Simultaneous classification of neuroactive compounds in zebrafish.

Douglas Myers-Turnbull[1,2], Jack C Taylor[2], Cole Helsell[1,2], Tia A Tummino[1,2,3], Matthew N McCarroll[2], Rebekah Alexander[2,4], Chris S Ki[2], Leo Gendelev[1,2], David Kokel[2,5]

[1]Quantitative Biosciences Consortium, University of California, San Francisco, California 94143, USA
[2]Institute for Neurodegenerative Diseases, University of California, San Francisco, California 94143, USA
[3]Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94158, USA
[4]Graduate School of Arts and Sciences, Georgetown University, Washington (DC) 20057, USA
[5]Department of Physiology, University of California, San Francisco, California 94158, USA

Abstract

Purpose: Compounds that act on the central nervous system (CNS) are crucial tools in drug discovery and neuroscience. To discover compounds with novel mechanisms of action, researchers have developed behavioral screens in larval zebrafish including various methods to identify and classify hit compounds. However, these methods typically do not admit intuitive numerical scores of screen performance. This study describes methods to classify compounds simultaneously in zebrafish and quantify screen performance.

Methods: We collected randomized, highly replicated data for two sets of compounds: 16 quality–control (QC) compounds and a reference set of 648 known CNS ligands. Machine learning models were trained to discriminate between compound-induced phenotypes, compare performance between protocols, and detect hit compounds.

Results: Classification accuracy on the QC set was 94.3%. In addition, 106 of 648 CNS ligands were identified as phenotypically active, and hits were enriched for dopaminergic and serotonergic targets. The raw data is included to facilitate replication and data mining.

Significance: This study describes methods to evaluate behavioral phenotyping assays, which can be used to facilitate comparison and standardization of data within the zebrafish phenotyping community.

Introduction

Disorders of the CNS affect 100 million Americans at an economic burden of $920 billion per year.[10] Despite this, CNS drug discovery rates have declined.[20] Most projects screen for high-affinity interaction with a single biological target.[26] Although extremely high-throughput, they require knowledge of disease pathogenesis to choose appropriate therapeutic targets. This knowledge is especially limited for CNS disorders.[1,32] Although most discovery projects are target-first, most first-in-class drugs approved by the FDA from 1999–2008 were discovered by phenotypic screening,[46] suggesting that many CNS drug discovery projects would benefit from phenotype-first screens.

In contrast to target-based screens, phenotypic screens require less understanding of pathogenesis and can identify compounds with previously unknown or multitarget pharmacologies. In many historical cases, a drug was discovered first, and its mechanism only later.[11,18] For example, the antidepressant activities of tricyclics and monoamine oxidase inhibitors were discovered in psychiatric hospitals by observing patients, and these discoveries implicated serotonin in depression and lent to the development of selective serotonin reuptake inhibitors.[36] Such phenomenological discoveries are responsible for most prototypical neuroactive drugs. Scaling this idea using animal models has yielded a powerful new approach to CNS drug discovery.

Zebrafish larvae and embryos have long been used to assay environmental toxicants.[2,3,8,30] They have also made waves in neuroscience as models for vision,[7,9,21,37,49] threat response,[38] memory,[52] algesia,[6,14,44] and sleep.[35,39,43]

In a rare example of bench-to-bedside, the FDA approved the drug lorcaserin as an antiepileptic, based mostly on evidence in zebrafish.[13] More recently, a zebrafish model was used in the life-saving treatment of a 12-year-old patient.[25] Genetic and compound-induced disease models in zebrafish larvae have shown promising consistency with rodent models, even for complex diseases like ALS.[29,42]

Zebrafish are also well-suited for *phenotypic profiling*, a quantitative, high-throughput approach to phenotype-first compound discovery.[34,39] Phenotypic profiles are quantitative readouts of the aggregate movement of many larvae on multiwell plates. Previous screens have identified new neuroactive compounds and predicted their targets, later supported by activity assays or knockout models.[16,23,24,28,39,45] Diverse compounds have been identified, including a photoactivatable TRPA1 ligand (optovin),[22] antiepileptics,[4] antipsychotics,[5] appetite modulators,[19] and anesthetic-like compounds.[27,53]

One way to understand how novel compounds are working is by association with a compound of known pharmacology, a *guilt-by-association* approach. These approaches link novel compounds to known ligands, but they require both *landmark* profiles for compounds with known pharmacology and a method to identify similar phenotypes. Such methods have been developed and applied in several of the aforementioned studies (Section S I.1). However, few have quantified the confidence and accuracy of the phenotypic associations. Such metrics would enable comparisons of performance, which could be used to optimize hardware, stimulus batteries, and computational methods.

Here, we describe approaches and metrics for quantifying the performance of behavioral screens using supervised machine learning. First, we assess performance on a set of 16 quality–control (QC) compounds, using this QC set as a benchmark to evaluate performance. Second, we describe a high-replicate set of landmark (reference) profiles for 648 known CNS ligands. We have made the data available as a resource that may be helpful in other studies.

Results

Screening instrument and pipeline.

We sought to develop a screening system to record movement behaviors in larval zebrafish in multiwell plates. We defined 5 criteria: high sensitivity to movement behaviors, support for continuous operation, sensor data and metadata for diagnostics; complete reproducibility of analyses; and extensibility to add or remove hardware components. For this, we modified an existing platform.[5]

The setup is shown in Figure 1a. Plates are positioned on a flat translucent acrylic stage, fixed in a shallow groove so that the sides and bottom of the plate contact the stage uniformly. The plates are then illuminated from the bottom with 760 nm infrared light through an acrylic diffuser and recorded with an overhead camera. Stimulus generators and low-distortion optics are used to perturb and record zebrafish locomotor activity (Figure 1a). The digital camera is mounted to a telecentric lens with an infrared filter. It captures 2 MP images at a preset frame rate of 100 Hz to 150 Hz. Nanosecond-resolved timestamps corresponding to the image sensor acquisition for each frame are recorded for precise synchronization with stimuli.

Light-based stimuli are delivered to the animals through 6 high-power LED arrays mounted overhead. Arbitrary acoustic stimuli are delivered through transducers mounted on the stage. Two push—pull solenoids are used to deliver secondary acoustic stimuli by forcefully tapping the stage surface. LED intensity and the force delivered by the solenoids are controlled by pulse-width modulation (PWM). A microphone, thermosensor, photosensor, and secondary camera verify the delivery of stimuli.

We use a 4-step workflow (Figure 1b). Animals, typically week-old, are anesthetized in cold egg water and dispensed into the wells of a multiwell plate (8 animals per well), dosed with compounds, and incubated for approximately 1 hr). The plates are then positioned in the instrument, and the animals are acclimated in darkness for 5 minutes. After this period, a battery of stimuli is applied. The video is then partitioned into a region of interest (ROI) for each well, and a simple time-series feature (*motion-trace*) is calculated per well to approximate the aggregate locomotor activity over time (Files 1 and 2). Figure 1c shows an example of motion-traces under a standard battery of stimuli for the atypical antipsychotic clozapine or vehicle (solvent).

The hardware is driven by cross-platform custom software that provides modes for running experiments, prototyping assays, and managing data. Post-processing of data is decoupled from capture, allowing many plates to be run without interruption. After a run completes, the videos are compressed and archived permanently, and data is inserted into a centrally located relational database. The database (MySQL) incorporates both course-grained information such as experimental purposes and low-grained information such as raw features, sensor readouts and metadata, compound batch information, and curated cheminformatics data. An accompanying website is used to design plate layouts, stimulus batteries, and experiments.

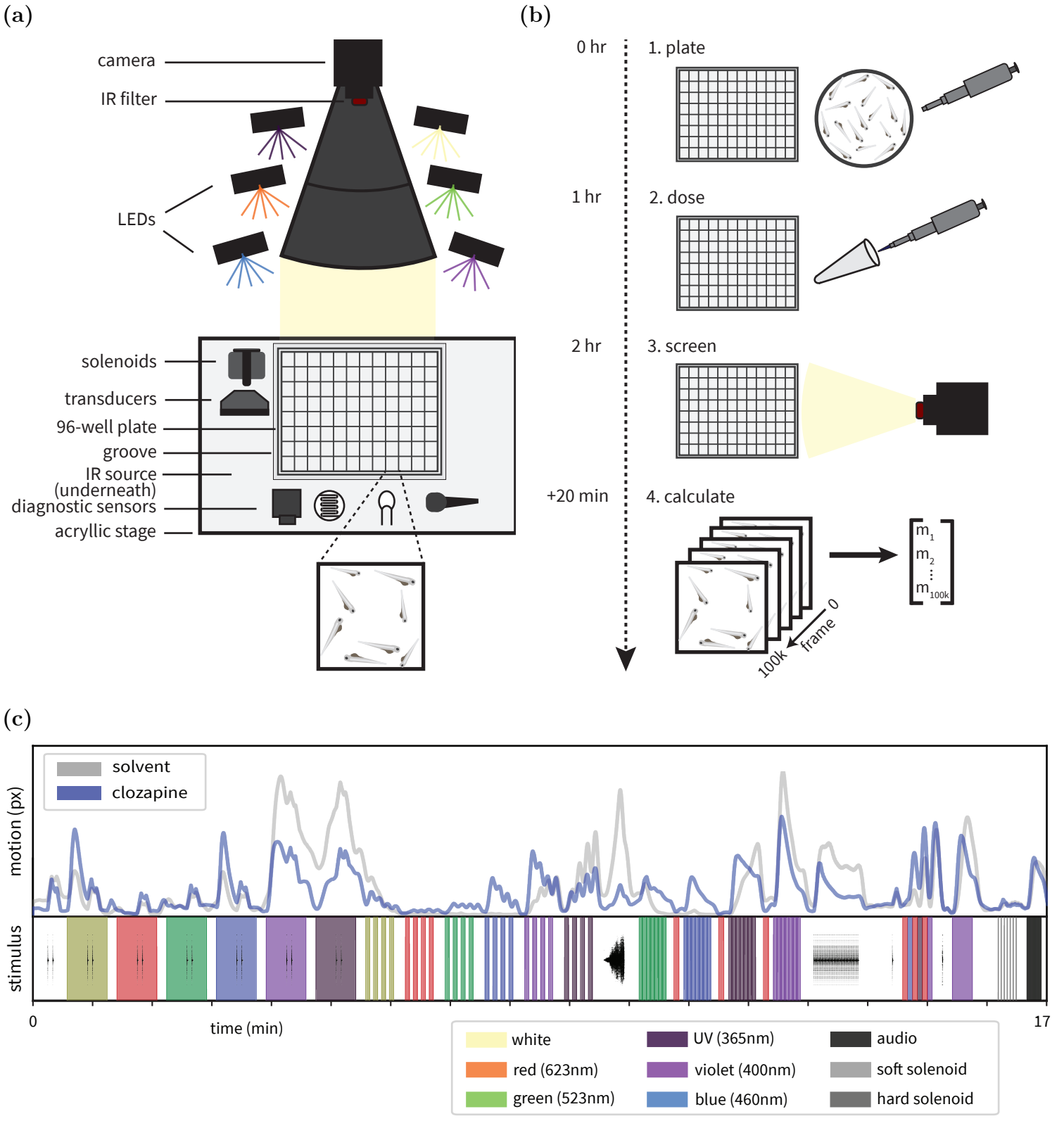

Figure 1: Overview of methods. **a:** Schematic of phenotypic profiling instrument. **b:** Experimental pipeline. **c:** Example motion-trace for wells treated with solvent (DMSO) or clozapine at 50 µM. Top: motion within the well as a function of time in the experiment, smoothed from 100 Hz to 10 Hz. Bottom: stimuli applied over time. The shaded colors represent high-intensity LEDs application; the black lines depict the waveforms of audio assays; and the gray lines at the end denote the delivery of acoustic stimuli by solenoids. (n = 12 wells/condition).

Although the hardware is larger than most commercial phenotyping systems at $61 \times 61 \times 114$ cm (Figure S1), the large size simplified construction and maintenance and enabled rapid iterations between collecting data, identifying phenotypes, and adapting hardware to capture them. We have built multiple machines to provide highly comparable data, which drove the need for quality–control (QC) protocols that would gauge phenotypic consistency.

Accurate discrimination of 16 QC treatments.

We wanted to evaluate the performance of the platform in phenotype-agnostic screens. Two criteria were defined: ability to identify phenotypically active compounds (*criterion (1)*), and ability to distinguish different compounds (*criterion (2)*). The first dictates the sensitivity to detect hits relative to controls, while the second reflects phenotypic resolution and is crucial for distinguishing between phenotypes and predicting mechanisms of action (MOAs). A good platform should meet both criteria.

As a first step toward evaluating system performance, we curated a set of 14 compounds that were structurally and mechanistically diverse and that appeared phenotypically active in prior data. An '*ideal*' (Section S R.2.4), non-lethal in-well concentration was fixed for each compound. Vehicle (solvent) control and lethal control were included (Table 1). The lethal control used a high dose of the anesthetic eugenol. These 16 treatments (14 compounds + 2 controls) formed the core *ideal-dose* QC set. The first experiment generated 54 wells per treatment across 9 plates and 3 weeks. All experiments were run 1 hr using the standard battery (Figure 1c).

For a preliminary analysis, we computed correlation distance between (single-well) motion-traces and visualized the results by t-distributed stochastic neighbor embedding (t-SNE). Each compound generated a cloud of replicate profiles generally separate from the controls and other compounds (Figure 2a). This motivated applying a method that could more robustly separate the treatments and provide a numerical metric of separation.

To get numerical metrics, machine learning classifiers were trained to label treatments from motion vectors. Random Forests (RF) were chosen because they perform and generalize well even without optimizing hyperparameters. RF is an ensemble model that averages the results of decision trees, each trained only using a subset of the data. RF classifiers can discover and combine informative features.As features, we used the ~100,000 elements of the motion vectors. We trained two types of classifiers (Figure 2b): first between solvent controls and compounds individually (*vs-solvent* or *vs-lethal*); second to distinguish between all treatments simultaneously (*vs-all*).

compound	conc. (µM)	primary MoA
almorexant	90	OX_1, OX_2 antagonist
bromocriptine	16	D_2R, D_3R agnoist
clozapine	50	D_2R, 5-HT_{2A} antagonist
donepezil	16	AChE inhibitor
endosulfan	0.32	$GABA_AR$ antagonist
etomidate	6.25	$GABA_AR$ agonist
haloperidol	25	D_2R antagnoist
indoxacarb	6.25	NaV inhibitor
(S)+ketamine	100	NMDAR antagonist
lidocaine	1200	NaV inhibitor
optovin	6.25	TRPA1 opener
(+)-sertraline	25	SERT inhibitor
tiagabine	100	GAT inhibitor
tracazolate	25	$GABA_AR$ modulator

Table 1: Quality–control compounds, their optimal concentrations in in 96-well plates (300 µL volume), and their primary mechanisms. Also see Table S1.

To quantify phenotypic strength, we trained vs-solvent models individually per compound. For each, 12 repeat models were trained, each assigned a balanced subsample of wells; this allowed us to compute the variance of accuracy across wells. As a negative control experiment, solvent–solvent models were trained by randomly false-labeling half of the solvent wells as 'a' and the other half as 'b'. This accuracy was 49%, reflecting no discrimination. Accuracy was 93% for lethal–solvent and 70–95% for all 14 compounds (Figure 2c). The accuracies were especially high for compounds that induced strong responses in the motion-traces: optovin, a violet/UV-light-induced transient receptor potential channel A1 (TRPA1) opener; endosulfan, a highly toxic $GABA_A$ ionotropic receptor ($GABA_AR$) antagonist and seizurogenic; and etomidate, a $GABA_AR$ agonist that induces a distinctive acoustic startle response[27] (Figure S2). These data supported an ability to distinguish active compounds (*criterion (1)*).

To evaluate the ability to distinguish compounds by their phenotypes (*criterion (2)*), we trained vs-all models and visualized the data as a confusion matrix (Figure 2d), which summarized how treatments were classified. The diagonal was high, reflecting accurate self-classification and phenotypic uniqueness, with a mean accuracy of 94.33%. Optovin and etomidate were the most accurately classified, and these compounds had distinct phenotypes (Figure S2). A prior experiment had similar results and showed pheno-

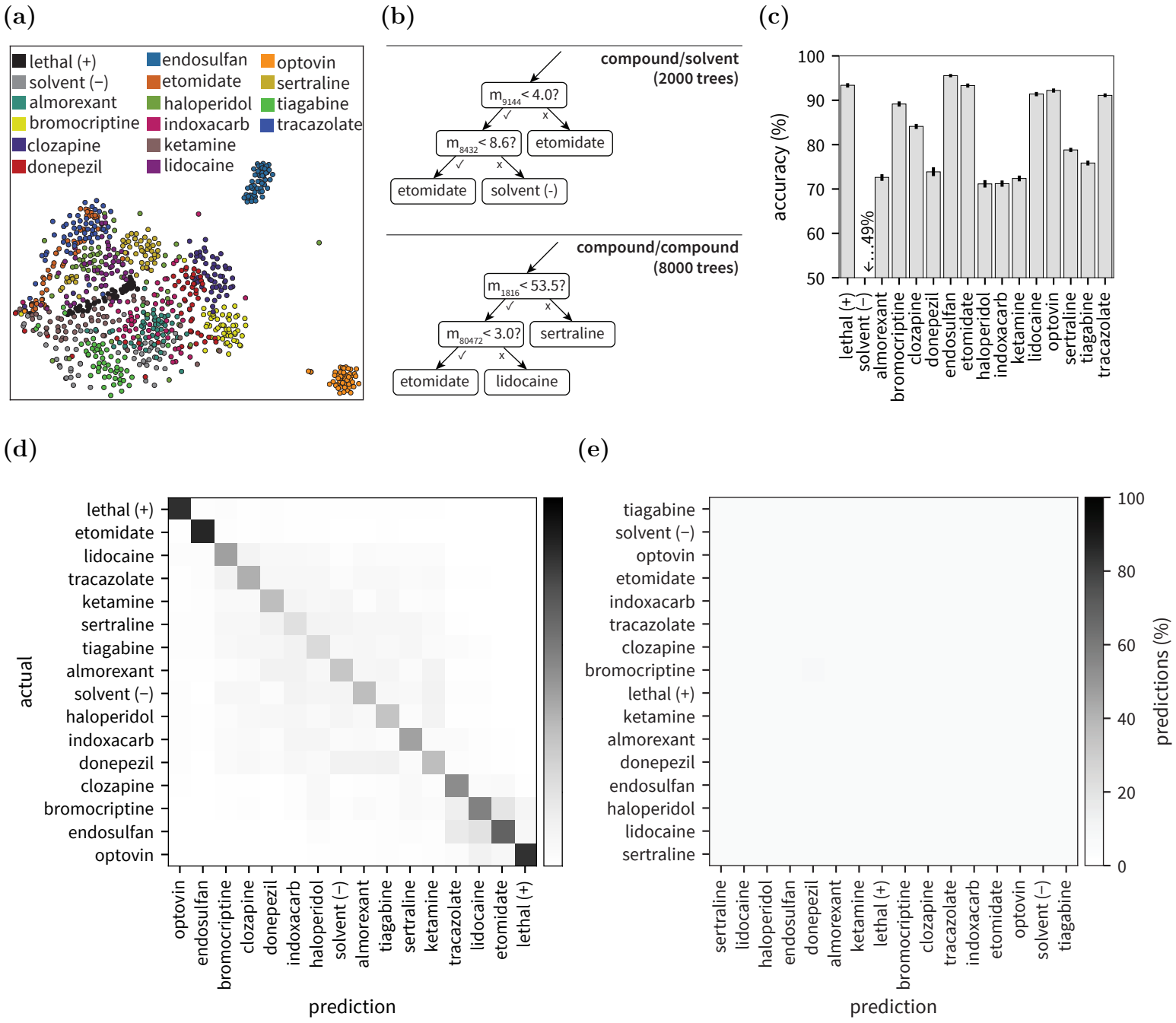

Figure 2: Results for ideal-dose QC experiments (n = 54 wells/condition). **a:** t-SNE projection of correlation distances between traces. Each point denotes one well. **b:** Illustration of RF applied to motion-traces. Both panels show an arbitrarily selected part of a decision tree for vs-solvent (*top*) and vs-all classification (*bottom*). m_i denotes the ith element of a trace corresponding to the ith video frame. **c:** Mean vs-solvent accuracy from RF. Error bars show a 90th-percentile confidence interval computed by repeat training on subsampled wells (n = 27+27=54 wells/model). **d:** Confusion matrix from a multiclass classification model on QC treatments. The mean accuracy was 94.33%. **e:** Confusion matrix from a multiclass model trained on false-labeled untreated wells (n = 18 wells/false-treatment).

typic similarity between etomidate and another $GABA_AR$ agonist, thiopental (Figure S8). For a second control experiment, we collected a dataset of only solvent-treated wells, false-labeled them to mimic the real dataset, and trained classifiers. Classifiers were unable to distinguish the false-labeled treatments (Figure 2e). Together, these data suggested resolution supporting *criterion (2)*.

To understand how phenotypes change with compound dose, we sought to assay complete concentration ranges, ranging from phenotypic inactivity to lethality. For each compound, we manually selected a logarithmic 5-point concentration range that was designed to range from 'just slightly active' to 'almost lethal' , estimated using earlier data (not shown). To control for plate and positional confounding, we plated controls and all 14 QC compounds at all 5 concentrations on each of 13 randomized plates, generating 13 replicates of each compound–concentration pair.

We plotted concentration–response curves, where the response was the classification accuracy. Due to the high dimensionality, such curves are not always sigmoidal or even monotone increasing. For most compounds, vs-solvent accuracy increased with concentration, while vs-lethal accuracy dropped sharply at high concentrations (Figures 3 and S11). Some compounds lacked these trends contrary to hypotheses *(1)* and *(2)*. This could be due to an insufficient sample size, failure of the models, or complex pharmacology. With qualifications, the data indicated that vs-solvent classification accuracy estimated phenotypic strength and that low vs-lethal accuracy signaled lethality.

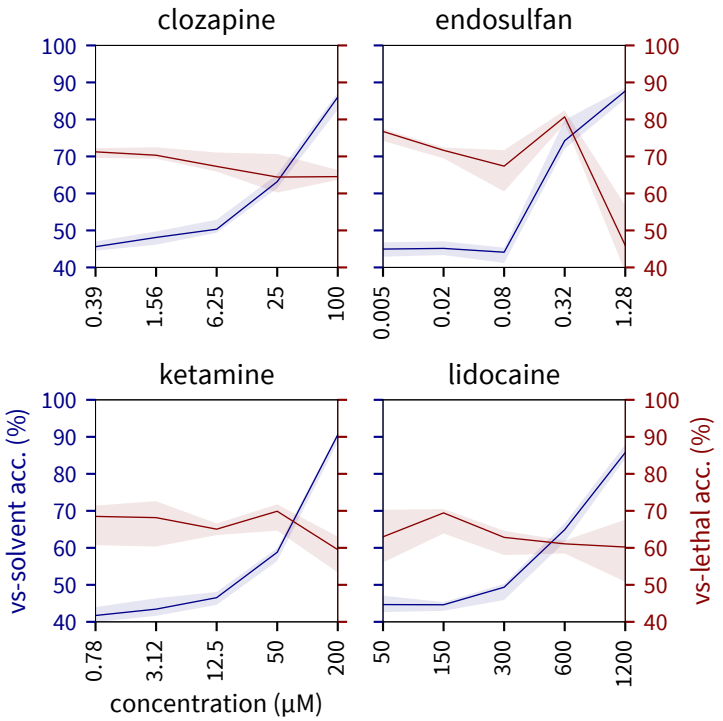

Figure 3: Concentration–response curves for vs-solvent (left axis; blue) and vs-lethal (right axis; red) accuracy. Opaque lines denote the median accuracy, and shaded regions denote a 95th percentile confidence interval by bootstrap. (n = 8 wells/condition).

Using the QC set to evaluate protocols.

In a second experiment, we explored the idea of using the QC set to compare and optimize experimental protocols. We first considered the number of animals per well. Plates were collected with 2, 3, 4, 6, 8, or 10 animals per well. Vs-all classification accuracy increased near-monotonically with more animals (Figure 4a). This trend could be explained by the increased total number of animals per condition or by a decrease in inter-well variance. Although accuracy was highest for 10 / well, this decreased survival prior to the run. We concluded that this approach could be applied to optimize other experimental parameters or assess the impact of potentially confounding variables.

We then considered using the QC set to optimize behavioral assays. To test this and better understand the classifier, we used the 9-plate QC data and analyzed the random forest vs-all weights by frame. The most heavily weighted frames occurred at the beginning or end of stimulus (Figure 5

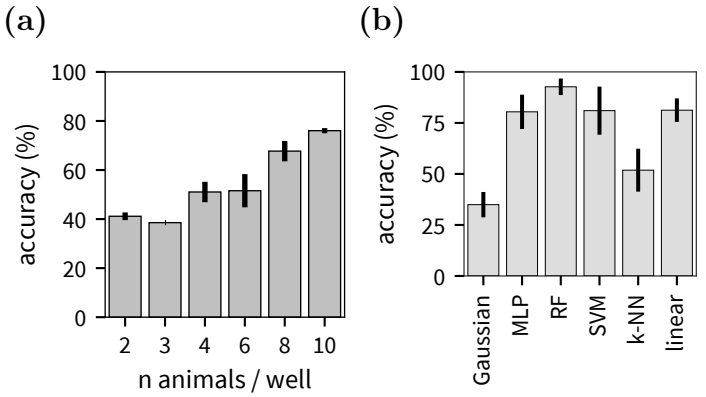

Figure 4: Example uses of the QC set to evaluate methods. **a:** X axis: number of animals per well (n = 2 plates, 12 wells/treatment). Y axis: Mean out-of-bag vs-all accuracy trained independently per plate (%). Error bars represent the values for the 2 plates. **b:** Mean vs-all accuracy by algorithm. Error bars show standard deviation over a set of hyperparameters.

and File 7), indicating that the stimuli were important. The standard battery was created using a similar approach using QC data under a large set of batteries (not shown). These approach could be used for other optimizations.

Scaling to 648 CNS ligands.

To predict MoAs for novel compounds, we sought to identify landmark profiles of compounds with known pharmacologies covering major compound classes. As an auxiliary goal, these would enable exploratory analyses of compound classes and phenotypes.

To achieve these goals, we screened the SCREEN-WELL Neurotransmitter library (Enzo Life Sciences), which contained 648 structurally and functionally diverse ligands acting on diverse targets. These were broadly grouped into 13 classes: adenosinergic, adrenergic, cholinergic, dopaminergic, GABAergic, glutamatergic (ionotropic), glutamatergic (metabotropic), histaminergic, melatonergic, opioidergic, purinergic, serotonergic, and σ (Figure 6a). We randomized treatments across plates and wells along with 8 solvent controls and 3 lethal controls per plate. We then ran approximately 7 replicates per compound at 33.3 µM per well. Deviation above or below 7 is due to a filtration step.

Vs-control classifiers were trained using methods similar to those for the QC set (Methods: Classification). We visualized the distribution of accuracies independently for treatment–solvent, solvent–solvent, and lethal–solvent comparisons (Figure 6c). Treatment–solvent accuracies were higher on average than solvent–solvent, which were centered near 50%. Noticeably, many treatment–solvent but no solvent–solvent were above 75%. Lethal–solvent comparisons were high on average, but the distribution was highly bimodal, with peaks near 78% and 95% (Figure S13). This could be explained by a sub-lethal dose of eugenol in some treatments or a confounding variable affecting water motion induced by acoustic assays. To call hits, we applied a threshold of accuracy that excluded all but 0.5% of solvent–solvent comparisons. This was 63%, yielding 106 hit compounds (Figure 6c) for a hit rate of 16.3%.

We examined the top 15 hits manually, which all had accuracy over 94% (Table 2, Figure S14, and File 10). The strongest, THDOC, was lethal. Despite diverse annotated mechanisms, 11 of the following 12 hits had very similar phenotypes in which light response was ablated but acoustic responses were preserved or only partly diminished. This included ivermectin, a toxic $GABA_AR$ antagonist and pesticide that was included in the QC set; and propofol, a $GABA_AR$ agonist and that causes sedation, paradoxical excitation, and a stereotyped acoustic startle response in zebrafish.[27] However, the remaining compounds, BRL-15,572 and ICI 199,441 had different phenotypes.

name	ChEMBL ID	acc.	primary MoA
THDOC**	1256760	98.2	$GABA_A$ ag.
DH 97	1327247	97.6	MT ant.
Riluzole	744	97.2	NaV inh.
Vanoxerine	281594	97.2	DAT inh.
Propofol	526	97.1	$GABA_A$ pot.
L-741,742	444309	97.1	D_4 ant.
Brexanolone	207538	97.0	$GABA_A$ pot.
GBR 13069	286991	96.9	DAT inh.
GBR 12935	26320	96.8	DAT inh.
CGS 12066B	27403	96.7	5-HT_1 ag.
Ivermectin	3349014	95.9	$GABA_A$ ant.
Naftopidil	142635	95.4	α_1-AR ant.
Butaclamol	8514	94.9	DAT inh.
BRL-15,572	534232	94.7	5-HT_{1D} ant.
ICI 199,441	320882	94.1	κ-opioid ag.

Table 2: Top 15 hits from NT-650 with their primary mechanism of action and vs-solvent accuracy (%). Abbreviations: *ag.*—agonist; *ant.*—antagonist; *inh.*—inhibitor; *pot.*—potentiator. ** THDOC was lethal.

We were interested to know which types of compounds induced the strongest phenotypes. To start, compounds were grouped to compare hit rates per class (Figure 6d). Dopaminergic and serotonergic targets were highly enriched and adenosinergic, purinergic, and glutamatergic depleted, though every class had at least one hit.

The compound classes did not distinguish between specific targets, so we increased the granularity of the annotations. Three independent sources were used to annotate compound mechanisms: mechanisms derived from those provided by the supplier, mechanisms from ChEMBL, and binding from ChEMBL (Methods: Annotations). Data were visualized in the same manner as before (Figures 6e and 6f and Table S4). Integrating the data, the dopamine, serotonin, and norepinephrine transporters (DAT, SERT, NET); most dopamine and serotonin receptors; the σ receptor; histamine receptor 1 (H_1); and metabotropic acetylcholine receptor (mAChR) were enriched.

We then aimed to understand the phenotypes in the dataset, as well as potential associations between targets and phenotypes. A vs-all classifier was trained to discriminate between the 106 hit compounds, and the results were vi-

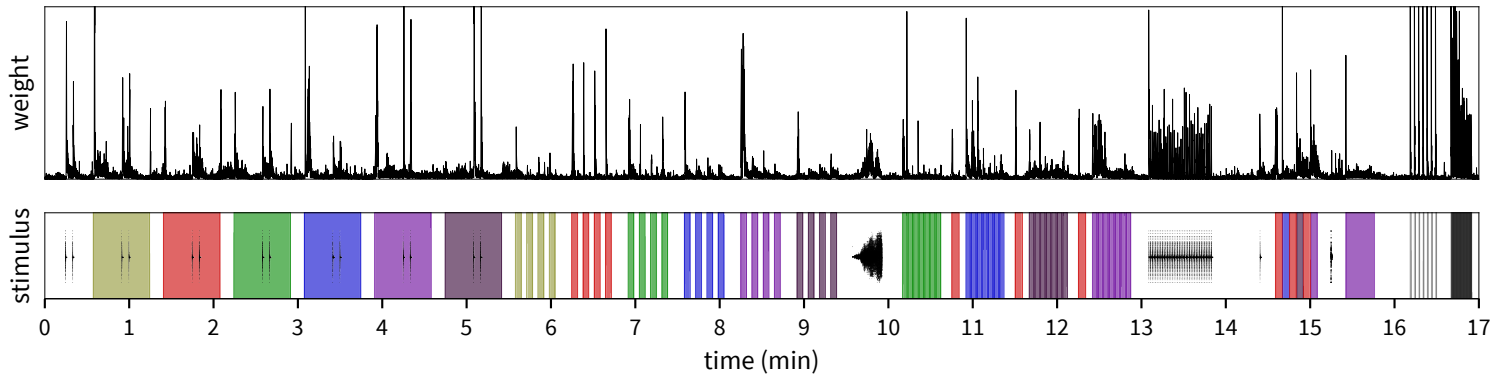

Figure 5: Feature weights by frame for vs-all ideal-dose comparisons.

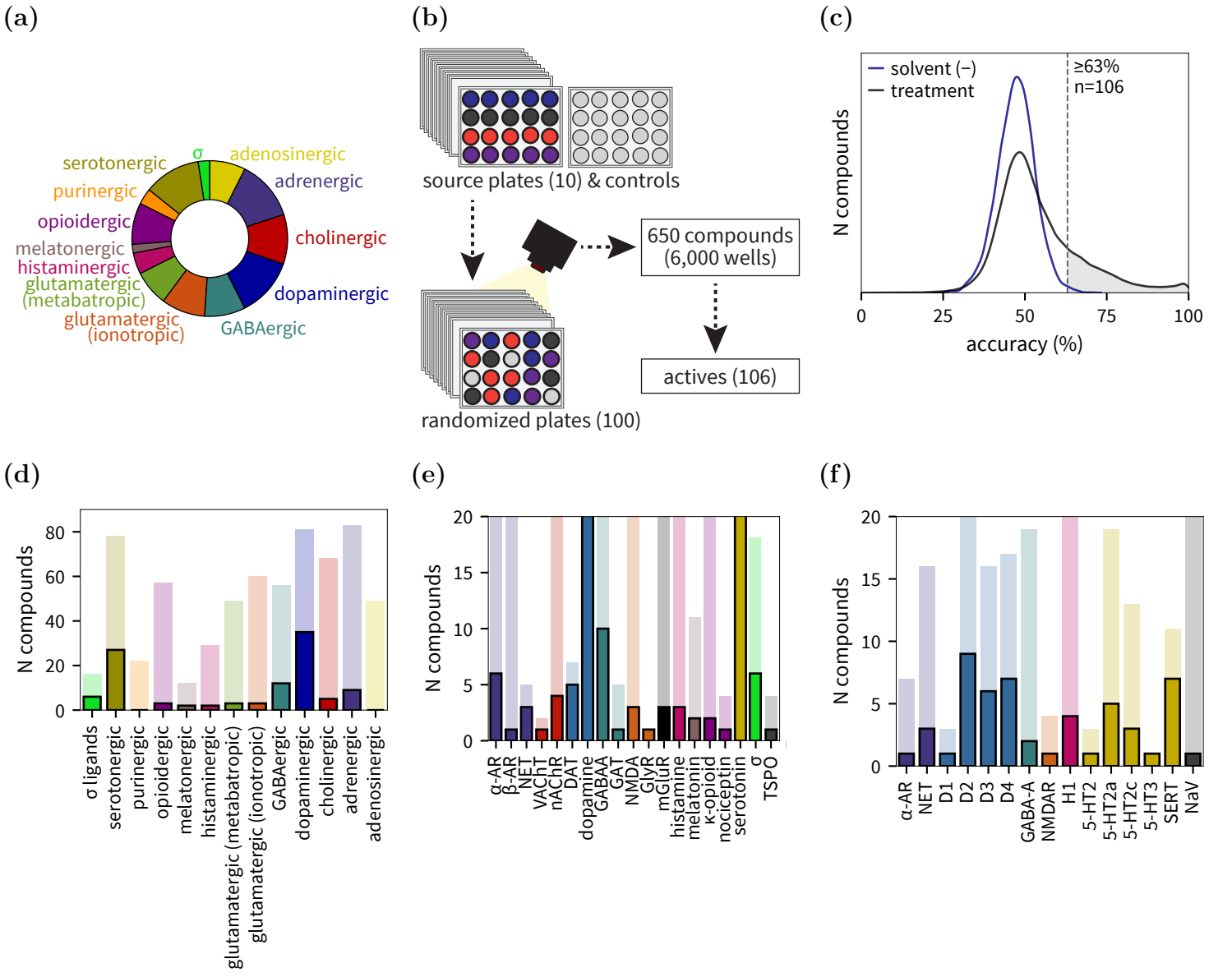

Figure 6: NT-650 screen and initial results. **a:** Distribution of compounds among the 13 classes. Arc length is proportional to number of compounds. **b:** Experimental design. 10 supplier-provided source plates were used to build a stack of 100 randomized daughter plates, run at $\tilde{7}$ replicates / well. Treatment–solvent classifiers identified 106 phenotypically active compounds. **c:** Distribution of vs-solvent accuracies for treatments (blue) and solvent (black). The threshold for determining hits is shown at x=63%. (n = 648 treatments, 200 solvent–solvent models, 200 lethal–solvent models). **d–f:** Distribution of hits (opaque) and total compounds (translucent) per class (**d**), supplier-provided target (**e:**), and ChEMBL-provided target (**f**). Colors correspond to those in **a**.

sualized as a lower-triangular affinity matrix to show the phenotypic similarity between compounds (Section 5.7). We sorted the compounds by class, then target, and then randomly (Figure 7).

The matrix had a strong diagonal, indicating that compounds were phenotypically coherent (self-similar). We also noticed that confusion with solvent was low, likely because compounds were restricted to hits. We also observed a structured, nonuniform distribution of off-diagonal elements. Several associations between targets and phenotypic clusters were visible, including for $GABA_AR$, dopamine transporter (DAT), N-methyl-D-aspartate receptor (NMDAR), metabotropic glutamate receptor (mGluR), and dopamine and melatonin receptor ligands. A number of unique corresponding phenotypes were visible in the motion-traces (File 10). These data suggested that at least 16% of compounds were phenotypically active and that many were distinguishable from others.

Discussion

Previous studies illustrate the power of phenotypic and behavioral profiling to discover and characterize neuroactive compounds. Different hardware, stimuli, zebrafish strains, and computational methods have been applied, suggesting utility in comparing performance. Here, we found that classification accuracy in a quality–control set provided a flexible and intuitive metric to summarize the performance of the system. This approach has immediate practical applications, such as quantifying the severity of confounding variables and designing optimal stimuli.

By simultaneously classifying compounds and controlling for confounding variables, we provide a reliable lower bound for the performance of a behavioral profiling system, and we hope this will invite comparisons using the same approach or the development of superior or complimentary benchmarks. We also note that plate and well-positional confounding can significantly affect results. We previously conducted a non-randomized screen that showed very strong grouping by the compound class. However, the SCREEN-WELL Neurotransmitter library provides compounds already arranged in plates according to their mechanism of action, making it easy to see how such confounding could solely explain a promising positive result.

We note several caveats. First, we did not confirm that the phenotypes were caused by the expected mechanisms. It is plausible that the compounds acted through mechanisms unrelated to their mechanisms in humans.Several data types could increase confidence, including phenotypes under genetic knockout of a target, activity at a target in vitro, expression or proteome changes, cardiotoxicity assays, and neuroimaging.

The data from 648 CNS ligands suggest that the space of compound-induced movement behaviors is moderately diverse, and several changes could easily be made to increase this lower bound. First, we used a concentration of 33 µM, but the dose–response QC experiments indicated that some compounds were phenotypically inactive below 100. Second, affecting complex behavioral states such as aggression, addiction, social behavior, or learning may improve resolution. We used a trivial readout for potentially high-dimensional movement behaviors, but tracking,[48] optic flow,[40] probabilistic models,[17] and deep learning[15] have been successful in analyzing similar data. Future studies will likely leverage advances in all these areas to improve the resolution of zebrafish behavioral profiling assays.

Methods

Animal husbandry

Zebrafish husbandry was as described.[51] Embryos were from group matings of wild-type zebrafish from Singapore and were raised on a 14/10-hour light/dark cycle at 28 °C until 7 dpf. Zebrafish experiments were performed in accordance with established protocols approved by UCSF's Institutional Animal Care Use Committee (IACUC) and in accordance with the Guide for the Care and Use of Laboratory Animals.[31]

Compound treatments

Healthy larvae were sorted and then immobilized with cold egg water[50] with 25 mL of 4 °C added to 12 mL room-temperature egg water containing about 1,000 fish. The larvae were then distributed by pipette into 96-well plates (GE Healthcare) with 8 fish per well in 300 µL aliquots. Plates were then incubated at room temperature for 1 hr, at which animals were mobile. Compound plates and aliquots were stored at −20 °C, except for the NT-650 plates which were stored at −80 °C. A Biomek liquid handler (Beckman Coulter) was used for randomization.

For QC treatments (Table S1), 2 µL of solvent-dissolved compound was added to each well. Solvents were DMSO except for donepezil (water). For the n-fish experiment, compounds were transferred from 2 manually randomized plates, for 2 plates / condition.

Figure 7:
Lower-triangular affinity matrix of the 106 NT-650 hit compounds. Values range from the 2nd to 98th percentile (white to black). Labels are CHEMBL IDs[a]. Label colors correspond to the classes in Figure 6c. Compounds are sorted by class, then target, then a random value[b,c]. Target labels are colored arbitrarily. (n = 106 compounds)

[a] n-butyl-β-carboline-3-carboxylate (NBBCC) (ID (2557)) was not linked to ChEMBL; 2557 is the internal ID.
[b] Dibenzepin (CHEMBL 442422) was annotated for *serotonin*, *histamine*, and *NET*; and strychnine (33495) for *GlyR* and *nAChR*.
[c] Clozapine (CHEMBL 42) was annotated for the dopaminergic and serotonergic classes and for targets *serotonin* and *dopamine*.

For NT-650 data, we purchased the SCREEN-WELL Neurotransmitter library (Enzo Life Sciences) library in 2014 and stored it at −80 °C. 1 µL of solvent-dissolved compound was added to each well to achieve 33 µM/well, except for peptides at 0.33 µM/well. Plates contained 14 DMSO, 8 water, and 6 lethal eugenol controls.

Instrument

The camera is a PointGrey Grasshopper GS3-U3-41C6M-C (FLIR Integrated Imaging Solutions). An infrared filter was used (LE8744 polyester #87, LEE Filters). Six LED arrays were positioned overhead, with 4 LEDs per array: red at 623 nm (1537-1041-ND, DigiKey), green at 525 nm (1537-1039-ND, DigiKey), blue at 460 nm (1537-1037-ND, DigiKey), white at 4000 K (416-0D0BN240E-SB01, Mouser), violet/UV at 400 nm (LZ4-40UB00-00U7, Mouser), and UV at 355 nm (416-LST101G01UV01, Mouser). Two surface transducers were fastened on the stage near the sides of the plate (5 W transducer, Generic) and used with a 150 W amplifier (APA150, Dayton Audio). Two 36 V push–pull solenoids (SparkFun Electronics) were positioned near the top of the plate, one contacting the stage directly, and the other contacting a 1 mm-deep strip of synthetic felt.

An Arduino Mega 2560 rev 3 (Arduino.cc) drove the LEDs, solenoids, and small sensors, while a separate computer directly controlled the microphone, transducers, and cameras. The camera streamed raw data to a high-performance M.2 SSD (970 PRO 2280 1TB, Samsung), which was needed to avoid throttling acquisition. 1600 × 1068, 8-bit grayscale videos were then trimmed, compressed with High-Efficiency Video Encoding (HEVC) with Constant Quantization Parameter (CQP) 15 and partitioned into identically sized ROIs for wells (Figure S15).

Four sensors were used for diagnostics: A H2a Hydrophone (Aquarian Hydrophones) with a rubber contact adapter on the stage surface; an overhead Logitech 1080p C930e webcam (Logitech); and a photoresistor and thermistor (Tinker Kit, SparkFun Electronics) under the stage.

Motion estimation

Motion was estimated as the count of pixels that changed from the previous frame by intensity $\geq 10/255$. The threshold was chosen by comparing a histogram of pixel intensity changes in wells with 8 fish and wells without fish. Image sensor acquisition timestamps were used to align the frames with the stimuli (Equation (S2)).

Data collection

All data were collected at 100 fps under a standard battery (Figure 1c and File 3). **QC data:** The 16-treatment set was replicated across 15 plates, applying 6 replicates of the 14 compounds and 2 controls ($16 \times 6 = 96$). We randomly distributed compounds into wells of 14 96-well plates, each containing 6 replicate wells of all 16 treatments. 5/14 ideal-dose plates and 1/9 dose–response plates were excluded by filtering for probable errors using integrated diagnostic sensors. Concentration–response plates were similarly designed. **NT-650 data:** 80 plates were collected across 2 months (Figure S12). 13/80 plates were excluded based on diagnostic sensor readout. After, we filtered 23/7680 wells that had insufficient volume of compound in the daughter plate. These compounds had fewer replicate wells (File 9).

Classification

All models used scikit-learn 0.21.1[33] with hyperparameters default except for the number of trees. For the QC data, this was 10,000 for vs-solvent and 80,000 for vs-all. These numbers were 4,000 and 32,000 for NT-650.

In the NT-650 analysis, 4 replicate treatment–solvent classifiers were trained per compound. For each classifier, all replicate treatment wells were compared against the same number of randomly sampled solvent wells. The solvent wells were restricted to the plates containing the compound treatment, and to the solvent corresponding to the treatment (DMSO or water). One compound (amoxapine; CHEMBL1113) used N-methyl-2-pyrrolidone (NMP) as a solvent; this was compared to DMSO.

The accuracy cutoff for active NT-650 compounds (63%) was $\tau^* = \lceil \max\{\tau\} \times 10 \rceil / 10$ s.t. $\sum_{p \in P} \mathbf{1}[p \geq \tau] \leq 0.05/200$, where P is the set of accuracies of the 200 solvent–solvent models and $\mathbf{1}$ is the indicator function.

Visualization

QC data: t-SNE parameters were left as scikit-learn defaults. To sort the confusion matrix, we applied confusion matrix ordering (CMO)[47] (Equation (S4)) as implemented by the author at https://github.com/MartinThoma/clana. The colors were linear from 0% (white) to 100% (black). Concentration–response curves (Figure 3) were for the median response and the 5th and 95th percentile of the median among 1,000 bootstrap samples.

NT-650 data: An affinity matrix was calculated as the lower triangle of mean of the confusion matrix and its transpose ($L = \mathrm{ltr}(\frac{1}{2}C + \frac{1}{2}C^T)$). Sorting was done incrementally by class, target (from the supplier), and a random number.

Values ranged from the 2nd to 98th percentile (pure white to pure black). In Figure 6c, a Gaussian kernel density estimate (KDE) was calculated using statsmodels 0.10:[41] `kdensityfft(kernel=gau, bw=normal_reference`.

Target annotations

All annotations are listed in File 8. Compound classes were provided the supplier, which also provided text descriptions of compound MoAs, which we converted into <compound> <binding mode / action> <target> triples by arranging words, standardizing vocabulary, and splitting disjoint annotations into multiple triples. No new information was introduced. For Figure 6, predicates were ignored and objects were grouped into the labels shown. Compound names were simplified from a prioritized list of resources.

Compounds were linked to ChEMBL 25.0[12] by exact InChIKeys, plus 28 linked to nearly identical ChEMBL IDs manually and 8 with no matches. ChEMBL target activities were restricted to measurements of K_i, potency, IC_{50}, or EC_{50} between 10e−7 and 10e4 (Section S M.9).

Some ChEMBL target identifiers were for multiple protein targets, which were split into one per target. For example, each "GABA A receptor alpha-6/beta-2/gamma-2" (CHEMBL2111365) was replaced with 3 new annotations. 'D2-like' dopamine receptors was split into D_2R, D_3R, and D_4R. Target records with identical or equivalent names were merged; for example, 'serotonin receptor' and '5-HT receptor'. Target names were subsequently abbreviated or simplified in other trivial ways.

Author contributions

DMT conceived of the study, designed the experiments, performed the analyses, and wrote the paper. DMT and CH developed the hardware and hardware drivers, with earlier work by DK and refinements by CK. DMT developed the data pipeline and computational methods. JT performed the behavioral experiments and assisted experimental design. LG, DMT, and JT designed the NT-650 layouts and randomization. RA, DMT, and TT collected initial QC sets on earlier hardware. JT and MM managed the animal husbandry along with technicians. All authors read, provided feedback on, and approved the manuscript.

Acknowledgments

The authors would like to thank Louie Ramos, Veronica Manzo, and Vy Nguyen for animal husbandry; Capria Rinaldi for collecting behavioral data referenced when selecting initial dose ranges; Eric Lam for component fabrication assistance; and Steven Chen and Michelle Arkin for Biomek access and handling. Also thanks to Michael J Keiser, Jason Gestwicki, and John Kornak for providing feedback on the experimental design. Funding was provided by the National Institute on Alcohol Abuse and Alcoholism, the Allen Family foundation, the Genentech Fellowship Program, and NIH training grant 4T32GM6754714.

References

[1] Yves Agid et al. "How can drug discovery for psychiatric disorders be improved?" *Nat. Rev. Drug Discov.* (Mar. 2007). DOI: 10.1038/nrd2217.

[2] Muhammad T Akhtar et al. "Developmental effects of cannabinoids on zebrafish larvae". *Zebrafish* (Sept. 2013). DOI: 10.1089/zeb.2012.0785.

[3] Shaukat Ali, Harald G J van Mil, and Michael K Richardson. "Large-Scale Assessment of the Zebrafish Embryo as a Possible Predictive Model in Toxicity Testing". *PLOS One* (June 2011). DOI: 10.1371/journal.pone.0021076.

[4] Scott C Baraban, Matthew T Dinday, and Gabriela A Hortopan. "Drug screening in Scn1a zebrafish mutant identifies clemizole as a potential Dravet syndrome treatment". *Nat. Commun.* (2013). DOI: 10.1038/ncomms3410.

[5] Giancarlo Bruni et al. "Zebrafish behavioral profiling identifies multitarget antipsychotic-like compounds". *Nat. Chem. Biol.* (2016). DOI: 10.1038/nchembio.2097.

[6] Andrew Curtright et al. "Modeling nociception in zebrafish: a way forward for unbiased analgesic discovery". *PLOS One* (Jan. 2015). DOI: 10.1371/journal.pone.0116766.

[7] Conor Daly et al. "A Brain-Derived Neurotrophic Factor Mimetic Is Sufficient to Restore Cone Photoreceptor Visual Function in an Inherited Blindness Model". *Sci. Rep.* (Sept. 2017). DOI: 10.1038/s41598-017-11513-5.

[8] Sudhakar Deeti, Sean O'Farrell, and Breandán N Kennedy. "Early safety assessment of human oculotoxic drugs using the zebrafish visualmotor response". *J. Pharmacol. Toxicol. Methods* (Jan. 2014). DOI: 10.1016/j.vascn.2013.09.002.

[9] Florian A Dehmelt et al. "Spherical arena reveals optokinetic response tuning to stimulus location, size and frequency across entire visual field of larval zebrafish". *bioRxiv* (Sept. 2019). DOI: 10.1101/754408.

[10] Chris Delvecchio, Jens Tiefenbach, and Henry M Krause. "The zebrafish: a powerful platform for in vivo, HTS drug discovery". *Assay Drug Dev. Technol.* (Aug. 2011). DOI: 10.1089/adt.2010.0346.

[11] J Drews. "Drug discovery: a historical perspective". *Science* (Mar. 2000). DOI: 10.1126/science.287.5460.1960.

[12] Anna Gaulton et al. "The ChEMBL database in 2017". *Nucleic Acids Res.* (Jan. 2017). DOI: 10.1093/nar/gkw1074.

[13] Aliesha Griffin et al. "Clemizole and modulators of serotonin signalling suppress seizures in Dravet syndrome". *Brain* (Mar. 2017). DOI: 10.1093/brain/aww342.

[14] Martin Haesemeyer et al. "A Brain-wide Circuit Model of Heat-Evoked Swimming Behavior in Larval Zebrafish". *Neuron* (May 2018). DOI: 10.1016/j.neuron.2018.04.013.

[15] Omer Ishaq, Sajith Kecheril Sadanandan, and Carolina Wählby. "Deep Fish". *SLAS Discov* (Jan. 2017). DOI: 10.1177/1087057116667894.

[16] Nathalie Jeanray et al. "Phenotype classification of zebrafish embryos by supervised learning". *PLOS One* (Jan. 2015). DOI: 10.1371/journal.pone.0116989.

[17] Robert Evan Johnson et al. "Probabilistic Models of Larval Zebrafish Behavior Reveal Structure on Many Scales". *Curr. Biol.* (Dec. 2019). DOI: 10.1016/j.cub.2019.11.026.

[18] Alan Wayne Jones. "Early drug discovery and the rise of pharmaceutical chemistry". *Drug Test. Anal.* (June 2011). DOI: 10.1002/dta.301.

[19] Josua Jordi et al. "High-throughput screening for selective appetite modulators: A multibehavioral and translational drug discovery strategy". *Sci Adv* (Oct. 2018). DOI: 10.1126/sciadv.aav1966.

[20] Aaron S Kesselheim, Thomas J Hwang, and Jessica M Franklin. "Two decades of new drug development for central nervous system disorders". *Nat. Rev. Drug Discov.* (2015). DOI: 10.1038/nrd4793.

[21] Andreas M Kist and Ruben Portugues. "Optomotor Swimming in Larval Zebrafish Is Driven by Global Whole-Field Visual Motion and Local Light-Dark Transitions". *Cell Rep.* (Oct. 2019). DOI: 10.1016/j.celrep.2019.09.024.

[22] David Kokel et al. "Photochemical activation of TRPA1 channels in neurons and animals". *Nat. Chem. Biol.* (2013). DOI: 10.1038/nchembio.1183.

[23] D Kokel et al. "Rapid behavior-based identification of neuroactive small molecules in the zebrafish". *Nat. Chem. Biol.* (2010). DOI: 10.1038/nchembio.307.

[24] Christian Laggner et al. "Chemical informatics and target identification in a zebrafish phenotypic screen". *Nat. Chem. Biol.* (Dec. 2011). DOI: 10.1038/nchembio.732.

[25] Dong Li et al. "ARAF recurrent mutation causes central conducting lymphatic anomaly treatable with a MEK inhibitor". *Nat. Med.* (July 2019). DOI: 10.1038/s41591-019-0479-2.

[26] Mark A Lindsay. "Target discovery". *Nat. Rev. Drug Discov.* (Oct. 2003). DOI: 10.1038/nrd1202.

[27] Matthew N McCarroll et al. "Zebrafish behavioural profiling identifies GABA and serotonin receptor ligands related to sedation and paradoxical excitation". *Nat. Commun.* (Sept. 2019). DOI: 10.1038/s41467-019-11936-w.

[28] Olivier Mirat et al. "ZebraZoom: an automated program for high-throughput behavioral analysis and categorization". *Front. Neural Circuits* (2013). DOI: 10.3389/fncir.2013.00107.

[29] Jessica R Morrice, Cheryl Y Gregory-Evans, and Christopher A Shaw. "Modeling Environmentally-Induced Motor Neuron Degeneration in Zebrafish". *Sci. Rep.* (Mar. 2018). DOI: 10.1038/s41598-018-23018-w.

[30] Anjali K Nath et al. "Chemical and metabolomic screens identify novel biomarkers and antidotes for cyanide exposure". *FASEB J.* (May 2013). DOI: 10.1096/fj.12-225037.

[31] National Research Council (US) Committee for the Update of the Guide for the Care and Use of Laboratory Animals. *Guide for the Care and Use of Laboratory Animals*. 8th ed. Washington (DC): National Academies Press (US), May 2011. DOI: 10.17226/12910.

[32] Menelas N Pangalos, Lee E Schechter, and Orest Hurko. "Drug development for CNS disorders: strategies for balancing risk and reducing attrition". *Nat. Rev. Drug Discov.* (July 2007). DOI: 10.1038/nrd2094.

[33] Fabian Pedregosa. "Scikit-learn: Machine Learning in Python". *J. Mach. Learn. Res.* (2011).

[34] Randall T Peterson and Mark C Fishman. "Discovery and use of small molecules for probing biological processes in zebrafish". *Methods Cell Biol.* (2004). DOI: 10.1016/s0091-679x(04)76026-4.

[35] David A Prober et al. "Hypocretin/orexin overexpression induces an insomnia-like phenotype in zebrafish". *J. Neurosci.* (Dec. 2006). DOI: 10.1523/JNEUROSCI.4332-06.2006.

[36] Chaitra T Ramachandraih et al. "Antidepressants: From MAOIs to SSRIs and more". *Indian J. Psychiatry* (Apr. 2011). DOI: 10.4103/0019-5545.82567.

[37] Owen Randlett et al. "Distributed Plasticity Drives Visual Habituation Learning in Larval Zebrafish". *Curr. Biol.* (Apr. 2019). DOI: 10.1016/j.cub.2019.02.039.

[38] Andrew J Rennekamp et al. "σ1 receptor ligands control a switch between passive and active threat responses". *Nat. Chem. Biol.* (July 2016). DOI: 10.1038/nchembio.2089.

[39] Jason Rihel et al. "Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation". *Science* (Jan. 2010). DOI: 10.1126/science.1183090.

[40] Tabitha S Rudin-Bitterli et al. "Combining Motion Analysis and Microfluidics – A Novel Approach for Detecting Whole-Animal Responses to Test Substances". *PLOS One* (Dec. 2014). DOI: 10.1371/journal.pone.0113235.

[41] Skipper Seabold and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python". *9th Python in Science Conference*. 2010.

[42] Matthew P Shaw et al. "Stable transgenic C9orf72 zebrafish model key aspects of the ALS/FTD phenotype and reveal novel pathological features". *Acta Neuropathol Commun* (Nov. 2018). DOI: 10.1186/s40478-018-0629-7.

[43] Chanpreet Singh, Grigorios Oikonomou, and David A Prober. "Norepinephrine is required to promote wakefulness and for hypocretin-induced arousal in zebrafish". *Elife* (2015). DOI: 10.7554/eLife.07000.001.

[44] Peter J Steenbergen and Nabila Bardine. "Antinociceptive effects of buprenorphine in zebrafish larvae: An alternative for rodent models to study pain and nociception?" *Appl. Anim. Behav. Sci.* (Mar. 2014). DOI: 10.1016/j.applanim.2013.12.001.

[45] Olivier Stern et al. "Zebrafish Skeleton Measurements using Image Analysis and Machine Learning Methods". *Benelearn 2011* (2011).

[46] D C Swinney. "Phenotypic vs. Target-Based Drug Discovery for First-in-Class Medicines". *Clinical Pharmacology & Therapeutics* (2013). DOI: 10.1038/clpt.2012.236.

[47] Martin Thoma. "Analysis and Optimization of Convolutional Neural Network Architectures" (July 2017). arXiv: 1707.09725 [cs.CV].

[48] Xiaoying Wang et al. "Automatic multiple zebrafish larvae tracking in unconstrained microscopic video conditions". *Sci. Rep.* (Dec. 2017). DOI: 10.1038/s41598-017-17894-x.

[49] Rebecca Ward et al. "Pharmacological restoration of visual function in a zebrafish model of von-Hippel Lindau disease". *Dev. Biol.* (Feb. 2019). DOI: 10.1016/j.ydbio.2019.02.008.

[50] Westerfield. "The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish". *http://zfin.org/zf_info/zfbook/zfbk.html* (2000).

[51] Monte Westerfield. *The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio Rerio)*. University of Oregon Press, 2000.

[52] Marc A Wolman et al. "Chemical modulation of memory formation in larval zebrafish". *Proc. Natl. Acad. Sci. U. S. A.* (Sept. 2011). DOI: 10.1073/pnas.1107156108.

[53] Xiaoxuan Yang et al. "High-throughput Screening in Larval Zebrafish Identifies Novel Potent Sedative-hypnotics". *Anesthesiology* (Sept. 2018). DOI: 10.1097/ALN.0000000000002281.