

Supporting Information for

Novel Inhibitors of Dengue Virus Methyltransferase: Discovery by in vitro-driven virtual screening on a Desktop Computer Grid.

*Michael Podvinec, Siew Pheng Lim, Tobias Schmidt, Marco Scarsi, Daying Wen, Sebastian Sonntag,
Paul Sanschagrin, Peter Shenkin and Torsten Schwede*

TABLE OF CONTENTS

Supporting Methods	S2
Aggregation prediction	S2
Test sets	S2
Decision-tree aggregation prediction.....	S2
Random Forest modeling of aggregation behavior	S3
Supporting tables	S4
Supporting Table S1: Decision-tree criteria, Seidler et al. and this work	S4
Supporting Table S2: Results of aggregator detection by decision tree method. Training and test sets are described in the publication's methods section.	S5
Supporting Table S3: Results of Random Forest-based prediction of aggregation behavior for three test sets. Training and test sets are described in the publication's methods section.....	S6
Supporting Table S4: Structure of compounds assayed in MTase inhibition assays.	S6
References.....	S11

SUPPORTING METHODS

Aggregation prediction

Test sets (A) Dengue MTase test set (DenV): The 263 compounds tested during the screening for dengue MTase inhibitors described above were classified as 237 non-aggregators and 25 aggregators, based on detergent sensitivity in the inhibition assay described below. Compounds losing their inhibitory activity in presence of 0.1% Triton-X 100 were classified as aggregators, whereas compounds either showing no inhibition or retaining their inhibitory activity in the presence of detergent were classified as non-aggregators.

(B) Medium-size test set (Med): Data on the aggregation behavior of 1030 molecules have been published¹⁻³ and the experimental results, based on dynamic light scattering and a high-throughput detergent sensitive inhibition assay against AmpC β -lactamase, have been made available online (<http://shoichetlab.ucsf.edu>). From the 1030 molecules, all compounds showing ambiguous aggregation behavior were removed, leading to a set of 653 non-aggregators and 263 aggregators.

(C) AmpC β -lactamase test set (AmpC): Recently, a set of 70'563 molecules from the National Institutes of Health Chemical Genomics Center (NCGC) library was assayed in a high-throughput screen for detergent-dependent inhibition of AmpC β -lactamase³. Out of the 70'563 molecules tested, 1204 were found to be unambiguously detergent-sensitive. From this dataset, obtained from <http://shoichetlab.ucsf.edu>, 402 non-aggregators and 82 aggregators were randomly picked as an additional test set.

Decision-tree aggregation prediction For a rapid attempt to predict aggregation behavior, we assembled a decision tree similar to that which Seidler, *et al.* derived using recursive partitioning⁴. Since we did not have access to machinery for generating the same set of descriptors, we employed descriptors that had similar meaning and used an iterative manual process to optimize the cut-off parameters for our own substitute descriptors to give the best agreement against the 111-compound

training set supplied by Seidler et al.

Supporting Table S1 compares the descriptors and parameters we used with those of Seidler et al. Supporting Table S2 shows the results we obtained on the training set of Seidler *et al.* using our descriptors with the parameters shown. With this decision tree, we achieved a prediction accuracy of 86.5%, which is not as good as the 93.4% accuracy reported by Seidler et al., but which was useful for our purposes. 43.2% of the compounds in the training set were aggregators; we predicted 42.3% aggregators, indicating that our false-positive and false-negative rates were approximately equal.

Random Forest modeling of aggregation behavior Compounds were predicted as aggregators or non-aggregators, based on calculated physicochemical descriptors, using a Random Forest (RF) model⁵. All compounds were prepared in their neutral form using the *LigPrep* protocol. For each compound, all 251 physicochemical descriptors available in MOE 2007.09 (Chemical Computing Group, Montreal, Canada) were calculated. For each test set described above, 70% of the compounds were randomly selected to train a RF model, which was then tested on the remaining 30% of the data. To determine average false positive and negative rates, 100 iterations were performed per test set, each time producing 1000 unpruned trees from subsets of the 251 descriptors. To correct for the imbalanced dataset, the majority class was down-sampled during training of the RF model. For cross-validation between test sets, a RF model was trained on all compounds of a test set and then used to predict aggregators in the other two test sets. 100 iterations were carried out for each test set as described above. All calculations were performed using R 2.5.1⁶.

SUPPORTING TABLES

Supporting Table S1: Decision-tree criteria, Seidler et al. and this work

Seidler et al. Criterion	Criterion, this work
Daylight clogp <= 3.633	QikProp clogPow < 3.1
Electrotopological S _{sssN} <= 2.287	Max Epik pKa for tertiary N < 7
Max_conj_path <= 18.5	Largest contiguous set of sp ² atoms < 19.5
Contains COOH	Contains COOH
Daylight clogP ≥ 5.389	QikProp clogPow ≥ 4.7

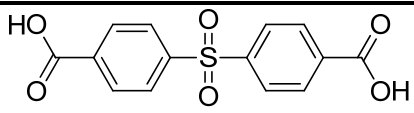
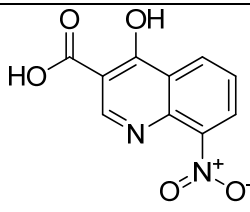
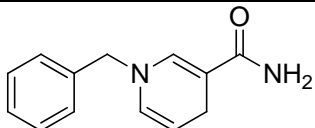
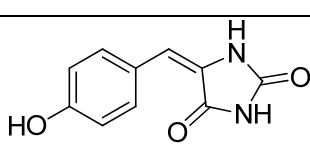
Supporting Table S2: Results of aggregator detection by decision tree method. Training and test sets are described in the publication's methods section.

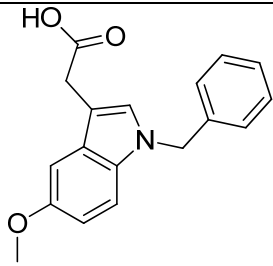
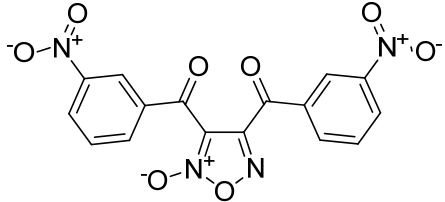
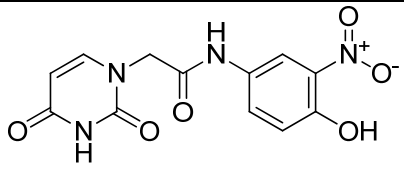
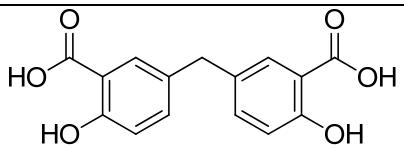
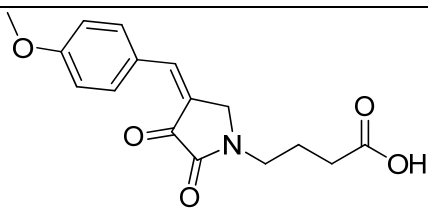
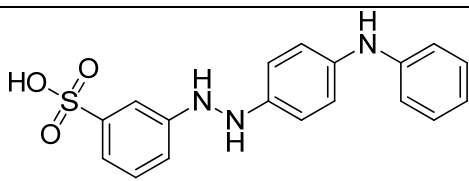
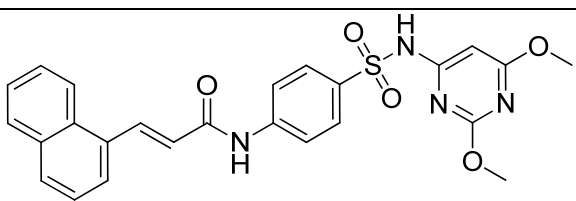
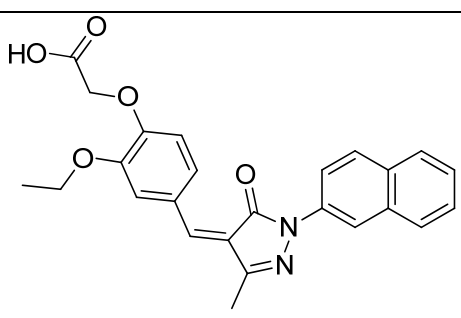
	Training	Med	AmpC	DenV
FP	0.17	0.35	0.29	0.21
FN	0.11	0.03	0.11	0.06
Sensitivity: TP/(TP+FN)	0.88	0.95	0.87	0.93
Specificity: TN/(TN+FP)	0.84	0.74	0.75	0.82

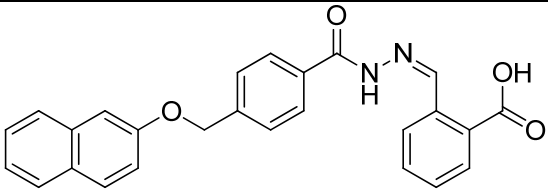
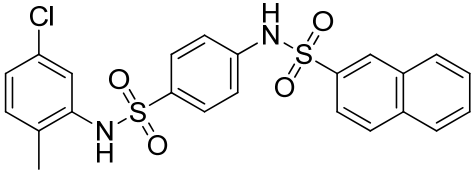
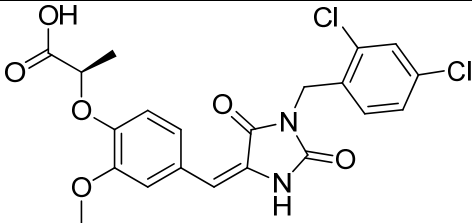
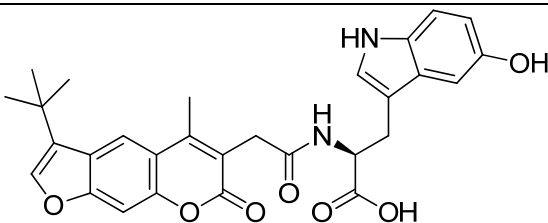
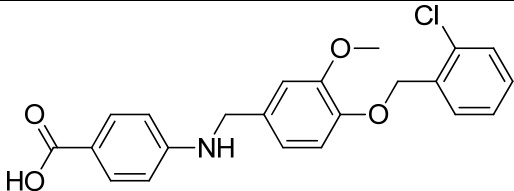
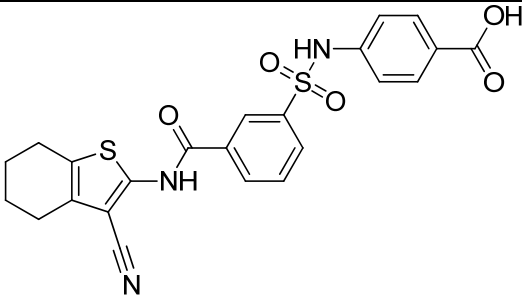
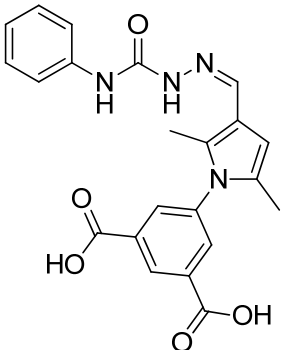
Supporting Table S3: Results of Random Forest-based prediction of aggregation behavior for three test sets. Training and test sets are described in the publication's methods section.

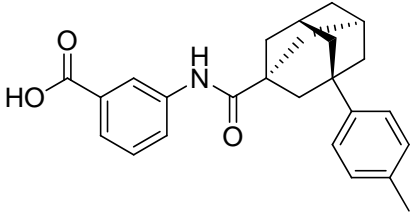
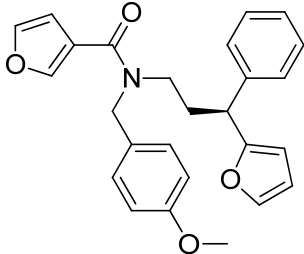
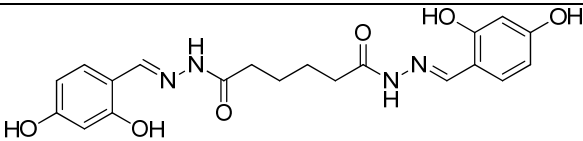
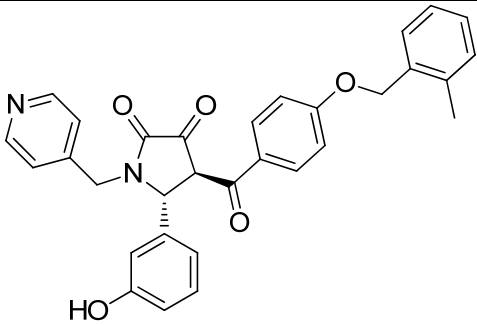
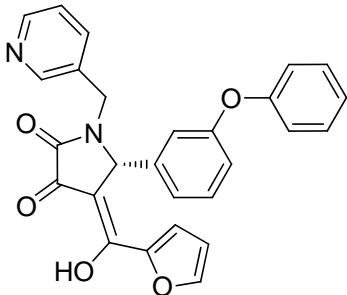
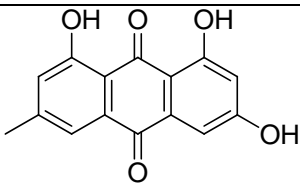
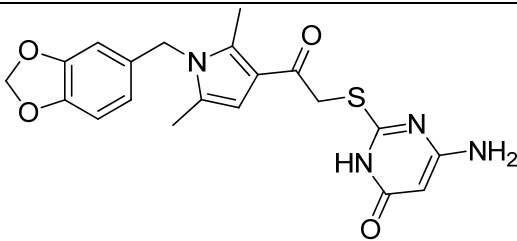
		test set						
		Med		AmpC		DenV		
		aver	stdev	aver	stdev	aver	stdev	
Training set	Med	FP	0.23	0.03	0.04	0.01	0.08	0.02
		FN	0.25	0.05	0.94	0.02	0.76	0.05
		Sensitivity: TP/(TP+FN)	0.76		0.50		0.55	
		Specificity: TN/(TN+FP)	0.77		0.56		0.76	
	AmpC	FP	0.88	0.01	0.42	0.05	0.95	0.01
		FN	0.03	0.01	0.39	0.11	0.06	0.04
		Sensitivity	0.82		0.60		0.43	
		Specificity	0.53		0.59		0.50	
	DenV	FP	0.59	0.03	0.39	0.02	0.38	0.06
		FN	0.09	0.01	0.62	0.04	0.37	0.22
		Sensitivity	0.82		0.50		0.63	
		Specificity	0.61		0.50		0.63	

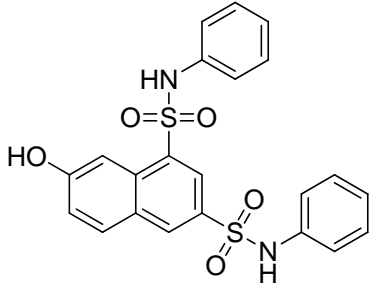
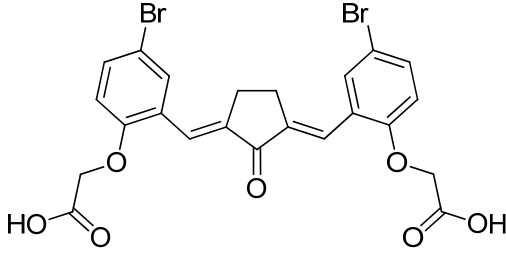
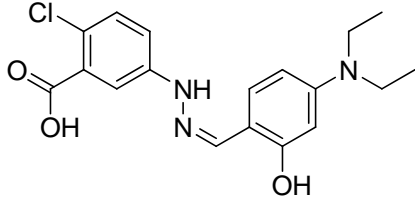
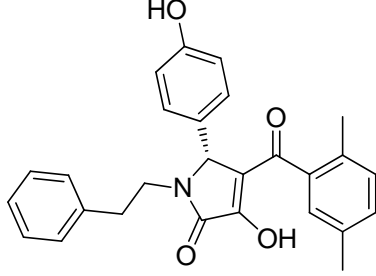
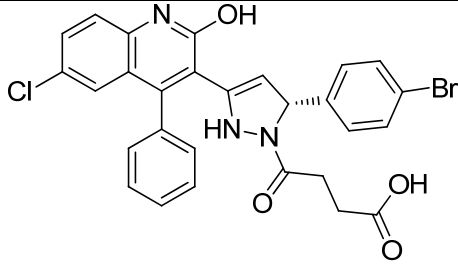
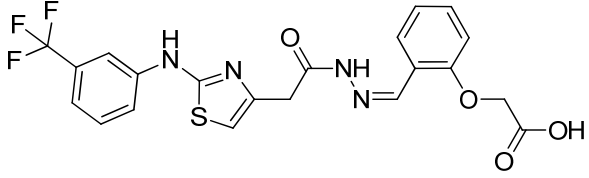
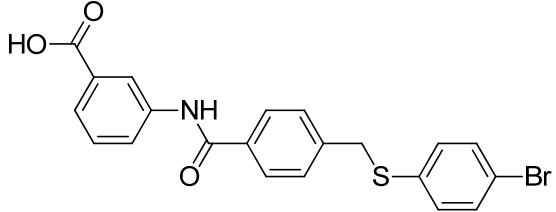
Supporting Table S4: Structure of compounds assayed in MTase inhibition assays.

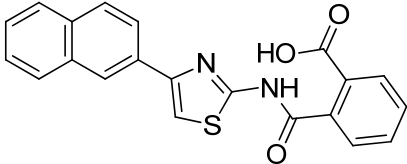
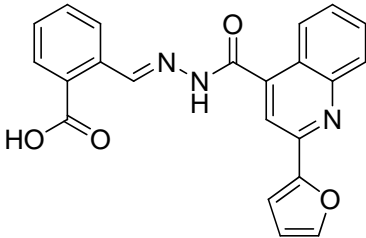
<i>Cpd</i>	<i>ID</i>	<i>Structure</i>
1	NSC12451	
2	NSC15765	
3	NSC26899	
4	NSC49419	

5	NSC54771	
6	NSC84407	
7	NSC91788	
8	NSC14778	
9	NSC140047	
10	ZINC02911543	
11	ZINC01174529	
12	ZINC03461039	

13	ZINC03287966	
14	ZINC01078518	
15	ZINC01138375	
16	ZINC02129857	
17	ZINC01112283	
18	ZINC02849675	
19	ZINC00632055	

20	ZINC01467812	
21	ZINC02826899	
22	ZINC01878835	
23	ZINC01758620	
24	ZINC00633950	
25	CACDB1751080	
26	ZINC02642996	

27	ZINC01226983	
28	ZINC02750651	
29	CACDB964942	
30	CACDB1563494	
31	ZINC01832826	
32	ZINC01078734	
33	ZINC01196449	

34	ZINC02379945	
35	ZINC03369470	

REFERENCES

1. Babaoglu, K.; Simeonov, A.; Irwin, J. J.; Nelson, M. E.; Feng, B.; Thomas, C. J.; Cancian, L.; Costi, M. P.; Maltby, D. A.; Jadhav, A.; Inglese, J.; Austin, C. P.; Shoichet, B. K., Comprehensive mechanistic analysis of hits from high-throughput and docking screens against beta-lactamase. *J Med Chem* **2008**, 51, (8), 2502-2511.
2. Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K., High-throughput assays for promiscuous inhibitors. *Nat Chem Biol* **2005**, 1, (3), 146-148.
3. Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P., A high-throughput screen for aggregation-based inhibition in a large compound library. *J Med Chem* **2007**, 50, (10), 2385-2390.
4. Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K., Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J Med Chem* **2003**, 46, (21), 4477-4486.
5. Breiman, L., Random forests. *Machine Learning* **2001**, 45, (1), 5-32.
6. Ihaka, R.; Gentleman, R., R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **1996**, 5, (3), 299-314.