Research Host: Dr. Alexander Rotenberg
Location: Neurology Research Department at Boston Children's Hospital

Project Title:

A Novel Machine-Learning Approach for Treatment Appraisal of Gene Therapy in SSADH Deficiency

Background:

ALDH5A1 is a gene encoding for succinic semialdehyde dehydrogenase (SSADH), an enzyme responsible for the catabolic break-down of γ-aminobutyric acid (GABA), the major inhibitory neurotransmitter in the brain. In the absence of SSADH, GABA and its metabolite γ-hydroxybutyrate (GHB) accumulate to pathologic levels in the brain. The *ALDH5A1* loss-of-function mutation triggering this harmful accumulation of GABA and GHB is known as SSADH deficiency (SSADHD), a rare autosomal neurodevelopmental disease characterized by epilepsy and autistic behaviors for which neither curative nor disease-modifying treatments exist at the clinical stage.

Recently, the Rotenberg Lab at Boston Children's Hospital has pioneered an adeno-associated-virus (AAV) mediated gene replacement therapy which has proven to be promising at the preclinical stage[1]. AAV gene replacement therapy works by selectively delivering a viral vector, which includes a wild-type copy of a therapeutic gene and its proper regulatory elements, through cell-type specific infection. In SSADH deficient mice, treatment is designed to drive expression of *ALDH5A1* in neuronal cells with a viral vector including the human-recombinant *ALDH5A1*-transgene and its full-length native promoter[1].

Before moving to clinical testing, it must be clearly determined that treatment is effective at the preclinical stage. To do so, I will develop a metric to appraise treatment efficacy systematically and without human bias.

The native exploratory and dwelling behavior in mice is a close analogue for complex human behavior, which might reflect treatment efficacy. Preliminary data suggests that mutant mice behavior is distinct from wild-type (normal) littermates, which is amenable to SSADH restoration across treatment groups (**Figure 1**). Meanwhile, other metrics such as survival rate and EEG provide opportunity to cross-validate treatment efficacy in mice.

As such, I propose implementing a machine-learning supervised procedure to quantify mouse behavior, which I believe is an effective proxy for treatment efficacy. A significant knowledge gap in the field exists on how to appraise the efficacy of a gene therapy treatment on SSADHD by an automated, unbiased approach through systematic behavioral analysis. I will attempt to bridge this gap by performing topological data analysis and higher-order Markovian modeling of behavioral datasets, which I will obtain from novel machine-learning techniques (**Figure 2**).

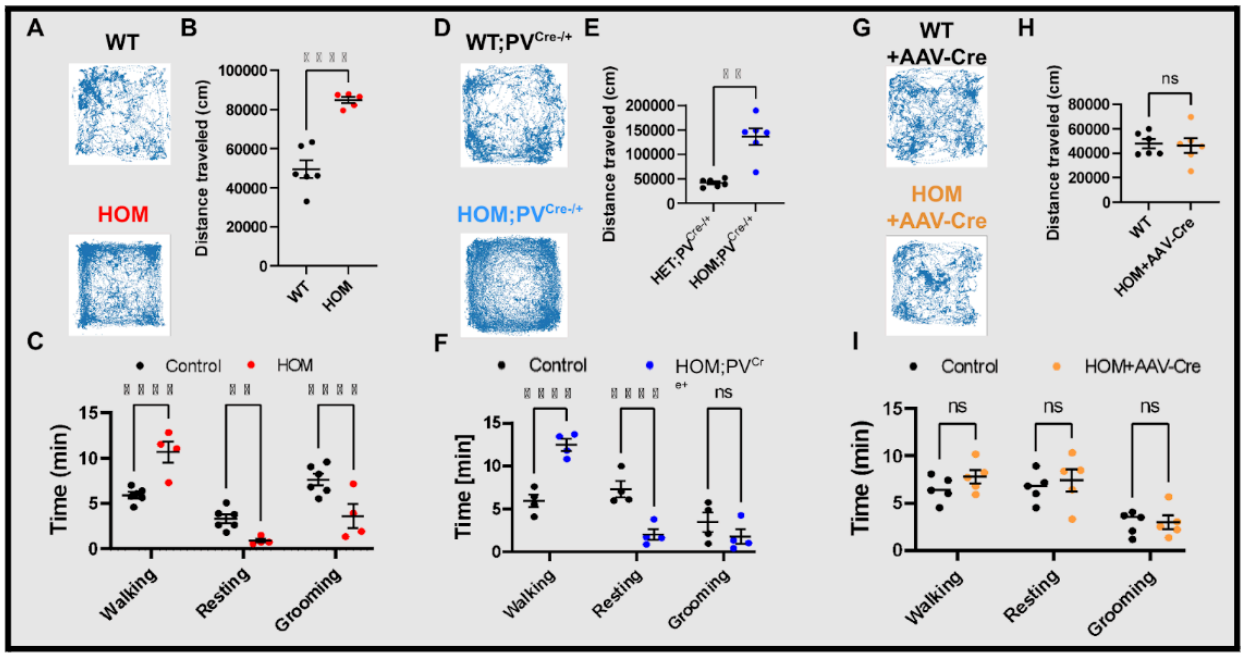

Figure 1. Behavioral reversal upon AAV-mediated brain-wide *ALDH5A1* restoration. Native exploratory behaviors in mice are captured by a high definition ventral view camera. Automated animal tracking was performed and distance traveled quantified. Detailed dwelling behaviors were analyzed manually, where behavioral episodes including walking, resting, and grooming were recorded and summarized. (A-C) Comparisons between untreated wild-type and homogeneous SSADH mice. (D-F) Mice with partial SSADH restoration by genetically targeting parvalbumin-expressing interneurons. (G-I) Mice treated with AAV-Cre targeting brain-wide SSADH; individual mouse data points are shown.[1]

Machine-Learning Pose Estimation and Behavioral-Motif Analysis:

Recent advances in machine-learning regarding keypoint-tracking algorithms such as DeepLabCut and motion-sequencing packages like MoSeq have enabled researchers to generate time-series data of behavioral-motifs without supervision[3].

Using video recordings of treated mice across various experimental conditions, I will first implement a reinforced-learning-from-human-feedback (RLHF) keypoint-algorithm in DeepLabCut. DeepLabCut returns continuous keypoint-data, which looks like a time-series of superimposed skeletons on mice. Using DeepLabCut's deep-learning inference, the mouse can be rapidly tracked without significant manual input. Manual analyses have already been performed, and these will cross validate my novel approach.

Next, I will apply the Keypoint MoSeq package to the keypoint-data from DeepLabCut. Through MoSeq, an unsupervised machine-learning package, I will extract a string of time-indexed behavioral patterns. The output of this pose analysis pipeline will be the time-series behavioral-motif sequence needed to perform my data analysis in the next section (**Figure 2**).

It is worth noting that the application of MoSeq to tracking-data is a new but proven-effective practice in the field. A 2023 study used this approach to determine hidden

behavioral-motifs underlying epilepsy, though depth-imaging rather than keypoint-data was used[4]. I hope to uncover similar insights regarding the underlying behaviors of differential efficacy in gene therapy.

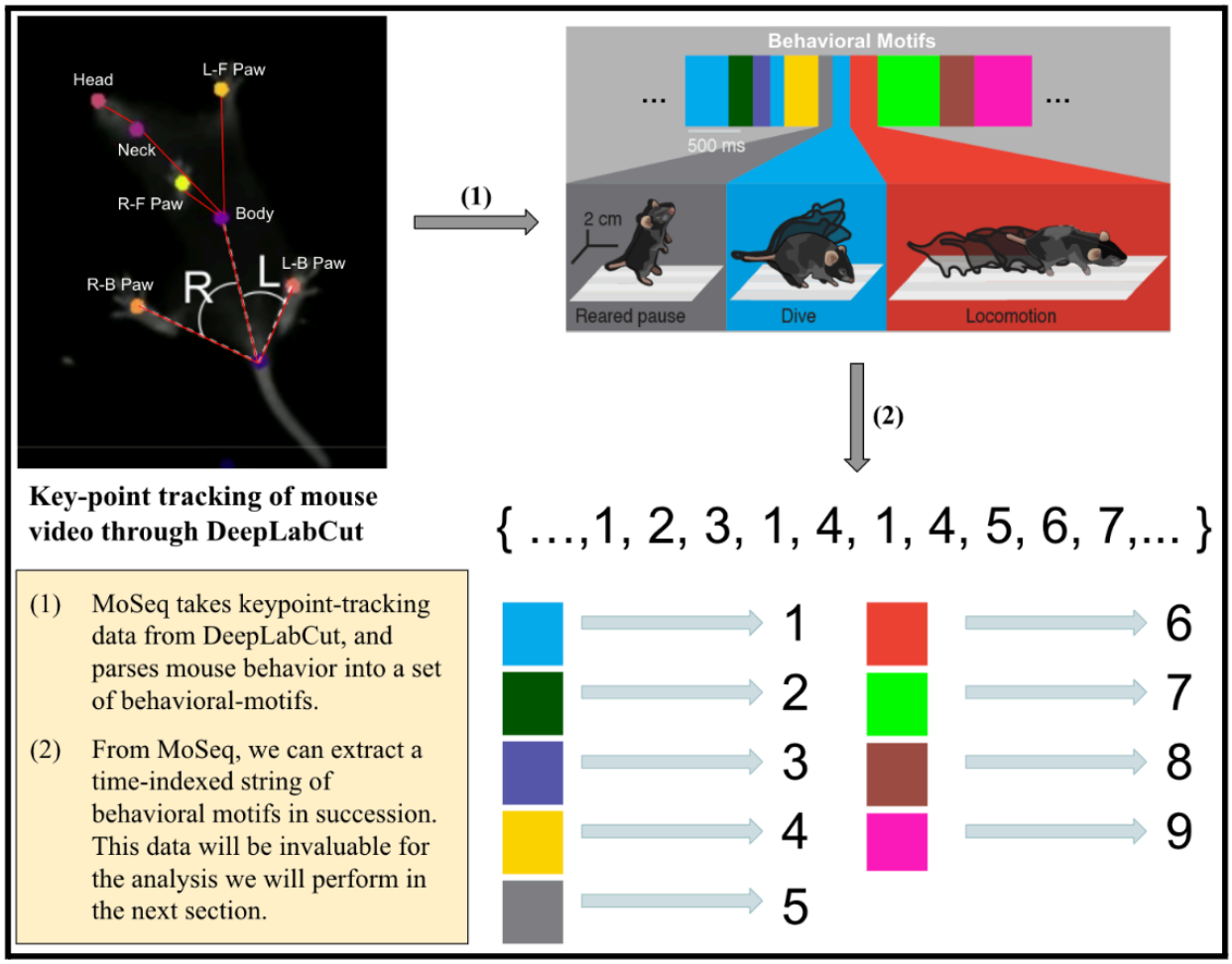

Figure 2. Pose-tracking keypoint data from DeepLabCut is fed into MoSeq, which then spits out a sequence of common behavioral motifs. We can then convert this sequence into a string by representing each behavioral motif as a character, on which we will perform subsequent data analysis. Image in the top right adapted from MoSeq.

Higher-Order Markov Chains and Topological Data Analysis:

For each treatment group, we now have time-series strings of behavioral-motifs. I present two approaches to analyzing this data:

(1) Modeling the behavioral-transition network with a higher-order Markov-chain, and
(2) Performing topological data analysis on a spring-embedded point-cloud.

Firstly, a study found that decision-making processes of mice self-generate stochasticity[5]. Thus, I decided a Markov chain would be appropriate. However, because decisions are still informed and do not fully exhibit memorylessness, I will implement a higher-order Markov chain[6].

Implementing (1) in Python is straightforward. Counting substring-to-substring transition frequencies by parsing the string, we can construct a transition matrix, where a_{ij} represents the i-to-j transition probability (**Figure 3**).

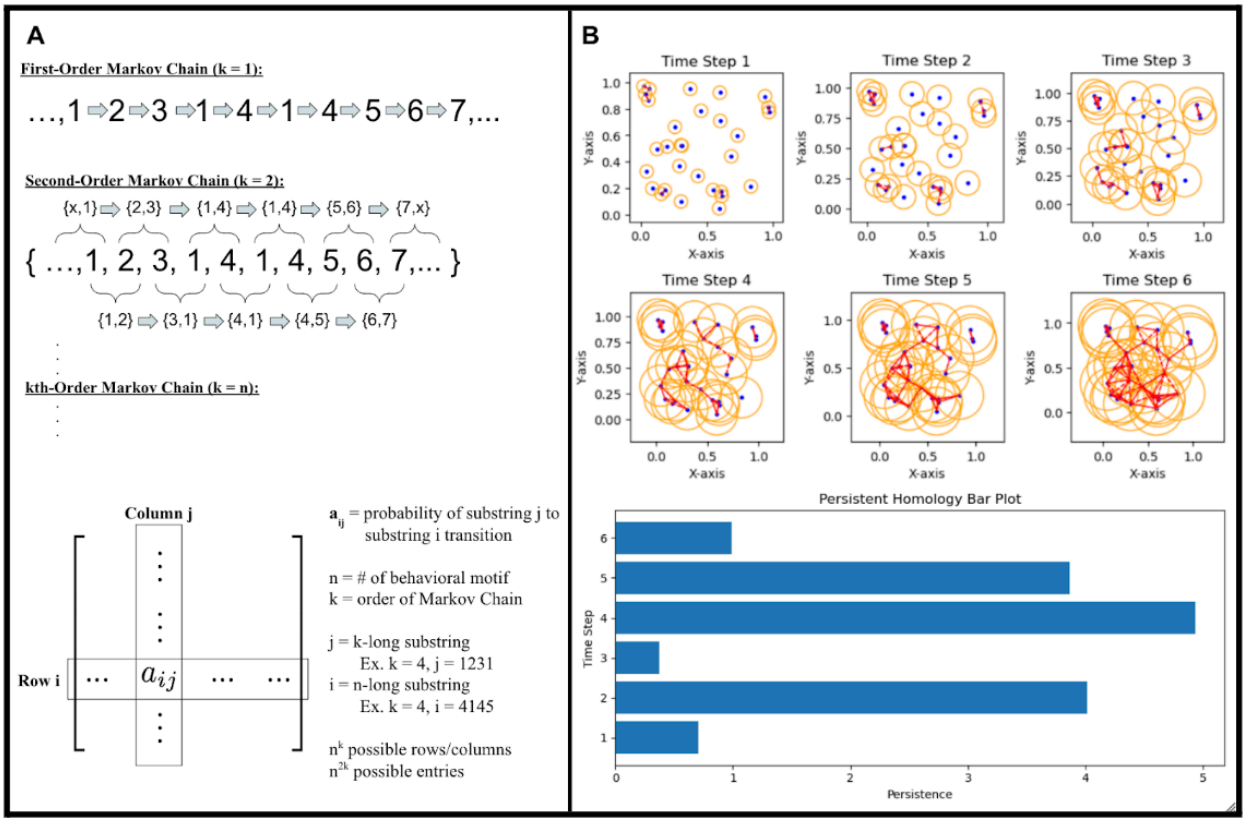

Figure 3. (A) Let k = the order of the Markov chain, where the next decision is informed by the mouse's previous k decisions, and n = the number of behavioral-motifs, treating each behavioral-motif as a node in the chain. We parse through our string of time indexed behavioral motifs for each length 2k substring and note the length-k substring to substring transition, iterating this from index 0 to index k-1. For example, with a k=2 Markov chain, I start at index 0 and count the frequency of every pairwise-transition (Figure 3), then we repeat starting from index 1. We then construct a transition matrix, where entry a_{ij} represents the probability of the substring-i to substring-j transition (i-to-j transition frequency/sum of all transitions from i). (B) Topological data analysis of a 2-dimensional point cloud. Points become ever-connected as resolution decreases across time-steps, and changing homology is noted on the bar graph below.

Matrix representation allows me to compare behavior across treatment groups quantitatively. For example, I will compare the steady-state vectors of each matrix by taking the

norm of their differences; higher values suggest stronger divergence between overarching behaviors. Additionally, I will use Kullback–Leibler divergence to compare steady-states as probability distributions, for which MatLab code is readily available[7].

Implementing (2) is more complicated. Let each k-length substring be represented as points in 3-dimensional space. We would like points with higher transition probabilities to be closer together. Thus, I will apply Tutte's spring embedding method, where we suppose every point is connected by spring constants proportional to the transition probabilities[8].

From this behavioral point-cloud, we can measure its persistent homology. A 2021 study has shown that persistent homology is an effective metric for distinguishing characteristic behaviors in worms, and I believe this will hold true for mice[9]. Put simply, persistent homology is an algebraic object which describes how topological invariants of a point-cloud persist over varying resolutions as it connects itself into manifolds (**Figure 3**). Persistent homology is a powerful tool for summarizing the shape of our data, and by analyzing the most persistent features of the point-cloud topology, we can distinguish the most characteristic behaviors.

Analyzing the topologies of behavioral state-spaces allows for a visually intuitive representation of our data. For example, suppose the behavioral manifold of one treatment group resembles a donut and another a pretzel. These tools will potentially serve as very powerful methods to quantify and infer treatment levels in new mouse data.

Impact and Innovations

Applications of Markovian and topological analysis to behavioral data in mice upon gene therapy have not been attempted. Establishing novel methodologies is a necessary step towards a standardized quantitative measure to assess treatment efficacy. We note that similar machine learning-assisted approaches might be translatable to the clinic, to better understand and analyze complex human behaviors applicable broadly to autism, epilepsy, and other neurologic disorders[10].

References:

(1) Liebhardt, Amanda. "Reduced Epilepsy Morbidity and Mortality upon Gene Replacement Therapy in Succinic Semialdehyde Dehydrogenase Deficiency SSADHD." aesnet.org/abstractslisting/reduced-epilepsy-morbidity-and-mortality-upon-gene-replacement-therapy-in-succinic-semialdehyde-dehydrogenase-deficiency-ssadhd.

(2) Unpublished data from the Rotenberg Lab.

(3) Weinreb, Caleb, et al. *Keypoint-MoSeq: Parsing Behavior by Linking Point Tracking to Pose Dynamics*, 17 Mar. 2023, https://doi.org/10.1101/2023.03.16.532307.

(4) Gschwind, Tilo, et al. "Hidden behavioral fingerprints in epilepsy." *Neuron*, vol. 111, no. 9, May 2023, https://doi.org/10.1016/j.neuron.2023.02.003.

(5) Belkaid, M., Bousseyrol, E., Durand-de Cuttoli, R. et al. Author Correction: Mice adaptively generate choice variability in a deterministic task. *Commun Biol* **3**, 54 (2020). https://doi.org/10.1038/s42003-020-0785-8.

(6) Wu, Sheng-Jhih, and Moody T. Chu. "Markov chains with memory, tensor formulation, and the dynamics of power iteration." *Applied Mathematics and Computation*, vol. 303, June 2017, pp. 226–239, https://doi.org/10.1016/j.amc.2017.01.030.

(7) Fass, David. "Kullback-Leibler Divergence." *MathWorks*, 2016, www.mathworks.com/matlabcentral/fileexchange/13089-kldiv.

(8) Urschel, John C., and Ludmil T. Zikatanov. "Discrete trace theorems and energy minimizing spring embeddings of planar graphs." *Linear Algebra and Its Applications*, vol. 609, Jan. 2021, pp. 73–107, https://doi.org/10.1016/j.laa.2020.08.035.

(9) Thomas A, Bates K, Elchesen A, Hartsock I, Lu H and Bubenik P (2021) Topological Data Analysis of *C. elegans* Locomotion and Behavior. *Front. Artif. Intell.* 4:668395. doi: 10.3389/frai.2021.668395.

(10) Farooq, M.S., Tehseen, R., Sabir, M. *et al.* Detection of autism spectrum disorder (ASD) in children and adults using machine learning. *Sci Rep* **13**, 9605 (2023). https://doi.org/10.1038/s41598-023-35910-1.