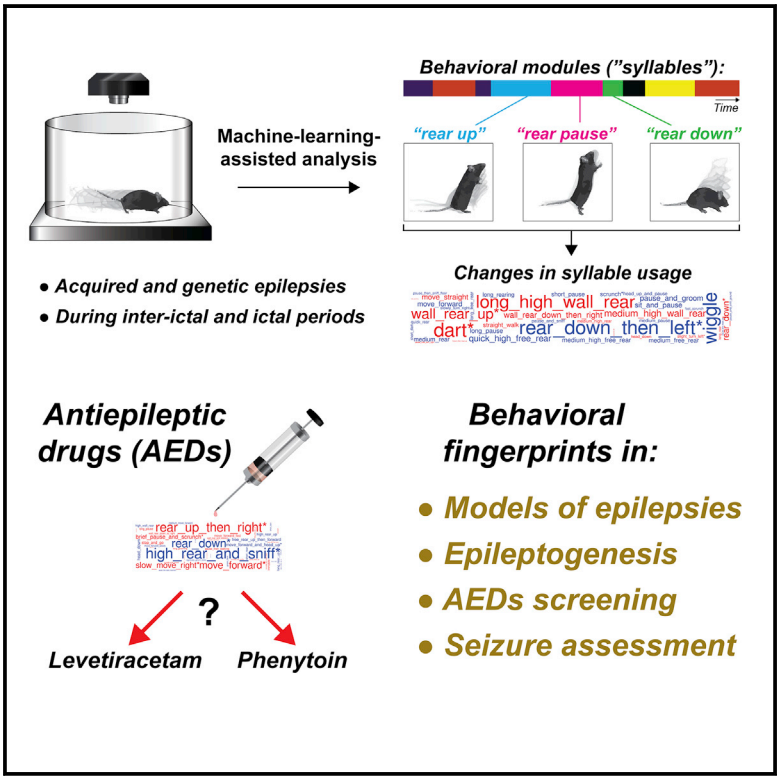
Neuron

Hidden behavioral fingerprints in epilepsy

Graphical abstract



Highlights

- Automated behavioral analysis enables high-throughput screening of epileptic mice
- Characteristic behavioral phenotypes are found in acquired and genetic epilepsies
- Inter-ictal behavior can be used to assess epileptogenesis and for drug screening
- A purely data-driven analysis can facilitate seizure assessment

Authors

Tilo Gschwind, Ayman Zeine, Ivan Raikov, ..., Lori L. Isom, Sandeep Robert Datta, Ivan Soltesz

Correspondence

gschwind@stanford.edu

In brief

Gschwind et al. show that machine learning-assisted behavioral analysis allows an unbiased assessment of epilepsy in animal models. The automated identification of behavioral phenotypes, including anti-epileptic drug responses, during the readily available inter-ictal periods bears great potential to accelerate rigorous, reproducible preclinical research into epilepsies.

Gschwind et al., 2023, Neuron *111*, 1440–1452
May 3, 2023
https://doi.org/10.1016/j.neuron.2023.02.003

CellPress

CellPress
OPEN ACCESS

Neuron

Article

Hidden behavioral fingerprints in epilepsy

Tilo Gschwind,[1,6,*] Ayman Zeine,[2] Ivan Raikov,[1] Jeffrey E. Markowitz,[2] Winthrop F. Gillis,[2] Sylwia Felong,[1] Lori L. Isom,[3,4,5] Sandeep Robert Datta,[2] and Ivan Soltesz[1]

[1]Department of Neurosurgery, Stanford University, Stanford, CA 94305, USA
[2]Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA
[3]Department of Pharmacology, University of Michigan, Ann Arbor, MI 48109, USA
[4]Department of Neurology, University of Michigan, Ann Arbor, MI 48109, USA
[5]Department of Molecular & Integrative Physiology, University of Michigan, Ann Arbor, MI 48109, USA
[6]Lead contact
*Correspondence: gschwind@stanford.edu
https://doi.org/10.1016/j.neuron.2023.02.003

SUMMARY

Epilepsy is a major disorder affecting millions of people. Although modern electrophysiological and imaging approaches provide high-resolution access to the multi-scale brain circuit malfunctions in epilepsy, our understanding of how behavior changes with epilepsy has remained rudimentary. As a result, screening for new therapies for children and adults with devastating epilepsies still relies on the inherently subjective, semi-quantitative assessment of a handful of pre-selected behavioral signs of epilepsy in animal models. Here, we use machine learning-assisted 3D video analysis to reveal hidden behavioral phenotypes in mice with acquired and genetic epilepsies and track their alterations during post-insult epileptogenesis and in response to anti-epileptic drugs. These results show the persistent reconfiguration of behavioral fingerprints in epilepsy and indicate that they can be employed for rapid, automated anti-epileptic drug testing at scale.

INTRODUCTION

There are 65 million people worldwide with epilepsy and 150,000 new cases of epilepsy are diagnosed every year in the US.[1] Treatment options for children and adults with epilepsies remain inadequate, as many patients suffer from uncontrolled seizures, cognitive or neurobehavioral comorbidities, and the negative side effects of treatment.

Preclinical-translational research in epilepsy relies on continuous video-electroencephalogram (EEG) monitoring for multiple stages of investigation (e.g., as a diagnostic tool, as well as to probe mechanistic insights and therapeutic outcomes), and yet, the toolbox falls short of capturing the complexity and heterogeneity of epilepsy. Continuous video-EEG can capture seizure burden across various timescales (e.g., ultradian and circadian dynamics[2,3]) and disease stages (e.g., epileptogenesis[4]). However, it is frequently accompanied by labor-intensive, inherently subjective scoring of a handful of pre-selected behavioral manifestations of epilepsy. These phenotypes range from subtle (e.g., repeated head nodding and forelimb clonus) to overt (e.g., frank motor seizures), which are quantified by human observers along the semi-quantitative "Racine scale," first introduced in 1972.[5] Furthermore, most behavioral investigations are restricted to the observable ictal periods, while inter-ictal periods in these datasets are largely ignored and lack any consensus annotation of epileptic behavior. Yet, there is emerging evidence that ictal and inter-ictal impairments often involve the same network[6–8] and careful decomposition of behavior can reveal novel insights into such functional networks.[9–11] Thus, the rich behavioral repertoire represents an untapped source of disease-relevant insights in epilepsies and enables testing of new therapeutics in a purely data-driven manner.

To progress toward evidence-based, reproducible preclinical-translational epilepsy research, we explored the transformative potential of state-of-the-art machine learning-assisted three-dimensional (3D) video analysis to phenotype mice with acquired and genetic epilepsies and screen for on- and off-target effects of anti-epileptic drugs (AEDs) in an automated and high-throughput manner. Specifically, we used the recently developed motion sequencing (MoSeq)[12] approach combining 3D imaging and unsupervised machine learning to deconstruct at sub-second timescales the behavior of different mouse models of epilepsies into stereotyped, recurrently occurring modules (sub-second behavioral "syllables") that are arranged according to specific transition rules ("grammar"). In contrast to the visual examination of complex behaviors by the experimenter subject to potential observer bias or to the arbitrary selection of a few particular behavioral measures (e.g., speed) for assessment, MoSeq analyzes behavior in a purely data-driven manner and can handle large datasets at high throughput. Taken together, our experiments revealed hidden behavioral phenotypes for different seizure and epilepsy types, epileptogenic time points, sex, drug type, and drug doses, allowing the automated assessment of epilepsy without the need for recording electrographic activity in each animal for prolonged periods of time. These

insights indicate that the persistently reconfigured behavioral syllables of epilepsy can be targeted for assessment at scale in acquired and genetic models to accelerate rigorous, reproducible preclinical research into epilepsies.

RESULTS

Automated analysis of sub-second 3D pose dynamics outperforms classical approaches in identifying and assessing epileptic mice during inter-ictal periods

First, we examined if we could uncover distinct phenotypes in mouse models of acquired and genetic epilepsies (Figure 1A) without the need for invasive EEG monitoring, batteries of behavioral testing, and seizure-susceptibility experiments. For acquired epilepsy, we used the unilateral intrahippocampal kainic acid (IHKA) model of chronic temporal lobe epilepsy (TLE), which is known to reproduce key histological, electrographic, and cognitive hallmarks of human TLE, the most prevalent form of intractable epilepsy in adults.[6,13] Our MoSeq-based analysis of the 3D video recordings from 60 min recording sessions during the inter-ictal period (i.e., without the occurrence of the rare overt behavioral seizures) revealed robust differences between the behavioral repertoires of sham-injected control (CON) and IHKA mice, with significant up- and downregulation of specific syllables such as "dart" (Video S1) and "scrunch," respectively (Figure 1B; see STAR Methods for syllable naming). Alongside the observed difference in syllable usage, control and epileptic mice also showed distinct changes in the transitions between two syllables (i.e., bigram transition probability), especially for syllables that were differentially expressed between the groups (Figure S1B). However, the overall predictability (i.e., entropy rate; see STAR Methods) of mouse behavior remained unchanged (Figure S1C), suggesting that IHKA did not profoundly alter the grammatical microstructure of behavior (see also Figure S1A).

Next, in order to determine whether the inter-ictal phenotype of IHKA mice is also evident to a human observer, four epilepsy researchers were shown video clips from epileptic and non-epileptic animals in a blinded manner and tasked to allocate them to the appropriate group. In a human vs. machine comparison, a linear classifier trained on MoSeq syllables ($F_1 = 0.79 \pm 0.29$) not only outperformed the classification based on common behavioral measures such as position ($F_1 = 0.47 \pm 0.35$), speed ($F_1 = 0.60 \pm 0.35$), or a combination of multiple 2D measures (length, height, speed, and position; "scalars"; $F_1 = 0.60 \pm 0.35$) but also surpassed these trained experimenters ($F_1 = 0.57 \pm 0.10$) in identifying epileptic from non-epileptic animals (see details about F_1 scores in STAR Methods; see also Figure 1C for classification matrices). This highlights the power of automated, objective behavioral analysis focusing on sub-second, stereotyped 3D pose dynamics to identify and assess epileptic animals even in the inter-ictal period, without the need to monitor animals for days or weeks to observe and count the usually infrequent behavioral seizures.

Behavioral syllables unveil a previously unrecognized sex-specific behavioral phenotype in an animal model of Dravet syndrome

Next, we examined the behavioral phenotype of a genetic epilepsy model, specifically a mouse model carrying a mutation in the sodium channel β1 subunit (*SCN1B*),[14,15] a model linked to Dravet syndrome, a form of severe developmental and epileptic encephalopathy in children. Interestingly, heterozygous mutations in *SCN1B* have been reported in patients with generalized epilepsy with febrile seizures plus type 1 (GEFS+1), exhibiting a mild-to-moderate range of seizure severity and a combination of seizure types that include febrile seizures, early-onset absence epilepsy, and focal epilepsies.[16–18] Although homozygous *Scn1b* null mice show severe ataxia, seizures, and early mortality (100% by postnatal day 25), heterozygous littermates lack any known discernible behavioral phenotype or change in seizure susceptibility, despite the significant reduction in total brain *Scn1b* mRNA.[14,19] In spite of the lack of previously reported differences in behaviors, MoSeq succeeded in revealing a selective change in behavior exclusively in female *SCN1b*$^{+/-}$ mice, with significant upregulation of syllables such as "head up" movements and downregulation of other syllables such as "dart" (Figure 1D). Male *SCN1b*$^{+/-}$ mice, on the other hand, showed no significant differences compared with their wild-type littermates (Figure 1E). These results reveal a previously unrecognized sex-specific behavioral alteration in mice with heterozygous deletion of *Scn1b*, illustrating the value of automated, data-driven behavioral analysis in genetic epilepsies.

Distinct behavioral phenotypes at different time points during epileptogenesis in a mouse model of temporal lobe epilepsy

We further determined whether there are distinguishable time-dependent behavioral phenotypes during the first few weeks after insult (Figure 2), which would complement long-standing efforts to study the mechanisms underlying the development of spontaneous recurrent seizures and identifying targeted treatment strategies for early interventions during epileptogenesis. As previously described,[3,20,21] the emergence of frequent spontaneous electrographic seizures in the murine IHKA model of chronic TLE is preceded by a process of pathogenesis with distinct electrographic signatures and histopathological changes. In short, IHKA injection causes an immediate *status epilepticus* (lasting up to 24 h), followed by a period of about 1–2 weeks characterized by hippocampal neuronal death, increased inflammation, synaptic reorganization, and the occurrence of isolated or grouped low voltage spikes and spikes-and-waves.[20,21] We tested whether the behavioral repertoire can be used to discriminate among time points after insult by comparing single 60 min 3D video recordings sampled every week after IHKA for a month (i.e., after 1–4 weeks post injection; Figure 2A; see Figure 2B for the syllable usage of kainic acid-injected mice, IHKA, and controls, CON). A linear classifier trained on syllable usage revealed that the modular description of behavior provided by MoSeq distinguished mouse behaviors during the different weeks after insult while also outperforming commonly used metrics such as position, speed, or a combination of different scalars (see Figure 2C for classification matrices). Specifically, syllable usage was found to be better than any of the scalar measures for discriminating between experimental conditions (CON or IHKA) and time points (1–4 weeks) (position $F_1 = 0.51 \pm 0.16$; speed $F_1 = 0.56 \pm 0.16$; scalars $F_1 = 0.51 \pm 0.16$; and MoSeq $F_1 = 0.69 \pm 0.15$; see Figure 2E for group-specific

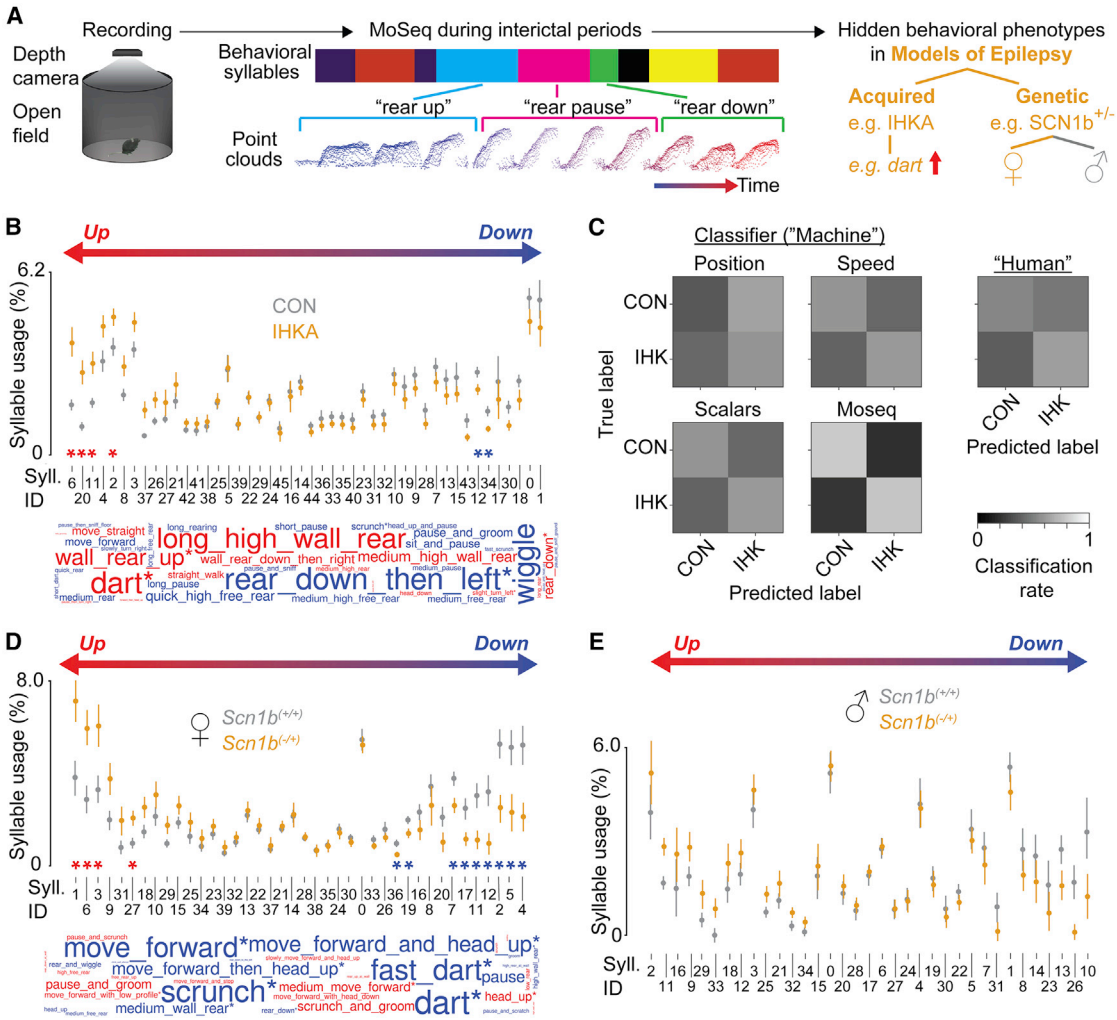


Figure 1. Hidden behavioral phenotypes in mouse models of acquired and genetic epilepsies during inter-ictal periods

(A) Experimental paradigm for artificial intelligence (AI)-guided behavioral phenotyping in epilepsy during inter-ictal periods. Mouse 3D pose dynamics (illustrated by point clouds) in an open-field assay are captured with a depth camera and analyzed with MoSeq. Behavioral syllables identified by MoSeq reveal hidden behavioral phenotypes in two distinct well-established mouse models of epilepsy, distinguishing sham mice from animals injected with intrahippocampal kainic acid (IHKA), a model of temporal lobe epilepsy (TLE), and identifying a previously unreported sex-specific phenotype in *Scn1b* mice, a model linked to Dravet syndrome. Note that syllable IDs are unique for each experiment and do not correspond to the same syllables across (B), (D), and (E).

(B) Syllable usage in mice (top), obtained four weeks after intrahippocampal injection with either saline (CON; n = 18) or kainic acid (IHKA; n = 20), is ordered by differential usage (arrow) with the most IHKA-upregulated (up) syllables on the left and IHKA-downregulated (down) syllables on the right. A word cloud (bottom) with syllable names color-coded (up and downregulated in IHKA, red and blue, respectively) and sized by the relative difference in syllable usage. In a model of chronic acquired epilepsy, IHKA mice exhibit a distinct behavioral repertoire compared with CON mice, with selective up- and downregulation of syllables such as "dart" (ID 20) and "scrunch" (ID 12), respectively. Asterisks indicate a significant change in syllable usage (Kruskal-Wallis and post hoc Dunn's two-sided test with permutation with Benjamini-Hochberg false discovery rate of $\alpha = 0.05$). Error bars indicate 95% bootstrap confidence intervals. See also Figure S1 and Video S1.

(C) Normalized classification matrices (across rows and columns) showing the performance of a linear classifier in discriminating between epileptic (IHKA) and non-epileptic animals (CON). Left: "machine." Classifiers were trained using one of the following behavioral measures: position, speed, combined 2D measures ("scalars"), or MoSeq syllables. Right: "human." The classification matrix summarizes the performance of four experimenters tasked to distinguish epileptic and non-epileptic animals (see results for details). The color bar is shared between the left ("machine") and right ("human"). An ideal classifier performance corresponds to a diagonal white with otherwise black fields (classification rate of 1).

(D) Same as in (B) but comparing female *Scn1b*$^{+/-}$ mice (n = 10) and *Scn1b*$^{+/+}$ littermates (n = 20). In a genetic model of developmental and epileptic encephalopathy, *Scn1b*$^{+/-}$ mice exhibit a distinct behavioral repertoire compared with *Scn1b*$^{+/+}$ littermates, with selective up- and downregulation of syllables such as "head up" and "dart," respectively. Kruskal-Wallis and post hoc Dunn's two-sided test with permutation were used, with Benjamini-Hochberg false discovery rate with $\alpha = 0.05$. Error bars indicate 95% bootstrap confidence intervals.

(E) Same as in (D) but comparing male *Scn1b*$^{+/-}$ mice (n = 14) and *Scn1b*$^{+/+}$ littermates (n = 14). Unlike female *Scn1b*$^{+/-}$ mice (D), male *Scn1b*$^{+/-}$ mice do not exhibit a distinct behavioral repertoire compared with control male littermates. Kruskal-Wallis and post hoc Dunn's two-sided test with permutation were used, with Benjamini-Hochberg false discovery rate with $\alpha = 0.05$. Error bars indicate 95% bootstrap confidence intervals.

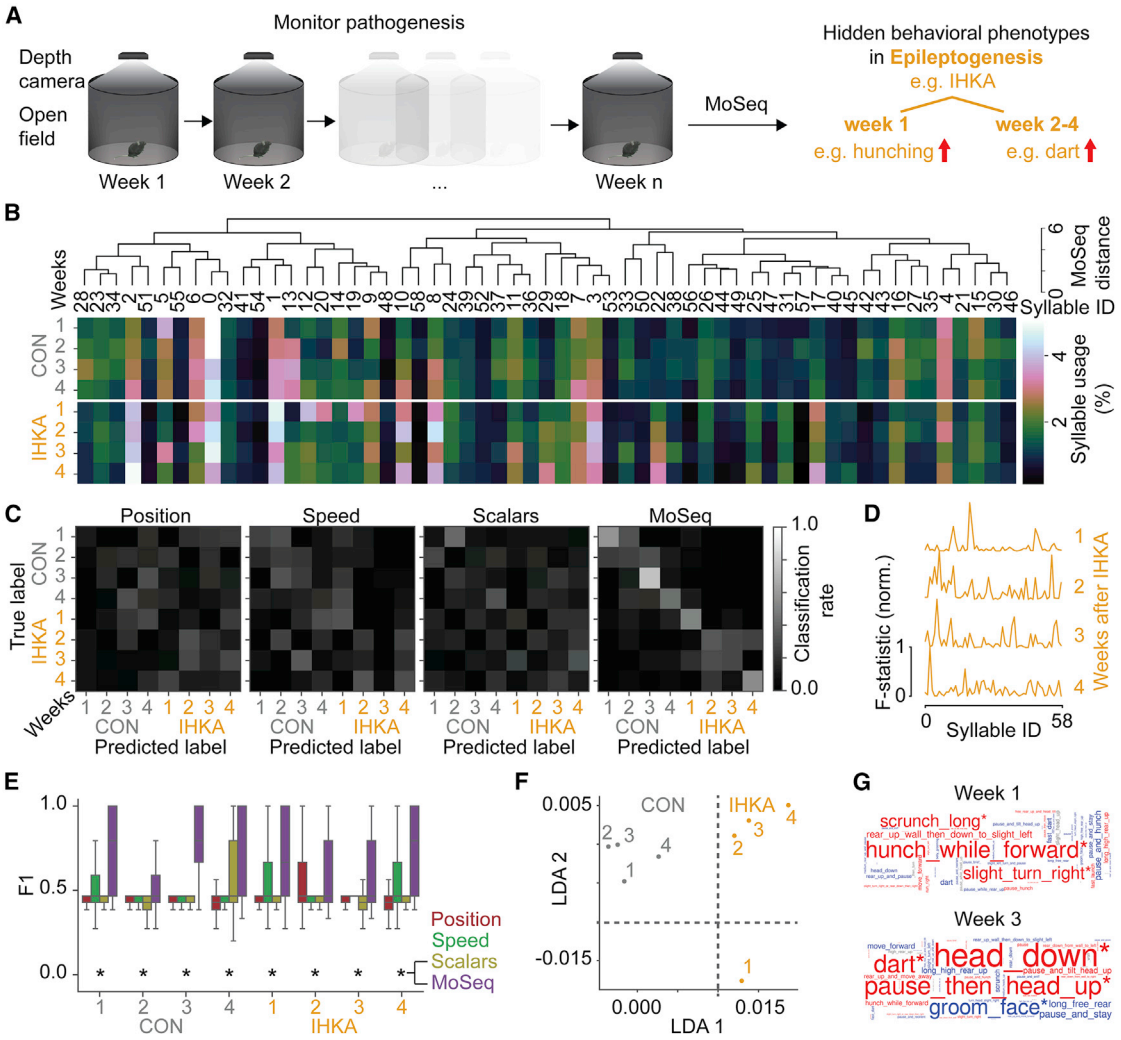


Figure 2. Distinct behavioral phenotypes during epileptogenesis in a mouse model of TLE

(A) Sketch illustrating the experimental paradigm to use AI-guided behavioral phenotyping to monitor the pathogenesis in animal models of epilepsies. Here, a 60 min 3D video was recorded every week after induction of *status epilepticus* in the IHKA model of TLE over the period of a month, revealing distinct phenotypes in the first and later weeks.

(B) Dendrogram of syllables indicating the MoSeq distance between syllables (top), and heatmap of syllable usage (bottom) in mice recorded 1–4 weeks after intrahippocampal injection with either saline (CON; n = 10) or kainic acid (IHKA; n = 10).

(C) Normalized classification matrices (across rows and columns) summarizing the performance of a linear classifier in distinguishing experimental conditions (CON or IHKA) and time points (1–4 weeks) based on different behavioral measures (position, speed, combined scalar measures, or MoSeq syllables). An ideal classifier performance corresponds to a diagonal white with otherwise black fields (classification rate of 1).

(D) Normalized *F* statistic highlighting the relevance of each indicated syllable for discriminating time points after IHK from CON (“behavioral fingerprint”; see STAR Methods for details).

(E) F_1 scores for linear classifiers distinguishing experimental conditions (CON or IHKA) and time points (1–4 weeks) based on different behavioral measures. Box plots represent the distribution across all cross-validation folds, with whiskers representing 1.5 times the inter-quartile range ($p < 0.01$, asterisks indicate significant differences between MoSeq and scalars, paired two-sided t test corrected with Holm-Bonferroni step-down procedure).

(F) Linear discrimination analysis (LDA) plot indicating the similarity of mean behavioral summaries of mice within conditions (CON or IHK) and within time points (1–4 weeks). Dashed lines highlight the separation between conditions (CON vs. IHK) and clusters of time points after IHKA (IHKA week 1 vs. IHKA weeks 2–4) along the LDA-1- and LDA-2-axis, respectively.

(G) Word cloud with syllable names color-coded (up- and downregulated in IHKA, red, and blue, respectively) and sized by the normalized *F* statistics in one vs. CON comparison (see Figure 2D). IHKA mice exhibit a distinct behavioral repertoire compared with CON mice during week 1 and, for example, week 3 (representative of later weeks 2–4), with selective upregulation of syllables such as “scrunch long” and “dart,” respectively. Asterisks indicate significant syllables (Holm-Bonferroni-corrected $p < 0.01$ from the two-sided *F*-test).

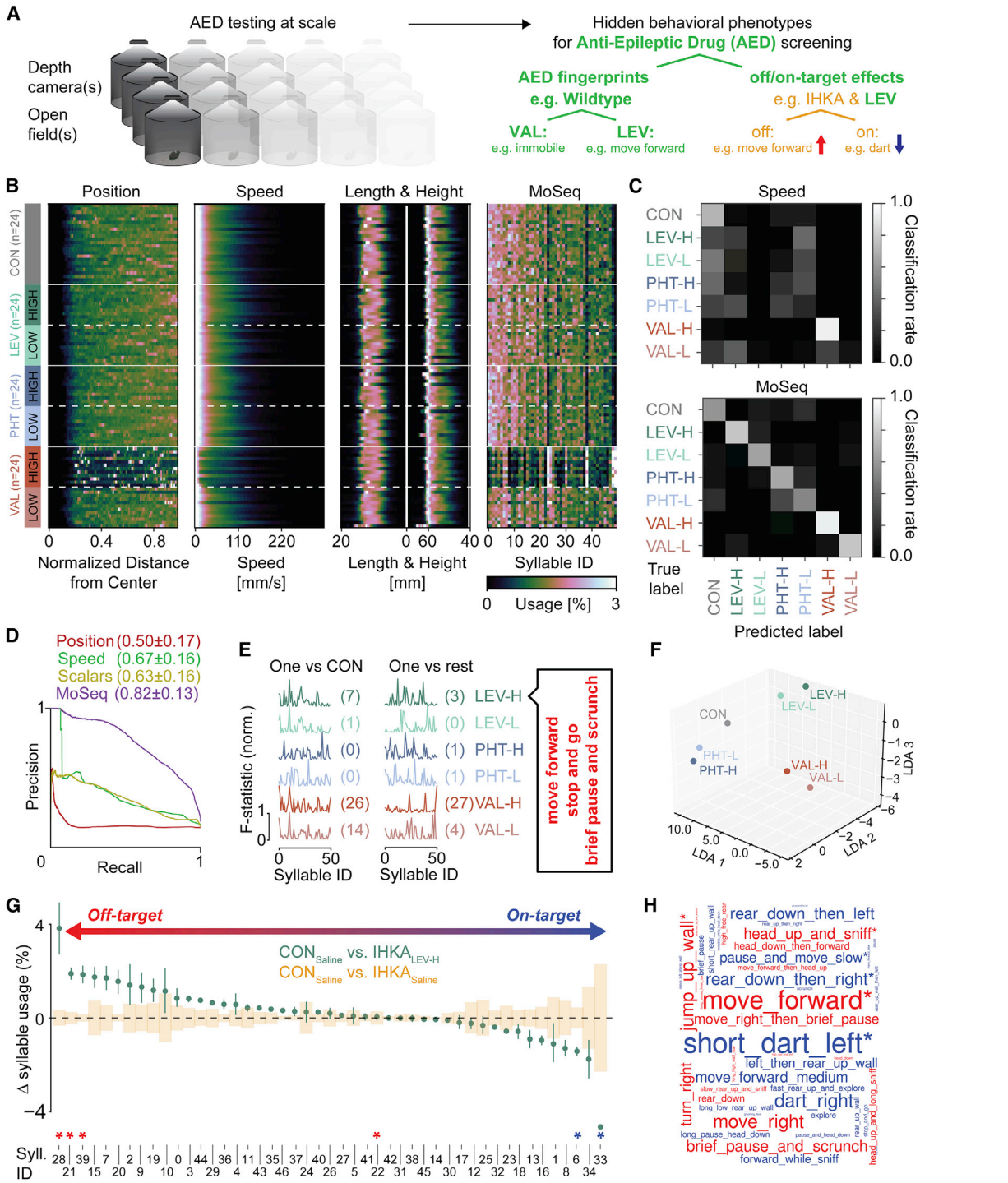


Figure 3. Hidden behavioral phenotypes in epilepsy for anti-epileptic drug screening at scale

(A) Schematic illustrating the pipeline for anti-epileptic drug (AED) screening at scale. Multiple open-field assays can be run in parallel (e.g., 4 setups in this study). Deploying this pipeline identified unique behavioral phenotypes for different drug-dose pairs, including levetiracetam (LEV), phenytoin (PHT), and valproic acid (VAL) in wild-type mice, while revealing on- and off-target effects of LEV in the IHKA mouse model of TLE. See also Figure S2 and Table S1.

(B) Behavioral summary for different AEDs. Wild-type mice were injected with an either high or low dose of levetiracetam (LEV-H or LEV-L; n = 12 each), phenytoin (PHT-H or PHT-L; n = 12 each), and valproic acid (VAL-H or VAL-L; n = 12 each) or control solution (CON; n = 24; see STAR Methods for details). From left to right, the position (normalized by the arena center position), velocity, length, and height, as well as MoSeq-identified syllable usages, were computed for each mouse (rows).

(legend continued on next page)

F_1 scores). To better understand and visualize the relationship between the behavioral repertoire of different groups, we embedded MoSeq behavioral summaries into a 2D space using linear discrimination analysis (LDA). Although CON and IHKA mice separated along one axis of the LDA space (LDA-1), recordings of IHKA mice from weeks 2–4 clustered together and separated along the other axis (LDA-2) from recordings from week 1 (Figure 2F). Similarly, when we compared the relevance of individual syllables for discriminating weeks after IHKA, the subsets of syllables identified by a normalized *F* statistic (the behavioral "fingerprint") that best summarized the behavioral phenotypes of weeks 2–4 were more similar between each other, and less so compared with those of week 1 after IHKA injection (Figure 2D). Discrimination-relevant syllables during the later weeks (e.g., week 3) included fast dart movements and the syllable "head down," pointing toward an impulsive phenotype, whereas syllables associated with more lethargic behaviors were characteristic of the first week after insult (e.g., long-lasting scrunches and hunching during a forward movement) (Figure 2G).

Behavioral fingerprints accelerate objective antiepileptic drug screening

Uncovering hidden behavioral phenotypes in animal models of epilepsies and at different time points during epileptogenesis in a fully automated and unbiased manner opens the opportunity to reliably test both established and candidate therapeutics rapidly and at scale. Building on the success of MoSeq in distinguishing different psychiatric medications,[22] we determined whether there are behavioral fingerprints of commonly used AEDs that not only correspond to drug types but perhaps even also to drug doses and on- and off-target effects (Figure 3A). We first tested if we could identify sets of syllables characteristic of established AEDs and doses, including levetiracetam (LEV), phenytoin (PHT), and valproic acid (VAL), in non-epileptic animals during random open-field exploration (Figures 3B–3F). In a comparison of linear classifiers trained on different behavioral metrics, MoSeq outperformed all scalar metrics in discriminating between different drugs (Figures S2A–S2C) and between drug-dose pairs (Figure 3D). While an obvious phenotype such as the one associated with high-dose valproic acid could easily be identified by both MoSeq and scalar measures (see VAL-H in the classification matrix of MoSeq and the best performing scalar speed in Figure 3C), scalar measures struggled with more subtle phenotypes (for a detailed comparison between MoSeq and scalars, see the overview of the different scalar measures and syllable usage of every mouse in Figure 3B).

We further explored to what extent behavioral syllables capture similarities and distinct characteristics of different AED dose pairs. By iteratively removing a single drug-dose pair from our dataset ("held-out" data) and training a linear classifier on the remaining data, we found that the highest classification rate for the held-out data remained within its drug label (e.g., held-out LEV-H predicted as LEV-L and vice versa) (Figure S2D), indicating that doses of the same drug elicited a similar syllable repertoire. The behavioral similarity between doses of a given drug was also evident in the clustering of drug-dose pairs within the 3D embedding space of an LDA (Figure 3F) as well as in the hierarchical clustering of pairwise cosine distances between the syllable repertoire of different drug-dose pairs (Figure S2E). To capture the distinct characteristics of a given AED treatment, we compared the behavioral fingerprints (i.e., the set of syllables identified by a normalized *F* statistic; Figure 3E) of different drug-dose pairs and found a set of significant syllables that discriminated a given pair from either the control treatment (one vs. CON; Figure 3E left) or all other treatments (one vs. rest; Figure 3E right) for the majority of drug-dose pairs. Behavioral fingerprints captured the phenotype of a given AED treatment and indicated similarities between doses of the same drug and highlighted differences between drugs, providing a comprehensive, yet intuitive description of the underlying behavioral phenotype (see plots of normalized *F* statistic in Figure 3E and word clouds in Figure S2F). For example, LEV-H could be distinguished from other AED dose pairs by only three syllables, including the syllable "move forward," which was significantly increased compared with controls (Figures 3E and S2F; see also Table S1 for additional information about the syllables in this dataset).

Next, we tested whether on- and off-target effects of AED treatment could be identified in a model of TLE (Figures 3G and 3H). The murine IHKA model with its high frequency of non-convulsive

(C) Normalized classification matrices (across rows and columns) representing the performance of a linear classifier (i.e., means of all cross-validation folds) for discriminating different drug-dose pairs based on speed (top) or MoSeq-identified syllable usage (bottom). Both classifiers—trained on either MoSeq or speed (the best performing scalar measure; see results)—showed high performance for high-dose valproic acid (VAL-H), which is known to induce an overt behavioral phenotype. However, MoSeq-based classification outperformed those of scalar measures otherwise (see results for details).

(D) Mean precision-recall curves and F_1 values (including standard error) for different behavioral measures across all drug treatments. The corresponding area under the curve (AUC) is 0.16 (Position), 0.39 (Speed), 0.35 (Scalars), and 0.76 (MoSeq).

(E) Normalized *F* statistics ("behavioral fingerprints"; see STAR Methods for details) highlighting the relevance of each indicated syllable for distinguishing a given drug-dose pair either from the control treatment (one vs. CON) or all other treatments (one vs. rest). The number of significant syllables is indicated in parentheses (Holm-Bonferroni-corrected $p < 0.01$ from the two-sided *F*-test). As an example, the text bubble names the three significantly upregulated syllables for LEV-H in the one vs. rest comparison that distinguish LEV-H from all other drugs.

(F) LDA plot indicating the similarity between the mean behavioral summaries of mice across drug-dose pairs.

(G) Difference in syllable usage between non-epileptic (CON) and chronically epileptic (IHKA) mice, which were intraperitoneally injected with either saline or high-dose levetiracetam (LEV-H; see STAR Methods for details). Values are normalized to the difference between CON_{Saline} and $IHKA_{Saline}$, which are aligned at zero (orange-shaded rectangles indicate the error bars). The syllables are ordered by the change in usage (arrow), where syllables usages in $IHKA_{LEV\text{-}H}$ that diverge even more from controls ("off-target" effect) are on the left and those which get closer to those of CON ("on-target" effect) are on the right. Kruskal-Wallis and post hoc Dunn's two-sided test with permutation were used, with Benjamini-Hochberg false discovery rate with $\alpha = 0.05$. Error bars indicate 95% bootstrap confidence intervals.

(H) Word cloud for data in (G), with syllable names color coded (up- and downregulated with red and blue, respectively) and sized by the relative difference in syllable usage. High-dose levetiracetam treatment in IHKA mice leads to an upregulation of syllables such as "move forward" (off-target effect) and to a downregulation of syllables related to "dart" (e.g., "short dart left" or "dart right"; on-target effect). Asterisks indicate significant syllables (see G).

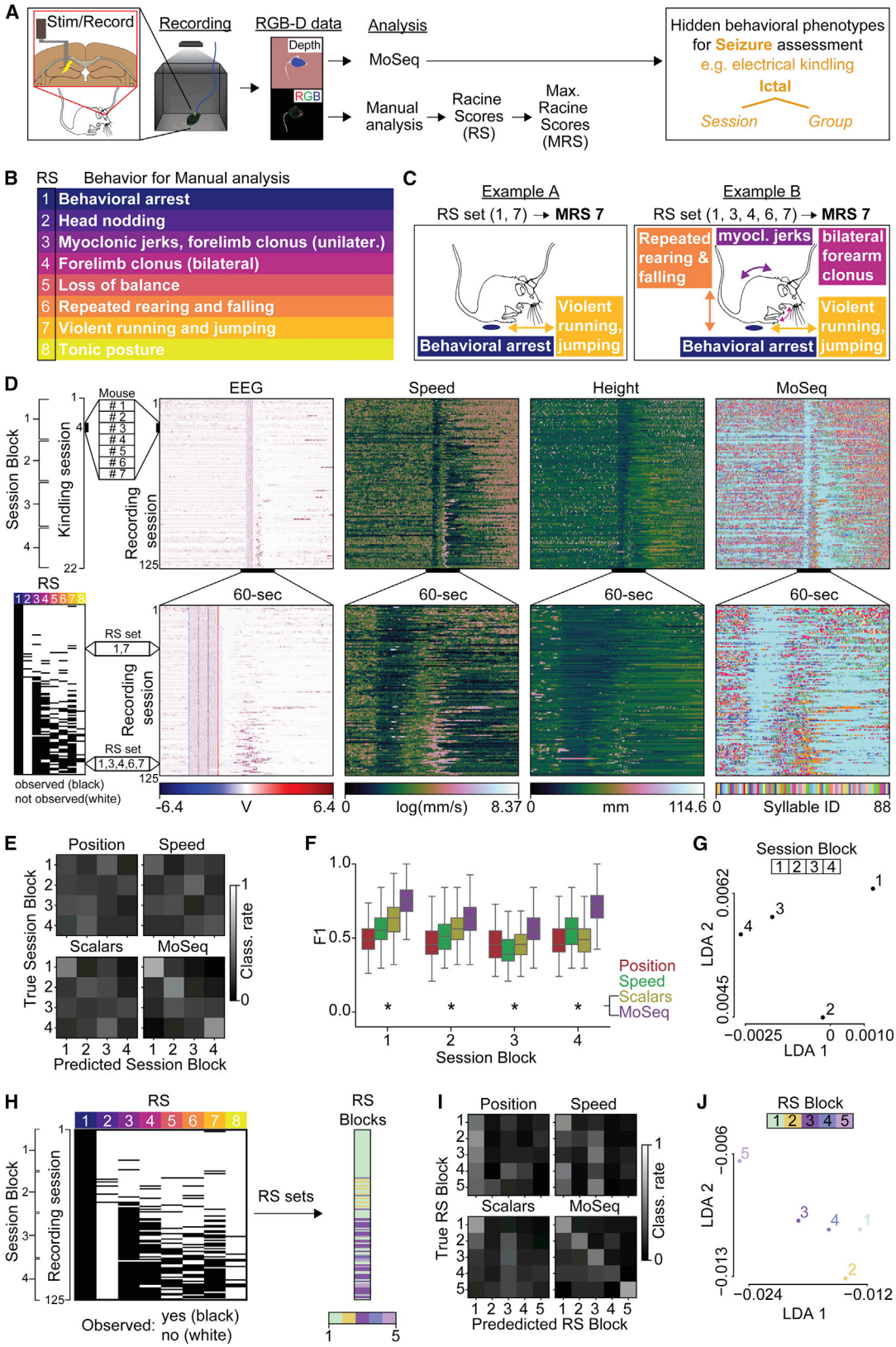
spontaneous recurrent seizures mimics refractory (AED-resistant) epilepsy and is thus widely used for AED screening and has previously been shown to respond to levetiracetam, but not other AEDs such as phenytoin or carbamazepine.[13] Therefore, we tested LEV, which is part of the latest generation of AEDs,[23] in IHKA mice. As expected, high-dose LEV in IHKA mice ($IHKA_{LEV-H}$) led to a normalization of some syllables toward control conditions (on-target effects). In other words, the difference in syllable usage between non-epileptic control, CON_{Saline}, and $IHKA_{LEV-H}$ was reduced compared with the difference between CON_{Saline} and saline-injected epileptic IHKA mice, $IHKA_{Saline}$ (Figure 3G). However, we also identified syllables in IHKA mice after LEV-H treatment that further diverged from control conditions (i.e., the difference between CON_{Saline} and $IHKA_{LEV-H}$ increased compared with the difference between CON_{Saline} and $IHKA_{Saline}$; Figure 3G). The latter can be considered off-target effects of high-dose LEV. Interestingly, some of these syllables that were upregulated in $IHKA_{LEV-H}$ and considered part of an off-target effect (e.g., syllable "move forward"; Figures 3G and 3H) were also upregulated in non-epileptic wild-type mice injected with high-dose LEV (e.g., syllable "move forward"; Figures 3E and S2F). Likewise, syllables similar to those identified as characteristic of IHKA mice in previous experiments (e.g., dart; see Figures 1B and 2G) were significantly reduced (e.g., short dart left; syllable ID 33) after LEV-H treatment in IHKA mice (Figures 3G and 3H), indicating an on-target effect (note, numerical syllable ID in Figures 1B and 3G are based on the frequency of the respective datasets).

Unsupervised segmentation of seizure behavior allows automated seizure assessment and links to traditional human-defined scoring

The results presented above showed that MoSeq was able to identify behavioral phenotypes in different epilepsy models and also allowed the study of AEDs in an automated manner free of human errors or bias, solely based on inter-ictal behaviors. The latter point is important since it is considerably easier to investigate the prolonged inter-ictal periods than the current practice of restricting the analysis of the behavioral manifestations of epilepsy (e.g., in AED testing) to the typically rather infrequent seizure (i.e., ictal) events. For these reasons, our focus so far in this study has been exclusively on the inter-ictal periods. In the last series of experiments, however, we set out to examine whether a similar MoSeq-based approach could also be used to discriminate between different ictal events and whether a MoSeq-based analysis of ictal periods could be related in a meaningful manner to the standard semi-quantitative, human observer-based seizure scoring strategies such as the Racine scale. However, it is non-trivial to relate MoSeq syllables to the Racine scale for various reasons. First, syllables identified by MoSeq typically last only for a few hundred milliseconds, whereas human observers using Racine scales most often focus on behaviors that last considerably longer, in the order of several seconds (e.g., rearing and falling). Second, traditional scoring scales heavily build on the scoring of distinct movements of individual body parts (e.g., forelimb clonus), whereas MoSeq takes the whole-body movement into account when assessing a behavioral state (e.g., forelimb clonus with rearing would be distinguished from forelimb clonus without rearing). Third, manual scoring of rodent seizure behavior by human observers utilizing Racine scales typically focuses on and reports only the most severe seizure behavior observed during the observation period, in contrast to the series of syllables automatically identified by MoSeq. In order to account for such differences between the MoSeq and Racine scale-based approaches, we aimed to establish if the syllable composition as determined by MoSeq reflects the composition of a manually observed series of behaviors during seizure events. Specifically, we sought to determine if MoSeq can classify and group together ictal events that were identified by human observers as being similar using Racine scale-based scoring.

In order to achieve the latter goal, we recorded the natural behavior of individual mice exploring an open field using an overhead camera in 5 min sessions. In the middle of the session at the 2.5 min time point, we delivered a focal electrical stimulation to the dorsal hippocampus to evoke a seizure (Figure 4A). Repeated delivery of such stimuli over time is known to lead to the appearance of more and more robust behavioral seizure events, a process referred to as "kindling."[24,25] For the traditional human observer-based "manual" analysis of the evoked behavioral seizure responses during kindling, we employed a version of the traditional Racine scale[26] that extends from simple indications of abnormal activity (e.g., behavioral arrest, which is assigned a score of 1) to modest manifestations of seizures (e.g., bilateral forelimb clonus, with a score of 4) to severe seizures with tonic postures (score of 8; see Figure 4B for the complete list). Figure 4C illustrates real-life examples from the experiments described below of observations of various ictal events during different kindling sessions, referred to as "Racine score (RS) sets," with the "maximal Racine score" (MRS) for each set also indicated. For example, when the human observer noted behavioral arrest followed by violent running and jumping, the resulting RS set based on the table in Figure 4B was (1 and 7), with the typically reported MRS = 7 (Example A in Figure 4C); for another animal that displayed a combination of behavioral arrest, myoclonic jerks, bilateral forearm clonus, repeated rearing, and falling as well as violent running and jumping, the RS set was (1, 3, 4, 6, and 7), with the MRS value being also 7 (Example B in Figure 4C). To avoid pitfalls associated with the typical practice of focusing only on the single behavior with the highest RS (i.e., the MRS), we utilized all human-observed Racine stage behaviors (the RS sets) noted during each post-stimulation period for every animal (Figure 4D, lower left). In addition to comparing MoSeq and human-based observations utilizing the Racine scale, we also carried out a quantitative comparison between the ability of the various automated behavioral measures (i.e., position, speed, scalars, and MoSeq) to distinguish ictal periods, using a 30 s time window after the electrical stimulation to approximate the windows used for the manual analysis. We trained a linear classifier for each automated behavioral measure to test its ability to classify stimulation sessions or RS sets (note that we used syllables exclusively in this 30 s post-stimulation time window for classification; see below).

Each kindling session included a single stimulation given to each of the n = 7 mice used in these experiments, and the mice were subjected to a total of 22 kindling sessions. Due to occasional experimental difficulties (e.g., detachment of a



(legend on next page)

cable), this resulted in a total of 125 recording sessions across the seven mice (instead of the expected total of 7 × 22 = 154). In order to reduce the number of session labels for classification, we grouped the 22 kindling sessions into 4 session blocks (Figure 4D, upper left panel).

For session classification, we evaluated the performance of different classifiers that were trained on the different behavioral measures to distinguish each session block. The results showed that MoSeq outperformed traditional 2D measures at identifying different session blocks (position F_1 = 0.22 ± 0.08; speed F_1 = 0.27 ± 0.09; scalars F_1 = 0.30 ± 0.09; and MoSeq F_1 = 0.49 ± 0.11; see also Figure 4E for classification matrices and Figure 4F for group-specific F_1 scores). Interestingly, embedding MoSeq behavioral summaries (i.e., syllable usages) into a 2D space using LDA clustered the later sessions (blocks 3 and 4) with more severe seizures together and separated the session blocks 1–4 in reverse numerical order along one axis of the LDA space (LDA-1; Figure 4G).

As we did for the session classification above, we reduced the number of unique RS sets for classification by grouping statistically similar RS sets together in one of five RS blocks (Figure 4H; see STAR Methods for details). Similar to the session classification results described above, the classifier trained on MoSeq behavioral summaries also performed best in distinguishing different RS blocks (position F_1 = 0.16 ± 0.07; speed F_1 = 0.25 ± 0.08; scalars F_1 = 0.26 ± 0.08; and MoSeq F_1 = 0.39 ± 0.13; see also Figure 4I for classification matrices and Figure S3 for group-specific F_1 scores). Moreover, LDA embedding of MoSeq summaries in 2D space grouped RS blocks based on severity along one axis (LDA-1; Figure 4J).

Together, these findings suggest that unsupervised segmentation of seizure behavior with MoSeq can classify seizure behaviors that were identified as similar based on human observations utilizing the Racine scale. These ictal period-focused automated behavioral assessments provide a link to traditional seizure scoring systems and indicate that MoSeq-based analysis can be applied to both inter-ictal periods as well as for the automated, unbiased assessment of seizures.

DISCUSSION

In this study, we explored the potential of machine learning-assisted 3D video analysis to phenotype mice with acquired and genetic epilepsies and screen for on- and off-target effects of AEDs in an automated and high-throughput manner. Our experiments revealed characteristic behavioral phenotypes during inter-ictal periods for different epilepsy types (acquired and genetic), distinct time points during epileptogenesis, as well as differences in behavioral fingerprints as a function of sex, drug type, and drug doses, allowing the automated, unbiased assessment of epilepsy without the need for recording electrographic activity in each animal for prolonged periods of time. In addition, our results showed that MoSeq-based approaches can be also used to classify ictal behaviors. These insights indicate that behavioral phenotypes can be targeted for assessment at scale in acquired and genetic models of epilepsies to accelerate rigorous, reproducible preclinical research into the epilepsies.

Toward rigorous behavioral fingerprinting in the epilepsies

Behavioral manifestations have long been recognized as an important feature of epilepsy disorders, as evidenced by the wealth of studies describing behavioral changes associated with seizures in human patients[27] as well as in a variety of animal models, including fruit flies,[28] zebrafish,[29–31] sea lions,[32] non-human primates,[33] and rodents (discussed below). However, our

Figure 4. Automated seizure assessment through unsupervised segmentation of behavior

(A) Experimental setup for assessing seizure behavior with AI-guided behavioral phenotyping. In a hippocampal kindling assay, intrahippocampal stimulation was combined with the synchronous acquisition of electroencephalographic (EEG) data and RGB-D data (i.e., red, green, and blue color data for manual analysis and depth data for MoSeq analysis). For manual analysis, an experimenter identified all behavior associated with different Racine scores (RSs), which for each seizure is commonly reported by selecting only the maximum Racine scores (MRSs; see results for details). Comparing MoSeq to manual analysis revealed that the syllable composition during seizures captures the aggravating nature of repeated kindling across sessions and can be used to identify different groups that share a similar composition of seizure behavior.

(B) Behavioral description of different seizure stages adapting a version of the traditional Racine scoring system.

(C) Example of two seizures with the same maximum Racine score but a different composition of observed behavior. For each seizure, all seizure-associated behavior was manually summarized in a set of observed RS behaviors (“RS set”). In example A, a human observer summarizes the behavior during a seizure with an RS set (1 and 7), which denotes behavioral arrest and violent running and jumping (see B), and would commonly only report the MRS value, which is 7. Similarly, in example B, another animal displaying a combination of behavioral arrest, myoclonic jerks, bilateral forearm clonus, repeated rearing, and falling as well as violent running and jumping (i.e., an RS set of [1, 3, 4, 6, and 7]) would also be reported as having an MRS of 7.

(D) EEG, speed, height, and MoSeq-identified syllables during kindling sessions. Top: each row represents the data of one 5 min recording session (total of 125 sessions in 7 mice; illustrative example shown for session 4 with mice 1–7 stacked on top of each other). Sessions are grouped into session blocks 1–4 for further analysis (see E–G below). Bottom: a zoomed-in 60 s window around the stimulation. On the left is a list of the RS sets for each of the 125 seizures, with observed Racine scores in black and not-observed ones in white. Two example RS sets are written out (same examples as in C). Note: a log scale was chosen for data “speed” due to the increase during seizures to improve the visualization for both ictal and inter-ictal periods.

(E) Normalized classification matrices (across rows and columns) representing the performance of a linear classifier for distinguishing different session blocks. Each classifier was trained on different behavioral measures (position, speed, combined scalar measures, or MoSeq syllables). An ideal classifier performance corresponds to a diagonal white with otherwise black fields (classification rate of 1).

(F) F_1 scores for linear classifiers discriminating between session blocks based on different behavioral measures (whiskers represent 1.5 times the inter-quartile range). Asterisks indicate significant differences between MoSeq and scalars ($p < 0.01$; paired two-sided t test corrected with Holm-Bonferroni step-down procedure).

(G) LDA plot indicating the similarity of mean MoSeq summaries (i.e., syllable usages) of mice within the same session block.

(H) Grouping RS sets into RS blocks for classification (see results).

(I) Same as (E), but to distinguish different RS blocks. See also Figure S3.

(J) Same as (G), but to distinguish different RS blocks.

understanding of behavioral manifestations of epilepsy remains most often confined to coarse assessments of relatively sparse ictal events (i.e., conspicuous motor seizures), which are readily discerned by the naked eye of human observers. The current gold standard method in the study of rodent models of epilepsy is the use of behavioral seizure scales,[5,34–39] where the manual assessments heavily depend on the expertise and intuition of the human observer. In addition, a fine-grained, sub-second behavioral analysis even by an experienced observer quickly becomes infeasible when large datasets need to be scored, and the strain on human resources further opens the door to problems with reproducibility and potential inter-observer biases.

In our study, we showed that unsupervised behavioral analysis allows an automated, unbiased assessment of epilepsy in animal models of epilepsies. We demonstrated that an experimenter with the help of MoSeq can reliably decompose behaviors from large datasets in different epilepsy models under various experimental conditions (over 300 h of data were acquired and analyzed in this study; Figure S4). Importantly, we also found that 60 min long recording of inter-ictal behavior per animal during random free exploration is sufficient to sort epileptic mice from non-epileptic littermates without requiring any additional information (e.g., 24/7 video-EEG or histopathology). Therefore, such unsupervised assessment of behavior with sub-second precision constitutes a type of behavioral fingerprinting for epilepsy that is effective even during the easily accessible inter-ictal periods. Because the approach takes into account subtle differences in pose dynamics that are not evident by eye, it can also yield unexpected insights. For example, we discovered a previously unrecognized sex-specific behavioral phenotype in mice with heterozygous deletion of *Scn1b* and identified subtle changes in behavioral repertoire at different time points during epileptogenesis. The latter could be particularly interesting for post-traumatic epilepsy research where a major current challenge is to find early biomarkers for individuals who are on track to eventually develop epilepsy, and an automated behavioral assessment that does not require ictal episodes may pave the way for an entirely novel way of forecasting disease progression after brain insults, yielding a "thermometer" for epilepsy in the form of artificial intelligence (AI)-assisted diagnostics. Furthermore, it is the inter-ictal periods that comprise most of the lives of patients with epilepsy, with cognitive and behavioral comorbidities often profoundly impairing the quality of life even during such nominally seizure-free periods. Therefore, capturing the behavioral manifestations of epilepsy during inter-ictal automated observation periods may be particularly important for the non-invasive assessment of meaningful biological variables indicating disease states.

Potential for rapid, scalable anti-epileptic drug testing

Our results also indicated that MoSeq can be used to automatically discriminate between mice injected with different AEDs by characterizing the underlying structure of their behavior during free exploration. We administered wild-type mice with one of three different AEDs (VAL, PHT, and LEV) at either high or low doses and showed that MoSeq outperformed traditional analysis methods (such as 2D measures, including the animal's position and speed) in predicting drug-dose pairs. As a strictly data-driven approach without human observers, MoSeq captured an unbiased spectrum of behavioral patterns that then could be used to identify on- and off-target effects of AEDs in epilepsy. Our results in a model of TLE indicate that our pipeline can be used to discriminate IHKA mice treated with a vehicle from those injected with LEV, at a dose previously shown to reduce the seizure frequency in this model.[13] Therefore, our findings highlight the potential of MoSeq for AED screening by measuring unrestrained naturalistic behavior in an automated fashion without the need of labor-intensive and expensive video-EEG monitoring.

Implications and outlook

In addition to gaining insights into the potential for automated behavioral assessment during the inter-ictal periods, we also acquired 3D videos of a variety of seizures in kindled mice to relate the automated MoSeq outcome measures to the currently widely used Racine scale-based seizure assessment approaches. Our results revealed that seizures of similar severity were composed of similar MoSeq-identified syllables, illustrating the potential for co-alignment of classical seizure scoring practices and MoSeq. These insights indicate that it should be possible to transition from the labor-intensive, human observation-based seizure analyses to a more automated, scalable approach while continuing to benefit from the accumulated knowledge about seizures acquired in the past decades relying on traditional Racine scales for assessment. It should be noted that although our data demonstrate that MoSeq-based analysis can be applied to both inter-ictal and ictal periods, clear links between MoSeq and traditional epilepsy-related expert annotations of behaviors could be established only for the ictal period. This is because, compared with the approach used in Figure 4 for seizures, there is no analogous way to meaningfully relate the differential expression of specific MoSeq-defined behavioral syllables (e.g., in epilepsy models, or as a result of AED effects) to changes in epileptic behavioral attributes that may be in principle observable by humans during the inter-ictal period. Indeed, consensus annotation for inter-ictal behavior does not exist, and even expert observers largely failed to correctly perform the related classification tasks when no overt motor seizures were present.

Importantly, recent results revealed that syllables identified by MoSeq in control mice are closely correlated with particular neuronal activity dynamics in the striatum (and perhaps also elsewhere) during naturalistic behaviors,[40] indicating that the behavioral syllables are anchored in actual patterns of circuit activity in the brain. Identification of the precise cell types and brain regions involved in the altered expression of behavioral syllables that we observed in our epilepsy models will be able to rely on recent advances in multi-site single unit recordings from thousands of neurons across the brain, likely yielding new insights into the nature of circuit plasticity associated with ictal and inter-ictal behaviors. It is interesting to note in this regard that the brain circuits responsible for generating and spreading epileptiform activity in epilepsy may also underlie comorbidities (e.g., impairments in memory and sleep disturbances) that persist during non-seizure periods.[6–8]

Future studies will be needed to determine whether epilepsy-associated syllables identified with MoSeq are shared in a variety of different animal models of epilepsy and whether such syllables

can be targeted for closed-loop interventions[6,41,42] where the electrical stimulation or optogenetic intervention would be triggered by the appearance of particular syllables in an on-demand manner. Such syllable-based on-demand, closed-loop, optogenetic interventions in the striatum of control mice have been recently demonstrated to result in changed syllable expression in spontaneously behaving animals in a persistent manner following the intervention.[43] Application of a similar approach to syllables whose expression have been changed in epilepsy will be an exciting undertaking in future projects. It is interesting to note that such an approach would not require invasive depth electrodes for detecting seizure onsets (since syllable detection is based on 3D video recordings and online analysis), and the intervention could target not only the epileptic focus (e.g., the hippocampal CA1 in the IHKA model of TLE) but also the more superficially located and thus more easily accessible, extra-focal regions such as the cerebellum that can have powerful effects on TLE as well as on-going behavior.[41,42] A closely related question is whether targeting particular syllables during the inter-ictal periods for closed-loop interventions would have beneficial effects on comorbidities and perhaps even on the rate of seizures as well.

In summary, although various novel tools bear great potential to create better animal models for a variety of epilepsies (e.g., CRISPR-Cas9[44]) and can also accelerate drug discovery on a molecular level (e.g., DeepMind's AlphaFold[45]), current behavioral assessment practices in epilepsy research constitute a major bottleneck for advancing mechanistic insights into epilepsies and screening for new AEDs in a reproducible fashion at scale. Our findings suggest that the unbiased detection of hidden inter-ictal and ictal behavioral phenotypes may begin to overcome this bottleneck and advance the field toward unbiased assessment approaches for epilepsies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Intrahippocampal injections
  - Electrical kindling seizure model
  - Behavioral data acquisition
  - Drug treatments
  - Behavioral recording
  - Electroencephalographic (EEG) recordings
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Motion Sequencing (MoSeq) - Data extraction
  - MoSeq - Data modeling
  - Behavioral summaries and wordclouds
  - MoSeq-based behavioral distance measurements
  - Linear classification of behavioral summaries
  - Cosine distance matrix for behavioral summary distance comparisons
  - Discrimination-relevant syllables ("behavioral fingerprint") and transitions
  - Visualizing behavioral summaries with low-dimensional embeddings
  - Human vs. Machine performance in identifying epileptic animals
  - Grouping of Racine score sets
  - Statistical tests

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2023.02.003.

ACKNOWLEDGMENTS

We thank current and former members of the Soltesz lab, in particular Xiaoyang Wang and Yatong Han for their technical input, and Rika Kumar for technical assistance. T.G. would also like to thank Dr. Frances Cho for useful discussions and constructive criticism. Research reported in this study was supported by the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health (NIH) under award nos. NINDSR01NS114020 (to I.S. and S.D.) and NINDSR37NS076752 (to L.L.I.), as well as by the Swiss National Science Foundation under award nos. 174811 and 186757 (to T.G.).

AUTHOR CONTRIBUTIONS

Conceptualization, T.G. and I.S.; resources, L.L.I.; investigation, T.G. and S.F.; software, T.G., A.Z., I.R., J.E.M., and W.F.G.; formal analysis, T.G.; funding acquisition, I.S., S.R.D., and L.L.I.; writing – original draft, T.G. and I.S.; writing – review & editing, T.G., A.Z., I.R., J.E.M., W.F.G., S.F., L.L.I., S.R.D., and I.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 28, 2022
Revised: November 11, 2022
Accepted: February 1, 2023
Published: February 24, 2023

REFERENCES

1. England, M.J., Liverman, C.T., Schultz, A.M., and Strawbridge, L.M. (2012). Epilepsy across the spectrum: promoting health and understanding. A summary of the Institute of Medicine report. Epilepsy Behav. *25*, 266–276. https://doi.org/10.1016/j.yebeh.2012.06.016.
2. Quigg, M., Straume, M., Menaker, M., and Bertram, E.H. (1998). Temporal distribution of partial seizures: comparison of an animal model with human partial epilepsy. Ann. Neurol. *43*, 748–755. https://doi.org/10.1002/ana.410430609.
3. Kim, H.K., Gschwind, T., Nguyen, T.M., Bui, A.D., Felong, S., Ampig, K., Suh, D., Ciernia, A.V., Wood, M.A., and Soltesz, I. (2020). Optogenetic intervention of seizures improves spatial memory in a mouse model of chronic temporal lobe epilepsy. Epilepsia *61*, 561–571. https://doi.org/10.1111/epi.16445.
4. Williams, P.A., White, A.M., Clark, S., Ferraro, D.J., Swiercz, W., Staley, K.J., and Dudek, F.E. (2009). Development of spontaneous recurrent seizures after kainate-induced status epilepticus. J. Neurosci. *29*, 2103–2112. https://doi.org/10.1523/JNEUROSCI.0980-08.2009.

5. Racine, R.J. (1972). Modification of seizure activity by electrical stimulation. II. Motor seizure. Electroencephalogr. Clin. Neurophysiol. *32*, 281–294. https://doi.org/10.1016/0013-4694(72)90177-0.

6. Bui, A.D., Nguyen, T.M., Limouse, C., Kim, H.K., Szabo, G.G., Felong, S., Maroso, M., and Soltesz, I. (2018). Dentate gyrus mossy cells control spontaneous convulsive seizures and spatial memory. Science *359*, 787–790. https://doi.org/10.1126/science.aan4074.

7. Gelinas, J.N., Khodagholy, D., Thesen, T., Devinsky, O., and Buzsáki, G. (2016). Interictal epileptiform discharges induce hippocampal-cortical coupling in temporal lobe epilepsy. Nat. Med. *22*, 641–648. https://doi.org/10.1038/nm.4084.

8. Holden, S.S., Grandi, F.C., Aboubakr, O., Higashikubo, B., Cho, F.S., Chang, A.H., Forero, A.O., Morningstar, A.R., Mathur, V., Kuhn, L.J., et al. (2021). Complement factor C1q mediates sleep spindle loss and epileptic spikes after mild brain injury. Science *373*, eabj2685. https://doi.org/10.1126/science.abj2685.

9. Krakauer, J.W., Ghazanfar, A.A., Gomez-Marin, A., MacIver, M.A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. Neuron *93*, 480–490. https://doi.org/10.1016/j.neuron.2016.12.041.

10. Datta, S.R., Anderson, D.J., Branson, K., Perona, P., and Leifer, A. (2019). Computational neuroethology: A call to action. Neuron *104*, 11–24. https://doi.org/10.1016/j.neuron.2019.09.038.

11. Dennis, E.J., El Hady, A., Michaiel, A., Clemens, A., Tervo, D.R.G., Voigts, J., and Datta, S.R. (2021). Systems neuroscience of natural behaviors in rodents. J. Neurosci. *41*, 911–919. https://doi.org/10.1523/JNEUROSCI.1877-20.2020.

12. Wiltschko, A.B., Johnson, M.J., Iurilli, G., Peterson, R.E., Katon, J.M., Pashkovski, S.L., Abraira, V.E., Adams, R.P., and Datta, S.R. (2015). Mapping sub-second structure in mouse behavior. Neuron *88*, 1121–1135. https://doi.org/10.1016/j.neuron.2015.11.031.

13. Klein, S., Bankstahl, M., and Löscher, W. (2015). Inter-individual variation in the effect of antiepileptic drugs in the intrahippocampal kainate model of mesial temporal lobe epilepsy in mice. Neuropharmacology *90*, 53–62. https://doi.org/10.1016/j.neuropharm.2014.11.008.

14. Chen, C., Westenbroek, R.E., Xu, X., Edwards, C.A., Sorenson, D.R., Chen, Y., McEwen, D.P., O'Malley, H.A., Bharucha, V., Meadows, L.S., et al. (2004). Mice Lacking Sodium Channel β1 Subunits Display Defects in Neuronal Excitability, Sodium Channel Expression, and Nodal Architecture. J. Neurosci. *24*, 4030–4042. https://doi.org/10.1523/JNEUROSCI.4139-03.2004.

15. Isom, L.L., De Jongh, K.S., Patton, D.E., Reber, B.F.X., Offord, J., Charbonneau, H., Walsh, K., Goldin, A.L., and Catterall, W.A. (1992). Primary structure and functional expression of the β1 subunit of the rat brain sodium channel. Science *256*, 839–842. https://doi.org/10.1126/science.1375395.

16. Audenaert, D., Claes, L., Ceulemans, B., Löfgren, A., Van Broeckhoven, C., and De Jonghe, P. (2003). A deletion in SCN1B is associated with febrile seizures and early-onset absence epilepsy. Neurology *61*, 854–856. https://doi.org/10.1212/01.WNL.0000080362.55784.1C.

17. Scheffer, I.E., Harkin, L.A., Grinton, B.E., Dibbens, L.M., Turner, S.J., Zielinski, M.A., Xu, R., Jackson, G., Adams, J., Connellan, M., et al. (2007). Temporal lobe epilepsy and GEFS+ phenotypes associated with SCN1B mutations. Brain *130*, 100–109. https://doi.org/10.1093/brain/awl272.

18. Wallace, R.H., Wang, D.W., Singh, R., Scheffer, I.E., George, A.L., Phillips, H.A., Saar, K., Reis, A., Johnson, E.W., Sutherland, G.R., et al. (1998). Febrile seizures and generalized epilepsy associated with a mutation in the Na+-channel β1 subunit gene SCN1B. Nat. Genet. *19*, 366–370. https://doi.org/10.1038/1252.

19. Patino, G.A., Claes, L.R.F., Lopez-Santiago, L.F., Slat, E.A., Dondeti, R.S.R., Chen, C., O'Malley, H.A., Gray, C.B.B., Miyazaki, H., Nukina, N., et al. (2009). A functional null mutation of SCN1B in a patient with Dravet syndrome. J. Neurosci. *29*, 10764–10778. https://doi.org/10.1523/JNEUROSCI.2475-09.2009.

20. Gschwind, T., Lafourcade, C., Gfeller, T., Zaichuk, M., Rambousek, L., Knuesel, I., and Fritschy, J.M. (2018). Contribution of early Alzheimer's disease-related pathophysiology to the development of acquired epilepsy. Eur. J. Neurosci. *47*, 1534–1562. https://doi.org/10.1111/ejn.13983.

21. Riban, V., Bouilleret, V., Pham-Lê, B.T., Fritschy, J.M., Marescaux, C., and Depaulis, A. (2002). Evolution of hippocampal epileptic activity during the development of hippocampal sclerosis in a mouse model of temporal lobe epilepsy. Neuroscience *112*, 101–111. https://doi.org/10.1016/s0306-4522(02)00064-7.

22. Wiltschko, A.B., Tsukahara, T., Zeine, A., Anyoha, R., Gillis, W.F., Markowitz, J.E., Peterson, R.E., Katon, J., Johnson, M.J., and Datta, S.R. (2020). Revealing the structure of pharmacobehavioral space through motion sequencing. Nat. Neurosci. *23*, 1433–1443. https://doi.org/10.1038/s41593-020-00706-3.

23. Löscher, W., Potschka, H., Sisodiya, S.M., and Vezzani, A. (2020). Drug resistance in epilepsy: clinical impact, potential mechanisms, and new innovative treatment options. Pharmacol. Rev. *72*, 606–638. https://doi.org/10.1124/pr.120.019539.

24. Goddard, G.V. (1967). Development of epileptic seizures through brain stimulation at low intensity. Nature *214*, 1020–1021. https://doi.org/10.1038/2141020a0.

25. Goddard, G.V., McIntyre, D.C., and Leech, C.K. (1969). A permanent change in brain function resulting from daily electrical stimulation. Exp. Neurol. *25*, 295–330. https://doi.org/10.1016/0014-4886(69)90128-9.

26. Velíšková, J., and Velíšek, L. (2017). Behavioral characterization and scoring of seizures in rodents. In Models of Seizures and Epilepsy, Second Edition (Elsevier Inc.), pp. 111–123. https://doi.org/10.1016/B978-0-12-804066-9.00009-2.

27. Fisher, R.S., Cross, J.H., French, J.A., Higurashi, N., Hirsch, E., Jansen, F.E., Lagae, L., Moshé, S.L., Peltola, J., Roulet Perez, E., et al. (2017). Operational classification of seizure types by the International League Against Epilepsy: position Paper of the ILAE Commission for Classification and Terminology. Epilepsia *58*, 522–530. https://doi.org/10.1111/epi.13670.

28. Parker, L., Howlett, I.C., Rusan, Z.M., and Tanouye, M.A. (2011). Seizure and Epilepsy: Studies of Seizure Disorders in Drosophila (Elsevier Inc.). https://doi.org/10.1016/B978-0-12-387003-2.00001-X.

29. Baraban, S.C., Taylor, M.R., Castro, P.A., and Baier, H. (2005). Pentylenetetrazole induced changes in zebrafish behavior, neural activity and c-fos expression. Neuroscience *131*, 759–768. https://doi.org/10.1016/j.neuroscience.2004.11.031.

30. Griffin, A., Hamling, K.R., Knupp, K., Hong, S.G., Lee, L.P., and Baraban, S.C. (2017). Clemizole and modulators of serotonin signalling suppress seizures in Dravet syndrome. Brain *140*, 669–683. https://doi.org/10.1093/brain/aww342.

31. Griffin, A., Carpenter, C., Liu, J., Paterno, R., Grone, B., Hamling, K., Moog, M., Dinday, M.T., Figueroa, F., Anvar, M., et al. (2021). Phenotypic analysis of catastrophic childhood epilepsy genes. Commun. Biol. *4*, 680. https://doi.org/10.1038/s42003-021-02221-y.

32. Gulland, F.M., Haulena, M., Fauquier, D., Langlois, G., Lander, M.E., Zabka, T., and Duerr, R. (2002). Domoic acid toxicity in Californian sea lions (Zalophus californianus): clinical signs, treatment and survival. Vet. Rec. *150*, 475–480. https://doi.org/10.1136/vr.150.15.475.

33. Gunderson, V.M., Dubach, M., Szot, P., Born, D.E., Wenzel, H.J., Maravilla, K.R., Zierath, D.K., Robbins, C.A., and Schwartzkroin, P.A. (1999). Development of a model of status epilepticus in pigtailed macaque infant monkeys. Dev. Neurosci. *21*, 352–364. https://doi.org/10.1159/000017385.

34. Jobe, P.C., Picchioni, A.L., and Chin, L. (1973). Role of brain norepinephrine in audiogenic seizure in the rat. J. Pharmacol. Exp. Ther. *184*, 1–10. https://jpet.aspetjournals.org/content/184/1/1.long.

35. Pinel, J.P., and Rovner, L.I. (1978). Electrode placement and kindling-induced experimental epilepsy. Exp. Neurol. *58*, 335–346. https://doi.org/10.1016/0014-4886(78)90145-0.

36. Pinel, J.P., and Rovner, L.I. (1978). Experimental epileptogenesis: kindling-induced epilepsy in rats. Exp. Neurol. *58*, 190–202. https://doi.org/10.1016/0014-4886(78)90133-4.

37. Pohl, M., and Mares, P. (1987). Effects of flunarizine on Metrazol-induced seizures in developing rats. Epilepsy Res. *1*, 302–305. https://doi.org/10.1016/0920-1211(87)90006-4.

38. Haas, K.Z., Sperber, E.F., and Moshé, S.L. (1990). Kindling in developing animals: expression of severe seizures and enhanced development of bilateral foci. Brain Res. Dev. Brain Res. *56*, 275–280. https://doi.org/10.1016/0165-3806(90)90093-e.

39. Velísková, J., Velísek, L., Mares, P., and Rokyta, R. (1990). Ketamine suppresses both bicuculline- and picrotoxin-induced generalized tonic-clonic seizures during ontogenesis. Pharmacol. Biochem. Behav. *37*, 667–674. https://doi.org/10.1016/0091-3057(90)90544-r.

40. Markowitz, J.E., Gillis, W.F., Beron, C.C., Neufeld, S.Q., Robertson, K., Bhagat, N.D., Peterson, R.E., Peterson, E., Hyun, M., Linderman, S.W., et al. (2018). The striatum organizes 3D behavior via moment-to-moment action selection. Cell *174*, 44–58.e17. https://doi.org/10.1016/j.cell.2018.04.019.

41. Stieve, B.J., Richner, T.J., Krook-Magnuson, C., Netoff, T.I., and Krook-Magnuson, E. (2023). Optimization of closed-loop electrical stimulation enables robust cerebellar-directed seizure control. Brain *146*, 91–108. https://doi.org/10.1093/brain/awac051.

42. Krook-Magnuson, E., Szabo, G.G., Armstrong, C., Oijala, M., and Soltesz, I. (2014). Cerebellar directed optogenetic intervention inhibits spontaneous hippocampal seizures in a mouse model of temporal lobe epilepsy. eNeuro *1*, 0005–14.2014. eNeuro. https://doi.org/10.1523/ENEURO.0005-14.2014.

43. Markowitz, J.E., Gillis, W.F., Jay, M., Wood, J., Harris, R.W., Cieszkowski, R., Scott, R., Brann, D., Koveal, D., Kula, T., et al. (2023). Spontaneous behaviour is structured by reinforcement without explicit reward. Nature *614*, 108–117. https://doi.org/10.1038/s41586-022-05611-2.

44. Das, A., Zhu, B., Xie, Y., Zeng, L., Pham, A.T., Neumann, J.C., Safrina, O., Benavides, D.R., Macgregor, G.R., Schutte, S.S., et al. (2021). Interneuron dysfunction in a new mouse model of SCN1A GEFS. eNeuro *8*, 1–16. https://doi.org/10.1523/ENEURO.0394-20.2021.

45. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

46. Musto, A., and Bazan, N.G. (2006). Diacylglycerol kinase epsilon modulates rapid kindling epileptogenesis. Epilepsia *47*, 267–276. https://doi.org/10.1111/j.1528-1167.2006.00418.x.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Levetiracetam	Gland Pharma Limited	NDC:0409-1886
Valproic Acid	Hikma Pharmaceutical USA Inc	NDC:0143-9785-01
Phenytoin	Hikma Pharmaceutical USA Inc	NDC:0641-2555-41
Kainic Acid	Tocris Bioscience	Cat# 0222
Deposited data		
Fingerprint data	This study	Zenodo: https://doi.org/10.5281/zenodo.7567521
Experimental models: Organisms/strains		
Mouse: C57Bl/6J	The Jackson Laboratory	#000664
Mouse: *Scn1b*$^{+/-}$ and *Scn1b*$^{+/+}$	Chen et al.[14]	N/A
Software and algorithms		
MoSeq	Wiltschko et al.[12]	https://dattalab.github.io/moseq2-website/index.html
MoSeq for drug effects	Wiltschko et al.[22]	https://github.com/dattalab/moseq-drugs
Fingerprint analysis	This study	Zenodo: https://doi.org/10.5281/zenodo.7567521
Python	Python Software Company	RRID:SCR_008394
Intan Recording System Software	Intan Technologies	https://intantech.com/downloads.html
Other		
Intan RHS Stim/Recording System	Intan Technologies	https://intantech.com/RHS_system.html
Kinect2	Microsoft	https://www.amazon.com/kinect-v2/s?k=kinect+v2
RealSense D415	Intel	https://www.intelrealsense.com/depth-camera-d415/
Open-field assay enclosure	US Plastics	https://www.usplastic.com/catalog/item.aspx?itemid=120721
Spray paint for OFA enclosure	Acryli-Quik	#132496

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Tilo Gschwind (gschwind@stanford.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data have been deposited onto Zenodo. The DOI is listed in the key resources table. Additional data (e.g., preprocessed datasets) will be available upon reasonable request from the lead contact.
- All code has been deposited onto Zenodo. The DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All procedures were approved and performed in accordance with the Administrative Panel of Laboratory Animal Care of Stanford University (Protocol 30183), and with the animal care guide- lines of the National Institutes of Health. For modeling electrical kindling, TLE and for the drug study, C57BL/6J male mice were used (The Jackson Laboratory, #000664). For modeling Dravet syndrome, both male and female *Scn1b*$^{+/-}$ and *Scn1b*$^{+/+}$ littermate mice, congenic on the C57BL/6J background for over 15 generations were used (for details, see[14]). Mice were 12-24 weeks old. Animals were given food and water ad libitum. All animals were group housed and in a 12-hour light/12-hour dark cycle.

METHOD DETAILS

Intrahippocampal injections

Stereotaxic injections were carried out as previously described.[3] Briefly, mice were anesthetized with 2-5% isoflurane and given local anesthetic (s.c. 0.5% bupivacaine). Kainic acid (70 nL, 20 mM in saline; Tocris Bioscience) or saline (sham controls) were injected into the left dorsal hippocampus (from bregma: 2.0 mm posterior, 1.25 mm left, 1.6 mm ventral). Mice were allowed to recover and then were returned to the vivarium for at least one week.

Electrical kindling seizure model

Teflon-coated twisted bipolar electrodes (tip separation of 0.5mm) were implanted chronically into the dorsal hippocampus of mice (from bregma: 2.0 mm posterior, 1.25 mm left, 1.6 mm ventral), similar to as described previously.[46] Following 1-2 weeks of recovery from surgery, seizures were elicited by electrical kindling stimulation (150–300μA of current delivered in 1ms biphasic pulses at 50Hz for 10 second). Kindling was achieved by stimulating 4–7 sessions daily for 4 days. All kindling sessions were performed in an enclosed glass arena and recorded with a side RGB camera for off-line manual seizure scoring, in addition to an overhead Intel RealSense D415 RGB-D camera which was used for both manual seizure scoring and for MoSeq analysis (described below). Electrical stimulation was delivered in the middle of the 5-min recording session.

Behavioral data acquisition

Behavioral data acquisition was performed similar to previous descriptions.[12,22,40] For Figures 1, 2, and 3, mice were placed in the center of a circular open-field assay (OFA) enclosure (18 inch diameter, 15 inch high; US Plastics) and recordings started immediately after. The opaque enclosure, which was painted black with spray paint (Acryli-Quik Ultra Flat Black; 132496) to avoid image artifacts, was illuminated with red light during the recording. Animals were allowed to freely explore the OFA enclosure for 60 min. For all experiments (Figures 1, 2, 3, and 4), the enclosure was cleaned between mice consecutively with 10% bleach, 1% alconox, and 70% ethanol. For the epileptogenesis experiments, mice were recorded weekly over four weeks after intrahippocampal injection. All data acquisition were obtained at the same time of day during the night phase, i.e., during the active period, of mice. For TLE, *Scn1b*, and drug study experiments, all mice were recorded for 60 minutes during inter-ictal periods. An experimenter observed the animal for 15 minutes prior during the recording session to ascertain the lack of overt motor seizures. No overt seizure behaviors were observed prior to the placement of the animal into acquisition setup nor during the recording. For electrical kindling experiments, electrical stimulation was delivered in the middle of the 5-min recording session to evoke seizures.

Drug treatments

Six drug-dose combinations of three well-characterized anti-epileptic drugs (AEDs) were used in this study: Levetiracetam (LEV), Phenytoin (PHT), and Valproic Acid (VAL), each at low and high concentrations. Individual drug-dose combinations were chosen based on previous studies[13,21] which used the same mouse model of TLE. Levetiracetam (Gland Pharma Limited; NDC 0409-1886-02) was dissolved in saline and used at 400 mg/kg and 800 mg/kg for LEV-Low (n=12 mice) and LEV-High (n=12 mice) respectively. Valproic Acid (Hikma Pharmaceutical USA Inc.; NDC 0143-9785-01) was dissolved in saline and used at 150 mg/kg and 300 mg/kg for VAL-Low (n=12 mice) and VAL-High (n=12 mice) respectively. Phenytoin (Hikma Pharmaceutical USA Inc.; NDC 0641-2555-41) was dissolved in saline and used at 20 mg/kg and 30 mg/kg for PHT-Low (n=12 mice) and PHT-High (n=12) respectively. Saline was used as vehicle for control experiments (n=24 mice). All drugs were prepared fresh before the recording, and each mouse was injected once with one drug-dose combination. For all drug and vehicle injections, an injection volume of 6 mL/kg body weight was delivered intraperitoneally. After the drug was administered (experimenter was not blinded to drug type and dose), the animal was immediately transferred to the recording box to start data acquisition. Data acquisition and analysis was performed in a blinded manner by the MoSeq pipeline, which uses a universally unique identifier (UUID), a 128-bit label, instead of other metadata (e.g., animal or group ID) to identify recording sessions. The use of the UUID obviates the need for any additional de-identification by an experimenter. All mice were recorded for 60 minutes, and all reported results are based on the assessment of AED effects throughout the entire recording duration. Time-dependent drug effects within the recording period were not assessed in this study.

Behavioral recording

Data acquisition was performed as previously described.[12,22,40] Briefly, each mouse was tracked in 3D with a Kinect 2 for Windows (Microsoft) or an Intel RealSense D415 stably suspended above the recording arena (i.e., the open-field assay enclosure), which provides a top-down view. The camera was placed at a working distance of approximate 0.65 m for optimal sensor position. Raw data from the camera were sent to an acquisition computer (i5-6400 Intel Quad Core with 16GB DDR4 RAM and NVIDIA GeForce GT1030 2GB graphics card) via USB 3.0 cables, and depth frames were retrieved (30 frames per second) and saved to disk in raw binary format using a custom user interface programmed in C#/C++.

Electroencephalographic (EEG) recordings

The bipolar electrode implanted for kindling experiments was also used to record EEG from the dorsal hippocampus. EEG was acquired with the Intan RHS Stim/Recording System (sampled at 30kHz). A subset of sessions were excluded because data acquisition was interrupted in the middle of the session.

QUANTIFICATION AND STATISTICAL ANALYSIS

Motion Sequencing (MoSeq) - Data extraction

Data preprocessing, extraction, and modeling pipeline (written in the Python programming language) were performed as previously described[12,22,40] on either local machines or the Stanford cluster computer (Sherlock 2). Briefly, custom mouse-tracking software was used to extract the mouse's position, orientation, and body morphometry from the raw depth data. The 3D image of the mouse was extracted and aligned using a previously published pipeline.[12,22,40] First, raw depth images were denoised through resampling and background subtraction (using a background image composed of the median value of the first 1000 frames). For kindling experiments, chroma keying was used for both cable removal and extraction of the mouse and combined with spatio-temporal filtering to fill in holes. For all experiments, the resulting images were composed of values indicating how high each pixel is relative to the background image, with all negative values and all values above a maximum height being set to zero. The orientation and center-of-mass of the mouse contour were calculated and used to extract a square (80x80 pixel) centered on the mouse in every frame. The resulting timeseries of aligned images was sampled at 30 frames per second.

MoSeq - Data modeling

Principal components analysis (PCA) was used to reduce the dimensionality of the extracted time-series, and covariance between the resulting principal components (PC) were removed by whitening the PC time-series across data obtained from all subjects. The whitened PCs were fit using an autoregressive hierarchical Dirichlet process hidden Markov model (AR-HMM), as described previously.[12] Separate models were used for different datasets (electrical kindling, IHKA, SCN1b female, SCN1b male, IHKA epileptogenesis, AED wildtype mice, and AED IHKA mice). The AR-HMM was used to identified syllables which captured 90-95% of the variance; the number of syllables ranged between 27-60 syllables (includes 5-min and 60-min recording sessions).

Behavioral summaries and wordclouds

Preprocessed behavioral recordings of mice in the open-field assay enclosure were summarized to compare their performance to distinguish different experimental groups. We used the following parameters, as described previously[22]: *position, speed, length, height, scalars*, and *MoSeq*. Position summaries were generated from histograms partitioning the normalized distance from the arena center into 90 bins. Speed summaries were obtained from the first derivative of the 2D position, which was used to generate a histogram with 90 bins spaced between 0 and 20 pixels. Length summaries were obtained from the major axis of the ellipse of the mouse body contour, which was used to generate a histogram with 45 bins spaced between 20 and 100 pixels. Height summaries were obtained from the maximum height of the extracted mouse image, which was used to generate a histogram with 45 bins spaced between 0 and 60 mm. Scalar summaries were generated by concatenating the length, height, speed and position summaries. MoSeq summaries were generated with histograms describing usage frequency of each syllable (numerical syllable ID) to which the (blinded) experimenter gave a semantic description (syllable names). Wordclouds with syllable names were generated in Python to visualize the relative usage frequency of syllables between experimental groups (available through github.com/amueller/word_cloud). Font size corresponds to difference in frequency between indicated groups, and color corresponds to whether syllable expression was increased or decreased (red or blue, respectively).

MoSeq-based behavioral distance measurements

We assessed similarity between pose trajectories of different syllables, as previously described.[22,40] Briefly, we used the autoregressive coefficients described by the AR-HMM model to simulate pose trajectories for each syllable over ten time-steps (corresponding to 300 ms). We then generated a distance matrix by computing the pairwise correlation distance between the top most-used syllables and represented this distance matrix as a dendrogram using the Voor Hees hierarchical clustering algorithm (scipy.cluster.hierarchy.linkage). Low distances (values near 0) represent similar syllables, and high distances (values near 6) represent dissimilar syllables.

Linear classification of behavioral summaries

We performed linear classification of behavioral summaries using logistic regression (Scikit-Learn Python package), as previously described.[22] We used a "one-vs-rest" formulation of multi-class classification, with an L2 weight penalty with an inverse regularization strength. We performed 500-fold cross-validation, using randomly shuffled folds with 10% of data held-out per fold; the relative proportion of each label were the same in both train and held-out sets using a stratified sampling. As is standard practice, the model uses user-provided group labels (e.g., IHKA vs. CON), and the model's performance is assessed by the concordance of the true labels with the model's predicted labels on the held-out dataset. A linear classifier model can have the following outcomes for each sample in the held-out set: true positive (model correctly predicts the positive "class", i.e., group label), true negative (model correctly predicts the negative class), false positive (model incorrectly predicts the positive class), and false negative (model incorrectly predicts the negative class). These outcomes are used to obtain the F_1 scores (described below). To classify drug identity alone, data from all doses (low and high) were merged. The mean and standard error of performance metrics are reported. We evaluated the performance of our linear classifier by computing confusion matrices, precision-recall curves, and F_1 scores.

Confusion matrices

Each confusion matrix is a square matrix where the number of rows and columns are equal to the number of possible target labels (e.g., six drug-dose pairs). Each square (indexed by i,j) represents the proportion of time a data point with the true label *i* was classified with label *j*. When *i=j*, the classifier correctly predicted the label; thus, an ideal classifier generates a matrix with a white diagonal amidst otherwise black fields (where classification rate is plotted on a scale of black to white, corresponding to 0 to 1). Matrices were normalized to one across each row and column to indicate the probability of classification or misclassification. The held-out confusion matrix (Figure S2D) was calculated by *N* repetitions of the training and evaluation process, where *N* is the number of treatment groups, in order to analyze the treatments which the classifier considers most similar to the target treatment class. For each iteration, one target class was removed from the training set and added into the held-out set of each fold, which forces the classifier to never correctly classify the held-out treatment class. The complete held-out confusion matrix was generated by repeating this process for all treatment groups.

Precision-recall (PR) curves

Precision and recall are calculated as follows:

$$Precision = \frac{t_p}{t_p + f_p} \;\&\; Recall = \frac{t_p}{t_p + f_n} \begin{cases} t_p = number\ of\ true\ positives \\ f_p = number\ of\ false\ positives \\ t_n = number\ of\ true\ negatives \end{cases}$$

The number of false positives, true positives, and true negatives are the outcomes of the linear classifier model, as described above. PR curves were generated to plot the precision and recall of linear classifier model as a decision threshold is varied, i.e., by measuring the false-positive and true-positive rates at a decision threshold for all data in the validation set. The number of false positives, true positives, and true negatives are the outcomes of the linear classifier model. The harmonic mean of precision and recall gives the F_1 score, which is a measure of binary classification performance:

$$F_1 = 2\,x\,\frac{precision\ x\ recall}{precision + recall}$$

F_1 values were calculated for each label class, and class-weighted averaging across F_1 score of all classes was used to generate a single mean F_1 score as a behavioral summary. Standard errors were also calculated.

Cosine distance matrix for behavioral summary distance comparisons

The cosine distance was computed between pairs of behavioral summaries (using the SciPy Python package). The cosine distance was chosen since it is well suited for high-dimensional data and allows a comparison between behavioral summaries with different units. To visualize the relationships between behavioral summaries, we show a reordered square matrix containing all pairwise cosine distances using hierarchical clustering (Ward's linkage).

Discrimination-relevant syllables ("behavioral fingerprint") and transitions

As previously described,[22] we used a *F* univariate statistical test to identify which syllables were most relevant for discriminating between conditions (e.g., between drug-dose pairs or between timepoints during epileptogenesis), with the reasoning that syllables whose usage frequency was highly statistically dependent on a given condition would be useful for linear classification, and therefore be considered characteristic of that condition. To represent the grammatical relationship between behavioral syllables, we performed an analysis of the bigram probabilities and the entropy rate as described previously.[12]

Visualizing behavioral summaries with low-dimensional embeddings

To visualize the relationship between conditions (e.g., between drug-dose pairs, between timepoints of epileptogenesis), we calculated low-dimensional (2D or 3D) embeddings from MoSeq behavioral summaries (i.e., mean syllable usages per session). We used linear discriminant analysis (LDA; from the Scikit-Learn Python package using the 'svd' – singular value decomposition – solver) to calculate a low-dimensional projection of the behavioral summaries which maximized linear separability between groups. Similar

separation was obtained with 2D or 3D embedding. The LDA input for Figure 2F with 8 groups was 59 syllables x 80 sessions, for Figure 3F with 7 groups, 50 syllables x 96 sessions, and for seizure behavior (30-sec windows) in Figure 4G with 4 groups, 27 syllables x 125 sessions, and in Figure 4J with 5 groups, 27 syllables x 125 sessions.

Human vs. Machine performance in identifying epileptic animals

To benchmark the ability of MoSeq to identify epileptic animals based on the inter-ictal phenotype, we compared its performance to that of four trained epilepsy researchers (blinded to experimental conditions) who were shown multiple 1-minute video clips from epileptic and non-epileptic mice and tasked with allocating them to the appropriate group. We trained a linear classifier on MoSeq-identified syllables or other common behavioral measures such as position, speed, or a combination of different scalars. We evaluated classifier performance as described above and reported the F_1 scores and classification matrices. The performance of human experimenters was evaluated by computing the mean of the F_1 scores of each experimenter. Note that the expert classification shown in Figure 1C had no influence on the grouping and was an additional experiment that was performed in order to compare MoSeq to human performance in identifying epileptic animals. For all experiments, the group labels were provided as metadata (see Figures S4B and S4C) at the time of data acquisition based on biological (e.g., genetic mutation) or physical variables (e.g., time after injection) in models of epilepsy (e.g., IHKA model of TLE).

Grouping of Racine score sets

To group Racine score sets (''RS sets'') into RS blocks, we first identified unique RS sets (30 unique RS sets in our dataset). Cosine distance was then used for grouping unique RS sets into RS blocks (see also section ''Cosine distance matrix for behavioral summary distance comparisons'' for details). The hierarchical clustering was thresholded at 3.5 using the fcluster function of SciPy Python package to reduce the number of clusters (i.e., number of RS blocks) for classification. For the kindling experiments, MoSeq extracted behavioral syllables that were similar to previous reports,[12,22,40] including syllables such as ''rearing'' and ''scrunching''. However, right after the kindling stimulus, MoSeq extracted previously unreported behavioral syllables that could be described with expressions found in traditional scoring systems such as ''wild running and jumping''.

Statistical tests

Error bars indicate 95% bootstrap confidence intervals. For statistical tests that assume normality, distributions were assumed to be normal but was not formally tested. In box-and-whisker plots of linear classifier performance, the box represents the distribution across 500 cross-validation folds and whiskers represent 1.5-times the inter-quartile range.

Behavioral syllables with differential usage across conditions (e.g. CON vs. IHKA, or *Scn1b*$^{+/-}$ vs. *Scn1b*$^{+/+}$) were identified using the Kruskal-Wallis test, post-hoc Dunn's two-sided tests with permutation, and Benjamini–Hochberg false discovery rate (FDR) of 0.05. In the Kruskal-Wallis test, the H-statistic (from the actual data) and the H-permutation (from permuted data in which group labels were randomly shuffled for all groups) were calculated for each syllable. Raw P-values were computed using the ratio of permutations where H-permutation is larger than H-data. These P-values were corrected using the Benjamini-Hochberg FDR across syllables, and syllables with FDR < 0.05 were identified as significant. For each syllable with FDR < 0.05 in the Kruskal-Wallis test, we performed a Dunn's post-hoc two-sided test by calculating z-data (the z-statistic from the actual data) and z-permutation (from permuted data in which group labels were shuffled). Raw P-values were computed using the ratio of permutations where z-permutation is larger than z-data. These P-values were corrected using Benjamini-Hochberg FDR across all pairwise comparisons, and syllables with FDR < 0.05 were identified as significant.

To assess differences in syllable usage between non-epileptic (CON) and chronically epileptic mice (IHKA) injected with either saline or high dose levetiracetam (LEV-H), we first normalized syllable usage of $IHKA_{LEV\text{-}H}$ mice to the difference between CON_{Saline} and $IHKA_{Saline}$. We reasoned that ''off-target'' effects should correspond to changes in $IHKA_{LEV\text{-}H}$ syllable usage which diverged from usage observed in controls, and ''on-target'' effects should correspond to changes in $IHKA_{LEV\text{-}H}$ syllable usage which moved closer to the usage observed in controls. We used the Kruskal-Wallis and post-hoc Dunn's two-sided test with permutation, with P-values corrected using the Benjamini–Hochberg FDR; syllables with FDR < 0.05 were identified as significant.

To identify syllables which distinguished a given drug-dose pair from the control treatment (i.e., one-vs-CON comparison) or from all other treatments (one-vs-rest), we compared syllable usage using the two-sided *F*-test with P-values corrected using the Holm-Bonferroni correction and significance set to P<0.01. These statistically significant syllables are indicated in wordclouds with their size scaled to the normalized *F* statistic and color-coded by whether the syllable was significantly up- or down-regulated (red or blue, respectively).

Differences in F_1 scores across behavioral summary types were tested for statistical significance using the paired two-sided t-test, corrected with Holm-Bonferroni step-down procedure, with significance set at P<0.01 after correction.