Neves, B.J.; Dantas, R.F.; Senger, M.R.; Melo-Filho, C.C.; Valente, W.C.; de Almeida, A.C.; Rezende-Neto, J.M.; Lima, E.F.; Paveley, R.; Furnham, N.; Muratov, E.; Kamentsky, L.; Carpenter, A.E.; Braga, R.C.; Silva-Junior, F.P.; Andrade, C.H. (2016) [Accepted Manuscript] Discovery of New Anti-Schistosomal Hits by Integration of QSAR-Based Virtual Screening and High Content Screening. Journal of medicinal chemistry. ISSN 0022-2623 DOI: https://doi.org/10.1021/acs.jmedchem.5b0203 (In Press)

Downloaded from: http://researchonline.lshtm.ac.uk/2965052/

DOI: 10.1021/acs.jmedchem.5b02038

Usage Guidelines

Please refer to usage guidelines at http://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by-nc-nd/2.5/

Discovery of New Anti-Schistosomal Hits by Integration of QSAR-Based Virtual Screening and High Content Screening

Bruno J. Neves,[†] Rafael F. Dantas,[‡] Mario R. Senger,[‡] Cleber C. Melo-Filho,[†] Walter C.G. Valente,[‡] Ana C. M. de Almeida,[‡] João M. Rezende-Neto,[‡] Elid F. C. Lima,[‡] Ross Paveley,[#] Nicholas Furnham,[#] Eugene Muratov,[§] Lee Kamentsky,[Ψ] Anne E. Carpenter,[Ψ] Rodolpho C. Braga,[†] Floriano P. Silva-Junior,[‡] Carolina Horta Andrade[†]**

[†]LabMol – Laboratory for Molecular Modeling and Drug Design, Faculdade de Farmácia, Universidade Federal de Goiás, Goiânia, Brazil

[‡]LaBECFar – Laboratório de Bioquímica Experimental e Computacional de Fármacos, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

[#]Department of Infection and Immunity, London School of Hygiene and Tropical Medicine, London, United Kingdom

[§]Laboratory for Molecular Modeling, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, USA

[Ψ]Imaging Platform, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA.

ABSTRACT

Schistosomiasis is a debilitating neglected tropical disease, caused by flatworms of *Schistosoma* genus. The treatment relies on a single drug, praziquantel (PZQ), making the discovery of new compounds extremely urgent. In this work, we integrated QSAR-based Virtual Screening (VS) of *S. mansoni* thioredoxin glutathione reductase (*Sm*TGR) inhibitors and high content screening (HCS) aiming to discover new antischistosomal agents. Initially, binary QSAR models for inhibition of *Sm*TGR were developed and validated using the Organization for Economic Co-operation and Development (OECD) guidance. Using these models, we prioritized 29 compounds for further testing in two HCS platforms based on image analysis of assay plates. Among them, 2-[2-(3-methyl-4-nitro-5-isoxazolyl)vinyl]pyridine and 2-(benzylsulfonyl)-1,3-benzothiazole, two compounds representing new chemical scaffolds, have activity against schistosomula and adult worms at low micromolar concentrations and therefore, represent promising antischistosomal hits for further hit-to-lead optimization.

KEYWORDS: schistosomiasis, thioredoxin glutathione reductase, quantitative structure-activity relationships, computer-assisted drug design, whole-organism phenotypic assay, high-content screening.

INTRODUCTION

Schistosomiasis is a neglected tropical disease caused by flatworms of the genus *Schistosoma*. These worms cause a chronic and often debilitating infection that impairs development and productivity, and exposure to these worms is strongly linked to extreme poverty.[1–4] Recent estimates of World Health Organization suggest that around 258 million people are infected resulting up to 200,000 deaths annually. Currently, schistosomiasis is endemic in 78 countries worldwide, mainly in sub-Saharan Africa, the Middle East, the Caribbean, and South America, where infections are mediated through poor knowledge about the disease, poor sanitation, and lack of effective health policies.[5]

In the absence of a vaccine, the control of schistosomiasis relies on a single drug, praziquantel (PZQ), which has been used in clinical practice for almost four decades.[6] However, because of high incidence of reinfection, the widespread and repeated use of this drug in endemic areas raises concerns about the development of drug resistance by the parasite.[7–11] This problem is further emphasized by the known lack of efficacy of PZQ against juvenile worms,[12] which is a potential cause of treatment failure in endemic areas. Hence, there is an urgent need for new anti-schistosomal drugs with novel mechanisms of action.

The complete genome sequencing of *S. mansoni*,[13,14] *S. japonicum*,[15] and *S. haematobium*[16] has provided new information on their biological pathways, identifying potentially relevant targets for therapeutic intervention.[17] Thioredoxin glutathione reductase (TGR) is one of these targets; it plays a crucial role in the redox homeostasis of the parasite.[18] TGR is a multifunctional enzyme that acts in the detoxification of reactive oxygen species (ROS) generated by digestion of red blood cells[19,20] and by the host immune system.[21,22] In mammalian cells there are two major systems to detoxify ROS, one is based on glutathione (GSH) and the other is based on

thioredoxin (Trx). In both systems, NADPH provides reducing equivalents via two specialized oxidoreductase flavoenzymes. Glutathione reductase (GR) reduces glutathione disulfide (GSSG) and drives the GSH-dependent systems, whereas Trx reductases (TR) are pivotal in the Trx-dependent system. On the other hand, in schistosomes, thiol redox homeostasis is completely dependent on TGR, which controls the NADPH reduction of GSSG and Trx in both systems.[23–25] Given these characteristics, it is expected that the maintenance of the homeostatic levels of Trx and GSH in schistosomes play a key role in a variety of cellular processes, such as defense against oxidative stress, DNA synthesis, detoxification, protein folding and repair.[26] Moreover, RNA interference studies have showed that inactivation of TGR of *S. mansoni* (*Sm*TGR)[18] and TGR of *S. japonicum* (*Sj*TGR)[27,28] has profound effects on worm survival rates both in culture medium and infected mice.

Due to the importance of TGR in parasite's redox balance, we hypothesized that known *Sm*TGR inhibitors listed on publicly available databases may serve as the chemical basis to discover new anti-schistosomal compounds by Virtual Screening (VS). Docking-based and pharmacophore-based approaches are the most popular VS strategies to identify putative hits in chemical libraries. However, in recent years, quantitative structure activity relationships (QSAR) models have been used widely in VS applications as well.[29–35]

The main goal of this study was the identification of new structurally dissimilar compounds with high anti-schistosomal activity. To achieve this goal, we designed a study with the following steps: (i) collection, rigorously curation, and integration of the largest possible dataset of *Sm*TGR inhibitors; (ii) development of rigorously validated and mechanistically interpretable binary QSAR models; (iii) application of generated models for VS of 3 subsets from ChemBridge library (~ 150,000 compounds); (iv) interpretation of developed models to derive

structural rules useful for targeted design of new inhibitors; and (v) experimental validation of prioritized/designed hits on live schistosomula and adult worms in two distinct HCS platforms. As a result of this study, we found that the QSAR models were efficient for prediction of new *Sm*TGR inhibitors and identified six novel anti-schistosomal hit compounds active against schistosomula and three hits active against adult worms. Among them, two hits, 2-[2-(3-methyl-4-nitro-5-isoxazolyl)vinyl]pyridine (**3**) and 2-(benzylsulfonyl)-1,3-benzothiazole (**4**), representing new chemical scaffolds structurally dissimilar to known inhibitors of *S. mansoni,* could be considered as promising anti-schistosomal agents.

RESULTS AND DISCUSSIONS

Dataset Balancing. Initially, thousands compounds with *Sm*TGR inhibition data were retrieved from the PubChem Bioassay database (AID: 485364) and used to build binary QSAR models. Further, un-curated chemical structures were standardized, duplicates were removed, and 2,854 compounds with reproducible potency (IC_{50}) ≤ 10 μM were considered as inhibitors, whereas the remaining 337,327 compounds were considered as non-inhibitors. Because the original dataset was highly unbalanced, i.e., 2,854 inhibitors and 337,327 non-inhibitors (1:118 ratio), it is not recommended building binary QSAR models for the entire data set. During model building, most machine learning methods need equal weighting of the classes in terms of both the number of instances and the level of importance (i.e., active class has the same importance as inactive class). Consequently, when trying to predict a minority class in an unbalanced dataset, machine learning methods are prone to assign most samples to the majority class, resulting in a large number of erroneous predictions for minority class.[36]

To reduce the number of the non-inhibitors and ideally maintain the "chemical space" of the original dataset, we evaluated the optimal number of representative compounds. To accomplish this task, we developed an under-sampling workflow based on *k*-nearest neighbors (*k*NN) distances of the each non-inhibitor to all inhibitors using the public available 166 substructures MACCS keys. We tested different sizes of the dataset by removing non-inhibitors, and changing the inhibitors-to-non-inhibitors ratios of 1:1 (balanced), 1:2, and 1:3.

In order to visualize the structural diversity of our dataset before and after balancing, we performed a principal component analysis (PCA). PCA transforms the original measured variables into new orthogonal variables called principal components, which are a linear combination of the original variables. Detailed results of structural diversity investigation are shown in Figure S1 (Supporting Information). The top two principal components retained 20% of the original information. Figure S1A represents the PCA plot of 2,854 inhibitors (blue dots) *vs.* all 337,327 non-inhibitors (grey dots). As we can see, the inhibitors are widely distributed across chemical space, reflecting significant chemical diversity. Figures S1B-D show the non-inhibitors selected with different ratios: 1:1 or 2,854 non-inhibitors, Figure S1B; 1:2 or 5,705 non-inhibitors, Figure S1C; and 1:3 or 8,562 non-inhibitors, Figure S1D. As we can see from the distribution of these dots, the most representative compounds were chosen that allowed minimal reduction of the original chemical space.

Performance of Individual QSAR Models. The balanced (ratio of 1:1) and unbalanced datasets (ratios of 1:2 and 1:3) were modeled by a combination of AtomPair,[37,38] MACCS[39–41] and Morgan fingerprints,[38,42] chemistry development kit (CDK)[43], and Dragon descriptors[44,45] along with eight machine learning methods leading to 120 different binary QSAR models (Tables S1, S2, and S3, Supporting Information). According to the statistical results of an 5-fold

external cross-validation procedure, we could draw three general conclusions: (i) random forest (RF), support vector machine (SVM), and gradient boosting machine (GBM) methods showed the best prediction ability among the eight tested machine learning methods; (ii) QSAR models built on balanced datasets are better than unbalanced (1:2 and 1:3 ratios) due to discrepant values between sensitivity (SE) and specificity (SP); the latter are prone to assign most samples as non-inhibitors, resulting in a large number of erroneous predictions; and (iii) the QSAR models which were built from the balanced dataset showed a high level of agreement between correct classification rate (CCR), SP, and SE values. Table 1 shows the detailed performances of the more predictive QSAR models derived from the balanced dataset.

Table 1. Summarized statistical characteristics of QSAR models developed with balanced dataset.

Model	CCR	*k*	SE	SP	Coverage
Morgan–RF	0.85	0.71	0.85	0.86	0.62
MACCS–RF	0.83	0.66	0.83	0.83	0.67
AtomPair–SVM	0.81	0.62	0.81	0.81	0.65
AtomPair–GBM	0.81	0.62	0.81	0.81	0.65
Dragon–SVM	0.85	0.70	0.85	0.84	0.69
Dragon–GBM	0.85	0.70	0.85	0.84	0.69
CDK–SVM	0.84	0.69	0.85	0.84	0.77
Consensus	0.87	0.74	0.87	0.88	1.00
Consensus rigor	0.91	0.81	0.96	0.87	0.38

RF: Random Forest; SVM: Support Vector Machine; GBM: Gradient Boosting Machine; CCR: correct classification rate; *k*: Cohen's kappa coefficient; SE: sensitivity; SP: specificity. Consensus and consensus rigor models were built by averaging the predicted values from the individual model for each machine learning technique (Morgan–RF, MACCS–RF, AtomPair–SVM, Dragon–SVM, and CDK–SVM).

The combination of Morgan fingerprints with RF (Morgan–RF), MACCS key with RF (MACCS–RF), AtomPair fingerprints with SVM (AtomPair–SVM) and GBM (AtomPair–GBM), Dragon descriptors with SVM (Dragon–SVM) and GBM (Dragon–GBM), and CDK

descriptors with SVM (CDK–SVM) led to more predictive QSAR models, with correct classification rate (CCR) ranging between 0.81–0.85 and coverage of 0.62–0.77 (Table 1). The best individual model was built using the combination of Morgan–RF (CCR = 0.85, SE = 0.85, and SP = 0.86).

To assure that the accuracy of the models was not due to chance correlation, 10 rounds of Y-randomization were performed for each dataset (Table S4). The results from this analysis (CCR values around 0.50) indicate that our models built using balanced dataset are statistically robust.

Performance of Consensus Models. Several individual QSAR models were generated using multiple machine learning algorithms and descriptors/fingerprints. However, our previous experience suggests that consensus models that combine individual QSAR models are advantageous[46–49] and naturally minimize prediction errors during a VS campaign. Therefore, consensus models were built by averaging the predicted values obtained after combining the individual models built using the balanced dataset. The detailed performances of 12 consensus models are given in Table S5. Among them, the consensus model built by combining the Morgan–RF, MACCS–RF, AtomPair–SVM, Dragon–SVM, and CDK–SVM (Tables 1 and S5, Supplementary Information) showed the best performance among all constructed consensus models (CCR = 0.87, SE = 0.87, and SP = 0.88). This consensus model discriminates inhibitors and non-inhibitors better than any of the individual QSAR models, with a 2% of increase in CCR, SE, and SP when compared with the best individual model (Morgan–RF).

In addition, the most rigorous consensus model (consensus rigor)[46] was built by combining five individual models with more restrictive conditions. A consensus rigor model only considers the outcome to be reliable when a compound was inside the applicability domain (AD) for the five models. If the compound was outside the AD for any model, then the outcome was specified

as unreliable. Expectedly, the combination of Morgan–RF, MACCS–RF, AtomPair–SVM, Dragon–SVM, and CDK–SVM models (Tables 1 and S5) also showed the best performance among all built consensus rigor models (CCR = 0.91, SE = 0.96, and SP = 0.87). In summary, the best consensus rigor model demonstrated better statistical results, with a 5% of increase in CCR, and 11% of increase in SE when compared with the best individual model (Morgan–RF). Although the AD of consensus rigor is limited only for certain chemical classes (coverage of 0.38), it has very high predictive power (CCR = 0.91).

Model interpretation. The Morgan-RF model exhibited the best predictive performance, and, consequently, it possesses the features that are best correlated with *Sm*TGR inhibition activity. Therefore, we translated its features (fingerprints) into predicted probability maps (PPMs) and visualized the atomic and fragment contributions predicted by the QSAR model (Figures 1 and 2). Atoms and fragments promoting the inhibition are highlighted by green (Figure 1); atoms and fragments decreasing the inhibitory potential are highlighted by purple (Figure 2) and gray lines (Figures 1 and 2) delimit the region of split between the favorable and the unfavorable contributions.[50]

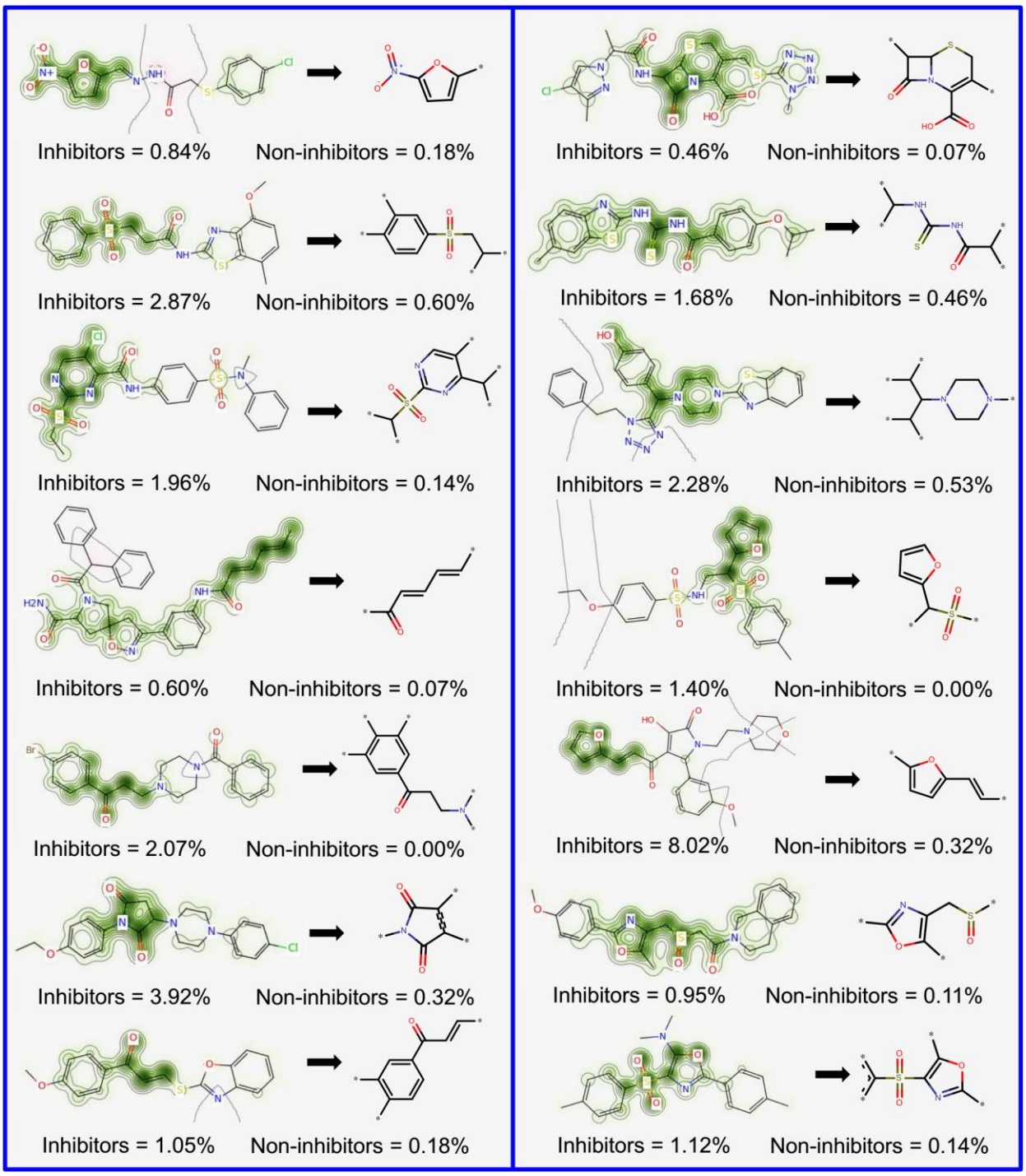


Figure 1. Favorable fragments (green) for *Sm*TGR inhibition predicted by the best individual QSAR model and their respective frequencies in inhibitors and non-inhibitors sets.

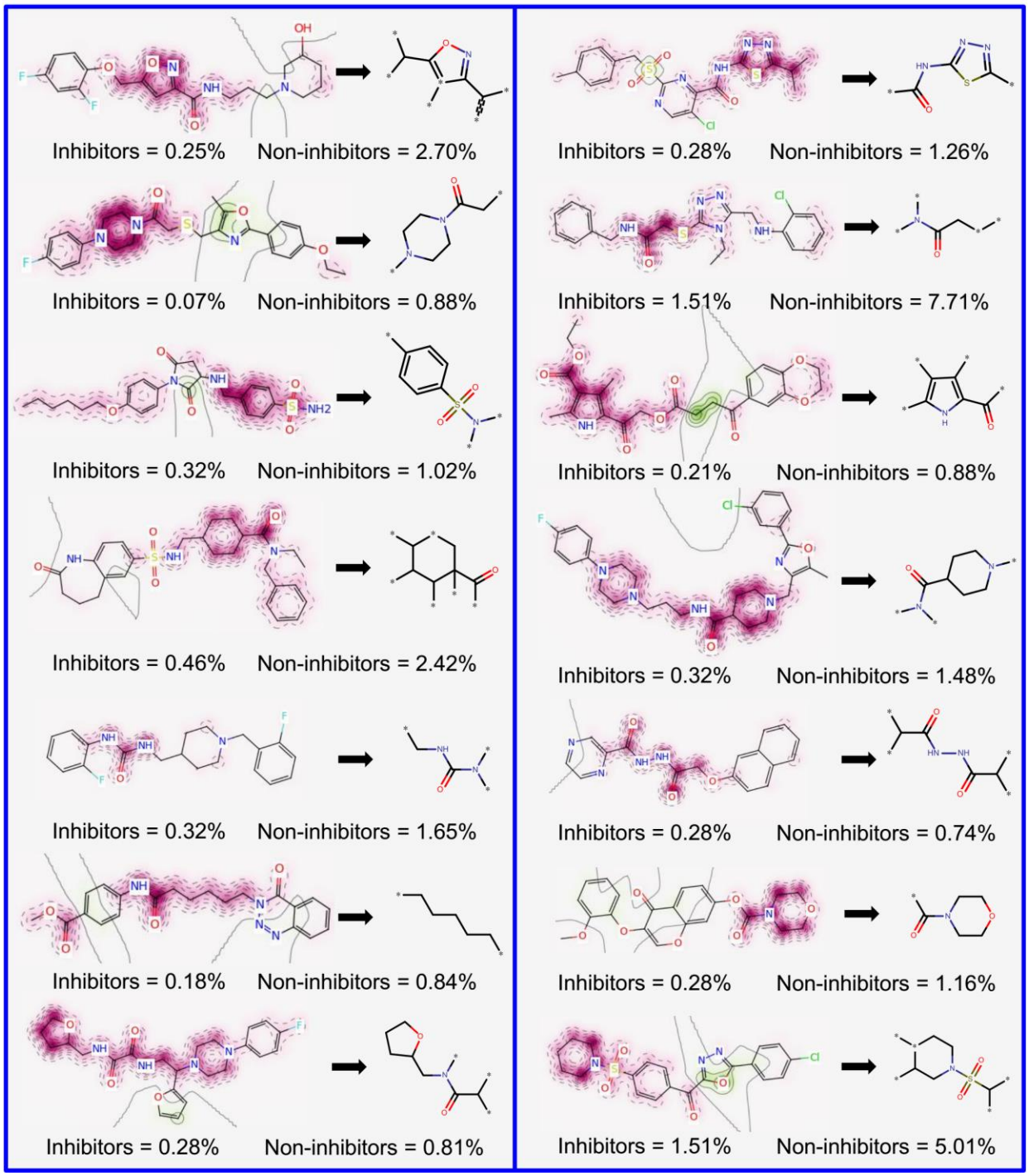


Figure 2. Unfavorable fragments (purple) for *Sm*TGR inhibition predicted by the best individual QSAR model and their respective frequencies in inhibitors and non-inhibitors sets.

Analyzing the fragments with favorable contributions highlighted by PPMs, we noticed that 14 fragments were more frequent in the inhibitors set and absent in the non-inhibitors set (Figure 1). Examples of favorable fragments for *Sm*TGR inhibition activity are nitrofuran, 2-ethenylfuran, (ethanesulfonyl) benzene, 2-(sulfonylmethyl) furan, carbonyl thiourea, and 4-methanesulfonyl-1,3-oxazole. By analyzing the fragments with unfavorable contribution into *Sm*TGR inhibition activity (Figure 2), several fragments such as benzylsulfonamide, methylurea, morpholine-4-carbonyl, piperidine-4-carboxamide, 1-methanesulfonylpiperidine, and cyclohexanecarbonyl, were more frequent in the non-inhibitors set. Compounds that contain these fragments may show a decreased *Sm*TGR inhibitory activity. This information could be useful for designing or optimizing new *Sm*TGR inhibitors by replacing unfavorable fragments by favorable fragments.

Reaction Mechanism of *Sm*TGR Inhibition. Although the inhibition mechanisms of most of the *Sm*TGR inhibitors are not well understood at the molecular level, the reaction mechanisms by which oxadiazole-2-oxides and cephalosporins operate could be identified according to a graphical interpretation of PPMs. However, for the best understanding of molecular inhibition mechanisms, it is important to highlight that the active site of *Sm*TGR is composed by a cysteine pair (Cys28/Cys31) in the glutharedoxin domain, a cysteine pair (Cys154/Cys159) in the thioredoxin domain, and a redox-active cysteine/selenocysteine pair (Cys596/Sec597) in the C-terminal tail. The latter should be highly mobile to accept electrons from the Cys154/Cys159 pair and to donate electron pairs to Cys28/Cys31 pair.[24] These amino acids provide the perfect chemical environment for covalent inhibition. The higher nucleophilicity and low pKa of the selenol group of Sec are thought to confer Sec a catalytic advantage over Cys at the attacking position.[51–53] Nonetheless, the thioredoxin domain contains His571 and Glu576, a catalytic dyad

that can facilitate proton abstraction of Cys159 thus impacting the catalytic efficiency of the thioredoxin domain of *Sm*TGR.[24]

We observed that the carbons 3 and 4 of the oxadiazole-2-oxide core presented the most important contributions for *Sm*TGR inhibition activity (Figure 3A−C). With PPMs information for this chemotype, a mechanistic rationale for inhibition was initiated through nucleophilic attack (presumably by a thiolate or selenoate of Cys or Sec, respectively) at either the position 3 or 4 of the oxadiazole ring and subsequent rearrangement of the heterocycle in a manner that allows release of the nitroxyl anion. An enzymatic oxidation is posited to transform this agent to nitric oxide (Figure 3D). These pieces of information corroborate with mechanism of inhibition proposed by Rai and colleagues[54] and mechanism of nitric oxide release in physiological solution under the action of thiols studied by Gasco and colleagues[55] In addition, PPMs indicated that the presence of amine-oxide group in core and electron-withdrawing substituents, such as carbonyl, at R1 and R2 positions are favorable for *Sm*TGR inhibition (Figure 3A), while removal of the amine-oxide group (Figure 3B) and presence of electron deficient substituents at R1 and R2 positions (Figure 3C) leads to modest potencies in terms of *Sm*TGR inhibition. These pieces of information corroborate with structure-activity relationships rules established by Rai and colleagues.[54]

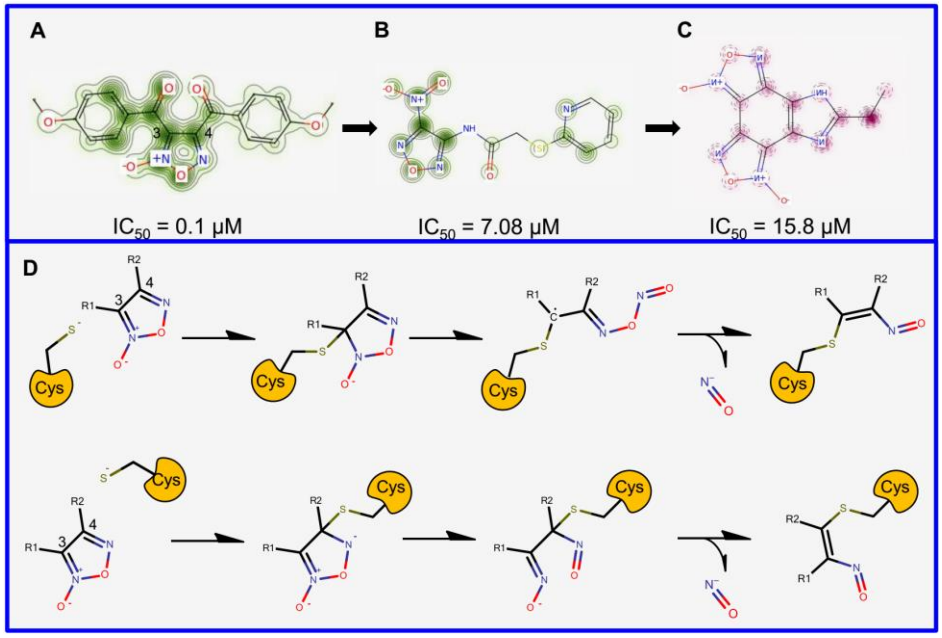


Figure 3. Predicted probability maps generated for oxadiazoles (A, B, and C) and their proposed reaction mechanism in the *Sm*TGR active site (D).

The reaction mechanism by which cephalosporins exert their *Sm*TGR inhibition activity was also proposed using the PPMs information (Figure S2A, Supporting Information). For both compounds, the PPMs picked up the positive contributions of the basic core structure of cephalosporines, more specifically carbon 8 and nitrogen 5 of β-lactam ring, and partially positive contribution of 1-methyl-5-tetrazolethione core for inhibition of *Sm*TGR. Based on these results, we suggest that inhibition of *Sm*TGR by cephalosporins may occur via a mechanism similar to proposed by Triboulet and colleagues,[56] i.e., a nucleophilic attack of Cys or Sec on β-lactam carbonyl carbon, with formation of a tetrahedral intermediate, which collapses with β-lactam ring opening by N5-C8 bond fission. Then, the acyl-enzyme intermediate could hydrolyze or react further, with expulsion of the 1-Methyl-5-tetrazolethione from carbon 3, generating a reactive methylene that could be trapped by other thiolate or selenoate (Figure S2B, Supporting Information).

QSAR-Based Virtual Screening. The QSAR-based VS was carried out following the workflow presented in Figure 4. Initially, 150,000 compounds available on PremiumSet, DIVERSet^{TM}-CL, and DIVERSet^{TM}-EXP libraries of ChemBridge were downloaded and prepared for VS. As drug-like ligands are highly desirable for the development of new leads with good oral bioavailability, we first filtered these libraries and excluded 1,285 compounds that violated Veber[57] and Lipinski's rules.[58] The remaining compounds were predicted by the consensus and consensus rigor models. To narrow down the compounds list and to obtain the highest level of confidence for each prediction, we took both the consensus score (average class prediction) and consensus model coverage into consideration. Consensus model coverage was defined as a fraction of individual models for which a compound was found to fall within the respective ADs. In that sense, introducing probability cutoffs can lead to predictions with higher confidence. Therefore, only putative hits with an average class number prediction of 1.0 and consensus model coverage over 50% were selected (470 putative hits). In addition, we removed compounds with previous bioactivity data reported against *Sm*TGR or *S. mansoni* and pan-assay interference compounds (PAINs)[59,60] so that selected compounds would be novel *Sm*TGR inhibitors and contain no PAINs structures. Finally, the compounds were evaluated by predicting a panel of properties including high aqueous solubility (CIQPlogS),[61] acceptable binding to human serum albumin (QPlogKhsa),[61] acceptable brain/blood partition coefficient (QPlogBB),[61] non-blocking or weak blocking of hERG channel,[46,47] absence of carcinogenicity and hepatotoxicity.[32] At the end of the VS workflow, 29 putative hits were visually inspected and acquired for biological evaluation (Table S6, Supporting Information).

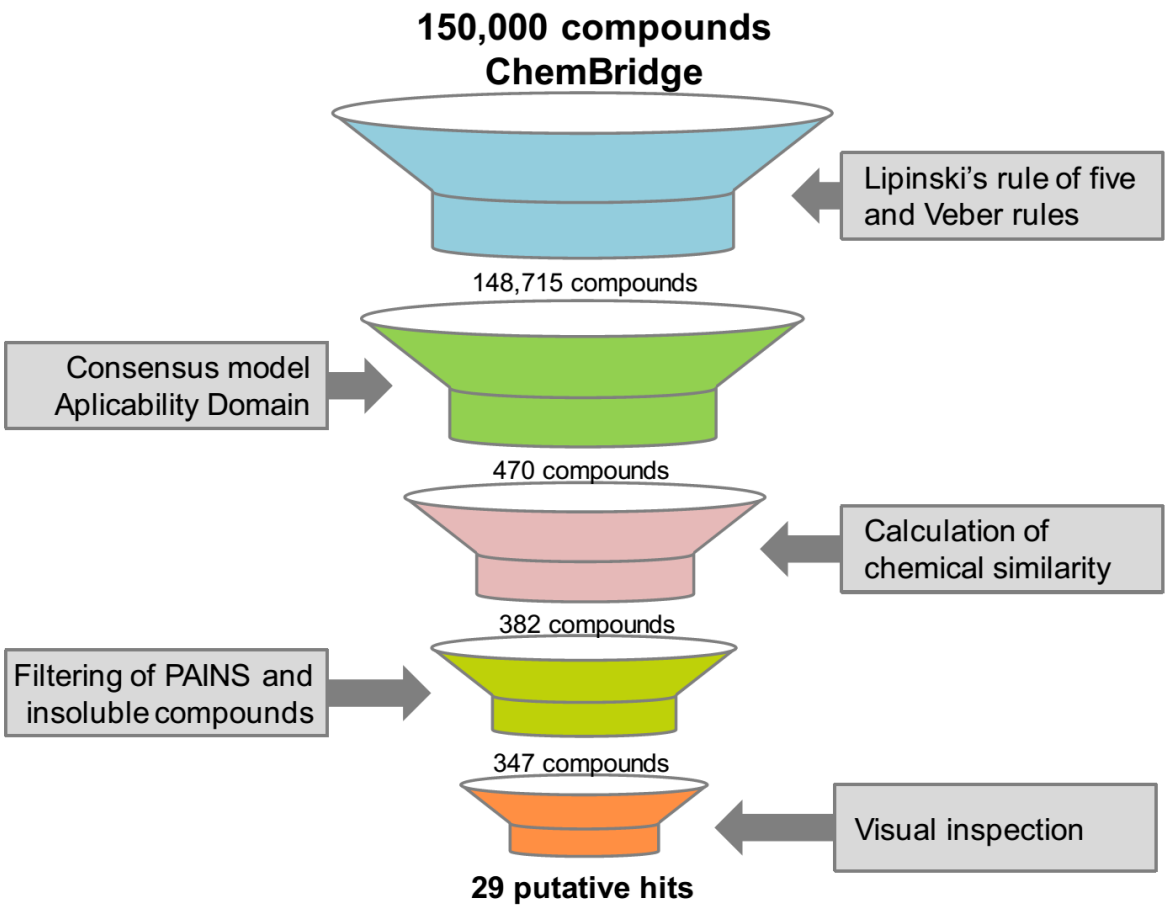


Figure 4. QSAR-based VS workflow used for identifying new compounds active against *S. mansoni*.

***Ex vivo* Activity Against Schistosomula.** Compared to target-based VS approaches, the traditional whole-organism schistosome screening approach (phenotypic screening) is an old but indispensable method to discover new anti-schistosomal agents. This phenotypic approach may be used to validate if the predicted *Sm*TGR-inhibitor interaction has anti-schistosomal activity. Moreover, a validated compound from a phenotypic assay must have been able to reach its target within the assayed organism only after crossing several biological membranes and resisting to degradation by detoxification enzymes. Hence, a hit coming from a phenotypic screen has much more biological value than one coming from a simple biochemical assay. Advances in automated

microscopes, liquid handling systems and computer-based image analysis programs have enabled the development of high-throughput phenotypic assays with cells or small whole organisms, a technique known as high-content screening (HCS).[62,63] HCS microscopes are able to capture high resolution images of live organisms in quick succession, a feature that has been explored to evaluate phenotypic and motility changes in schistosomula[64] or adult worms.[65,66]

Therefore, we employed a HCS assay to evaluate the biological activity of the selected compounds from virtual screening against the *S. mansoni* schistosomula. Assaying against this larval stage is commonly used as an initial screening step in antischistosomal drug discovery campaigns[67–72] since schistosomula are easier to obtain in larger numbers than adult worms. Of the 29 compounds tested against schistosomula, six were declared confirmed actives based on motility and phenotype scores at 20 µM after 48h of exposure (Table S6). The chemical structures of the six primary hits are shown in Figure 5.

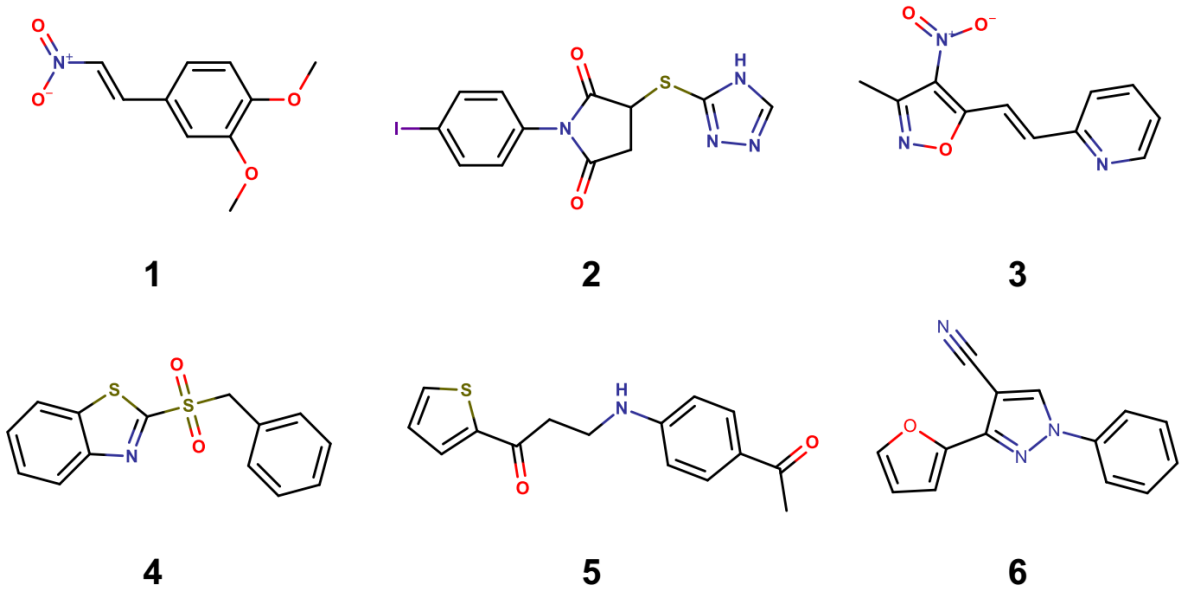


Figure 5. Chemical structures of six priority hits selected for further follow up.

Following the initial screening on schistosomula, the six primary hits were selected for determining half-maximal motility concentration (EC_{50}) at 0.31–20 µM range (Table 2 and Figure S3). Among primary hits, 1,2-dimethoxy-4-(2-nitrovinyl)benzene (**1**), 1-(4-iodophenyl)-3-(4H-1,2,4-triazol-3-ylthio)-2,5-pyrrolidinedione (**2**), 3-[(4-acetylphenyl)amino]-1-(2-thienyl)-1-propanone (**5**), and 3-(2-furyl)-1-phenyl-1H-pyrazole-4-carbonitrile (**6**) only showed inhibition activity at the highest tested concentration (>20 µM). On the other hand, 2-[2-(3-methyl-4-nitro-5-isoxazolyl)vinyl]pyridine (**3**) and 2-(benzylsulfonyl)-1,3-benzothiazole (**4**) showed efficacy in the same range of activity of the reference drug PZQ (EC_{50} = 1.90 µM), with EC_{50} values of 3.23 µM and 2.62 µM, respectively. This is an important feature for a new anti-schistosomal drug since modern lead discovery pipelines prioritizes compounds that possess bioactivity across the entire developmental cycle of the parasite in the mammalian host.[73,74]

Table 2. Biological activity data for hits of interest.

Compound	**Schistosomula EC_{50} (µM)**	**Adult EC_{50} (µM)**		**WSS-1 CC_{50} (µM)**	**Papain IC_{50} (µM)**
		Male	**Female**		
1	>20	29.8	5.77	17.48	>100
2	>20	10.2	17.9	133.40	>100
3	3.23	6.43	5.68	16.38	>100
4	2.62	21.1	4.91	28.49	>100
5	>20	N.D.	N.D.	N.D.	N.D.
6	>20	N.D.	N.D.	N.D.	N.D.
PZQ	1.90	0.22*	0.64	>400	N.D.

N.D. – not determined. * EC_{50} values produced for adult male after 72 h of exposure. WSS-1 human kidney epithelial cells were used to evaluate cytotoxicity.

Analysis of Phenotypic Profile. Compounds **3** and **4** promoted the internal disruption of larvae as evidenced by the appearance of multiple vacuoles as well as the rounding and darkening of the schistosomula (Figure 6). In order to evaluate if schistosomula response profile towards hits resemble those observed in the presence of known anti-schistosomal drugs (OLT, PZQ, dihydro-artemisinin, methyl-clonazepam, Ro15-5458, and oxamniquine), we applied a Bayesian treatment class model using phenotype scores.[64] This analysis indicated a shared target and/or mechanism of action between OLT and hits, and therefore, all six hits were classified as OLT-like compounds. At least in part, these results could be related to *Sm*TGR inhibition since OLT has already been identified as a noncompetitive inhibitor of this enzyme. It is also important to note that these phenotypic profile has been also observed after *Sm*TGR gene knockout.[18]

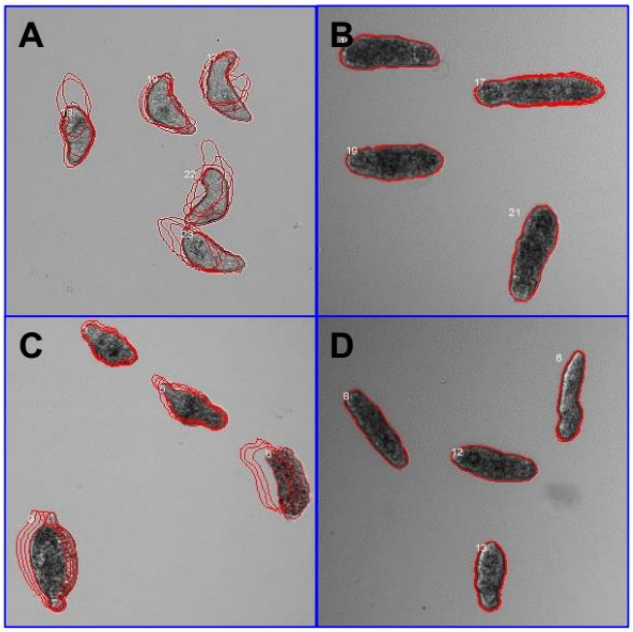


Figure 6. Phenotypes of schistosomula exposed for 48h to 0.625 % DMSO (control, A), 20 µM of **4** (B), and 10 µM of PZQ (C) and OLT (D). The outlines represent the position of each parasite over 5 time points (11s interval).

***Ex vivo* Activity on Adult Worms.** Our next step was to investigate if the compounds identified as hits for schistosomula also had an effect on adult *S. mansoni* worms.[65,66] Therefore, we employed a new HCS platform recently developed by our group that allows for systematic evaluation of gender-, dose-, and time-dependent drug effects on individual male and female parasites by measuring over 100 image features related to worm motility and morphology. Previously, we have demonstrated the successful application of this platform in identification of potent antischistosomal hit compounds.[65,66] In this study, four compounds (**1**−**4**) were screened at 0.1−100 μM concentrations for incubation times varying from 0h (immediately after compound addition to culture medium) to 72h.

Inspection of the measured features suggested that at least three features were able to distinguish active from inactive compound concentrations or the DMSO control: the Overlap_RandIndex feature, which is related to motility, the intensity and the area of the identified worm object. Figure 7 shows a 3D plot of these relevant features for individual female worms exposed to the investigated compounds at 20 μM concentration as well as for the PZQ and negative control (treated with 0.1% DMSO) after 48h incubation. The sample images are shown to exemplify the phenotypes that can be captured by these features. In general, the feature most correlated to the antischistosomal activity of these compounds was the Overlap_RandIndex, which roughly measures the difference in worm position from one time-lapse frame to the next and is inversely proportional to worm motility in a scale varying from 0 to 1. For simplicity, we hereafter refer to this feature as the “motility score”.

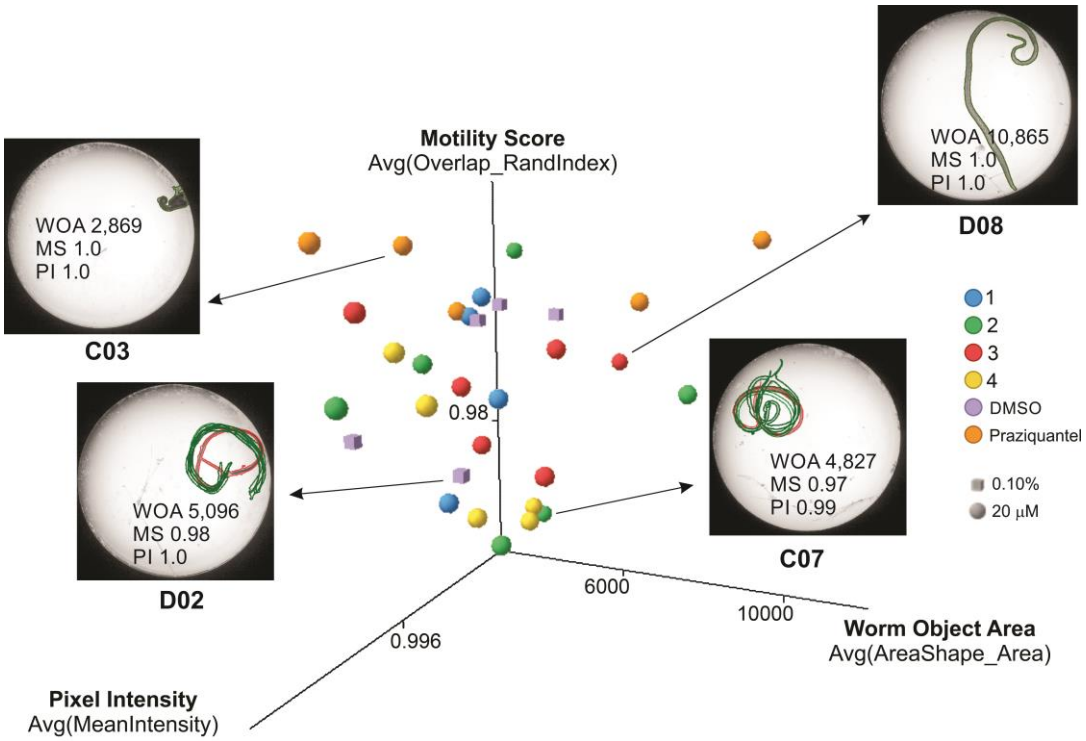


Figure 7. 3D scatter plot of the top three image features correlated to antischistosomal activity of the investigated compounds on female *S. mansoni* worms after 48 drug exposure. Each point in the graph represent a well/condition in the assay. Sample images are shown for selected wells to illustrate the different phenotypes captured by these three parameters (Object Area; Mobility Score; and Pixel Intensity of the worm object). The green outlines represent the position of each parasite over 5 time points (3s interval) overlaid on the initial position (red outline).

In order to determine the potency of the hit compounds against adult worms with the reference drug PZQ, we have determined EC_{50} values from dose response curves against male and female worms with varying incubation times (Table S7 and Figure S7, Supporting information). Compounds showed motility inhibition potencies against adult worms ranging from 4.91 to 35 µM, depending on incubation time and gender (Table 2). Overall, inhibition was fully achieved after 48h of incubation (Table 2). Compound **3** was the most active with EC_{50} around 6.00 µM

for both genders. Compounds **1** and **4** showed satisfactory potencies (<10 μM) for females worms, with EC_{50} = 5.77 μM and 4.91 μM, respectively, but not for male worms. Compound **2** was the less potent with EC_{50} values of 10.2 and 17.9 for male and female, respectively. Despite the satisfactory potencies displayed, all compounds had a less pronounced effect on adult worms than PZQ at all incubation times (EC_{50} values ≤ 0.66 μM, see Table S2). Results also indicated that female worms and schistosomula are slightly more sensitive to compounds action since they showed on EC_{50} values up to 5−8 times lower than those determined in males. In part, this could be due to a gender-specific expression pattern of *Sm*TGR and immature antioxidant system of the schistosomula. In fact, schistosomula express lower levels of *Sm*TGR than adults, which make them more susceptible to oxidative damage caused by inhibitors.[75,76]

Cytotoxicity Against Human Cells. Compounds **1**−**4** and PZQ were further evaluated for its cytotoxicity against human epithelial cells (WSS-1) from human kidney using a resazurin-based viability assay (Table 2). PZQ showed the lowest cytotoxicity, exhibiting half-maximal cytotoxic concentration (CC_{50}) above 400 μM. Compounds **2**, **3**, and **4** only were cytotoxic in concentrations higher than those necessary for antischistosomal activity. Compound **2** was the least cytotoxic compound (CC_{50} = 133.40 μM), followed by **4** (CC_{50} = 28.49 μM), **1** (CC_{50} = 17.48 μM), and **3** (CC_{50} = 16.38 μM).

Controls for Non-specific Inhibition and Off-Target Effects. Colloidal aggregates have long plagued early drug discovery. When a colloid is formed, membrane and soluble proteins adsorb to its surface and are partially denatured, leading to nonspecific inhibition and occasionally activation.[77,78] Therefore, adult worms were co-incubated with investigated compounds (at 20 μM and 100 μM) and detergent Triton X-100 (0.01%) and their antischistosomal effect was compared with activities obtained without detergent for excluding a

possible promiscuous colloidal aggregate effect. No significant differences were observed after comparison of inhibition activities of both groups, showing that antischistosomal activity of the hit compounds is related with specific inhibition (Figure S5). Further, we also investigated possible off-target effects of the hit compounds towards nucleophilic thiols in a papain inhibition assay. Again, none of the anti-schistosomal hits showed significant inhibition of papain at 100 μM while positive control E-64 fully inhibits this enzyme at 20 μM concentration (Table 2 and Figure S6).

CONCLUSIONS

To the best of our knowledge, this is the first study integrating QSAR-based VS and HCS methods to discover new anti-schistosomal agents. We have developed robust and predictive QSAR models for anti-schistosomal activity. Developed models were used in the most conservative way, i.e., in consensus fashion with the strictest AD criteria, for virtual screening of three ChemBrigde datasets: DIVERSet^{TM}-CL, DIVERSet^{TM}-EXP, and PremiumSet. As a result, 470 putative *Sm*TGR inhibitors were identified. Then, 29 compounds were selected and tested against *S. mansoni* schistosomula using a HCS platform and six of them showed significant inhibition activities at 20 μM. Among them, compounds **3** and **4** showed inhibitory effect equivalent to PZQ, with EC_{50} values around 2.50 μM. Both hits were also classified as OLT-like compounds, indicating a shared target with OLT, which has already been identified as an inhibitor of *Sm*TGR.[79] The results of gender-, dose-, and time-dependent inhibitory effect indicated that adult female worms of *S. mansoni* are slightly more sensitive than males to compounds action. Compounds **3** and **4** showed satisfactory potencies for female worms, with EC_{50} values around 6.00 μM. Both compounds also demonstrated low cytotoxicity to WSS-1

mammalian cells (CC_{50} > 16 μM) and inhibition of papain only in concentrations >100 μM. Finally, both compounds represent new chemical scaffolds, which are structurally dissimilar to known inhibitors of *S. mansoni,* and thus can be considered as new hit compounds for further chemical optimization.

EXPERIMENTAL SECTION

Computational

Dataset. The QSAR models were developed according to best practices of predictive QSAR modeling,[80,81] which is fully compliant to Organization for Economic Co-operation and Development (OECD) guidance on development and validation of QSAR models, such as (*i*) a defined end point, (*ii*) an unambiguous algorithm, (*iii*) a defined domain of applicability, (*iv*) appropriate measures of goodness-of-fit, robustness, and predictivity, and (*v*) mechanistic interpretation.[82] All *in silico* steps developed in this study were implemented in a publicly available KSAR workflow (http://labmol.farmacia.ufg.br/ksar). The KSAR workflow is tightly integrated with R and KNIME and includes many modules, such as the module for preparing the data, PCA, building of QSAR models, and VS.[46,83] We first retrieved 359,841 compounds containing half maximal inhibitory concentration (IC_{50}) data for the *Sm*TGR enzyme from the PubChem BioAssay database (AID: 485364). Compounds with inconclusive IC_{50} results were considered experimental errors and thus were not included in this study to avoid noise in model building. A total of 2,854 out of these 359,841 compounds had reproducible potency ($IC_{50} \leq 10$ μM) and were considered as inhibitors, whereas the remaining 356,987 compounds were considered as non-inhibitors.

Dataset Curation. Each compound of dataset was carefully standardized according to the protocol proposed by Fourches and colleagues.[84,85] Briefly, explicit hydrogens were added, salts were removed, whereas specific chemotypes such as aromatic and nitro groups were normalized using ChemAxon Standardizer (v.6.1.2, ChemAxon, Budapest, Hungary, http://www.chemaxon.com). Polymers, inorganic salts, organometallic compounds, and mixtures were also removed. In addition, 4,437 compounds with multiple *Sm*TGR measurements were identified during analyses of duplicates. Further analysis showed high concordance (99.9%) of duplicated records. In addition, 345 compounds with molecular weight greater than 700 Daltons were removed. In the end, the prepared dataset contained 2,854 inhibitors and 337,327 non-inhibitors.

Molecular Fingerprints and Descriptors. Three different types of fingerprints were used in this study: the Morgan fingerprint, a RDKit implementation[38] of the extended-connectivity fingerprints,[42] with radius of 2 and bit vector of 1024 bits; the molecular access system (MACCS) structural key fingerprints;[39–41] and the AtomPair fingerprints (RDKit implementation[38] of the Carhart's atom pairs)[37] with bit vector of 1024 bits. All the fingerprints were calculated by the open-source cheminformatics toolkit RDKit v.2.4.0.[86] A brief description of Morgan, AtomPair and MACCS fingerprints is available in Supporting Information.

The Chemistry Development Kit (CDK, v.1.4.19, GNU Lesser General Public License) descriptors and 0-2D descriptors were calculated using the PaDEL-Descriptor program[43] and DRAGON (v.5.5, Talete SRL, Milan, Italy) programs, respectively. The complete list of CDK descriptors and a detailed discussion for DRAGON descriptors can be found elsewhere.[44,45] The descriptors matrix was then normalized and constant/near constant and highly correlated ($r \geq 0.9$) descriptors were removed.

Dataset analysis and under-sampling. Because the original library was highly unbalanced (2,854 inhibitors and 337,327 non-inhibitors), it is not recommended building binary QSAR models for the entire data set. Thus, we decided to balance the dataset. Unlike the traditional under-sampling methods which randomly balance the dataset, our linear under-sampling strategy retains most of the representative structures of the non-inhibitors set, thus ensuring as high as possible coverage of original chemical space. The basic principle here is to measure the whole inhibitors matrix represented by the MACCS key fingerprints evaluating the Euclidean distance to the MACCS key fingerprints of each non-inhibitor using a *k*NN method,[87] implemented in R software v.3.0.3.[88] Then, the samples on non-inhibitors set were linearly extracted over the whole set by using *k*-distances and were used to generate balanced and partially balanced datasets. Finally, we generated three under-sampled datasets with inhibitor-to-non-inhibitor ratios of 1:1 (2,854 inhibitors and 2,854 non-inhibitors), 1:2 (2,854 inhibitors and 5,705 non-inhibitors), and 1:3 (2,854 inhibitors and 8,562 non-inhibitors).

Machine Learning Implementation. The building and optimization of statistically acceptable QSAR models requires a close combination between chemical information (i.e., fingerprints or descriptors) and several machine learning classifiers. For this reason, eight different machine learning classifiers, including the SVM with the radial basis Kernel function,[89] the RF,[90] GBM,[91] and partial least squares discriminant analysis (PLS-DA)[92] approaches, classification and regression trees (CART),[93] *k*NN with Euclidean distance,[87] multi-layer perceptron (MLP),[94] and multivariate adaptive regression splines (MARS)[95] were used. All machine learning classifiers were implemented using the R v.3.0.3.[88] A brief description about the theory of each machine learning method is described in Supporting Information.

5-fold external cross-validation. The full dataset of compounds with known inhibition activities is randomly divided into five subsets of equal size; then one of these subsets (20% of all compounds) is set aside as an external validation set and the remaining four sets together form the modeling set (80% of the full set). This procedure is repeated five times allowing each of the five subsets to be used as external validation set. Models are built using the modeling set only, and it is important to emphasize that the compounds in momentary external set (fold) are not employed either to build or select the models.

Applicability Domain. The AD for each descriptor or fingerprint type was estimated based on the Euclidean distances among the training set of each QSAR model generated in the external 5-fold cross-validation procedure. The distance of a test compound to its nearest neighbor in the training set was compared to the predefined AD threshold level. If the distance was greater than this threshold level, the prediction was considered to be less trustworthy.[96] In this study, we defined AD as a distance threshold D_T between a compound under prediction and its closest nearest neighbors of the training set. It was calculated as follows:

$$D_T = \bar{y} + Z\sigma \qquad (1)$$

Here, $\bar{y}$ is the average Euclidean distance of the k nearest neighbors of each compound within the training set, σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. We set the default value of this parameter Z at 0.5. Thus, if the distance of the external compound from all of its nearest neighbors in the training set exceeds this threshold, the prediction is considered unreliable.

Evaluation of Performance and Robustness. To access the predictive performance of the binary QSAR models, SE, SP, and CCR were used. These statistic metrics are calculated by the following equations:

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{3}$$

$$\mathrm{CCR} = \frac{\mathrm{SE} + \mathrm{SP}}{2} \tag{4}$$

Here, N denotes the total number of compounds, TP and TN represent the number of true positives (correct classifications of inhibitors) and true negatives (correct classifications of non-inhibitors), respectively, while FP and FN represent the number of false positives (incorrect classifications of inhibitors) and false negatives (incorrect classifications of non-inhibitors), respectively.

In addition to the above model evaluation metrics, Cohen's kappa (k) was used to measure the agreement between model predictions and experimental data.[97] This statistical parameter is calculated by the following equations:

$$\mathrm{Pr}(a) = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{N}} \tag{5}$$

$$\mathrm{Pr(e)} = \frac{(\mathrm{TP} + \mathrm{FP})\ \mathrm{x}\ (\mathrm{TP} + \mathrm{FN}) + (\mathrm{TN} + \mathrm{FN})\ \mathrm{x}\ (\mathrm{TN} + \mathrm{FP})}{\mathrm{N}} \tag{6}$$

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (7)$$

Here, Pr(a) represents the relative observed agreement between the predicted classification of the model and the known classification, and Pr(e) is the hypothetical probability of chance agreement. In the end, k analysis returns values between −1.0 (no agreement) and 1.0 (complete agreement), but values between 0.6 and 1.0 denote that the model is predictive. Finally, to further assure that the robustness of the models was not due to chance correlation, 10 rounds of Y-randomization were performed for each constructed model.

Consensus Modeling. After the building of QSAR models using all pairwise combinations of different types of chemical descriptors/fingerprints and various machine learning methods, the best models were used for consensus modeling, which can be derived by calculating an average for individual models. In consensus modeling, the final predicted value for each compound is estimated by including an average of the predicted values from the set of QSAR models. Thus, the averaged predicted activity for each compound is in the [0, 1] range. Formally, compounds with the predicted activity higher than 0.5 are classified as inhibitors, and those with the predicted activity lower than 0.5 are classified as non-inhibitors. Obviously, the closer the average predicted value is to 1 or 0, the higher the concordance among all models and the higher our confidence is in the classification of compounds as inhibitors or non-inhibitors, respectively.

Mechanistic Interpretation. To explore favorable or unfavorable structural fragments for *Sm*TGR inhibition, the PPMs were generated to visualize the atomic and fragment contributions predicted by the best QSAR model.[50]

Virtual Screening. The purpose of VS is to identify in a library of chemicals a subset of compounds with the desired properties based on computational calculations. Here the DIVERSet^{TM}-CL, DIVERSet^{TM}-EXP, and PremiumSet diversity datasets taken from the ChemBridge database were screened to identify inhibitors of *Sm*TGR. Prior to screening, the datasets were curated in the same way as modeling set (see Data Curation section) and filtered using the Veber[57] and Lipinski's rules[58] to obtain drug-like compounds. Fingerprints and molecular descriptors were generated for all compounds and normalized (except fingerprints) based on the minimum and maximum values of each descriptor of the modeling set. Then, best consensus and consensus rigor models were used to predict the *Sm*TGR inhibition activity of compounds. The prediction results were accepted only when the compound was found within the applicability domains of more than 50% of all models used in consensus prediction. In addition, to estimate the structural novelty of putative hits, we calculated the pairwise Tanimoto coefficients (using MACCS key fingerprints) between each screened putative hit and compounds in the full dataset of *Sm*TGR inhibitors. Then, putative hits with previous bioactivity data against *Sm*TGR or *S. mansoni* were identified and PAINS were removed using a workflow developed by Saubern and colleagues.[98] Finally, hits were imported into Maestro workspace v.9.3 and their aqueous solubility (CIQPlogS), binding to human serum albumin (QPlogKhsa), brain/blood partition coefficient (QPlogBB) properties were predicted using QikProp v.3.4[61] and hERG inhibition, carcinogenicity and hepatotoxicity were predicted using the Pred-hERG server,[46,47,99] admetSAR server,[100,101] and PaDEL-DDPredictor program,[102,103] respectively.

Experimental

Materials. Investigated compounds were purchased from ChemBridge (San Diego-CA, USA), resuspended in 100% DMSO and used immediately in the assays. It is important to mention that

all chemical structures were confirmed using proton (^{1}H) NMR spectra at 300/400 MHz and Liquid Chromatography–Mass Spectrometry (LC-MS) analysis with evaporative light scattering and ultraviolet detectors confirmed a minimum purity of 95% for all compounds (spectrums of compounds are listed in Supporting Information). DMEM and M169 media were purchased from Vitrocell Embriolife (Campinas-SP, Brazil). All other reagents were purchased from Sigma-Aldrich (St. Louis-MO, USA).

Automated *ex vivo* Larval *S. mansoni* HCS Assay. Cercarie (*S. mansoni*, BH strain) were vortexed at maximum speed for 5 minutes for tail shedding and transformation into schistosomula by an adapted method from literature.[104,105] Briefly, schistosomula were resuspended in Medium 169, placed in 384 well plates (120 per well) and maintained in an incubator with 5% CO_2 overnight before compound addition. The worms were then incubated with investigated compounds and PZQ at 0.31–20 µM concentrations or DMSO (0.625%). The effect of the compounds on schistosomula motility and phenotypes was assessed at 48h after compound addition using an automated Bright-field ImageXpressMicro HCS microscope (IXM; Molecular Devices, Wokingham, UK). For motility analysis 5 x 11 sec interval time-lapse images were collected using a 4x objective. For detailed morphology, a 10x objective was used to collect 4 adjacent images fields from within a well in order to increase the number of schistosomula for phenotype analysis. Analysis of both the larval phenotype and motility was then carried out in Pipeline Pilot 9 as described by Paveley and colleagues.[64] Phenotype analysis of individual parasites was carried out by a two class Laplacian-modified Bayesian categorization analysis of 80 image descriptors which constituted shape, size, image intensity, and texture statistics and compared to a training set of data comprising 20,000 parasites. Motility analysis of individual parasites was also carried out by the average object displacement from the

origin point in subsequent 4x image across the time-frame series. Both the Bayesian phenotype and motility scores were subsequently adjusted to the control wells (DMSO treated) on each plate.[64]

Automated *ex vivo* Adult *S. mansoni* HCS Assay. After 42-49 days post percutaneously infection of infant Swiss mice with 150 ± 10 *S. mansoni* cercariae (BH strain), animals were euthanized, and worms perfused from portal hepatic and mesenteric veins. Male and female parasites were rinsed and individually transferred into 96 well plates with complete DMEM media (i.e. DMEM plus 10% fetal calf serum, 2mM L-glutamine, 100 µM/ml penicillin, 100 µg/ml streptomycin). The plates were maintained overnight at 37 °C in a humidified atmosphere of 5% CO_2. Further, worms were then incubated up to 72h with 0.10−100 µM of selected compounds and PZQ or negative control DMSO at 0.1%. The effect of the compounds on adult worm motility or phenotype was assessed either immediately 24, 48 or 72h after compound addition using a newly developed HCS assay. Our method uses 100 time-lapse images captured every 250-300 ms with an automated bright-field microscope using a 2x objective lens (ImageXpress Micro XLS, Molecular Devices, CA). Subsequent quantitative image analysis used a custom-developed pipeline for detecting changes in parasite motility and morphology using the open-source CellProfiler software v. 2.1.2.[106] The pipeline along with its validation will be thoroughly described in a subsequent publication and the pipeline itself is freely available (www.cellprofiler.org/published_pipelines.shtml). Briefly, our strategy for motility measurement was based on sequential pairwise comparison of the 100 captured time-lapse images. The motility measurement called "AdjustedRandIndex" is calculated by comparing worm objects identified on images captured at times t_n and t_{n-1} with CellProfiler's CalculateImageOverlap module. This measure ranges from 0 to 1, with 1 meaning two objects are perfectly aligned (no

movement). In addition to the "Overlap" mobility score, over 100 features related to size, shape, intensity, texture, and granularity are calculated for worm objects identified in the image analysis pipeline and saved in a database. These features are expected to describe different parasite phenotypes in response to drug exposure.

Cytotoxicity Assay. WSS-1 [WS-1](ATCC®CRL-2029™) epithelial cells derived from human kidney were grown in DMEM medium, supplemented with 4.5 g/L glucose, 50 µg/mL gentamicin and 10% fetal bovine serum, and seeded into 96-well microplates at 5 x 10^4 cells/mL. Twenty hours later, cells were exposed to 0.2 - 400 µM of PZQ, OLT, and LabMol compounds and kept under a humidified atmosphere (37°C, 5% CO_2) for 48h. In order to evaluate the cytotoxic effects of the compounds, the fluorescent viability dye resazurin was added to each well at a final concentration of 0.01 mg/mL 4 hours before the end of the incubation. Resorufin fluorescence readings (λ_{ex} = 560 nm, λ_{em} = 590 nm) were performed immediately and 4 hours after resazurin addition in a FlexStation 3 Benchtop Multi-Mode Microplate Reader (Molecular Devices, Sunnyvale, CA). The percentage of viable cell was calculated using cells treated only with DMSO (0.2 – 0.8%) as controls.

Colloidal aggregation assay. Adult worms were co-incubated with compounds (at 20 µM and 100 µM) and detergent Triton X-100 (0.01%). The, motility measurements were performed after 48h and 72h and their antischistosomal effect was compared with activities obtained without detergent.

Papain inhibition assay. Enzymatic assay was performed at 37ºC in 100 mM sodium acetate buffer, pH 3.5. Positive control E-64 and compounds were incubated at 20 µM and 100 µM concentrations for 5 min with papain (5 µg/mL) and the reaction was initiated with the addition of 50 µM Z-FR-AMC fluorogenic peptide substrate.

Statistical analysis. One-way ANOVA followed by Tukey's multiple comparisons test was performed using GraphPad Prism v.5.00 (GraphPad Software, La Jolla California USA, www.graphpad.com). The EC_{50} and CC_{50} values were determined by four parameter logist curve function using the same software. EC_{50} values obtained for adult worms were calculated using TIBCO Spotfire software (Boston, MA).

Ethics Statement. Animal's maintenance and experiments were carried out in accordance with the Institutional Ethics Committee for Laboratory Animal Use at the Oswaldo Cruz Foundation (CEUA/FIOCRUZ, Brazil; license number, L-044/15).

ASSOCIATED CONTENT

Supporting Information. More computational details regarding molecular fingerprints calculation and QSAR model development are available in the Supporting Information, as well as additional tables and figures of experimental results. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*LabMol, Laboratory for Molecular Modeling and Drug Design, Faculdade de Farmácia, Universidade Federal de Goiás, Rua 240, Qd.87, Setor Leste Universitário, Goiânia – GO 74605-170, Brazil. Tel: + 55 62 3209-6451; Fax: +55 62 3209-6037; E-mail: carolina@ufg.br.

* LaBECFar, Laboratório de Bioquímica Experimental e Computacional de Fármacos, Fundação Oswaldo Cruz, Instituto Oswaldo Cruz, Av. Brasil, 4365, Rio de Janeiro, RJ 21040-900, Brazil. Tel: + 55 21 3865 8248; Fax: +55 21 2590 3495. E-mail: floriano@ioc.fiocruz.br

Author Contributions

BJN, RFD, and MRS contributed equally.

Funding Sources

BJN was supported by a fellowship from the Coordination for the Improvement of Higher Education Personnel (CAPES). This work has been funded by the National Counsel of Technological and Scientific Development (CNPq), the State of Goiás Research Foundation (FAPEG), and State of Rio de Janeiro Research Foundation (FAPERJ). C.H.A. and F.P.S.Jr are CNPq productivity fellows. LEK and AEC were supported by a grant from the U.S. National Institutes of Health (GM095672). E.M. also acknowledge NIH (grants GM66940 and GM096967), CNPq (grant 400760/2014-2), and UNC for Junior Faculty Development Award for partial financial support. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors would like to thank Brazilian funding agencies, CNPq, CAPES, FAPEG, FAPERJ and FIOCRUZ for financial support and fellowships. We are grateful to ChemAxon for providing academic license of their program. Authors also thank the Malacology Laboratory (Dr. Silvana C. Thiengo) from IOC/FIOCRUZ for providing *S. mansoni* cercariae, and the Bioassays and Drug Screening Platform (FIOCRUZ RPT11-I subunit), and NIH R01 GM095672 (Automated image analysis for high-throughput phenotypic screening in *C. elegans*) for technological support.

ABBREVIATIONS

AD, applicability domain; CART, classification and regression trees; CC_{50}, half-maximal cytotoxic concentration; CDK, chemistry development kit; EC_{50}, half-maximal motility concentration; FN, false negatives; FP, false positives; GBM, gradient boosting machine; GR, glutathione reductase; GSH, glutathione; GSSG, glutathione disulfide; HCS, high content screening; IC_{50}, half maximal inhibitory concentration; *k*NN, *k*-nearest neighbors; MACCS, Molecular ACCess System (MACCS) keys; MARS, multivariate adaptive regression splines; MLP, multi-layer perceptron; NADPH, nicotinamide adenine dinucleotide phosphate; OECD, Organization for Economic Cooperation and Development; OLT, oltipraz; PCA, principal component analysis; PLS-DA, partial least squares discriminant analysis; PPMs, predicted probability maps; PZQ, praziquantel; QSAR, quantitative structure-activity relationships; RF, random forest; *S. mansoni*, *Schistosoma mansoni*; SAR, structure-activity relationships; SE, sensitivity; SMARTS, SMILES arbitrary target specification; *Sm*TGR, *S. mansoni* TGR; SP, specificity; SVM, support vector machine; TGR, thioredoxin glutathione reductase; TN, true negatives; TP, true positives; TR, thioredoxin reductase; Trx, thioredoxin; and VS, virtual screening.

REFERENCES

(1) Colley, D. G.; Bustinduy, A. L.; Secor, W. E.; King, C. H. Human Schistosomiasis. *Lancet* **2014**, *383* (9936), 2253–2264.

(2) Ross, A. G. P.; Bartley, P. B.; Sleigh, A. C.; Olds, G. R.; Li, Y.; Williams, G. M.; McManus, D. P. Schistosomiasis. *N. Engl. J. Med.* **2002**, *346* (16), 1212–1220.

(3) King, C. H. Toward the Elimination of Schistosomiasis. *N. Engl. J. Med.* **2009**, *360* (2), 106–109.

(4) Gryseels, B.; Polman, K.; Clerinx, J.; Kestens, L. Human Schistosomiasis. *Lancet* **2006**, *368* (9541), 1106–1118.

(5) World Health Organization. Schistosomiasis. Available from: http://www.who.int/mediacentre/factsheets/fs115/en. (Accessed: June 13, 2016).

(6) Gönnert, R.; Andrews, P. Praziquantel, a New Board-Spectrum Antischistosomal Agent. *Z. Parasitenkd.* **1977**, *52* (2), 129–150.

(7) Ismail, M.; Metwally, A.; Farghaly, A.; Bruce, J.; Tao, L. F.; Bennett, J. L. Characterization of Isolates of Schistosoma Mansoni from Egyptian Villagers That Tolerate High Doses of Praziquantel. *Am. J. Trop. Med. Hyg.* **1996**, *55* (2), 214–218.

(8) Melman, S. D.; Steinauer, M. L.; Cunningham, C.; Kubatko, L. S.; Mwangi, I. N.; Wynn, N. B.; Mutuku, M. W.; Karanja, D. M. S.; Colley, D. G.; Black, C. L.; Secor, W. E.; Mkoji, G. M.; Loker, E. S. Reduced Susceptibility to Praziquantel among Naturally Occurring Kenyan Isolates of Schistosoma Mansoni. *PLoS Negl. Trop. Dis.* **2009**, *3* (8), e504.

(9) Fallon, P. G.; Sturrock, R. F.; Niang, A. C.; Doenhoff, M. J. Short Report: Diminished Susceptibility to Praziquantel in a Senegal Isolate of Schistosoma Mansoni. *Am. J. Trop. Med. Hyg.* **1995**, *53* (1), 61–62.

(10) Hagan, P.; Appleton, C. C.; Coles, G. C.; Kusel, J. R.; Tchuem-Tchuenté, L.-A. Schistosomiasis Control: Keep Taking the Tablets. *Trends Parasitol.* **2004**, *20* (2), 92–97.

(11) Loukas, A.; Bethony, J. M. New Drugs for an Ancient Parasite. *Nat. Med.* **2008**, *14* (4), 365–367.

(12) Wang, W.; Wang, L.; Liang, Y. Susceptibility or Resistance of Praziquantel in Human Schistosomiasis: A Review. *Parasitol. Res.* **2012**, *111* (5), 1871–1877.

(13) Berriman, M.; Haas, B. J.; LoVerde, P. T.; Wilson, R. A.; Dillon, G. P.; Cerqueira, G. C.; Mashiyama, S. T.; Al-Lazikani, B.; Andrade, L. F.; Ashton, P. D.; Aslett, M. a; Bartholomeu, D. C.; Blandin, G.; Caffrey, C. R.; Coghlan, A.; Coulson, R.; Day, T. a; Delcher, A.; DeMarco, R.; Djikeng, A.; Eyre, T.; Gamble, J. a; Ghedin, E.; Gu, Y.; Hertz-Fowler, C.; Hirai, H.; Hirai, Y.; Houston, R.; Ivens, A.; Johnston, D. a; Lacerda, D.; Macedo, C. D.; McVeigh, P.; Ning, Z.; Oliveira, G.; Overington, J. P.; Parkhill, J.; Pertea, M.; Pierce, R. J.; Protasio, A. V; Quail, M. a; Rajandream, M.-A.; Rogers, J.; Sajid, M.; Salzberg, S. L.; Stanke, M.; Tivey, A. R.; White, O.; Williams, D. L.; Wortman, J.; Wu, W.; Zamanian, M.; Zerlotini, A.; Fraser-Liggett, C. M.; Barrell, B. G.; El-Sayed, N. M. The Genome of the Blood Fluke Schistosoma Mansoni. *Nature* **2009**, *460* (7253), 352–358.

(14) Protasio, A. V; Tsai, I. J.; Babbage, A.; Nichol, S.; Hunt, M.; Aslett, M. a; De Silva, N.; Velarde, G. S.; Anderson, T. J. C.; Clark, R. C.; Davidson, C.; Dillon, G. P.; Holroyd, N. E.; LoVerde, P. T.; Lloyd, C.; McQuillan, J.; Oliveira, G.; Otto, T. D.; Parker-Manuel, S. J.; Quail, M. a; Wilson, R. A.; Zerlotini, A.; Dunne, D. W.; Berriman, M. A Systematically Improved High Quality Genome and Transcriptome of the Human Blood Fluke Schistosoma Mansoni. *PLoS Negl. Trop. Dis.* **2012**, *6* (1), e1455.

(15) Zhou, Y.; Zheng, H.; Chen, Y.; Zhang, L.; Wang, K.; Guo, J.; Huang, Z.; Zhang, B.;

Huang, W.; Jin, K.; Dou, T.; Hasegawa, M.; Wang, L.; Zhang, Y.; Zhou, J.; Tao, L.; Cao, Z.; Li, Y.; Vinar, T.; Brejova, B.; Brown, D.; Li, M.; Miller, D. J.; Blair, D.; Zhong, Y.; Chen, Z.; Liu, F.; Hu, W.; Wang, Z.-Q.; Zhang, Q.-H.; Song, H.-D.; Chen, S.; Xu, X.; Xu, B.; Ju, C.; Huang, Y.; Brindley, P. J.; McManus, D. P.; Feng, Z.; Han, Z.-G.; Lu, G.; Ren, S.; Wang, Y.; Gu, W.; Kang, H.; Chen, J.; Chen, X.; Chen, S.; Wang, L.; Yan, J.; Wang, B.; Lv, X.; Jin, L.; Wang, B.; Pu, S.; Zhang, X.; Zhang, W.; Hu, Q.; Zhu, G.; Wang, J.; Yu, J.; Wang, J.; Yang, H.; Ning, Z.; Beriman, M.; Wei, C.-L.; Ruan, Y.; Zhao, G.; Wang, S.; Liu, F.; Zhou, Y.; Wang, Z.-Q.; Lu, G.; Zheng, H.; Brindley, P. J.; McManus, D. P.; Blair, D.; Zhang, Q.; Zhong, Y.; Wang, S.; Han, Z.-G.; Chen, Z.; Wang, S.; Han, Z.-G.; Chen, Z. The Schistosoma Japonicum Genome Reveals Features of Host–parasite Interplay. *Nature* **2009**, *460* (7253), 345–351.

(16) Young, N. D.; Jex, A. R.; Li, B.; Liu, S.; Yang, L.; Xiong, Z.; Li, Y.; Cantacessi, C.; Hall, R. S.; Xu, X.; Chen, F.; Wu, X.; Zerlotini, A.; Oliveira, G.; Hofmann, A.; Zhang, G.; Fang, X.; Kang, Y.; Campbell, B. E.; Loukas, A.; Ranganathan, S.; Rollinson, D.; Rinaldi, G.; Brindley, P. J.; Yang, H.; Wang, J.; Wang, J.; Gasser, R. B. Whole-Genome Sequence of Schistosoma Haematobium. *Nat. Genet.* **2012**, *44* (2), 221–225.

(17) Ferreira, L. G.; Oliva, G.; Andricopulo, A. D. Target-Based Molecular Modeling Strategies for Schistosomiasis Drug Discovery. *Future Med. Chem.* **2015**, *7* (6), 753–764.

(18) Kuntz, A. N.; Davioud-Charvet, E.; Sayed, A. A.; Califf, L. L.; Dessolin, J.; Arnér, E. S. J.; Williams, D. L. Thioredoxin Glutathione Reductase from Schistosoma Mansoni: An Essential Parasite Enzyme and a Key Drug Target. *PLoS Med.* **2007**, *4* (6), e206.

(19) TIMMS, A. R.; BUEDING, E. Studies of a Proteolytic Enzyme from Schistosoma

Mansoni. *Br. J. Pharmacol. Chemother.* **1959**, *14* (1), 68–73.

(20) Hall, S. L.; Braschi, S.; Truscott, M.; Mathieson, W.; Cesari, I. M.; Wilson, R. A. Insights into Blood Feeding by Schistosomes from a Proteomic Analysis of Worm Vomitus. *Mol. Biochem. Parasitol.* **2011**, *179* (1), 18–29.

(21) LoVerde, P. . Do Antioxidants Play a Role in Schistosome Host–Parasite Interactions? *Parasitol. Today* **1998**, *14* (7), 284–289.

(22) Callahan, H. L.; Crouch, R. K.; James, E. R. Helminth Anti-Oxidant Enzymes: A Protective Mechanism against Host Oxidants? *Parasitol. today* **1988**, *4* (8), 218–225.

(23) Alger, H. M.; Williams, D. L. The Disulfide Redox System of Schistosoma Mansoni and the Importance of a Multifunctional Enzyme, Thioredoxin Glutathione Reductase. *Mol. Biochem. Parasitol.* **2002**, *121* (1), 129–139.

(24) Huang, H.; Day, L.; Cass, C. L.; Ballou, D. P.; Williams, C. H.; Williams, D. L. Investigations of the Catalytic Mechanism of Thioredoxin Glutathione Reductase from Schistosoma Mansoni. *Biochemistry* **2011**, *50* (26), 5870–5882.

(25) Bonilla, M.; Denicola, A.; Marino, S. M.; Gladyshev, V. N.; Salinas, G. Linked Thioredoxin-Glutathione Systems in Platyhelminth Parasites: Alternative Pathways for Glutathione Reduction and Deglutathionylation. *J. Biol. Chem.* **2011**, *286* (7), 4959–4967.

(26) Williams, D. L.; Bonilla, M.; Gladyshev, V. N.; Salinas, G. Thioredoxin Glutathione Reductase-Dependent Redox Networks in Platyhelminth Parasites. *Antioxid. Redox Signal.* **2013**, *19* (7), 735–745.

(27) Song, L.; Li, J.; Xie, S.; Qian, C.; Wang, J.; Zhang, W.; Yin, X.; Hua, Z.; Yu, C.

Thioredoxin Glutathione Reductase as a Novel Drug Target: Evidence from Schistosoma Japonicum. *PLoS One* **2012**, *7* (2), e31456.

(28) Han, Y.; Fu, Z.; Hong, Y.; Zhang, M.; Han, H.; Lu, K.; Yang, J.; Li, X.; Lin, J. Inhibitory Effects and Analysis of RNA Interference on Thioredoxin Glutathione Reductase Expression in Schistosoma Japonicum. *J. Parasitol.* **2014**, *100* (4), 463–469.

(29) Zhang, L.; Fourches, D.; Sedykh, A.; Zhu, H.; Golbraikh, A.; Ekins, S.; Clark, J.; Connelly, M. C.; Sigal, M.; Hodges, D.; Guiguemde, A.; Guy, R. K.; Tropsha, A. Discovery of Novel Antimalarial Compounds Enabled by QSAR-Based Virtual Screening. *J. Chem. Inf. Model.* **2013**, *53* (2), 475–492.

(30) Neves, B. J.; Bueno, R. V; Braga, R. C.; Andrade, C. H. Discovery of New Potential Hits of Plasmodium Falciparum Enoyl-ACP Reductase through Ligand- and Structure-Based Drug Design Approaches. *Bioorg. Med. Chem. Lett.* **2013**, *23* (8), 2436–2441.

(31) Bueno, R. V; Toledo, N. R.; Neves, B. J.; Braga, R. C.; Andrade, C. H. Structural and Chemical Basis for Enhanced Affinity to a Series of Mycobacterial Thymidine Monophosphate Kinase Inhibitors: Fragment-Based QSAR and QM/MM Docking Studies. *J. Mol. Model.* **2013**, *19* (1), 179–192.

(32) Braga, R. C.; Alves, V. M.; Silva, A. C.; Nascimento, M. N.; Silva, F. C.; Liao, L. M.; Andrade, C. H. Virtual Screening Strategies in Medicinal Chemistry: The State of the Art and Current Challenges. *Curr. Top. Med. Chem.* **2014**, *14* (16), 1899–1912.

(33) Melo-Filho, C. C.; Braga, R. C.; Andrade, C. H. 3D-QSAR Approaches in Drug Design: Perspectives to Generate Reliable CoMFA Models. *Curr. Comput. Aided. Drug Des.*

2014, *10* (2), 148–159.

(34) Artemenko, A. G.; Muratov, E. N.; Kuz'min, V. E.; Kovdienko, N. A.; Hromov, A. I.; Makarov, V. A.; Riabova, O. B.; Wutzler, P.; Schmidtke, M. Identification of Individual Structural Fragments of N,N'-(bis-5-Nitropyrimidyl)dispirotripiperazine Derivatives for Cytotoxicity and Antiherpetic Activity Allows the Prediction of New Highly Active Compounds. *J. Antimicrob. Chemother.* **2007**, *60* (1), 68–77.

(35) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N.; Volineckaya, I. L.; Makarov, V. A.; Riabova, O. B.; Wutzler, P.; Schmidtke, M. Quantitative Structure-Activity Relationship Studies of [(Biphenyloxy)propyl]isoxazole Derivatives. Inhibitors of Human Rhinovirus 2 Replication. *J. Med. Chem.* **2007**, *50* (17), 4205–4213.

(36) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem. *J. Chem. Inf. Model.* **2014**, *54* (3), 705–712.

(37) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25* (4), 64–73.

(38) Riniker, S.; Landrum, G. a. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminform.* **2013**, *5* (1), 26.

(39) 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.* **2002**, *42* (6), 1273–1280.

(40) Dill, J. D. .; Hounshell, W. D. .; Marson, S. .; Peacock, S. .; Wipke, W. T. Search and

Retrieval Using an Automated Molecular Access System. In *182nd National Meeting of the American Chemical Society*; New York, 1981; pp 23–28.

(41) Anderson, S. Graphical Representation of Molecules and Substructure-Search Queries in MACCStm. *Journal of Molecular Graphics*. 1984, pp 83–90.

(42) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

(43) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466–1474.

(44) PaDEL-Descriptor. Http://padel.nus.edu.sg/software/padeldescriptor/ (Accessed January 5, 2015).

(45) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, 1st ed.; Todeschini, R., Consonni, V., Eds.; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2000.

(46) Braga, R. C.; Alves, V. M.; Silva, M. F. B.; Muratov, E.; Fourches, D.; Tropsha, A.; Andrade, C. H. Tuning HERG out: Antitarget QSAR Models for Drug Development. *Curr. Top. Med. Chem.* **2014**, *14* (11), 1399–1415.

(47) Braga, R. C.; Alves, V. M.; Silva, M. F. B.; Muratov, E.; Fourches, D.; Lião, L. M.; Tropsha, A.; Andrade, C. H. Pred-hERG: A Novel Web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Mol. Inform.* **2015**, *34* (10), 698–701.

(48) Alves, V. M.; Muratov, E.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; Tropsha, A. Predicting Chemically-Induced Skin Reactions. Part I: QSAR Models of Skin

Sensitization and Their Application to Identify Potentially Hazardous Compounds. *Toxicol. Appl. Pharmacol.* **2015**, *284* (2), 262–272.

(49) Alves, V. M.; Muratov, E.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; Tropsha, A. Predicting Chemically-Induced Skin Reactions. Part II: QSAR Models of Skin Permeability and the Relationships between Skin Permeability and Skin Sensitization. *Toxicol Appl Pharmacol* **2015**, *284* (2), 273–280.

(50) Riniker, S.; Landrum, G. a. Similarity Maps - A Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J. Cheminform.* **2013**, *5* (9), 1–7.

(51) Bonilla, M.; Denicola, A.; Novoselov, S. V; Turanov, A. A.; Protasio, A.; Izmendi, D.; Gladyshev, V. N.; Salinas, G. Platyhelminth Mitochondrial and Cytosolic Redox Homeostasis Is Controlled by a Single Thioredoxin Glutathione Reductase and Dependent on Selenium and Glutathione. *J. Biol. Chem.* **2008**, *283* (26), 17898–17907.

(52) Nauser, T.; Steinmann, D.; Koppenol, W. H. Why Do Proteins Use Selenocysteine instead of Cysteine? *Amino Acids* **2012**, *42* (1), 39–44.

(53) Johansson, L.; Gafvelin, G.; Arnér, E. S. J. Selenocysteine in Proteins-Properties and Biotechnological Use. *Biochim. Biophys. Acta* **2005**, *1726* (1), 1–13.

(54) Rai, G.; Sayed, A. A.; Lea, W. A.; Luecke, H. F.; Chakrapani, H.; Prast-Nielsen, S.; Jadhav, A.; Leister, W.; Shen, M.; Inglese, J.; Austin, C. P.; Keefer, L.; Arnér, E. S. J.; Simeonov, A.; Maloney, D. J.; Williams, D. L.; Thomas, C. J. Structure Mechanism Insights and the Role of Nitric Oxide Donation Guide the Development of Oxadiazole-2-Oxides as Therapeutic Agents against Schistosomiasis. *J. Med. Chem.* **2009**, *52* (20),

6474–6483.

(55) Gasco, A.; Fruttero, R.; Sorba, G.; Di Stilo, A.; Calvino, R. NO Donors: Focus on Furoxan Derivatives. *Pure Appl. Chem.* **2004**, *76* (5), 973–981.

(56) Triboulet, S.; Dubée, V.; Lecoq, L.; Bougault, C.; Mainardi, J.-L.; Rice, L. B.; Ethève-Quelquejeu, M.; Gutmann, L.; Marie, A.; Dubost, L.; Hugonnet, J.-E.; Simorre, J.-P.; Arthur, M. Kinetic Features of L,D-Transpeptidase Inactivation Critical for β-Lactam Antibacterial Activity. *PLoS One* **2013**, *8* (7), e67831.

(57) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623.

(58) 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.* **1997**, *23* (1-3), 3–25.

(59) Baell, J.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513* (7519), 481–483.

(60) Baell, J. B.; Holloway, G. a. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53* (7), 2719–2740.

(61) QikProp version 3.4. Schrödinger Inc., New York, USA, http://www.schrodinger.com/, 2011.

(62) Zanella, F.; Lorens, J. B.; Link, W. High Content Screening: Seeing Is Believing. *Trends*

Biotechnol. **2010**, *28* (5), 237–245.

(63) Neves, B. J.; Muratov, E.; Machado, R. B.; Andrade, C. H.; Cravo, P. V. L. Modern Approaches to Accelerate Discovery of New Antischistosomal Drugs. *Expert Opin. Drug Discov.* **2016**, *11* (6), 557–567.

(64) Paveley, R.; Mansour, N. R.; Hallyburton, I.; Bleicher, L. S.; Benn, A. E.; Mikic, I.; Guidi, A.; Gilbert, I. H.; Hopkins, A. L.; Bickle, Q. D. Whole Organism High-Content Screening by Label-Free, Image-Based Bayesian Classification for Parasitic Diseases. *PLoS Negl. Trop. Dis.* **2012**, *6* (7), e1762.

(65) Melo-Filho, C. C.; Dantas, R. F.; Braga, R. C.; Neves, B. J.; Senger, M. R.; Valente, W. C. G.; Rezende-Neto, J. M.; Chaves, W. T.; Muratov, E. N.; Paveley, R. A.; Furnham, N.; Kamentsky, L.; Carpenter, A. E.; Junior, F. P. S.; Andrade, C. H. QSAR-Driven Discovery of Novel Chemical Scaffolds Active Against Schistosoma Mansoni. *J. Chem. Inf. Model.* [Online early access]. DOI: 10.1021/acs.jcim.6b00055. Published Online: June 2, 2016.

(66) Neves, B. J.; Dantas, R. F.; Senger, M. R.; Valente, W. C. G.; Rezende-Neto, J. de M.; Chaves, W. T.; Kamentsky, L.; Carpenter, A.; Silva-Junior, F. P.; Andrade, C. H. The Antidepressant Drug Paroxetine as a New Lead Candidate in Schistosome Drug Discovery. *Med. Chem. Commun.* [Online early access]. DOI: 10.1039/C5MD00596E. Published Online: April 25, 2016.

(67) Panic, G.; Flores, D.; Ingram-Sieber, K.; Keiser, J. Fluorescence/luminescence-Based Markers for the Assessment of Schistosoma Mansoni Schistosomula Drug Assays. *Parasit. Vectors* **2015**, *8* (8), 624.

(68) Cowan, N.; Keiser, J. Repurposing of Anticancer Drugs: In Vitro and in Vivo Activities against Schistosoma Mansoni. *Parasit. Vectors* **2015**, *13* (8), 417.

(69) Panic, G.; Vargas, M.; Scandale, I.; Keiser, J. Activity Profile of an FDA-Approved Compound Library against Schistosoma Mansoni. *PLoS Negl. Trop. Dis.* **2015**, *9* (7), e0003962.

(70) Ingram-Sieber, K.; Cowan, N.; Panic, G.; Vargas, M.; Mansour, N. R.; Bickle, Q. D.; Wells, T. N. C.; Spangenberg, T.; Keiser, J. Orally Active Antischistosomal Early Leads Identified from the Open Access Malaria Box. *PLoS Negl. Trop. Dis.* **2014**, *8* (1), e2610.

(71) Paveley, R. A.; Bickle, Q. D. Automated Imaging and Other Developments in Whole-Organism Anthelmintic Screening. *Parasite Immunol.* **2013**, *35* (9-10), 302–313.

(72) Abdulla, M.-H.; Ruelas, D. S.; Wolff, B.; Snedecor, J.; Lim, K.-C.; Xu, F.; Renslo, A. R.; Williams, J.; McKerrow, J. H.; Caffrey, C. R. Drug Discovery for Schistosomiasis: Hit and Lead Compounds Identified in a Library of Known Drugs by Medium-Throughput Phenotypic Screening. *PLoS Negl. Trop. Dis.* **2009**, *3* (7), e478.

(73) Mansour, N. R.; Paveley, R.; Gardner, J. M. F.; Bell, A. S.; Parkinson, T.; Bickle, Q. High Throughput Screening Identifies Novel Lead Compounds with Activity against Larval, Juvenile and Adult Schistosoma Mansoni. *PLoS Negl. Trop. Dis.* **2016**, *10* (4), e0004659.

(74) Long, T.; Neitz, R. J.; Beasley, R.; Kalyanaraman, C.; Suzuki, B. M.; Jacobson, M. P.; Dissous, C.; McKerrow, J. H.; Drewry, D. H.; Zuercher, W. J.; Singh, R.; Caffrey, C. R. Structure-Bioactivity Relationship for Benzimidazole Thiophene Inhibitors of Polo-Like Kinase 1 (PLK1), a Potential Drug Target in Schistosoma Mansoni. *PLoS Negl. Trop. Dis.*

2016, *10* (1), e0004356.

(75) Mkoji, G. M.; Smith, J. M.; Prichard, R. K. Antioxidant Systems in Schistosoma Mansoni: Evidence for Their Role in Protection of the Adult Worms against Oxidant Killing. *Int. J. Parasitol.* **1988**, *18* (5), 667–673.

(76) Nare, B.; Smith, J. M.; Prichard, R. K. Schistosoma Mansoni: Levels of Antioxidants and Resistance to Oxidants Increase during Development. *Exp. Parasitol.* **1990**, *70* (4), 389–397.

(77) Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58* (17), 7076–7087.

(78) Owen, S. C.; Doak, A. K.; Wassam, P.; Shoichet, M. S.; Shoichet, B. K. Colloidal Aggregation Affects the Efficacy of Anticancer Drugs in Cell Culture. *ACS Chem. Biol.* **2012**, *7* (8), 1429–1435.

(79) Kuntz, A. N.; Davioud-Charvet, E.; Sayed, A. A.; Califf, L. L.; Dessolin, J.; Arnér, E. S. J.; Williams, D. L. Thioredoxin Glutathione Reductase from Schistosoma Mansoni: An Essential Parasite Enzyme and a Key Drug Target. *PLoS Med.* **2007**, *4* (6), e206.

(80) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29* (6-7), 476–488.

(81) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.;

Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010.

(82) Organisation for Economic Co-operation and Development. OECD principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationship models. Available from: http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf. (Accessed June 13, 2016).

(83) Neves, B. J.; Braga, R. C.; Bezerra, J. C. B.; Cravo, P. V. L.; Andrade, C. H. In Silico Repositioning-Chemogenomics Strategy Identifies New Drugs with Potential Activity against Multiple Life Stages of Schistosoma Mansoni. *PLoS Negl. Trop. Dis.* **2015**, *9* (1), e3435.

(84) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204.

(85) Fourches, D.; Muratov, E.; Tropsha, A. Curation of Chemogenomics Data. *Nat. Chem. Biol.* **2015**, *11* (8), 535.

(86) Landrum, G. RDKit: Open-Source Cheminformatics. 2014, p http://www.rdkit.org/.

(87) Altman, N. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46* (3), 175–185.

(88) R Development Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(89) Vapnik, V. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.

(90) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.

(91) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29* (5), 1189–1232.

(92) Barker, M.; Rayens, W. Partial Least Squares for Discrimination. *J. Chemom.* **2003**, *17*, 166–173.

(93) Breiman, L.; Friedman, J.; Olshen, R. A.; Charles, J. S. *Classification and Regression Trees*, 1nd ed.; Breiman, L., Ed.; Wadsworth & Brooks/Cole Advanced Books & Software: Monterey, 1984.

(94) Rosenblatt, F. *Principles of Neurodynamics; Perceptrons and the Theory of Brain Mechanisms*, 1nd ed.; Rosenblatt, F., Ed.; Spartan Books: Washington, 1962.

(95) Friedman, J. H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19* (1), 1–67.

(96) Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. *J. Chem. Inf. Model.* **2006**, *46* (5), 1984–1995.

(97) Cohen, J. A Coefficient of Agreement of Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46.

(98) Saubern, S.; Guha, R.; Baell, J. B. KNIME Workflow to Assess PAINS Filters in

SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Mol. Inform.* **2011**, *30* (10), 847–850.

(99) Alves, V. M.; Braga, R. C.; Silva, M. B.; Muratov, E.; Fourches, D.; Tropsha, A.; Andrade, C. H. Pred-hERG: A Novel Web-Accessible Computational Tool for Predicting Cardiac Toxicity of Drug Candidates. In *Abstracts of Papers, 248th ACS National Meeting & Exposition, San Francisco, CA, United States, August 10-14, 2014*; American Chemical Society, 2014; p CINF – 40.

(100) Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P. W.; Tang, Y. admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *J. Chem. Inf. Model.* **2012**, *52* (11), 3099–3105.

(101) Lagunin, A.; Filimonov, D.; Zakharov, A.; Xie, W.; Huang, Y.; Zhu, F.; Shen, T.; Yao, J.; Poroikov, V. Computer-Aided Prediction of Rodent Carcinogenicity by PASS and CISOC-PSCT. *QSAR Comb. Sci.* **2009**, *28* (8), 806–810.

(102) He, Y.; Liew, C. Y.; Sharma, N.; Woo, S. K.; Chau, Y. T.; Yap, C. W. PaDEL-DDPredictor: Open-Source Software for PD-PK-T Prediction. *J. Comput. Chem.* **2013**, *34* (7), 604–610.

(103) Liew, C. Y.; Lim, Y. C.; Yap, C. W. Mixed Learning Algorithms and Features Ensemble in Hepatotoxicity Prediction. *J. Comput. Aided. Mol. Des.* **2011**, *25* (9), 855–871.

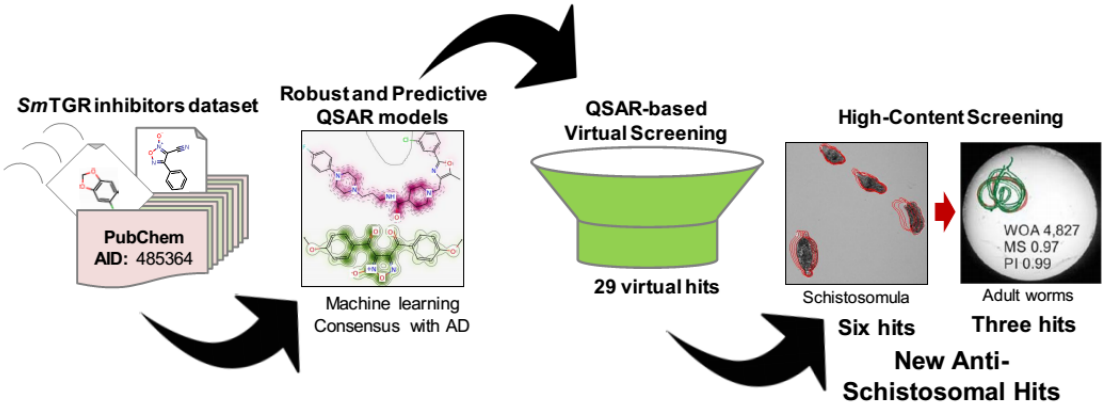
(104) Marxer, M.; Ingram, K.; Keiser, J. Development of an in Vitro Drug Screening Assay Using Schistosoma Haematobium Schistosomula. *Parasit. Vectors* **2012**, *5* (1), 165.

(105) Mansour, N. R.; Bickle, Q. D. Comparison of Microscopy and Alamar Blue Reduction in

a Larval Based Assay for Schistosome Drug Screening. *PLoS Negl. Trop. Dis.* **2010**, *4* (8), e795.

(106) Kamentsky, L.; Jones, T. R.; Fraser, A.; Bray, M.-A.; Logan, D. J.; Madden, K. L.; Ljosa, V.; Rueden, C.; Eliceiri, K. W.; Carpenter, A. E. Improved Structure, Function and Compatibility for CellProfiler: Modular High-Throughput Image Analysis Software. *Bioinformatics* **2011**, *27* (8), 1179–1180.

Table of Contents Graphic (TOC)


*Sm*TGR inhibitors dataset
PubChem
AID: 485364
Robust and Predictive
QSAR models
Machine learning
Consensus with AD
QSAR-based
Virtual Screening
29 virtual hits
High-Content Screening
Schistosomula
WOA 4,827
MS 0.97
PI 0.99
Adult worms
Six hits
Three hits
New Anti-
Schistosomal Hits

List of Figures

Figure 1. Favorable fragments for *Sm*TGR inhibition predicted by the best individual QSAR model and their respective frequencies in inhibitors and non-inhibitors sets.

Figure 2. Unfavorable fragments for *Sm*TGR inhibition predicted by the best individual QSAR model and their respective frequencies in inhibitors and non-inhibitors sets.

Figure 3. Predicted probability maps generated for oxadiazoles (A, B, and C) and their probable mechanism of action in the *Sm*TGR active site (D).

Figure 4. QSAR-based VS workflow used for identifying new hit compounds against *S. mansoni*.

Figure 5. Chemical structures of six priority hits selected for further follow up.

Figure 6. Phenotypes of schistosomula exposed for 48h to 0.625 % DMSO (control, A), 20 μM of **4** (B), and 10 μM of PZQ (C) and OLT (D). The outlines represent the position of each parasite over 5 time points (11s interval).

Figure 7. 3D scatter plot of the top three image features correlated to antischistosomal activity of the hits on female *S. mansoni* worms after 48 drug exposure. Each point in the graph represent a well/condition in the assay. Sample images are shown for selected wells to illustrate the different phenotypes captured by these three parameters (Object Area; Mobility Score; and Pixel Intensity of the worm object). The green outlines represent the position of each parasite over 5 time points (3s interval) overlaid on the initial position (red outline).