Weber *et al. BMC Biology* (2022) 20:232
https://doi.org/10.1186/s12915-022-01434-9

BMC Biology

METHODOLOGY ARTICLE · Open Access

Deep learning-based behavioral profiling of rodent stroke recovery

Rebecca Z. Weber[1,2], Geertje Mulders[3], Julia Kaiser[4], Christian Tackenberg[1,2*†] and Ruslan Rust[1,2*†]

Abstract

Background: Stroke research heavily relies on rodent behavior when assessing underlying disease mechanisms and treatment efficacy. Although functional motor recovery is considered the primary targeted outcome, tests in rodents are still poorly reproducible and often unsuitable for unraveling the complex behavior after injury.

Results: Here, we provide a comprehensive 3D gait analysis of mice after focal cerebral ischemia based on the new deep learning-based software (DeepLabCut, DLC) that only requires basic behavioral equipment. We demonstrate a high precision 3D tracking of 10 body parts (including all relevant joints and reference landmarks) in several mouse strains. Building on this rigor motion tracking, a comprehensive post-analysis (with >100 parameters) unveils biologically relevant differences in locomotor profiles after a stroke over a time course of 3 weeks. We further refine the widely used ladder rung test using deep learning and compare its performance to human annotators. The generated DLC-assisted tests were then benchmarked to five widely used conventional behavioral set-ups (neurological scoring, rotarod, ladder rung walk, cylinder test, and single-pellet grasping) regarding sensitivity, accuracy, time use, and costs.

Conclusions: We conclude that deep learning-based motion tracking with comprehensive post-analysis provides accurate and sensitive data to describe the complex recovery of rodents following a stroke. The experimental set-up and analysis can also benefit a range of other neurological injuries that affect locomotion.

Keywords: DeepLabCut, Deep learning, Automated behavior analysis, Photothrombotic stroke, Brain injury, Locomotor profile, Behavioral tests, Ischemic stroke, Mouse

Background

Stroke is a leading cause of disability and death worldwide. Over 13.7 million strokes occur each year, and one in four people over 25 years of age will experience a stroke in their lifetime [1]. The presence of life-saving medicines allows timely intervention, which has significantly decreased mortality following a stroke [2, 3]. However, acute treatments are not applicable in most patients, mainly because of the narrow therapeutic time window, leaving five million patients permanently disabled every year [4, 5]. To promote recovery outside the confines of conventional therapies, a variety of experimental treatments in rodents have emerged targeting neuroprotection [6], therapeutic angiogenesis [7–10], axonal sprouting [11], or cell-based therapies [12–14]. In most of these studies, behavioral evaluation is the primary outcome and ultimately provides evidence that functional impairment can be corrected by the experimental treatment. However, behavioral tests in rodents have proved difficult: (1) test results are often poorly reproducible and (2) the task is limited to a specific sensorimotor outcome, thus ignoring most of the other biologically relevant parameters of functional recovery after stroke [15].

Advances in high-speed video equipment have enabled scientists to record massive datasets of animal behavior

[†] Christian Tackenberg and Ruslan Rust contributed equally to this work.

*Correspondence: christian.tackenberg@irem.uzh.ch; ruslan.rust@irem.uzh.ch

[1] Institute for Regenerative Medicine (IREM), University of Zurich, Campus Schlieren, Wagistrasse 12, 8952 Schlieren, Switzerland
Full list of author information is available at the end of the article

in exquisite detail, and commercial software solutions including Ethovision (Noldus), AnyMaze (Stoelting Co.), and Top Scan (CleverSys Inc.) have assisted with vision-based tracking and analysis. However, these technologies offer little methodological transparency, are not affordable for many laboratories [16], and are often designed to study pre-specified modules within one particular paradigm (e.g., the Morris water maze or the open field test) rather than discover new behavioral patterns. Although the concept of quantitative behavioral analysis has already been implemented in some cases before [17], the introduction of machine learning algorithms has recently reached various sectors of life and provided a new set of tools ideally suited for behavior analysis [18–27]. These algorithms, referred to as deep learning models, offer user-defined feature tracking with greater flexibility, as well as reduced software and hardware acquisition costs [28]. One of the latest contributions to this toolbox is the open-source software DeepLabCut (DLC) [29], which uses convolutional neural networks to automatically capture movements and postures directly from images and without requiring active or passive markers. DLC is a modified version of a state-of-the-art algorithm for tracking human movement, DeeperCut [30], and can be used in a broad range of study systems with near human-level accuracy [31, 32]. Typically, such algorithms are seen as "data-hungry"; algorithms must be trained first by showing thousands of hand-labeled frames, an effort that requires an enormous amount of time. DLC, however, is pre-trained on ImageNet, a large database of images used for image recognition research [33]. With that pretraining in place, DLC only needs a few training examples (typically 50 - 200 frames) to achieve human-level accuracy, making it a highly data-efficient software [29, 34]. DLC has already been implemented in different research fields including neuroscience for general pose estimation and injury prediction [32, 35–38].

In this study, we developed a modular experimental set-up to identify biologically relevant parameters to reveal gait abnormalities and motor deficits in rodents after a focal ischemic stroke. We trained the neural networks to recognize mice of different fur colors from three perspectives (left, bottom, and right) and to label 10 body parts with high accuracy. A detailed comprehensive post hoc script allows analysis of a wide range of anatomical features within basic locomotor functions, vertical and horizontal limb movements, and coordinative features using the freeware software environment R. We detect distinct changes in the overall mouse gait affecting, e.g., step synchronization, limb trajectories, and joint angles after ischemia. These changes are distinct at acute and chronic time points and primarily (but not exclusively) affect the body parts contralateral to the lesion. We further refine the conventional ladder rung tests with DLC (e.g., for detection of foot placements) and compare the deep learning-assisted analysis with widely used behavioral tests for stroke recovery that use human annotations, the gold standard. We detect similar levels of accuracy, less variation, and a considerable reduction in time using the DLC-based approach. The findings are valuable to the stroke field to develop more reliable behavioral readouts and can be applied to other neurological disorders in rodents involving gait abnormalities.

Results

Generation of a comprehensive locomotor profile using deep learning-based tracking

Our aim was to develop a sensitive and reliable profiling of functional motor recovery in mice after stroke using the open-access deep learning software, DLC. Unraveling the complexity of changes in locomotion is best approached via generation of gait parameters [39]. Therefore, we customized a free walking runway with two mirrors that allowed 3D recording of the mice from the lateral/side and down perspectives. The runway can be exchanged with an irregular ladder rung to identify fine-motor impairments by paw placement (Fig. 1A). The dimensions of the set-up were adapted from the routinely used MotoRater (TSE Systems) [40]. After adaptation to the set-up, non-injured mice were recorded from below with a conventional GoPro Hero 8 camera during the behavioral tasks. The DLC networks were trained based on ResNet-50 by manually labeling 120 frames from randomly selected videos of different mice. Individual body parts were selected according to previous guidelines to enable a comprehensive analysis of coordination, movement, and relative positioning of the mouse joints from all three perspectives and included tail base, iliac crest, hip, back ankles, back toe tip, shoulder, wrist, elbow, front toe tip, and head (Fig. 1B, C) [41].

Next, we applied the neural network model to detect and extract the relevant body coordinates in each frame of all recorded videos to generate a 3D walking profile (Additional file 1: Fig. S1A, B). A training set of six videos proved sufficient to achieve a cross-entropy loss of < 0.1% indicating a marginal predicted divergence from the actual label after 500,000 iterations (Fig. 2A). We achieved an average tracking performance, estimated from the root-mean-square error (RMSE), on the test set that only deviated 5.5 pixels (≈ 0.14 cm) in the runway and 4.7 pixels (≈ 0.12 cm) in the rung walk from the human-annotated ground truth. The RMSE differed between individually tracked body parts ranging between 0.05 and 0.3 cm in the test set (Fig. 2B, Additional file 1: Fig. S2A, B). A more detailed analysis revealed that >99% of all predicted body part

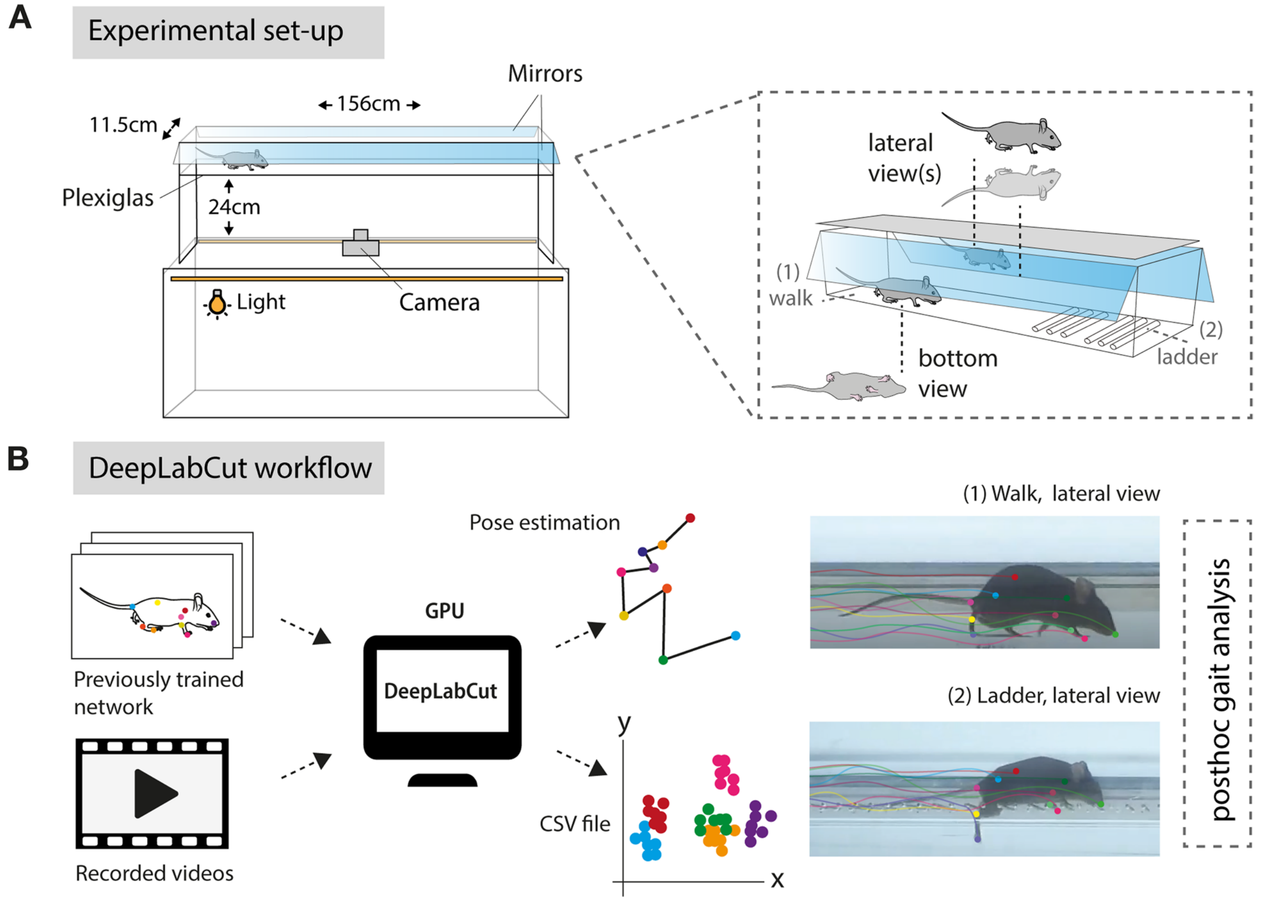


Fig. 1 Experimental workflow to perform deep learning-based gait analysis. **A** Schematic view of the dimensions of experimental set-up. **B** Workflow to identify and label anatomical landmarks of mice for pose estimation. **C** Overview of labeled body parts from side and down perspective

labels fell within the confidence threshold of 17 pixels (≈ 0.45 cm) to the ground truth (Fig. 2C). The ratio of confident labels (labels with a likelihood to appear within the confidence threshold to the ground truth of >95%) to total labels ranged for individual body parts from 96 to 100% for the runway and between 89 and 100% for the rung walk (Fig. 2C, D). In both set-ups, we observed the highest variability for the front and back toe tip labels. For further analysis, all data points that did not pass the likelihood of detection threshold of 95% were excluded. The remaining data generated a full 3D profile of each animal during the behavioral task (Additional file 1: Fig. S1A, B).

Next, we used the same trained networks to reliably label body parts of (a) the same mice 3 days after stroke, (b) different mice with the same genetic background (C57BL/6J, black fur), and (c) mice with a different genetic background (NOD, white fur). We achieved similar confidence in labeling for mice after stroke (95–100%) and mice with the same genotype (97–100%) after minor refinement of the network (see the "Methods" section, Additional file 1: Fig. S2C, D). However, we were unable to successfully refine the pre-existing network to track mice of a different strain with white fur (0–41%). We then created an entirely new training set for these mice with the same training parameters and reached similar levels

(See figure on next page.)

Fig. 2 DeepLabCut enables markerless 3D tracking of mouse body parts. **A** Training efficiency of neural networks. **B** Root mean square error (RMSE) of individual mouse body parts during runway (left) and ladder rung test (right). **C** Likelihood of a confident labeling for individual body parts from down view (left) side view (right) in the runway and **D** during the ladder rung walk. Each dot represents an anatomical landmark of individual image frames in a video. The red dotted line represents the confidence threshold of 95% likelihood for confident labeling

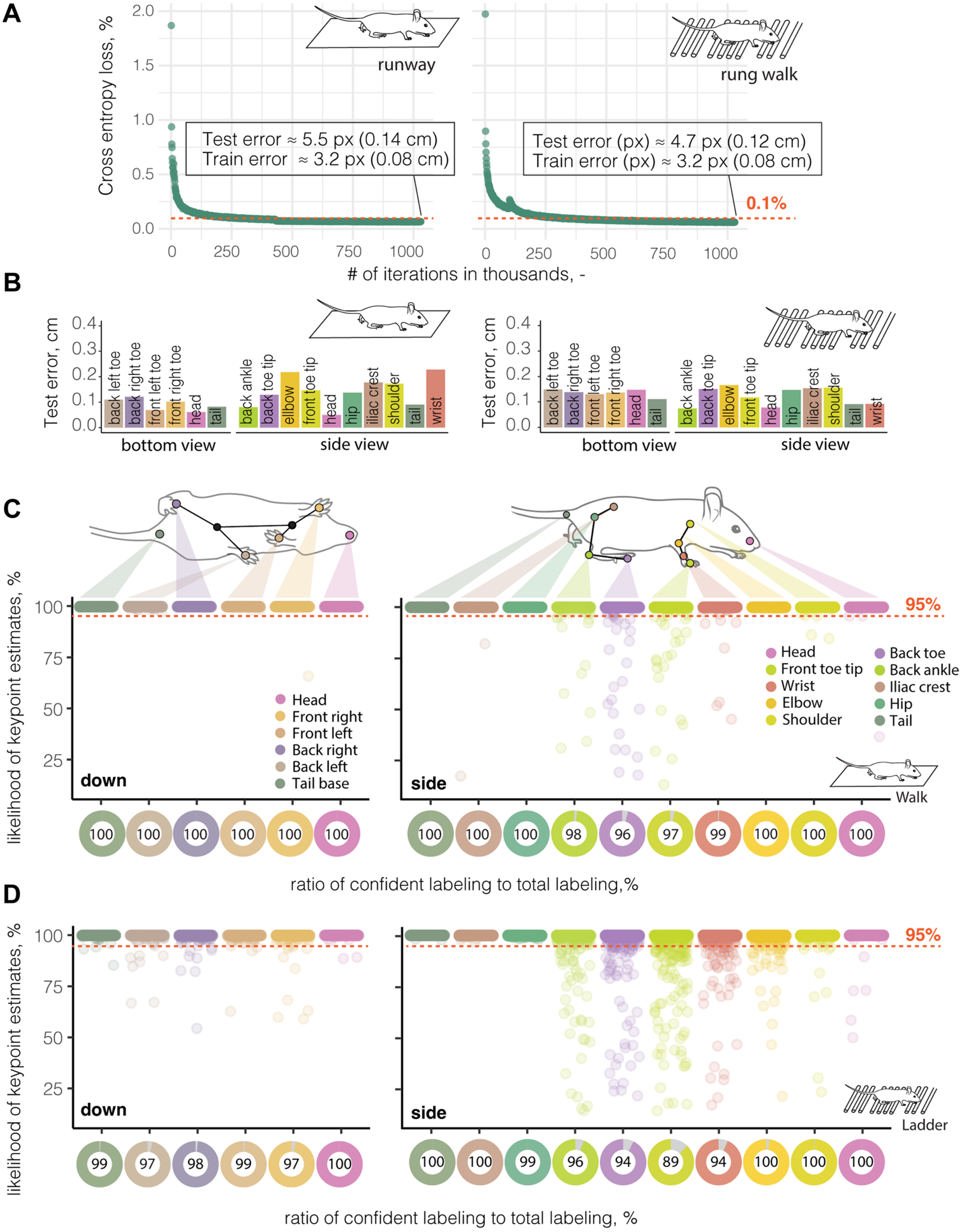


Fig. 2 (See legend on previous page.)

of confident labels (94–100%) to the original training set (Additional file 1: Fig. S2E, F).

Overall, we demonstrated successful labeling and generation of 3D locomotor coordinates in non-injured and injured mice of different genetic backgrounds and fur colors for both the runway and ladder rung walk using deep learning.

Deep learning trained networks detect distinct gait abnormalities following stroke

To identify stroke-related gait abnormalities across a specific time period, we induced a photothrombotic stroke in the right hemisphere of the sensorimotor cortex (Fig. 3A, B) [7, 42]. We confirmed successful induction of ischemia with a reduction of cerebral blood flow only in the ipsilesional right side (right: −72.1 ± 11.5%, p < 0.0001, left: −3.2 ± 8.6%, p = 0.872) using laser Doppler imaging 24 h after injury (Fig. 3C). Three weeks after injury, mice had histological damage in all cortical layers, which was accompanied by a microglial activation and glial scar formation on the ipsilesional hemisphere while sparing subcortical regions and the contralesional side. The injured tissue extended from +2 to −2 mm anterior-posterior related to bregma, and the average stroke volume was 1.3 ± 0.2 mm^3 (Fig. 3D, E).

We began the motion tracking analysis by assessing the overall gait at baseline and after injury over a 3-week period. Individual steps were identified by the movement speed of each limb between frames as filmed from below (Fig. 4A, B). In uninjured animals, the footfall pattern showed a typical gait synchronization [43] of opposing front and back paws (Fig. 4C). Normalizing the data to a single step cycle revealed that this pattern was severely altered acutely after injury as shown by single-animal data (Fig. 4D). We noticed that the asynchronization

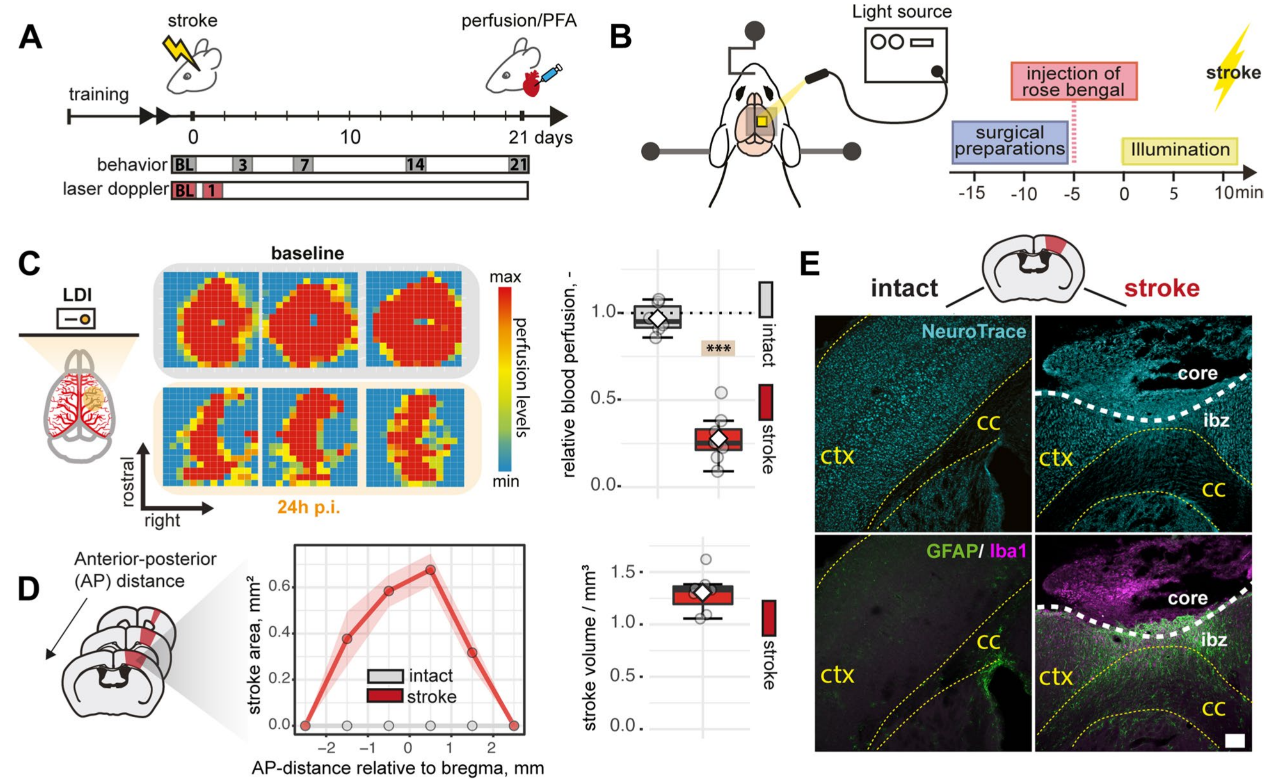


Fig. 3 Induction of photothrombotic stroke leads to permanent focal ischemia in the cortex. **A** Schematic time course of experimental interventions. **B** Schematic representation of stroke procedure. **C** Laser Doppler imaging (LDI) of three representative baseline and stroked brains 24 h after injury. **D** Quantification of stroke area and stroke volume at 21 dpi. Stroke area: Each dot represents the mean of the corresponding subgroup (red: stroke, grey: intact). **E** Representative histological overview of cortical damage (Neurons, cyan), inflammatory infiltration (Iba1+, magenta), and scar formation (GFAP$^+$, green) at 21 dpi, scale bar: 100 μm. Data are shown as mean distributions where the white dot represents the mean. Boxplots indicate the 25 to 75% quartiles of the data. For boxplots: each dot in the plots represents one animal. Line graphs are plotted as mean ± sem. Significance of mean differences between the groups (baseline hemisphere, contralesional hemisphere, and ipsilesional hemisphere) was assessed using Tukey's HSD. Asterisks indicate significance: ****P* < 0.001. ctx, cortex; cc, corpus callosum; ap, anterior posterior; p.i., post injury; ibz, ischemic border zone

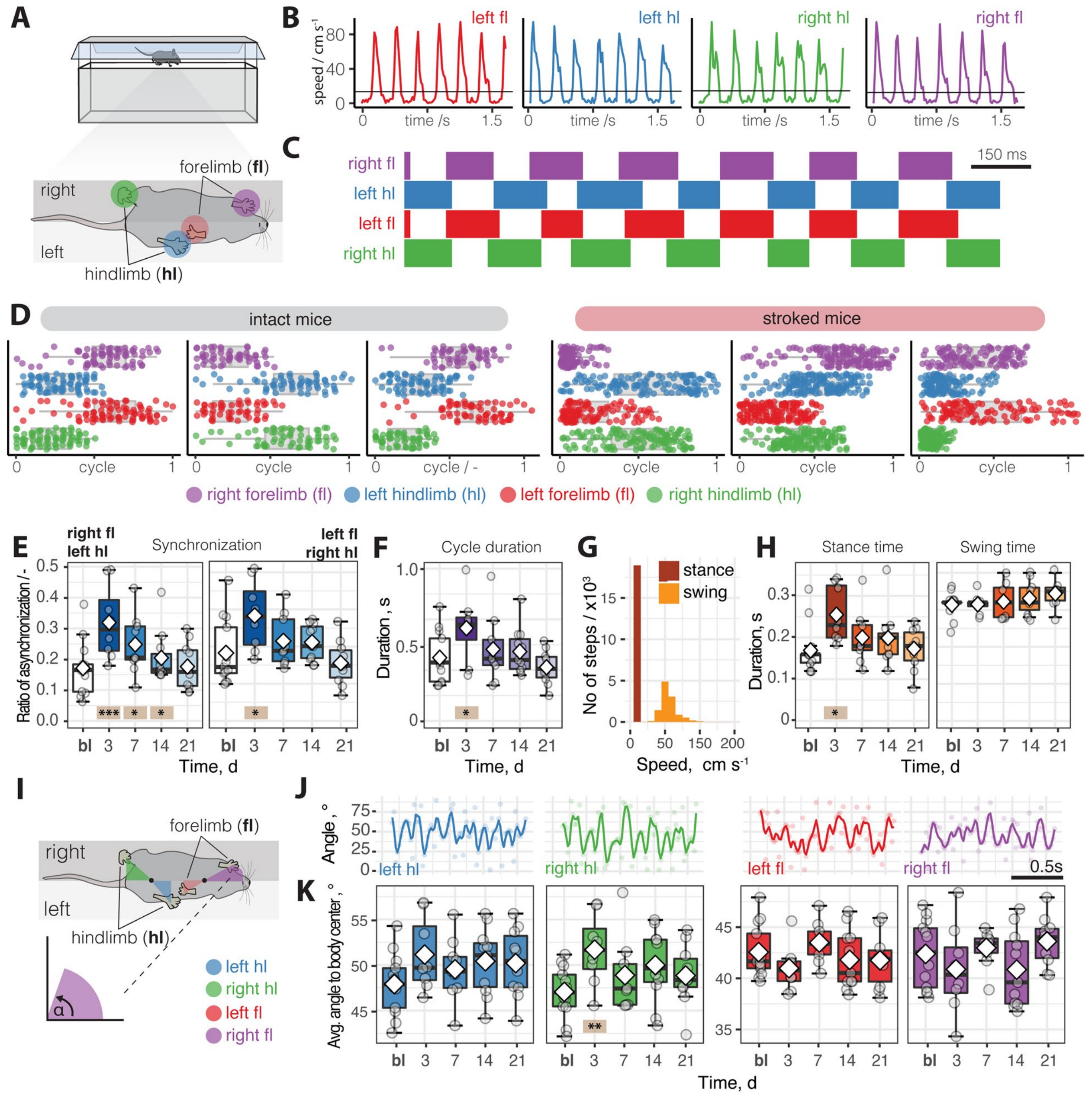


Fig. 4 Gait changes in footfall pattern in spontaneous walk after stroke. **A** Schematic set-up of runway walks from bottom perspective. **B** Movement speed of individual fore- and hindlimbs during spontaneous walk. **C** Footfall profile of single mouse without injury. **D** Footfall profiles of a normalized locomotor cycle showing the stance and phase start and end of three individual control mice (left) and stroked mice (right). **E** Ratio of asynchronization at baseline, 3, 7, 14, 21 dpi. **F** Duration of a cycle. **G**, **H** Comparison of cycle duration between stance and swing time in a time course. **I** Schematic view on analysis of positioning paws to body centers. **J** Profile of paw angles relative to body center of an individual animal. **K** Comparison of angles of individual paws to body center in a time course. Data are shown as mean distributions where the white dot represents the mean. Boxplots indicate the 25 to 75% quartiles of the data. Each dot in the plots represents one animal and significance of mean differences between the groups was assessed using repeated ANOVA with post hoc analysis. Asterisks indicate significance: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$

between the paws is acutely increased after injury ($p < 0.001$, 3 dpi (days post-injury); $p = 0.029$, 7 dpi; p = 0.031, 14 dpi) but recovered to baseline over 21 days ($p = 0.81$, 21 dpi; Fig. 4E). Furthermore, acutely injured mice walked slower as the step cycle duration was increased compared to intact mice ($p = 0.05$, 3 dpi, Fig. 4F). While the swing duration did not differ at any time point (all $p > 0.05$), stroked mice had a longer stance duration ($p =$

0.04, 3dpi, Fig. 4G, H). These alterations in the footfall pattern were associated with changes in the positioning of the paws during a step (Fig. 4I, J). The angle amplitude of the ipsilesional hindlimb relative to the body center increased acutely after injury ($p = 0.003$, 3dpi) while the angle of the front limbs remained unchanged (all $p > 0.05$, Fig. 4K).

Overall, the synchronization of the footfall pattern was severely altered during the acute phase of stroke but returned to a normal pattern in the long-term.

The kinematics of a spontaneous walk were then compared by tracking the fore- and hindlimb joints from the left- and right-side perspectives (Fig. 5A, B). First, we analyzed the average height and total vertical movement of each joint involved in the hindlimb (iliac crest, hip, back ankle, and back toe tip) and forelimb movement (shoulder, elbow, wrist, and front toe tip, Fig. 5C, D). We identified alterations in the total vertical movement and average height of the fore- and hindlimb joints that was, as expected, more prominent on the contralateral left side. Most notably, the total vertical movement decreased in the contralateral left back and front toe tips, shoulder, and wrist at 3 dpi (all $p < 0.05$) with a partial but incomplete recovery over time (Fig. 5E, Additional file 1: Fig. S3). Interestingly, we also observed compensatory changes in the vertical movement on the ipsilateral right side most prominent in the back toe tip, back ankle, elbow, and wrist (Fig. 5E, Additional file 1: Fig. S3). Next, we checked for alterations in the horizontal movement determining the average step length, as well as protraction, and retraction of the individual paws. At 3 dpi both retraction and protraction length are reduced in stroked mice. These changes remained more pronounced in the hind limbs during retraction whereas protractive changes returned to normal throughout the time course (Fig. 5F). Like the vertical movement, we also observed compensatory changes in protraction in the ipsilateral right hindlimb at later time points. Then, the joint positions were used to extract the angles of the hindlimbs (iliac crest-hip-ankle; hip-ankle-toe tip) and forelimbs (shoulder-elbow-wrist; elbow-wrist-toe tip). The angular variations were acutely unchanged after stroke and showed a similar profile throughout the time course (Fig. 5G, H, Additional file 1: Fig. S4).

To understand the individual importance of the >100 measured parameters (Additional file 2: Table S1) during kinematics analysis, we applied a random forest classification to the data of all animals based on all extracted parameters throughout the entire time course (Fig. 5I). The highest Gini impurity-based feature importance between the groups was observed for parameters: left front and back toe heights as well as protractive and total horizontal back toe movements. Since many studies in stroke research only focus on either acute or long-term deficits, a separate analysis was performed between baseline and acutely injured mice at 3 dpi and mice with long-term deficits at 21 dpi (Additional file 1: Fig. S5). In these subgroup analyses, we evaluated the performance of the random forest model with a confusion matrix. We were able to predict the acute injury status with 90% accuracy and long-term deficits with 85% accuracy. The overlap between the 20 highest Gini impurity-based feature importance parameters (top 10% of all measured parameters) in acute and chronic time points was 20% further confirming the need to consider the entirety of the gait to understand the complexity of functional recovery over time (Additional file 1: Fig. S5). We then used a principal component analysis (PCA) to reduce the dimensions of our data and determine the differences between the groups (Fig. 5J). We found that data from later time points after injury cluster closer to the baseline suggesting that the recovery effects can be ascertained based on kinematic parameters. The separation expands when comparing only data from 3 dpi and 21 dpi to baseline (Additional file 1: Fig. S5). Importantly, these changes were not observed in non-stroked control mice throughout the time course (Additional file 1: Fig. S6).

Next, we considered whether deep networks can also be applied to conventional behavioral tests to detect fine motor deficits in a ladder rung test (Fig. 6A). Tracking the fore- and hindlimbs during the ladder rung recordings enabled the identification of stepping errors in the side view (Fig. 6B, C). We identified a 106% increase of

(See figure on next page.)

Fig. 5 Kinematic changes in runway walk after stroke. **A** Schematic overview of analysis from the sides and time course of the experiment. **B** Stick profile of fore and hindlimb movement in individual mouse. **C, D** Walk profile of hindlimb and forelimb joints in intact and stroked mice; *x*-axis represents relative height and *y*-axis represents time across multiple steps (grey shades). **E** Absolute height of selected joints at baseline and 3, 7, 14, 21 dpi. **F** Experimental design of runway testing. **G** Protraction and retraction of joints throughout a time course. **H** Angular variability between front and hindlimb joints. **I** Random Forest classification of most important parameters. **J** Principal component analysis of baseline 3, 7, 14, 21 dpi. Data are shown as mean distributions where the white dot represents the mean. Boxplots indicate the 25 to 75% quartiles of the data. For boxplots: each dot in the plots represents one animal. Line graphs are plotted as mean ± sem. For line graphs: the dots represent the mean of the data. Significance of mean differences between the groups was assessed using repeated ANOVA with post hoc analysis. Asterisks indicate significance: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. i-h-a, iliac crest-hip-ankle; h-a-t, hip-ankle-toe; s-e-w, shoulder-elbow-wrist; e-w-t, elbow-wrist-toe; PC, principal component

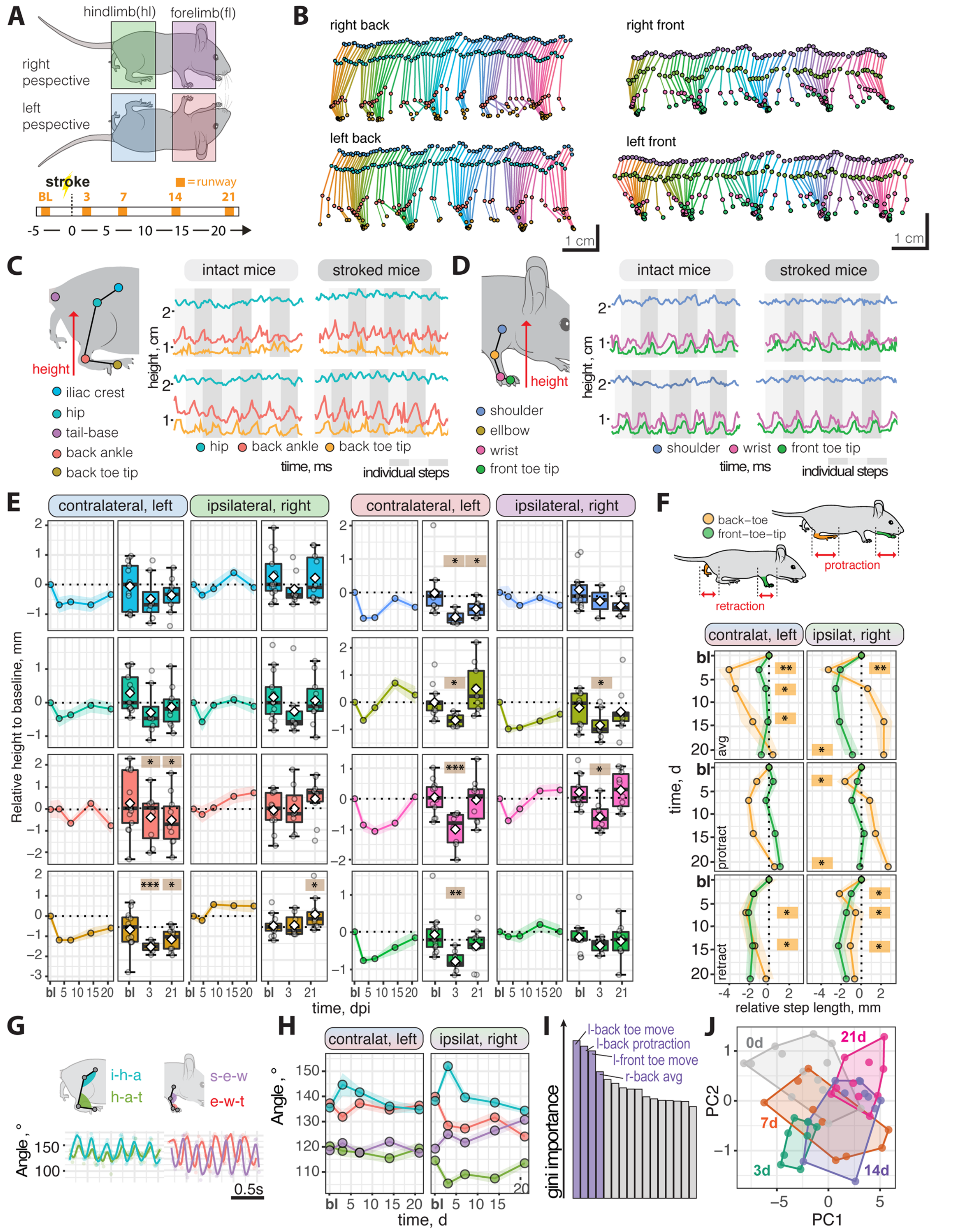
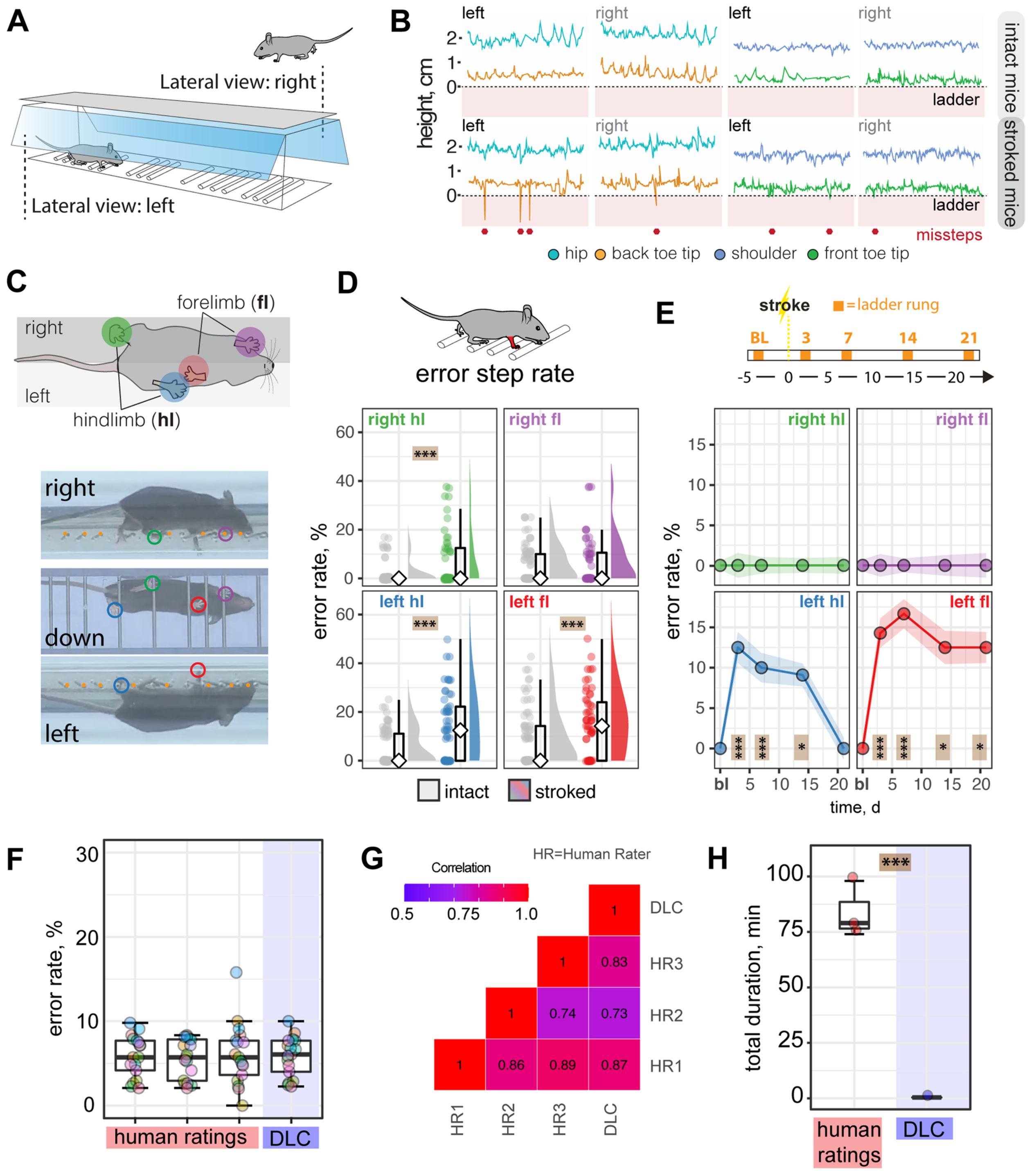


Fig. 5 (See legend on previous page.)



	HR1	HR2	HR3	DLC
DLC				1
HR3			1	0.83
HR2		1	0.74	0.73
HR1	1	0.86	0.89	0.87

Fig. 6 DeepLabCut assisted analysis of horizontal rung test after stroke. **A** Schematic view of ladder rung test. **B** Side view of step profile of hips, back toes, shoulder, and front toes in individual animals at baseline and 3 dpi. **C** Photographs of mouse from three perspectives. **D** Overall success and error rate in the contralesional and ipsilesional hemisphere of all paws. **E** Time course of error rate during ladder rung test in the individual paws. **F** Comparison of error rate scores in selected videos between three human observers and DLC. **G** Correlation matrix between human observers and DLC. **H** Duration of analysis for ladder rung test for 20× 10s videos. Data are shown as mean distributions where the white dot represents the mean. Boxplots indicate the 25 to 75% quartiles of the data. For boxplots: each dot in the plots represents one animal. Line graphs are plotted as mean ± sem. For line graphs: the dots represent the mean of the data. Significance of mean differences between the groups was assessed using repeated ANOVA with post hoc analysis. Asterisks indicate significance: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$

the overall error rate in injured animals compared with their intact controls (intact 5.27 ± 8.4%, stroked: 10.9 ± 12.6%, $p < 0.001$) at 3 dpi. This increased error rate after acute stroke was most pronounced on the contralesional side (left front paw: +182%; left back paw: +142%, both p < 0.001) but also marginally detectable in the ipsilesional site (right front paw: +21%, $p = 0.423$; right back paw: +17%, $p < 0.001$; Fig. 6D). In a time course of 3 weeks, we detected a marked increase of footfall errors in both the contralateral left front and back toe (all $p < 0.001$) compared to baseline. Although the error rate returned to baseline for the back paw ($p = 0.397$), it does not fully recover for front paw ($p = 0.017$), as previously observed [7] (Fig. 6E).

In a subset of 20 randomly selected videos, we cross-verified the error rates by a blinded observer and compared the variability between the DLC-approach and the manual assessment of the parameters regarding (1) variability of the analysis and (2) duration of the analysis. We did not detect a difference in the scoring accuracy between the manual assessment and DLC-assisted analysis (Fig. 6F). The error rate in individual videos highly correlated between human and the machine learning evaluation (Fig. 6G, Additional file 1: Fig. S7A-F). A frame-by-frame analysis revealed that footfall errors that were identified by all human raters were recognized in the 22/23 (95%) frames by the DLC-assisted analysis further confirming the correct classification of the DLC model (Additional file 1: Fig. S7G, H). The automated analysis is also considerably faster after the initial effort has been overcome (Additional file 1: Fig. S8A). We estimate that DLC-assisted analysis is 200 times faster with regular use, once the neural networks are established (human: 4.18 ± 0.63 min; DLC: 0.02 min; for a 10s video, $p < 0.0001$; Fig. 6F–H, Additional file 1: Fig. S7).

Overall, these results suggest that DLC-assisted analysis of the ladder rung test achieves human-level accuracy, while saving time and avoiding variability between human observers.

Comparison of deep learning-based tracking to conventional behavioral tests for stroke-related functional recovery

Finally, we benchmarked DLC-tracking performance against popular functional tests to detect stroke-related functional deficits. We performed a rotarod test with the same set of animals and analyzed previously acquired data from a broad variety of behavioral tasks routinely used in stroke research including neurological scoring, cylinder test, the irregular ladder rung walk, and single pellet grasping (Fig. 7A, B).

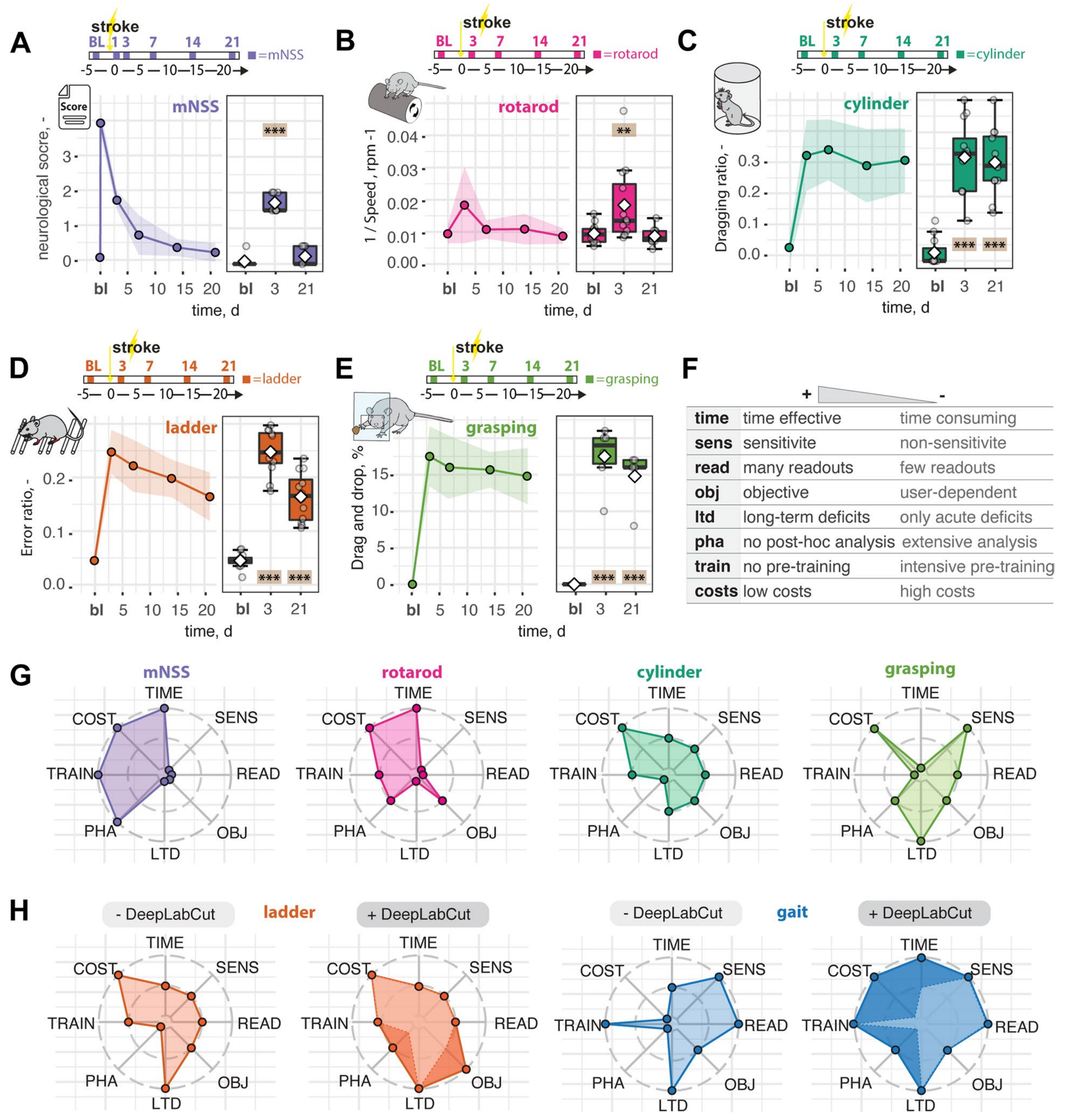
In all behavioral set-ups, we identified initial deficits after stroke (rotarod: $p = 0.006$, all other tests: $p < 0.001$). While the neurological deficit score (21 dpi, $p = 0.97$) and the rotarod (21 dpi, $p = 0.99$) did not provide much sensitivity beyond the acute phase (Fig. 7A, B), the ladder rung test, cylinder test, and single pellet grasping were suitable to reveal long-term impairments in mouse stroke models (21 dpi, all $p < 0.001$, Fig. 7C–E, Additional file 1: Fig. S9A).

These functional tests were then further compared in a semi-quantitative spider diagram regarding (1) duration to perform the task, (2) objectivity, (3) post hoc analysis, (4) requirement of pre-training, and (5) costs (Fig. 7G, Additional file 1: Fig. S9B, C). Despite the simple performance, the neurological scoring, rotarod, and cylinder tests have the drawback of a relatively low sensitivity and objectivity. On the other hand, more sensitive tests such as the pellet grasping test require intense pre-training of the animals, or the manual post-analysis of a ladder rung test can be tedious and suffers from variability between investigators. Many conventional tests only provide a very low number of readouts, which may not capture the entire complexity of the acute injury and subsequent recovery.

More advanced analysis including kinematic tracking offers the advantage of generating a variety of parameters but the high costs for the set-up and the commercial software are disadvantageous (Fig 7H). The DLC-assisted tracking presented here provides an open-source solution that is available at negligible costs and can be set up easily. The experiment duration is shortened, and animal welfare is improved since the test does not require marking the mouse joints beforehand. Most importantly, using our comprehensive post-analysis, the set-up reduces analysis time while minimizing observer biases during the evaluation.

Discussion

Preclinical stroke research heavily relies on rodent behavior when assessing functional recovery or treatment efficacy. Nonetheless, there is an unmet demand for comprehensive unbiased tools to capture the complex gait alterations after stroke; many conventional methods either do not have much sensitivity aside from identifying initial injures or require many resources and a time-consuming analysis. In this study, we used deep learning to refine 3D gait analysis of mice after stroke. We performed markerless labeling of 10 body parts in uninjured control mice of different strains and fur color with 99% accuracy. This allowed us to describe a set of >100 biologically meaningful parameters for examining, e.g., synchronization, spatial variability, and joint angles during a spontaneous walk, and that showed differential importance for acute and long-term deficits. We refined our deep learning analysis for use with the ladder rung test,



	+	-
time	time effective	time consuming
sens	sensitivite	non-sensitivite
read	many readouts	few readouts
obj	objective	user-dependent
ltd	long-term deficits	only acute deficits
pha	no post-hoc analysis	extensive analysis
train	no pre-training	intensive pre-training
costs	low costs	high costs

Fig. 7 Functional assessment of recovery after stroke using conventional behavioral tests. **A** Neurological score, **B** rotarod test, **C** dragging during cylinder test, **D** missteps in ladder rung test, and **E** drag and drop in single pallet grasping. **F** Semi-quantitative measure of relevant parameters for behavioral tests (time, sensitivity, readouts, objectivity, long-term deficits, post-hoc analysis, pre-training, and costs). **G** Spider chart of conventional behavioral tests. **H** Spider chart of behavioral tests without and with DLC assistance. Scale: +: high , 0: neutral, –: low. Dara are shown as line graphs and are plotted as mean ± sem. Significance of mean differences between the groups was assessed using repeated ANOVA with post hoc analysis. Asterisks indicate significance: $^{**}P < 0.01$, $^{***}P < 0.001$

which achieved outcomes comparable to manual scoring accuracy. We found that our DLC-assisted tracking approach, when benchmarked to other conventionally used behavior tests in preclinical stroke research, outperformed those based on measures of sensitivity, time-demand, and resources.

The use of machine learning approaches has dramatically increased in life sciences and will likely gain importance in the future. The introduction of DeepLabCut considerably facilitated the markerless labeling of mice and expanded the scope of kinematic tracking software [29, 34]. Although commercial attempts to automate behavioral tests eliminated observer bias, the analyzed parameters are often pre-defined and cannot be altered. Especially for customized set-ups, DLC has been shown to reach human-level accuracy while outperforming commercial systems (e.g., EthoVision, TSE Multi Conditioning system) at a fraction of the cost [32]. These advantages may become more apparent in the future since unsupervised machine learning is beginning to reveal the true complexity of animal behavior and may allow recognition of behavioral sequences not detectable by humans. On the other hand, execution and interpretation of unsupervised tracking are often beyond the reach of many basic research labs and require the necessary machine learning knowledge [15].

Consequently, deep learning-based quantitative analysis has been successfully performed in various behavioral set-ups often on rodents for data-driven unbiased behavior evaluation [22, 24, 25, 27, 32]. Some data-driven approaches also used acute brain injuries and neurological disorders to reliably predict the extent of injury [38, 44]. For instance, spinal cord injuries and traumatic brain injuries could be reliably quantified over a time course of 3 weeks using an automated limb motion analysis (ALMA) that was also based on DLC [44]. Other studies applied DLC to stroked mice and rats to a set of behavioral tests including set-ups analyzed here such as the horizontal ladder rung test and single pellet grasping [38]. Although the set-up provides an excellent tool for predicting moving deficits in specific stroke-related behavioral tasks (e.g., paw movement during pellet grasp or missteps during horizontal rung test), the tool was not tested for general pose estimation and gait analysis after stroke. Nevertheless, these methods achieved a comparable level of reliability to the present study in tracking and identifying injury status and may also be refined to perform similar post hoc analysis. We did not perform a quantitative comparison of the performance among the developed data-driven behavioral analysis tools. It is challenging to compare or determine the most appropriate method for functional stroke outcome as it is likely to depend on a variety of factors, such as the general experimental set-up, the species used, and the biological question. We believe that our study is particularly useful for groups that may not have the methodological machine learning experience to apply a generalized pose estimation tool to their stroke model. In addition, research groups may complement their already established network with our post hoc analysis with pre-defined and biologically relevant parameters.

Many neurological disorders (e.g., multiple sclerosis, Huntington’s, and spinal cord injury) result in pronounced motor deficits in patients, as well as in mouse models, with alterations in the general locomotor pattern. These alterations are usually readily identifiable, especially in the acute phase, and excellent automated tools have recently been developed to track the motor impairments [44]. In contrast, deficits following cortical stroke in mice often do not reveal such clear signs of injury and require higher levels of sensitivity to identify the motor impairment [7, 8, 45]. The degree of functional motor deficits after stroke is highly dependent on corticospinal tract lesions that often result in specific deficits, e.g., impairment of fine motor skills [46]. Moreover, a stroke most commonly affects only one body side; therefore, an experimental set-up that contains 3D information is highly valuable, as it enables the detection of contra- and ipsilateral trajectories of each anatomical landmark. Accordingly, our set-ups enabled us to also detect intra-animal differences that may be important to distinguish between normal and compensatory movements throughout the recovery time course [47].

Compensatory strategies (e.g., avoiding the use of the impaired limb or relying on the intact limb) are highly prevalent in rodents and in humans [48, 49]. Although functional recovery is generally observed in a variety of tests, it is important to distinguish between compensatory responses and “true” recovery. These mechanisms are hard to dissect in specific trained tasks (e.g., reaching during single pellet grasping). Therefore, tasks of spontaneous limb movements and many kinematic parameters are valuable to distinguish these two recovery mechanisms [39]. Interestingly, we observed alterations in several ipsilateral trajectories during the runway walk affecting the vertical positioning as well as protractive and retractive movement (although less prominent than in the contralateral paws) that suggest a compensatory movement. Similar gait alterations have been previously reported in a mouse model of distal middle cerebral artery occlusion, another common model of ischemia in mice [50]. These compensatory movements are predominantly caused by either plastic change by the adjacent areas of cortex or through support from anatomical reorganization of the contralesional hemisphere [51–53] and, therefore, could provide valuable information about the therapeutic effects of a drug or a treatment.

Apart from general kinematic gait analysis, variations of the horizontal ladder rung/foot fault or grid tests remain one of the most reproducible tasks to assess motor skill in rodents after injury, including stroke. However, these tests often remain unused in many experimental stroke studies, most likely due to the associated time-consuming analysis. DLC-assisted refinement might allow future studies to incorporate this important assessment into their analyses given the striking decrease in time investment. We have demonstrated that DLC-assisted refinement of these conventional tests represents a striking decrease in time consumption. Therefore, it is conceivable that some of these conventional tests and other assessment methods (e.g., single pellet grasping) may profit from the advancements of deep learning and will not be fully replaced by kinematic gait analysis.

Interestingly, some of the assessed parameters showed an impairment after stroke only in the acute phase (e.g., synchronization, cycle duration, hip movement), while some parameters showed an initial impairment after injury followed by a partial or full recovery (e.g., wrist height, toe movement, and retraction) and others showed no recovery in the time course of this study. Given the number of parameters raised in this setting, this approach might be particularly suited to assess treatment efficacy of drug interventions in preclinical stroke research. Overall, we found a strong separation of parameters in the acute vs. chronic phase in the PCA and random forest analysis, making this approach suitable to assess both the acute phase as well as the chronic phase. It will be of interest in the future to assess the presented approach in different models of stroke as well as in additional neurological conditions such as spinal cord injury, ALS, cerebral palsy, or others [39].

Notably, a detailed kinematic analysis required optional recording settings to generate a high contrast between animal and background. In our experience, these parameters needed to be adapted to the fur color in the animals. Although we reached almost equivalent tracking accuracy of 99.4% (equivalent to losing 6 in every 1000 recorded frames), mice with black and white fur could not be tracked based on the same neural network and required two training sessions, which may show slight differences in the analysis. Moreover, the high accuracy in our experimental set-up was achieved by recording only mice with smooth runs without longer interruptions. In the future, these limitations could be overcome by combining DLC-tracking with a recently developed unsupervised clustering approach to reveal grooming or other unpredictable stops during a run [54, 55].

Conclusions

Taken together, in this study, we developed a comprehensive gait analysis to assess stroke impairments in mice using deep learning. The developed set-up requires minimal resources and generates characteristic multifaceted outcomes for acute and chronic phases after stroke. Moreover, we refined conventional behavioral tests used in stroke assessment at human-level accuracy that may be expanded for other behavioral tests for stroke and other neurological diseases affecting locomotion.

Methods

Study design

The goal of the study was to develop a comprehensive, unbiased analysis of post-stroke recovery using deep learning over a time course of 3 weeks. Therefore, we performed a large photothrombotic stroke in the sensorimotor cortex of wildtype and NSG mice (male and female mice were used). We assessed a successful stroke using laser Doppler imaging directly after surgery and confirmed the stroke volume after tissue collection at 3 weeks post-injury. Animals were subjected to a series of behavioral tests at different time points. The here used tests included the (1) runway, (2) ladder rung test, (3) rotarod test, (4) neurological scoring, (5) cylinder test, and (6) single pallet grasping. All tests were evaluated at baseline and 3, 7, 14, and 21 after stroke induction. Video recordings from the runway and ladder rung test were processed by a recently developed software DeepLabCut (DLC, v. 2.1.5), a computer vision algorithm that allows automatic and markerless tracking of user-defined features. Videos were analyzed to plot a general overview of the gait. Individual steps were identified within the run by the speed of the paws to identify the "stance" and "swing" phase. These steps were analyzed (from the bottom perspective for, e.g., synchronization, speed, length, and duration from the down view over a time course. From the lateral/side view, we next measured, e.g., average, and total height differences of individual joins (*y*-coordinates) and the total movement, protraction, and retraction changes per step (x-coordinates) over the time course. All >100 generated parameters were extracted to perform a random forest classification to determine the importance for determining accuracy of the injury status. The most five important parameters were used to perform a principal component analysis to demonstrate separation of these parameters.

We compared DLC-tracking performance against popular functional tests to detect stroke-related functional deficits including neurological score, rotarod test, dragging during cylinder test, missteps in ladder rung test, and drag and drop in single pallet grasping.

Animals

All procedures were conducted in accordance with governmental, institutional (University of Zurich), and ARRIVE guidelines and had been approved by the Veterinarian Office of the Canton of Zurich (license: 209/2019). In total, 33 wildtype (WT) mice with a C57BL/6 background mice and 12 non-obese diabetic SCID gamma

(NSG) mice were used (female and male, 3 months of age). Mice were housed in standard type II/III cages on a 12h day/light cycle (6:00 A.M. lights on) with food and water ad libitum. All mice were acclimatized for at least a week to environmental conditions before set into experiment. All behavioral analysis were performed on C57BL/6 mice. NSG mice were only used to evaluate the ability of the networks to track animals with white fur.

Photothrombotic lesion

Mice were anesthetized using isoflurane (3% induction, 1.5% maintenance, Attane, Provet AG). Analgesic (Novalgin, Sanofi) was administered 24 h prior to the start of the procedure via drinking water. A photothrombotic stroke to unilaterally lesion the sensorimotor cortex was induced on the right hemisphere, as previously described [56–59]. The stroke procedure was equivalent for all mouse genotypes [60]. Briefly, animals were placed in a stereotactic frame (David Kopf Instruments), the surgical area was sanitized, and the skull was exposed through a midline skin incision. A cold light source (Olympus KL 1,500LCS, 150W, 3,000K) was positioned over the right forebrain cortex (anterior/posterior: −1.5 to +1.5 mm and medial/lateral 0 to +2 mm relative to Bregma). Rose Bengal (15 mg/ml, in 0.9% NaCl, Sigma) was injected intraperitoneally 5 min prior to illumination and the region of interest was subsequently illuminated through the intact skull for 12 min. To restrict the illuminated area, an opaque template with an opening of 3 × 4 mm was placed directly on the skull. The wound was closed using a 6/0 silk suture and animals were allowed to recover. For postoperative care, all animals received analgesics (Novalgin, Sanofi) for at least 3 days after surgery.

Blood perfusion by laser Doppler imaging

Cerebral blood flow (CBF) was measured using laser Doppler imaging (LDI, Moor Instruments, MOORLDI2-IR). Animals were placed in a stereotactic frame; the surgical area was sanitized and the skull was exposed through a midline skin incision. The brain was scanned using the repeat image measurement mode. All data were exported and quantified in terms of flux in the ROI using Fiji (ImageJ). All mice receiving a stroke were observed with LDI directly after injury to confirm a successful stroke. A quantification of cerebral blood perfusion 24 h after injury was performed in $N = 8$ mice.

Perfusion with paraformaldehyde (PFA) and tissue processing

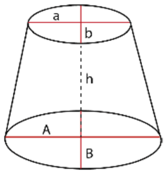
On post-stroke day 21, animals were euthanized by intraperitoneal application of pentobarbital (150mg/kg body weight, Streuli Pharma AG). Perfusion was performed using Ringer solution (containing 5 ml/l Heparin, B.Braun) followed by paraformaldehyde (PFA, 4% in 0.1 M PBS, pH 7.5). For histological analysis, brains were rapidly harvested, post-fixed in 4% PFA for 6 h, subsequently transferred to 30% sucrose for cryoprotection and cut (40 µm thick) using a sliding microtome/Microm HM430, Leica). Coronal sections were stored as free-floating sections in cryoprotectant solution at −20°.

Lesion volume analysis

A set of serial coronal sections (40 µm thick) were immunostained for NeuroTracer (fluorescent Nissl) and imaged with a 20×/0.8 objective lens using an Axio Scan.Z1 slide scanner (Carl Zeiss, Germany). The sections lie between 200 and 500 µm apart; the whole brain was imaged. The cortical lesion infarct area was measured (area, width, depth) on FIJI using the polygon tool, as defined by the area with atypical tissue morphology including pale areas with lost NeuroTracer staining.

The volume of the injury was estimated by calculating an ellipsoidal frustum, using the areas of two cortical lesions lying adjacent to each other as the two bases and the distance between the two areas as height. The following formula was used:

$$V = h * \frac{\pi}{3} * \frac{(A*B^2 - a*b^2)}{(B-b)}$$



The single volumes (*frustums*) were then summed up to get the total ischemic volume. The stroke analysis was performed 21 days after stroke in a subgroup of $N = 8$ mice.

Immunofluorescence

Brain sections were washed with 0.1M phosphate buffer (PB) and incubated with blocking solution containing donkey serum (5%) in PB for 30 min at room temperature. Sections were incubated with primary antibodies (rb-GFAP 1:200, Dako, gt-Iba1, 1:500 Wako, NeuroTrace™ 1:200, Thermo Fischer) overnight at 4°C. The next day, sections were washed and incubated with corresponding secondary antibodies (1:500, Thermo Fischer Scientific). Sections were mounted in 0.1 M PB on Superfrost PlusTM microscope slides and coverslipped using Mowiol.

Behavioral studies

Animal were subjected to a series of behavioral tests at different time points. The here used tests included the

(1) runway, (2) ladder rung test, (3) the rotarod test, (4) neurological scoring, (5) cylinder test, and (6) single pallet grasping. All tests were evaluated at baseline and 3, 7, 14, and 21 after stroke induction. Animals used for deep learning-assisted tests (runway, ladder rung) represent a different cohort of animals to the remaining behavior tasks.

Runway test

A runway walk was performed to assess whole-body coordination during overground locomotion. The walking apparatus consisted of a clear Plexiglas basin, 156 cm long, 11.5 cm wide, and 11.5 cm high (Fig. 1). The basin was equipped with two ~ 45° mirrors (perpendicularly arranged) to allow simultaneous collection of side and bottom views to generate three-dimensional tracking data. Mice were recorded crossing the runway with a high-definition video camera (GoPro Hero 7) at a resolution of 4000 × 3000 and a rate of 60 frames per second. Lighting consisted of warm background light and cool white LED stripes positioned to maximize contrast and reduce reflection. After acclimatization to the apparatus, mice were trained in two daily sessions until they crossed the runway at constant speed and voluntarily (without external intervention). Each animal was placed individually on one end of the basin and was allowed to walk for 3 minutes.

Ladder rung test

The same set-up as in the runway was used for the ladder rung test, to assess skilled locomotion. We replaced the Plexiglas runway with a horizontal ladder (length: 113 cm, width: 7 cm, distance to ground: 15 cm). To prevent habituation to a specific bar distance, bars were irregularly spaced (1-4 cm). For behavioral testing, a total of at least three runs per animal were recorded. Kinematic analysis of both tasks was based exclusively on video recordings and only passages with similar and constant movement velocities and without lateral instability were used. A misstep was defined when the mouse toe tips reached 0.5 cm below the ladder height. The error rate was calculated by errors/total steps × 100.

Rotarod test

The rotarod test is a standard sensory-motor test to investigate the animals' ability to stay and run on an accelerated rod (Ugo Basile, Gemonio, Italy). All animals were pre-trained to stay on the accelerating rotarod (slowly increasing from 5 to 50 rpm in 300s) until they could remain on the rod for > 60 s. During the performance, the time and speed were measured until the animals fell or started to rotate with the rod without running. The test was always performed three times and means were used for statistical analysis. The recovery phase between the trials was at least 10 min.

Neurological score/Bederson score

We used a modified version of the Bederson (0–5) score to evaluate neurological deficits after stroke. The task was adapted from Biebet et al. The following scoring was applied: (0) no observable deficit; (1) forelimb flexion; (2) forelimb flexion and decreased resistance to lateral push; (3) circling; (4) circling and spinning around the cranial-caudal axis; and (5) no spontaneous movement/ death.

Cylinder test

To evaluate locomotor asymmetry, mice were placed in an opentop, clear plastic cylinder for about 10 min to record their forelimb activity while rearing against the wall of the arena. The task was adapted from Roome et al. [61]. Forelimb use is defined by the placement of the whole palm on the wall of the arena, which indicates its use for body support. Forelimb contacts while rearing were scored with a total of 20 contacts recorded for each animal. Three parameters were analyzed which include paw preference, symmetry, and paw dragging. Paw preference was assessed by the number of impaired forelimb contacts to the total forelimb contacts. Symmetry was calculated by the ratio of asymmetrical paw contacts to total paw contacts. Paw dragging was assessed by the ratio of the number of dragged impaired forelimb contacts to total impaired forelimb contacts.

Single pellet grasping

All animals were trained to reach with their right paw for 14 days prior to stroke induction over the left motor cortex. Baseline measurements were taken on the day before surgery (0dpo) and test days started at 4 dpo and were conducted weekly thereafter (7, 14, 21, 28 dpo). For the duration of behavioral training and test periods, animals were food restricted, except for 1 day prior to 3 days post-injury. Body weights were kept above 80% of initial weight. The single pellet reaching task was adapted from Chen et al. [62]. Mice were trained to reach through a 0.5-cm-wide vertical slot on the right side of the acrylic box to obtain a food pellet (Bio-Serv, Dustless Precision Pellets, 20 mg) following the guidelines of the original protocol. To motivate the mice to not drop the pellet, we additionally added a grid floor to the box, resulting in the dropped pellets to be out of reach for the animals. Mice were further trained to walk to the back of the box in between grasps to reposition themselves as well as to calm them down in between unsuccessful grasping attempts. Mice that did not successfully learn the task during the 2 weeks of shaping were excluded from the task ($n = 2$). During each experiment session, the grasping success was scored for 30 reaching attempts or for a maximum of 20 min. Scores for the grasp were as follows: "1" for a successful grasp and retrieval of the pellet (either on first attempt or after several attempts); "0" for a complete miss in which the pellet was misplaced and not retrieved into the box; and "0.5" for drag or drops, in which the animal successfully grasped the pellet but dropped it during the retrieval phase. The success rate was calculated for each animal as end score = (total score/number of attempts × 100).

DeepLabCut (DLC)

Video recordings were processed by DeepLabCut (DLC, v. 2.1.5), a computer vision algorithm that allows automatic and markerless tracking of user-defined features. A relatively small subset of camera frames (training dataset) is manually labeled as accurately as possible (for each task and strain, respectively). Those frames are then used to train an artificial network. Once the network is sufficiently trained, different videos can then be directly input to DLC for automated tracking. The network predicts the marker locations in the new videos, and the 2D points can be further processed for performance evaluation and 3D reconstruction.

Dataset preparation

Each video was migrated to Adobe Premiere (v. 15.4) and optimized for image quality (color correction and sharpness). Videos were split into short one-run-sequences (left to right or right to left), cropped to remove background and exported/compressed in H.264 format. This step is especially important when analyzing the overall gait performance of the animals because pauses or unexpected movements between the steps may influence the post-hoc analysis.

Training

The general networks for both behavioral tests were trained based on ResNet-50 by manually labeling 120 frames selected using k-means clustering from multiple videos of different mice ($N = 6$ videos/network). An experienced observer labeled 10 distinct body parts (head, front toe tip, wrist, shoulder, elbow, back toe, back ankle, iliac crest, hip, tail; Additional file 1: Fig. S10) in all videos of mice recorded from side views (left, right) and 8 body parts (head, right front toe, left front toe, center front, right back toe, left back toe, center back, tail base) in all videos showing the bottom view, respectively (for details, see Additional file 1: Fig. S1, 2). We then randomly split the data into training and test set (75%/25% split) and allowed training to run for 1,030,000 iterations (DLC's native cross-entropy loss function plateaued between 100,000 and 300,000 iterations). Labeling accuracy was calculated using the root-mean-squared error (RMSE) in pixel units, which is a relevant performance metric for assessing labeling precision in the train and test set. This function computes the Euclidean error between human-annotated ground truth data and the labels predicted by DLC averaged over the hand locations and test images. During training, a score-map is generated for all keypoints up to 17 pixels (≈0.45 cm, distance threshold) away from the ground truth per body part, representing the probability that a body part is at a particular pixel [27].

Refinement

Twenty outlier frames from each of the training videos were manually corrected and then added to the training dataset. Locations with a p <0.9 were relabeled. The network was then refined using the same numbers of iterations (1,030,000). For the ladder rung test, frames were manually selected with footfalls to ensure that DLC reliably identifies missteps as they occur rarely in healthy mice and are important for the analysis.

All experiments were performed inside the Anaconda environment (Python 3.7.8) provided by DLC using NVIDIA GeForce RTX 2060.

Data processing with R

Video pixel coordinates for the labels produced by DLC were imported into R Studio (Version 4.04 (2021-02-15) and processed with custom scripts that can be assessed here: https://github.com/rustlab1/DLC-Gait-Analysis [63]. Briefly, the accuracy values of individual videos were evaluated and data points with a low likelihood were removed. Representative videos were chosen to plot a general overview of the gait. Next, individual steps were identified within the run by the speed of the paws to identify the "stance" and "swing" phase. These steps were analyzed for synchronization, speed, length, and duration from the down view over a time course. Additionally, the angular positioning between the body center and the individual paws was measured. From the lateral/side view, we next measured average and total height differences of individual joins (y-coordinates) and the total movement, protraction, and retraction changes per step (x-coordinates) over the time course. Next, we measured angular variability (max, average, min) between neighboring joints including (hip-ankle-toe, iliac crest-hip-back-ankle, elbow-wrist-front toe, shoulder-elbow-wrist). More details on the parameter calculation can be found in Additional file 2: Table S1.

All >100 generated parameters were extracted to perform a random forest classification with scikit-learn [64] (ntree = 100, depth = max). We split our data into a training set and a test set (75%/25% split) and determined the Gini impurity-based feature importance. To evaluate the prediction accuracy that was generated on the training data, we cross-validated the predictions on the test data with a confusion matrix. The same procedure was also applied in a subgroup analysis between baseline vs. 3 dpi (acute injury) and baseline vs. 21 dpi (long-term recovery). The most five important parameters were used to perform a principal component analysis to demonstrate separation of these parameters.

Parameter calculations for pose estimation

Bottom analysis

Speed of steps The speed of a step was defined by the horizontal distance covered between two frames. Pixel units (1 cm = 37.79528 pixel) were converted in cm and frames converted to time (60 frames = 1s). Within a step, we classified stance and swing periods. A swing period was defined when the speed of a step was higher than 10 cm/s. The speed of steps was also used to determine individual steps within a run.

Duration of steps We calculated the average duration of steps from the number of frames it took to finish a step cycle. We further calculated the duration for each individual paw from the number of frames it took to start and finish a step cycle for each individual paw. The duration of a swing and a stance phase was determined by calculating the number of frames until a swing period was replaced by a stance period and vice versa.

Stride length The stride length was calculated as an average horizontal distance that was covered between two steps.

Synchronization We assume that for proper synchronization the opposite front and back paws (front-left and back right and front-right and back left) should be simultaneously in the stance position. We calculated the total time of frames when the paws are synchronized (totalSync) and the total time when the paws are not synchronized (totalNotSync). Then we used the formula: 1 – [totalSync/(totalSync + totalNotSync)]. A full synchronization would be the value of 0, the more it goes towards 1 the steps become asynchronous.

Side-view analysis

Average height and total vertical movement of individual body parts We calculated the height (differences in the *y* axis) of each tracked body part during a step from the left and right perspective. For the average height, we calculated the average value for the relative *y* coordinate within a step, whereas for the total vertical movement we subtracted the highest *y*-value from the lowest *y*-value during a course of a step.

Step length, length of protraction, and retraction We calculated the step length (differences in the *x* axis) of each tracked body part during a step from the left and right perspective. For the average step length, we calculated the average distance in the *x* coordinate covered by a step, whereas for the total horizontal movement we subtracted the highest *x*-value from the lowest *x*-value during a course of a step. We defined the phase of protraction when the *x*-coordinates of a paw between 2 frames was positive, and retraction was defined when the *x*-coordinate of a paw between 2 frames was negative. Then, we calculated the maximum distance (*x*-coordinate) between the beginning of the protraction and retraction.

Angles between body parts

The angles were calculated based on three coordinates. The position of body part 1 (P1), body part 2 (P2), and body part 3 (P3). The angle can be calculated using arctan formula using the *x* and *y* coordinates of each point, for example, the position of the left elbow (P1), left shoulder (P2), and left wrist (P3). The angle can be calculated using arctan function: Angle= atan2(P3.y − P1.y, P3.x − P1.x) − atan2(P2.y − P1.y, P2.x − P1.x). Details to all other angles between body parts can be found in Additional file 2: Table S1.

Additional details for each individual parameter calculation can be found in Additional file 2: Table S1 for each and in the Github code https://github.com/rustlab1/DLC-Gait-Analysis [63].

Statistical analysis

Statistical analysis was performed using RStudio (4.04 (2021-02-15). Sample sizes were designed with adequate power according to our previous studies [7, 42, 57] and to the literature [8, 12]. Overview of sample sizes can be found in Additional file 2: Table S2. All data were tested for normal distribution by using the Shapiro-Wilk test. Normally distributed data were tested for differences with a two-tailed unpaired one-sample *t*-test to compare changes between two groups (differences between ipsi- and contralesional sides). Multiple comparisons were initially tested for normal distribution with the Shapiro-Wilk test. The significance of mean differences between normally distributed multiple comparisons was assessed using repeated measures ANOVA with post-hoc analysis (p adjustment method = holm). For all continues measures: Values from different time points after stroke were compared to baseline values. Variables exhibiting a skewed distribution were transformed, using natural logarithms before the tests to satisfy the prerequisite assumptions of normality. Data are expressed as means ± SD, and statistical significance was defined as $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$. Boxplots indicate the 25 to 75% quartiles of the data (IQR). Each whisker extends to the furthest data point within the IQR range. Any data point further was considered an outlier and was indicated with a dot. Raw data, summarized data, and statistical evaluation can be found in the supplementary information (Additional file 2: Tables S3-S37).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-022-01434-9.

Additional file 1: Fig. S1. Walking profile of mice following DeepLabCut tracking. **Fig. S2.** Tracking of body parts in injured mice and mice with different genotypes. **Fig. S3.** Kinematic changes in spontaneous walk after stroke. **Fig. S4.** Angular variability between body center and front and hind paws. **Fig. S5.** Subgroup analysis for random forest classification and principal component analysis. **Fig. S6.** Principal component analysis and random forest classification of uninjured control mice. **Fig. S7.** Correlation plots between human annotators and DeepLabCut evaluation of the ladder rung test. **Fig. S8.** Overview of overhead for establishing DLC-assisted analysis. **Fig. S9.** Functional assessment of recovery after stroke using conventional behavioral tests. **Fig. S10.** Manual labeling of joints and body parts for DLC.

Additional file 2: Table S1. Name and description of all generated parameters following runway performance. **Table S2.** Overview of sample size for each experiment. **Tables S3-S37.** Raw data, summarized data sets, and statistical data of data sets generated in this study.

Acknowledgements

The authors acknowledge the initial help to set-up the DeepLabCut software by Dr. Oliver Sturman and Prof. Johannes Bohacek (ETH Zurich).

Authors' contributions

R.Z.W, C.T., and R.R. designed the study. R.Z.W., G.M., J.K., and R.R. performed the experiments and generated and analyzed data. R.Z.W., C.T., and R.R. wrote the manuscript. R.Z.W., J.K., C.T., and R.R. revised the manuscript. All authors edited the manuscript and approved the final version.

Funding

The authors RR and CT acknowledge support from the Mäxi Foundation and the 3R Competence Center (OC-2020-002). RR acknowledges the funding of SNF Swiss National Science Foundation (CRSK-3_195902) and Vontobel Foundation (1346/2021).

Availability of data and materials

The code with sample data are available at https://github.com/rustlab1/DLC-Gait-Analysis [63].

Declarations

Ethics approval and consent to participate

All animal studies were approved by Veterinarian Office of the Canton of Zurich (license: 209/2019).

Consent for publication

All authors consent to the publication of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute for Regenerative Medicine (IREM), University of Zurich, Campus Schlieren, Wagistrasse 12, 8952 Schlieren, Switzerland. [2]Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich, Switzerland. [3]Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland. [4]Burke Neurological Institute, White Plains, NY, USA.

Received: 24 March 2022 Accepted: 5 October 2022
Published online: 15 October 2022

References

1. GBD 2016 Stroke Collaborators. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019;18:439–58.
2. Thomalla G, Simonsen CZ, Boutitie F, Andersen G, Berthezene Y, Cheng B, et al. MRI-guided thrombolysis for stroke with unknown time of onset. N Engl J Med. 2018;379:611–22.
3. Nogueira RG, Jadhav AP, Haussen DC, Bonafe A, Budzik RF, Bhuva P, et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. N Engl J Med. 2018;378:11–21.
4. Katan M, Luft A. Global burden of stroke. Semin Neurol. 2018;38:208–11.
5. McDermott M, Skolarus LE, Burke JF. A systematic review and meta-analysis of interventions to increase stroke thrombolysis. BMC Neurol. 2019;19:86.
6. Chamorro Á, Dirnagl U, Urra X, Planas AM. Neuroprotection in acute stroke: targeting excitotoxicity, oxidative and nitrosative stress, and inflammation. Lancet Neurol. 2016;15:869–81.
7. Rust R, Grönnert L, Gantner C, Enzler A, Mulders G, Weber RZ, et al. Nogo-A targeted therapy promotes vascular repair and functional recovery following stroke. Proc Natl Acad Sci. 2019;116(28):14270–9.
8. Nih LR, Gojgini S, Carmichael ST, Segura T. Dual-function injectable angiogenic biomaterial for the repair of brain tissue following stroke. Nat Mater. 2018;17:642.
9. Rust R. Insights into the dual role of angiogenesis following stroke. J Cereb Blood Flow Metab Off J Int Soc Cereb Blood Flow Metab. 2020;40(6):1167–71.
10. Rust R, Kirabali T, Grönnert L, Dogancay B, Limasale YDP, Meinhardt A, et al. A practical guide to the automated analysis of vascular growth, maturation and injury in the brain. Front Neurosci. 2020;14:244.
11. Cheatwood JL, Emerick AJ, Schwab ME, Kartje GL. Nogo-A expression after focal ischemic stroke in the adult rat. Stroke. 2008;39:2091–8.
12. Wang Y, Zhao Z, Rege SV, Wang M, Si G, Zhou Y, et al. 3K3A-APC stimulates post-ischemic neuronal repair by human neural stem cells in mice. Nat Med. 2016;22:1050–5.
13. Kokaia Z, Llorente IL, Carmichael ST. Customized brain cells for stroke patients using pluripotent stem cells. Stroke. 2018;49:1091–8.
14. Rust R, Tackenberg C. Stem cell therapy for repair of the injured brain: five principles. Neuroscientist. 2022. https://doi.org/10.1177/10738584221110100.
15. von Ziegler L, Sturman O, Bohacek J. Big behavior: challenges and opportunities in a new era of deep behavior profiling. Neuropsychopharmacology. 2021;46:33–44.
16. Dell AI, Bender JA, Branson K, Couzin ID, de Polavieja GG, Noldus LPJJ, et al. Automated image-based tracking and its application in ecology. Trends Ecol Evol. 2014;29:417–28.
17. Basso DM, Beattie MS, Bresnahan JC. A sensitive and reliable locomotor rating scale for open field testing in rats. J Neurotrauma. 1995;12:1–21.
18. Kabra M, Robie AA, Rivera-Alba M, Branson S, Branson K. JAABA: interactive machine learning for automatic annotation of animal behavior. Nat Methods. 2013;10:64–7.
19. Berman GJ, Choi DM, Bialek W, Shaevitz JW. Mapping the stereotyped behaviour of freely moving fruit flies. J R Soc Interface. 2014;11:20140672.
20. Machado AS, Darmohray DM, Fayad J, Marques HG, Carey MR. A quantitative framework for whole-body coordination reveals specific deficits in freely walking ataxic mice. eLife. 2015;4:e07892.

21. Wiltschko AB, Johnson MJ, Iurilli G, Peterson RE, Katon JM, Pashkovski SL, et al. Mapping sub-second structure in mouse behavior. Neuron. 2015;88:1121–35.
22. Ben-Shaul Y. OptiMouse: a comprehensive open source program for reliable detection and analysis of mouse body and nose positions. BMC Biol. 2017;15:41.
23. Markowitz JE, Gillis WF, Beron CC, Neufeld SQ, Robertson K, Bhagat ND, et al. The striatum organizes 3D behavior via moment-to-moment action selection. Cell. 2018;174:44–58.e17.
24. Graving JM, Chae D, Naik H, Li L, Koger B, Costelloe BR, et al. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. eLife. 2019;8:e47994.
25. Pereira TD, Aldarondo DE, Willmore L, Kislin M, Wang SS-H, Murthy M, et al. Fast animal pose estimation using deep neural networks. Nat Methods. 2019;16:117–25.
26. Torabi R, Jenkins S, Harker A, Whishaw IQ, Gibb R, Luczak A. A neural network reveals motoric effects of maternal preconception exposure to nicotine on rat pup behavior: a new approach for movement disorders diagnosis. Front Neurosci. 2021;15:686767.
27. Inayat S, Singh S, Ghasroddashti A, Qandeel, Egodage P, Whishaw IQ, et al. A Matlab-based toolbox for characterizing behavior of rodents engaged in string-pulling. eLife. 2020;9:e54540.
28. Forys BJ, Xiao D, Gupta P, Murphy TH. Real-time selective markerless tracking of forepaws of head fixed mice using deep neural networks. eNeuro. 2020;7(3):ENEURO.0096-20.202.
29. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci. 2018;21:1281–9.
30. Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. ArXiv160503170 Cs; 2016.
31. Labuguen R, Matsumoto J, Negrete SB, Nishimaru H, Nishijo H, Takada M, et al. MacaquePose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture. Front Behav Neurosci. 2021;14:581154.
32. Sturman O, von Ziegler L, Schläppi C, Akyol F, Privitera M, Slominski D, et al. Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. Neuropsychopharmacology. 2020;45:1942–52.
33. Deng J, Dong W, Socher R, Li L-J, Li K, Li F-F. ImageNet: a large-scale hierarchical image database; 2009. p. 248–55.
34. Nath T, Mathis A, Chen AC, Patel A, Bethge M, Mathis MW. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nat Protoc. 2019;14:2152–76.
35. Williams S, Zhao Z, Hafeez A, Wong DC, Relton SD, Fang H, et al. The discerning eye of computer vision: can it measure Parkinson's finger tap bradykinesia? J Neurol Sci. 2020;416:117003.
36. Worley NB, Djerdjaj A, Christianson JP. Convolutional neural network analysis of social novelty preference using DeepLabCut. bioRxiv. 2019:736983. https://www.biorxiv.org/content/10.1101/736983v2.
37. Clemensson EKH, Abbaszadeh M, Fanni S, Espa E, Cenci MA. Tracking rats in operant conditioning chambers using a versatile homemade video camera and DeepLabCut. J Vis Exp. 2020. https://doi.org/10.3791/61409.
38. Ryait H, Bermudez-Contreras E, Harvey M, Faraji J, Mirza Agha B, Gomez-Palacio Schjetnan A, et al. Data-driven analyses of motor impairments in animal models of neurological disorders. PLoS Biol. 2019;17:e3000516.
39. Corbett D, Carmichael ST, Murphy TH, Jones TA, Schwab ME, Jolkkonen J, et al. Enhancing the alignment of the preclinical and clinical stroke recovery research pipeline: consensus-based core recommendations from the Stroke Recovery and Rehabilitation Roundtable translational working group. Int J Stroke. 2017;12:462–71.
40. Zörner B, Filli L, Starkey ML, Gonzenbach R, Kasper H, Röthlisberger M, et al. Profiling locomotor recovery: comprehensive quantification of impairments after CNS damage in rodents. Nat Methods. 2010;7:701–8.
41. Preisig DF, Kulic L, Krüger M, Wirth F, McAfoose J, Späni C, et al. High-speed video gait analysis reveals early and characteristic locomotor phenotypes in mouse models of neurodegenerative movement disorders. Behav Brain Res. 2016;311:340–53.
42. Weber RZ, Grönnert L, Mulders G, Maurer MA, Tackenberg C, Schwab ME, et al. Characterization of the blood brain barrier disruption in the photothrombotic stroke model. Front Physiol. 2020;11:586226.
43. Bellardita C, Kiehn O. Phenotypic characterization of speed-associated gait changes in mice reveals modular organization of locomotor networks. Curr Biol. 2015;25:1426–36.
44. Aljovic A, Zhao S, Chahin M, de la Rosa C, Steenbergen VV, Kerschensteiner M, et al. A deep-learning-based toolbox for Automated Limb Motion Analysis (ALMA) in murine models of neurological disorders. Commun Biol. 2022;5(1):131.
45. Metz GA, Whishaw IQ. The ladder rung walking task: a scoring system and its practical application. J Vis Exp. 2009;(28):1204.
46. Zhu LL, Lindenberg R, Alexander MP, Schlaug G. Lesion load of the corticospinal tract predicts motor impairment in chronic stroke. Stroke J Cereb Circ. 2010;41:910–5.
47. Levin MF, Kleim JA, Wolf SL. What do motor "recovery" and "compensation" mean in patients following stroke? Neurorehabil Neural Repair. 2009;23:313–9.
48. MacLellan CL, Langdon KD, Botsford A, Butt S, Corbett D. A model of persistent learned nonuse following focal ischemia in rats. Neurorehabil Neural Repair. 2013;27:900–7.
49. Cai S, Li G, Zhang X, Huang S, Zheng H, Ma K, et al. Detecting compensatory movements of stroke survivors using pressure distribution data and machine learning algorithms. J Neuroeng Rehabil. 2019;16:131.
50. Caballero-Garrido E, Pena-Philippides JC, Galochkina Z, Erhardt E, Roitbak T. Characterization of long-term gait deficits in mouse dMCAO, using the CatWalk system. Behav Brain Res. 2017;331:282–96.
51. Lindau NT, Bänninger BJ, Gullo M, Good NA, Bachmann LC, Starkey ML, et al. Rewiring of the corticospinal tract in the adult rat after unilateral stroke and anti-Nogo-A therapy. Brain. 2014;137:739–56.
52. Kaiser J, Maibach M, Salpeter I, Hagenbuch N, de Souza VBC, Robinson MD, et al. The spinal transcriptome after cortical stroke: in search of molecular factors regulating spontaneous recovery in the spinal cord. J Neurosci. 2019;39:4714–26.
53. Murphy TH, Corbett D. Plasticity during stroke recovery: from synapse to behaviour. Nat Rev Neurosci. 2009;10:861–72.
54. Hsu AI, Yttri EA. An open source unsupervised algorithm for identification and fast prediction of behaviors. Nat Commun. 2021;12(1):5188.
55. Wiltschko AB, Tsukahara T, Zeine A, Anyoha R, Gillis WF, Markowitz JE, et al. Revealing the structure of pharmacobehavioral space through motion sequencing. Nat Neurosci. 2020;23:1433–43.
56. Labat-gest V, Tomasi S. Photothrombotic ischemia: a minimally invasive and reproducible photochemical cortical lesion model for mouse stroke studies. J Vis Exp. 2013. https://doi.org/10.3791/50370.
57. Rust R, Weber RZ, Grönnert L, Mulders G, Maurer MA, Hofer A-S, et al. Anti-Nogo-A antibodies prevent vascular leakage and act as pro-angiogenic factors following stroke. Sci Rep. 2019;9:1–10.
58. Weber R, Bodenmann C, Uhr D, Zurcher K, Wanner D, Generali M, et al. Intracerebral transplantation and in vivo bioluminescence tracking of human neural progenitor cells in the mouse brain | Protocol. J Vis Exp. 2022; (in press).
59. Rust R, Weber RZ, Generali M, Kehl D, Bodenmann C, Uhr D, et al. Xeno-free induced pluripotent stem cell-derived neural progenitor cells for in vivo applications. 2022.
60. Weber RZ, Mulders G, Perron P, Tackenberg C, Rust R. Molecular and anatomical roadmap of stroke pathology in immunodeficient mice; 2022. p. 2022.07.28.501836.
61. Roome RB, Vanderluit JL. Paw-dragging: a novel, sensitive analysis of the mouse cylinder test. J Vis Exp. 2015:e52701. https://doi.org/10.3791/52701.
62. Chen C-C, Gilmore A, Zuo Y. Study motor skill learning by single-pellet reaching tasks in mice. J Vis Exp. 2014:e51238. https://doi.org/10.3791/51238.
63. Rust R. rustlab1/DLC-Gait-Analysis: Zenodo; 2022. https://zenodo.org/record/6499925#.YxnxsexByrM. Accessed 8 Sept 2022
64. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

Publisher's Note