i, y_t^i, z_t^i)$ into a global component f_t^g and a local component $f_t^{l,i}$, where $f_t^i = f_t^g + f_t^{l,i}$ for each i^{th} -key joint. The global component represents a motion of the body center, implying a change in the absolute position of the 3D pose representation. It is modeled as the combination of the coordinates of the body center and the body’s bounding cuboid. Hence, we defined the global components as $f_t^g = (x^g, y^g, z^g, w, l, h)$. Each feature of the global component was calculated with the following equations:

$$x^g = \frac{\max(x^i) + \min(x^i)}{2}; y^g = \frac{\max(y^i) + \min(y^i)}{2}; z^g = \frac{\max(z^i) + \min(z^i)}{2} \tag{21}$$

$$w = \max(x^i) - \min(x^i); l = \max(y^i) - \min(y^i); h = \max(z^i) - \min(z^i) \tag{22}$$

We defined the local component as a relative motion of key joints with respect to the body center. It was calculated from the global component as follows:

$$x^{l,i} = \frac{x^i - x^g}{w}; y^{l,i} = \frac{y^i - y^g}{l}; z^{l,i} = \frac{z^i - z^g}{h} \tag{23}$$

We introduced encoder Gated Recurrent Units (GRU) modules and decoder GRU modules for both the global and local components to implement MPED-RNN. Each module sends messages to its counterpart and receives messages for the calculation of the latter state. The messages exchanged in the module with current state h were computed as specified below, where σ denotes the activation function, W represents the weight, and b is the bias term:

$$\text{message}_t^{l \rightarrow g} = \sigma\left(W^{l \rightarrow g} h_{t-1}^l + b_t^{l \rightarrow g}\right) \tag{24}$$

$$\text{message}_t^{g \rightarrow l} = \sigma\left(W^{g \rightarrow l} h_{t-1}^g + b_t^{g \rightarrow l}\right) \tag{25}$$

Then, to generate training segments, we used a utilized sliding-window strategy as follows:

$$\text{segment}_i = \{f_t\}_{t=b_i..e_i} \tag{26}$$

for $b_i = s \times i, e_i = s \times i + T$, where the sliding stride $s=10$ and the segment length $T=60$. In model training, we measured reconstruction and prediction loss using the loss function suggested by Morais, Romero, et al.:

$$L_s(\text{segment}_i) = \frac{1}{2} \left(\frac{1}{T} \sum_{t=b_i}^{e_i} \|f_t^* - f_t^*\|_2^2 + \frac{1}{P} \sum_{t=e_i+1}^{e_i+P} \|f_t^* - f_t^*\|_2^2 \right) \tag{27}$$

where P represents prediction length, which was set to 30, and $*$ denotes either l or g. The training and validation data sets were split into an 80:20 ratio.

To calculate the frame $_i$ anomaly score $\text{anomaly_score}(\text{frame}_i)$, we used the following equation:

$$\text{anomaly_score}(\text{frame}_i) = \frac{\sum_{\text{segment} \in S_i} \sum_{\{l, g\}} L(\text{segment})}{T} \quad (28)$$

where $S_i = \{\text{segment}_j | i - T + 1 \leq j \leq i + T - 1\}$ denotes the set of segments containing frame_i . For simplicity, we only considered frames belonging to T segments and analyzed the distribution of frame-wise anomaly scores for the evaluation of abnormality in the disease group compared to its control group.

Visualization. To visualize the embedding spaces of the global and local components, we leveraged an unsupervised deep anomaly detection approach called LatentOut, which was developed by Fabrizio Angiulli et al.²⁹. With the mean μ and standard deviation for an anomaly score distribution σ , we normalized anomaly scores with the following equation:

$$\text{anomaly_score}^{\text{normalized}}(\text{frame}_i) = \frac{\text{anomaly_score}(\text{frame}_i) - \mu}{\sigma} \quad (29)$$

Then, we projected the results for the embedding structure of global and local components using PCA onto a 3D coordinate system with the z-axis defined as

$$z(\text{frame}_i) = \frac{\sum_{\text{frame} \in N_k(\text{frame}_i)} \text{anomaly_score}^{\text{normalized}}(\text{frame})}{k} \quad (30)$$

where $N_k(\cdot)$ is the k-nearest-neighbor using default parameters ($k = 5$). In the nearest neighbor module, distance was defined as Euclidean distance between projected cartesian coordinates and the normalized anomaly score.

Hierarchical clustering

To estimate the overall clustering structure of the CLOSER framework, we performed a hierarchical recategorization of the embedding. An average embedding z_c for behavioral cluster c was measured according to

$$z_c = \frac{\sum z_c^i}{N} \quad (31)$$

where z_i is the embedding information of the i^{th} segment in cluster c , and N is the number of segments in cluster c . We leveraged the Scikit-learn (sklearn) module for agglomerative clustering to group clustered objects based on similarity using complete linkage.

Dimensionality reduction

Principal component analysis (PCA). To extract the features of high-dimensional behavioral sequences using variance-maximizing projection⁵⁴, we employed the Scikit-learn implementation of PCA with 13 components, with default parameters: (i) $\text{whiten} = \text{False}$, which preserves the original variance scales, (ii) $\text{svd_solver} = \text{'auto'}$, which determines the most efficient singular value decomposition strategy based on the input dimensions and the chosen number of components.

Uniform manifold approximation and projection (UMAP). We introduced UMAP to perform nonlinear dimensionality reduction, with the preservation of the topological structure of behavioral data within a low-dimensional embedding⁵⁵. We leveraged the UMAP implementation in 13-dimension, with default settings: (i) $n_{\text{neighbors}} = 15$, (ii)

$\text{min_dist} = 0.1$, and (iii) $\text{metric} = \text{'euclidean'}$ in pairwise distance measurements.

Benchmarking

To estimate the quality of the embedding precisely, we compared it to that of existing motion analysis frameworks, VAME and keypoint-MoSeq.

Variational animal motion embedding (VAME). For the VAME, we modified the input structure to accept 3D behavioral segments as input through the extension of the number of features, from 12 to 27, and the embedding dimension, from 30-dimension to 108-dimension. In model training and inference, we adjusted the frame lengths of reconstruction and prediction methods to ten and five frames, respectively. All 3D pose trajectories were concatenated across animals and conditions and then windowed into 10-frame segments. Each segment was normalized and fed into the VAME encoder in batches of 4096 segments. From each batch, the VAME encoder produced a 4096×108 array of latent means (μ), followed by the concatenation of all batches to reconstruct a full embedding matrix. As the variational autoencoder architecture of the VAME consists of an encoder and a decoder, each using GRUs, we trained the model with the given parameters for reconstruction, prediction, and KL divergence, which shows the maximal performance in behavioral classification. For the i -th segment segment_i , we generated the action representation.

$$z_i = \text{VAME}^{\text{encoder}}(\text{segment}_i) \quad (32)$$

We applied both a Hidden Markov Model and k-means clustering after majority voting to assign frame-wise cluster labels.

Keypoint-MoSeq. For the keypoint-MoSeq, we configured the model to align with its key joints and connections, maintaining consistency with the variational autoencoder structure. The 3D pose trajectories were loaded and directly subjected to PCA, while omitting the noise-calibration step, which had been applied in AVATAR calibration. We then initialized and fit an autoregressive hidden Markov model on the PCA projections, followed by fitting the MoSeq model. Calculated inferred state sequences were exported to obtain frame-wise behavioral syllable labels for downstream analysis.

The details of hyperparameters for both VAME and keypoint-MoSeq are listed in the Supplementary Information (Supplementary Table 5).

Investigation of representative behavior phenotypes

A behavior segmentation system requires the ability to cluster motion sequences into discrete modules of action for further ethological interpretation. Hence, we inspected exemplary behavior phenotypes (i.e., walking, turning, rearing, and falling) from automatically categorized motion segments to identify the semantic implementation of the framework. For behavior segments longer than 10 frames, we equally divided the segments into 10 sections and averaged each metric over a single section to summarize the behavioral features on an identical time scale. The set of frames F_n involved in cluster n was defined as

$$F_n = \left\{ (x_i^j, y_i^j, z_i^j) \mid j \in \text{Joint} \mid s \leq i \leq e \right\} \quad (33)$$

with start frame s and end frame e for key joints Joint.

The i -th behavior segment in “walking” clusters was evaluated with moving distance d_i , where $\text{dist}(\cdot, \cdot)$ denotes the Euclidean distance between two body points from consecutive time sections t and $t - 1$.

$$d_i = \frac{\sum \text{dist}((x_t^j, y_t^j, z_t^j), (x_{t-1}^j, y_{t-1}^j, z_{t-1}^j))}{N}, \quad (34)$$

$$N = \frac{e-s+1}{10}, s + (i - 1)N + 1 \leq t \leq s + iN$$

To measure the degree of “turning”, we measured the angle

$$a_i = \frac{\sum \text{angle}(v1_t, v2_t)}{N}; N = \frac{e-s+1}{10}; s+(i-1)N+1 \leq t \leq s+iN \quad (35)$$

between the anus-body center vector, $v1$, and the body center head vector, $v2$, where $\text{angle}(v1, v2) = \text{acos} \frac{v1 \cdot v2}{|v1||v2|}$. For the qualification of behavior segments in “rearing” and “falling”, we measured the height of the nose as follows:

$$h_i = \frac{\sum z_t^{\text{nose}}}{N} \quad (36)$$

for $N = \frac{e-s+1}{10}, s+(i-1)N+1 \leq t \leq s+iN$.

Normalization

We normalized the frequency and transition probabilities for the behavioral clusters using the initial behavior structure while considering individual variance in the exploratory behavior paradigms throughout the subjects. We implemented fold change of a cluster frequency $\text{Freq}_{\text{Norm}}$ and transition probability p_{Norm} for the k -th segment x_k :

$$\text{Freq}_{\text{Norm}}(c_i, \text{week}_m, \text{mouse}_n) = \frac{|\{x_k | x_k \in \text{week}_m \cap \text{mouse}_n, \text{Label}(x_k) = c_i\}|}{|\{x_k | x_k \in \text{week}_0 \cap \text{mouse}_n, \text{Label}(x_k) = c_i\}|} \quad (37)$$

$$p_{\text{Norm}}(c_i \rightarrow c_j, \text{week}_m, \text{mouse}_n) = \frac{|\{(x_{k-1}, x_k | x_{k-1}, x_k \in \text{week}_m \cap \text{mouse}_n, \text{Label}(x_{k-1}) = c_i, \text{Label}(x_k) = c_j)\}|}{|\{(x_{k-1}, x_k | x_{k-1}, x_k \in \text{week}_0 \cap \text{mouse}_n, \text{Label}(x_{k-1}) = c_i, \text{Label}(x_k) = c_j)\}|} \quad (38)$$

Linear discriminant analysis (LDA)

To calculate the class-separability through a dimensionality reduction, we projected the training configuration of LinearDiscriminantAnalysis (in sklearn) results into the plane

$$(x, y)^{\text{projected}} = \text{LDA}(\text{ST} - \text{GCN}^{\text{joint}}(x_i)) \quad (39)$$

for labeled data $\{x = (\frac{t_{s+e}-t}{2}, \frac{t_{s+e}+t}{2}), y = c\}$.

Pearson correlation coefficient

To evaluate the linear relationships between embedding vectors, we calculated the Pearson Correlation Coefficient (PCC) by dividing the covariance of the variables by the product of their standard deviations, yielding a value between -1 and 1, as follows:

$$r_{xy} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (40)$$

where X and Y are latent vectors, μ_X and μ_Y are respective means, and σ_X and σ_Y are standard deviations.

k-nearest neighbor classifier

For quality evaluation of latent embedding, we employed a non-parametric, k-Nearest Neighbor (k-NN) classifier that predicts the label of an input instance by identification of the k closest labeled neighbors within the feature space and assignment of label using majority voting. We utilized the KNeighborsClassifier from the Scikit-learn library with default parameters: five neighbors ($N_{\text{neighbors}} = 5$) with uniform voting weights, using the Minkowski distance ($p = 2$) corresponding to the Euclidean metric. The classifier was configured to automatically select an optimal search algorithm (algorithm = 'auto') with a leaf size of 30, without additional metric parameters.

Random forest classifier

To verify the behavioral correlation between free movement and depressive-like behaviors, we obtained the confusion matrix for control and depressive-like behavior model mice utilizing the Random Forest Classifier (from sklearn) with default settings. The Random Forest Classifier was trained with fivefold validation. To prepare the dataset, we partitioned 20-min sample recording videos into 5-min sections and calculated fold changes for cluster frequencies and transition probabilities as described in “Normalization”. To validate the size invariance of CLOSER and other benchmarking algorithms, eighty male and female mice were divided into five weight categories, spanning $18 \times g$ to $28 \times g$ in $2 \times g$ increments. The Random Forest classifier was trained to predict the weight categories from the embedding vectors extracted in non-overlapping 100-frame strides from behavioral sequences. We trained the classifier with 5-fold cross-validation in the default setting and evaluated the performance using precision, recall, and F1 score.

SHapley additive exPlanations (SHAP) analysis

To investigate the influence of behavioral cluster frequencies on classifier decisions, we applied a SHAP analysis⁵⁶ using the scikit-learn library. Initially, we trained a RandomForestClassifier without employing k-fold validation followed by TreeSHAP⁵⁷, which defines the value function through conditional expectation to extract SHAP values. Given the feature set N , a subset S , the classifier function f , and the input x , $f_S(x)$ denotes the classifier output restricted to S , which can be further described as the conditional expectation $f_S(x) = E[f(x)|x_S]$. The SHAP value for the i -th feature was then calculated as follows.

$$\phi_i = \sum_{S \subset N(i)} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (f_{S \cup \{i\}}(x) - f_S(x)) \quad (41)$$

Leveraging the additivity of the SHAP value, average absolute SHAP values for overall feature importance were calculated and depicted as cumulative bar graphs, while SHAP values for individual groups were represented using beeswarm plots.

Transition probability

The transition probability from behavioral cluster c_i to behavioral cluster c_j was generated from the conditional probability of the $\text{Label}(x_k) = c_j$ transition given $\text{Label}(x_{k-1}) = c_i$, where $\text{Label}(x_k)$ defines the behavioral cluster of the k -th segment x_k :

$$p_{c_i \rightarrow c_j} = \frac{|\{(x_{k-1}, x_k) | \text{Label}(x_{k-1}) = c_i, \text{Label}(x_k) = c_j\}|}{|\{(x_{k-1}, x_k) | \text{Label}(x_{k-1}) = c_i\}|} \quad (42)$$

To compute the change in transition probability for a behavioral cluster, we estimated the difference of each transition probability and visualized leveraging the network analysis framework NetworkX. Only transitions whose alteration was significant ($p < 0.05$) were displayed in transition graphs.

Statistics

Statistical analysis and graphs were produced in Python (3.10.12) and GraphPad Prism 10. Time-series datasets were analyzed by two-way ANOVA with Geisser-Greenhouse corrections. Post hoc analysis for multiple comparisons leveraged the Benjamini-Hochberg test to control the false discovery rate, which we fixed at 0.05. For the estimation of significance in groups of more than two, we estimated one-way ANOVAs with a 95% confidence interval and Dunnett’s post hoc analysis for multiple comparisons to the control group. Two-group comparisons, including those for syllable usage and syllable subset analysis, were performed using two-tailed unpaired t -tests (parametric) if the group passed the Anderson-Darling (AD) normality test. Otherwise, we used a Mann-Whitney test (non-parametric) to deal

with a non-Gaussian distribution. Classification performance of k-NN and random forest classifiers was summarized with average values and standard deviation of precision, recall, and F1 scores. Box graphs show the 25th–75th percentiles for the whole dataset with a line indicating the median. Bar graphs and timelines present means \pm standard error of the mean. Asterisks denote levels of statistical significance as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, and **** $p < 0.001$. More detailed descriptions are summarized in Supplementary Data 1, with statistical significance marked with yellow blocks.

Reporting summary

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

Data availability

Details for experimental batches and sample sizes for all experiments are provided in Supplementary Table 4. Statistical tests are clarified in Supplementary Data 1. The mouse 3D action skeleton data associated with the study are available at <https://zenodo.org/records/13585818> (<https://doi.org/10.5281/zenodo.13585818>)⁵⁸. Source data are provided with this paper.

Code availability

The open-source code and documentation of the CLOSER framework are available at GitHub (<https://github.com/ChoiSian/CLOSER-for-tailored-diagnosis-of-depression-in-mice>), Zenodo (<https://doi.org/10.5281/zenodo.13585818>), and from the corresponding author on request.

References

- Organization, W. H. *Depression and Other Common Mental Disorders: Global Health Estimates* (World Health Organization, 2017).
- Lang, U. E. & Borgwardt, S. Molecular mechanisms of depression: perspectives on new treatment strategies. *Cell Physiol. Biochem.* **31**, 761–777 (2013).
- Schrijvers, D., Hulstijn, W. & Sabbe, B. G. C. Psychomotor symptoms in depression: a diagnostic, pathophysiological and therapeutic tool. *J. Affect Disord.* **109**, 1–20 (2008).
- Gupta, R. K. Major depression: an illness with objective physical signs. *World J. Biol. Psychiatr.* **10**, 196–201 (2009).
- Buyukdura, J. S., McClintock, S. M. & Croarkin, P. E. Psychomotor retardation in depression: Biological underpinnings, measurement, and treatment. *Prog. Neuro-Psychoph* **35**, 395–409 (2011).
- Bennabi, D., Vandel, P., Papaxanthis, C., Pozzo, T. & Haffen, E. Psychomotor retardation in depression: a systematic review of diagnostic, pathophysiological, and therapeutic implications. *Biomed. Res. Int.* **2013**, 158746 (2013).
- Petit-Demouliere, B., Chenu, F. & Bourin, M. Forced swimming test in mice: a review of antidepressant activity. *Psychopharmacology* **177**, 245–255 (2005).
- Shemesh, Y. & Chen, A. A paradigm shift in translational psychiatry through rodent neuroethology. *Mol. Psychiatr.* **28**, 993–1003 (2023).
- Trunnell, E. R. & Carvalho, C. The forced swim test has poor accuracy for identifying novel antidepressants. *Drug Discov. Today* **26**, 2898–2904 (2021).
- Huys, Q. J. M., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).
- Allen, G., Buxton, R. B., Wong, E. C. & Courchesne, E. Attentional activation of the cerebellum independent of motor involvement. *Science* **275**, 1940–1943 (1997).
- Takakusaki, K. Functional neuroanatomy for posture and gait control. *J. Mov. Disord.* **10**, 1–17 (2017).
- Schraa-Tam, C. K. L. et al. fMRI Activities in the emotional cerebellum: a preference for negative stimuli and goal-directed behavior. *Cerebellum* **11**, 233–245 (2012).
- Zhang, X. Y. et al. A role for the cerebellum in motor-triggered alleviation of anxiety. *Neuron* **112**, 1165–1181 e1168 (2024).
- Sheppard, K. et al. Stride-level analysis of mouse open field behavior using deep-learning-based pose estimation. *Cell Rep.* **38**, 110231 (2022).
- Gschwind, T. et al. Hidden behavioral fingerprints in epilepsy. *Neuron* **111**, 1440–1452 (2023).
- Wiltshcko, A. B. et al. Revealing the structure of pharmacobehavioral space through motion sequencing. *Nat. Neurosci.* **23**, 1433–1433 (2020).
- von Ziegler, L., Sturman, O. & Bohacek, J. Big behavior: challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacol* **46**, 33–44 (2021).
- McCullough, M. H. & Goodhill, G. J. Unsupervised quantification of naturalistic animal behaviors for gaining insight into the brain. *Curr. Opin. Neurobiol.* **70**, 89–100 (2021).
- Thoker, F. M., Doughty, H. & Snoek, C. G. M. Skeleton-contrastive 3D action representation learning. In *Proceedings of the 29th Acm International Conference on Multimedia, Mm 2021* (eds Heng Tao Shen et al.), 1655–1663 (2021).
- Yang, X. et al. Hierarchical contrastive motion learning for video action recognition. Preprint at <https://arxiv.org/abs/2007.10321> (2020).
- Lin, L. L., Song, S. J., Yang, W. H. & Liu, J. Y. MSL: multi-task self-supervised learning for skeleton based action recognition. In *Mm '20: Proceedings of the 28th ACM International Conference on Multimedia* (eds Chang Wen Chen et al.) 2490–2498 (2020).
- Kim, D.-G., Shin, A., Jeong, Y.-C., Park, S. & Kim, D. AVATAR: AI vision analysis for three-dimensional action in real-time. Preprint at <https://www.biorxiv.org/content/10.1101/2021.12.31.474634v1> (2022).
- Marks, M. et al. A closer look at benchmarking self-supervised pre-training with image classification. Preprint at <https://arxiv.org/abs/2407.12210> (2024).
- Zhao, S., Zhang, B., Yang, J., Zhou, J. & Xu, Y. Linear discriminant analysis. *Nat. Rev. Methods Prim.* **4**, 70 (2024).
- Tseng, Y. T. et al. Systematic evaluation of a predator stress model of depression in mice using a hierarchical 3D-motion learning framework. *Transl. Psychiatry* **13**, 178 (2023).
- Zhao, Y. et al. A mouse model of depression induced by repeated corticosterone injections. *Eur. J. Pharm.* **581**, 113–120 (2008).
- Hyeon, B. et al. Integrating artificial intelligence and optogenetics for Parkinson's disease diagnosis and therapeutics in male mice. *Nat Commun* **16**, 7797 (2025).
- Angiulli, F., Fassetti, F. & Ferragina, L. Latent out: an unsupervised deep anomaly detection approach exploiting latent space distribution. *Mach. Learn.* **112**, 4323–4349 (2023).
- Liu, J. et al. The neuroprogressive nature of major depressive disorder: evidence from an intrinsic connectome analysis. *Transl. Psychiatr.* **11**, 102 (2021).
- Dulawa, S. C., Hollick, K. A., Gundersen, B. & Hen, R. Effects of chronic fluoxetine in animal models of anxiety and depression. *Neuropsychopharmacol* **29**, 1321–1330 (2004).
- Quevedo, J. et al. Acute administration of ketamine induces antidepressant-like effects in the forced swimming test and increases BDNF levels in the rat hippocampus. *Bipolar Disord.* **10**, 84–84 (2008).
- Strekalova, T., Gorenkova, N., Schunk, E., Dolgov, O. & Bartsch, D. Selective effects of citalopram in a mouse model of stress-induced anhedonia with a control for chronic stress. *Behaviour. Pharmacol.* **17**, 271–287 (2006).
- Belozertseva, I. V., Kos, T., Popik, P., Danysz, W. & Bespalov, A. Y. Antidepressant-like effects of mGluR1 and mGluR5 antagonists in the rat forced swim and the mouse tail suspension tests. *Eur. Neuropsychopharm* **17**, 172–179 (2007).
- Wei, Y. C. et al. Medial preoptic area in mice is capable of mediating sexually dimorphic behaviors regardless of gender. *Nat. Commun.* **9**, 279 (2018).

36. Karigo, T. et al. Distinct hypothalamic control of same- and opposite-sex mounting behaviour in mice. *Nature* **589**, 258–263 (2021).
37. Northoff, G. et al. Catatonia as a psychomotor syndrome: a rating scale and extrapyramidal motor symptoms. *Mov. Disord.* **14**, 404–416 (1999).
38. Sobin, C. & Sackeim, H. A. Psychomotor symptoms of depression. *Am. J. Psychiatry* **154**, 4–17 (1997).
39. Yin, R. et al. Lipopolysaccharide-induced depression-like model in mice: meta-analysis and systematic evaluation. *Front. Immunol.* **14**, 1181973 (2023).
40. O'Connor, J. C. et al. Lipopolysaccharide-induced depressive-like behavior is mediated by indoleamine 2,3-dioxygenase activation in mice. *Mol. Psychiatry* **14**, 511–522 (2009).
41. Shin, J. et al. Dysregulation of FGFR1 signaling in the hippocampus facilitates depressive disorder. *Exp. Mol. Med.* <https://doi.org/10.1038/s12276-025-01519-9> (2025).
42. Hartley, R. & Zisserman, A. *Multiple View Geometry in Computer Vision* (Cambridge University Press, 2000).
43. Can, A. et al. The mouse forced swim test. *Jove-J. Vis. Exp.* e3638 <https://doi.org/10.3791/3638> (2012).
44. Meylan, E. M. et al. Involvement of the agmatineric system in the depressive-like phenotype of the knockout mouse model of depression. *Transl. Psychiat.* **6**, e852 (2016).
45. Podkowa, K. et al. Antidepressant-like effects of scopolamine in mice are enhanced by the group II mGlu receptor antagonist LY341495. *Neuropharmacology* **111**, 169–179 (2016).
46. Palucha-Poniewiera, A., Podkowa, K., Lenda, T. & Pilc, A. The involvement of monoaminergic neurotransmission in the antidepressant-like action of scopolamine in the tail suspension test. *Prog. Neuro-Psychoph* **79**, 155–161 (2017).
47. Yang, B. et al. Comparison of R-ketamine and rapastinel antidepressant effects in the social defeat stress model of depression. *Psychopharmacol. (Berl.)* **233**, 3647–3657 (2016).
48. Yang, J. J., Lu, H. F., Li, C. M., Hu, X. P. & Hu, B. Data augmentation for depression detection using skeleton-based gait information. *Med. Biol. Eng. Comput.* **60**, 2665–2679 (2022).
49. Mao, Y. et al. l^2 MD: 3D action representation learning with inter- and intra-modal mutual distillation. Preprint at <https://arxiv.org/abs/2310.15568> (2023).
50. Yu, B., Yin, H. T. & Zhu, Z. X. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (ed Jérôme Lang), 3634–3640 (2018).
51. van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at <https://arxiv.org/abs/1807.03748> (2018).
52. Weinreb, C. et al. Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. *Nat. Methods* **21**, 1329–1339 (2024).
53. Morais, R. et al. Learning regularity in skeleton trajectories for anomaly detection in videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (Cvpr 2019)* (eds Larry Davis et al.), 11988–11996 (2019).
54. Greenacre, M. et al. Principal component analysis. *Nat. Rev. Method Prime* **2**, 100 (2022).
55. Healy, J. & McInnes, L. Uniform manifold approximation and projection. *Nat. Rev. Method Prime* **4**, 82 (2024).
56. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4768–4777 (2017).
57. Yang, J. Fast treeSHAP: accelerating shap value computation for trees. Preprint at <https://arxiv.org/abs/2109.09847> (2021).
58. Oh, H. et al. Contrastive Learning-based Observer-free analysis of Spontaneous behavior for Ethogram Representation. Zenodo <https://doi.org/10.5281/zenodo.13585818> (2024).

Acknowledgements

The authors thank professor Daesoo Kim, Daegun Kim (ACTNOVA) and Junesu Lee for their guidance and for maintaining the AVATAR studio. This research was supported by the Samsung Science and Technology Foundation project SSTF-BA1902-06 (W.D.H.) and a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (2020R1A2C301474213 and RS-2024-00439379, W.D.H.). In addition, this research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2024-00440398) (B.H.). Finally, we would like to thank Sangkyu Lee, Yeonji Jeong, Dong-gyun Kim, and Chihong Shin for stimulating discussions.

Author contributions

H.O., S.C., and W.D.H. conceived the project and directed the research. H.O., S.C., J.L., and W.D.H. designed experiments. S.C. implemented the CLOSER module. H.O. and S.C. performed the experiments. H.O., S.C., and W.D.H. wrote the manuscript. H.L., J.S., and S.S. assisted with experimental design and discussed the results. B.H. contributed to the AVATAR system setup and discussions on behavioural analysis frameworks, and provided unpublished 3D motion data from A53T Parkinson's disease model mice.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-67559-x>.

Correspondence and requests for materials should be addressed to Won Do Heo.

Peer review information *Nature Communications* thanks Pavol Bauer, Jeffrey Markowitz, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025