Behavioral fingerprints predict insecticide and anthelmintic mode of action

Adam McDermott-Rouse^{1,2*}, Eleni Minga^{1,2*}, Ida Barlow^{1,2}, Luigi Feriani^{1,2}, Philippa H Harlow³, Anthony J Flemming³, André EX Brown^{1,2}

¹MRC London Institute of Medical Sciences, London, UK

²Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London, UK

³Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire, UK

andre.brown@imperial.ac.uk

*These authors contributed equally

Abstract

Novel invertebrate-killing compounds are required in agriculture and medicine to overcome resistance to existing treatments. Because insecticides and anthelmintics are discovered in phenotypic screens, a crucial step in the discovery process is determining the mode of action of hits. Visible whole-organism symptoms are combined with molecular and physiological data to determine mode of action. However, manual symptomology is laborious and requires symptoms that are strong enough to see by eye. Here we use high-throughput imaging and quantitative phenotyping to measure *C. elegans* behavioral responses to compounds and train a classifier that predicts mode of action with an accuracy of 88% for a set of ten common modes of action. We also classify compounds within each mode of action to discover pharmacological relationships that are not captured in broad mode of action labels. High-throughput imaging and automated phenotyping could therefore accelerate mode of action discovery in invertebrate-targeting compound development and help to refine mode of action categories.

Introduction

Invertebrate pests including insects, mites, and nematodes damage crops, decrease livestock productivity, and cause disease in humans. Nematodes alone infect over 1 billion people and lead to the loss of 5 million disability adjusted life years annually¹. In livestock, they infect sheep, goats, cattle and horses causing gastroenteritis that leads to diarrhea, reduced growth, and weight loss. Nematodes that parasitize crops have been estimated to cause well over \$100 billion in annual crop losses². Crop loss due to insects is measured in tens of metric megatons and is predicted to increase due to climate change³. Compounds that kill or impair invertebrates are one of the primary means of defense in human and veterinary medicine and in crop protection. However, resistance is widespread in nematodes and insects and drives continuing efforts to discover new invertebrate-targeting compounds^{4,5}.

To date, most currently approved treatments for infections in humans and livestock as well as for crop protection in the field have been discovered through phenotypic screens^{6,7}. That is, compounds are first screened for the ability to kill or impair a target species without any hypothesized molecular target. A critical problem is then determining hit compounds' mode of action which is important for understanding resistance mechanisms, avoiding pathways where resistance is already common, and subsequent lead optimization. Despite advances in biochemical and genetic methods for determining mode of action, direct observation of the symptoms induced by compounds remains a key step in mode of action discovery⁶. Because most insecticides and anthelmintics target the neuromuscular system, behavioral symptoms are a particularly important class of phenotypes to consider, but manual observation of behavior is time consuming, insensitive to subtle phenotypes, and prone to inter-operator variability and bias⁸. We therefore sought to develop more automated and quantitative methods to do mode of action prediction from phenotypic screens of freely behaving invertebrates.

Pioneering work in zebrafish showed that behavioral fingerprints can be used to discover neuroactive compounds and that behavioral fingerprints correlate with compound mode of action^{9–11}. However, this approach has not yet been applied to invertebrate animals—the targets of insecticides and anthelmintics—at a large scale. Furthermore, although previous zebrafish screens were high throughput, their spatial resolution was low and phenotypes were limited to activity levels in response to stimuli. Recent work in computational ethology has shown the power of moving beyond point representations of animal behavior to include information on posture^{12–15}. From previous symptomology work, it is clear that detailed postural information can be useful for resolving mode of action^{16,17}. We chose *C. elegans* as our model system because it is small and compatible with multiwell plates and automated liquid handling. It is sensitive to anthelmintics and insecticides and has played an important role in mode of action discovery in the past^{16,18–21}.

To combine the benefits of high throughput and high resolution, we used megapixel camera arrays to record the behavioral responses of worms to a library of 110 compounds covering 22 distinct modes of action. We simultaneously recorded all of the wells of 96-well plates with sufficient resolution to extract the pose of each animal and a high-dimensional behavioral fingerprint that captures aspects of posture, motion, and path. We show that worms have diverse dose-dependent behavioral responses to insecticides and anthelmintics and develop a machine learning approach that shares information across replicates and doses to accurately predict the mode of action of previously unseen test compounds. Furthermore, we show that a novelty detection algorithm can be used to identify compounds with modes of action not contained in the training data suggesting a way to prioritize lead compounds with

novel modes of action early in the development process. These results demonstrate that high-throughput phenotyping in *C. elegans* is a promising approach for assisting target deconvolution in anthelmintic and insecticide discovery. Finally, we show that our prediction accuracy may not be limited by noise or phenotypic dimensionality. Rather, we show that we can classify compounds even within a mode of action class, suggesting that there are limitations in our knowledge of the relevant pharmacology rather than limitations in our ability to reproducibly detect compound-induced phenotypes.

Results

Insecticides affect phenotypes in multiple behavioral dimensions

We assembled a library of 110 insecticides and anthelmintics with diverse targets to sample a range of modes of action used medically and commercially (see Table S1 for full list). The modes of action represented in the library cover 70% of the market of insecticides used in the field⁵ and several important classes of anthelmintics used in veterinary and human medicine⁴. To quantify the effects of the compounds on behavior we recorded worms using megapixel camera arrays that simultaneously image all of the wells of 96-well plates (Fig. 1a). We recorded at least 10 replicates at 3 doses for each compound with enough resolution to extract high-dimensional behavioral fingerprints following segmentation, pose estimation, and tracking (Fig. 1a). The behavioral fingerprints are vectors of posture and motion features that are subdivided by body segment and motion state including, for example, ‘midbody curvature during forward crawling’ or ‘angular velocity of the head with respect to the tail while the worm is paused’. We have previously shown that similar features can detect even subtle behavioral differences that can be difficult to detect by eye²² and that the combined feature set has sufficient dimensionality to accurately classify worms with diverse behavioral differences caused by genetic variation and optogenetic perturbation^{23,24}.

As expected, several compound classes have strong visible effects on *C. elegans* behavior including, for example, the glutamate-gated chloride channel activator emamectin benzoate, the spiroindoline vesicular acetylcholine transporter inhibitor SY1713, and the serotonin receptor antagonist mianserin. All three compounds can be distinguished from DMSO controls and from each other in a simple two-dimensional space defined by speed and body curvature (Fig. 1b). The large differences in curvature and in motion caused by some compounds are observable by eye, as shown in the inset images and in Fig. 1c. However, not all compounds are well separated in these two dimensions (gray points, Fig. 1b). To find which features are different for the full compound set, we compared the behavioral fingerprints of treated worms with DMSO controls for each feature and corrected for multiple comparisons using the Benjamini-Yekutieli procedure²⁵. To account for random day-to-day variation in the experiments, we used a Linear Mixed Model for these statistical tests, where the fixed effect is the drug dose and the day of the experiment is added as a random effect. The number of features that are significantly different at a false discovery rate of 1% between the behavior of worms treated with each compound and the DMSO controls is summarized in a heatmap (Fig. 1d).

To further increase the dimensionality of the behavioral phenotypes, we included a blue light stimulation protocol. Each tracking experiment is divided into three parts: 1) a 5-minute pre-stimulus recording, 2) a 6-minute stimulus recording with three 10-second blue light pulses starting at 60, 160, and 260 seconds, and 3) a 5-minute post-stimulus recording. Blue light is

Figure 1: Insecticides affect phenotypes in multiple behavioral dimensions. **A)** We image the entire 96 well plates with a megapixel camera array with enough resolution and high enough framerate to track, segment, and estimate the posture of *C. elegans* over time. **B)** Several compounds, including the serotonin receptor antagonist mianserin (blue), glutamate-gated chloride channel activator emamectin benzoate (purple), and vesicular acetylcholine transporter inhibitor SY1713 (red), have a strong effect on the worms' behavioral phenotype. They can be distinguished from the DMSO control (black) and from each other based on speed and tail curvature alone. Not all compounds are well separated in these two dimensions (gray points). Inset images are samples that show postural differences. **C)** Sample worm skeletons over time show the effect of the compounds highlighted in b) on motion. **D)** Number of features significantly different (at a false discovery rate of 1%) from the DMSO control for each compound, grouped by mode of action. The pre-stimulus, blue light stimulus, and post-stimulus data are shown separately. The percentage of significantly different features is highest for the blue light stimulus recording.

aversive to *C. elegans*²⁶ and so it can help to distinguish between animals that are simply pausing and those that are not able to move²⁷. Behavioral differences are observed in each assay period but the stimulus period shows the most differences (Fig. 1d). Even within mode of action classes, compound potency can be highly variable. The largest potency difference is observed for the Octopamin agonists where amidine affects 0.08% of features and oxazoline affects 75% of features. Overall, 86% of compounds have a detectable effect on behavior in at least one feature. Compounds with no detectable effect were not included in subsequent analysis.

Compounds with the same mode of action have similar effects on behavior

Having established that *C. elegans* shows diverse behavioral responses to insecticides and anthelmintics, we next sought to determine to what extent the responses are mode of action specific. For all clustering and classification tasks we first z-normalize each feature to put them on a common scale and to prevent arbitrary choices of units from impacting the analysis. We used hierarchical clustering to visualize the relationships between the behavioral

Figure 2: Cluster and dose response of behavioral fingerprints. **A)** Hierarchical clustering of behavioral fingerprints highlights structure in the responses to different compounds. Clear clusters can be observed for some compound classes, e.g. AChE inhibitors, vAChT inhibitors, GluCl agonists, and mAChR agonists. **B)** Cluster purity as a function of the hierarchical cluster distance shows that the degree of mode of action clustering (red) is greater than expected by chance for random clusters (gray). **C)** Compounds in the same class can have different dose-response curves. The three mitochondrial inhibitors cyazofamid, rotenone, and SY1048 all decrease angular velocity, but the concentration at which their effect is measurable is not conserved across compounds. Different spiroindolines affect body curvature differently. SY1786's dose-response curve is non-monotonic.

responses to different compounds (Fig. 2a). Several of the compound classes show clear clustering, including the AChE inhibitors, vAChT inhibitors, GluCl agonists, and mAChR agonists. The degree of mode of action clustering is greater than expected by chance, which can be seen in a plot of the cluster purity observed in the data compared to random clustering (Fig. 2b). However, the distance between compounds that share the same mode of action can be large, even for classes that cluster well over all, in part because behavioral fingerprints change with dose. It is also not always possible to align feature vectors using doses because compounds can have very different potencies: a low dose for one compound could be a high dose for another.

These effects can be seen in dose-response plots for individual features. The three mitochondrial inhibitors in Fig. 2c all decrease angular velocity, but they do it at different doses. At 3 μ M, only SY1048 has a strong effect, while at 30 μ M, rotenone has a similarly strong effect. Clustering based on angular velocity would lead to qualitatively different conclusions about nearest neighbors at these different doses. For the spiroindolines, similar differences in dose-response are observed for body curvature with the added difference that the effect of SY1786 is non-monotonic and returns to baseline at high doses. These non-monotonic effects can be due to compounds precipitating from solution at high doses or due to intrinsically complex compound effects such as a compound that causes an increase in speed at low doses but is lethal at high doses. Regardless of the cause, complex dose-response curves present challenges for mode of action prediction since supervised machine learning algorithms rely on differences in feature distributions to learn decision boundaries and dose-response effects spread out the distributions and increase the overlap between classes.

Combining classifiers by voting enables mode of action prediction

We take advantage of the fact that several replicates are recorded per condition and resample with replacement from the multiple replicates for each dose to create a set of average behavioral fingerprints. This effectively smooths the data and provides a simple method for balancing classes before classifier training. For classes with fewer compounds, we resample more times so that each class contains the same number of points. To partially mitigate the effect of compound potency, we then normalize each behavioral fingerprint to unit magnitude. This normalization is done row-wise on each sample in contrast to the z-normalization described above which is done column-wise on each feature. Rescaling in this way brings compounds with similar effect profiles but different potencies closer together in feature space (Fig. 3a), but because of nonlinearities in the dose response profiles, the overlap is not perfect even after rescaling.

Predictions must be combined across doses and replicates to make a single prediction for the mode of action of a given compound. Inspired by an analogy with the multi-sensor fusion problem²⁸, we use a voting procedure to make a final prediction. However, in contrast to multi-sensor fusion, we cannot train different classifiers for each dose because 1 μ M for one compound is not equivalent to 1 μ M for another compound. Instead, we train a single classifier for all doses and make predictions for each replicate. Each replicate contributes a vote for a compound's class and the class with the most votes wins.

We split our data into a training/tuning set consisting of 60 compounds and a hold-out/reporting dataset consisting of 16 compounds containing at least one compound for each mode of action class. We used the training set to determine an appropriate classifier, select

Figure 3: Classifiers trained on behavioral fingerprints can predict the mode of action of unseen test compounds. **A)** Toy data illustrating the potential benefit of normalization in correcting for potency differences within mode of action classes. Following normalization, each behavioral fingerprints exists on a sphere in phenotype space regardless of effect size in the original space. Nonlinear dose-response curves will not collapse perfectly following normalization which is a linear transformation. **B)** The confusion matrix for the multinomial logistic regression classifier following hyper-parameter tuning and feature selection on the training data. **C)** The confusion matrix for the classifier trained in B applied to previously unseen test compounds without any further tuning. **D)** The novelty score assigned to novel compounds with a mode of action not seen during training compared to the novelty score of compounds from the test set in c). Novel compounds tend to have higher novelty scores than compounds from previously seen modes of action. The non-novel compounds with high novelty scores include the two incorrectly classified test compounds (in red box).

features, and tune hyperparameters using cross-validation. We achieved the highest cross-validation accuracy using multinomial logistic regression with 1024 features selected using recursive feature elimination (Fig. 3b). To determine whether the classifier could generalise to unseen compounds, we applied it to the test data without further tuning. The classifier predicted the correct mode of action for the unseen compounds 88% of the time (Fig. 3c).

In addition to the ten modes of action in our dataset that were represented by at least five compounds, we also included 19 compounds representing an additional 11 modes of action. We used these additional compounds to simulate another use-case for our approach: detecting screening hits that represent potentially novel modes of action that do not fall into known

classes. Using a novelty detection algorithm²⁹, we assigned a novelty score to each of the new compounds based on their affinity to each of the existing classes. To obtain the novelty score we use an ensemble of support vector machine (SVM) classifiers that flag novel compounds based on the confidence values of the main multinomial logistic regression classifier used for the predictions of known classes. The ensemble of SVM classifiers is trained using partitions of the training set into presumed-known and presumed-unknown classes. The novelty score is defined as the weighted average of the output of this ensemble. Most of the truly novel compounds were assigned novelty scores above 0.8. Several of the non-novel compounds—those that come from a class that is present in the training data—have high novelty scores but this includes the two test compounds that were incorrectly classified. To explore the origin of the high novelty score for the incorrectly classified compounds, we looked for differences between the effects of compounds within a class.

Mode of action deconvolution within classes

Although compounds are categorized into broad mode of action classes, most compounds will have some degree of off-target engagement. If the off-target effects are different for compounds within a mode of action class or if the compounds have differences in pharmacokinetics, they may lead to different phenotypes. In this case, it may be possible to use behavioral fingerprinting to further deconvolve mode of action classes revealing hidden compound heterogeneity. To test for phenotypic differences within mode of action classes, we trained a classifier to distinguish the replicates from each compound within a class from the replicates of the other compounds in the class. We then used cross-validation accuracy to quantify the distinguishability of the compounds within a class. We hypothesized that for classes without mechanistic sub-classes, the classifier would perform similarly to random guessing. In contrast, if compounds with the same mode of action had different off-target profiles or different pharmacokinetics, the classifier would be able to reliably distinguish individual compounds or subsets of compounds with the same broad mode of action.

In all classes, the compound-level classifiers performed better than random guessing, in some cases by a large margin. One of the incorrectly classified compounds in the test set was ritanserin, which was also assigned a high novelty score. The within-class classifier shows that it is indeed clearly distinguishable from the other 5-HT receptor antagonists (Fig. 4a). Although ritanserin is known to be a 5-HT receptor antagonist, it is also known to affect multiple other targets. In addition to detecting outlying compounds, the deviations from random guessing revealed substructures within the classes. For example, one group contains the two antidepressants, which are nearest neighbors in terms of structural similarity (atom pair Tanimoto coefficient of 0.44 between the antidepressants compared to a Tanimoto coefficient of 0.20 ± 0.09 (mean $\pm$ standard deviation) for the other pairwise comparisons within the class). As with ritanserin, the other four compounds in this class have known polypharmacology which could be driving their clustering. Another class with interesting substructure is that consisting of mitochondrial inhibitors which also separates into distinguishable groups. In this case, the phenotypically distinct groups separate the complex I inhibitors from the complex II and III inhibitors which appear phenotypically more similar. The mectins we tested are structurally similar and are known to share the same binding site, suggesting they would be difficult to separate into subgroups. Consistent with this expectation, the compound-level mectin classifier performs only slightly better than chance (Fig. S1).

Figure 4: Mode of action can be resolved within compound classes. **A)** The confusion matrix showing cross-validation performance of a classifier trained to distinguish serotonin receptor antagonists from each other. Ritanserin is distinguishable from all other compounds and the two structurally similar antidepressants (mianserin and methiothepin) are somewhat mutually confused by the classifier but are distinguishable from the non-antidepressants. **B)** The confusion matrix for the mitochondrial inhibitors also shows some substructure: complex II and III inhibitors (cyazofamid, antimycin, cyenopyrafen) are phenotypically similar, and distinct from the complex I inhibitors.

Discussion

We have shown that worms have diverse responses to insecticides and anthelmintics and that behavioral fingerprints can be used to cluster compounds with similar modes of action. With appropriate normalization and by combining information across doses and replicates through voting, we can also accurately predict the mode of action of previously unseen compounds. Given the wide variety of modes of action included in this study, it might be expected that the dimensionality of worm behavior would be limiting. In other words, it could have been the case that worm behavior can only vary in a small number of ways and that this would limit the number of distinct classes that are distinguishable using tracking data alone. On the contrary, we found that in addition to predicting mode of action across ten classes, we could often classify at the level of individual compounds and that the within-mode-of-action confusion matrices revealed subgroups that related to finer-scale mechanistic divisions. This finding suggests there is sufficient dimensionality in worm behavioral responses to detect mode of action as well as more detailed pharmacological differences such as the spectrum of a compound's off-target effects or differences in target engagement. These results belie worms' superficial simplicity but are consistent with genetic findings that mutations can lead

to multiple types of uncoordination¹⁸ and with tracking results suggesting that even wild type spontaneous locomotion is surprisingly complex^{30,31}.

The most obvious limitation of our method is that not all compounds affect the N2 strain of *C. elegans*. Only a small minority of the compounds we assayed had no detectable effect, but there will be entire classes of compounds that are not expected to have an effect on *C. elegans* because their targets are not conserved, such as the pyrethroid insecticides that target voltage-gated sodium channels³². Expanding the range of organisms included in the training data is one way to address this limitation. There are methods for deriving multidimensional behavioral fingerprints that incorporate postural information from flies³³ and zebrafish larvae³⁴ and both organisms are compatible with high throughput screening. Because our approach already involves the fusion of data from multiple samples, additional species level classifiers could be seamlessly incorporated into the voting procedure to arrive at a final prediction. Results from ensemble learning in diverse fields suggest a further benefit of a multi-species approach: increasing the votes from independent classifiers should increase classifier accuracy if the predictions from different species are partially uncorrelated³⁵. The sample principle applies beyond behavioral phenotypes. The results of automated symptomology can also be combined—within the same analysis framework—with data from non-animal species including bacteria and fungi as well as other assays commonly used in mode of action identification including genetic and biochemical assays.

Strains of *C. elegans* that have been recently isolated from the wild are readily available³⁶ and have given insight into anthelmintic resistance³⁷. Using wild isolated strains of *C. elegans* would increase the diversity of compound responses without having to modify the experimental protocols for screening or downstream analysis. Moving beyond *C. elegans* to other nematode species, including parasitic nematodes, could provide further independent phenotypes for improving mode of action prediction. However, particularly for animal parasites, different morphologies might require alternative analysis approaches^{21,38,39}.

Phenotypes derived from *in vitro* microscopy of cells in culture have also been extensively studied for mode of action prediction^{40–42}. Human cells in culture would provide an additional phylogenetically distinct species to combine with invertebrate screening. Human cell responses may also give an indication of mammalian toxicity⁴³. However, there are drawbacks with respect to insecticide and anthelmintic discovery. The first, which is shared with zebrafish, is that the goal of these compounds is to affect invertebrates without affecting vertebrates so there may be a lower response rate than for pharmaceutical compounds. Cultured insect cells⁴⁴ could be used in similar assays as human cells but may provide more relevant information for insecticide mode of action. The second drawback, which applies to most culture systems, is that some compounds act on super-cellular structures. For example, acetylcholinesterase inhibitors act by causing a buildup of acetylcholine at synapses or neuromuscular junctions. Whole-animal phenotypic screening is therefore likely to continue to play a role in phenotypic screening of neuroactive compounds both for discovery and mode of action prediction.

Symptomology is an important technique for mode of action determination in insecticide discovery. Our method would therefore fit straightforwardly into existing discovery pipelines. Since phenotypic screening in pest species is the primary means of lead identification, an intriguing possibility would be to adapt our method to primary screening data so that mode of action prediction is not only available for selected hits but can already be included at the earliest stages of decision making. Given advances in computer vision and

pose estimation^{45–47}, it should be possible to track pest species in complex media such as leaf sections with existing technology. Improved phenotyping in primary screens might also reduce false negative rates by picking up compounds with unique phenotypic effects that are too weak to register as hits using current metrics but might provide useful starting points for optimization.

Methods

Worm husbandry

The N2 Bristol *C. elegans* strain was obtained from the CGC (*Caenorhabditis* Genetics Center) and cultured on Nematode Growth Medium at 20°C and fed with *E. coli* (OP50) following standard procedures¹⁸.

Worm preparation

Synchronized populations of young adult worms for imaging were cultured by bleaching unsynchronized gravid adults, and allowing refed L1 diapause progeny to develop for two and a half days (detailed protocol: <https://dx.doi.org/10.17504/protocols.io.2bzgap6>). On the day of imaging, young adults were washed in M9 (detailed protocol: <https://dx.doi.org/10.17504/protocols.io.bfqbjmsn>) and transferred to the prepared drug plates (3 worms per well) via the COPAS 500 Flow Pilot (detailed protocol: <https://dx.doi.org/10.17504/protocols.io.bfc9jiz6>) and returned to a 20°C incubator for 3.5 hours. Plates were then transferred onto the multi-camera tracker for another 30 minutes to habituate prior to imaging so that the total drug exposure time was 4 hours.

Plate preparation

Low peptone (0.013%) nematode growth medium (detailed protocol: <https://dx.doi.org/10.17504/protocols.io.2rcgd2w>) was prepared as follows: 20g agar (Difco), 0.13g Bactopeptone, and 3g NaCl were dissolved in 975ml of milliQ water. After autoclaving, 1ml of 10mg/ml cholesterol was added along with 1ml of 1M CaCl₂, 1ml 1M MgSO₄ and 25ml of 1M KPO₄ buffer (pH 6.0). Molten agar was cooled to 50–60°C and 200 µl was dispensed into each well of 96-square well plates (WHAT-7701651) using an Integra VIAFILL (detailed protocol: <https://dx.doi.org/10.17504/protocols.io.bmxbk7in>). Poured plates were stored agar-side up at 4°C until required.

Prior to applying compounds, plates were placed without lids in a drying cabinet to lose 3–5% weight by volume and then stored with lids (WHAT-77041001) on at room temperature until required.

Compound preparation

Compounds were prepared for screening by dissolving in DMSO at 1000X their final imaging plate concentration (so that final concentration of DMSO in imaging plates was 0.01%). The results presented here are pooled from two rounds of experiments that used two different methods of adding compounds.

For the first round (December 2019), the library was stored in 56 ‘master’ 96-well plates in which eight replicates of each compound solution were in a single column of a 96-well plate and lower concentrations were made by serial dilution of the highest concentration using DMSO. There were up to six doses per drug and one column was reserved for DMSO and one column for no compound controls on each master plate, which were stored at -20°C. The day prior to imaging, ‘source’ plates were prepared using an Integra VIAFLO 96-

channel pipette by adding 7 μ l water and 0.5 μ l 1000X drug and mixing by pipetting up and down. 5 μ l water was added to the surface of each well of an imaging plate to facilitate compound transfer and prevent agar damage from pipette contact. Then, 3 μ l of compound was transferred from the source plates to the destination imaging plates using an Opentrons liquid handling robot, which randomly shuffled the column order during the transfer.

For the second round (January 2020), ‘master’ 96-well plates were filled sequentially with three doses of each drug (made by serial dilution of the highest concentration using DMSO) so that the entire library fitted into three and a half 96-well plates. Three sets of shuffled ‘library’ plates were created with randomized column orders using an Opentrons robot so that three replicates per drug per dose fitted into 10.5 96-well plates. All plates were stored at -20°C. The day prior to imaging, ‘source’ plates were prepared using the VIAFLO by transferring 1.4 μ l of 1000X drug from the shuffled ‘library’ plates into 96 well (PCR) plates filled with 19 μ l of water and pipetted up and down to mix. 5 μ l of water was added to the imaging plates to facilitate compound transfer and then 3 μ l of the pre-diluted compounds were then transferred to imaging plates using the VIAFLO.

After the drug had absorbed into the agar, imaging plates were seeded with 5 μ l OP50 diluted 1:10 in M9 solution using an Integra VIAFILL, and left overnight at room temperature in the dark. Full protocols are available at <https://doi.org/10.17504/protocols.io.9vqh65w>, <https://doi.org/10.17504/protocols.io.bn5zmg76>.

Image acquisition

Image acquisition was performed on custom-built tracking rigs (LoopBio GMBH, Vienna) equipped with six Basler acA4024 cameras to simultaneously image all wells of a 96-well plate. Each tracking experiment is divided into three parts, which are run in series by a script: 1) a 5-minute pre-stimulus recording, 2) a 6-minute stimulus recording with three 10-second blue light pulses starting at 60, 160, and 260 seconds, and 3) a 5-minute post-stimulus recording. Blue light was provided using four Luminus CBT-90 TE light-emitting diodes with a peak wavelength of 456nm and peak radiometric flux of 10.3W each.

Image processing and quality control

Segmentation and skeletonization were performed using Tierpsy tracker²³ (<https://github.com/Tierpsy/tierpsy-tracker>). Each video was manually checked using the Tierpsy tracker viewer and any wells that had precipitation, excess unabsorbed liquid that led to swimming worms, or damaged agar were marked as bad and excluded from the analysis.

Feature extraction and pre-processing

Python code to reproduce the analysis in the paper is available on GitHub (<https://github.com/em812/moaclassification>). We used Tierpsy tracker to obtain summarized behavioural features for each screened well²⁴. These include morphological features such as length and width, postural features such as curvature and features describing movement such as speed and angular velocity. A total of 3020 features is obtained. We derived summarized features separately for the pre-stimulus period, for the period with blue light stimuli and for the post-stimulus period resulting in a total of 9060 features for each well.

In addition to the manual quality control described above we used the filtering options in Tierpsy tracker to filter out tracked objects that have average width smaller than 20 μ m or larger than 500 μ m and average length smaller than 200 μ m or larger than 2000 μ m. Features

that have more than 5% NaN values in the entire dataset are removed entirely. Any remaining NaN values are imputed to the mean of the given feature calculated across all wells. Compounds with very low effect compared to DMSO are also removed from the dataset. A compound is considered to have very low effect when none of the features shows significant differences across doses (including DMSO as dose 0) in a statistical test based on a Linear Mixed Model with random intercept where the dose is the fixed effect and the day of the experiment is the random effect. The Benjamini-Yekutieli method²⁵ with 1% false discovery rate was used to correct for multiple comparisons. After cleaning up the dataset, we standardize the features (to mean 0 and standard deviation 1) to bring them on a common scale and prevent the unit differences from influencing the analysis.

For the classification task, we use a bootstrap method that simultaneously smooths the data and balances the classes. We bootstrap the compound dose replicates (i.e. we resample with replacement until we get a sample 0.6 times the size of the initial sample) multiple times and each time we derive the mean feature values. In this way, we replace the drug dose replicates with bootstrapped averages which smooths the data reducing the effect of outliers. This method also gives an easy way to balance the classes. The number of bootstrapped averages per dose in the training set depends on how well-populated the mode of action class is. In the class with the most members (AChE inhibitors with 9 compounds), we get 20 bootstrap averages per drug dose. In the rest of the classes, the number of averages is proportional to the ratio between the max number of members in a class (9) and the number of members in the given class. Finally, to partially mitigate the effect of compound potency, we normalize each behavioral fingerprint to unit L2-norm. Rescaling in this way brings compounds with similar effect profiles across features but different potencies closer together in feature space.

Hierarchical clustering

To investigate how well the compounds of the same mode of action cluster together in an unsupervised way, we use hierarchical clustering. For this task, we create a matrix where each row is the average of a drug dose and each column is one of the 256 features in the tierpsy_256 feature set²⁴. We also include 5 rows with the average of 5 random partitions of the DMSO control replicates. We use the hierarchical clustering algorithm implemented in the seaborn⁴⁸ clustermap function with complete linkage and cosine distances. To assess the quality of the clustering, we use the row linkage dendrogram and compare the cluster purity at each level of the dendrogram with the purity of random clusters. To get the random clusters we permute the cluster labels derived from the dendrogram. We repeat the permutation 1000 times.

Train, test, novel split

For the classification task, we created a training set and a test set. We considered that we need at least 4 compounds per class in the training set for cross validation, so classes with less than 5 compounds were not included. As a result, we used ten classes for the classification task. The split in training set and test set was done using stratified split as implemented in scikit-learn⁴⁹ with 20% of the compounds of each class included in the test set. Using this method, 60 compounds were assigned to the training set, while 16 compounds spanning all nine classes were assigned to the test set. The compounds of the sparsely-populated classes that are not included in the classification task were considered ‘novel’ compounds and were used for novelty detection. Eleven classes with a total of 17 compounds are included in the novel compounds dataset. The training set was used for feature selection and hyperparameter selection. The test set was only used to test the classification accuracy of the selected trained model.

Classification by mode of action

The classification of compounds into modes of action is a classification of ‘bags’ of data points, as we know that we have the same label across doses and across replicates of the same compound. To get compound-level predictions, we use the following approach. We train a single classifier with the rescaled bootstrapped averages from all the compounds and then we combine the predictions across all the data points of the same compound with a simple voting procedure. Each bootstrapped average point contributes one vote for the class of the compound. The predicted class for the compound is the one with the most votes. We initially tested three types of classifiers: logistic regression, random forest and XGBoost. Logistic regression performed better than the ensemble methods in terms of cross-validation accuracy in the training set, so it was adopted for the entire pipeline.

As we have thousands of highly correlated features, we performed feature selection with recursive feature elimination using the training set. The estimator used was multinomial logistic regression with elastic net. We tested different sizes of selected feature sets and chose the set with 1024 features which resulted in the highest 4-fold cross-validation accuracy. The full list of selected features is shown in Table S2.

Using only the 1024 selected features, we optimized the hyperparameters of the estimator using stratified 4-fold cross validation. We tested the following parameter grid:

Parameter	Values option 1	Values option 2
penalty	'l1', 'l2'	'elasticnet'
C	0.01, 0.1, 1, 10, 100, 1000	0.01, 0.1, 1, 10, 100, 1000
l1_ratio		0.1, 0.5, 0.9
multi_class	'multinomial', 'ovr'	'multinomial', 'ovr'
solver	'saga'	'saga'

The optimal parameters for the logistic regression classifier are penalty='l2', C=10, multi_class='multinomial'. The cross-validation accuracy presented in Figure 3b is the result obtained with the optimal parameters. Finally, using the optimal feature set and the optimal hyperparameters we trained a classifier with the entire training set and predicted the class of the unseen compounds in the held-out test set.

Novelty detection

To detect potentially novel modes of action that are not part of the known classes seen by the trained classifier, we use an adjusted version of the novelty detection algorithm for multiclass problems proposed in Ref 29. The algorithm uses θ scores to assess the affinity of a compound to the known classes. The θ score is defined as the ratio between the confidence of the classifier in the most likely class and the confidence in the second most likely class. Using our logistic regression classifier, the θ score of a compound is given by:

$$\theta = \frac{\text{sort} \left(\text{mean}_{i \in S}(\text{class_probas}) \right) [0]}{\text{sort} \left(\text{mean}_{i \in S}(\text{class_probas}) \right) [1]}$$

where S is the set of data points belonging to the given compound and *class_probas* are the probabilities of each class predicted by the trained classifier. These θ scores are fed to an

ensemble of binary support vector machines (SVMs) to flag novel compounds. The ensemble is trained with the following procedure. First, we partition the training set ten times, each time leaving one class out as presumed-novel. For each partition, we train a logistic regression classifier and we get the θ scores for the compounds of the known classes and for the compounds of the presumed-novel class. Using the θ scores and the mean θ scores in the respective class as input features, we train a binary SVM to label compounds with 0 if they are known and with 1 if they are novel. This procedure gives us an ensemble of ten SVMs that take as input the θ score of a compound and the mean θ score of its assigned class and give as output a prediction of whether the compound is novel. The final novelty score of the compound is given by the average of the ten predictions weighted by the cross-validation accuracy of the SVM in its respective training set.

Within mode of action classification

To test for phenotypic differences within mode of action classes, we trained a classifier to distinguish the replicates from each compound within a class from the replicates of the other compounds in the class. In this case, we did not use the bootstrapped averages, but the standardized and normalized raw data. For each of the ten modes of action included in the main classification task, we pool all the compounds from the training set and the test set. Each compound is considered a separate class. We use stratified 4-fold cross validation to include replicates of every compound at every dose both in the training and the test set in each fold. We then pool together the predictions for each test fold to create a confusion matrix for all the replicates. Finally, we cluster the confusion matrix using the spectral co-clustering algorithm as implemented in scikit-learn to reveal internal structure within the mode of action.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 714853) and was supported by the Medical Research Council through grant MC-A658-5TY30. AMR was supported by a BBSRC CASE studentship part-funded by Syngenta.

References

1. Pullan, R. L., Smith, J. L., Jasrasaria, R. & Brooker, S. J. Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. *Parasit. Vectors* **7**, 37 (2014).
2. Elling, A. A. Major Emerging Problems with Minor *Meloidogyne* Species. *Phytopathology*® **103**, 1092–1102 (2013).
3. Deutsch, C. A. *et al.* Increase in crop losses to insect pests in a warming climate. *Science* **361**, 916–919 (2018).
4. Nixon, S. A. *et al.* Where are all the anthelmintics? Challenges and opportunities on the path to new anthelmintics. *Int. J. Parasitol. Drugs Drug Resist.* **14**, 8–16 (2020).
5. Sparks, T. C. & Nauen, R. IRAC: Mode of action classification and insecticide resistance management. *Pestic. Biochem. Physiol.* **121**, 122–128 (2015).
6. Wing, K. D. Pharmaceutical technologies with potential application to insecticide discovery[†]. *Pest Manag. Sci.* ps.6075 (2020) doi:10.1002/ps.6075.
7. Geary, T. G., Sakanari, J. A. & Caffrey, C. R. Anthelmintic Drug Discovery: Into the Future. *J. Parasitol.* **101**, 125–133 (2015).
8. Garcia, V. A., Crispim Junior, C. F. & Marino-Neto, J. Assessment of observers' stability and reliability — A tool for evaluation of intra- and inter-concordance in animal behavioral recordings. in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* 6603–6606 (IEEE, 2010). doi:10.1109/IEMBS.2010.5627131.
9. Rihel, J. *et al.* Zebrafish Behavioral Profiling Links Drugs to Biological Targets and Rest/Wake Regulation. *Science* **327**, 348–351 (2010).

10. Kokel, D. *et al.* Rapid behavior-based identification of neuroactive small molecules in the zebrafish. *Nat. Chem. Biol.* **6**, 231–237 (2010).
11. Laggner, C. *et al.* Chemical informatics and target identification in a zebrafish phenotypic screen. *Nat. Chem. Biol.* **8**, 144–146 (2011).
12. Anderson, D. J. & Perona, P. Toward a Science of Computational Ethology. *Neuron* **84**, 18–31 (2014).
13. Berman, G. J. Measuring behavior across scales. *BMC Biol.* **16**, 23 (2018).
14. Egnor, S. E. R. & Branson, K. Computational Analysis of Behavior. *Annu. Rev. Neurosci.* **39**, 217–236 (2016).
15. Brown, A. E. X. & de Bivort, B. Ethology as a physical science. *Nat. Phys.* (2018) doi:10.1038/s41567-018-0093-0.
16. Sluder, A. *et al.* Spiroindolines Identify the Vesicular Acetylcholine Transporter as a Novel Target for Insecticide Action. *PLoS ONE* **7**, e34712 (2012).
17. Salgado, V. L. Insect TRP channels as targets for insecticides and repellents. *J. Pestic. Sci.* **42**, 1–6 (2017).
18. Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
19. Hahnel, S. R., Dilks, C. M., Heisler, I., Andersen, E. C. & Kulke, D. *Caenorhabditis elegans* in anthelmintic research – Old model, new perspectives. *Int. J. Parasitol. Drugs Drug Resist.* S2211320720300312 (2020) doi:10.1016/j.ijpddr.2020.09.005.
20. Burns, A. R. *et al.* *Caenorhabditis elegans* is a useful model for anthelmintic discovery. *Nat. Commun.* **6**, 7485 (2015).
21. Buckingham, S. D., Partridge, F. A. & Sattelle, D. B. Automated, high-throughput, motility analysis in *Caenorhabditis elegans* and parasitic nematodes: Applications in the search for new anthelmintics. *Int. J. Parasitol. Drugs Drug Resist.* **4**, 226–232 (2014).
22. Yemini, E., Jucikas, T., Grundy, L. J., Brown, A. E. X. & Schafer, W. R. A database of *Caenorhabditis elegans* behavioral phenotypes. *Nat. Methods* **10**, 877–879 (2013).
23. Javer, A. *et al.* An open source platform for analyzing and sharing worm behavior data. (2018) doi:10.1101/377960.
24. Javer, A., Ripoll-Sánchez, L. & Brown, A. E. X. Powerful and interpretable behavioural features for quantitative phenotyping of *Caenorhabditis elegans*. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170375 (2018).
25. Benjamini, Y. & Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing under Dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
26. Edwards, S. L. *et al.* A Novel Molecular Solution for Ultraviolet Light Detection in *Caenorhabditis elegans*. *PLoS Biol.* **6**, e198 (2008).
27. Churgin, M. A. *et al.* Longitudinal imaging of *Caenorhabditis elegans* in a microfabricated device reveals variation in behavioral decline during aging. *eLife* **6**, (2017).
28. Singh, M., Singh, R. & Ross, A. A comprehensive overview of biometric fusion. *Inf. Fusion* **52**, 187–205 (2019).
29. Vinokurov, N. & Weinshall, D. Novelty Detection in MultiClass Scenarios with Incomplete Set of Class Labels. *ArXiv160406242 Cs* (2016).
30. Schwarz, R. F., Branicky, R., Grundy, L. J., Schafer, W. R. & Brown, A. E. X. Changes in Postural Syntax Characterize Sensory Modulation and Natural Variation of *C. elegans* Locomotion. *PLOS Comput. Biol.* **11**, e1004322 (2015).
31. Gomez-Marin, A., Stephens, G. J. & Brown, A. E. X. Hierarchical compression of *Caenorhabditis elegans* locomotion reveals phenotypic differences in the organization of behaviour. *J. R. Soc. Interface* **13**, 20160466 (2016).
32. Vijverberg, H. P. M. & vanden Bercken, J. Neurotoxicological Effects and the Mode of Action of Pyrethroid Insecticides. *Crit. Rev. Toxicol.* **21**, 105–126 (1990).
33. Berman, G. J., Choi, D. M., Bialek, W. & Shaevitz, J. W. Mapping the stereotyped behaviour of freely moving fruit flies. *J. R. Soc. Interface* **11**, 20140672–20140672 (2014).
34. Marques, J. C., Lackner, S., Félix, R. & Orger, M. B. Structure of the Zebrafish Locomotor Repertoire Revealed with Unsupervised Behavioral Clustering. *Curr. Biol.* (2018) doi:10.1016/j.cub.2017.12.002.

35. Kittler, J., Hatef, M., Duin, R. P. W. & Matas, J. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 226–239 (1998).
36. Cook, D. E., Zdraljevic, S., Roberts, J. P. & Andersen, E. C. CeNDR, the *Caenorhabditis elegans* natural diversity resource. *Nucleic Acids Res.* gkw893 (2016) doi:10.1093/nar/gkw893.
37. Ghosh, R., Andersen, E. C., Shapiro, J. A., Gerke, J. P. & Kruglyak, L. Natural Variation in a Chloride Channel Subunit Confers Avermectin Resistance in *C. elegans*. *Science* **335**, 574–578 (2012).
38. Marcellino, C. *et al.* WormAssay: A Novel Computer Application for Whole-Plate Motion-based Screening of Macroscopic Parasites. *PLoS Negl. Trop. Dis.* **6**, e1494 (2012).
39. Partridge, F. A. *et al.* An automated high-throughput system for phenotypic screening of chemical libraries on *C. elegans* and parasitic nematodes. *Int. J. Parasitol. Drugs Drug Resist.* **8**, 8–21 (2018).
40. Perlman, Z. E. Multidimensional Drug Profiling By Automated Microscopy. *Science* **306**, 1194–1198 (2004).
41. Ljosa, V. *et al.* Comparison of Methods for Image-Based Profiling of Cellular Morphological Responses to Small-Molecule Treatment. *J. Biomol. Screen.* **18**, 1321–1329 (2013).
42. Caicedo, J. C. *et al.* Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).
43. Kleinstreuer, N. C. *et al.* Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms. *Nat. Biotechnol.* **32**, 583–591 (2014).
44. Douris, V. *et al.* Stably Transformed Insect Cell Lines: Tools for Expression of Secreted and Membrane-anchored Proteins and High-throughput Screening Platforms for Drug and Insecticide Discovery. in *Advances in Virus Research* vol. 68 113–156 (Elsevier, 2006).
45. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
46. Pereira, T. D. *et al.* Fast animal pose estimation using deep neural networks. *Nat. Methods* **16**, 117–125 (2019).
47. Graving, J. M. *et al.* DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife* **8**, e47994 (2019).
48. Waskom, M. & team, the seaborn development. *mwaskom/seaborn*. (Zenodo, 2020). doi:10.5281/zenodo.592845.
49. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

cross validation accuracy = 26.6%