

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 Kametsky,^ψ Anne E. Carpenter,^ψ Rodolpho C. Braga,[†] Floriano P. Silva-Junior,^{‡} Carolina H. 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

^ψImaging Platform, Broad Institute of Massachusetts Institute of Technology and Harvard,
Cambridge, Massachusetts, USA.

Table Contents

Molecular Fingerprints and Machine Learning Details.....S3

Table S1.....S8

Table S2.....S9

Table S3.....S10

Table S4.....S11

Table S5.....S11

Table S6.....S12

Table S7.....S16

Figure S1.....S17

Figure S2.....S18

Figure S3.....S18

Figure S4.....S19

Figure S5.....S19

Figure S6.....S20

Figure S7.....S21

Nuclear Magnetic Resonance (NMR) and Purity Data.....S22

REFERENCES.....S37

Molecular Fingerprint Details. In this study, three 2D fingerprint types were used: dictionary-based fingerprints, circular fingerprints, and path-based fingerprints. The public MACCS structural keys are a collection of 166 predefined substructures associated with the SMILES arbitrary target specification (SMARTS) pattern and belonging to the dictionary-based fingerprint class.¹ Morgan is a type of circular fingerprint that encodes circular atom environments up to determined radius from a central atom using Morgan algorithm and feature invariants. Atom features, such as chirality, atom, and bond types were used for generating Morgan fingerprints.²⁻⁴ AtomPair fingerprints, implemented by Carhart and colleagues, are defined as a pair of atoms (AT) or descriptor centers separated by a fixed topological distance: $AT_i-AT_j-Dist_{ij}$, where $Dist_{ij}$ is the shortest path (the number of bonds) between AT_i and AT_j . In addition, AtomPair fingerprints account for the information about element type, the number of bonded non-hydrogen neighbors and the number of π electrons.^{5,6}

Machine Learning Details. *Classification and Regression Trees (CART).* The CART algorithm, introduced by Breiman and colleagues,⁷ is a non-parametric decision tree learning method that produces either classification or regression trees, depending on whether the dependent variable is binary or continuous, respectively. In this method, tree is built by dividing the root node, containing all samples, in two child nodes based on a split value for one descriptor/fingerprint of the matrix of dependent variables. Each split of descriptors within a tree is created based on the best partitioning that is possible. Then, child nodes become parents and each new parent node can give rise to two child nodes, etc. Nodes that are not split anymore (e.g. because they are homogeneous) are called terminal nodes or leaves. In general, the building of a CART model contains two steps. First, a big tree is built. Each node is split until pure terminal nodes are found. The tree obtained has a large number of terminal nodes and describes the

Supporting Information

training set almost perfectly, but provides a poor predictive ability for new samples. To solve this problem, a last step, known as pruning, is performed. The pruning step consists of cutting away branches of the big tree to find smaller trees with improved predictive ability.

Random Forest (RF). The RF⁸ algorithm is a tree bagging method that creates a large collection of decorrelated decision trees, and the final prediction is defined by majority voting from an ensemble of decision trees. In each tree, 1/3 of the training set is randomly extracted, while the remaining 2/3 of the training set is used for model building. Then, each tree in the forest is built by CART method,⁷ and best split generated, among the randomly investigated descriptor in each node, is chosen. Each tree is grown to the largest possible extent without pruning. Last, the trained forest is then used to predict test set. The predicted classification values are defined by majority voting for one of the classes. The proportion of votes cast for a class may provide an indication of the probability of a label being correctly assigned, or of confidence in a prediction, but this should be considered an informal estimate only.

Gradient Boosting Machine (GBM). The GBM method⁹ differs from bagging methods through the base learners, here classification or regression trees, are trained and combined sequentially. The principle idea behind this algorithm is generate models by computing a sequence of trees, in which each successive tree is built from the prediction residuals of the preceding tree. A simple (best) partitioning of the data is determined at each step in the boosting tree algorithm, and the deviations of the observed values from the respective residuals for each partition are computed. Given the preceding sequence of trees, the next tree will then be fitted to the residuals in order to find another partition that will further reduce the residual (error) variance for the data.

Supporting Information

Multivariate Adaptive Regression Splines (MARS). MARS¹⁰ is a multivariate nonparametric regression technique that can be extended to handle classification problems. This operating method is based on divide-and-conquer strategy partitioning the training data sets into separate regions, each of which gets its own classification. This makes MARS particularly suitable for problems with high input dimensions.

Support Vector Machine (SVM). SVM¹¹ is a kernel based approach first developed by Vapnik as a general data modeling methodology, aiming at minimizing structural risk and statistical learning theory. Briefly, SVM maps the data into a high-dimensional hyper plane (e.g. descriptors or fingerprints), using a kernel function that is typically linear, radial, or polynomial. The SVM seeks to find an optimal separation between two classes (e.g. inhibitors and non-inhibitors), such that each in their entirety lie on opposite sides of a separating hyper plane. Thus, SVM minimizes the empirical classification error and maximizes the geometric margin. This margin is defined as the distance from the separating hyper plane to its nearest sample. The hyper plane that defines such margin is called support hyper planes, and the data points that lie on these hyper planes are called support vectors. Thus, SVM is also known as a maximum margin classifier.

Partial Least Squares – Discriminant Analysis (PLS-DA). PLS-DA¹² is a linear and parametric method based on the PLS model in which the dependent variable is chosen to represent the class membership (e.g. inhibitors or non-inhibitors). First a classical PLS model is built with a training set. In PLS, the number of variables is reduced using PCA by creating new latent variables which maximize the covariance between the original variables and the response. Using the optimal number of latent variables, to build a linear regression model should provide the best predictive model. Opposite to classical PLS, where the response is quantitative and continuous, the

responses in PLS-DA are qualitative, discrete and coded in a vector with a class member. For an unknown sample, the predicted value obtained with the PLS-DA model is normally distributed around 0 or 1. To determine the limit from which a sample is considered to be in the inhibitors or non-inhibitors class, a threshold 0.5 is determined. When a value above the threshold is predicted, a sample is considered to belong to the inhibitors class, while a value below the threshold indicates that the sample belong to the non-inhibitors class.

Multi-Layer Perceptron (MLP). The MLP method¹³ is a network of simple neurons called perceptrons. The basic concept of a single perceptron was introduced by Rosenblatt in 1958. The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then possibly putting the output through some nonlinear activation function. A typical MLP network consists of a set of source nodes forming the input layer (e.g. descriptors or fingerprints), one or more hidden layers of computation nodes, and an output layer related directly to the activity being predicted. The hidden layers and their number can generally vary depending on the problem at hand. To each of the hidden and output neurons, one virtual neuron can be assigned, called bias used as reference to activate or deactivate a neuron. In addition, MLPs are considered feed-forward neural networks since input signal propagates in only one direction, from the input to output layers. Finally, to train a network to predict values for given arguments, an iterative process that has information fed back from the output neurons to neurons in some layer before was performed, to enable further processing and adjustment of weights on the connections. In this study, training of MLP was performed with the back-propagation algorithm.

k-Nearest Neighbors (kNN). The k NN¹⁴ is a non-parametric method that classifies samples based on a similarity measure. In this method, a sample could be classified by a majority vote of

Supporting Information

its neighbors, with the sample investigated being assigned to the class most common amongst its k nearest neighbors measured by a distance function. This distance is usually taken to be the Euclidean distance, though other metrics such as the Jaccard distance could be used. If $k = 3$, then the case is simply assigned to the class of its three nearest neighbors in a feature space. The samples, which in chemical applications are typically compounds, are described as position vectors in the feature space, which is usually of high dimensionality. It is helpful to scale the features so that distances measured in different directions in the space are comparable.

Table S1. Statistical characteristics of QSAR models developed with balanced dataset.

Model	CCR	k	SE	SP	Coverage
Morgan-GBM	0.84	0.68	0.85	0.84	0.62
Morgan-SVM	0.84	0.69	0.83	0.85	0.62
Morgan-RF	0.85	0.71	0.85	0.86	0.62
Morgan-PLS-DA	0.81	0.62	0.81	0.81	0.62
Morgan- k NN	0.78	0.55	0.90	0.66	0.62
Morgan-CART	0.73	0.47	0.73	0.73	0.62
Morgan-MARS	0.76	0.54	0.77	0.77	0.62
Morgan-MLP	0.82	0.65	0.84	0.81	0.62
CDK-GBM	0.84	0.67	0.84	0.83	0.77
CDK-SVM	0.84	0.69	0.85	0.84	0.77
CDK-RF	0.82	0.64	0.85	0.80	0.77
CDK-PLS-DA	0.77	0.54	0.76	0.78	0.77
CDK- k NN	0.78	0.57	0.78	0.78	0.77
CDK-CART	0.70	0.40	0.71	0.69	0.77
CDK-MARS	0.77	0.53	0.76	0.77	0.77
CDK-MLP	0.78	0.56	0.79	0.77	0.77
Dragon-GBM	0.85	0.70	0.85	0.84	0.69
Dragon-SVM	0.85	0.70	0.85	0.84	0.69
Dragon-RF	0.84	0.69	0.85	0.83	0.69
Dragon-PLS-DA	0.80	0.61	0.80	0.81	0.69
Dragon- k NN	0.80	0.59	0.83	0.76	0.69
Dragon-CART	0.76	0.53	0.79	0.74	0.69
Dragon-MARS	0.80	0.60	0.81	0.80	0.69
Dragon-MLP	0.80	0.61	0.81	0.79	0.69
MACCS-GBM	0.83	0.65	0.83	0.82	0.67
MACCS-SVM	0.83	0.65	0.82	0.83	0.67
MACCS-RF	0.83	0.66	0.83	0.83	0.67
MACCS-PLS-DA	0.76	0.52	0.76	0.76	0.67
MACCS- k NN	0.76	0.53	0.81	0.72	0.67
MACCS-CART	0.73	0.46	0.74	0.72	0.67
MACCS-MARS	0.74	0.48	0.73	0.75	0.67
MACCS-MLP	0.78	0.57	0.79	0.78	0.67
AtomPair-GBM	0.81	0.62	0.81	0.81	0.65
AtomPair-SVM	0.81	0.62	0.81	0.81	0.65
AtomPair-RF	0.80	0.61	0.79	0.82	0.65
AtomPair-PLS-DA	0.74	0.47	0.74	0.73	0.65
AtomPair- k NN	0.74	0.47	0.74	0.73	0.65
AtomPair-CART	0.69	0.37	0.72	0.65	0.65
AtomPair-MARS	0.70	0.41	0.70	0.70	0.65
AtomPair-MLP	0.76	0.52	0.77	0.75	0.65

CCR: correct classification rate; k : Cohen's kappa; SE: sensitivity; SP: specificity

Table S2. Statistical characteristics of QSAR models developed with unbalanced dataset (1:2).

Model	CCR	k	SE	SP	Coverage
Morgan-GBM	0.85	0.72	0.78	0.93	0.67
Morgan-SVM	0.86	0.73	0.77	0.94	0.67
Morgan-RF	0.86	0.74	0.76	0.95	0.67
Morgan-PLS-DA	0.81	0.64	0.71	0.91	0.67
Morgan- k NN	0.83	0.67	0.76	0.90	0.67
Morgan-CART	0.67	0.38	0.44	0.90	0.67
Morgan-MARS	0.75	0.53	0.62	0.88	0.67
Morgan-MLP	0.84	0.68	0.77	0.90	0.67
CDK-GBM	0.82	0.66	0.71	0.93	0.79
CDK-SVM	0.84	0.69	0.75	0.92	0.79
CDK-RF	0.82	0.66	0.70	0.94	0.79
CDK-PLS-DA	0.73	0.49	0.57	0.89	0.79
CDK- k NN	0.79	0.60	0.69	0.89	0.79
CDK-CART	0.70	0.41	0.52	0.87	0.79
CDK-MARS	0.76	0.53	0.63	0.88	0.79
CDK-MLP	0.78	0.56	0.70	0.86	0.79
Dragon-GBM	0.84	0.70	0.76	0.93	0.70
Dragon-SVM	0.85	0.72	0.78	0.93	0.70
Dragon-RF	0.84	0.74	0.74	0.94	0.70
Dragon-PLS-DA	0.78	0.57	0.65	0.90	0.70
Dragon- k NN	0.81	0.63	0.72	0.90	0.70
Dragon-CART	0.74	0.51	0.60	0.89	0.70
Dragon-MARS	0.78	0.58	0.68	0.89	0.70
Dragon-MLP	0.80	0.61	0.73	0.88	0.70
MACCS-GBM	0.82	0.66	0.74	0.91	0.64
MACCS-SVM	0.82	0.66	0.72	0.92	0.64
MACCS-RF	0.82	0.66	0.72	0.92	0.64
MACCS-PLS-DA	0.72	0.46	0.55	0.88	0.64
MACCS- k NN	0.77	0.54	0.67	0.86	0.64
MACCS-CART	0.70	0.42	0.53	0.87	0.64
MACCS-MARS	0.70	0.43	0.52	0.88	0.64
MACCS-MLP	0.79	0.58	0.72	0.86	0.64
AtomPair-GBM	0.81	0.64	0.70	0.92	0.66
AtomPair-SVM	0.81	0.65	0.71	0.92	0.66
AtomPair-RF	0.79	0.62	0.64	0.94	0.66
AtomPair-PLS-DA	0.72	0.47	0.58	0.87	0.66
AtomPair- k NN	0.78	0.55	0.71	0.85	0.66
AtomPair-CART	0.67	0.37	0.48	0.87	0.66
AtomPair-MARS	0.65	0.39	0.50	0.86	0.66
AtomPair-MLP	0.78	0.56	0.70	0.86	0.66

CCR: correct classification rate; k : Cohen's kappa; SE: sensitivity; SP: specificity

Table S3. Statistical characteristics of QSAR models developed with unbalanced dataset (1:3).

Model	CCR	k	SE	SP	Coverage
Morgan-GBM	0.85	0.73	0.74	0.95	0.65
Morgan-SVM	0.85	0.73	0.74	0.96	0.65
Morgan-RF	0.84	0.72	0.71	0.97	0.65
Morgan-PLS-DA	0.79	0.61	0.64	0.94	0.65
Morgan- k NN	0.82	0.64	0.72	0.91	0.65
Morgan-CART	0.72	0.49	0.51	0.93	0.65
Morgan-MARS	0.73	0.50	0.53	0.93	0.65
Morgan-MLP	0.83	0.66	0.73	0.92	0.65
CDK-GBM	0.80	0.65	0.65	0.95	0.80
CDK-SVM	0.83	0.69	0.71	0.95	0.80
CDK-RF	0.78	0.64	0.60	0.97	0.80
CDK-PLS-DA	0.69	0.44	0.45	0.94	0.80
CDK- k NN	0.78	0.59	0.62	0.93	0.80
CDK-CART	0.68	0.41	0.44	0.92	0.80
CDK-MARS	0.72	0.49	0.52	0.92	0.80
CDK-MLP	0.77	0.55	0.65	0.90	0.80
Dragon-GBM	0.83	0.71	0.71	0.96	0.71
Dragon-SVM	0.85	0.72	0.74	0.95	0.71
Dragon-RF	0.82	0.70	0.68	0.97	0.71
Dragon-PLS-DA	0.73	0.51	0.51	0.94	0.71
Dragon- k NN	0.79	0.60	0.65	0.93	0.71
Dragon-CART	0.69	0.43	0.43	0.94	0.71
Dragon-MARS	0.75	0.55	0.58	0.93	0.71
Dragon-MLP	0.80	0.61	0.69	0.91	0.71
MACCS-GBM	0.81	0.66	0.68	0.95	0.69
MACCS-SVM	0.81	0.66	0.68	0.95	0.69
MACCS-RF	0.81	0.67	0.67	0.96	0.69
MACCS-PLS-DA	0.68	0.40	0.42	0.93	0.69
MACCS- k NN	0.76	0.55	0.60	0.92	0.69
MACCS-CART	0.68	0.41	0.43	0.93	0.69
MACCS-MARS	0.66	0.37	0.40	0.93	0.69
MACCS-MLP	0.78	0.58	0.66	0.91	0.69
AtomPair-GBM	0.78	0.62	0.61	0.96	0.67
AtomPair-SVM	0.75	0.58	0.53	0.97	0.67
AtomPair-RF	0.75	0.58	0.53	0.97	0.67
AtomPair-PLS-DA	0.69	0.43	0.45	0.93	0.67
AtomPair- k NN	0.77	0.57	0.61	0.92	0.67
AtomPair-CART	0.64	0.34	0.36	0.93	0.67
AtomPair-MARS	0.64	0.34	0.36	0.93	0.67
AtomPair-MLP	0.75	0.53	0.60	0.90	0.67

CCR: correct classification rate; k : Cohen's kappa; SE: sensitivity; SP: specificity

Table S4. Statistical characteristics of Y-randomization models developed with balanced dataset.

Model	CCR (SD)	k (SD)	SE (SD)	SP (SD)
Morgan-RF	0.51 (0.01)	0.01 (0.02)	0.50 (0.01)	0.51 (0.01)
MACCS-RF	0.51 (0.01)	0.02 (0.02)	0.51 (0.04)	0.51 (0.04)
AtomPair-SVM	0.50 (0.01)	0.01 (0.01)	0.49 (0.01)	0.50 (0.01)
AtomPair-GBM	0.51 (0.01)	0.01 (0.01)	0.50 (0.02)	0.51 (0.02)
Dragon-SVM	0.50 (0.01)	0.01 (0.01)	0.51 (0.02)	0.50 (0.02)
Dragon-GBM	0.51 (0.01)	0.01 (0.02)	0.51 (0.02)	0.51 (0.02)
CDK-SVM	0.50 (0.01)	0.00 (0.01)	0.51 (0.01)	0.50 (0.01)

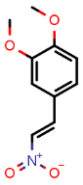
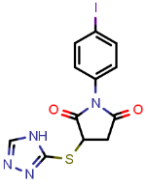
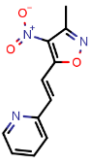
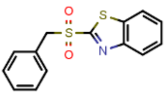
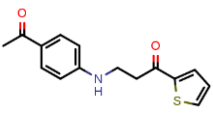
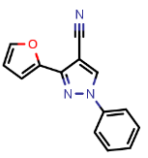
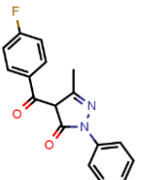
CCR: correct classification rate; k : Cohen's kappa; SE: sensitivity; SP: specificity

Table S5. Statistical characteristics of consensus and consensus rigor models developed.

Model	CCR	k	SE	SP	Coverage	Combination used
Consensus	0.87	0.74	0.87	0.88	1.00	Morgan-RF + MACCS-RF + AtomPair-SVM + Dragon-SVM + CDK-SVM
Consensus rigor	0.91	0.81	0.96	0.87	0.38	
Consensus	0.87	0.73	0.86	0.87	1.00	Morgan-RF + MACCS-RF + AtomPair-GBM + Dragon-GBM + CDK-GBM
Consensus rigor	0.91	0.80	0.95	0.87	0.38	
Consensus	0.87	0.73	0.86	0.87	1.00	Morgan-RF + MACCS-RF + AtomPair-RF + Dragon-SVM + CDK-SVM
Consensus rigor	0.91	0.81	0.95	0.87	0.38	
Consensus	0.85	0.71	0.86	0.85	1.00	Morgan-RF + MACCS-RF + AtomPair-RF + Dragon-RF + CDK-RF
Consensus rigor	0.89	0.77	0.95	0.84	0.38	
Consensus	0.87	0.74	0.87	0.87	1.00	Morgan-RF + MACCS-RF + AtomPair-SVM + Dragon-GBM + CDK-GBM
Consensus rigor	0.91	0.80	0.95	0.87	0.38	
Consensus	0.87	0.74	0.87	0.88	1.00	Morgan-RF + MACCS-RF + AtomPair-GBM + Dragon-SVM + CDK-SVM
Consensus rigor	0.91	0.81	0.96	0.86	0.38	
Consensus	0.87	0.74	0.87	0.87	1.00	Morgan-RF + MACCS-RF + AtomPair-GBM + Dragon-GBM + CDK-SVM
Consensus rigor	0.91	0.81	0.95	0.87	0.38	
Consensus	0.87	0.74	0.87	0.87	1.00	Morgan-GBM + MACCS-GBM + AtomPair-SVM+ Dragon-SVM + CDK-SVM
Consensus rigor	0.91	0.81	0.96	0.87	0.38	
Consensus	0.87	0.74	0.87	0.87	1.00	Morgan-GBM + MACCS-RF + AtomPair-SVM+ Dragon-SVM + CDK-SVM
Consensus rigor	0.91	0.81	0.96	0.86	0.38	
Consensus	0.87	0.74	0.87	0.87	1.00	Morgan-RF + MACCS-GBM + AtomPair-SVM+ Dragon-SVM + CDK-SVM
Consensus rigor	0.91	0.81	0.96	0.87	0.38	
Consensus	0.87	0.73	0.86	0.87	1.00	Morgan-GBM + MACCS-GBM + AtomPair-GBM+ Dragon-GBM + CDK-GBM
Consensus rigor	0.90	0.78	0.95	0.86	0.38	
Consensus	0.87	0.74	0.86	0.88	1.00	Morgan-SVM + MACCS-SVM + AtomPair-GBM+ Dragon-GBM+ CDK-GBM
Consensus rigor	0.91	0.80	0.95	0.88	0.38	

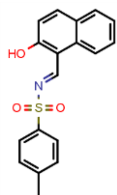
CCR: correct classification rate; k : Cohen's kappa; SE: sensitivity; SP: specificity

Table S6. The predictions for the PZQ, OLT, and 29 putative hits using more predictive consensus model and its motility and phenotype adjusted index values obtained for *S. mansoni* schistosomula exposed for 48h at a 20 μM concentration. The positive controls PZQ and OLT were tested at 10 μM concentration.

Compound	Chemical structure	Prob.	AD	Motility adjusted index (mean \pm SD)	Phenotype adjusted index (mean \pm SD)
1,2-dimethoxy-4-(2-nitrovinyl)benzene (LabMol-23, 1)		1.0	0.6	-0.74 \pm 0.08	-0.64 \pm 0.01
1-(4-iodophenyl)-3-(4H-1,2,4-triazol-3-ylthio)-2,5-pyrrolidinedione (LabMol-28, 2)		1.0	0.6	-0.61 \pm 0.09	-0.48 \pm 0.05
2-[2-(3-methyl-4-nitro-5-isoxazolyl)vinyl]pyridine (LabMol-37, 3)		1.0	0.6	-0.92 \pm 0.04	-0.58 \pm 0.01
2-(benzylsulfonyl)-1,3-benzothiazole (LabMol-49, 4)		1.0	0.8	-0.93 \pm 0.02	-0.61 \pm 0.09
3-[(4-acetylphenyl)amino]-1-(2-thienyl)-1-propanone (LabMol-50, 5)		1.0	0.8	-0.89 \pm 0.02	-0.40 \pm 0.03
3-(2-furyl)-1-phenyl-1H-pyrazole-4-carbonitrile (LabMol-51, 6)		1.0	1.0	-0.33 \pm 0.13	-0.24 \pm 0.05
4-(4-fluorobenzoyl)-5-methyl-2-phenyl-2,4-dihydro-3H-pyrazol-3-one (LabMol-24, 7)		1.0	0.8	0.15 \pm 0.05	-0.17 \pm 0.04

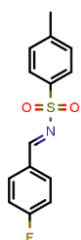
Supporting Information

N-[(2-hydroxy-1-naphthyl)methylene]-4-methylbenzenesulfonamide (LabMol-25, 8)



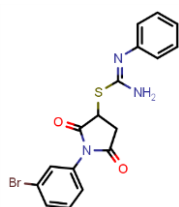
1.0 0.6 -0.12 ± 0.09 -0.08 ± 0.00

N-(4-fluorobenzylidene)-4-methylbenzenesulfonamide (LabMol-26, 9)



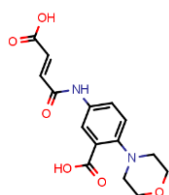
1.0 0.6 -0.04 ± 0.04 -0.05 ± 0.02

1-(3-bromophenyl)-2,5-dioxo-3-pyrrolidinyl N'-phenylimidothiocarbamate (LabMol-27, 10)



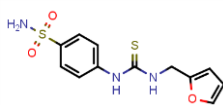
1.0 0.6 -0.49 ± 0.06 -0.08 ± 0.04

5-[(3-carboxyacryloyl)amino]-2-(4-morpholinyl)benzoic acid (LabMol-29, 11)



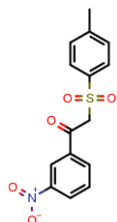
1.0 0.6 0.34 ± 0.10 -0.10 ± 0.14

4-(((2-furylmethyl)amino)carbonothioyl)amino)benzenesulfonamide (LabMol-30, 12)



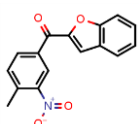
1.0 0.6 0.13 ± 0.01 -0.15 ± 0.04

2-[(4-methylphenyl)sulfonyl]-1-(3-nitrophenyl)ethanone (LabMol-31, 13)



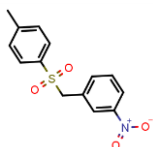
1.0 0.6 0.22 ± 0.04 -0.11 ± 0.03

1-benzofuran-2-yl(4-methyl-3-nitrophenyl)methanone (LabMol-32, 14)



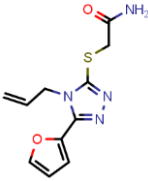
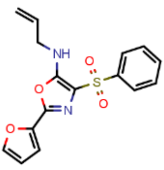
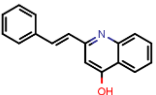
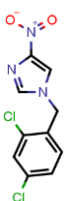
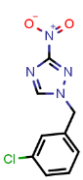
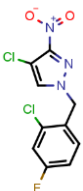
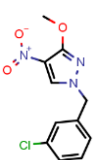
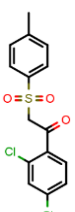
1.0 0.8 0.06 ± 0.05 -0.12 ± 0.01

1-(((4-methylphenyl)sulfonyl)methyl)-3-nitrobenzene (LabMol-33, 15)

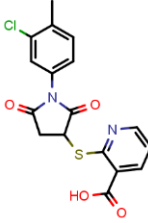
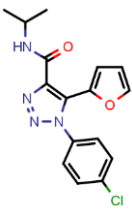
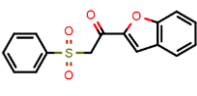

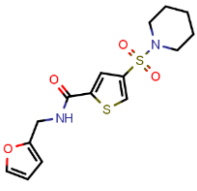
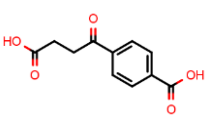
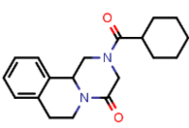
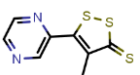


1.0 0.8 -0.07 ± 0.08 -0.12 ± 0.06

Supporting Information

2-{{[4-allyl-5-(2-furyl)-4H-1,2,4-triazol-3-yl]thio}acetamide (LabMol-34, 16)		1.0	0.6	0.17 ± 0.13	0.02 ± 0.15
N-allyl-2-(2-furyl)-4-(phenylsulfonyl)-1,3-oxazol-5-amine (LabMol-35, 17)		1.0	0.8	-0.52 ± 0.02	-0.08 ± 0.02
2-(2-phenylvinyl)-4-quinolinol (LabMol-36, 18)		1.0	1.0	-0.41 ± 0.10	-0.15 ± 0.02
1-(2,4-dichlorobenzyl)-4-nitro-1H-imidazole (LabMol-38, 19)		1.0	0.6	0.00 ± 0.15	-0.12 ± 0.07
1-(3-chlorobenzyl)-3-nitro-1H-1,2,4-triazole (LabMol-39, 20)		1.0	0.6	0.02 ± 0.06	-0.04 ± 0.02
4-chloro-1-(2-chloro-4-fluorobenzyl)-3-nitro-1H-pyrazole (LabMol-40, 21)		1.0	0.6	-0.28 ± 0.01	-0.13 ± 0.02
1-(3-chlorobenzyl)-3-methoxy-4-nitro-1H-pyrazole (LabMol-41, 22)		1.0	0.8	-0.26 ± 0.05	-0.19 ± 0.04
1-(2,4-dichlorophenyl)-2-[(4-methylphenyl)sulfonyl]ethanone (LabMol-42, 23)		1.0	0.8	-0.05 ± 0.19	-0.15 ± 0.08

Supporting Information

2-[[1-(3-chloro-4-methylphenyl)-2,5-dioxo-3-pyrrolidinyl]thio}nicotinic acid (LabMol-43, 24)		1.0	0.8	0.26 ± 0.10	-0.13 ± 0.01
1-(4-chlorophenyl)-5-(2-furyl)-N-isopropyl-1H-1,2,3-triazole-4-carboxamide (LabMol-44, 25)		1.0	0.8	0.20 ± 0.13	-0.10 ± 0.02
1-(1-benzofuran-2-yl)-2-(phenylsulfonyl)ethanone (LabMol-45, 26)		1.0	0.8	0.18 ± 0.12	-0.14 ± 0.02
N-[2-(3-pyridinyl)-2-(2-thienylsulfonyl)ethyl]-2-furamide (LabMol-46, 27)		1.0	0.8	0.35 ± 0.04	0.00 ± 0.12
N-(2-furylmethyl)-4-(1-piperidinylsulfonyl)-2-thiophenecarboxamide (LabMol-47, 28)		1.0	0.8	0.10 ± 0.05	-0.11 ± 0.03
4-(3-carboxypropanoyl)benzoic acid (LabMol-48, 29)		1.0	0.6	0.12 ± 0.07	-0.10 ± 0.01
Praziquantel (PZQ)		0.0	0.8	-0.48 ± 0.04	-0.17 ± 0.02
Oltipraz (OLT)		0.8	1.0	-0.90 ± 0.04	-0.34 ± 0.07

Prob.: probability; AD: applicability domain coverage; The identified hits are highlighted in bold fonts

Table S7. EC₅₀ values for the effect of investigated compounds and PZQ on the motility of male and female *S. mansoni* exposed for up to 72h.

Time (h)	Compound	EC ₅₀ (μM)	
		Male	Female
24h	PZQ	0.26	0.28
	1	9.59	12.3
	2	17.0	19.8
	3	17.5	16.1
	4	35.1	21.9
48h	PZQ	No fit	0.64
	1	29.8	5.77
	2	10.2	17.9
	3	6.43	5.68
	4	21.1	4.91
72h	PZQ	0.22	0.59
	1	28.1	No Fit
	2	11.4	No Fit
	3	No fit	5.84
	4	20.4	5.83

Supporting Information

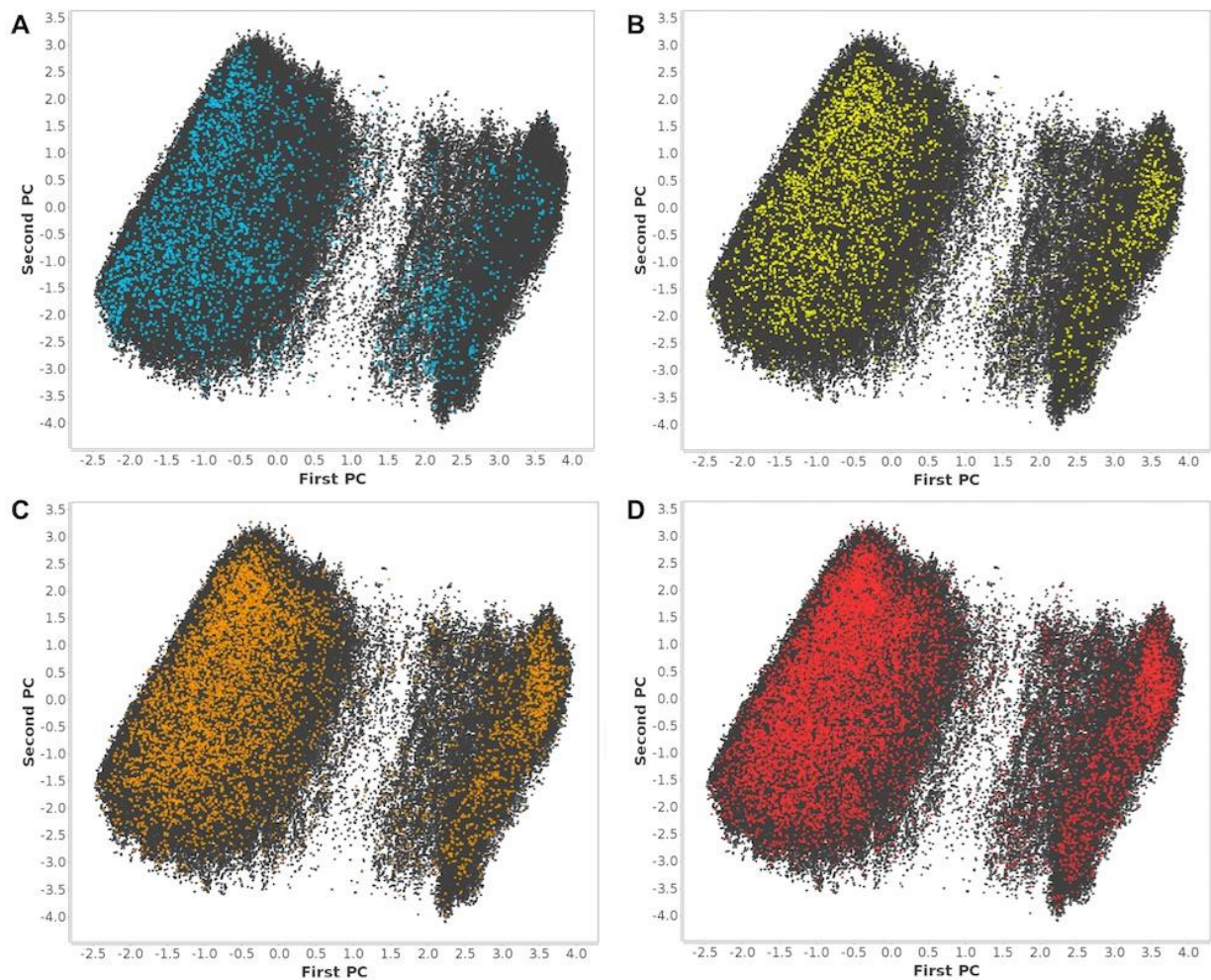


Figure S1. Chemical space of *SmTGR* inhibitors and non-inhibitors into first two PCs computed using MACCS keys; blue dots representing the 2,854 inhibitors at 10 μM threshold (A); Yellow (B), orange (C), and red (D) dots represent non-inhibitors selected for dataset balancing with ratios of 1:1, 1:2, and 1:3 correspondingly; grey dots represent remaining non-inhibitors.

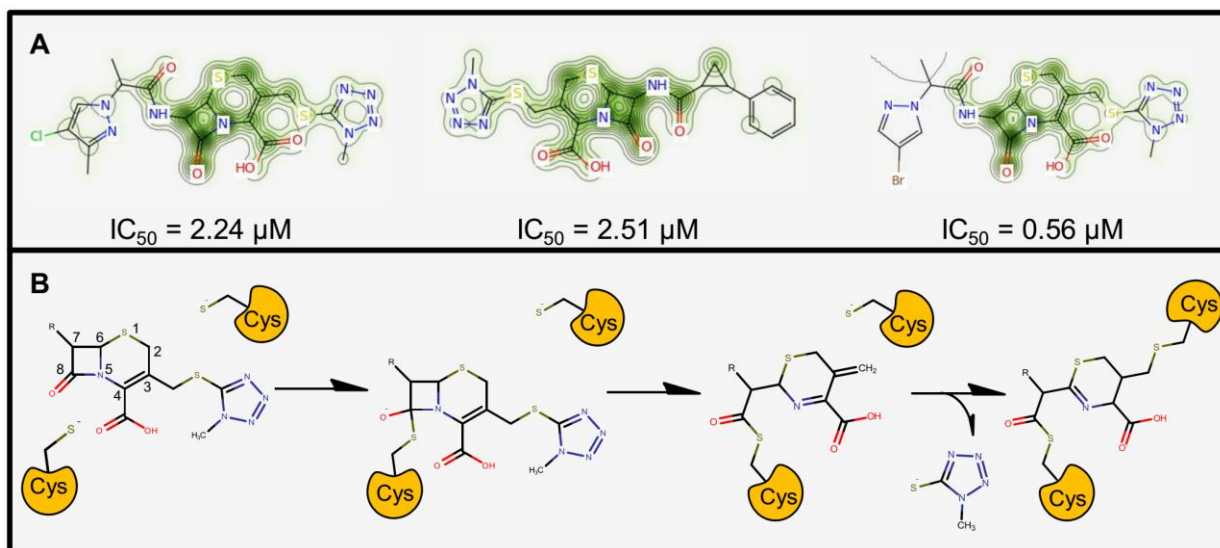


Figure S2. Predicted probability maps generated for cephalosporins (A) and their probable reaction mechanism in the *SmTGR* active site (B).

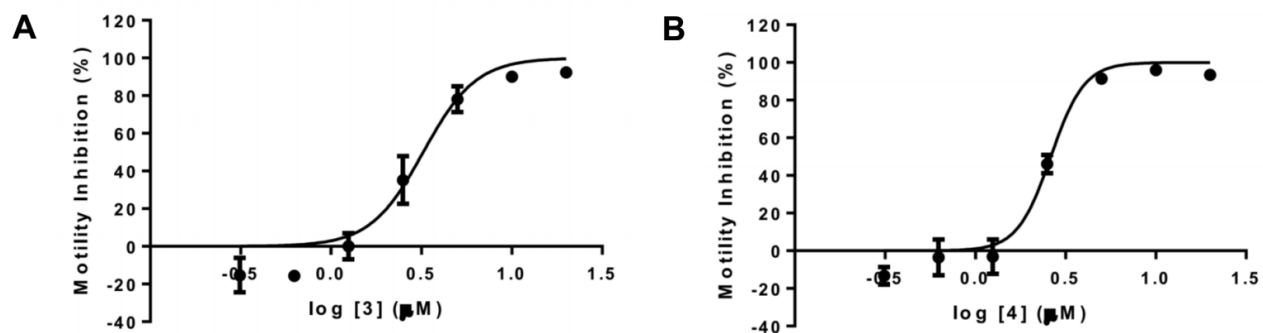


Figure S3. Motility dose-response curves for compounds **3** (A) and **4** (B) against *S. mansoni* larvae after 48h of incubation.

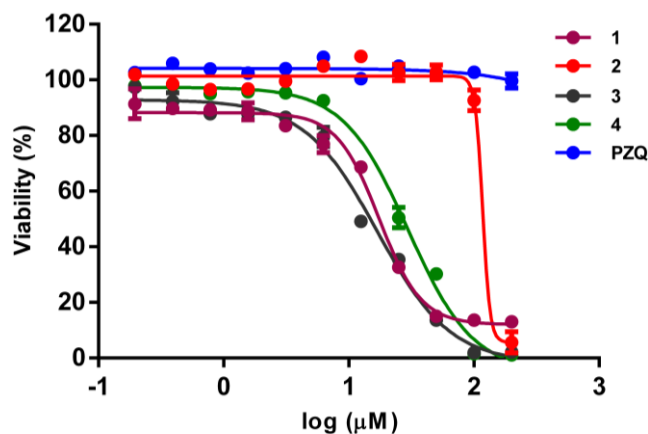


Figure S4. Viability dose-responses curves for compounds **1–4** and PZQ against WSS-1 human cells after 48h of incubation.

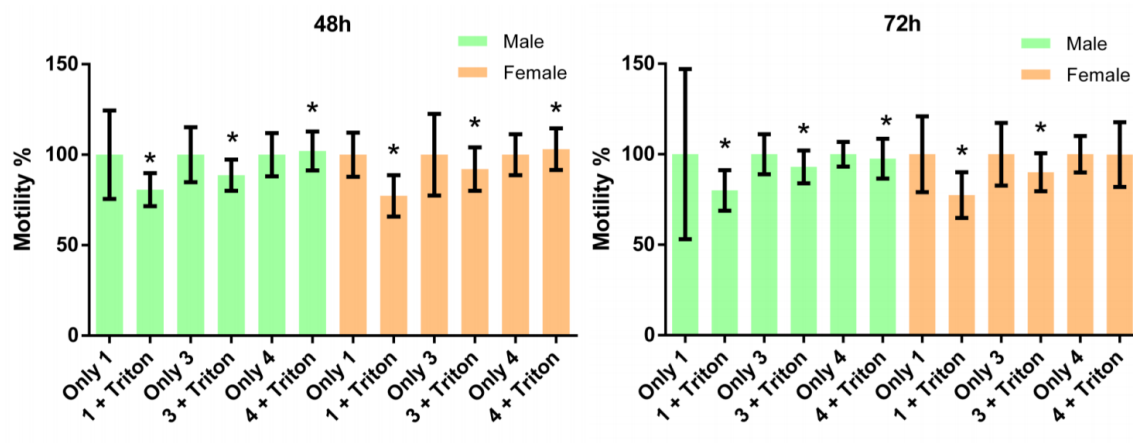


Figure S5. Effect of 0.01% Triton X-100 co-incubated with compounds **1, 3** and **4** on adult male and female *S. mansoni* worms. Motility measurements were performed after 48h and 72h, and compared with group without detergent. Data expressed as mean \pm standard deviation. * $P \leq 0.05$ using student-*t* test.

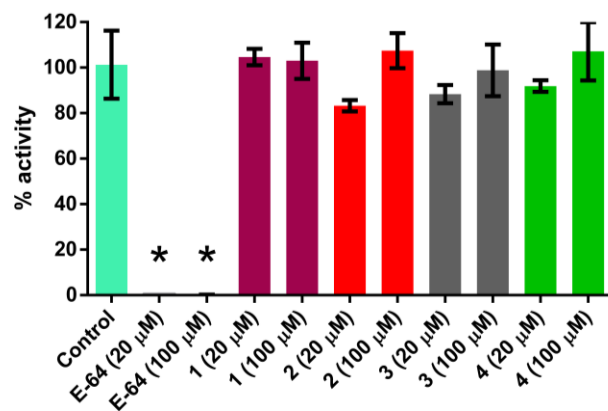


Figure S6. Papain activity in the presence of compounds 1–4 and E-64. Data expressed as mean \pm standard deviation. * $P \leq 0.05$ relative to control using ANOVA followed by Dunnett’s test.

Supporting Information

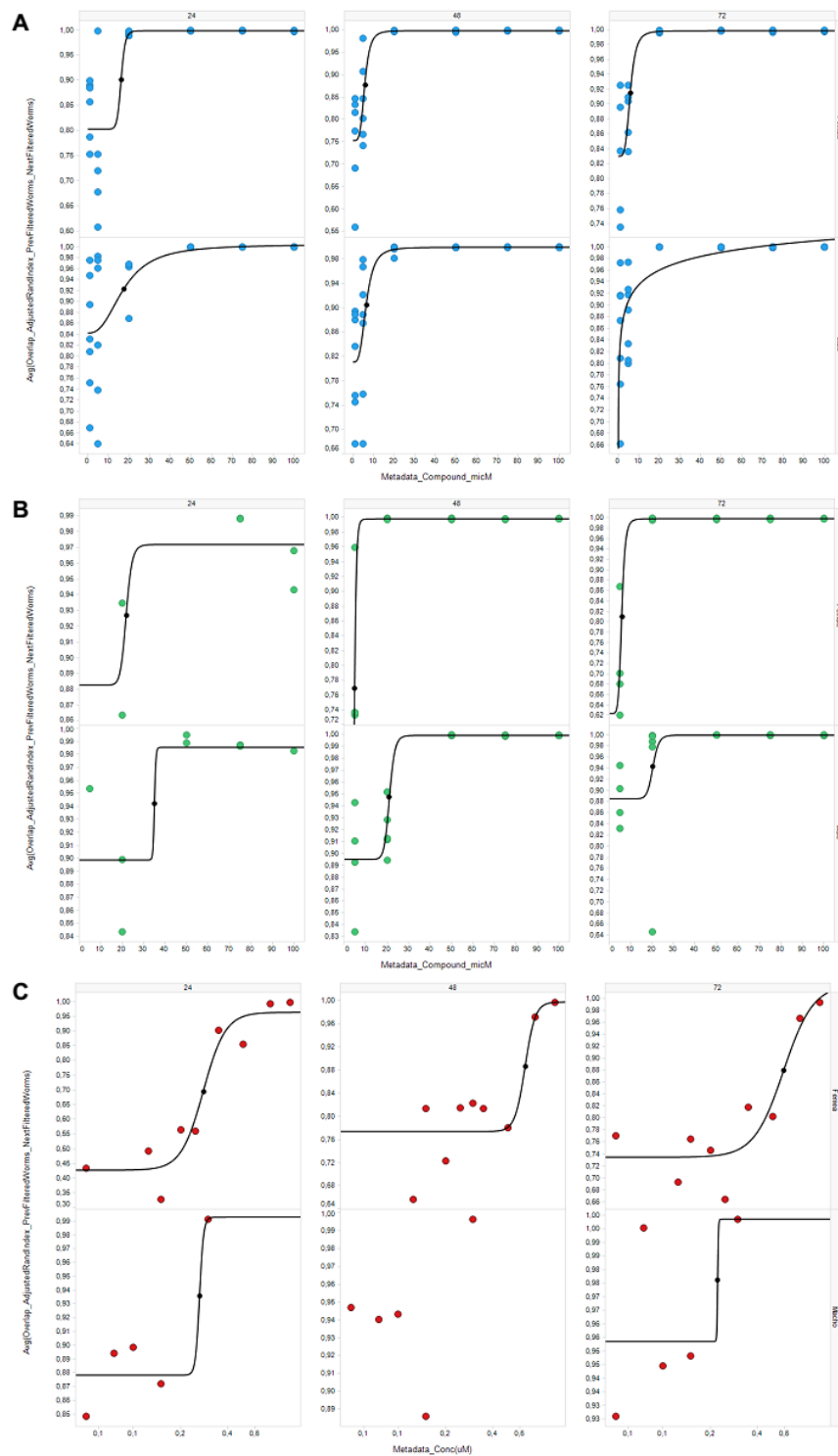
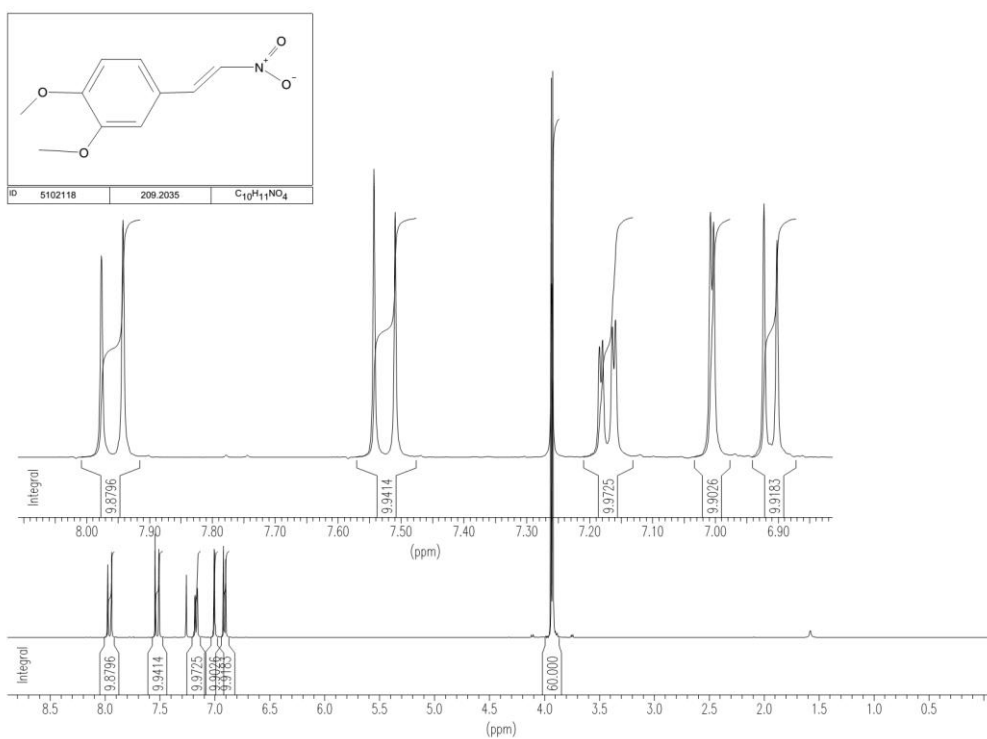


Figure S7. Dose-response curves for compounds 3 (A), 4 (B) and PZQ (C) on the motility of male and female *S. mansoni* exposed for up to 72h of incubation.

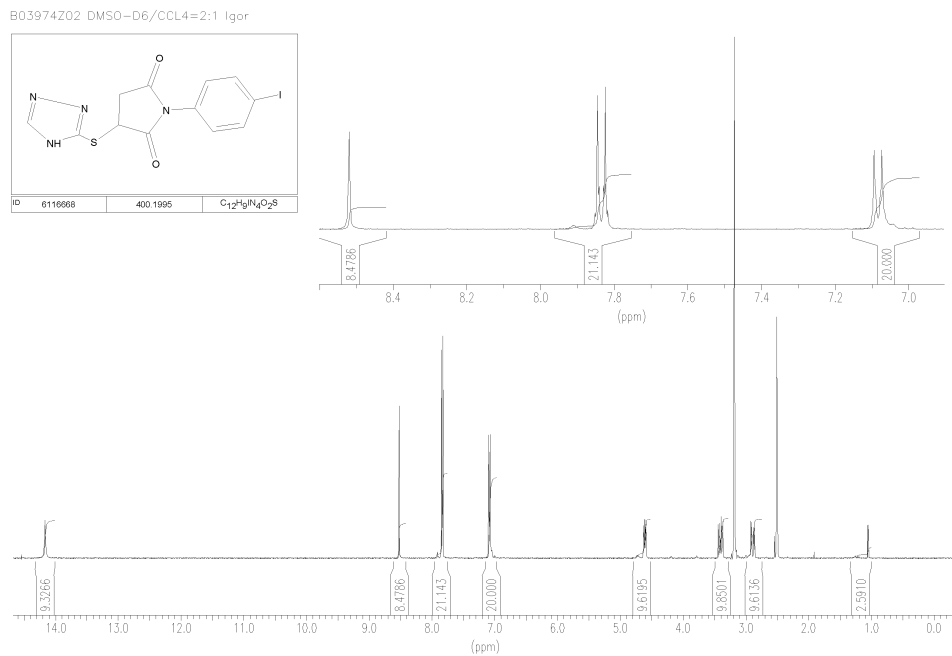
Nuclear Magnetic Resonance (NMR) and Purity Data. The chemical structure of all compounds purchased from ChemBridge was confirmed using proton (^1H) NMR spectra at 300/400 MHz. The ^1H RMN spectrums of compounds are listed below. The Liquid Chromatography–Mass Spectrometry (LC-MS) analysis with evaporative light scattering and ultraviolet detectors confirmed a minimum purity of 95% for all samples.

1,2-dimethoxy-4-(2-nitrovinyl)benzene (LabMol-23, **1**);

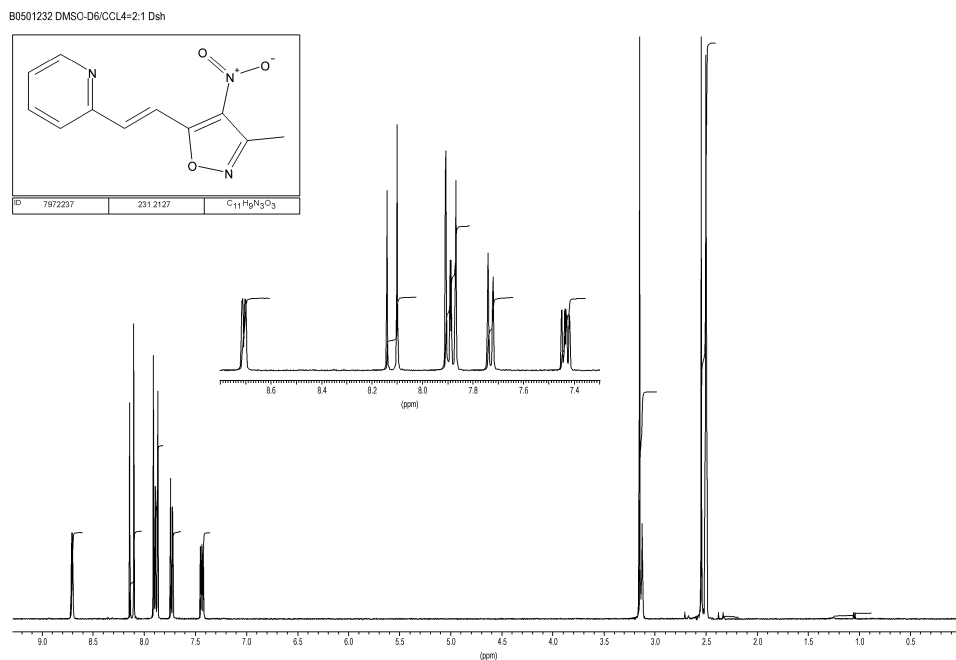


Supporting Information

1-(4-iodophenyl)-3-(4H-1,2,4-triazol-3-ylthio)-2,5-pyrrolidinedione (LabMol-28, **2**);

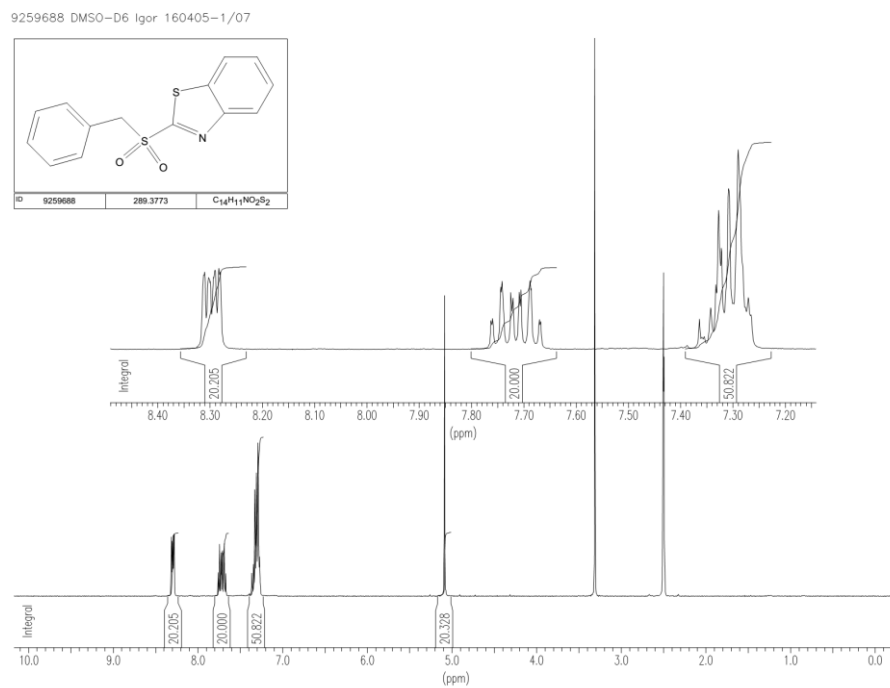


2-[2-(3-methyl-4-nitro-5-isoxazolyl)vinyl]pyridine (LabMol-37, **3**);

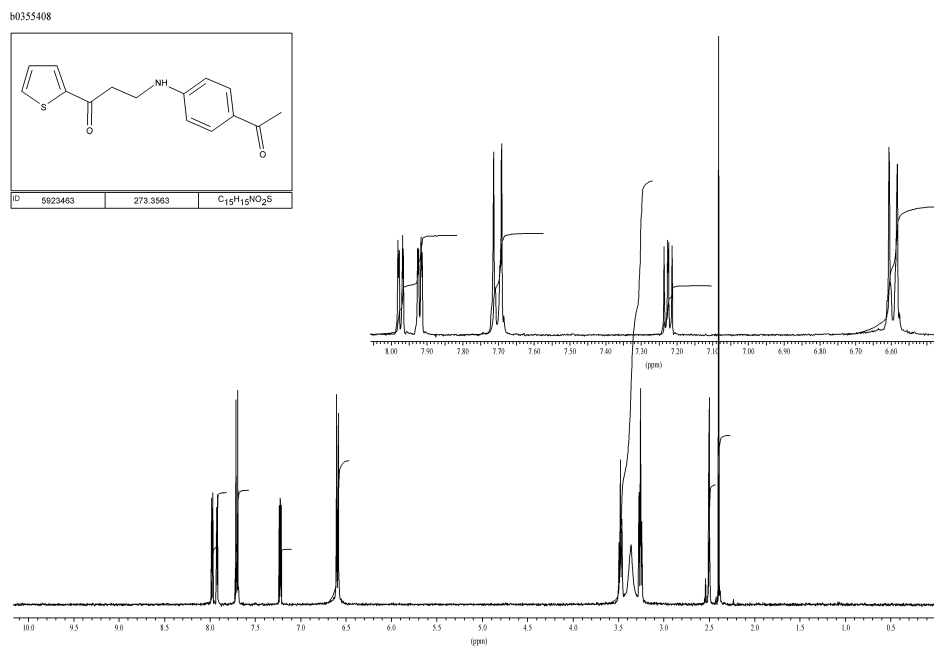


Supporting Information

2-(benzylsulfonyl)-1,3-benzothiazole (LabMol-49, **4**);

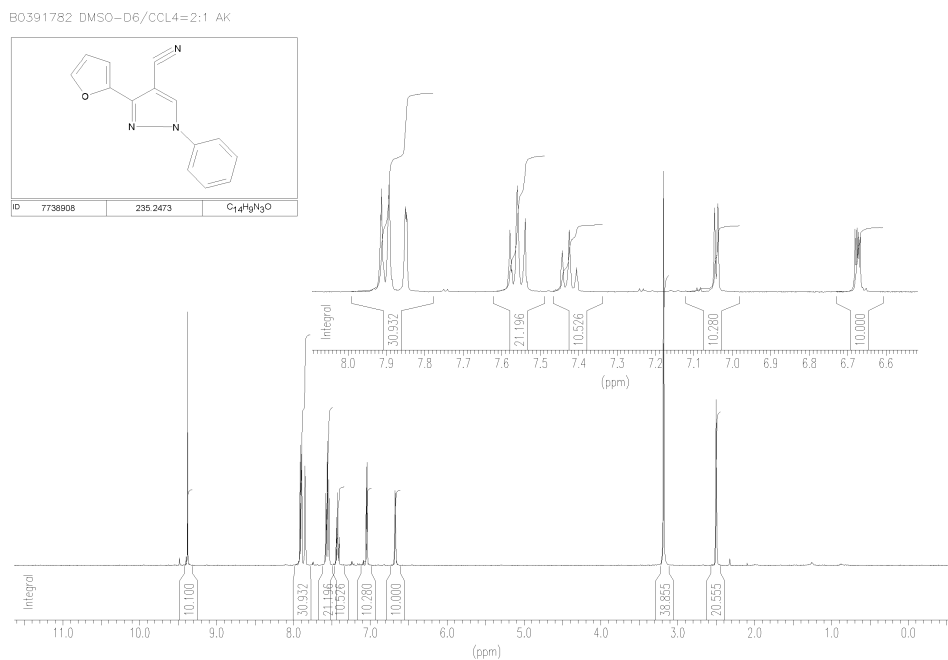


3-[(4-acetylphenyl)amino]-1-(2-thienyl)-1-propanone (LabMol-50, **5**);

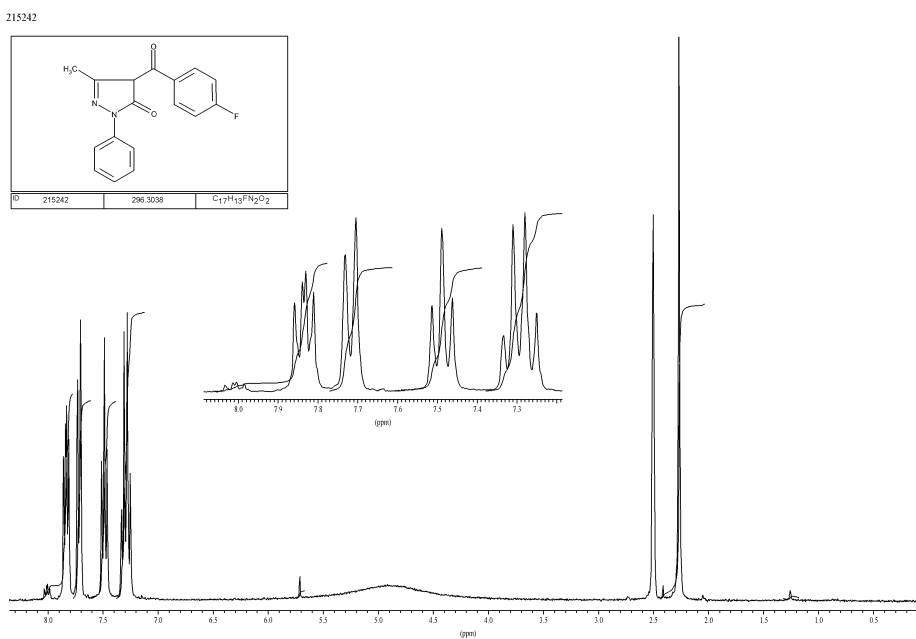


Supporting Information

3-(2-furyl)-1-phenyl-1H-pyrazole-4-carbonitrile (LabMol-51, **6**);

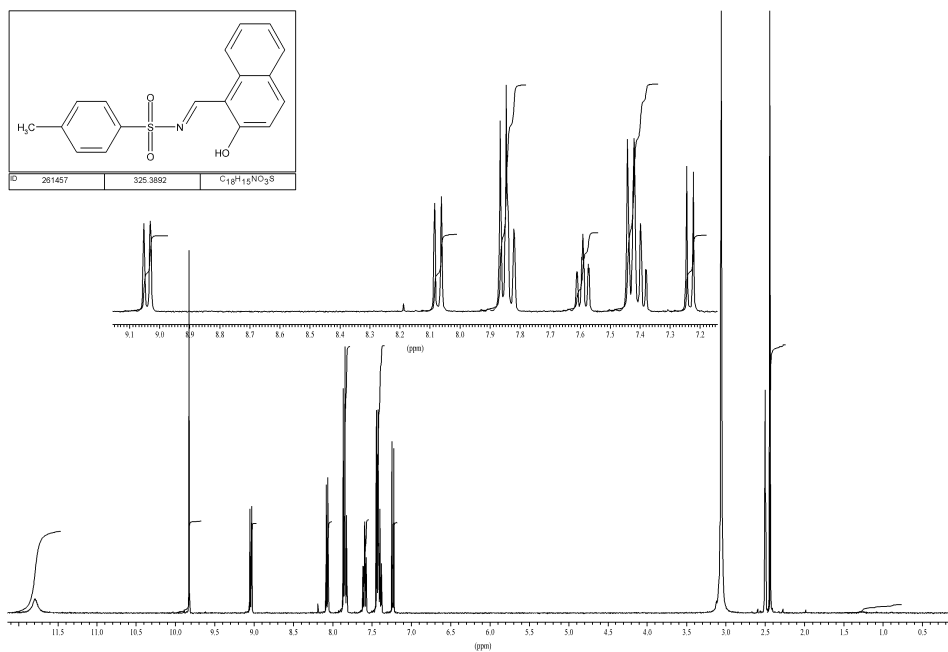


4-(4-fluorobenzoyl)-5-methyl-2-phenyl-2,4-dihydro-3H-pyrazol-3-one (LabMol-24, **7**);

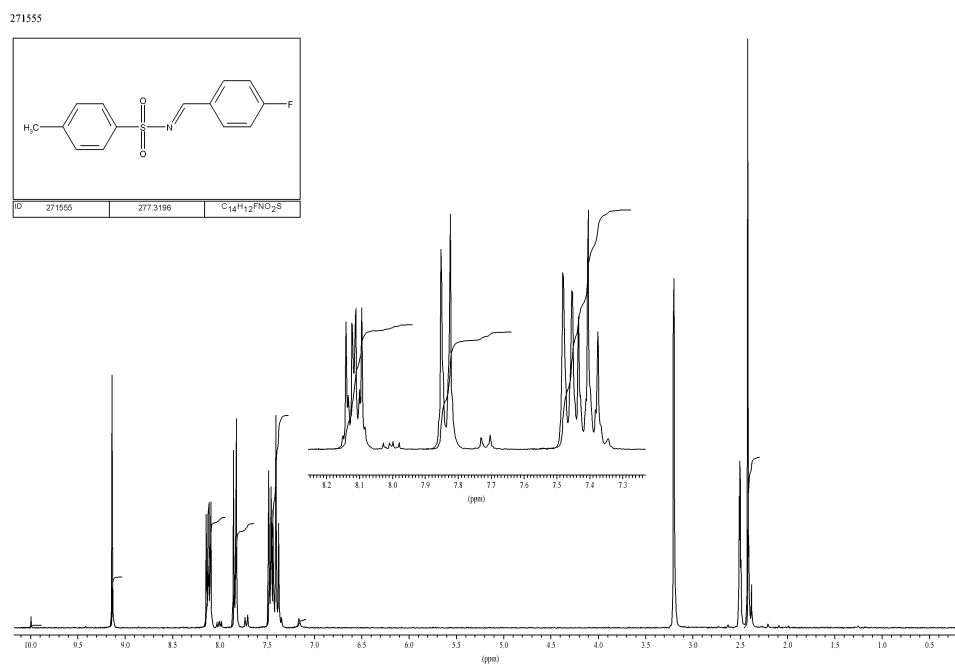


Supporting Information

N-[(2-hydroxy-1-naphthyl)methylene]-4-methylbenzenesulfonamide (LabMol-25, **8**);



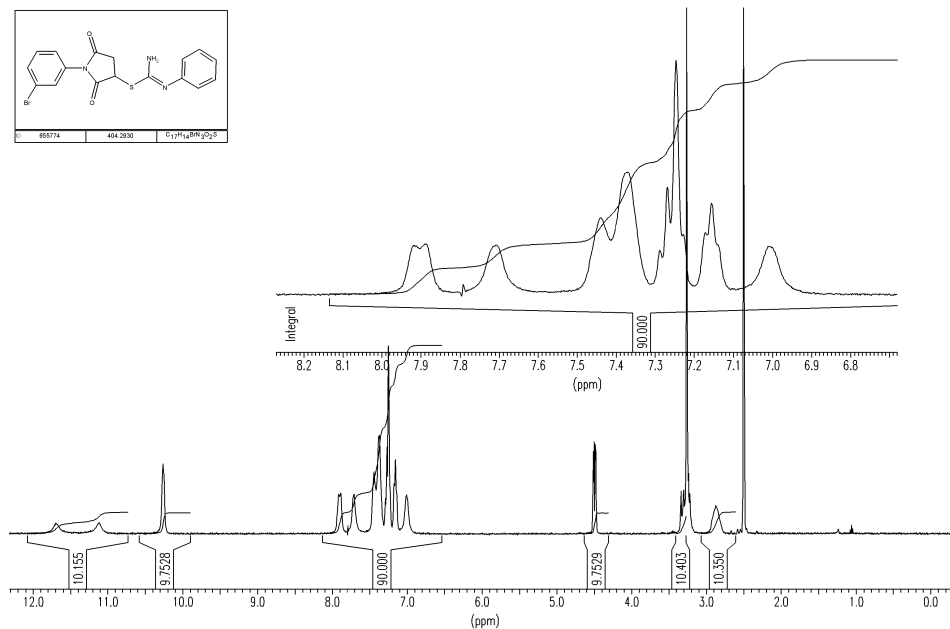
N-(4-fluorobenzylidene)-4-methylbenzenesulfonamide (LabMol-26, **9**);



Supporting Information

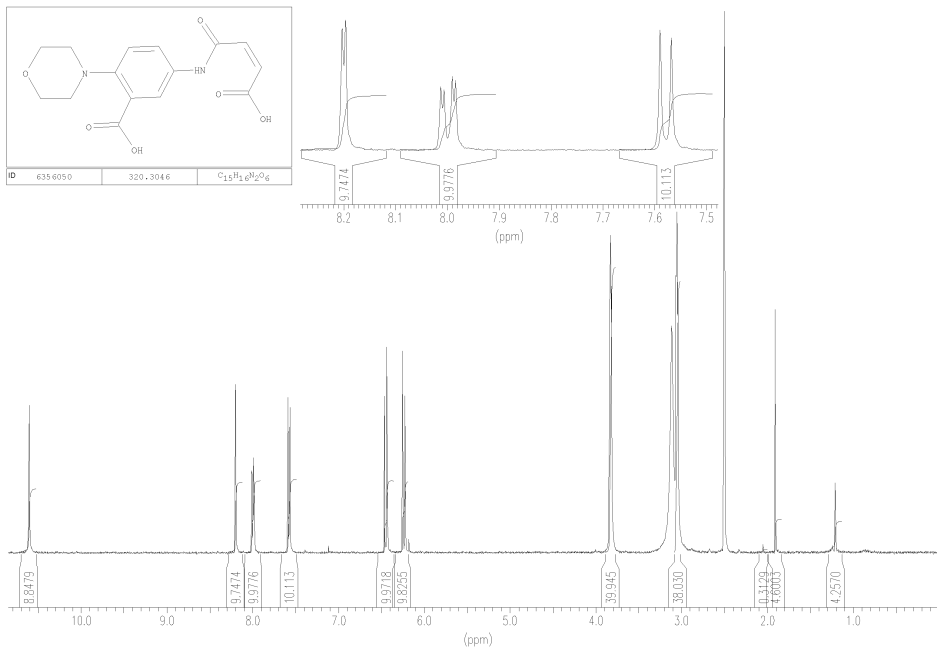
1-(3-bromophenyl)-2,5-dioxo-3-pyrrolidinyl N'-phenylimidithiocarbamate (LabMol-27, **10**);

655774x1 in DMSO.



5-[(3-carboxyacryloyl)amino]-2-(4-morpholinyl)benzoic acid (LabMol-29, **11**);

B1161-72

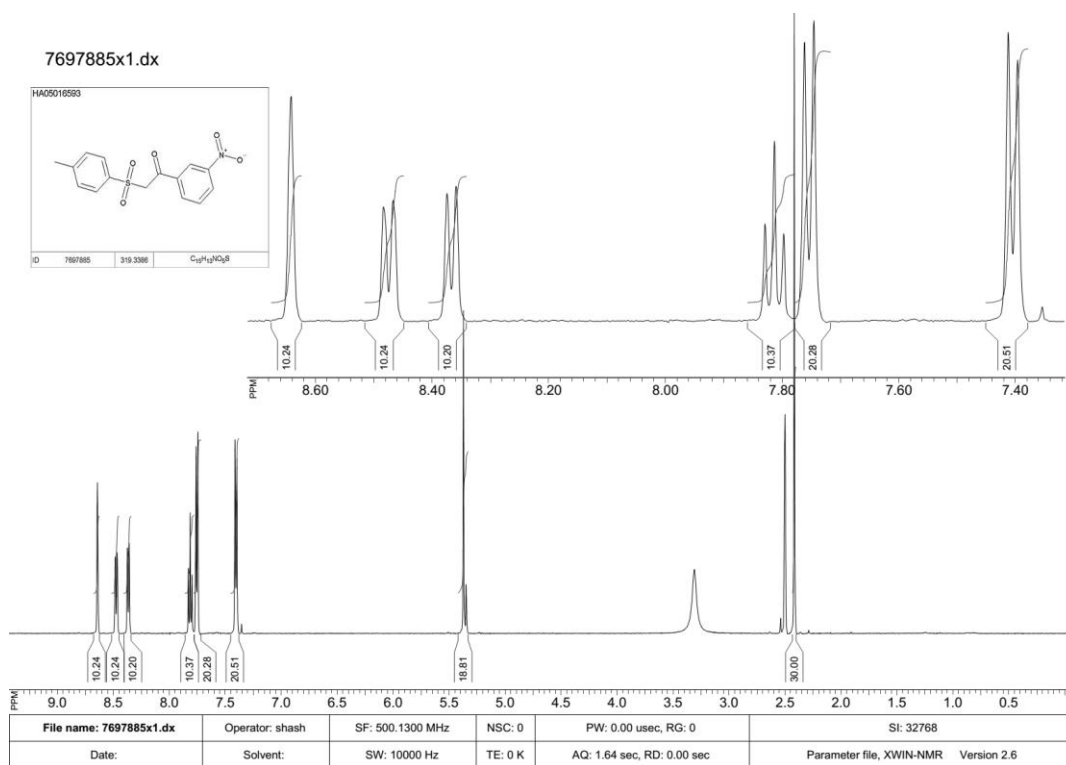


Supporting Information

4-({[(2-furylmethyl)amino]carbonothioyl} amino)benzenesulfonamide (LabMol-30, **12**);

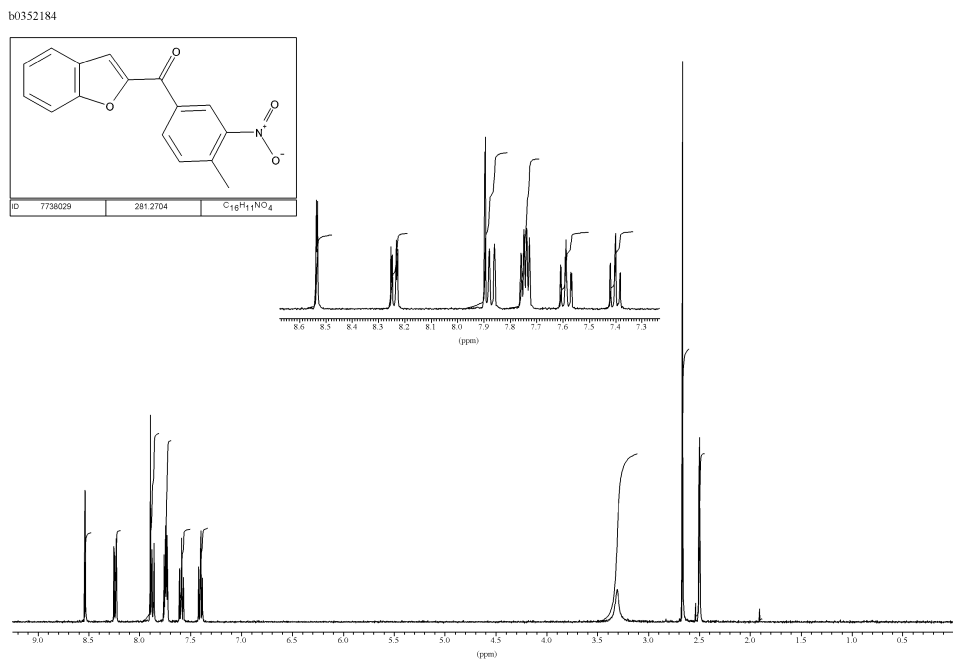


2-[(4-methylphenyl)sulfonyl]-1-(3-nitrophenyl)ethanone (LabMol-31, **13**);

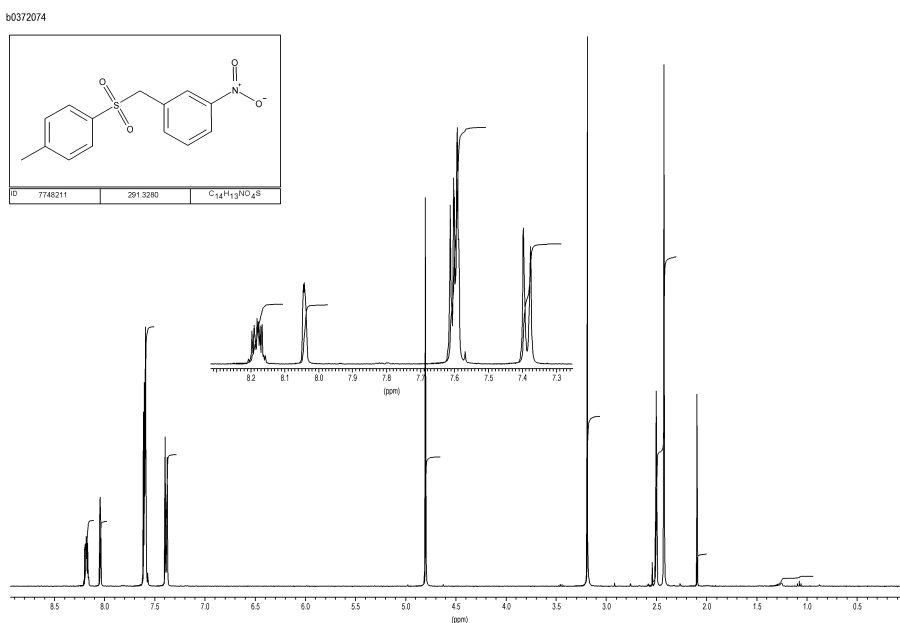


Supporting Information

1-benzofuran-2-yl(4-methyl-3-nitrophenyl)methanone (LabMol-32, **14**);

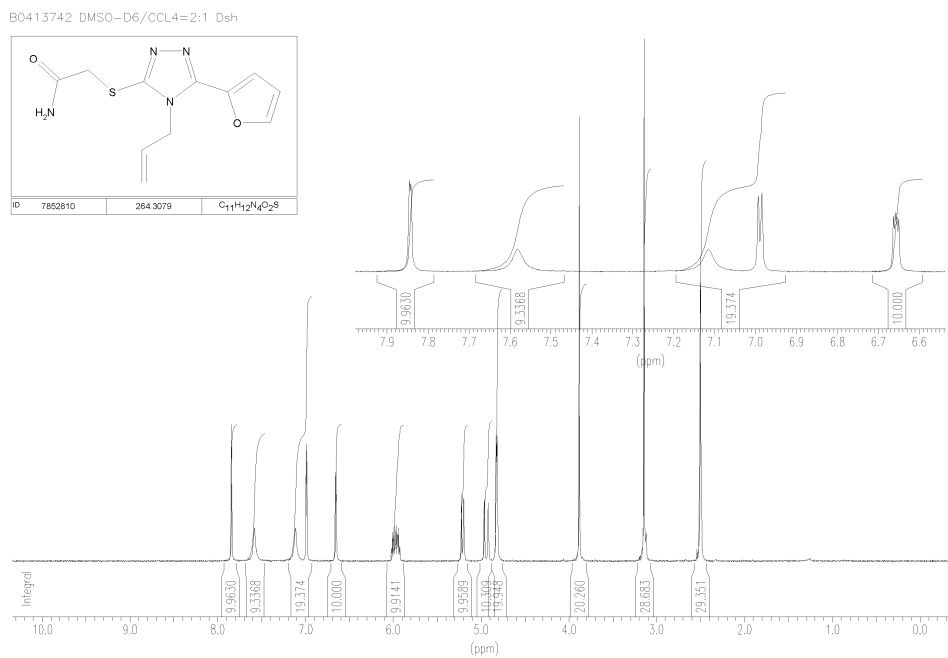


1-{[(4-methylphenyl)sulfonyl]methyl}-3-nitrobenzene (LabMol-33, **15**);

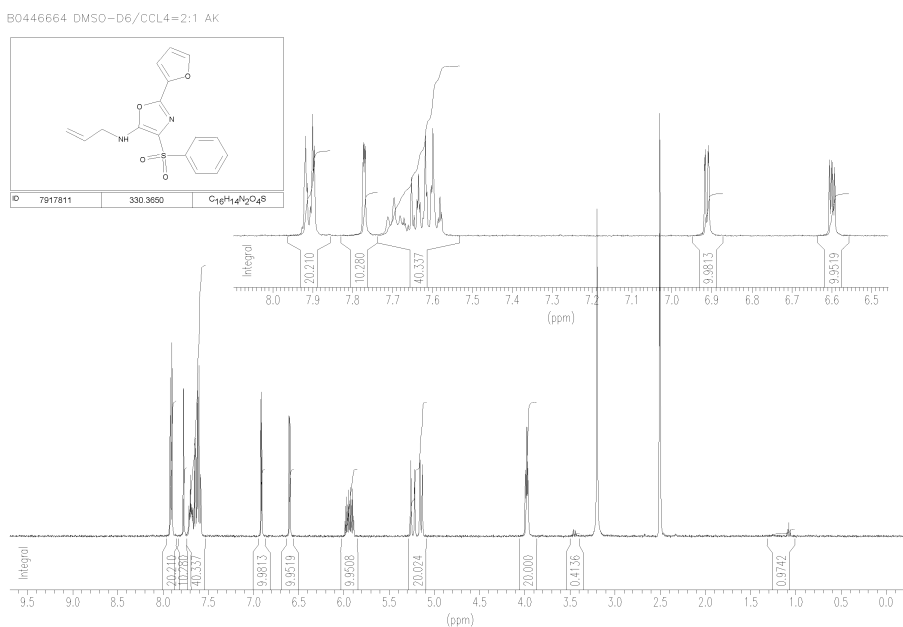


Supporting Information

2-{[4-allyl-5-(2-furyl)-4H-1,2,4-triazol-3-yl]thio}acetamide (LabMol-34, **16**);



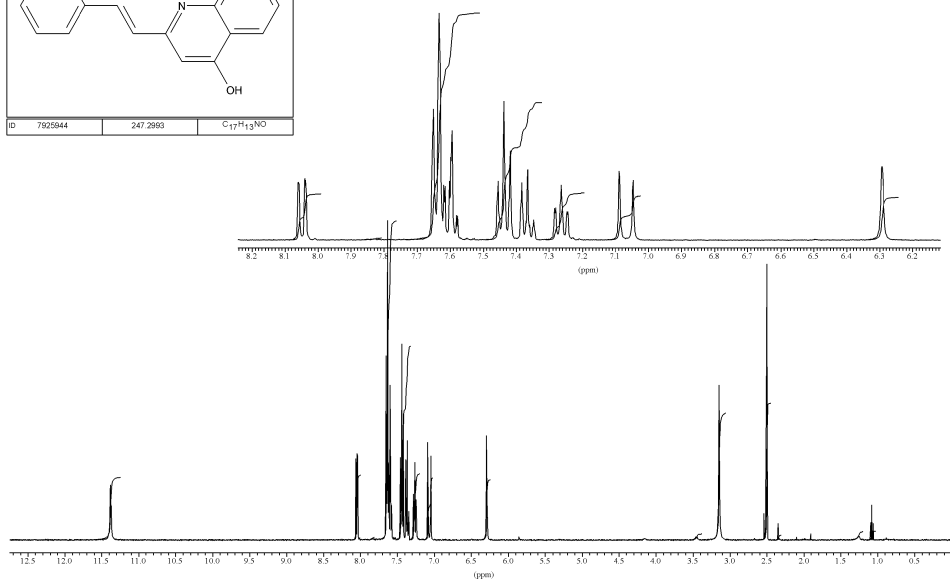
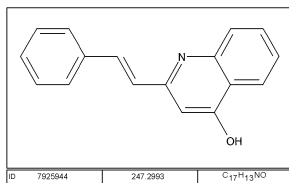
N-allyl-2-(2-furyl)-4-(phenylsulfonyl)-1,3-oxazol-5-amine (LabMol-35, **17**);



Supporting Information

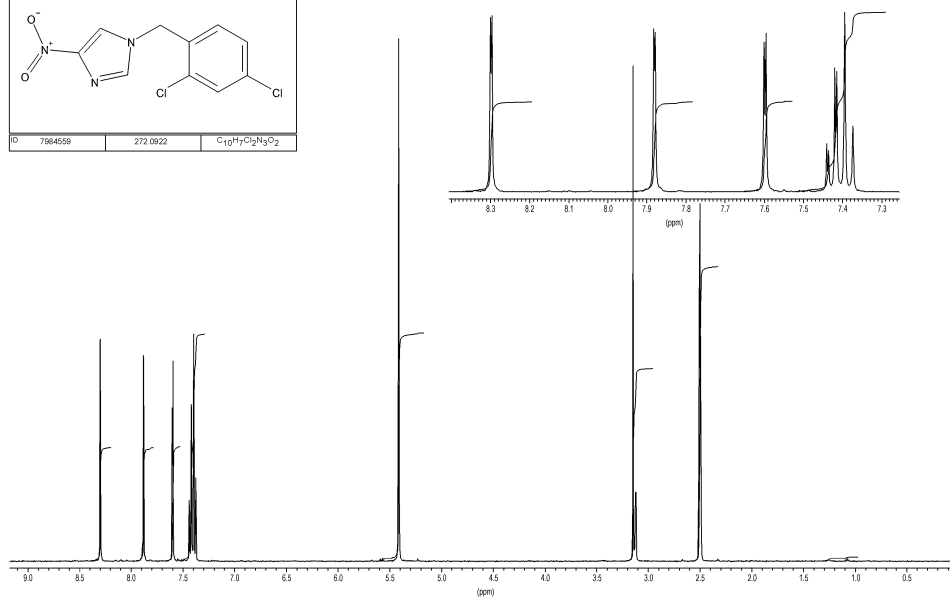
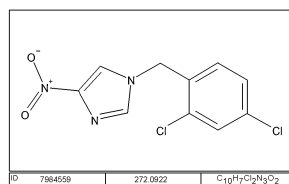
2-(2-phenylvinyl)-4-quinolinol (LabMol-36, **18**);

B0425855 DMSO-D6/CCL4=2:1 AK



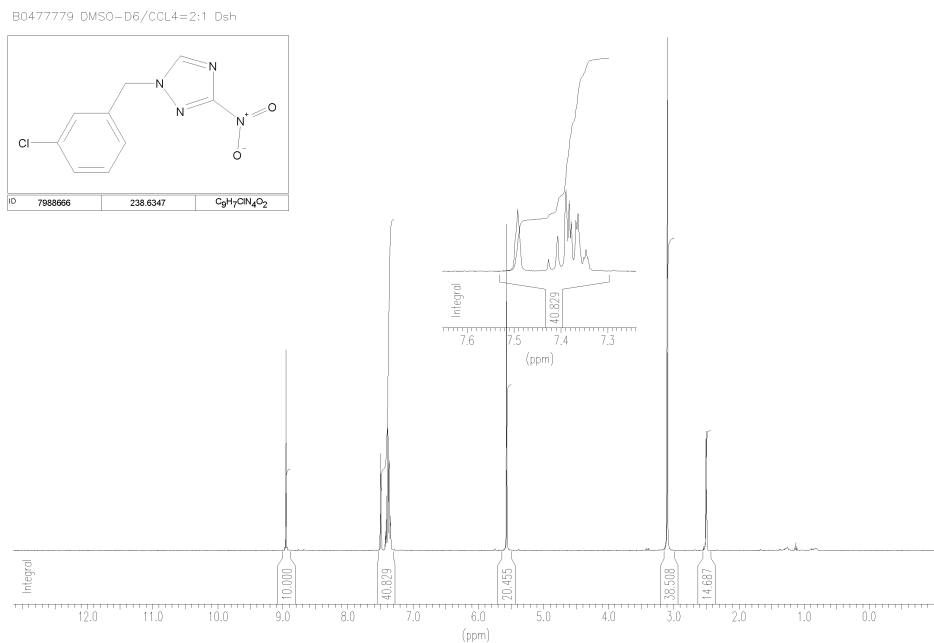
1-(2,4-dichlorobenzyl)-4-nitro-1H-imidazole (LabMol-38, **19**);

B0503023 DMSO-D6/CCL4=2:1 Dsh

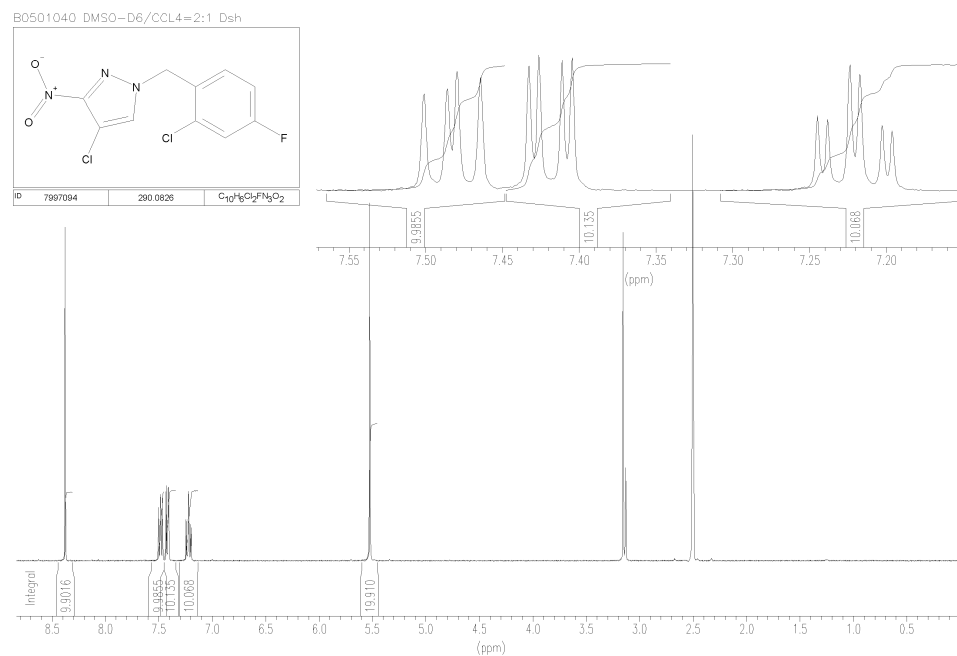


Supporting Information

1-(3-chlorobenzyl)-3-nitro-1H-1,2,4-triazole (LabMol-39, **20**);

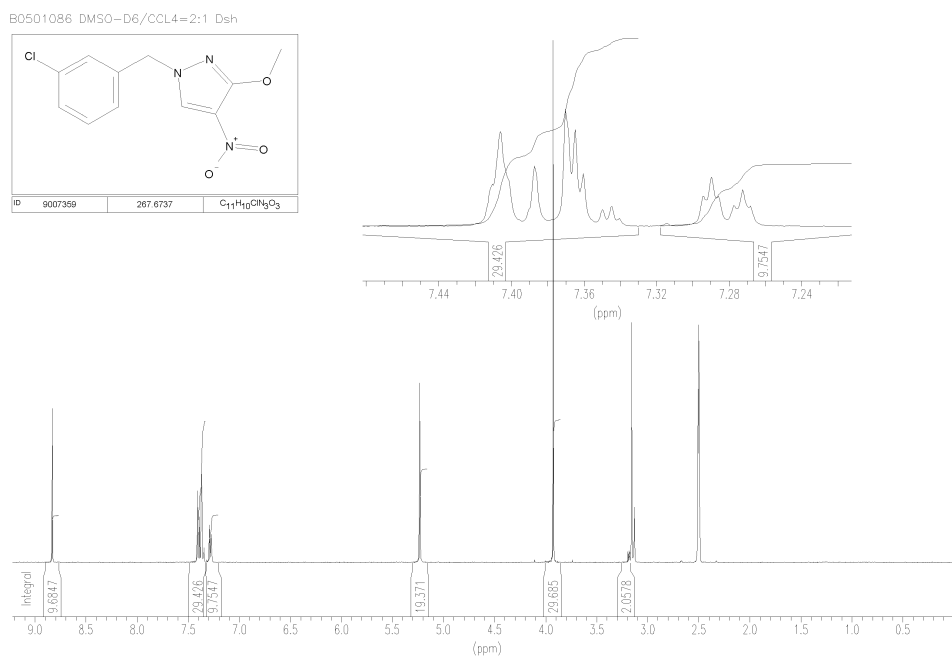


4-chloro-1-(2-chloro-4-fluorobenzyl)-3-nitro-1H-pyrazole (LabMol-40, **21**);

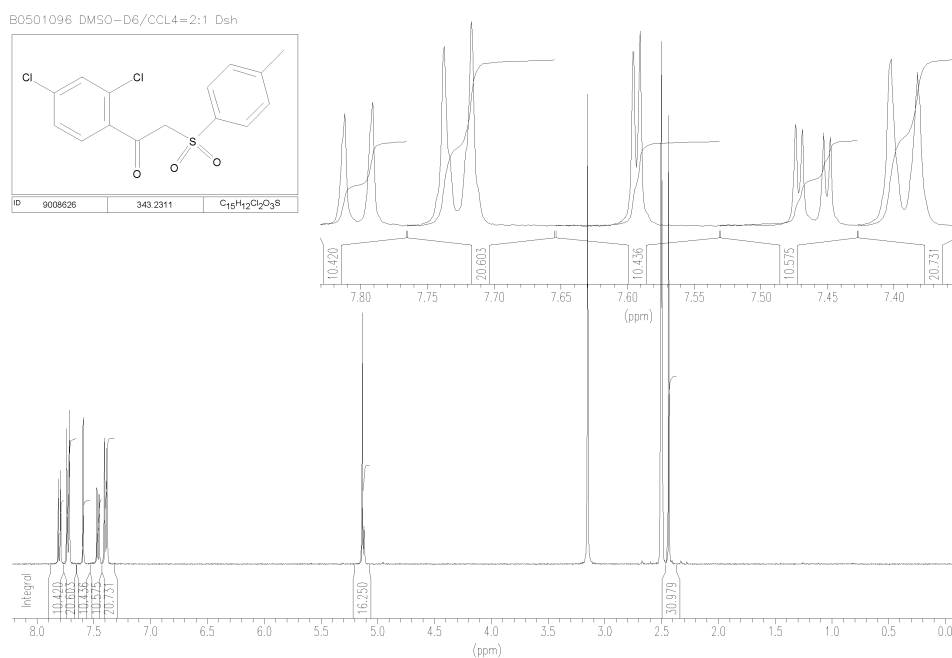


Supporting Information

1-(3-chlorobenzyl)-3-methoxy-4-nitro-1H-pyrazole (LabMol-41, **22**);

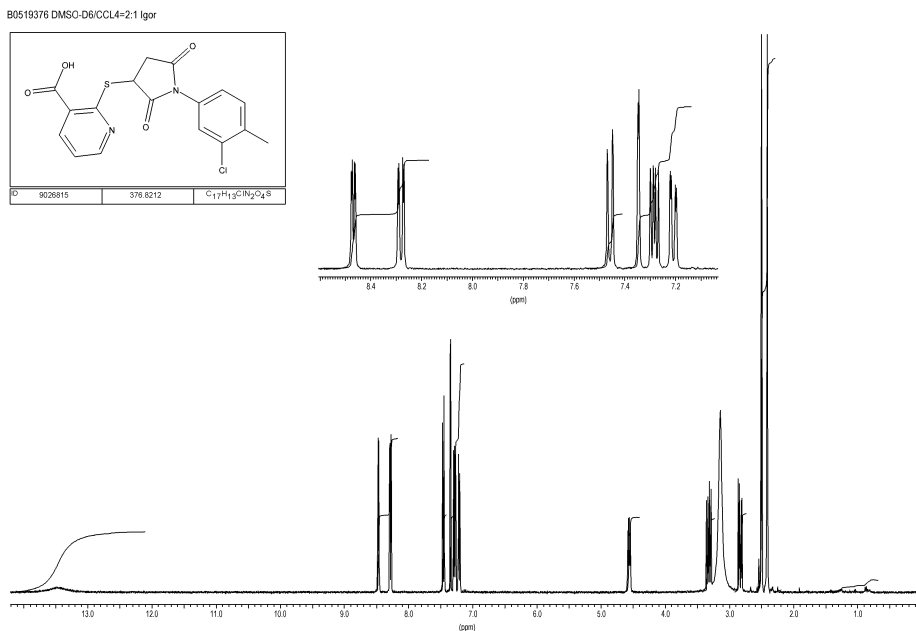


1-(2,4-dichlorophenyl)-2-[(4-methylphenyl)sulfonyl]ethanone (LabMol-42, **23**);

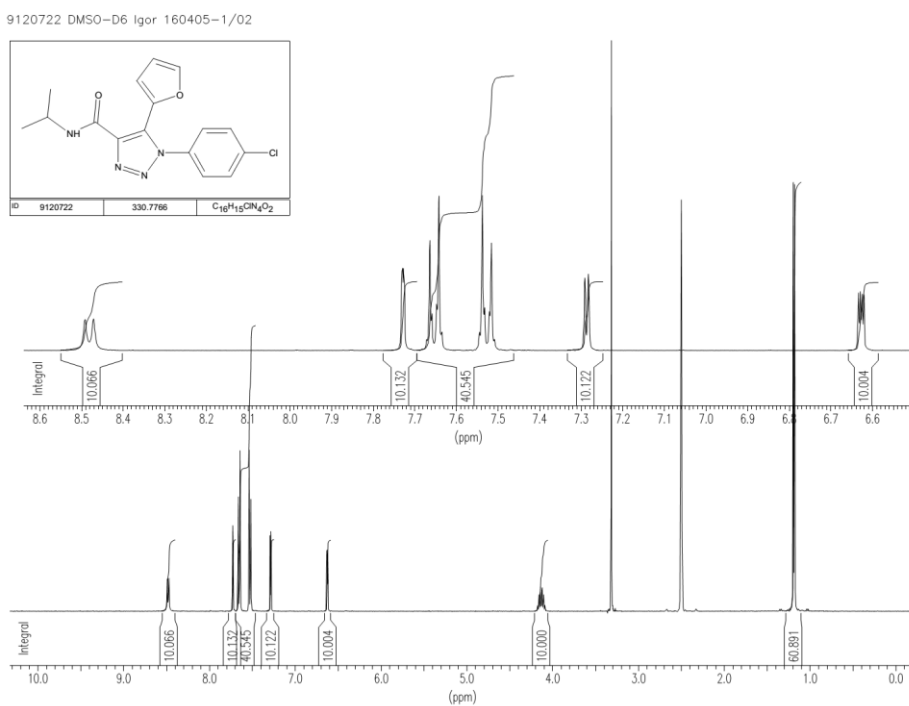


Supporting Information

2-[[1-(3-chloro-4-methylphenyl)-2,5-dioxo-3-pyrrolidinyl]thio]nicotinic acid (LabMol-43, 24);

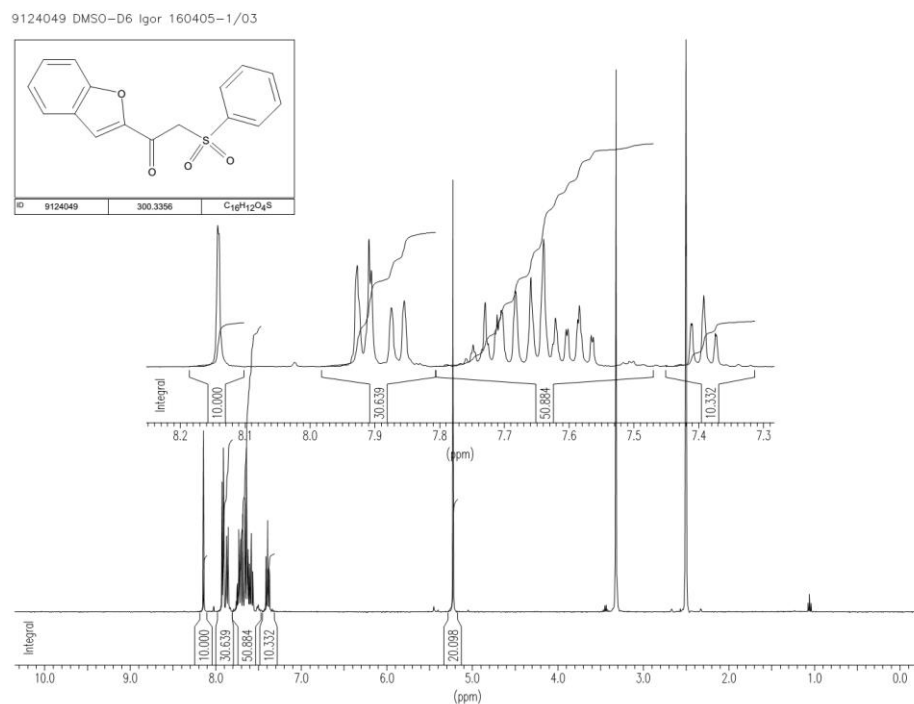


1-(4-chlorophenyl)-5-(2-furyl)-N-isopropyl-1H-1,2,3-triazole-4-carboxamide (LabMol-44, 25);

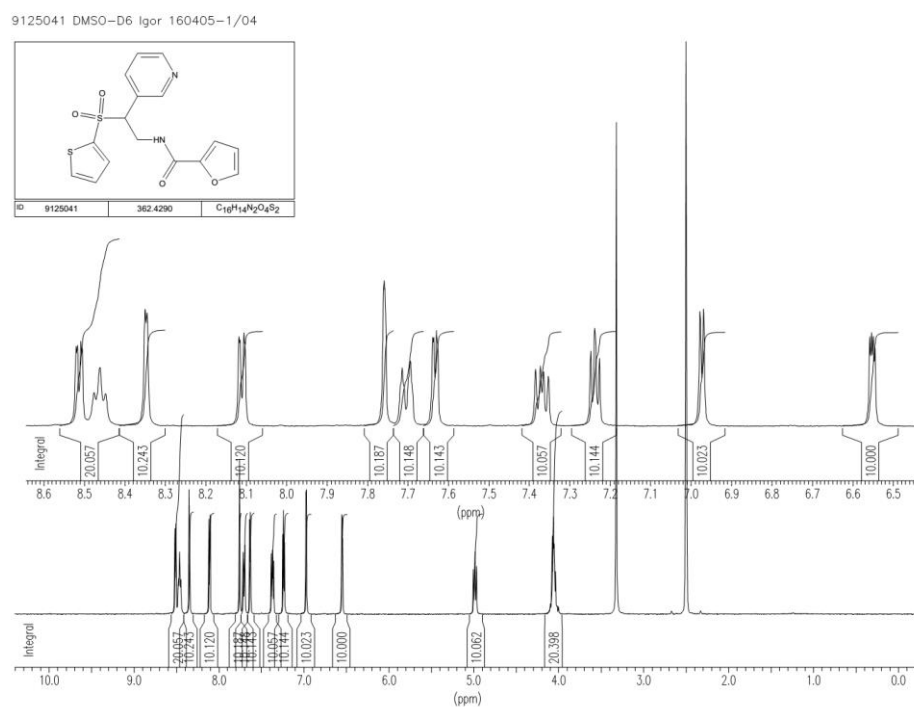


Supporting Information

1-(1-benzofuran-2-yl)-2-(phenylsulfonyl)ethanone (LabMol-45, **26**);

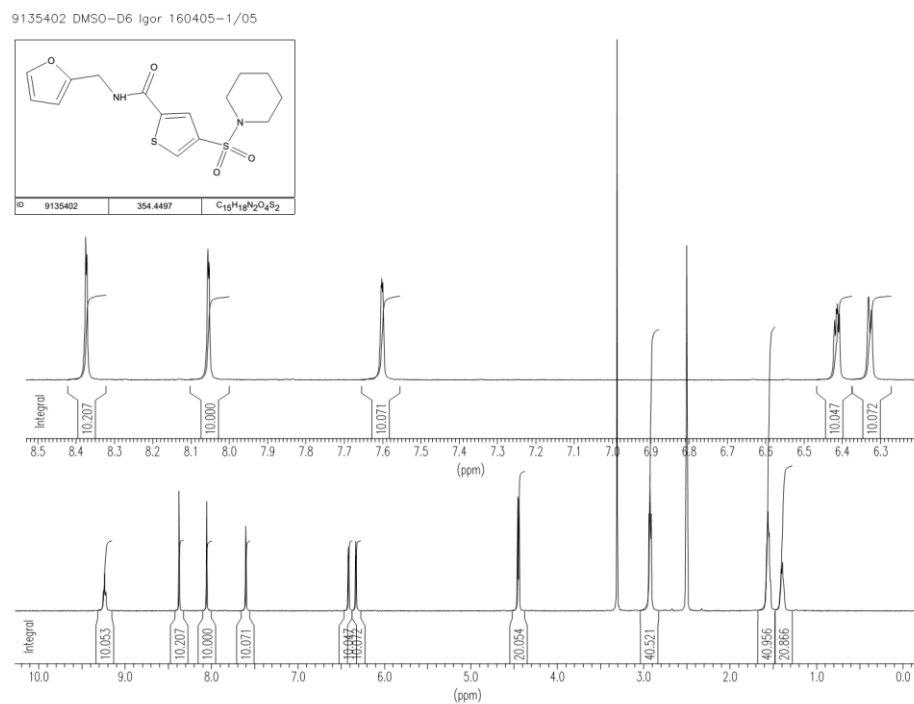


N-[2-(3-pyridinyl)-2-(2-thienylsulfonyl)ethyl]-2-furamide (LabMol-46, **27**);

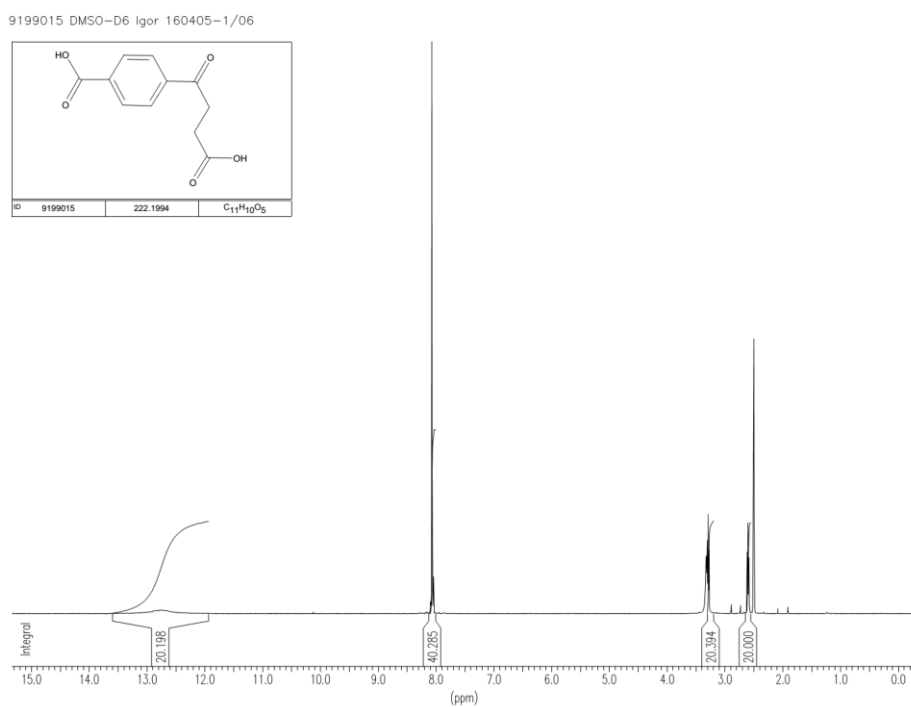


Supporting Information

N-(2-furylmethyl)-4-(1-piperidinylsulfonyl)-2-thiophenecarboxamide (LabMol-47, **28**);



4-(3-carboxypropanoyl)benzoic acid (LabMol-48, **29**);



REFERENCES

- (1) MACCS Structural Keys. 2013, Accelrys, San Diego, CA.
- (2) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107–113.
- (3) Riniker, S.; Landrum, G. a. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminform.* **2013**, *5* (1), 26.
- (4) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.
- (5) Baskin, I.; Varnek, A. Building a Chemical Space Based on Fragment Descriptors. *Comb. Chem. High Throughput Screen.* **2008**, *11* (8), 661–668.
- (6) 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.
- (7) 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.
- (8) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.

Supporting Information

- (9) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29* (5), 1189–1232.
- (10) Friedman, J. H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19* (1), 1–67.
- (11) Vapnik, V. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.
- (12) Barker, M.; Rayens, W. Partial Least Squares for Discrimination. *J. Chemom.* **2003**, *17*, 166–173.
- (13) Rosenblatt, F. *Principles of Neurodynamics; Perceptrons and the Theory of Brain Mechanisms*, 1nd ed.; Rosenblatt, F., Ed.; Spartan Books: Washington, 1962.
- (14) Altman, N. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46* (3), 175–185.