Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Social microbehavioral signatures across sex, partner type, and developmental insult
Principal Investigator:	Sarah Ferri, PhD (Pediatrics) and Shane Heiney, PhD (Iowa Neuroscience Institute)
Prepared by (IIAI):	Zhi Chen
Other investigators:	
Date:	9/2/2025
Were specific aims fulfilled:	Y
Readiness for extramural proposal?	Y
If yes … Planned submission date	Fall 2025
Funding agency	NIH
Grant mechanism	R01
If no … Why not? What went wrong?	Additional data collected and to be analyzed

Brief summary of accomplished results:

We developed a segmentation-based pipeline for automated detection of mouse social sniffing behaviors. Using Annolid, we generated high-quality semi-automated annotations, which were further refined and used to train a fully automated nnU-Net 3D full-resolution segmentation model. The trained model successfully segmented mice and their body parts across five testing videos, with post-processing steps improving temporal consistency. By skeletonizing and subdividing segmentations into anatomically meaningful regions, we enabled robust detection of specific directed sniffing behaviors. The approach achieved 70–91% overall accuracy, demonstrating reliable detection of general interaction patterns while revealing the challenges of classifying rarer or ambiguously defined behaviors.

Research report:

Aims (provided by PI):

To use machine learning to identify individual complex microbehaviors during free interaction of mouse pairs and determine differences in patterns across sex and dyad type (familiar, aggressive, etc).

Data:

Our dataset consists of grayscale videos, each 10 minutes in duration, recorded at a resolution of 1280 × 960 pixels and 25 frames per second (15,000 frames per video). Figure 1 depicts a single frame from one of these videos. Each video captures two mice freely interacting in a rectangular arena, recorded in the dark under infrared illumination. Ground-truth labels of microbehaviors, provided by expert manual annotation, are available for 7 videos. One of these labeled videos was excluded from analysis due to overexposure, which resulted in characteristics that differed from the rest of the dataset.

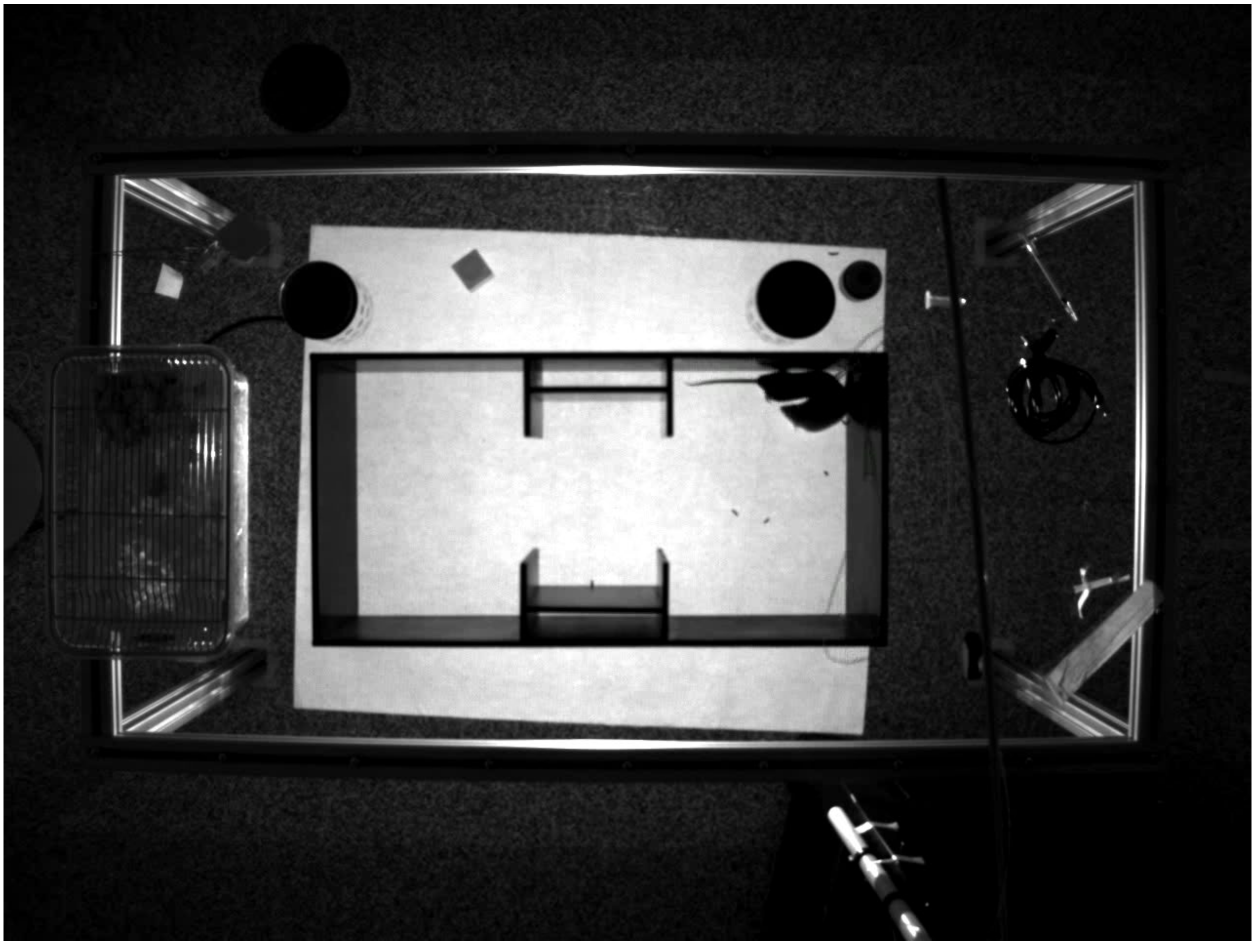


Figure 1. Representative frame from social interaction video.

AI/ML Approach:

We utilized the Annolid platform to trace the outlines of two mice. Annolid leverages state-of-the-art machine learning for video object segmentation and tracking, combining instance segmentation with the Cutie model to enable robust, markerless tracking of multiple animals from minimal annotations. It further integrates Segment Anything and Grounding-DINO to support automatic text-prompted masking and segmentation, reducing reliance on manual labeling.

In addition, the **nnU-Net** platform was employed to perform fully automated segmentation of the two mice. nnU-Net is a self-configuring deep learning framework based on the U-Net architecture, which adapts its preprocessing, network architecture, training pipeline, and postprocessing automatically to the given dataset. It employs convolutional neural networks optimized for biomedical image segmentation tasks, using techniques such as data-driven normalization, patch-based training, and extensive data augmentation.

Experimental methods, validation approach:

1. Preprocessing:

As shown in Figure 1, many irrelevant objects outside the mice's movement area could interfere with subsequent segmentation and classification tasks. To minimize this noise and normalize the input videos, we applied a preprocessing pipeline. A central frame was first extracted from each video to serve as a reference image, on which bounding boxes were manually defined by selecting four corner points. These bounding boxes were then applied to crop the original videos to the regions of interest, with background subtraction, downsampling, and smoothing, producing standardized cropped video outputs (500 x 250 pixels) for further analysis (Figure 2).

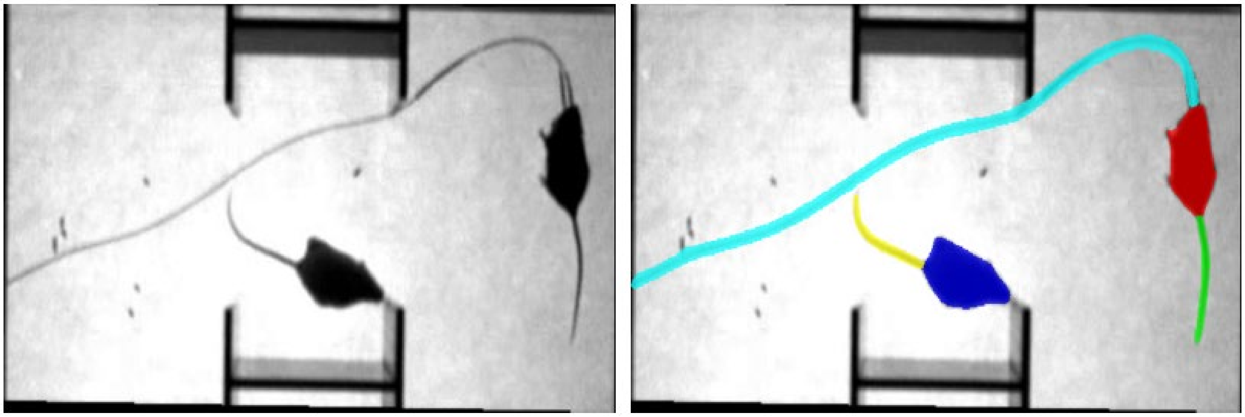


Figure 2. Example video frame after cropping and smoothing. Left: region of interest only. Right: outlines of five labeled objects—FP mouse (red), FP tail (green), ST mouse (blue), ST tail (yellow), and cable (cyan).

2. Semi-automated annotation of mice:

A total of 15,000 frames from one video were semi-automatically annotated using the Annolid platform with five labels: FP mouse, FP tail, ST mouse, ST tail, and cable (Figure 2). After loading the video, the expert selected the first frame in which both mice and all relevant objects were visible. Annolid, powered by Meta's Segment Anything Model, automatically generated object outlines with a single click, which the expert refined and assigned to one of the five labels. Once all objects were annotated in the initial frame, the Cutie model was applied to propagate predictions across subsequent frames. The expert then iteratively reviewed these predictions, made corrections as needed, and relaunched the prediction process from the corrected frame, repeating this cycle until all frames were fully annotated.

The annotation masks were stacked into 3D volumes in NIfTI format for compatibility with nnU-Net. To improve label quality, the masks were cleaned to retain only the largest connected components for major structures, thereby removing noise and artifacts. The resulting labels were independently reviewed and refined by a second expert to ensure accuracy, and the finalized annotations were then used as the training dataset for subsequent automated segmentation.

3. Fully automated segmentation of mice:

All videos were converted into 3D NIfTI volumes with isotropic voxel spacing and normalized using z-score normalization based on frame intensity statistics. We trained nnU-Net in its 3D full-resolution configuration, using a patch size of 256×64×128 voxels and a batch size of 2, optimized for the large spatial dimensions of our data. The model architecture was a convolutional encoder–decoder U-Net with deep supervision and residual connections. Training incorporated extensive data augmentation, including spatial resampling, intensity normalization, and patch-based sampling, to enhance robustness.

For testing, each of the five input videos was split into grayscale frames and converted into 3D NIfTI volumes, divided into ten parts per video for efficient processing. The trained nnU-Net 3D full-resolution model was then applied to generate segmentation predictions for every frame. To improve consistency, the raw predictions were post-processed: label cleaning was performed to retain only the largest connected components for the mice (FP and ST) and their tails, while removing redundant detections such as cables, which are unrelated to mouse behaviors. Finally, missing labels in frames where an object was not detected were interpolated by estimating motion vectors from the previous five frames and shifting the most recent valid mask to the predicted position.

4. Segmentation-based behavior detection:

To detect social sniffing behaviors, the segmentation outputs were post-processed into anatomically meaningful subregions. First, the masks of each mouse were combined with their corresponding tail masks and skeletonized to capture the midline structure. For each skeleton, the point nearest to the tail was designated as the body's "bottom," and distances from this anchor point were used to divide the skeleton into three functional regions: head, body/flank, and anogenital. Based on expert observation, the frontmost 5% of pixels were assigned to the head, the rearmost 30% to the anogenital region, and the remaining pixels to the flank. The original segmentations of the FP and ST mice were then reassigned to these skeleton-derived regions using nearest-neighbor mapping, producing refined labels (FP head, FP flank, FP anogenital; ST head, ST flank, ST anogenital). These anatomically enhanced labels (Figure 3) provided the foundation for detecting directed sniffing interactions.

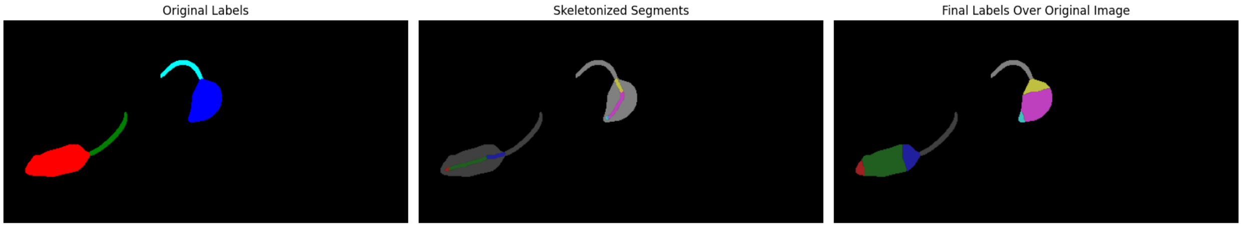


Figure 3. Example of segmentation-based body part refinement. *Left:* Original segmentation labels for the FP and ST mice along with their tails. *Middle:* Skeletonized representations divided into head, flank,

and anogenital regions based on distance from the tail. *Right:* Final anatomically enhanced labels overlaid on the original masks, producing refined regions for subsequent behavior detection.

Using these part-based labels, frame-by-frame interactions were classified by detecting spatial overlaps within a small neighborhood around the head regions. Binary masks of the FP head were compared with the ST head, flank, and anogenital regions to identify nose-to-nose, nose-to-flank, and nose-to-anogenital sniffing events, respectively. Overlaps between the ST head and any region of the FP mouse were classified as stimulus-to-FP sniffing. To account for variability in contact distance, the head masks were dilated with a structuring element of 4 pixels before overlap detection, ensuring that near-contacts were also captured.

Results:

The resulting frame-level behavior predictions were exported as labeled time series and evaluated against expert-annotated ground truth. Performance was quantified using confusion matrices, with class-wise sensitivity and specificity as well as overall accuracy reported.

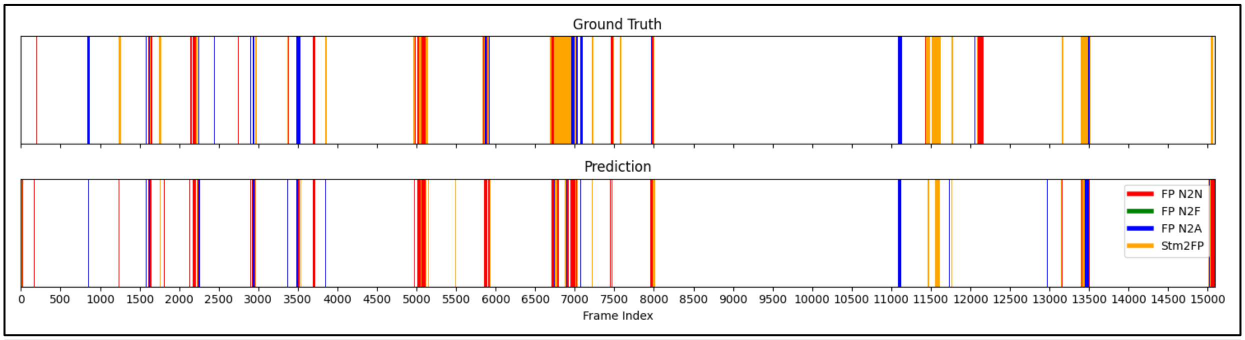
Across all 5 testing videos (shown in Table 1), segmentation-based behavior detection achieved high accuracy in identifying frames without mouse interactions, with sensitivities ranging from ~81% to 99%. Detection of specific sniffing interactions was more variable. Nose-to-nose sniffing from FP to ST was recognized with moderate-to-high sensitivity (53–78%) and high specificity (>91%). Nose-to-flank sniffing showed moderate sensitivity (41–63%) with consistently high specificity (>91%). Nose-to-anogenital sniffing proved the most challenging, with lower sensitivities (33–59%) despite strong specificity (>96%). Detection of stimulus-to-FP sniffing was weakest, likely due to its ambiguous definition that does not explicitly categorize interaction parts, yielding sensitivities of only 8–30% but specificity above 95% in all cases.

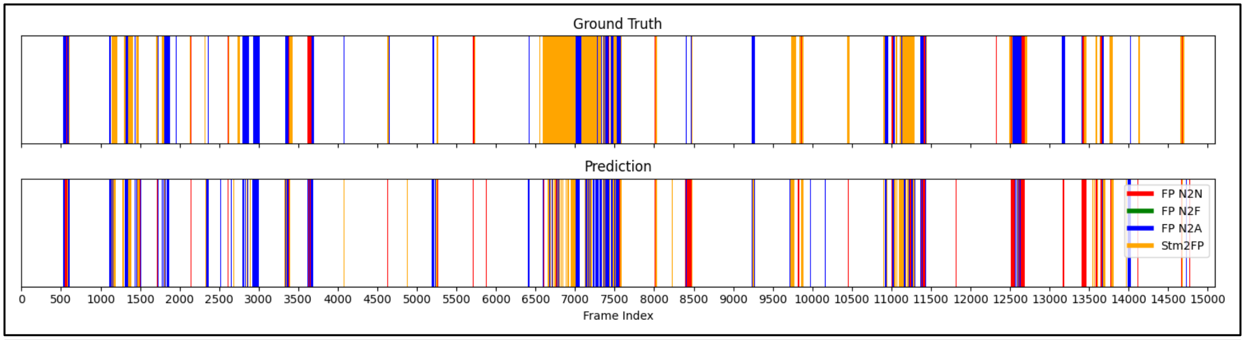
	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Accuracy	88.3%	85.2%	90.7%	70.5%	84.9%
No Interaction Sens	98.5%	95.1%	97.8%	80.5%	90.8%
No Interaction Spec	68.1%	72.9%	62.0%	82.1%	91.6%
Nose-to-Nose Sens	63.7%	64.7%	52.6%	64.0%	77.8%
Nose-to-Nose Spec	97.3%	96.9%	97.2%	91.6%	95.5%
Nose-to-Flank Sens	62.9%	59.0%	40.7%	51.2%	57.6%
Nose-to-Flank Spec	97.8%	96.4%	98.9%	91.2%	93.5%
Nose-to-Anogenital Sens	32.9%	47.7%	58.7%	53.1%	46.5%
Nose-to-Anogenital Spec	99.2%	97.9%	99.7%	96.2%	97.6%
Stimulus-to-FP Sens	22.7%	16.7%	27.4%	7.8%	30.0%
Stimulus-to-FP Spec	99.7%	99.2%	99.1%	95.9%	99.1%

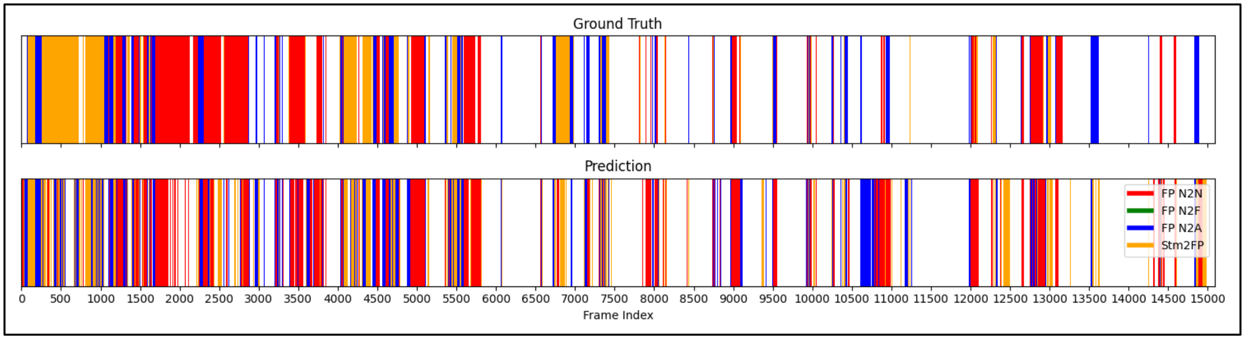
Table 1. Segmentation-based behavior detection performance across five datasets, showing accuracy, and sensitivity/specificity for each interaction type.

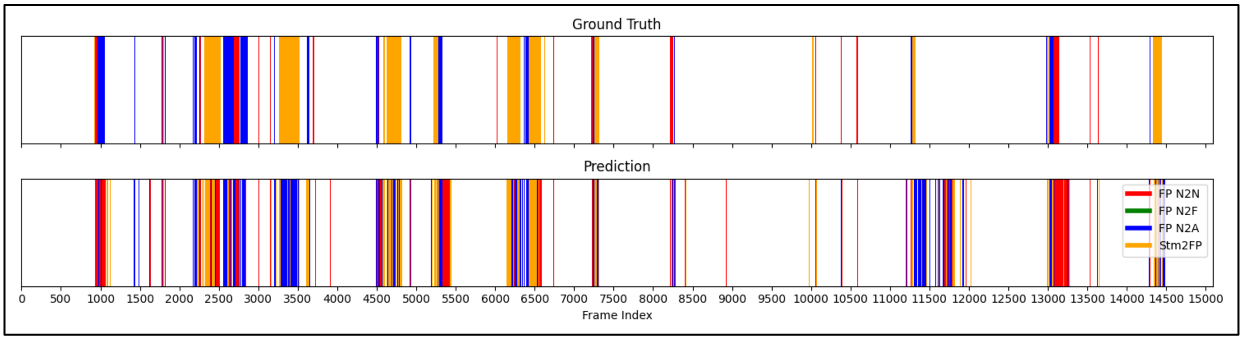
Overall accuracy across the 5 videos ranged from 70% to 91%, reflecting robust performance in distinguishing sniffing versus no-sniffing frames but highlighting the difficulty of reliably classifying less frequent or ambiguous social interactions.

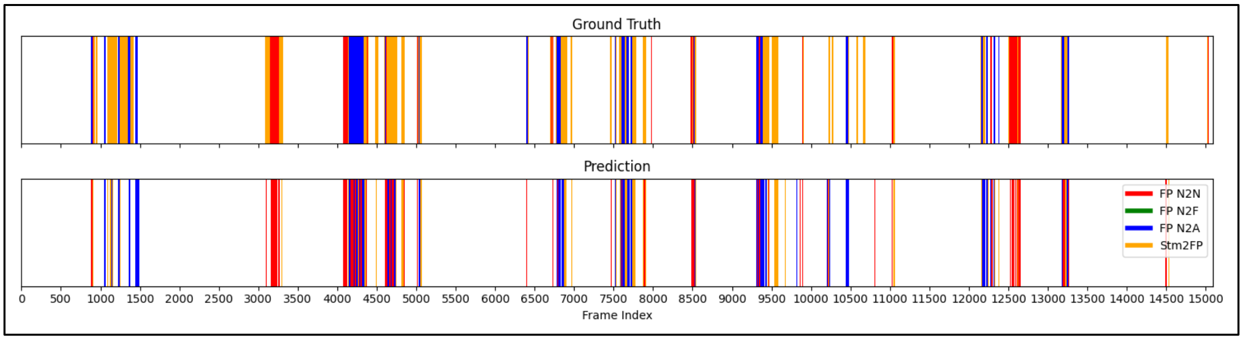
For each video, a figure below presents a frame-by-frame comparison between expert-annotated ground truth (top) and segmentation-based predictions (bottom) of sniffing behaviors across 15,000 frames. Colored bars denote different interaction types: nose-to-nose (red), nose-to-flank (green), nose-to-anogenital (blue), and stimulus-to-FP sniffing (orange). The visualization highlights both concordance and discrepancies between annotations and predictions, showing that the model reliably captured the overall timing and distribution of sniffing events. Occasional mismatches and misclassifications were observed, particularly for rarer behaviors and in cases where subjective inconsistencies in manual labeling were identified upon further review.











Ideas/aims for future extramural project:

- Enhancing ground truth annotations: Improve the quality and resolution of ground truth labels by explicitly categorizing stimulus-to-FP (ST-to-FP) behaviors. This will reduce ambiguity in labeling and enable more precise evaluation of directed interactions.
- Expanding and diversifying the training dataset: Incorporate a larger set of annotated videos that capture a wider range of behaviors and environmental conditions (e.g., different lighting, camera angles, and instances of overexposure). This will help increase the robustness and generalizability of the segmentation model across diverse experimental settings.
- Broadening behavior categories: Extend the behavioral annotation framework beyond sniffing to include transitional behaviors such as pre-sniffing approaches and post-sniffing withdrawals. These additions will allow for a more complete characterization of social interaction dynamics.
- Characterizing solo behaviors: Develop classification methods to detect solo behaviors of FP and ST mice (e.g., grooming, rearing) independently of social interaction frames. This will complement interaction-based segmentation and provide a comprehensive view of individual activity.
- Incorporating motion dynamics: Integrate temporal motion analysis across neighboring frames to determine which mouse initiates and terminates an interaction. This will help resolve object-swapping errors during close contacts and improve the accuracy of labelling social interactions.

Publications resulting from project:

N/A