

A predictive model for drug bioaccumulation and bioactivity in *Caenorhabditis elegans*

Andrew R Burns^{1,2}, Iain M Wallace², Jan Wildenhain^{3,4}, Mike Tyers^{1,3,4}, Guri Giaever^{1,2,5}, Gary D Bader^{1,2,6}, Corey Nislow^{1,2,6}, Sean R Cutler⁷ & Peter J Roy^{1,2,*}

The resistance of *Caenorhabditis elegans* to pharmacological perturbation limits its use as a screening tool for novel small bioactive molecules. One strategy to improve the hit rate of small-molecule screens is to preselect molecules that have an increased likelihood of reaching their target in the worm. To learn which structures evade the worm's defenses, we performed the first survey of the accumulation and metabolism of over 1,000 commercially available drug-like small molecules in the worm. We discovered that fewer than 10% of these molecules accumulate to concentrations greater than 50% of that present in the worm's environment. Using our dataset, we developed a structure-based accumulation model that identifies compounds with an increased likelihood of bioavailability and bioactivity, and we describe structural features that facilitate small-molecule accumulation in the worm. Preselecting molecules that are more likely to reach a target by first applying our model to the tens of millions of commercially available compounds will undoubtedly increase the success of future small-molecule screens with *C. elegans*.

The many virtues of the tiny nematode *C. elegans* make it a promising model system for the discovery and characterization of novel bioactive compounds. Because of the worm's rapid life cycle, small size and hermaphroditism, libraries of small molecules can be screened for bioactivity in the context of the whole animal and over its entire life cycle in a high-throughput fashion¹. Furthermore, the power of *C. elegans* genetic analysis has repeatedly uncovered the mechanism of action of both small-molecule tools^{2–5} and novel anthelmintics^{6,7}. The worm also shares extensive genetic conservation with more complex animals, and the growing list of human disease models in the worm further distinguish *C. elegans* as a unique tool for the discovery of novel therapeutics⁸.

Unfortunately, *C. elegans* is relatively resistant to perturbation by pharmacologically active molecules. For example, pharmacological agents must often be applied to the worm at concentrations that are orders of magnitude higher than those used in mammalian cell culture^{2,9–11}. Moreover, we found that only 2% of pharmacologically active compounds can induce a robust phenotype in the worm when screened at a concentration of 25 μM ^{1,2}. Screening compounds at a higher concentration may overwhelm the worm's xenobiotic defenses, but doing so can be prohibitively costly and would likely result in molecules precipitating out of solution in the screening medium. Circumventing the resistance of *C. elegans* to bioactive compounds would increase its utility as a small-molecule screening tool.

C. elegans has extensive physical and enzymatic xenobiotic defenses that may render many pharmacological tools ineffective. The physical barriers include a four-layered cuticle that lines its exterior and oral and rectal cavities¹², as well as an intestine through which solutes are rapidly pumped¹³. The worm's genome is replete with predicted xenobiotic detoxification enzymes, including 86 cytochrome P450s, and 60 ATP-binding cassette transporters, many

of which likely function as xenobiotic efflux pumps¹⁴. Compounds that are ineffective when applied to whole animals can readily antagonize their targets if they are provided with direct access^{2,15,16}. Hence, it is likely that *C. elegans* is generally resistant to exogenously applied pharmacologicals because they fail to accumulate to effective concentrations within its tissues. Modeling the properties of small molecules that promote accumulation in the worm would greatly facilitate the discovery of new biological probes and drug leads.

Typically, only a small fraction of the more than 13 million commercially available small molecules¹⁷ is screened by any one academic laboratory. Many bioactive compounds therefore remain undiscovered within the unscreened fraction of purchasable chemical space. For instance, the largest *C. elegans* chemical screen to date used 88,000 compounds³, which covers less than 0.7% of available chemical space. Even this relatively large-scale screen leaves more than 99% of all possible bioactives undiscovered within the unscreened fraction of available chemical space. Hence, it is essential to maximize the number of bioactive compounds in the fraction of molecules that is screened. Given that bioavailability is a prerequisite to bioactivity, one way to increase the hit rate of a chemical screen is to develop a property-based computer model that predicts small-molecule accumulation within the screening system of choice. The model can then be used to select molecules with an increased likelihood of bioavailability from the purchasable chemical space and improve the chances of finding a hit.

Here, we describe a high-throughput HPLC-based technique to measure the bioaccumulation of exogenous compounds and their metabolites in *C. elegans*. We use this method to assay the accumulation of more than 1,000 commercially available drug-like molecules in whole animals. We use these data to build a predictive model that distinguishes accumulating from non-accumulating compounds on the basis of their structural properties. When applied to two

¹Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. ²Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada. ³Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada.

⁴Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh, United Kingdom. ⁵Department of Pharmaceutical Sciences, University of Toronto, Toronto, Ontario, Canada. ⁶Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada. ⁷Center for Plant Cell Biology, Department of Botany and Plant Sciences, University of California, Riverside, California, USA. *e-mail: peter.roy@utoronto.ca

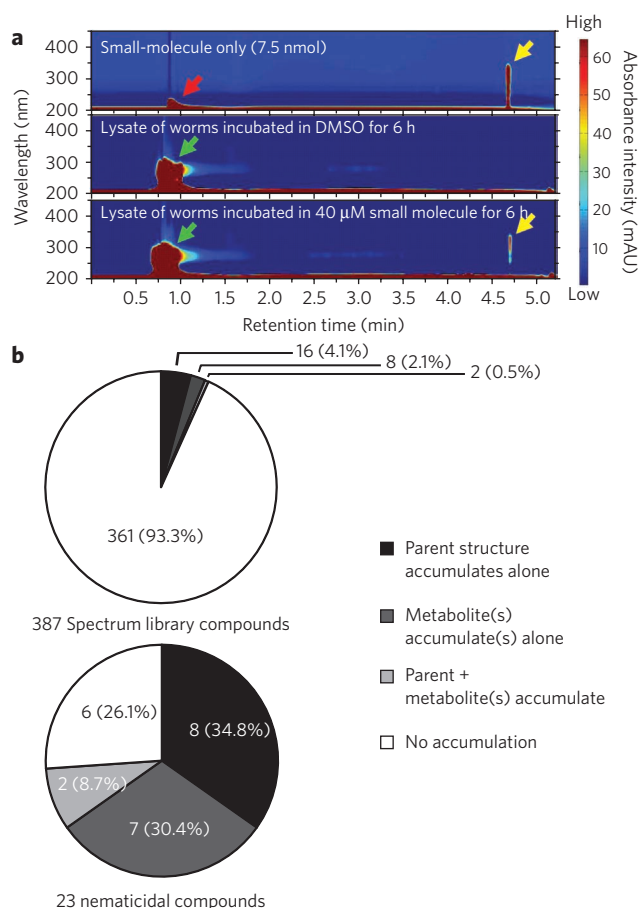


Figure 1 | A survey of exogenous drug-like molecule accumulation in *C. elegans*. (a) Heat-mapped HPLC-DAD chromatograms. Retention time is shown on the x axis, and absorbance wavelength is shown on the y axis. The scale of absorbance intensity, in milli-absorbance units (mAU), is shown on the right. The DMSO peak (red arrow), peak of worm contents (green arrows) and small-molecule peak (yellow arrows) are indicated. (b) Pie charts showing the fraction of accumulating structures for 387 compounds from the Spectrum library (Microsource) and for 23 nematicides derived from a 10K DIVERSet library (Chembridge) and the 1K HitsKit library (Maybridge).

naive chemical libraries, our predictive model of bioaccumulation enriches for compounds with diverse bioactivities in the worm.

RESULTS

An HPLC-based small-molecule accumulation assay for worms

As a first step toward developing a predictive model of bioaccumulation in *C. elegans*, we surveyed over 1,000 commercially available drug-like molecules for their ability to accumulate in worm tissue. To do this, we modified our previously described HPLC-based assay² to a 96-well-plate format. Briefly, we incubated ~5,000 fourth-larval-stage worms in 40 μM of compound for 6 h in filter plates. We then drained the incubation buffer from the wells, washed the worms, re-suspended them in fresh buffer, transferred them to new plates and then lysed the worms chemically. We used a reverse-phase HPLC system coupled with a diode array detector (HPLC-DAD) to separate and visualize the components of the worm lysate (Fig. 1a). This method measures small-molecule bioaccumulation in *C. elegans*, as opposed to bioavailability, because it detects compounds that are taken up by the worms during incubation and then remain in the worms after washing. Similar HPLC approaches have been developed to identify drugs and drug metabolites for clinical and forensic

purposes^{18,19}. Though the use of a diode array detector limits our detection space to compounds with conjugated systems of pi bonds (for example, aromatics), we chose this approach over a mass spectrometry-based approach because of the cost effectiveness and relatively high-throughput nature of the HPLC-DAD technique.

Here, we define an accumulating molecule as one that (i) is detectable by our HPLC protocol in at least two of three replicate lysates, (ii) is undetectable in the no-worm sham trials (to control for compounds that precipitate in the filter-plate wells) and (iii) remains detectable after the worms are washed with the highest concentration of sodium dodecyl sulfate (SDS) that does not result in any obvious physiological changes (to control for small-molecule association with the cuticle). Molecules that accumulate as putative metabolites (see below) are reprocessed using dead worms to ensure that the putative metabolites are dependent on living worms (to control for accumulated contaminants or spontaneous oxidation products of the parent molecule).

As a proof of principle, we used our high-throughput (HT) HPLC method to examine the accumulation of 21 members of the 1,4-dihydropyridine (DHP) family of L-type calcium channel antagonists after either a 0.5- or 6-h co-incubation with worms (Supplementary Fig. 1). Using a low-throughput approach, we previously examined the accumulation of 12 of these DHPs after a 2 h incubation period and found that four of the 12 DHPs accumulate, including the three bioactive molecules nemadipine-A, nemadipine-B and felodipine, and one inactive DHP called analog 7 (ref. 2). Using our HT-HPLC approach, all three bioactive DHPs were found to accumulate at both time points, as well as DHP analog 7 at the 0.5-h time point and DHP analog 1 at the 6-h time point. Hence, our HT-HPLC protocol can robustly detect the accumulation of exogenous small molecules in the worm, and the results obtained using the HT method correlate well with the low-throughput approach.

A survey of drug-like molecule accumulation in *C. elegans*

There are currently more than 13 million commercially available small molecules¹⁷, a quantity that greatly exceeds what could likely be sampled in any single *C. elegans* screen. Therefore, it is important to carefully select the compounds that will be screened in order to maximize the number of bioactives obtained. Property-based modeling of small-molecule accumulation in worms could be used to prioritize purchasable chemical space so as to enrich for compounds that have a greater likelihood of accessing a biologically relevant target. Even though the overwhelming majority of commercially available compounds obey Lipinski's rule-of-five and are considered "drug-like"^{17,20}, the rate at which molecules are identified as bioactive against *C. elegans* remains relatively low^{1,2}. To better understand the molecular properties of commercially available drug-like molecules that facilitate their accumulation in *C. elegans*, we used our HT-HPLC assay to measure the accumulation of the Spectrum collection of 2,000 pharmacologically active compounds (MicroSource Inc.) in whole worms. We chose the Spectrum library because it is enriched for popular small bioactive compounds that include human drugs and pharmacological tools²¹, and more than 90% of these compounds have drug-like properties (Supplementary Fig. 2). In addition, we have previously investigated this collection for bioactivity in wild-type worms^{1,2}, which allowed us to compare bioaccumulation and bioactivity. Of these 2,000 Spectrum molecules, 1,096 were detectable when we processed 7.5 nmol of each molecule using our HPLC-DAD system.

We next determined which of the 1,096 detectable Spectrum molecules accumulate in wild-type worms after a 6-h incubation. We found that 96 compounds (9.3%) accumulate out of a set of 1,027 (69 molecules were eliminated from consideration because the sample was either lost during processing or considered a false positive after control experiments). We call this dataset the "complete Spectrum dataset" (Supplementary Fig. 3a). The majority of accumulating

compounds (64%) are found in worms at concentrations that are less than the 40- μ M assay concentration, with a median concentration of 26 μ M (Supplementary Fig. 4). We discovered a bias in our accumulation dataset such that molecules with relatively low limits of detection are enriched for accumulating compounds, and molecules with relatively high detection limits are enriched for non-accumulating compounds (see Supplementary Methods for the approach taken to estimate the limits of detection for the compounds in the dataset and for our analysis of bias). We therefore applied a 19- μ M detection-limit cutoff to the complete Spectrum dataset to make an unbiased assessment of the fraction of small drug-like molecules that accumulate in worms (see Supplementary Methods for the derivation and application of this cutoff). We found that only 26 (6.7%) of the 387 molecules that passed the cutoff accumulate using this unbiased criterion (Fig. 1b). These data support the conclusion that worms are generally resistant to the accumulation of exogenous drug-like molecules and provide a foundation to investigate the properties that govern small-molecule accumulation in worms.

Exogenous compounds can accumulate as metabolites

Of the molecules that we have found to accumulate in worms (Supplementary Fig. 3), 36% accumulate as metabolites. The majority of these metabolites have a similar spectral absorption profile as their respective parental compounds but typically elute from the reverse-phase HPLC column faster than the parental molecule. To verify that the metabolites are bona fide derivatives of the parental compounds, we examined 17 metabolites from 12 parental compounds (1–12) by mass spectrometry and tandem MS (MS-MS; Table 1 and Supplementary Fig. 5). For 11 of the 12 compounds, our analyses show that the metabolites are parent derived, indicating that the majority of the novel peaks observed on the HPLC chromatograms represent bona fide metabolites of the parental compound. For one of the 12 compounds (compound 12 in Table 1), we could not identify abundant masses specific to the purified metabolite fractions relative to the DMSO control fractions.

We deduced the modification(s) made to several of the parental compounds on the basis of the mass of the metabolites and their fragmented derivatives in the MS-MS analysis (Table 1 and Supplementary Fig. 5). The modifications made to parental compounds 4, 5 and 6 are consistent with drug metabolism in mammals whereby phase I enzymes functionalize the drug through demethylation or reduction for subsequent metabolism by phase II conjugating enzymes^{22,23}. The diol metabolite of compound 8 likely results from the activity of the phase I enzymes cytochrome P450 and epoxide hydrolase. For compounds 1, 2, 3 and 7, only conjugated derivatives were found, including glucosidated and sulfated derivatives, suggesting that these compounds are modified directly by phase II-like enzymes. To verify the correctness of our proposed biotransformations, accurate mass determinations were performed for four representative metabolites (indicated in Table 1). For all four metabolites the measured masses are within 5 p.p.m. of the calculated monoisotopic masses, indicating that the elemental compositions are correct (see Supplementary Table 1). This work provides the first *in vivo* survey of xenobiotic metabolism in *C. elegans*, to our knowledge, and reveals many similarities of drug metabolism between worms and mammals.

A predictive structure-based accumulation model (SAM)

The majority of the purchasable drug-like compounds we have so far tested fail to accumulate in worms (Supplementary Fig. 3). Given that most commercially available chemical libraries are designed to have similar drug-like properties as the molecules we have tested¹⁷, we sought to better understand the attributes of these compounds that influence their accumulation in worms and to develop a generally applicable model that can predict bioavailability on the basis of these attributes.

To build a predictive small-molecule structure-based accumulation model (SAM) for worms, we consolidated all of the 1,132 molecules we have assayed for accumulation into one dataset (Supplementary Fig. 3). After applying the 19- μ M detection limit cutoff and removing duplicate structures, 483 molecules remained in the SAM training set, of which 74 accumulate in worms (Fig. 2a). We used the 483 compounds to train an ECFP_4-Naive Bayesian classifier to distinguish accumulating from non-accumulating molecules on the basis of their structural features (Fig. 2b). ECFP_4 is a two-dimensional circular fingerprint descriptor that represents each compound as a series of small fragments that are built by starting from each heavy atom in the compound and extending out up to four bond lengths. In total, 4,698 fragments were derived from the 483 compounds in the training set using the ECFP_4 fingerprint descriptor. The Naive Bayesian classifier identifies the fragments that are over-represented and under-represented in the accumulating subset of molecules^{24,25}. Overrepresented fragments receive positive scores, and under-represented fragments receive negative scores. The accumulation score of a molecule is then calculated by summing up the scores of its respective fragments. There was relatively little overlap in the scores of accumulating and non-accumulating subsets of the 483-molecule training set (Fig. 2c), indicating that there are fragments that distinguish accumulating from non-accumulating structures in worms.

The predictive power of the SAM was estimated by five independent five-fold cross-validation experiments (Fig. 2d). On average, the SAM performs about three times better than random; the top scoring 5% and 10% of molecules are 3.5- and 2.9-fold enriched for accumulation compared to randomly selected compounds, respectively, indicating that the SAM can be successfully applied to independent datasets. When compared to three additional Naive Bayesian models, trained using distinct molecular descriptors (MDL PublicKeys²⁶, physicochemical properties and Lipinski properties²⁰), the SAM outperforms all three as determined by five-fold cross-validation (Supplementary Fig. 6).

To biologically validate the SAM and to assess its general applicability, we used it to rank the 50,000 compounds of a 50K DIVERSet library (Chembridge Corp.) and experimentally tested its predictions. First, we applied an in-house predictive model of UV-visible absorbance to the 50K DIVERSet library to increase the probability that any chosen molecule could be detected by our diode array detector (see Supplementary Methods). Second, we excluded library molecules from consideration that were >85% similar to any molecule in the SAM training set. After applying these two filters, we applied our SAM to the remaining 4,993 DIVERSet molecules. We randomly selected 23 molecules from the top-scoring 5%, 29 molecules from the bottom-scoring 5% and 28 molecules from the entire set (which we refer to as random-scoring molecules). We assayed the accumulation of these 80 molecules in worms using our HT-HPLC method. After controls, and after applying the 19- μ M detection limit cutoff, 15 of the top-scoring, 25 of the bottom-scoring and 18 of the random-scoring compounds remained in the test set. We found that only 1 out of 25 (4%) of the bottom-scoring molecules accumulate, representing a 2.8-fold under-enrichment relative to the random-scoring compounds (Fig. 2e). By contrast, we found that 9 out of 15 (60%) of the top-scoring molecules accumulate, representing a 5.5-fold enrichment compared to the random-scoring compounds and a 15-fold enrichment relative to the bottom-scoring molecules (Fig. 2e). These results indicate that our structure-accumulation model can be applied to diverse libraries and successfully predict the accumulation of exogenous drug-like compounds in worms on the basis of their structural features.

Structural features that influence accumulation

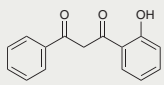
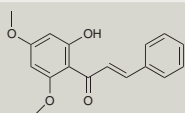
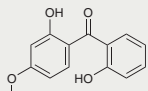
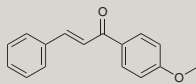
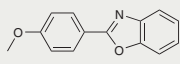
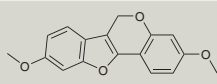
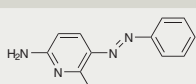
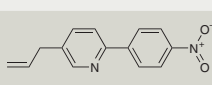
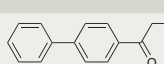

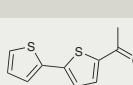
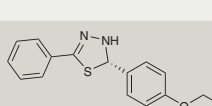
To identify which of the 4,698 ECFP_4 fragments most strongly influence the accumulation of exogenous compounds in worms we compiled the top-scoring and bottom-scoring five structurally

non-redundant and non-overlapping fragments (Fig. 3 and Supplementary Fig. 7). Strikingly, the top five and bottom five fragments are present in ~50% of the accumulating and non-accumulating compounds, respectively (Fig. 3a,b).

To interpret how the top five features facilitate accumulation, we considered the structural scaffolds from which they were derived within our SAM training set. Four scaffold types enriched for accumulating

molecules were identified: (i) the ‘unfused’ biaryl scaffold, (ii) the methyl piperazine scaffold, (iii) the 2- or 3-phenyl-chromen-4-one scaffolds and (iv) the ‘fused’ biaryl scaffold (Fig. 3c–f). Seventy-two distinct compounds in the training set are composed of these four scaffolds, and they account for 41% of the accumulating compounds in the training set. Notably, 10 out of 12 accumulating ‘unfused’ biaryl structures accumulate as metabolites in the worm, and all but one

Table 1 | Characterization of *C. elegans* xenobiotic metabolites

Compound	Parent structure	Parent (P)		Metabolite 1 (M1)		Metabolite 2 (M2)	
		Molecular weight	Mass	Enzymatic modification	Mass	Enzymatic modification	
1		240	402	O-hexosidation ^a [P + 162]			
2		284	446	O-glucosidation ^b [P + 162]			
3		244	406	O-hexosidation [P + 162]			
4		238	386	Demethylation; O-glucosidation ^{a,b} [P - 14 + 162]			
5		225	211	Demethylation [P - 14]	373	O-glucosidation ^b [M1 + 162]	
6		282	284	Reduction ^c [P + 2]	446	O-glucosidation ^{b,d} [M1 ₂₈₄ + 162]	
7		213	455	N-glucosidation; N-sulfation [P + 162 + 80]			
8		240	274	Aryl epoxidation; epoxide hydrolysis ^a (diol formation) [P + 16 + 18]			
9		210	—	—			
10		196	—	—			
11		208	—	—			
12		284	n/a	n/a	n/a	n/a	

Metabolite masses and corresponding biotransformations, inferred from MS and MS-MS analysis, are listed (see Supplementary Fig. 5 for supporting MS data). A dash indicates that the metabolite is likely a bona fide derivative of the parent molecule but the respective biotransformation could not be unambiguously determined. For the metabolites of compound 12, abundant masses specific to the metabolite fractions, compared to controls, could not be identified.

^aAccurate mass was measured for metabolites 1_M1, 4_M1, 6_M2₂₆₈ and 8_M1 (Supplementary Table 1). All four metabolites have measured accurate masses within 5 p.p.m. of their respective calculated monoisotopic masses. ^bThe metabolites of compounds 2, 4, 5 and 6 were confirmed to be O-linked β-glucosides by β-glucosidase digestions (Supplementary Fig. 10). ^cMetabolites M1₂₈₄ and M1₂₆₈ derived from compound 6 had the same retention time, when processed using the HT-HPLC method, and were co-purified. ^dMetabolites M2₂₈₄ and M2₂₆₈ derived from compound 6 had the same retention time, when using the HT-HPLC method, and were co-purified.



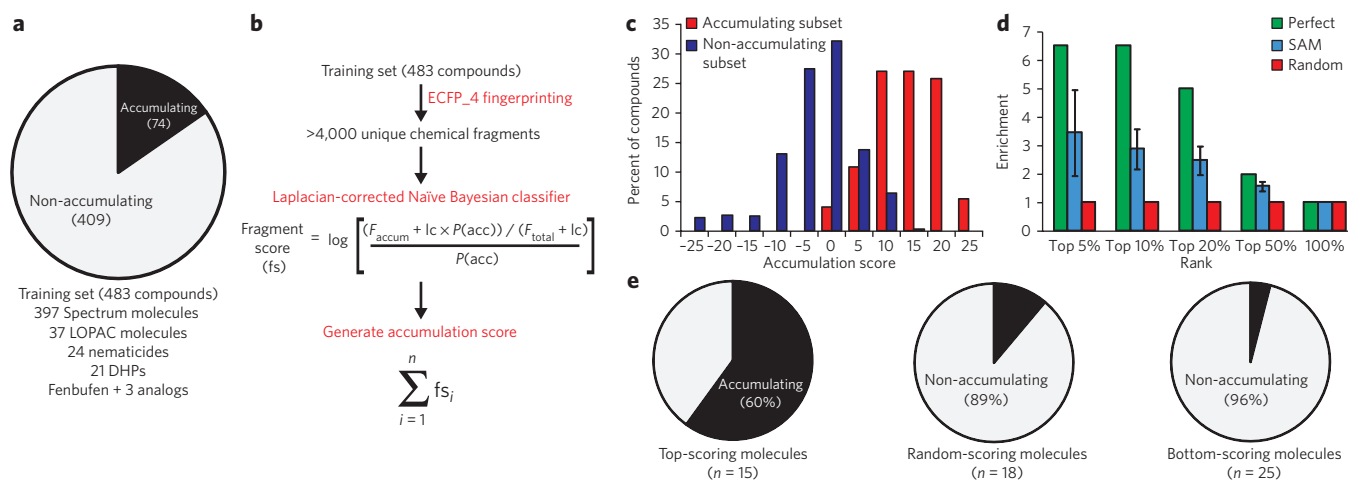


Figure 2 | A machine-learned structure-based model predicts the accumulation of drug-like molecules in *C. elegans*. (a) A 483-molecule training set was used for machine learning. (b) Building the ECFP₄-Bayesian structure-based accumulation model (SAM). F_{accum} is the number of accumulating compounds in which the fragment is present, F_{total} is the total number of compounds in which the fragment is present, $P(\text{acc})$ is the probability that a compound accumulates in the training set, and the Laplacian correction ($lc = P(\text{acc})^{-1}$). An accumulation score is generated for any molecule by summing the scores for each fragment in the molecule. (c) The distribution of accumulation scores for the accumulating and non-accumulating subsets of molecules in the training set. (d) Five-fold cross-validation of the model. Test sets were ranked by the model and the enrichment of accumulating compounds, relative to random, was plotted for the top-scoring 5%, 10%, 20%, 50% and 100% of compounds. Enrichment for a perfect model was also plotted relative to random. Error bars represent ± 1 s.d. (e) Biological validation of the SAM. A set of compounds from a 50K DIVERSet library (Chembridge) was ranked using the SAM, and three subsets were chosen at random from the top-scoring 5%, the entire set and the bottom-scoring 5%, respectively. These molecules were tested for accumulation in worms using our HT-HPLC method. Pie charts show the fraction of accumulating molecules for each subset.

induce penetrant lethality at 25 μM , suggesting that this scaffold may be a useful building block for the development of novel nematocides.

Three of the bottom five fragments (B1, B2 and B4 in Fig. 3b) are derived from compounds that contain a carboxylic acid group, an aliphatic hydroxyl group or a sulfonyl group. It is not surprising that these fragments are over-represented in the non-accumulating subset of molecules, for a number of reasons. First, these features contain hydrogen and oxygen atoms that can make hydrogen bond contacts with water molecules, promoting solubility in aqueous buffer and preventing cell membrane permeability. Second, these functional groups are known to be sites of phase II conjugation reactions in mammals, such as glucuronidation, glycosylation and sulfation²³. Thus, compounds with these features are more likely to be modified and excreted by the organism. In agreement with these data, an analysis of the Lipinski property distributions for the accumulating and non-accumulating subsets of our training set shows that the accumulating compounds generally have fewer hydrogen bond donors and acceptors than the non-accumulating molecules and that accumulating structures typically have a greater LogP than non-accumulating structures (Supplementary Fig. 8). Despite these differences, the Lipinski properties perform the least well of all the approaches tested to predict accumulation (see Supplementary Fig. 6). It is presently unclear how the remaining two fragments (B3 and B5 in Fig. 3b) might antagonize small-molecule accumulation in the worm.

We observed that compounds with scaffolds that facilitate accumulation will generally accumulate in worms unless they have one or more of the bottom-scoring fragments B1–B5 (Fig. 3c–f). Hence, the features that antagonize small-molecule accumulation generally act as the ‘master’ determinants of accumulation. An example of how the negative features influence bioaccumulation is provided by our study of the biphenyl fenbufen, a nonsteroidal anti-inflammatory drug²⁷, and three fenbufen analogs (Supplementary Fig. 9). Fenbufen and fenbufen analog 3 both contain carboxylic acid groups and fail to accumulate in worms. By contrast, fenbufen analogs 1 and 2 both

lack carboxylic acid groups and accumulate as metabolites (compounds 9 and 10 in Table 1). The fenbufen analogs that accumulate as metabolites also induce penetrant lethality at 25 μM , but the non-accumulating analogs do not.

The SAM enriches for compounds with distinct bioactivities

We previously found that the only DHPs that show bioactivity are those that can accumulate in worm tissue (Supplementary Fig. 1), suggesting that bioaccumulation is generally required for bioactivity. Additional observations made here further support this idea. First, only those fenbufen analogs that accumulate are bioactive (Supplementary Fig. 9). Second, only two molecules from our unbiased Spectrum dataset are bioactive in the worm, and both accumulate (Supplementary Fig. 3a). Finally, 17 out of the 23 nematocides that we previously described² accumulate to concentrations greater than 19 μM in the worm (Fig. 1b).

Given that bioaccumulation is correlated with bioactivity, we anticipated that our SAM would also enrich for bioactive molecules in *C. elegans*. To test whether our SAM can enrich for molecules that are bioactive in the worm, we used it to score the 10K DIVERSet library (Chembridge Inc.) that we previously screened for the induction of gross phenotypes in wild-type worms^{1,2}. We then ranked the molecules that are structurally distinct from those in the SAM’s training set ($n = 9,740$) and determined whether the top-scoring molecules are enriched for phenotype. Thirty percent (14 out of 47) of the bioactive molecules in the library were present in the top-scoring 5% of the compounds, representing a six-fold enrichment of bioactives ($P < 3.5 \times 10^{-8}$, Fig. 4a). To further test our SAM’s ability to enrich for compounds with distinct bioactivities, we used it to rank 1,040 compounds that were screened by the National Institute of Neurological Disorders and Stroke (NINDS) for the correction of neuronal defects in a worm model of Huntington’s disease (PubChem BioAssay, AID: 1599). The top 5% of molecules that are structurally distinct from the compounds in our SAM’s training set ($n = 826$) are 3.8-fold enriched for bioactive molecules

relative to random ($P < 0.02$, Fig. 4a). Hence, applying our SAM to naive libraries can greatly increase the efficiency by which novel bioactive molecules are identified, which validates the model and the HT-HPLC method used to generate it.

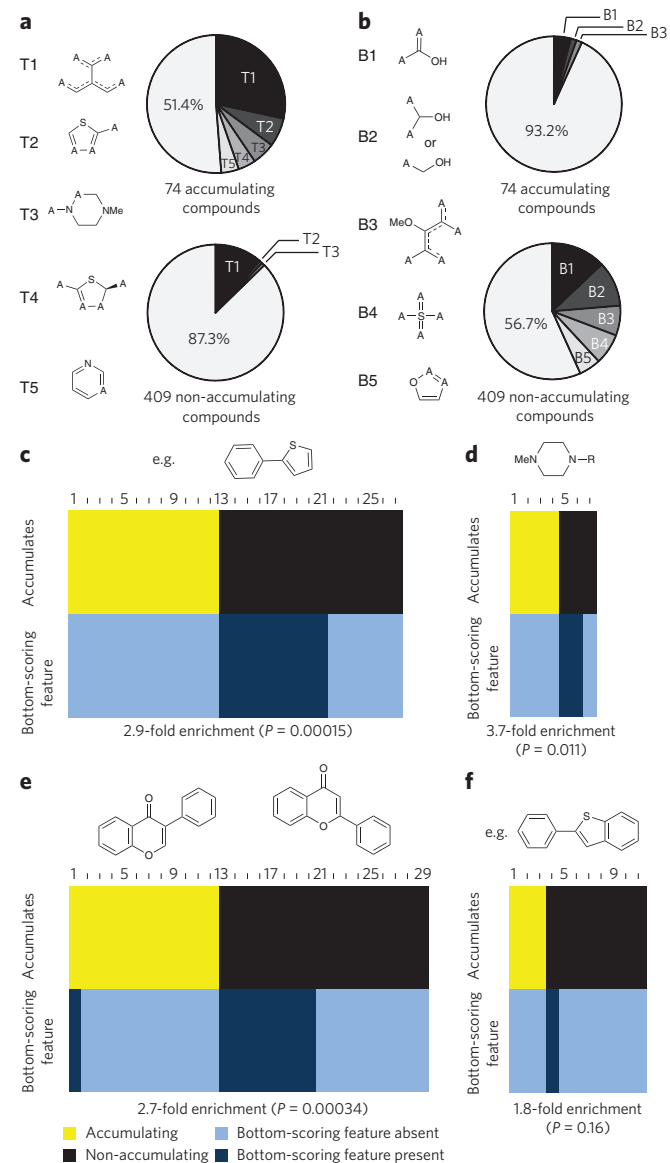


Figure 3 | Prominent substructures that influence small-molecule accumulation in *C. elegans*.

(a) The top five non-overlapping fragments learned by the structure-based accumulation model (T1–T5). An atom denoted with an ‘A’ could be either C, N, O or S (where appropriate). Dashed lines indicate double-bond character. Pie charts show the fraction of accumulating and non-accumulating compounds in the training set with each of the five top-scoring non-overlapping fragments. (b) The bottom five non-overlapping fragments learned by the structure-based accumulation model (B1–B5). Pie charts show the fraction of accumulating and non-accumulating compounds in the training set with each of the five bottom-scoring non-overlapping fragments. (c–f) Clustergrams of compounds containing the ‘unfused’ biaryl scaffold (c), the methyl piperazine scaffold (d), the 2- or 3-phenyl-chromen-4-one scaffolds (e) and the ‘fused’ biaryl scaffold (f). Each column, represented by a number or a dash on top of the box, represents a distinct molecule of the class depicted by the structural scaffold above. The enrichment of accumulating compounds and the hypergeometric P value associated with that enrichment are shown at the bottom of each clustergram.

Diversity analysis of SAM-ranked structures

Applying structure-based filters to a chemical library will inevitably narrow its structural diversity. We therefore investigated the impact of our SAM on structural diversity in three different ways using the Tanimoto molecular similarity scoring system with ECFP₄ fingerprints (see Methods). Pairwise Tanimoto coefficients range from 0, which indicates the absence of structural similarity, to 1, which indicates identical structures. In practice, Tanimoto scores ≤ 0.2 are so low that they do not represent any meaningful structural similarity^{28,29}.

In our first approach to assessing the diversity of SAM-selected molecules, we calculated all pairwise Tanimoto coefficients for the top-scoring 5% and a random-scoring 5% of compounds from each of the 10K DIVERSet and NINDS libraries (the distributions are shown in Fig. 4b). The average pairwise similarity scores for the top-scoring compounds from the DIVERSet and NINDS libraries are 0.138 and 0.112, respectively, with 87% and 94% of the pairwise scores being ≤ 0.2 , respectively. The average pairwise similarity scores for the random-scoring compounds from the DIVERSet and NINDS libraries are 0.110 and 0.089, respectively, with 96% and 97% of the pairwise scores being ≤ 0.2 , respectively. Although the random-scoring molecules are expectedly more diverse, our SAM does not simply reduce the library to a small number of structurally similar compounds.

We next investigated the diversity of the core scaffolds present in the SAM-selected compounds. Murcko scaffolds³⁰, which retain the ring systems and linkers of molecules but eliminate the side chains, were generated for the top-scoring 5% of molecules and three random sets of an equivalent number of molecules from each of the 10K DIVERSet and NINDS libraries. Similarity networks were created for the top-scoring and the random-scoring sets of scaffolds, in which scaffolds are connected if they have a pairwise ECFP₄ or Tanimoto score ≥ 0.7 (the top-scoring scaffold similarity networks are shown in Fig. 4c). Scaffold networks generated in this way have been used to explore the scaffold composition of purchasable screening libraries³¹. We then counted the number of unique scaffold clusters in each network, including unconnected singletons. The top-scoring DIVERSet scaffolds are comprised of 290 distinct scaffold clusters (Fig. 4c), and the random-scoring DIVERSet scaffolds are comprised of an average of 380 (± 3.5) distinct scaffold clusters. The top-scoring NINDS scaffolds are composed of 33 distinct scaffold clusters (Fig. 4c), and the random-scoring NINDS scaffolds are composed of an average of 30 (± 5.0) distinct scaffold clusters. Hence, our analysis shows that although the application of our SAM can reduce the scaffold diversity of a given library, the top-scoring 5% of molecules (~500 structures) from our analyses still represent hundreds of unique scaffolds.

Finally, we assessed the structural diversity of the bioactive molecules in the top-scoring 5% of molecules from the 10K DIVERSet and NINDS libraries. We built similarity networks for these bioactive compounds by linking compounds with a Tanimoto coefficient of 0.3 or more, as a high degree of structure-phenotype concordance has been observed in cell-based assays for pairs of molecules that have a Tanimoto coefficient of 0.3 or greater^{32,33}. Nine out of 14 (64%) active compounds in the top-scoring DIVERSet subset are singletons in the network, and two out of four (50%) actives in the top-scoring NINDS subset are singletons in the network (Fig. 4d). Furthermore, the majority of the bioactive compounds retrieved in the top-scoring 5% of the 10K DIVERSet (11 out of 14) and NINDS (four out of four) libraries have distinct core scaffolds (Fig. 4c). Hence, our analysis of the structural diversity of the SAM-selected compounds suggest that the SAM enriches for both accumulating and bioactive molecules without dramatically limiting the structural diversity of the compounds screened or the hits obtained.

For the benefit of the chemical biology community, we have ranked all of the >13 million purchasable molecules of the ZINC

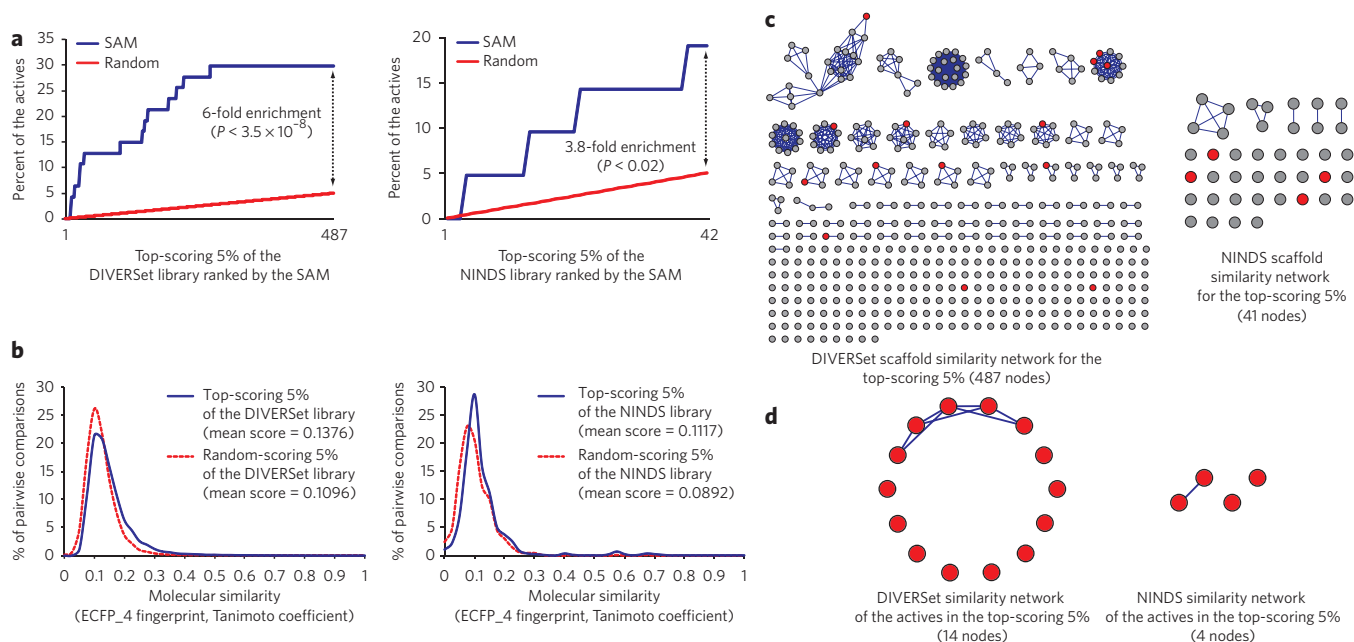


Figure 4 | The *C. elegans* structure-based accumulation model (SAM) enriches for structurally diverse compounds with distinct bioactivities in the worm. (a) Phenotype enrichment curves for the molecules in the top-scoring 5% of the 10K DIVERSet (Chembridge) and NINDS (PubChem BioAssay, AID: 1599) libraries ranked by the SAM. The percentage of actives obtained as the compounds are traversed from the highest- to the lowest-ranking compound is shown. The enrichment value and the hypergeometric P value associated with the enrichment are indicated for the top-scoring 5%. (b) Structural diversity of the compounds shown in (a). The distributions of all pairwise Tanimoto similarity scores are shown for the top-scoring 5% of compounds, as well as a random-scoring 5%, from the 10K DIVERSet and NINDS libraries. Three different random-scoring distributions were generated for each library; because of considerable overlap, only one distribution is shown for clarity. (c) Scaffold composition of the compounds shown in a. Murcko scaffold similarity networks are shown for the compounds in the top-scoring 5% of the 10K DIVERSet and NINDS libraries. Scaffolds (nodes) are connected if they have a pairwise ECFP₄-Tanimoto similarity score ≥ 0.7 . Red nodes indicate active compound scaffolds, whereas gray nodes indicate inactive compound scaffolds. (d) Similarity networks for the active compounds in the top-scoring 5% of the 10K DIVERSet and NINDS libraries. Compounds (red nodes) are connected if they have a pairwise ECFP₄-Tanimoto similarity score ≥ 0.3 . Network visualization was performed using Cytoscape⁴³.

database using our SAM. These data are available as a zipped 1.1-GB .tab file (tab separated text file) at <http://142.150.52.140:10080/ZINCSAM.zip>. We have made the top-scoring 5% of the ZINC database available as an Excel file (**Supplementary Data 1**).

DISCUSSION

The general resistance of *C. elegans* to pharmacological perturbation has impeded its utility as a chemical-genetic model system. Our systematic analysis of drug-like compound accumulation has shown for the first time that poor bioavailability is a major contributing factor to this resistance. Through property-based modeling, we are able to identify molecules that have an increased likelihood of accumulating in worms and, by extension, have an increased likelihood of affecting a biologically relevant target. Hence, by screening molecules that are more likely to accumulate, we are able to circumvent many of the xenobiotic defenses of the worm to identify new biological probes and potential drug leads. Our structure-based accumulation model now provides a tool by which commercially available molecules can be prioritized to increase the probability of identifying structurally distinct compounds with diverse bioactivities in worms, resulting in more efficient screens. The approach of generating computer-based models to identify molecules with a higher likelihood of bioactivity can be applied to other systems as well (such as planaria, *Drosophila* and zebrafish) and could serve as a new paradigm for the design of model organism chemical screens.

Given the resulting enrichment of bioactive molecules upon the application of our SAM, the resulting minimal loss in library diversity is acceptable for our purposes. However, it is up to the individual screener to decide what constitutes an acceptable loss in diversity.

Personalized diversity filters can be applied to SAM-selected compounds to ensure adequate structural uniqueness.

There are additional important considerations when applying our model to novel compound sets in the future. First, the model will work best when applied to libraries with feature distributions that are similar to those of the training set used to generate the model. In the same vein, the structural fragments learned by our model do not represent an exhaustive list, and there are likely other structural features that influence small-molecule bioavailability in worms that were not sampled in our analysis. Sampling of more compounds from an increasingly diverse chemical space will undoubtedly improve the prediction coverage of our model. Finally, our enrichment-of-bioactives and diversity analyses were performed on only two libraries—the only two publicly accessible, large-scale screening datasets for *C. elegans*. When applying our model to new compound sets, the enrichment rates and structural diversities of top-scoring compounds will depend on the characteristics of the compound sets being ranked and the phenotypes assayed.

The structural scaffolds we have identified provide the first guidelines to better design new small-molecule libraries intended for chemical screens with *C. elegans* and perhaps other nematodes. Notably, the biaryl scaffold, the piperazine scaffold and the chromenone substructure in the 2- or 3-phenyl-chromen-4-one scaffolds have all been previously identified as privileged substructures^{34,35}. A privileged substructure is defined as “a single molecular framework able to provide ligands for diverse receptors”^{34,36}. Specificity can be achieved by varying the substituents that decorate the privileged scaffold^{37–39}. For example, the biphenyl scaffold, which accounts for

one-third of the accumulating 'unfused' biaryls in worms, is found in 4.3% of all known drugs, representing molecules from diverse therapeutic classes^{34,38}. Indeed, statistical analysis of NMR-derived binding data for 10,080 compounds (represented by 104 sub-structural fragments) and 11 protein targets identified the biphenyl scaffold as a privileged substructure that preferentially binds proteins³⁸. These results are encouraging, as molecular scaffolds that promote bioavailability in *C. elegans* also promote physical interactions with diverse protein targets.

Our analysis of nematicide accumulation in worms revealed that 40% of these compounds accumulate as metabolites in the worm. This result suggests that compounds may be processed to their bioactive form by the worm, which is analogous to how prodrugs like heroin and levodopa are metabolized to their bioactive form in humans^{40,41}. Hence, for a given bioactive compound discovered through an *in vivo* screen, it is not a certainty that the parent structure is the bioactive species. This insight is especially relevant to those who seek the target of a compound using biochemical approaches, which may not be fruitful if the bioactive agent is a metabolite of the parental compound.

Here, we have established our high-throughput HPLC-based accumulation assay as a powerful approach to investigate the interaction of a whole animal with a constellation of chemical structures in its environment. Relative to the HPLC protocols used with other model systems⁴², the chromatographic output of our worm HT-HPLC assay yields a high signal-to-noise ratio because the majority of endogenous worm material elutes in the flowthrough. In combination with the wide spectral range of the diode array detector, the low background of our assay allows for the easy identification and quantification of small molecules and their metabolites from worm lysates. As a result, we now have a better understanding of the property space occupied by drug-like molecules that accumulate in *C. elegans* and have devised methods to circumvent its xenobiotic defenses.

METHODS

HPLC-based small-molecule accumulation assay. Late-stage fourth-larval-stage worms, grown from synchronized hatchlings at 25 °C for 45 h on NA22 *Escherichia coli*, were used for the accumulation assay. The worms were harvested, washed at least twice and re-suspended in enough M9 buffer¹ for a final concentration of ~10 worms per μl . Five hundred microliters of this worm suspension was added to each well of Pall Acroprep 96-well filter plates (0.45- μm GHP membrane, 1-ml well volume). Chemicals were added to each well to a final concentration of 40 μM (0.4% DMSO, v/v). Worms were incubated in the small-molecule solutions at 20 °C for 6 h with aeration, after which the incubation buffer was drained from the wells by vacuum (6 h is the longest time allowed before the filter membranes weaken). The worms were then washed three times with 500 μl of M9 buffer. After washing, the worms were resuspended in 50 μl of M9 buffer, transferred to new 96-well solid-bottom plates and stored frozen at -20 °C. The samples were later lysed by adding 50 μl of a 2 \times lysis solution (100 mM KCl, 20 mM Tris, pH 8.3, 0.4% SDS, 120 $\mu\text{g ml}^{-1}$ proteinase K) to each well and incubating the plates at 60 °C for 1 h with agitation. After lysis, the plates were stored frozen at -80 °C for later processing by HPLC (see **Supplementary Methods** for the full HPLC methods).

Cheminformatics and machine learning. The cheminformatic package in Pipeline Pilot version 6.1 (Scitegic Inc. Accelrys) was used to standardize the representation of all compounds studied, including removing inorganic compounds, salts and duplicates. Pipeline Pilot was also used for all Naive Bayes statistical model building. The Naive Bayesian structure-based accumulation model (SAM) was built using the Extended Connectivity Fingerprints (ECFP₄) method²⁵, in which the compounds are represented by overlapping fragments of a diameter of up to four bond lengths. The SAM was validated using a five-fold cross validation procedure, where four-fifths of the data are used to train the model and the remaining one-fifth are used to test the model. This procedure is run five times, and each compound appears in the test set once and the training set four times. The accuracy of the SAM was measured using the enrichment rate. The enrichment rate is calculated by ranking all compounds in the test set using the model, and then comparing the number of actives found in the top *n*% to the number of expected actives for *n*%. The final model was built using all of the compounds in the training set. When ranking other datasets with the SAM, compounds were filtered from the dataset before ranking if they had >85% structural similarity, using the Tanimoto

score with ECFP₄ fingerprints (see below), to any compound in the training set used to build the model.

All similarity calculations were carried out using the Tanimoto coefficient with the ECFP₄ fingerprinting method. The Tanimoto coefficient is the number of fragments in common between two compounds, divided by the total number of fragments present in both compounds. Murcko scaffolds³⁹ were generated using the 'Generate Fragments' protocol in Pipeline Pilot.

The SAM can be used directly by opening the script found in **Supplementary Data 2** using Pipeline Pilot. Alternatively, the training set (**Supplementary Data 3**) can be used in conjunction with Open Source software (see **Supplementary Data 4**) to build a similar model.

Analysis of metabolites by LC-MS. Metabolites were HPLC-purified from worm lysates and dried using a Savant DNA120 SpeedVac (acid was not added to the HPLC solvents). Chromatographic separations of the purified metabolites for LC-MS were performed using a nano-AQUITY Ultra Performance Liquid Chromatography (UPLC) system (Waters Corp.)—see **Supplementary Methods** for full methods. Mass spectrometry was performed using a Micromass Quadrupole-Time-of-Flight Premiere instrument (Waters Corp.). The data acquisition software used was MassLynx NT, version 4.0. Mass spectra were acquired in positive ion mode using a nano-ESI with capillary voltage and sample cone voltage set to 3,000 V and 20 V, respectively. The MS acquisition rate was set to 1.0 s, with a 0.1-s interscan delay. Ninety-eight percent argon gas was employed as the collision gas with collision energy varying from 13–46 V for the mass range of 100–1,000 *m/z*. Ions selected for LC-MS-MS were identified after manual analysis of original LC-MS runs, and a corresponding inclusion list was generated for targeted data-dependent acquisition experiments.

Received 5 October 2009; accepted 26 April 2010;
published online 30 May 2010

References

- Burns, A.R. *et al.* High-throughput screening of small molecules for bioactivity and target identification in *Caenorhabditis elegans*. *Nat. Protoc.* **1**, 1906–1914 (2006).
- Kwok, T.C.Y. *et al.* A small-molecule screen in *C. elegans* yields a new calcium channel antagonist. *Nature* **441**, 91–95 (2006).
- Petrasccheck, M., Ye, X. & Buck, L.B. An antidepressant that extends lifespan in adult *Caenorhabditis elegans*. *Nature* **450**, 553–556 (2007).
- Kokel, D., Li, Y., Qin, J. & Xue, D. The nongenotoxic carcinogens naphthalene and para-dichlorobenzene suppress apoptosis in *Caenorhabditis elegans*. *Nat. Chem. Biol.* **2**, 338–345 (2006).
- Kwok, T.C. *et al.* A genetic screen for dihydropyridine (DHP)-resistant worms reveals new residues required for DHP-blockage of mammalian calcium channels. *PLoS Genet.* **4**, e1000067 (2008).
- Jones, A.K., Buckingham, S.D. & Sattelle, D.B. Chemistry-to-gene screens in *Caenorhabditis elegans*. *Nat. Rev. Drug Discov.* **4**, 321–330 (2005).
- Kaminsky, R. *et al.* A new class of anthelmintics effective against drug-resistant nematodes. *Nature* **452**, 176–180 (2008).
- Kaletta, T. & Hengartner, M.O. Finding function in novel targets: *C. elegans* as a model organism. *Nat. Rev. Drug Discov.* **5**, 387–398 (2006).
- Broeks, A., Janssen, H.W., Calafat, J. & Plasterk, R.H. A P-glycoprotein protects *Caenorhabditis elegans* against natural toxins. *EMBO J.* **14**, 1858–1866 (1995).
- Rand, J.B. & Johnson, C.D. Genetic pharmacology: interactions between drugs and gene products in *Caenorhabditis elegans*. In *Methods in Cell Biology*, **48** (eds Epstein, H.F. & Shakes, D.C.) 187–204 (Academic, San Diego, 1995).
- Choy, R.K. & Thomas, J.H. Fluoxetine-resistant mutants in *C. elegans* define a novel family of transmembrane proteins. *Mol. Cell* **4**, 143–152 (1999).
- Cox, G.N., Kusch, M. & Edgar, R.S. Cuticle of *Caenorhabditis elegans*: its isolation and partial characterization. *J. Cell Biol.* **90**, 7–17 (1981).
- Avery, L. & Shtonda, B.B. Food transport in the *C. elegans* pharynx. *J. Exp. Biol.* **206**, 2441–2457 (2003).
- Lindblom, T.H. & Dodd, A.K. Xenobiotic detoxification in the nematode *Caenorhabditis elegans*. *J. Exp. Zool. A Comp. Exp. Biol.* **305**, 720–730 (2006).
- Jospin, M., Jacquemond, V., Mariol, M.C., Segalat, L. & Allard, B. The L-type voltage-dependent Ca²⁺ channel EGL-19 controls body wall muscle function in *Caenorhabditis elegans*. *J. Cell Biol.* **159**, 337–348 (2002).
- Franks, C.J. *et al.* Ionic basis of the resting membrane potential and action potential in the pharyngeal muscle of *Caenorhabditis elegans*. *J. Neurophysiol.* **87**, 954–961 (2002).
- Irwin, J.J. & Shoichet, B.K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).
- Herre, S. & Pragst, F. Shift of the high-performance liquid chromatographic retention times of metabolites in relation to the original drug on an RP8 column with acidic mobile phase. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **692**, 111–126 (1997).

19. Herzler, M., Herre, S. & Pragst, F. Selectivity of substance identification by HPLC-DAD in toxicological analysis using a UV spectra library of 2682 compounds. *J. Anal. Toxicol.* **27**, 233–242 (2003).
20. Lipinski, C.A., Lombardo, F., Dominy, B.W. & Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).
21. Kocisko, D.A. *et al.* New inhibitors of scrapie-associated prion protein formation in a library of 2000 drugs and natural products. *J. Virol.* **77**, 10288–10294 (2003).
22. Eddershaw, P. & Dickins, M. Phase I metabolism. in *A Handbook of Bioanalysis and Drug Metabolism* (ed. Evans, G.) 208–221 (CRC Press, Boca Raton, Florida, USA, 2004).
23. Manchee, G., Dickins, M. & Pickup, E. Phase II enzymes. in *A Handbook of Bioanalysis and Drug Metabolism* (ed. Evans, G.) 222–243 (CRC Press, Boca Raton, Florida, USA, 2004).
24. Xia, X., Maliski, E.G., Gallant, P. & Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **47**, 4463–4470 (2004).
25. Rogers, D., Brown, R.D. & Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* **10**, 682–686 (2005).
26. Durant, J.L., Leland, B.A., Henry, D.R. & Nourse, J.G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280 (2002).
27. Kerwar, S.S. Pharmacologic properties of fenbufen. *Am. J. Med.* **75**, 62–69 (1983).
28. Flower, D.R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **38**, 379–386 (1998).
29. Hert, J., Irwin, J.J., Laggner, C., Keiser, M.J. & Shoichet, B.K. Quantifying biogenic bias in screening libraries. *Nat. Chem. Biol.* **5**, 479–483 (2009).
30. Bemis, G.W. & Murcko, M.A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
31. Shelat, A.A. & Guy, R.K. Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.* **3**, 442–446 (2007).
32. Hoon, S. *et al.* An integrated platform of genomic assays reveals small-molecule bioactivities. *Nat. Chem. Biol.* **4**, 498–506 (2008).
33. Young, D.W. *et al.* Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* **4**, 59–68 (2008).
34. Horton, D.A., Bourne, G.T. & Smythe, M.L. The combinatorial synthesis of bicyclic privileged structures or privileged substructures. *Chem. Rev.* **103**, 893–930 (2003).
35. Klekota, J. & Roth, F.P. Chemical substructures that enrich for biological activity. *Bioinformatics* **24**, 2518–2525 (2008).
36. Evans, B.E. *et al.* Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **31**, 2235–2246 (1988).
37. Mason, J.S. *et al.* New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **42**, 3251–3264 (1999).
38. Hajduk, P.J., Bures, M., Praestgaard, J. & Fesik, S.W. Privileged molecules for protein binding identified from NMR-based screening. *J. Med. Chem.* **43**, 3443–3447 (2000).
39. Chen, Y. & Shoichet, B.K. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat. Chem. Biol.* **5**, 358–364 (2009).
40. Garzon-Aburbeh, A., Poupaert, J.H., Claesen, M. & Dumont, P. A lymphotropic prodrug of L-dopa: synthesis, pharmacological properties, and pharmacokinetic behavior of 1,3-dihexadecanoyl-2-[(S)-2-amino-3-(3,4-dihydroxyphenyl)prop anoyl]propane-1,2,3-triol. *J. Med. Chem.* **29**, 687–691 (1986).
41. Inturrisi, C.E. *et al.* Evidence from opiate binding studies that heroin acts through its metabolites. *Life Sci.* **33** Suppl 1: 773–776 (1983).
42. Hou, B., Lim, E.K., Higgins, G.S. & Bowles, D.J. N-glycosylation of cytokinins by glycosyltransferases of *Arabidopsis thaliana*. *J. Biol. Chem.* **279**, 47822–47832 (2004).
43. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).

Acknowledgments

We thank C. Cummins for critical comments on the manuscript; V. Wong and G. Selman for technical assistance early on in the project; S. Pan at the University of California, Riverside, for MS and MS-MS analyses; and A. Young of the Advanced Instrumentation for Molecular Structure Mass Spectrometry Laboratory at the University of Toronto for accurate mass MS analyses. This work was supported by Canadian Institutes of Health Research (CIHR) operating grants to P.J.R. (grant number 68813) and G.G. and C.N. (MOP-81340), Natural Sciences and Engineering Research Council of Canada support to G.D.B., a Natural Sciences and Engineering Research Council of Canada Graduate Scholarship doctoral award to A.R.B. and a Marie Curie Fellowship to I.M.W. G.G. and P.J.R. are Canadian Research Chairs in Chemical Biology and Molecular Neurobiology, respectively.

Author Contributions

A.R.B. did the wet-lab work and analyzed the MS data with guidance from S.R.C. and P.J.R. I.M.W. did the computational analysis with guidance from A.R.B., J.W., M.T., G.D.B., G.G., C.N. and P.J.R. The project was conceived by P.J.R., S.R.C. and A.R.B., and the paper was written by A.R.B. and P.J.R.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information and chemical compound information is available online at <http://www.nature.com/naturechemicalbiology/>. Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>. Correspondence and requests for materials should be addressed to P.J.R.