Deep phenotypic profiling of neuroactive drugs in larval zebrafish

Leo Gendelev[1], Jack Taylor[1,5], Douglas Myers-Turnbull[1], Steven Chen[2], Matthew N. McCarroll[1,2], Michelle R. Arkin[2], David Kokel[1 *], Michael J. Keiser[1,2,3,4 *]

1. Institute for Neurodegenerative Diseases, University of California, San Francisco, San Francisco, CA, USA
2. Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, USA
3. Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA
4. Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA
5. UCSF Weill Institute for Neurosciences Memory and Aging Center, University of California, San Francisco, CA, USA

* Correspondence to: dave.kokel@gmail.com, keiser@keiserlab.org

Abstract

Behavioral larval zebrafish screens leverage a high-throughput small molecule discovery format to find neuroactive molecules relevant to mammalian physiology. We screened a library of 650 central nervous system active compounds in high replicate to train a deep metric learning model on zebrafish behavioral profiles. The machine learning initially exploited subtle artifacts in the phenotypic screen, necessitating a complete experimental re-run with rigorous well-wise randomization. These large matched phenotypic screening datasets (initial and well-randomized) provided a unique opportunity to quantify and understand shortcut learning in a full-scale, real-world drug discovery dataset. The final deep metric learning model substantially outperforms correlation distance–the canonical way of computing distances between profiles–and generalizes to an orthogonal dataset of novel druglike compounds. We validated predictions by prospective *in vitro* radio-ligand binding assays against human protein targets, achieving a hit rate of 58% despite crossing species and chemical scaffold boundaries. These newly discovered neuroactive compounds exhibited diverse chemical scaffolds, demonstrating that zebrafish phenotypic screens combined with metric learning achieve robust scaffold hopping capabilities.

Introduction

The mechanism of action of central nervous system (CNS) drugs remains poorly understood, even for those used for decades (e.g., ketamine[1]). The complex nature of G-Protein Coupled Receptor (GPCR) and ion channel-mediated pathways of the vertebrate nervous system[2–11] exacerbate the problem. Because of the prevalence of polypharmacology in neuroactive drugs[12], a "magic bullet" single-target approach to drug discovery[13] falls short[14]. Phenotypic screening circumvents these problems by identifying compounds that may interact with individual or multiple targets[15,16]. These screens prioritize desired and often biologically complex readouts of induced phenotypes on higher-level model systems. Despite historically limited throughput, rapid phenotypic profiling of thousands of compounds in vivo is now possible using larval zebrafish[17–21]. These vertebrates have high levels of shared genetics[22,23] and CNS anatomy[24] (with humans) and scale to high-throughput testing of complex behavioral readouts[2–11]. Phenotypic screening in larval zebrafish, combined with human-target-based cheminformatic methods such as the Similarity Ensemble Approach (SEA[25,26]), and enrichment factor (EF) calculations[27,10,28,29], have enabled novel drug discovery and target deconvolution for neuroactive phenotypes in mammals.

However, high-content zebrafish behavioral screening data are both a blessing and a curse for pharmacological studies because of the challenges in extracting and comparing features in the collected video data. In previous work, larval zebrafish, plated on 96-well plates, were treated with various compounds, and various stimuli — including acoustic stimuli and high-intensity light of different colors — elicited a broad spectrum of behavioral responses in the fish [10,28]. Videos recorded each well, from which a "motion-index" (MI) time series is computed to measure bulk motion over time (Figure 1a-d, [7] Figure 1b and Eq 1). Traditionally, the phenotypic distance between MI time series is computed using correlation distance[28]. Other approaches have included classification and video analysis using machine learning[2,30–34]. Correlation distance reliably discriminates antipsychotic[28] and anesthetic[10,28,29] phenotypes but fails to distinguish more subtle phenotypes. Indeed, fish rarely respond to stimuli in a one-to-one video frame correspondence when each frame is 1/30th of a second, breaking a basic assumption

of how these MI distances are traditionally computed. In an experiment with various assays, the strength of the response to each assay may vary in drug-treated fish; however, correlation distance values all frame contributions equally.

We sought a distance metric that leveraged zebrafish phenotypic screens for a broader range of induced behaviors. Specifically, we used a class of neural networks uniquely suited for learning distances between pairs of inputs (twin neural networks, twin-NNs, Figure 1e). These models were initially developed for biometric fingerprint verification[35], subsequently finding use in many machine learning (ML) tasks, such as "one-shot learning" on small datasets for image classification[36,37].

In this work, we screened a library of 650 ligands (from the SCREEN-WELL Neurotransmitter Set, "NT-650", Methods) in high-replicate and trained twin-NNs to relate drugs via the phenotypes they induce in larval zebrafish. We constructed the screens from the ground up with ML model training in mind, but the models still exploited unanticipated artifacts in the resulting screen dataset via an undesirable process known more broadly as "shortcut learning" [38]. We studied the effect of retraining the deep metric learning models on synthetically randomized datasets that we designed to test for confounding effects, ultimately driving the redesign and re-collection of a new experimental screen. Models trained on the revised screen cluster diverse neuroactive compounds in a way that corresponds strikingly well with known neuroactive biology, and they phenotypically link structurally distinct compounds by scaffold hopping[39,40]. Finally, the learned distance metric generalized to a screening dataset of novel compounds unseen during training, automating the discovery of novel neuroactive compounds active on human receptors when tested prospectively *in vitro*.

Results

Twin neural networks identify drug replicates from complex behavioral readouts

We collect a high-throughput phenotypic dataset based on the NT-650 neurotransmitter library screened in high replicate (7-10 replicates per drug) for training machine learning models.

We plate larval zebrafish onto 96-well plates (8 fish per well) and dose wells with drugs at a 10 µM concentration, a reasonable dose for in-vivo primary screening of neuroactives (Figure 1). Various stimuli, such as acoustic sounds, light stimulus, and physical tapping of the multi-well plate stage, are performed to elicit diverse behavioral responses in the fish, as optimized previously[7]. We record videos of the fish's behavior throughout the experiment. For each well, we encode and convert videos of larval fish into aggregate motion over time, resulting in a time-series vector, or "motion index" (MI).

We evaluate how well twin-NNs can identify whether two MI profiles, such as those shown in Figure 3a-b, originate from the same category — specifically, whether they are caused by the same drug. This is in contrast to other correlation metrics that often fail to reliably recognize when different samples have been affected by the same drug, especially when the resulting phenotypic changes are subtle. By necessity, a Twin-NN must learn which time points are most informative and how to correct for slightly or partially misaligned MI traces to correctly group same-drug replicates accurately across a diverse range of pharmacology and their concomitant behavioral traces. Twin-NNs consist of "twin" encoding layers, which share model weights and operate on a pair of different inputs (MI traces) to output a distance reflecting whether the MIs represent replicates of the same drug (distance = 0) or traces from mismatched drugs (distance > 1)[35].

We filter the dataset to remove human drugs that do not alter zebrafish behavior, namely those whose MI traces cannot be distinguished from vehicle controls with a simple random forest model (see Methods, Supplementary Figure S1). Drugs can fail to induce strong behavioral responses in zebrafish due to many factors, including differences in cross-species biology, concentration, incubation time, absorption route, or other factors. The neural network embedding architecture of each half of the model is a design choice; we implement a fully connected multi-layer perceptron (Twin-NN) (Figure 1e) as a baseline model and a second architecture motivated by DenseNet [41] (Twin-DN) as a more computationally expressive alternative. We explore recurrent architectures – neural networks designed to operate on sequences, such as the LSTM[42] or GRU[42,43] – but find that the relatively long length of our time series limited the feasibility of these approaches (not shown).

Prior work using zebrafish behavioral MI for scaffold-hopping and phenotypic drug discovery predominantly uses vector distance calculations to compare MI traces without warping, alignment, or relative weighting of individual time points. Compared to these conventional correlation and Euclidean distance approaches, both the Twin-NN and Twin-DN models discern positive (matched) and negative (mismatched) drug replicate pairs with drastically improved performance (Figure 2a), with Twin-DN scores achieving 0.97 ROC-AUC and 0.98 PRC-AUC (Figure 2b-c). We observe a near-perfect ability of the learned distance metric to discern MIs of replicates from the same drug from MIs of different drugs. Further, a Uniform Manifold Approximation and Projection (UMAP) [44] plot (Figure 2d), calculated over the means of the time series for all replicates of all 650 drugs in the screen, yields pronounced, discrete, and localized clusters. While heartening, this performance was substantially higher than we had anticipated and to a suspicious degree: many compounds do not reliably induce larval zebrafish behavioral phenotypes. Nevertheless, these results suggested that nearly all the compound library's experimental replicates could be grouped by the Twin-DN model with near-perfect fidelity. We wondered whether the model's exceptional performance might rely instead on shortcut learning[38] or the exploitation of hidden artifactual cues encoded within the data that were invisible to human researchers but perceivable by the deep learning model.

Machine learning exploits high-frequency components and plate-location effects

We evaluate the presence of shortcut learning in our model by testing how well the learned distance metric generalizes to an archival quality control (QC) screen (Methods) performed at an earlier time on data never seen by the metric-learning models. For this screen, we select 14 drugs with diverse mechanisms of action (MOAs) and assay them, also in high replicate, along with a vehicle (dimethyl sulfoxide; DMSO) control and lethal control (eugenol). To test generalizability, we train a k-nearest neighbors classifier on the QC replicates using one of three distance metrics: Twin-NN, Twin-DN, or correlation distance (Methods). For most drugs, the Twin-DN distance metric underperforms correlation distance (Figure 2e, orange bars). We suspected that the expressive Twin-DN models readily memorize the high-frequency

components in the time series, which may come from artifacts such as plate vibrations or high-frequency noise in the imaging sensor. Indeed, when we ablate the high-frequency components of the time series (with a Hanning smoothing filter[45] as implemented in the scipy package[46]), Twin-DN performance drops precipitously on the original NT-650 screen, with Twin-SN performance likewise dropping, but to a lesser extent (Figure 2 c-d). These results indicate that the learned distance metrics exploit the high-frequency components of the NT-650 screen data and that these hidden "shortcut" patterns do not generalize to the separate QC screen.

Per standard practice, we had already attempted to address potential overfitting or data leakage by splitting the training and validation sets by drug - leaving no replicates of any one drug in common between train and validation splits. Additionally, we performed a set of further machine-learning soundness checks (scrambling data labels and randomizing the training data features (Supplementary Figure S2), so the Twin-DN model's evident exploitation of high-frequency signal was initially surprising. However, all compounds in the NT-650 library arrived from the supplier in preset layouts, meaning all replicates of the same compounds (or all positive pairs) in the dataset always corresponded to identical plate locations. In contrast, mismatched pairs could come from any combination of compound locations across and within the plates. Our original experimental design for the screen did not control for the potentially confounding layout effect. Even a simple machine-learning architecture might be able to learn light-based patterns for distinguishing different location pairs. The Twin-DN model performance took the biggest hit with data smoothing, suggesting that the more expressive a model is, the more readily it can exploit feature shortcuts.

An optimized experimental screening design

To unequivocally control for within-plate positional confounding effects, we perform a second high replicate screen of NT-650, but this time with the treatments fully robotically randomized across plates and wells (Methods). We also include wells treated with a high dose of the anesthetic eugenol as a control[7] baseline for lethality. We take a near-identical approach to that of pre filter for drugs without effect as in the original screen, except that we use the random

forest model to label the MI profile into three possible bins: “active,” “inactive,” and “lethal” (Methods, Supplemental Figure S1).

We train new Twin-NN and Twin-DN models on the newly collected experimentally randomized NT-650 dataset (NT-650-revised). While the SNN models achieved slightly lesser performance on the randomized dataset than on the original non-randomized (NT-650-naive) well layout dataset (e.g., 0.84 vs. 0.89 AUROC for Twin-NN and 0.84 vs. 0.97 AUROC for Twin-DN), their performance still dramatically exceeded that of correlation distance and Euclidean distance approaches (0.66 and 0.62 AUROC, respectively). Striking differences in the distribution of Twin-NN and Twin-DN distances for the positive and negative pairs (Figure 3a) agree with greater ROC-AUC and PRC-AUC performance (Figure 3b-c). Fast-DTW, a popular dynamic time-warping approach for time series prediction that optimizes the alignment between time series [47], marginally improves on Euclidean distance and falls short of correlation distance in classifying positive versus negative pairs. Training these baseline models on the NT-650-revised screen with computationally smoothed high-frequency components did not significantly reduce performance, indicating the models no longer rely on high-frequency feature components.

Another way of assessing model performance is by measuring its ability to identify replicates of a compound. In the ideal case, a model can identify all replicates for compounds that induce significant behavioral responses. In practice, experimental replicates will often be ineffective for many reasons. However, the better a model performs, the more replicates across drugs it can identify. Although correlation distance identified one replicate for most drugs, it rarely identified three or more replicates, whereas Twin-NN did so frequently and sometimes picked up all 7–8 replicates (Figure 3d). This effect is emphasized by the early plateauing of the cumulative count curve for correlation distance.

Mapping a larval zebrafish “behaviorome”

Using the Twin-NN learned distance metric, we cluster the compounds’ MI traces from the fully randomized NT-650 screen and visualize the resulting phenotypic landscape by UMAP [44] (Figure 4a). We observed defined clustering and structure within this “behaviorome” view,

representing a behavior-based pharmacological map of 650 known human drugs in larval zebrafish. Ineffective drugs populated the yellow region, while drugs inducing behavioral readout changes favored the violet region. The most robust phenotypes appeared towards the bottom of the plot, falling in the negative value range for principal component 2 (y-axis) of the UMAP. One qualitative way to assess the behaviorome layout is whether drugs with similar known MOAs and indications group together. We highlight several such drugs in this plot with labels. The SSRIs fluoxetine and paroxetine clustered (labels 2,13) but distinctly separated from the tricyclic antidepressant clomipramine (label 4), although these shared a broader neighborhood as expected. This observation is consistent with the intuition that behaviors based on different classes of antidepressants should be more closely related to each other than to other classes of neuroactive drugs, such as stimulants. The dopamine $D_{2/3}$ agonists lisuride and PD 128,907 also appeared in a similar region of space (labels 10,12). Antipsychotics clozapine and mianserin appeared closely in phenotypic space (labels 1,7). Through the lens of correlation distance instead (Supplementary Figure S3), we see some similar high-level patterns but less behaviorome structure. For example, mianserin and loxapine are no longer neighbors; indeed, mianserin (label 1) appears closer to paroxetine (label 2) than to loxapine (label 7). In other words, correlation distance places an antipsychotic closer to an SSRI than another antipsychotic, suggesting a lower clustering quality based on this region's canonical MOAs and indications.

Model generalization and novel drug discovery

We test the ability of the machine learning model to generalize to a library of “novel” compounds from the DIVERSet, which had been screened months before the NT-650 set. The goal of this prior screen was novel neuroactive compound discovery rather than quality control or model training; thus, it traded fewer replicates for greater compound diversity. We previously used a similar library to discover novel compounds that cause paradoxical excitation in larval zebrafish [10,28,29]. In that study, we performed the phenotypic screen with the novel library, and the resulting MI traces were compared (using correlation distance) against a reference drug, etomidate, that consistently induced a strong phenotype in the larval zebrafish.

Here, we investigate if the learned distance model outperforms correlation distance in identifying novel compounds from the DIVERSet library that cause similar phenotypes. Instead of focusing on a single known reference drug, we calculate a distance matrix for every known compound in the NT-650 set against every novel compound in the DIVERSet library. Figure 5a shows an example of the top 5 compounds matching fluoxetine's phenotype. Evaluating performance in this context was challenging, as the novel compounds lack known bioactivity ground-truth labels.

As one means of assessment, we use an established systems pharmacology tool, the similarity ensemble approach (SEA[25]), to predict MOAs for all novel DIVERSet compounds from their chemical structures alone and compare these predictions against the established MOAs of NT-650 compounds that were their closest neighbors in the learned distance-metric space (see Methods). While SEA predictions are not perfect, they illuminate an otherwise dark MOA landscape of chemical matter. In a "phenosearch" approach, we rank-order and select the top 500 novel compounds by phenotypic distance to each known drug using correlation distance and the Twin-NN models.

We observe a striking enrichment for known-target MOAs for the Twin-NN distance over correlation distance (Figure 5b, d) based on the phenotypic associations for these novel compounds. Twin-NN identifies more novel compounds with similar MoAs to the drug queries than correlation distance (Figure 5b). Unexpectedly, random selection (as a null hypothesis; Figure 5b, violet line) typically outperforms correlation distance at identifying shared MOAs, highlighting the limitations of correlation distance as a metric for time-series data such as MI. We also compare distance metrics by examining how often negative control wells match up with known drugs. With correlation distance, negative controls frequently rank in the top-500 phenosearch list for known drugs, but not by Twin-NN distance (Supplementary Figure S4 a-b). These findings suggest that the Twin-NNs are more effective than correlation distance at discovering novel compounds that induce similar phenotypes to known drugs and improve scaffold-hopping and neuroactive drug discovery for novel chemical matter.

Experimental validation of learned-distance metric in finding novel compounds with shared pharmacology

Since the Twin NN models consistently enriched for predicted MOAs of novel compounds shared with known compounds (Figure 5b), we sought to experimentally test these learned-distance predictions prospectively in a scaffold-hopping drug-discovery scenario. We select 12 neuroactive drugs from diverse regions in the behaviorome UMAP (Figure 4a). We purchase the top 5 novel DIVERSet compounds ranked by Twin-NN distance for each drug (60 compounds in total). We hypothesized that the novel compounds acted through the same protein targets as those known for the drugs that the novel compounds mimicked phenotypically.

This was a straightforward logic in some cases: for example, IMETIT is a human Histamine H_3 agonist with activity at Histamine H_4[48]. We purchase and test its five novel "pheno-matched" compounds for direct binding to human Histamine H_3 and H_4, discovering binding of three of the novels to H_3, at 1.1μM, 0.99μM, and 2.7μM (binding affinity K_i), and of one novel to H_4 at 5.4 μM (K_i) (Figure 6c). In other cases, the choice of test targets was more complex, such as for the tricyclic antidepressant clomipramine, an inhibitor of serotonin and norepinephrine transporters with additional activity against other GPCRs, including serotonergic, dopaminergic, adrenergic, and histaminergic receptors. Furthermore, a compound's most potent activity in humans may not always account for its observed behavior in zebrafish. Off-target or side activities might cause the most pronounced response in the fish; this is an inherent limitation in the cross-organism study's design for polypharmacological drugs. Clomipramine's phenotypic location being closer to chlorpromazine than to the SSRIs fluoxetine and paroxetine in the behaviorome (Figure 4a), illustrates one such case. In humans, the clinical timescales involved in serotonin reuptake for behavioral modification are much longer [49] than the 1-hour treatment duration used in our phenotypic screening, so we reasoned that the novel compounds phenomatched with fluoxetine might have acted through a subset of the targets it shares with chlorpromazine, such as serotonin 2B (5-HT_{2B}) [50,51]. Accordingly, two of clomipramine's top 5 pheno-matched novel compounds achieve affinity (K_i) of 33 nM and 1.9 μM K_is at 5-HT_{2B} in prospective testing (Figure 6c).

We test 216 new compound-target pairs based on 60 unique compounds and 17 unique protein targets. Of these, 8.3% are active at 10 μM or better Ki (Figure 6a-b; Supp Table T2). IMETIT has the highest hit rate; 3 of its top 5 novel compounds have at least 50% inhibition at 10 μM or better against at least one of the targets; in the dose-response assays, two yield $K_i < 10$ μM, and the most potent, compound 58040, has a K_i=0.99 μM for Histamine H_3. Overall, 7 of the 12 drug queries yield at least one novel hit for a 58% per-query hit rate; this corresponded to a 22% hit rate on a per-compound basis. All the dose-response binding curves from the secondary assays for the hits are provided (Supplementary Figs S5-S13). Where the tests failed, we may have picked the wrong subset of a query drug's protein targets to test against its novel compounds. For instance, clomipramine has known activities at a substantially wider range of targets than we could empirically test within the scope of this study, and this may account for mechanisms of action for those of its novel compounds that did not bind to 5-HT_{2B}.

Learned phenotypic distances enable chemical scaffold hopping

Despite strikingly different chemical structures, the learned distance metric identified compounds that induced a similar behavioral phenotype in the case studies. We explored this idea further by comparing ECFP4[52] (chemical fingerprint Tanimoto distance) versus Twin-NN phenotypic distance for all possible combinations of two drugs from the randomized highly-replicated library used for training (Figure 7). Here we define four quadrants: top-left (low Tanimoto distance, high Twin-NN distance), top-right (high Tanimoto and high Twin-NN), bottom-left (low Tanimoto and low Twin-NN), and bottom-right (high Tanimoto and low Twin-NN). We color the dots by the average "drug-likeness" of the compound pair, which we define as the strength of the phenotype (distance from control) minus the toxicity score (yellow for non-drug-like/toxic to purple for very drug-like). Of potential interest in drug discovery efforts, the bottom-right region (dark) highlights where the commonly used cheminformatic means of comparing two molecules fail, but the Twin-NN distance succeeds.

Tanimoto chemical-structure distance does not correlate with phenotypic distance, except for isolated cases in the lower left (Figure 7). Most pairs have Tanimoto chemical-structure distances greater than 0.4, despite sometimes inducing similar phenotypes through putatively

shared MOAs. Dot size reflects observed MOA similarity, computed as a separate Tanimoto distance between the vectors of known-target activities for the two drugs derived from the ChEMBL 23 pharmacology database [48], Methods). The highest concentration of high-target-similarity compound pairs (large dots) favors regions where phenotypic distance is low and chemical structure distance is average (0.3–0.7). This enrichment of known-MOA matches in the presence of good (low) Twin-NN phenotypic distance pairs is consistent with learned phenotypic distance predicting shared biological mechanisms. We hone in on these known-drug pairs with several thresholds (>0.2 ChEMBL target-activity similarity, a chemical-structure distance >0.5, and a Twin-NN phenotypic distance <0.3), which yields 51 known-drug pairs that we rank by biological target similarity (full table provided in Supplementary Table T3). The drugs in the top pair (7-OH-DPAT and ropinirole) are potent Dopamine D_3 agonists and antiparkinsonian agents. Thus our Twin-NN phenotypic distance associates known drugs with a shared mechanism of action but high chemical structure distance, highlighting its usefulness for scaffold hopping.

On the other hand, some pairs of drugs with high phenotypic similarity and middling structural similarity lack shared MOAs (small dots), which suggests these drugs induce similar phenotypic effects in larval zebrafish through different, parallel, or unstudied MOAs. These pairs correspond to a region of the known-drug space of particular interest for drug discovery, and further studies might explore why these pairs of known drugs are linked phenotypically in our study through potentially underexplored mechanisms.

Discussion

Deep metric learning models trained on high-replicate phenotypic larval zebrafish screens identify pairs of drug-like compounds despite experimental variability, group human drugs based on zebrafish effect, find connections among compounds that traditional chemical data analyses fail to make, and group structurally distinct novel compounds by biological MOAs. These observations support using metric learning on large phenotypic screening datasets for drug discovery and scaffold hopping. Moreover, our first implementation of these complex learned-distance models fell prey to "shortcut learning,"[38] wherein they exploited experimental

artifacts in the screening dataset to achieve misleadingly high performance that did not generalize to similar but independent zebrafish screens. This deep *mis*-learning was nuanced and eluded conventional cross-validation, soundness checks, and exploratory data analysis tests. We believe the strategies described here to detect, correct, and stress-test the experimental screening datasets and revised models will find use in other studies that combine complex biological data with deep learning models.

Straightforward measurement methods like correlation, Euclidean, or dynamic time-warping distance fall short when identifying drugs whose replicates induced perceptible but subtle changes in zebrafish behavior (Figure 3a-b). The main issue is that these methods cannot differentiate between irrelevant random variations and meaningful changes that illuminate the underlying pharmacology. Conventional metrics take all time points into account without weighting their importance. On the other hand, contrastive metric learning models disregard irrelevant parts of the data (features) and concentrate on the segments that display significant behavioral differences. For instance, clozapine- and DMSO-treated zebrafish exhibit periods of reduced motion (Figure 1d, time points 600-700). Clozapine can look like a negative control by standard correlation methods, which attribute equal importance to periods of inactivity and activity. In an extreme example, correlation distance scores two traces as almost identical when comparing a drug that sedates the fish except for a sudden movement spike versus a lethal control such as eugenol. However, a metric learning model learns that sudden motion spikes matter in differentiating drugs.

At a more global level, we construct a “behaviorome” - a visual map of drug similarity based on zebrafish behavior. This landscape, created by pairing zebrafish phenotype with an appropriate distance metric, reveals relationships between known neuroactive drugs and identifies underexplored areas with potential for drug discovery. From high-throughput behavioral screening data and the learned distance metric, we link human drugs directly to the *in vivo* vertebrate behaviors they induce. Classical informatic methods falter on diverse chemical structures, as they rely by necessity on the similar property principle of chemical informatics.[53] This is particularly true at activity cliffs,[54] where slight chemical structure changes drastically affect bioactivity. Phenotypic screening, using behavior, circumvents these limitations. Different

compounds triggering similar zebrafish behaviors may interact with the same targets and pathways. The learned distance metric complements raw structural similarity (Figure 7), underlining traditional cheminformatics limitations and opportunities for drug discovery and scaffold hopping.

We attempted to automate the discovery of novel drug hits for disease-related mechanisms and pathways in the CNS. For new compounds, such as those from the Chembridge DIVERSet library, the Similarity Ensemble Approach[25,26,54] predicted unknown experimental MOAs. The metric learning models identified library compounds with marked enrichment in their predicted MOAs to the MOAs of known drugs, indicating pharmacological similarities (Figure 5b and d). Instead of relying on *in silico* validation, we experimentally tested the predicted MOAs *in vitro* via prospective radio-ligand binding assays. We found that neuroactive drugs successfully linked to novel library compounds by phenotype and MOA 58% of the time. This hit rate surpassed early drug discovery hit rates using high throughput screening (HTS, 0.01-0.14%) or virtual screening (VS, 1-40%).[55] Unlike typical HTS or VS hits, behavioral hits may offer more robust lead series starting points because they, by definition, already trigger an *in vivo* effect in zebrafish and show animal tolerance. Many *in vitro* hits fail *in vivo* due to absorption, distribution, metabolism, excretion (ADME) issues, and pharmacodynamic/kinetic properties such as blood-brain barrier penetration are crucial for neuroactive drugs. However, deep metric learning on behavioral screening data quickly identified hits that could circumvent these issues.

In an unintended but instructive project outcome, we grappled with the first metric learning models silently exploiting shortcut learning on the original dataset, which had not used randomized plate layouts. Despite passing conventional soundness check analyses, including label randomization and scrambling input features (y- and x-scrambling), the learning models exploited subtle experimental dataset artifacts. Pre-determined plate layouts from drug suppliers might inadvertently teach the models positional effects by exploiting slight irregularities in the experimental setup, such as minor differences in distance to directional light and sound sources (Figure 1a). These effects, imprinted in high-frequency components of time-series traces, were imperceptible to humans but perceptible to deep learning models. This generalizability limitation

was not an overfitting issue and could not be rectified by refining training-test set splits, such as scaffold-splitting drugs or time-series trace clustering. Consequently, we re-ran the full-scale experimental screen with robotically randomized plate layouts on the same compound library to assess this challenge unequivocally. Indeed, models trained on the original dataset deteriorated when we computationally smoothed high-frequency components of the motion index traces (Figure 2b), but those trained on the fit-to-purpose randomized screen remained unaffected (Figure 3b). While we might instead have attempted to train generative adversarial networks (GANs)[56] to remove shortcut signals computationally,[57] complex models such as GANs can be brittle, and we sought a definitive analysis. As an intriguing challenge, follow-up studies by those interested in mitigating shortcut learning might find value in comparing new algorithmic versus the experimental plate-effect removal strategies on these two datasets.

We faced several practical caveats in the metric learning training procedures. Particularly, mismatched compound pairs within the same pharmacological class may trigger similar behaviors in zebrafish. We considered using Anatomical Therapeutic Chemical (ATC)[58] class or predicted protein target activity profiles by the Similarity Ensemble Approach (SEA)[25,26] to exclude misleading false-negative compound pairs from model training. However, ATC classes operate across a hierarchy of varying branch depths, and it is likewise not clear what threshold to use for SEA-prediction similarity, given the ~2,000 proteins in a target profile. Conversely, we might incorrectly label compound pairs as positive (false-positives) if they do not elicit a strong behavioral response. Inactive compounds could result from biological differences between humans and zebrafish, inactive concentrations, or limited effects on zebrafish behavior in our particular assay conditions. To tackle this, we deployed a separate random forest (Methods) to remove inactive traces from positive-pair candidacy before metric learning training as a provisional solution. Consequently, the ground truth labels of compound phenotypic similarity used during model training are imperfect and noisy. Whereas improving these weak labels[59] may be an avenue for further refinement, we found them sufficient to train distance metrics robust to this biological label noise.

Comparing our success rates with conventional single-target-based high throughput screening (HTS) or virtual screening (VS) presents different hurdles. Since we do not know the

protein MOA for novel neuroactive compounds *a priori*, we tested novel compounds against multiple predicted protein targets. Consequently, we calculate a "best-of" hit rate, which provides more identification opportunities than a single-target screen. However, the lack of knowledge about which protein targets the novel compounds impact makes direct comparisons with per-target success rates problematic. Our overall hit rate was 58%, which implies a 42% chance that a given query using a known drug would result in no protein-matched hits. These represent missed opportunities more than methodological failure. Here, errors in MOA prediction for novel compounds or cryptic but shared protein off-targets may cause unexpected associations between known and novel compounds. However, as cheminformatic target prediction accuracy improves, this will further complement the phenotype-based metric learning approach. Finally, we acknowledge that the larval zebrafish animal model for studying neuroactive drugs has limitations due to genetic, anatomical and behavioral complexity differences with mammals. Consequently, we must carefully vet the compounds, pathways, or behavioral phenomena identified in the larval zebrafish in more advanced animal models and humans to establish their therapeutic import, which is beyond our scope. This study's blend of screening technology and metric learning is thus a tool to complement but not replace accepted animal models and methods.

Deep metric learning models with high-replicate behavioral zebrafish screens directly reveal scaffold hopping opportunities. These models outperform traditional distance metrics and cheminformatics methods, accurately classifying and grouping compounds from their zebrafish behavior alone. Prospective testing confirmed most of the predicted neuroactive MOAs using human receptors *in vitro*. Deep metric learning enriches phenotypic screening, yielding novel compounds with actionable neuroactive effects despite different chemical structures. Despite the challenges presented by experimental variability and shortcut learning—where models exploited experimental artifacts—we successfully redeployed the screen and stress-tested the models, creating a robust approach applicable to diverse investigations that pair complex biological data with deep learning. Closely integrating fit-to-purpose larval zebrafish behavioral screening with deep metric learning is an efficient and robust way to identify new neuroactive compounds in vertebrates.

Methods

Animal husbandry

Animal husbandry was performed according to UCSF's Institutional Animal Care Use Committee (IACUC) and the Guide to Care and Use of Laboratory Animals (National Institutes of Health 1996). Eggs from a wild-type "Singapore" strain were collected by group matings and raised on a 14/10-hour light/dark cycle at 28°C in egg water (GCULA) until 7 dpf. 8 healthy larvae were then distributed by pipette into the wells of 96-well plates. They were then incubated pre-treatment for 1 hr, dosed, incubated post-treatment for 1 hr, and then screened in the behavioral instrument.

Chemical libraries

Two chemical libraries were used in our study: the SCREEN-WELL Neurotransmitter Set (Enzo Life Sciences, Farmingdale, USA https://www.enzolifesciences.com/BML-2810/screen-well-neurotransmitter-library-10-plate-set/ and the ChemBridge DIVERSet Screening Library https://www.chembridge.com/screening_libraries/diversity_libraries/.

Screening platform

The screening platform is described in detail in doi:10.1101/2020.01.01.891432 (Figure 1 and Methods). The QC set screening methods are also described in that study. For the randomized experiments using the Screen-Well Neurotransmitter Set, we randomized the plate layouts with a custom code provided with this study. We transformed the physical layout of the plates accordingly using a BioMek robot in the Arkin lab at UCSF.

Data collection

Larval zebrafish are plated onto 96-well plates (8 fish per well), and wells are dosed with drugs at 10 µM for primary screening (Figure 1). Various stimuli such as acoustic sounds and

physical tapping of the plate platforms are performed to elicit diverse behavioral responses in the fish, as optimized by Myers-Turnbull et al.[7] Videos are recorded of the fish behavior for the duration of the experiment, which for the screens discussed in this work is around 14 minutes. For each well, the videos are encoded and converted into bulk motion over time, resulting in a one-dimensional time series or "motion index" (MI). Specifically, we used pre-interpolation m' values[7] defined by:

$$m'(I^t) = \sum_{ij} \mathbb{1}\left|I_{ij}^t - I_{ij}^{t-1} \geq 10\right|$$

where I^t is the grayscale image matrix at 1-indexed frame t. These videos represent the average motion across all 8 fish in each well. Since zebrafish movement can be uncoordinated, averaging over multiple fish can greatly improve the signal-to-noise ratio for many classes of drugs[7,10].

Filtering ineffective and lethal compounds by Random Forest

For the first high-replicate screen, we trained a random forest classifier to identify ineffective compounds (mimicking DMSO) using sci-kit learn[60]. The inputs are the MI of drugs and DMSO-treated wells, and the output is a binary label (effective or ineffective). We first split the entire dataset using an 80/20 train/test split. There were fewer DMSO-treated examples, so we randomly undersampled from the drug-treated wells to match the number of DMSO-treated wells. This resulted in 556 examples from each class in the training set, and 180 examples in each class in the test set. For the randomized high-replicate screen, we include positive controls (eugenol) which is lethal to the larval zebrafish. Here, the random forest is trained to label MI into one of 3 possible bins ("effective," "ineffective," or "toxic"). As before, we first split the entire dataset using an 80/20 train/test split. There were fewer "toxic" examples than in the other 2 classes, so we randomly undersampled from those classes to match the number of "toxic" examples. This resulted in 100 examples for each class in the training set and 28 examples for each class in the test set (using an 80/20 train/test split).

Conventional metrics used to calculate phenotypic distance

We used the sci-py package [46] to compute correlation, euclidean, and the "fast-dtw"[61] python library to compute the dynamic time warping distance between the 10500 long MI time series traces.

Classifying quality control drugs with distance metrics

We trained a kNN (k-Nearest-Neighbors) [62] algorithm as implemented in the sklearn KNeighborsClassifier package to classify the quality control (QC) 16 neuroactive drugs based on their motion-index time-series traces using the implementation in the scikit-learn package [46,60]. Each QC compound was screened in replicates of 10, so we split the dataset into train and validation splits (8 train, 2 validation) replicates for each drug. The task of the KNN was to predict, for a given time series, which one of the 16 drugs it most closely corresponds to. We chose 15 for the number of nearest neighbors parameter.

Training deep metric learning models for phenotypic distance:

In general, a dataset of positive and negative pairs is required to train a Twin-NN or Twin-DN model. To construct such a dataset, we screened the SCREEN-WELL Neurotransmitter Set ("NT-650") in high replicate. We define any two replicates of the same compound to be positive pairs, while a replicate of one compound paired with a replicate of another compound was a negative pair. Each plate of 96 drugs was replicated 7-10 times, creating a sizable dataset of positive and negative pairs. In practice, not all compounds will exhibit an observable effect in larval zebrafish, so it can be misleading to label replicates of such drugs as "positive" (see Discussion). Thus, we filtered all pairs where at least one pair member was "ineffective" or "lethal" by the random forest. We trained the second round of Twin-NNs similarly but using the fully randomized dataset instead.

To create the dataset of pairs, we first split the dataset by drug to minimize memorization and over-fitting, allocating 80% of drugs for training and 20% for the test set. This scheme naturally presents a generalizability challenge for the models, since phenotypes induced by the training drugs might not be induced in the test drugs. Data splitting encourages the models to learn fundamental features that are independent of drug or phenotype, which can lead to much

better generalizability. Next, we performed a class balancing procedure. There are many more pairs that can be enumerated across different drugs (negative pairs) than pairs from the same drug (positive pairs), but this could lead to class-imbalance issues during training. Hence we randomly subsampled the negative pairs to match the number of positive pairs for both the training and test sets. Next we aimed to include additional commonly encountered pair types. For positive pairs, we often encountered control-control and tox-tox pairs, and for the negative pairs, we often encountered control-drug, control-tox, and tox-drug pairs. To ensure the models were exposed to enough of these pairs, we ensured that 25% of the total positive pairs in the dataset came from the control-control or tox-tox classes (allocated evenly), and that 25% of the total negative pairs in the dataset came from the control-drug, control-tox, or tox-drug classes (allocated evenly across these 3 classes). We saved pairs to numpy arrays of indices that corresponded to indices of the time series data, and provided both the pair arrays and raw data in our online data repository.

We used PyTorch to train the Twin-NN and Twin-DN models (see github repository). Briefly, we loaded the positive and negative pairs using a Pytorch Dataloader, randomly swapped the pair order, sampled at every 5th frame and min-max normalized them. Then we passed these pairs of motion index time series through the MLP architecture (Twin-NN) or Dense architecture (Twin-DN). The Twin architecture was a 6-layer feed-forward neural network. The input layer size was 20250 (length of the input); each subsequent layer was 4000, 500, 250, 100, and 10, respectively. After each linear layer (except the last) we performed batch normalization and ReLU activation. We passed each time series from an input pair this feed-forward architecture, after which we computed a contrastive loss from the 2 outputs (vectors of length 10 each). We used a margin of 0.5 for the negative pairs in the contrastive loss. We back-propagated the contrastive loss and updated model weights after each batch, until reaching convergence or the maximum epoch count. We used the same training procedure for the Twin-DN models, except that we based the architecture on DenseNets instead. The final output for each input from the DenseNet was also a vector of length 10. To train the models, we used a learning rate of 5e-4 with the Adam optimizer and weight decay set to 1e-6. For the Twin-NN model, we used a batch size of 32. For the Twin-DN model, we used a batch size of 8, as this was the largest batch size

to fit in GPU memory. We used an NVIDIA GeForce GTX 1080 Ti GPU on a CentOS Linux kernel 3.10.0 operating system with an x86-64 architecture.

Stress-testing the metric learning models

We performed high-frequency signal filtering using a Hanning smoothing filter[45] as implemented in the scipy package[46], using a window size of 11. We tried three adversarial controls on our Twin-NN models: label-shuffling, input randomization, and predicting well distance (Supplementary Figure S2). For label-shuffling, we randomly shuffled the labels while keeping the input fixed. For input randomization, we generated random MI vectors of the original length using the python numpy package[63], while holding the labels fixed. For well distance, we used the Twin-NN models to predict well distance (computed as the Euclidean distance between pairs of wells). We define as "same" those pairs that are neighbors below a certain distance cutoff (2 and 5.2 plate distance units), and as "different" those pairs that are distant from each other (above the chosen cutoffs). The Twin-NN models appear to have no ability to distinguish between pairs in this context at either of these cutoffs, suggesting a lack of strong signal in the motion index time series attributable to plate location alone.

Chemical informatics and bioactivity prediction

Chemical structure similarity is computed from ECFP4 fingerprints and tanimoto similarity using the rdkit package. Bioactivities of compounds of unknown indication or mechanism of action (as for hits from the SCREEN-WELL Neurotransmitter Set) are computed using the similarity ensemble approach (SEA [25,26]) together with version 23 of ChEMBL [48]. We use a SEA p-value cutoff of 1e-25 for bioactivities.

In vitro binding assays against human receptors

The 60 novel compounds from the DIVERSet screen were tested in radioligand binding assays (performed as previously described [64–66]) against human receptors at the National Institute of Mental Health Psychoactive Drug Screening Program at UNC (PDSP). Detailed assay protocols and procedures are also available from the NIMH PDSP homepage

(https://pdsp.unc.edu/pdspweb/?site=assays). Primary inhibition screens were performed at the final dose of 10 μM and compounds passing a threshold of 50% inhibition were subjected to secondary dose-response assays.

Use of Large Language Models (LLMs)

We used OpenAI ChatGPT 3.5 Turbo and ChatGPT 4 as scientific editing tools when writing the manuscript. We prompted the LLMs to suggest revisions to our manually drafted text for improved clarity and conciseness at the sentence and paragraph levels. We did not ask the LLMs to generate content *de novo*. An example of a prompt we used was, "You are helping edit papers for a broad scientific audience, emphasizing clarity and conciseness. Revise the following paragraph: <draft text here>." We manually reviewed the LLMs' suggested revised text word-by-word and decided whether to include parts, all, or none.

Figure Legends

Figure 1. Zebrafish behavioral screening and architecture

Diagrammatic representation of zebrafish experimental screening setup, motion index calculations, and deep metric learning model architectures. **(a)** Simplified representation of zebrafish screening platform, with larval zebrafish in 96-well plates under a camera subject to varied stimuli such as blue light, purple light, acoustic stimuli, and physical tapping. **(b)** Example of a representative video frame. **(c)** Zoom view of a single well. **(d)** Example motion index (MI) time series for clozapine-treated fish and negative control (DMSO) wells. The MIs are averaged across all the drug- and control-treated well replicates. **(e)** Deep metric learning model architectures: Twin-NN (top) and Twin-DN (bottom). In both models, the input is a *pair* of MI time-series vectors passing through multiple neural net layers. A contrastive loss function [67] scores the two learned output vectors (y1 and y2) distances based on whether the input MI vectors were from the same or different treatments.

Figure 2. Metric learning models exploit high-frequency components of time-series signals in an initial non-randomized screen

We compare the Twin-NN and Twin-DN models against traditional methods such as correlation and Euclidean distance on both raw and smoothed motion index (MI) time series data and examine how well the models cluster drugs and generalize to a separate dataset not used for training. **(a)** Separation of positive (treatment replicate) and negative (mismatched replicates) MI vector pairs using the Twin-DN model (left), the Twin-NN model (2nd column), correlation distance ("Correlation," 3rd column), and Euclidean distance ("Euclidean," right). The Twin-DN and Twin-NN models exhibit a drastically improved ability to separate positive from negative pairs, as evidenced by the strong distance separation (x-axis) between the positive and negative pair distributions. Euclidean and correlation distances fail to separate the same- from mismatched replicates for most MI pairs, except for those with minimal distances (e.g., the most phenotypically similar pairs). **(b)** Receiver operator characteristic plots. Twin-DN achieves the

best area under the curve (AUC=0.97), followed by Twin-NN (0.89). Performance dropped drastically for both learning models using MI time-series inputs smoothed with a Hanning window[45] of size 11, particularly Twin-DN (from 0.97 to 0.78). Correlation and Euclidean distance were robust to Hanning smoothing. **(c)** Precision recall curves, showing trends consistent with (c). **(d)** A UMAP[44] using the Twin-DN distances reveals extreme clustering with distinct phenotypic islands; as many drugs are unlikely to induce strong phenotypes in the fish, this was an unexpected and suspicious result. **(e)** We trained k-Nearest-Neighbor (kNN) classifiers using scikit learn[60] on a separate high-replicate MI trace dataset of 16 "quality control" drugs never used for training or model evaluation. For many drugs (e.g., haloperidol), the Twin-DN-based distance underperforms the zero-baseline defined by kNNs using correlation distance. Twin-NN distance outperforms correlation distance on a few drugs (e.g., tiagabine and lidocaine) and always matches or exceeds the Twin-DN model.

Figure 3. Metric learning operates actionably on a fully randomized screen

We investigate models trained on the second, fully randomized screen. **(a)** Separation of positive and negative motion index (MI) trace pairs from the fully randomized screen with Twin-NN (left), Twin-DN (2nd column), correlation (3rd column), euclidean (4th column), and Fast-DTW (right) distances. Assessed as in Figure 2a, the revised deep learning models significantly outperform correlation, euclidean, and fast-DTW distances. **(b)** Twin-NN and Twin-DN receiver operator characteristic performance is similar (AUC=0.84 and 0.79, respectively) and significantly exceeds correlation, euclidean, and fast-DTW (0.66, 0.62, and 0.64). Notably, models trained with and without Hanning smoothing no longer differ. **(c)** Precision recall curves are consistent with (b). **(d)** The Twin-NN model identifies matched drug replicates more effectively than correlation distance, which typically starts to fail beyond one replicate.

Figure 4. A learned phenotypic distance identifies islands of drugs by protein target profile that correlation distance cannot

We investigate how well the learned phenotypic distance meaningfully clusters drugs using a UMAP [44] on Twin-NN distances between the average time-series across all replicates of each drug of the fully randomized NT-650 screen. Labeled example drugs represent anchor points across the phenotypic landscape. Dot color changes by phenotype strength as determined by a separate random forest classifier employed earlier in the dataset construction process (Methods). A UMAP on correlation distances of the same data (Figure S3) fails to form meaningful phenotypic clusters.

Figure 5. Learned distance is a good proxy for the target bioactivity profile of novel compounds

Assessed in a scaffold-agnostic screening paradigm, we compare motion index (MI) traces of NT-650 "query" compounds against a screened library of "novel" compounds (Chembridge DIVERSet) using the Twin-NN learned distance and correlation distance, versus a random baseline wherein the "matched" traces are randomly selected. **(a)** As an example, the fluoxetine MI trace (purple) from the NT-650 agrees well with the top 5 matched library compound traces (gold) ranked by Twin-NN distance. **(b)** We use a separate chemical informatic method, the Similarity Ensemble Approach (SEA[25,26]), to assess the library compound hits. Ranked by the similarity of their phenotypes to drugs from the NT-650 screen, we would expect that the likelihood of SEA target profiles between a "query" (NT-650) and its closest-match "library" (Chembridge) compounds will increase with the quality of the phenotypic distance metric. "Hits" (y-axis) are the number of novel compounds in a given sample that match by their separate SEA profiles. "Sample" (x-axis) is the percentage of the novel library examined, where the analysis is limited to the top 500 matches from the library. The learned distance metric enriches for SEA hits better than correlation and the random baseline across the entire range of the screen. **(c)** Similar to (b); but for specific NT-650 compounds selected by phenotypic strength (see Methods). Learned distance outperforms correlation and random distance, as with pindolol,

imetit, and chlorpromazine. Correlation distance has significantly better enrichment for only one NT-650 compound, MDL 72832 (4th row, 4th column in grid plot).

Figure 6. Prospective experimental validation against human receptors *in vitro*

Using a cheminformatic protein target prediction approach (SEA[25,26]) with the Twin-NN phenotypic distance, we make mechanism of action predictions for NT-650 compounds and test them experimentally by radioligand binding assays. **(a)** Left: Top 5 novel compounds (represented by their motion index time-series, rows 2-6, gold) matched by Twin-NN distance to the NT-650 drug Imetit (top row, purple). Right: Diagram of this "phenoblast" approach. NT-650 drugs are columns. DIVERSet novel compounds are rows, ordered by Twin-NN phenotypic distance. Supplemental Table T1 maps compound IDs to supplier IDs. **(b)** Primary radioligand binding assays (binding inhibition at 10 μM, %) for 7 known drugs. The heatmap shows 5 novel compounds selected for testing (rows), with the SEA-predicted human protein targets as columns. **(c)** Same as (b) for secondary assays (dose-response radioligand binding experiments). **(d)** Representative dose-response curves from (c) for selected novel compounds tested against two human targets: the histamine H3 receptor (left) and the 5-hydroxytryptamine 2B receptor (right). Results (mean ± SEM) from a minimum of 3 independent assays (each in triplicate) were normalized, pooled, and fitted to the built-in one-site competition binding function in the GraphPad Prism V10.

Figure 7. Phenotypic screening with learned distance reveals scaffold hopping and drug prospecting opportunities

Learned phenotypic distance (Twin-NN, y-axis) complements conventional chemical informatic distance based on chemical structure (Tanimoto coefficient on ECFP4 circular fingerprints, x-axis). Scatterplot contains all pairwise combinations of 83 NT-650 compounds (each pair is a dot), calculated from their average MI traces and chemical structures. Where the phenotypic distance is low (< 0.3) but the tanimoto distance is average or high (> 0.4), molecular structure dissimilarity misses neuroactive similarity. We illustrate each quadrant with examples. Bottom left: low tanimoto and phenotypic distance (both metrics agree that molecules are

similar). Bottom right: low phenotypic distance and high tanimoto distance (scaffold hopping opportunity). Top-left: low tanimoto distance, but high phenotypic distance (classic “activity cliff”: disparate activity despite high structural similarity). Top right: high tanimoto and phenotypic distances (both metrics agree that molecules are unrelated).

Figures

Figure 1

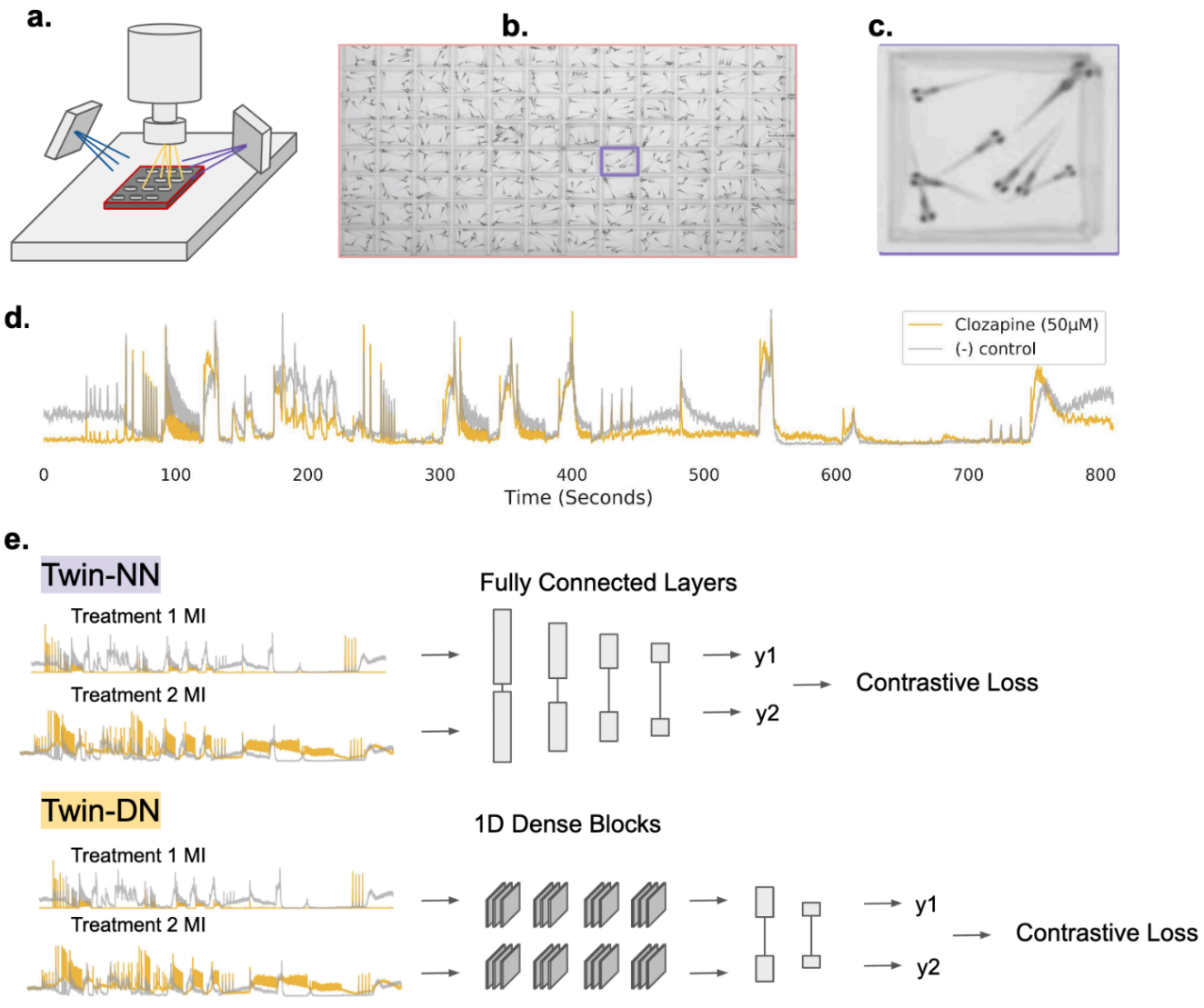


Figure 2

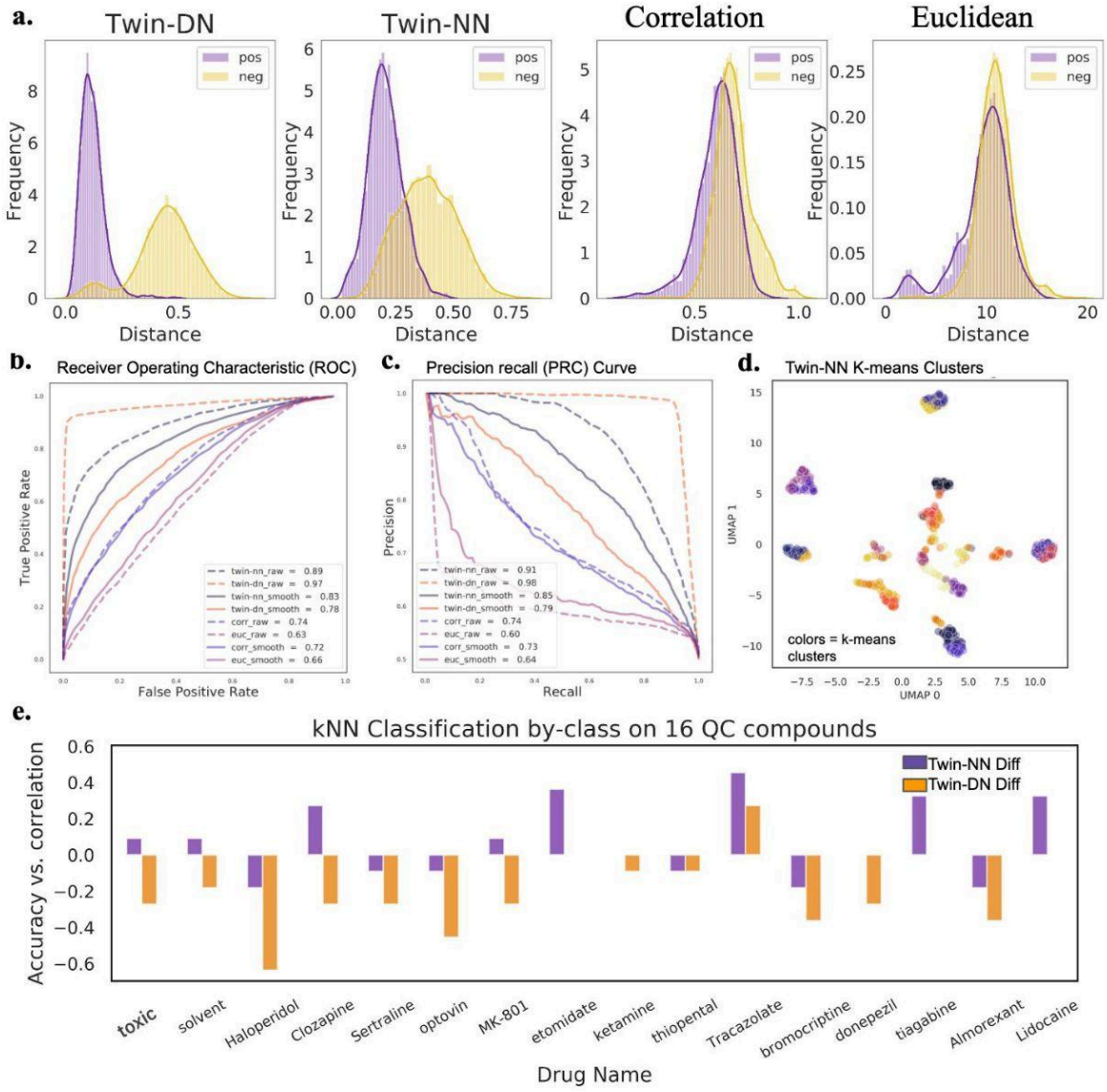


Figure 3

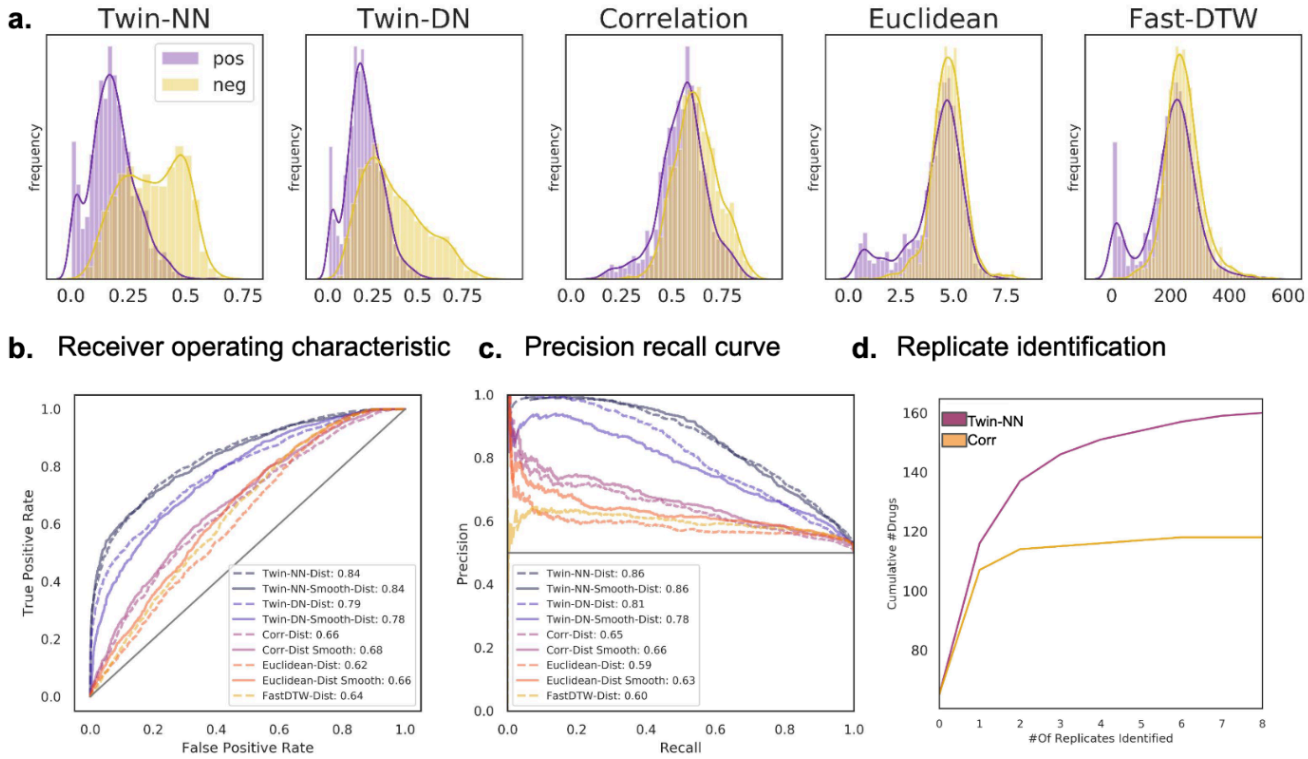


Figure 4

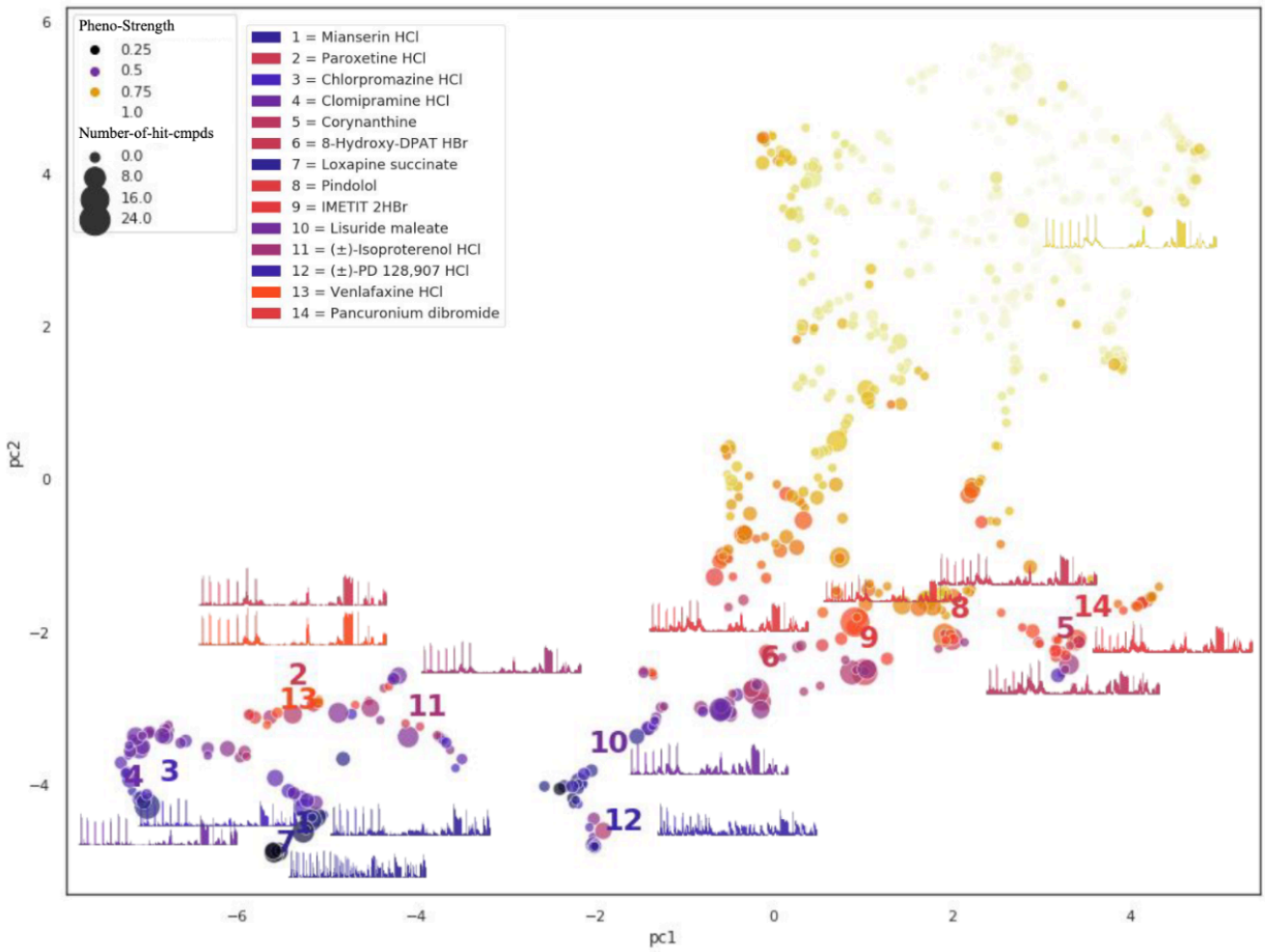


Figure 5

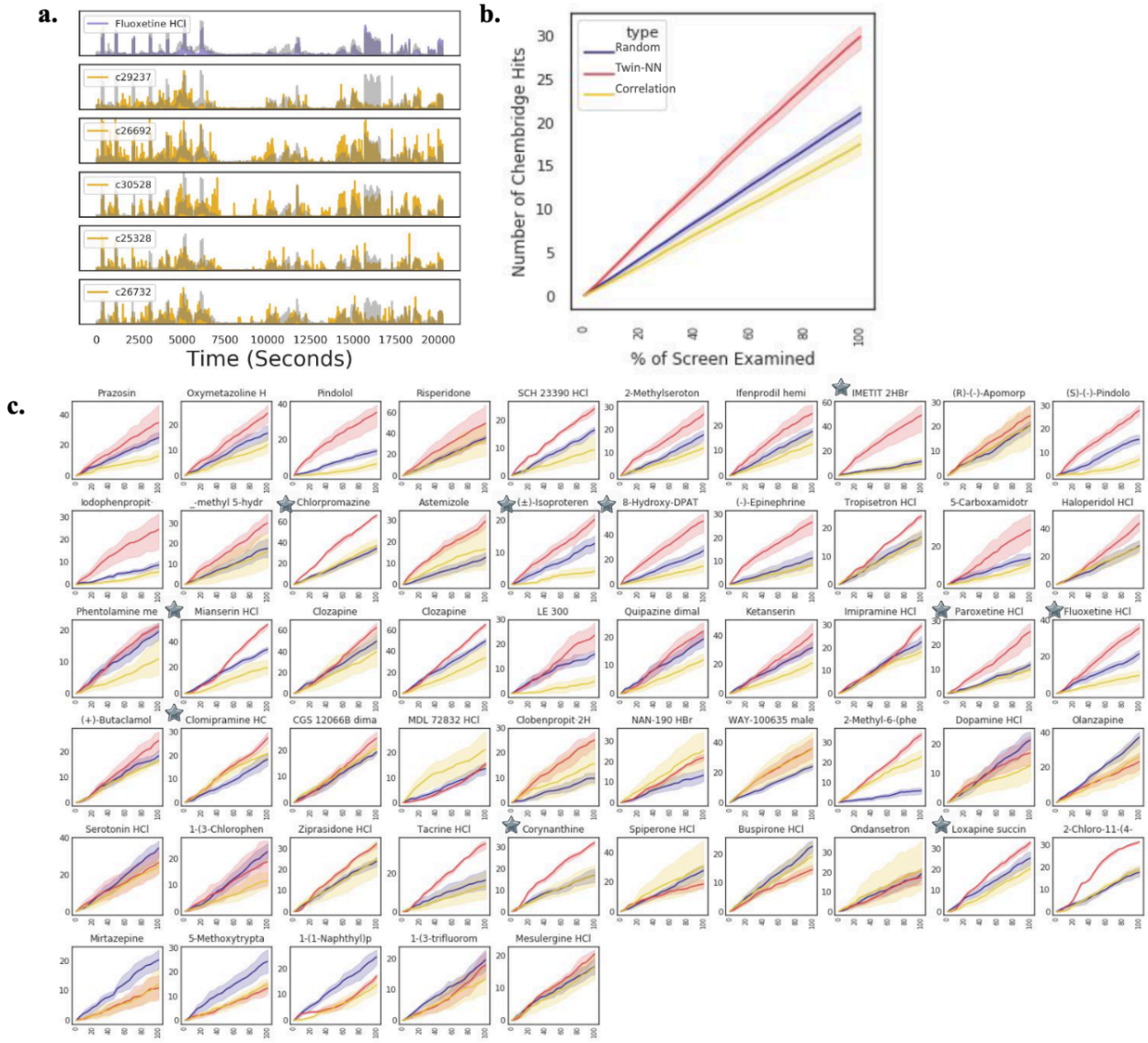


Figure 6

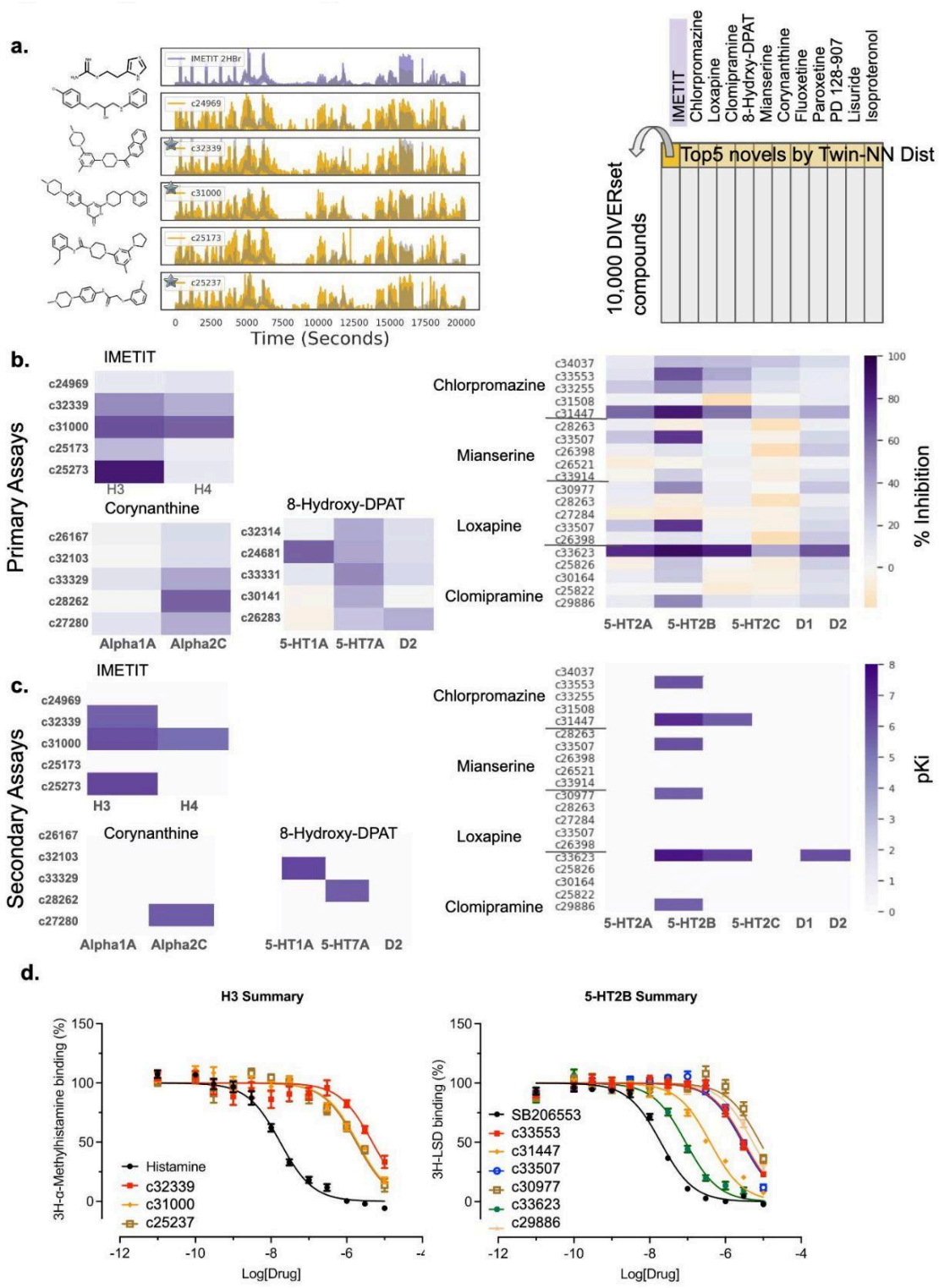
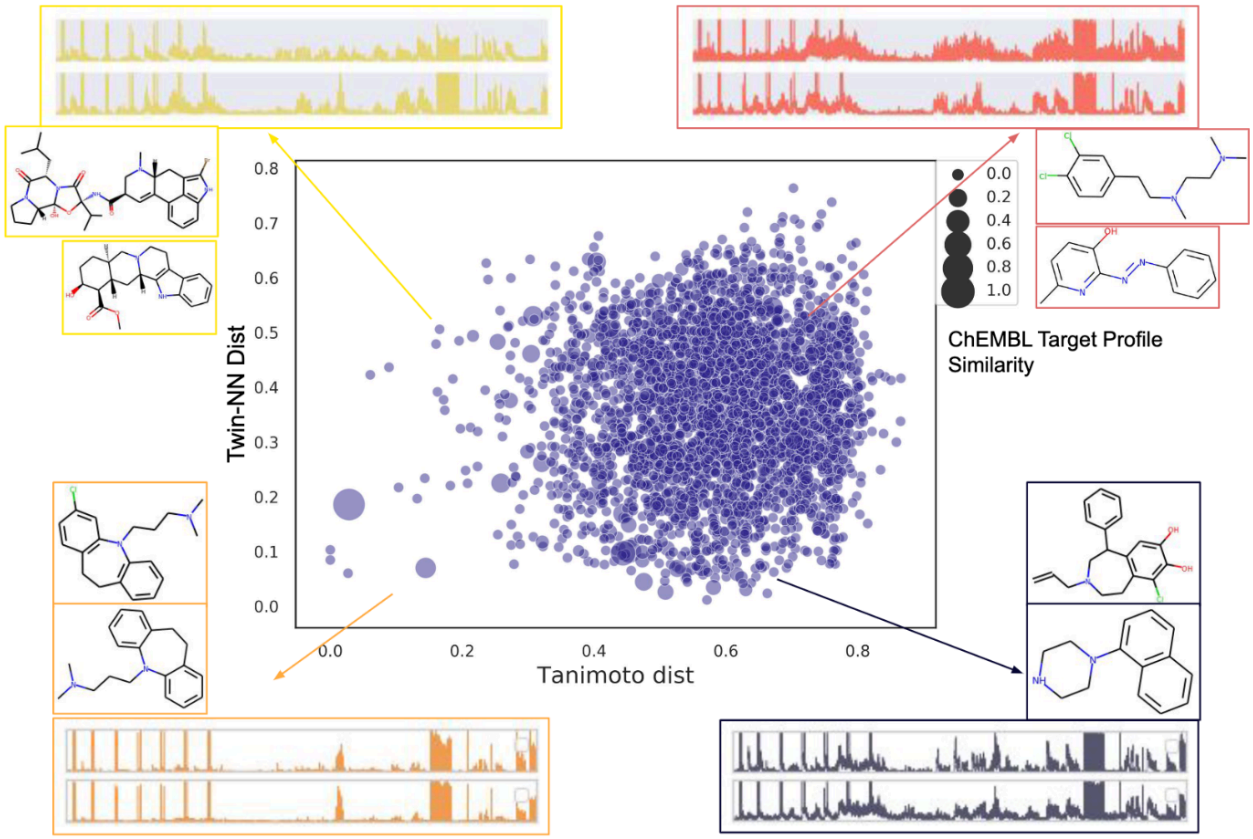


Figure 7



Authors and Affiliations

Institute for Neurodegenerative Diseases, University of California, San Francisco, CA, USA

Leo Gendelev, David Kokel, Michael J. Keiser, Jack Taylor, Douglas Myers-Turnbull

Matthew N. McCarroll

Department of Pharmaceutical Chemistry, University of California, San Francisco, CA, USA

Michael J. Keiser, Matthew N. McCarroll, Michelle R. Arkin, Steven Chen

Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA

Michael J. Keiser

Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA

Michael J. Keiser

UCSF Weill Institute for Neurosciences Memory and Aging Center, University of California, San Francisco, CA, USA

Jack C. Taylor

Author Contributions

Conceptualization, L.G., M.J.K.; Methodology L.G., M.J.K.; Software, L.G., D.M.T.; Validation, L.G., M.J.K.; Formal Analysis, L.G., D.M.T.; Investigation, J.T., D.M.T., M.N.M., S.C.; Resources, D.K., M.R.A.; Data Curation, D.M.T.; Writing - Original Draft, L.G.; Writing - Reviewing and Editing, L.G., D.M.T., and M.J.K., Visualization, L.G.; Supervision, M.J.K., D.K., M.R.A., Project Administration, L.G., M.J.K.; Funding Acquisition, D.K., M.J.K.

Corresponding Authors

Correspondence to David Kokel (dave.kokel@gmail.com) and Michael J. Keiser (keiser@keiserlab.org).

Acknowledgments

This work was supported by grant DAF2018-191905 (https://doi.org/10.37921/550142lkcjzw) from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation (funder https://doi.org/10.13039/100014989) (M.J.K.), the Paul G. Allen Family Foundation (D.K. and M.J.K.), and the US National Institutes of Health (NIH) grant R01AA022583 (D.K.). Binding results (primary and secondary K_i determinations) were generously provided by the National Institute of Mental Health's Psychoactive Drug Screening Program, Contract # HHSN-271-2018-00023-C (NIMH PDSP). The NIMH PDSP is Directed by Bryan L. Roth MD, PhD at the University of North Carolina at Chapel Hill and Project Officer Jamie Driscoll at NIMH, Bethesda MD, USA.

Ethics Declarations

Competing Interests

The authors declare the following competing interests: DK is employed at BioSymetrics, a company that uses AI/ML for drug discovery using phenotypic and diverse data modalities.

Data Availability

All motion index time series data and weights for the trained Twin-NN and Twin-DN models used in this study are available at: https://zenodo.org/records/10652682 (https://zenodo.org/doi/10.5281/zenodo.10652681).

Code Availability

The source code underlying the Twin-NN and Twin-DN models is available at:

https://github.com/keiserlab/deepfish.

References

1. Zanos, P. *et al.* Ketamine and Ketamine Metabolite Pharmacology: Insights into Therapeutic Mechanisms. *Pharmacol. Rev.* **70**, 621–660 (2018).
2. Jeanray, N. *et al.* Phenotype classification of zebrafish embryos by supervised learning. *PLoS One* **10**, e0116989 (2015).
3. Maximino, C. *et al.* Fingerprinting of psychoactive drugs in zebrafish anxiety-like behaviors. *PLoS One* **9**, e103943 (2014).
4. Ali, S., Champagne, D. L. & Richardson, M. K. Behavioral profiling of zebrafish embryos exposed to a panel of 60 water-soluble compounds. *Behav. Brain Res.* **228**, 272–283 (2012).
5. Bandara, S. B. *et al.* Susceptibility of larval zebrafish to the seizurogenic activity of GABA type A receptor antagonists. *Neurotoxicology* **76**, 220–234 (2020).
6. Cheng, D., McCarroll, M. N., Taylor, J. C., Wu, T. & Kokel, D. Identification of compounds producing non-visual photosensation via TRPA1 in zebrafish. 2020.06.10.111203 (2020) doi:10.1101/2020.06.10.111203.
7. Myers-Turnbull, D. *et al.* Simultaneous analysis of neuroactive compounds in zebrafish. *bioRxiv* 2020.01.01.891432 (2022) doi:10.1101/2020.01.01.891432.
8. Jordi, J. *et al.* High-throughput screening for selective appetite modulators: A multibehavioral and translational drug discovery strategy. *Sci Adv* **4**, eaav1966 (2018).
9. Lopez-Luna, J., Al-Jubouri, Q., Al-Nuaimy, W. & Sneddon, L. U. Impact of analgesic drugs on the behavioural responses of larval zebrafish to potentially noxious temperatures. *Appl. Anim. Behav. Sci.* **188**, 97–105 (2017).

10. McCarroll, M. N. *et al.* Zebrafish behavioural profiling identifies GABA and serotonin receptor ligands related to sedation and paradoxical excitation. *Nat. Commun.* **10**, 4078 (2019).
11. Dinday, M. T. & Baraban, S. C. {Large-Scale} {Phenotype-Based} Antiepileptic Drug Screening in a Zebrafish Model of Dravet Syndrome(1,2,3). *eNeuro* **2**, (2015).
12. Peters, J.-U. Polypharmacology - foe or friend? *J. Med. Chem.* **56**, 8955–8971 (2013).
13. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Reviews Drug Discovery* vol. 3 353–359 Preprint at https://doi.org/10.1038/nrd1346 (2004).
14. Prior, M. *et al.* Back to the future with phenotypic screening. *ACS Chem. Neurosci.* **5**, 503–513 (2014).
15. Pham, T.-H., Qiu, Y., Zeng, J., Xie, L. & Zhang, P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nat Mach Intell* **3**, 247–257 (2021).
16. Zoffmann, S. *et al.* Machine learning-powered antibiotics phenotypic drug discovery. *Sci. Rep.* **9**, 5013 (2019).
17. Oprişoreanu, A.-M. *et al.* Drug screening in zebrafish larvae reveals inflammation-related modulators of secondary damage after spinal cord injury in mice. *Theranostics* **13**, 2531–2551 (2023).
18. Baraban, S. C. A zebrafish-centric approach to antiepileptic drug development. *Dis. Model. Mech.* **14**, (2021).
19. Griffin, A., Anvar, M., Hamling, K. & Baraban, S. C. Phenotype-Based Screening of

Synthetic Cannabinoids in a Dravet Syndrome Zebrafish Model. *Front. Pharmacol.* **11**, 464 (2020).

20. Lubin, A. *et al.* A versatile, automated and high-throughput drug screening platform for zebrafish embryos. *Biol. Open* **10**, (2021).
21. Čapek, D. *et al.* EmbryoNet: using deep learning to link embryonic phenotypes to signaling pathways. *Nat. Methods* **20**, 815–823 (2023).
22. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).
23. MacRae, C. A. & Peterson, R. T. Zebrafish as tools for drug discovery. *Nat. Rev. Drug Discov.* **14**, 721–731 (2015).
24. Panula, P. *et al.* The comparative neuroanatomy and neurochemistry of zebrafish CNS systems of relevance to human neuropsychiatric diseases. *Neurobiol. Dis.* **40**, 46–57 (2010).
25. Keiser, M. J. *et al.* Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197–206 (2007).
26. Keiser, M. J. *et al.* Predicting new molecular targets for known drugs. *Nature* **462**, 175–181 (2009).
27. Lounkine, E. *et al.* Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361–367 (2012).
28. Bruni, G. *et al.* Zebrafish behavioral profiling identifies multitarget antipsychotic-like compounds. *Nat. Chem. Biol.* **12**, 559–566 (2016).
29. McCarroll, M. N., Gendelev, L., Keiser, M. J. & Kokel, D. Leveraging Large-scale Behavioral Profiling in Zebrafish to Explore Neuroactive Polypharmacology. *ACS Chem.*

Biol. **11**, 842–849 (2016).

30. McGuirl, M. R., Volkening, A. & Sandstede, B. Topological data analysis of zebrafish patterns. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 5113–5124 (2020).
31. Breier, B. & Onken, A. Analysis of Video Feature Learning in Two-Stream CNNs on the Example of Zebrafish Swim Bout Classification. (2019).
32. Zienkiewicz, A., Barton, D. A. W., Porfiri, M. & di Bernardo, M. Data-driven stochastic modelling of zebrafish locomotion. *J. Math. Biol.* **71**, 1081–1105 (2015).
33. Hughes, G. L. *et al.* Machine learning discriminates a movement disorder in a zebrafish model of Parkinson's disease. *Dis. Model. Mech.* **13**, (2020).
34. Yang, P., Takahashi, H., Murase, M. & Itoh, M. Zebrafish behavior feature recognition using three-dimensional tracking and machine learning. *Sci. Rep.* **11**, 13492 (2021).
35. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R. Signature verification using a 'Siamese' time delay neural network. in *Proceedings of the 6th International Conference on Neural Information Processing Systems* 737–744 (Morgan Kaufmann Publishers Inc., 1993).
36. Koch, G. Siamese Neural Networks for One-Shot Image Recognition. https://www.cs.utoronto.ca/~gkoch/files/msc-thesis.pdf (2015).
37. Koch, G., Zemel, R. & Salakhutdinov, R. Siamese neural networks for one-shot image recognition. in *ICML deep learning workshop* vol. 2 (2015).
38. Nauta, M., Walsh, R., Dubowski, A. & Seifert, C. Uncovering and Correcting Shortcut Learning in Machine Learning Models for Skin Cancer Diagnosis. *Diagnostics (Basel)* **12**, (2021).
39. Böhm, H.-J., Flohr, A. & Stahl, M. Scaffold hopping. *Drug Discov. Today Technol.* **1**,

217–224 (2004).

40. Schneider, G., Schneider, P. & Renner, S. Scaffold-Hopping: How Far Can You Jump. *ChemInform* vol. 38 Preprint at https://doi.org/10.1002/chin.200715270 (2007).
41. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. *arXiv [cs.CV]* (2016).
42. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
43. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv [cs.NE]* (2014).
44. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
45. Elliott, D. F. *Handbook of digital signal processing: Engineering applications*. (Academic Press, 2014). doi:10.1016/c2009-0-21739-9.
46. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
47. Salvador, S. & Chan, P. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* **11**, 561–580 (2007).
48. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–7 (2012).
49. Chu, A. & Wadhwa, R. Selective Serotonin Reuptake Inhibitors. in *StatPearls* (StatPearls Publishing, 2022).
50. DrugMatrix/ToxFX. https://ntp.niehs.nih.gov/data/drugmatrix/.

51. Hajjo, R. *et al.* Development, validation, and use of quantitative structure-activity relationship models of 5-hydroxytryptamine (2B) receptor ligands to identify novel receptor binders and putative valvulopathic compounds among common drugs. *J. Med. Chem.* **53**, 7573–7586 (2010).

52. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

53. Johnson, M. A. & Maggiora, G. M. *Concepts and Applications of Molecular Similarity*. (Wiley, 1990).

54. Cruz-Monteagudo, M. *et al.* Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* **19**, 1069–1080 (2014).

55. Zhu, T. *et al.* Hit identification and optimization in virtual screening: practical recommendations based on a critical literature analysis. *J. Med. Chem.* **56**, 6560–6572 (2013).

56. Goodfellow, I. J. *et al.* Generative Adversarial Networks. *arXiv [stat.ML]* (2014).

57. Upadhyay, U. & Jain, A. Removal of Batch Effects using Generative Adversarial Networks. *arXiv [cs.LG]* (2019).

58. Anatomical therapeutic chemical (ATC) classification. https://www.who.int/tools/atc-ddd-toolkit/atc-classification.

59. Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl Sci Rev* **5**, 44–53 (2017).

60. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

61. fastdtw. *PyPI* https://pypi.org/project/fastdtw/.

62. Fix, E. & Hodges, J. L. Discriminatory analysis: Nonparametric discrimination: Consistency properties. *PsycEXTRA Dataset* Preprint at https://doi.org/10.1037/e471672008-001 (1951).

63. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

64. Barnea, G. *et al.* The genetic design of signaling cascades to record receptor activation. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 64–69 (2008).

65. Xiao, Y. *et al.* Rat alpha3/beta4 subtype of neuronal nicotinic acetylcholine receptor stably expressed in a transfected cell line: pharmacology of ligand binding and function. *Mol. Pharmacol.* **54**, 322–333 (1998).

66. Xiao, Y. *et al.* Sazetidine-A, A Novel Ligand That Desensitizes α4β2 Nicotinic Acetylcholine Receptors without Activating Them. *Mol. Pharmacol.* **70**, 1454–1460 (2006).

67. Chopra, S., Hadsell, R. & LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* vol. 1 539–546 vol. 1 (IEEE, 2005).