BRIDGING SCALES BETWEEN CHEMICAL SPACE AND BEHAVIORAL PHENOTYPE

Adrien Jouary[1] **, J. Miguel Mata**[2]**, Dean Rance**[1]**,**
Gonzalo G. de Polavieja[1]**, Christian K. Machens**[1]**, Michael B. Orger**[1]
[1] Department of Neuroscience, Champalimaud Foundation
[2] Department of Medicinal Chemistry, Leiden Academic Centre for Drug Research, Universiteit Leiden
adrien.jouary@research.fchampalimaud.org

1 INTRODUCTION

Machine learning in drug discovery has predominantly focused on molecular or cellular-level interactions, such as predicting binding affinities between small molecules and a protein (Vamathevan et al., 2019). Organism-level behavioral screening offers a powerful complementary approach. Behavior provides a holistic view of compound effects by capturing interactions across biological pathways and enabling validation of therapeutic outcomes at the whole-organism level. In model organisms such as *C. elegans* (O'Brien et al., 2025) and zebrafish (Kokel et al., 2010; Rihel et al., 2010), behavioral screening can be performed at scale ($>$10,000 compounds), making it viable for machine learning. When combined with genetic perturbations (Harpaz et al., 2021; O'Brien et al., 2023), these methods become a powerful tool for drug discovery.

Two recent advances are particularly promising for improving behavioral drug screening:

- Molecular embeddings can capture complex chemical properties (Chithrananda et al., 2020; Suryanarayanan et al., 2024).
- High-resolution behavioral analysis in zebrafish larvae can identify subtle changes in locomotion (Marques et al., 2018).

Here, we leverage these advances in two complementary studies: (1) establishing a cross-modal mapping between molecular structure and behavior, and (2) enhancing the sensitivity of behavioral phenotyping for drug screening.

2 USING MULTIVIEW REPRESENTATION TO LINK MOLECULAR AND BEHAVIORAL SPACE

To connect molecular structure with behavioral outcomes, we use multiview representation learning, a framework that learns shared representation across multiple sources simultaneously. In particular, we apply Canonical Correlation Analysis (CCA) (Chapman & Wang, 2021), a linear method that identifies projections maximizing correlation between two datasets.

We analyzed the published dataset of Gendelev et al. (2024), where zebrafish larvae (8 per well) were exposed to 653 CNS-targeting compounds. Behavior was quantified as the average time series of the motion index (overall pixel intensity changes over time) during exposure to various auditory and visual stimuli (see A.1 for details). Molecular embeddings were derived from the Simplified Molecular-Input Line-Entry System (SMILES) using pre-trained transformer models. We benchmarked three models: ChemBERTa (Chithrananda et al., 2020), MMELON (Suryanarayanan et al., 2024) and Unikei (unikei, 2025). For each model, we performed dimensionality reduction with Principal Component Analysis (PCA) to retain 50 components (capturing over 94% of the variance for molecular embedding and and 69% for the behavioral time series), and then applied CCA. In the remainder of this section, we focus on the Unikei embedding, which achieved the highest CCA correlation (see 1). CCA identified two statistically significant projection dimensions (Supp. Fig. 3 and Supp. Fig. 4). The first CCA dimension meaningfully separated different neurotransmitter class: dopaminergic/serotonergic ligands clustered distinctly from purines and metabotropic glutamate ligands (Fig. 1a). The corresponding behavioral projections revealed distinct sensitivity to the stimuli: high projections were associated with stronger responses to salient stimuli, while low projections

corresponded to higher baseline activity (Fig. 1b). These results demonstrate that, despite the complexity of linking molecular structure to organism-level behavior, multiview learning can uncover meaningful relationships.

As a direction for future work, exploring nonlinear multiview representations may help capture additional relationship between compounds and behavior. Another potential improvement is to refine behavioral quantification beyond the coarse motion index. In the next section we show that high-resolution behavioral analysis enhances our ability to distinguish drug phenotypes.

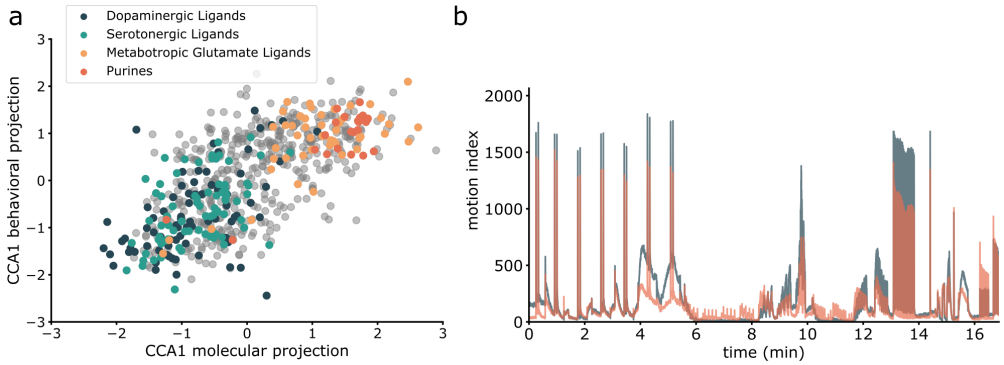

Figure 1: **a.** First CCA dimension for molecular and behavioral projection, color-coded by drug class. The drug class labels were not used to compute the representation **b.** Behavioral profiles for fish with the highest (top 10% in orange) and lowest (bottom 10% in green) projections on the first behavioral CCA axis, highlighting the extremes of the distribution.

3 Improving the sensitivity of behavioral phenotyping

Recent work in mice (Wiltschko et al., 2020) has shown that fine-grained behavioral analysis enhances the detection of drug effects. However, these approaches remain constrained to small-scale experiments, as scaling up rodent studies to high-throughput remains impractical. In zebrafish, behavioral screens have faced a similar trade-off: high-throughput assays rely on coarse metrics (e.g., motion index), while detailed analyses sacrifice scale. Advances in high-resolution imaging hardware, such as gigapixel cameras (Thomson et al., 2022), and machine learning pipelines, including Megabouts for pose estimation and action classification (Jouary et al., 2024), now resolve this conflict. Here, we assess how improved behavioral quantification enhances our ability to distinguish between drug phenotypes (see Fig. 2).

We recorded 162 zebrafish larvae exposed to 9 pharmacological compounds across diverse contexts (spontaneous behavior, photomotor/optomotor responses or visually driven escapes, see A.2 for details). To predict drug identity from behavioral time series, we used a MiniRocket classifier (Dempster et al., 2021) and evaluated performance using 10-fold cross-validation (Supp. Fig. 5).

The classifier achieved a test accuracy of 31% using a binary time series (movement vs. no movement) and 40% accuracy with the locomotion speed as input. Performance increased to 52% accuracy when using action categories (one-hot encoded tail movement categories), compared to a chance level of 11%. These results indicate that improving the quantification of behavior significantly boost the sensitivity of pharmacological phenotyping.

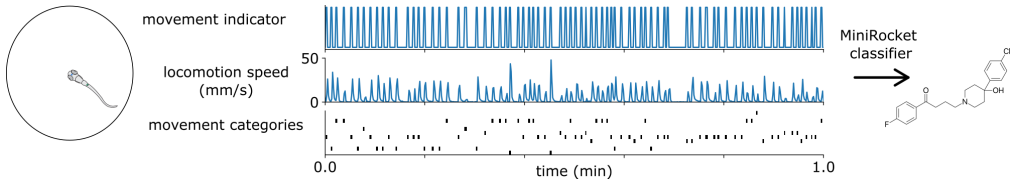

Figure 2: Classifying treatments based on zebrafish behavior measured via a binary movement indicator, locomotion speed, or the time series of tail movement categories.

MEANINGFULNESS STATEMENT

Meaningful life representations should capture hierarchical biological processes, bridging molecular to organism levels. Zebrafish behavioral screening offers a powerful and scalable readout of small molecule effects. However, current methods struggle to generalize across chemical space and rely on coarse behavioral quantification. We address these by (1) using pretrained molecular embeddings to map drug compounds to behaviors, successfully generalizing to new compounds, and (2) employing high-resolution behavioral data to enhance drug effect detection sensitivity. These advances highlight the value of integrating molecular and behavioral data to enhance drug discovery.

ACKNOWLEDGMENTS

C.K.M. acknowledges support from the Simons Collaboration on the Global Brain (543009); C.K.M and A.J. received support from Fundação para a Ciência e a Tecnologia (FCT-PTDC/BIA-OUT/32077/2017-IC&DT-LISBOA-01-0145-FEDER). M.B.O received support from the European Research Council (ERC NEUROFISH 773012) and Volkswagen Stiftung "Life?" initiative. A.J was supported by a Fellowship from the La Caixa Foundation (LCF/BQ/PR20/11770007). G.G.P acknowledges support from Fundação para a Ciência e a Tecnologia (FCT-PTDC/BIA-COM/5770/2020). Fish husbandry was supported by the Champalimaud Foundation Fish Facility, which received support from the research infrastructure CONGENTO, co-financed by the Lisboa Regional Operational Programme (Lisboa2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), and the Fundação para a Ciência e Tecnologia (Portugal) under the project LISBOA-01-0145-FEDER-022170.

REFERENCES

James Chapman and Hao-Ting Wang. Cca-zoo: A collection of regularized, deep learning based, kernel, and probabilistic cca methods in a scikit-learn style framework. *Journal of Open Source Software*, 6(68):3823, 2021.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Angus Dempster, Daniel F Schmidt, and Geoffrey I Webb. Minirocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 248–257, 2021.

Leo Gendelev, Jack Taylor, Douglas Myers-Turnbull, Steven Chen, Matthew N McCarroll, Michelle R Arkin, David Kokel, and Michael J Keiser. Deep phenotypic profiling of neuroactive drugs in larval zebrafish. *Nature Communications*, 15(1):9955, 2024.

Roy Harpaz, Ariel C Aspiras, Sydney Chambule, Sierra Tseng, Marie-Abèle Bind, Florian Engert, Mark C Fishman, and Armin Bahl. Collective behavior emerges from genetically controlled simple behavioral motifs in zebrafish. *Science advances*, 7(41):eabi7460, 2021.

Adrien Jouary, Pedro TM Silva, Alexandre Laborde, J Miguel Mata, João C Marques, Elena MD Collins, Randall T Peterson, Christian K Machens, and Michael B Orger. Megabouts: a flexible pipeline for zebrafish locomotion analysis. *bioRxiv*, pp. 2024–09, 2024.

David Kokel, Jennifer Bryan, Christian Laggner, Rick White, Chung Yan J Cheung, Rita Mateus, David Healey, Sonia Kim, Andreas A Werdich, Stephen J Haggarty, et al. Rapid behavior-based identification of neuroactive small molecules in the zebrafish. *Nature chemical biology*, 6(3): 231–237, 2010.

João C Marques, Simone Lackner, Rita Félix, and Michael B Orger. Structure of the zebrafish locomotor repertoire revealed with unsupervised behavioral clustering. *Current Biology*, 28(2): 181–195, 2018.

Thomas J O'Brien, Ida L Barlow, Luigi Feriani, and André EX Brown. High-throughput tracking enables systematic phenotyping and drug repurposing in c. elegans disease models. *eLife*, 12: RP92491, 2025.

Thomas J O'Brien, Ida L Barlow, Luigi Feriani, and André EX Brown. Systematic creation and phenotyping of mendelian disease models in c. elegans: towards large-scale drug repurposing. *bioRxiv*, pp. 2023–08, 2023.

Jason Rihel, David A Prober, Anthony Arvanites, Kelvin Lam, Steven Zimmerman, Sumin Jang, Stephen J Haggarty, David Kokel, Lee L Rubin, Randall T Peterson, et al. Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation. *Science*, 327(5963):348–351, 2010.

Parthasarathy Suryanarayanan, Yunguang Qiu, Shreyans Sethi, Diwakar Mahajan, Hongyang Li, Yuxin Yang, Elif Eyigoz, Aldo Guzman Saenz, Daniel E Platt, Timothy H Rumbell, et al. Multi-view biomedical foundation models for molecule-target and property prediction. *arXiv preprint arXiv:2410.19704*, 2024.

Eric E Thomson, Mark Harfouche, Kanghyun Kim, Pavan C Konda, Catherine W Seitz, Colin Cooke, Shiqi Xu, Whitney S Jacobs, Robin Blazing, Yang Chen, et al. Gigapixel imaging with a novel multi-camera array microscope. *Elife*, 11:e74988, 2022.

unikei. Bidirectional transformer pretrained on smiles. `https://huggingface.co/unikei/bert-base-smiles`, 2025. Accessed: 2025-02-12.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Alexander B Wiltschko, Tatsuya Tsukahara, Ayman Zeine, Rockwell Anyoha, Winthrop F Gillis, Jeffrey E Markowitz, Ralph E Peterson, Jesse Katon, Matthew J Johnson, and Sandeep Robert Datta. Revealing the structure of pharmacobehavioral space through motion sequencing. *Nature neuroscience*, 23(11):1433–1443, 2020.

A Appendix

A.1 Canonical Correlation Analysis Pipeline

SMILES embedding	train		test	
	CC1	CC2	CC1	CC2
ChemBERTa	0.77	0.61	0.65	0.28
MMELON	0.76	0.63	0.66	0.21
unikei*	0.78	0.64	0.68	0.24

Table 1: Table of Canonical Correlations. First canonical correlations of SMILES embeddings with motion indices, CCA fitted on training data and evaluated on both train and test data.

Motion Indices

The dataset consisted of 6144 time series (or "motion indices") of 17 min sampled at 100 Hz. To avoid contamination from high-frequency artifact, we reduced the sampling by 10 fold using max-pooling. We then averaged all the motion indices from the same drug exposure, to obtain an average response per drug, and dropped all motion indices for which we had no SMILES representation information. This resulted in 653 motion indices, each corresponding to a unique drug; this data was then split 80%-20% into train and test sets, and all further fitting (PCA, CCA) were done on the training set. The dimensionality of the motion indices was then reduced to 50 via Principal Component Analysis (PCA), capturing 69% of the total variance.

SMILES Embeddings

For the purpose of finding a vector for each drug, we embedded their SMILES. We tried three pretrained models, with results shown in Table 1. No further training was done on the models.

In each case, we then used PCA with 50 components to further reduce the dimensionality of the embeddings.

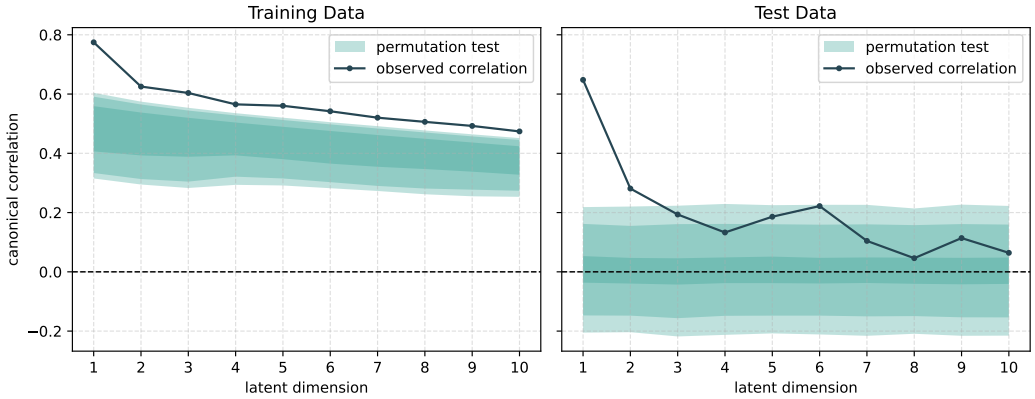

Figure 3: Canonical Correlations and Permutation Test Distribution. Canonical correlations for each latent dimension as computed on the training (left) and test (right) datasets. The shaded regions show the 1-99, 5-95, and 32-68 inter-percentile regions (with increasing opacity) of the permutation test distribution.

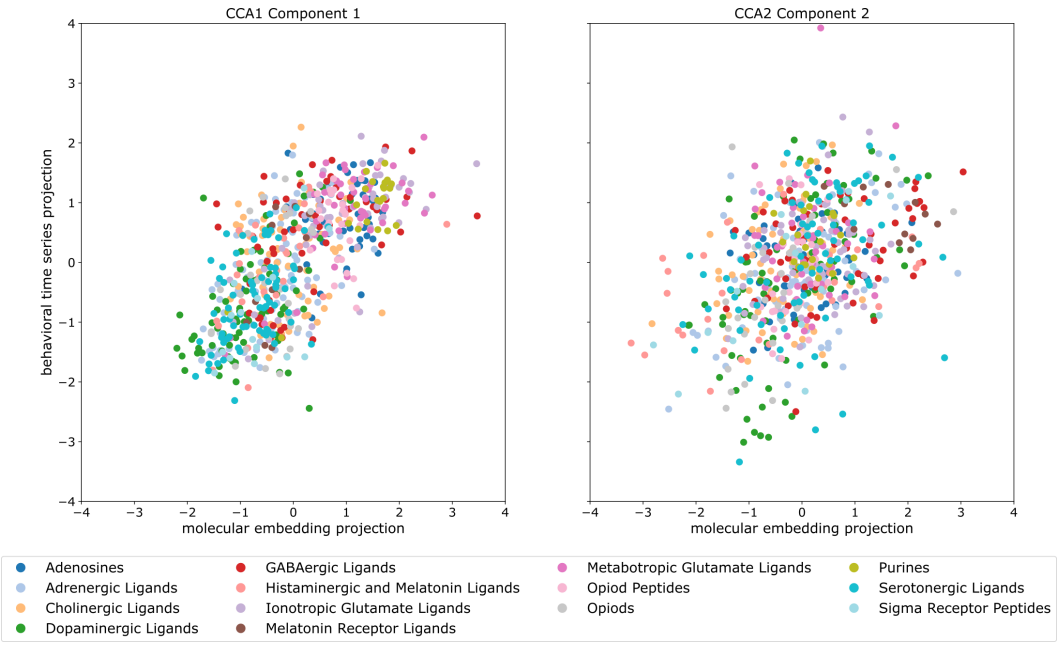

Figure 4: First (left) and second (right) CCA dimension for molecular and behavioral projection, color-coded by drug class. The drug class labels were not used to compute the representation.

REGULARISED CCA

We used regularised CCA Chapman & Wang (2021) which entails selecting a pair of regularisation parameters $c_1, c_2 \in [0, 1]$. To this end, we used grid search with 4-fold cross validation to select these parameters.

PERMUTATION TEST

To determine the validity of our results, we didn't solely rely on the test canonical correlations but also performed permutation tests. To this end, we permuted, and refit our models on, the training dataset to determine a distribution of canonical correlations when the data are artificially decorrelated (see Figure 3).

A.2 HIGH-RESOLUTION BEHAVIOR QUANTIFICATION DETAILS

FISH CARE

Adult fish were maintained at 28°C on a 14:10 hour light cycle. Embryos were collected and larvae were raised at 28°C in E3 embryo medium (5 mM NaCl, 0.17 mM KCl, 0.33 mM CaCl2 and 0.33 mM MgSO4) in groups of 25. Sexual differentiation occurs at a later stage, and therefore the sex of the animals cannot be reported. Behavioral experiments were conducted using the wild-type line Tubingen (Tu) between 5 and 7 days post fertilization (dpf). Larvae were fed with rotifers (*Brachionus sp.*) from 5 dpf onwards. All experimental procedures were approved by the Champalimaud Foundation Ethics Committee and the Portuguese Direcção Geral Veterinária, and were performed according to the European Directive 2010/63/EU.

SETUP

Larvae swimming behavior was recorded in an array of 20 wells, with each larva placed in an individual circular acrylic arena (44 mm diameter, 4 mm depth). Fish were imaged using an EoSens 4CXP Camera (MC4086, Mikrotron) operating at 400 fps. The frame grabber used was a Silicon Software GmbH AQ8 CXP6D. Fish were illuminated by an infrared array of 64 LEDs at 850 nm (TSHG6400, Vishay Semiconductor), controlled by a T-Cube LED Driver (LEDD1B, Thorlabs), with spacing sufficient to cover the full field of view. Imaging was performed using a fixed focal length lens (Xenoplan 2.0/28, Schneider Optische Werke) and an infrared long-pass filter (LP780-37, 780 nm, VisionLightTech).

For visual stimulation, a video projector (ML750e, Optoma) and a cold mirror (64-452, Edmund Optics) projected images onto a diffuser screen (three layers of Rosco Cinegel White Diffuser #3000) positioned 5 mm below the larva.

QUANTIFICATION OF BEHAVIOR

Larval position and orientation were tracked in real time using custom C# software. From these tracking datasets, we used the Megabouts Python package. Using the *trajectory-tracking* pipeline, we segmented the trajectory into individual tail bout movements and classified each tail bout into 11 categories, excluding prey-capture movements.

STIMULI PRESENTATION

- Approaching Dot: A black disk (1 mm radius) started 2 cm from the fish, approached at 5 mm/s, and remained beneath the fish for 1 second, moving perpendicular to the fish's head vector ($\pm 90°$).
- Directional Optomotor Response: A moving grating (10 mm period) moved at 10 mm/s at angles ranging from 0° to 315° relative to the fish's heading. It remained stationary for 5 seconds before moving for 10 seconds.
- Light/Dark Transition: The projector switched from 1000 lux to 0 for 30 seconds.

Stimuli were displayed at 60 fps using a custom OpenTK/OpenGL rendering engine. Stimuli were presented in pseudo-randomized blocks over a 2-hours experiment.

DRUG TREATMENT

Fish were treated with nine neuroactive drugs:

- Ketanserin (serotonin 5-HT2 antagonist, 8.5 µM)
- Quipazine (serotonin 5-HT$2/3$ agonist, 25 µM)
- Trazodone (serotonin antagonist and reuptake inhibitor, 25 µM)
- Quinpirole (selective D_2/D_3 receptor agonist, 25 µM)
- Clozapine (atypical antipsychotic, 8.5 µM)
- Fluoxetine (selective serotonin reuptake inhibitor (SSRI), 8.5 µM)

- Haloperidol (D_2 receptor antagonist, 8.5 µM)
- Apomorphine (non-selective dopamine agonist, 25 µM)
- Valproic acid (GABAergic voltage-gated sodium channel blocker, 25 µM)

Stock solutions were prepared by dissolving the compounds in autoclaved Milli-Q water. For drugs with low water solubility, an initial solution in 10% (v/v) DMSO was prepared, ensuring that the final DMSO concentration in the stock did not exceed 0.3%, a threshold below which no behavioral changes have been reported. Drug stocks were stored at -20°C after confirming stability via UV-Vis spectroscopy.

To prepare the arena solution, the appropriate stock volume was diluted in 600 mM Tris-buffered E3 medium to a final volume of 5 mL. All arena solutions had a final pH of 7 ± 0.3. The concentration for each drug was set to half of the maximum non-lethal dose. Larvae were allowed to habituate to the arena and drug solution for 1 hour before data collection. Each drug and concentration was tested on 18 fish.

Drug Classification Analysis

We applied the Mini-Rocket classifier to three different time-series: binary movement indicators, locomotion speed, and a multivariate one-hot encoded time series of movement categories. We employed a 10-fold stratified cross-validation. Each dataset was transformed using MiniRocket-Multivariate with 5000 kernels. The transformed data was then classified using a ridge classifier.

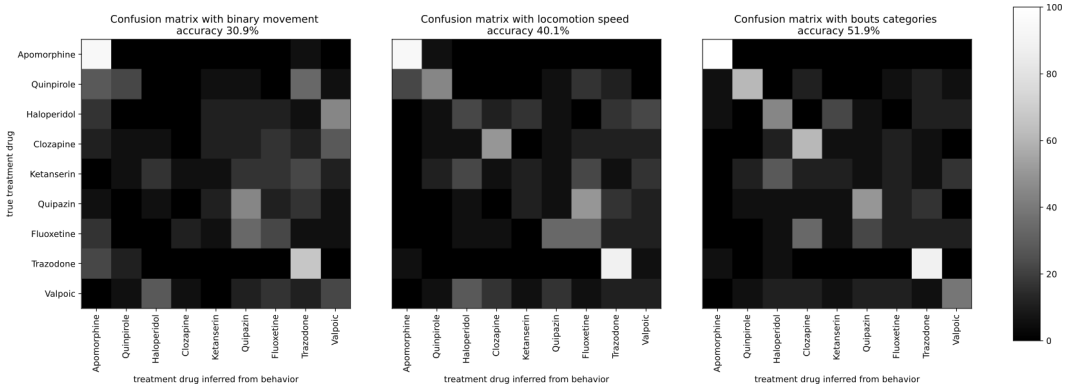

Figure 5: Confusion matrix displaying the classification performance of the Mini-Rocket model across different quantification of behavior.