Bayesian models trained with HTS data for predicting β-haematin inhibition and *in vitro* antimalarial activity

Kathryn J. Wicht[a], Jill M. Combrinck[a,b], Peter J. Smith[b] and Timothy J. Egan[a,*]

[a]Department of Chemistry, University of Cape Town, Rondebosch 7701, South Africa

[b]Division of Pharmacology, Department of Medicine, Faculty of Health Sciences, University of Cape Town, Observatory 7925, South Africa

SUPPLEMENTARY DATA

Table S1: Cross-validation result of the Bayesian models for predicting βH inhibition and whole-cell parasite activity.

Cross-Validation Result									
Model Name	**ROC Score (leave-one-out)**	**ROC Rating**	**True Positive[a]**	**False Negative[a]**	**False Positive[a]**	**True Negative[a]**	**Sensitivity[a]**	**Specificity[a]**	**Concordance[a]**
βH inhib. (cut-off = 100 µM)	0.915	Excellent	1944	169	8531	55586	0.920	0.867	0.869
Whole-cell (cut-off = 2 µM)	0.992	Excellent	804	13	1394	40334	0.984	0.967	0.967

[a] 5-fold cross-validation

Figure S1 (a) 5-Fold cross-validation ROC curve for βH inhibition activity (best Bayesian probability score cut-off = -4.920) and (b) whole-cell parasite activity (best Bayesian probability score cut-off = -8.580).

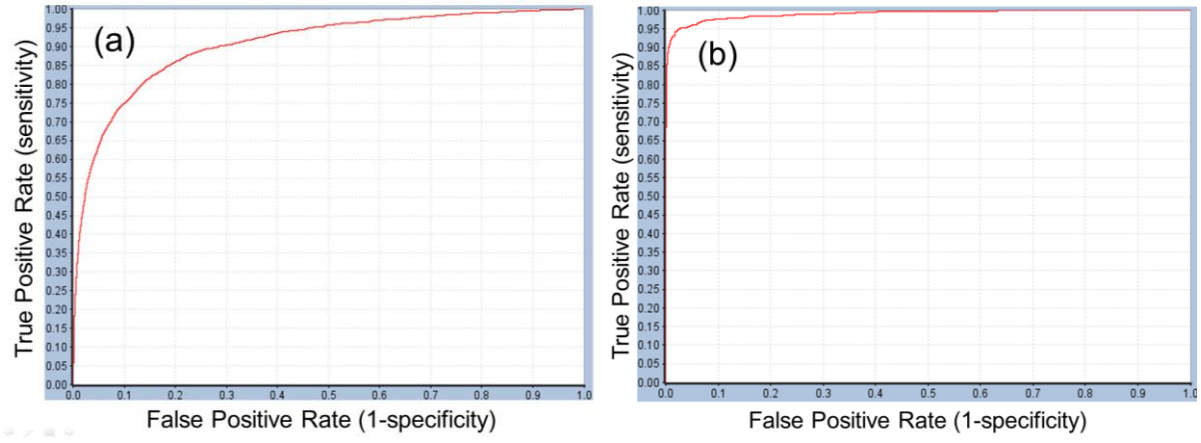
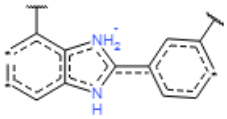
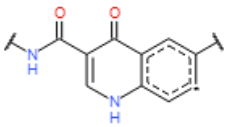
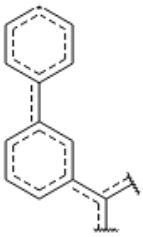
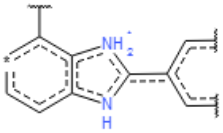
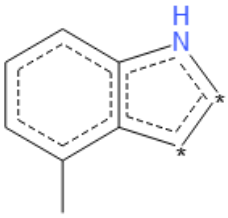
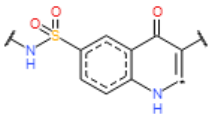
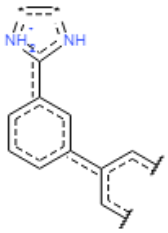
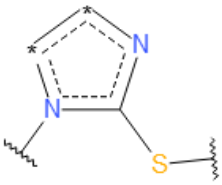
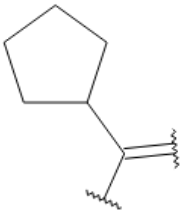
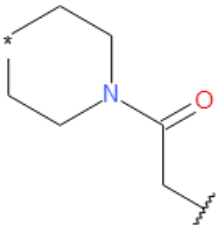
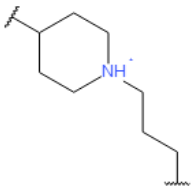
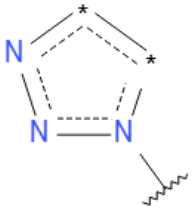
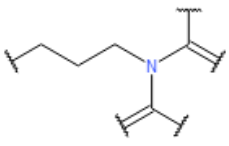
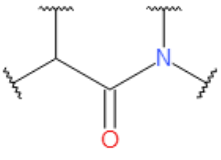
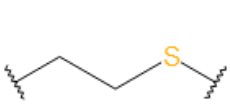
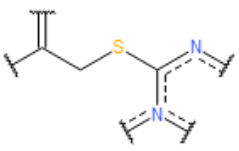
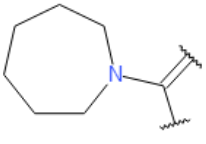


Figure S2 (a) Good and bad fingerprints for βH model (cut-off of 100 μM)

Good features from ECFP_6

			
G1: -601866137 103 out of 115 good Bayesian Score: 2.999	G3: -589198510 95 out of 121 good Bayesian Score: 2.877	G6: 1339619580 127 out of 176 good Bayesian Score: 2.850	G7: 1323548531 123 out of 170 good Bayesian Score: 2.848
			
G11: -8962239 154 out of 224 good Bayesian Score: 2.829	G13: -1821955803 80 out of 106 good Bayesian Score: 2.814	G14: -429210291 217 out of 334 good Bayesian Score: 2.808	

Bad features from ECFP_6

				
B1: 1986731747 0 out of 683 good Bayesian Score: -3.253	B2: -19155222 0 out of 495 good Bayesian Score: -2.945	B3: -1059921813 1 out of 982 good Bayesian Score: -2.911	B4: 294592992 0 out of 438 good Bayesian Score: -2.830	B5: 912011250 0 out of 431 good Bayesian Score: -2.814
				
B6: -2125181541 0 out of 418 good Bayesian Score: -2.786	B7: 1659031914 0 out of 402 good Bayesian Score: -2.749	B8: 1924960193 0 out of 395 good Bayesian Score: -2.733	B9: 802681296 0 out of 383 good Bayesian Score: -2.704	B10: 1601239950 0 out of 364 good Bayesian Score: -2.657

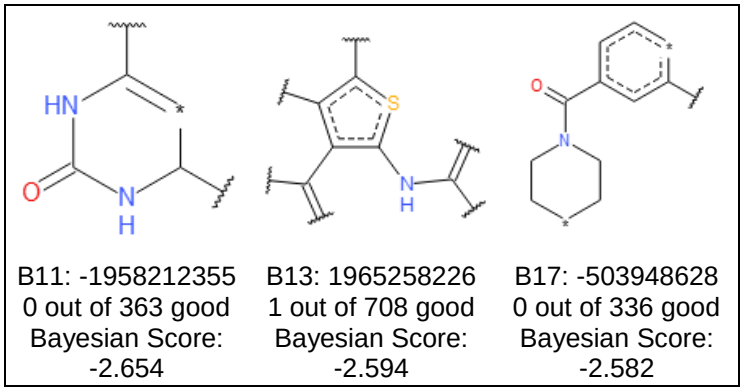
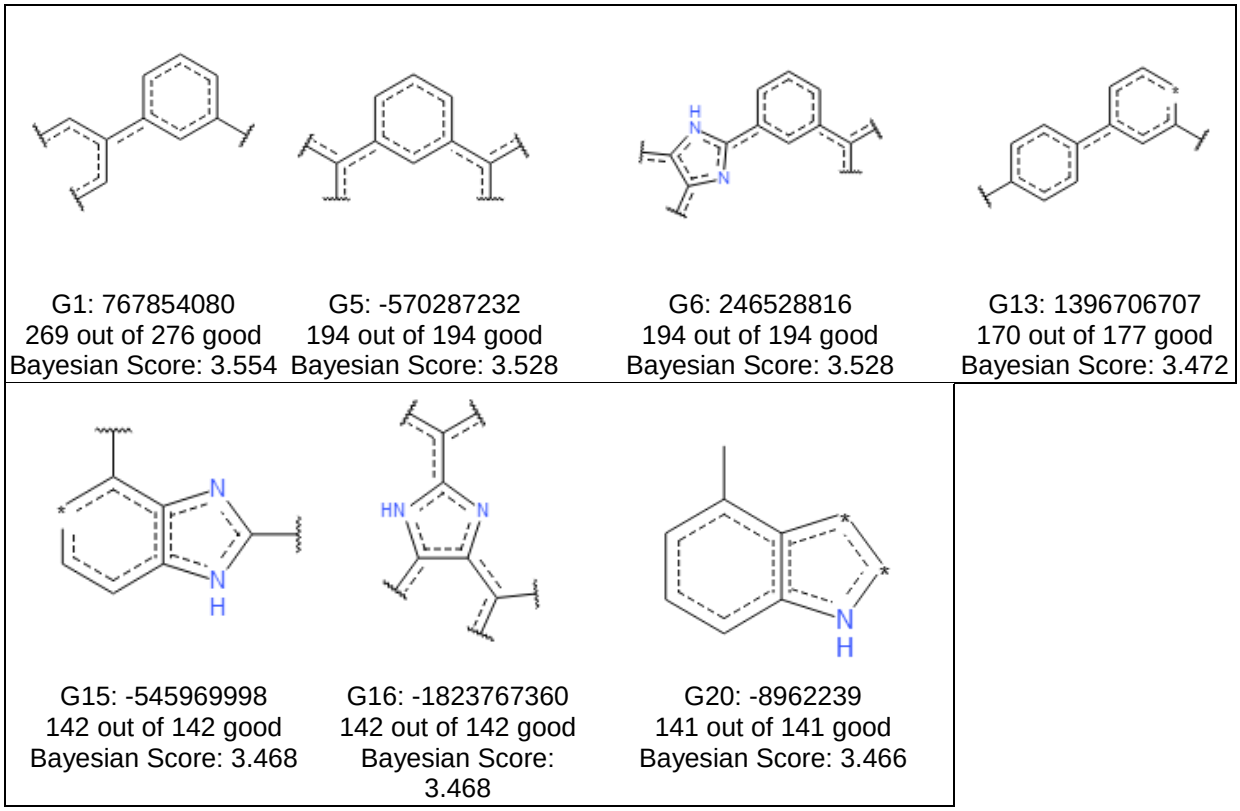


Figure S2 (b) Good and bad fingerprints for parasite activity model (cut-off of 2 µM)

Good Features from ECFP_6



Bad features from ECFP_6

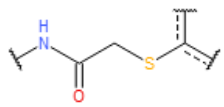
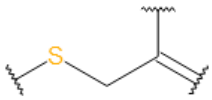
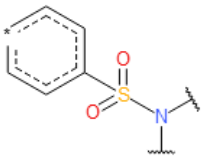
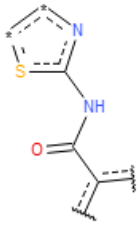
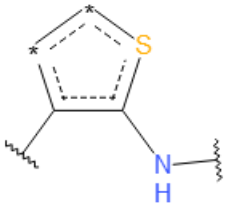
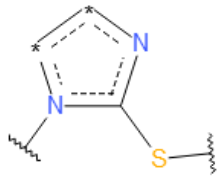
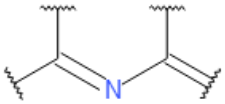
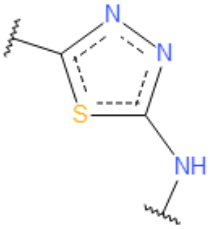
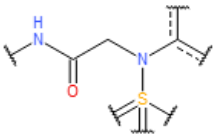
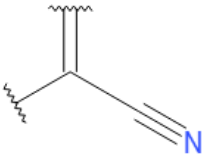
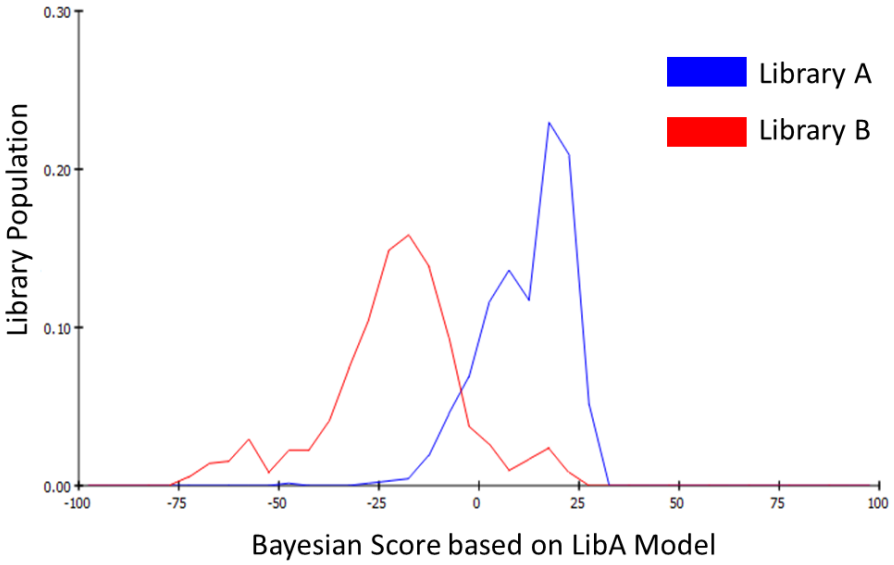
				
B1: 0 out of 3251 good Bayesian Score: -4.385	B3: 1 out of 4330 good Bayesian Score: -3.975	B4: 0 out of 2111 good Bayesian Score: -3.960	B5: 0 out of 2087 good Bayesian Score: -3.948	B7: 0 out of 1772 good Bayesian Score: -3.788
				
B8: 0 out of 1763 good Bayesian Score: -3.783	B9: 0 out of 1740 good Bayesian Score: -3.770	B11: 0 out of 1663 good Bayesian Score: -3.726	B14: 0 out of 1598 good Bayesian Score: -3.687	B17: 0 out of 1423 good Bayesian Score: -3.574

Table S2: Coefficients for PCA of TCAMS compounds. The largest coefficient (absolute value) of each specific feature is given in bold, indicating the component to which it makes the largest contribution.

(a) PCA for TCAMS compounds			
Equation Terms	**GSK_PC1**	**GSK_PC2**	**GSK_PC3**
Constant	-7.535	-2.173	-1.67
ALogP	-0.0297	**0.3191**	-0.07758
Molecular_Weight	**0.004961**	0.001587	-0.002147
Num_H_Donors	**0.3419**	-0.2197	0.1379
Num_H_Acceptors	**0.2155**	-0.1538	-0.04567
Num_RotatableBonds	0.1401	0.00571	**-0.1751**
Num_Rings	0.2899	**0.3646**	0.3234
Num_AromaticRings	0.2157	**0.2792**	0.438
Molecular_FractionalPolarSurfaceArea	1.717	**-6.678**	6.176
(b) PCA for TCAMS, VU and OU compounds			
Equation Terms	**GSK_VU_OU PC1**	**GSK_VU_OU PC2**	**GSK_VU_OU PC3**
Constant	-6.926	1.34	-2.264
ALogP	0.1477	**0.325**	-0.05444
Molecular_Weight	**0.004804**	-0.001017	-0.001884
Num_H_Donors	0.2157	**-0.2864**	0.2792
Num_H_Acceptors	0.1471	**-0.2593**	-0.06759
Num_RotatableBonds	0.146	-0.06714	**-0.1623**
Num_Rings	**0.3852**	0.1654	0.3146
Num_AromaticRings	**0.3179**	0.1867	0.4487
Molecular_FractionalPolarSurfaceArea	-1.953	**-6.19**	5.602
(c) PCA for TCAMS, VU, OU, FDA and Aldrich compounds			
Equation Terms	**PCA_PC1**	**PCA_PC2**	**PCA_PC3**
Constant	-4.299	-2.03	0.7441
ALogP	0.1367	**-0.1948**	0.02947
Molecular_Weight	**0.003753**	0.001072	0.0009481
Num_H_Donors	0.144	**0.2699**	-0.1638
Num_H_Acceptors	0.09671	**0.2059**	0.03007
Num_RotatableBonds	0.1043	0.06799	**0.2143**
Num_Rings	**0.2972**	-0.07645	-0.236
Num_AromaticRings	**0.2733**	-0.121	-0.2679
Molecular_FractionalPolarSurfaceArea	-1.398	**3.891**	-3.011

Figure S3: Population graphs of Bayesian scores for Library A (βH hits with >60% inhibition at 19 μM) and Library B (βH non-hits with <40% inhibition). Bayesian models LibA and LibB were created using samples in Library A and samples in Library B respectively. A score was computed for all samples in both libraries using (a) LibA and (b) LibB models respectively. The Bayesian distance was calculated as ***ScoreAA*** + ***ScoreBB*** - ***ScoreAB*** – ***ScoreBA***, where ***ScoreAA*** = average LibA score for the molecules in Library A, ***ScoreBB*** = average LibB score for the molecules in Library B, ***ScoreAB*** = average LibB score for the molecules in Library A, ***ScoreBA*** = average LibA score for the molecules in Library B. The larger the distance, the more dissimilar the libraries. The Bayesian distance for βH hits vs non-hits was 72.3, indicating a large difference in the data sets.

(a)



(b)

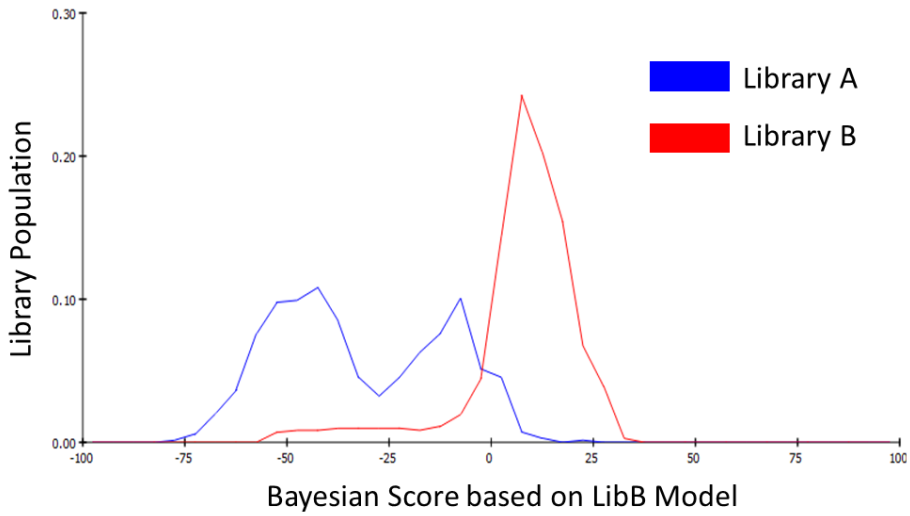
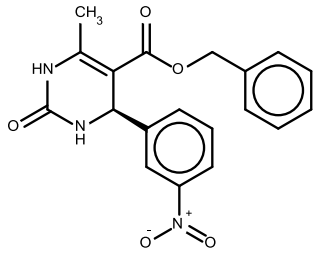
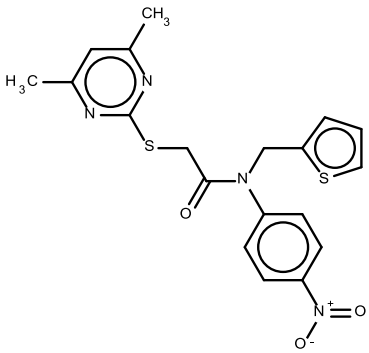
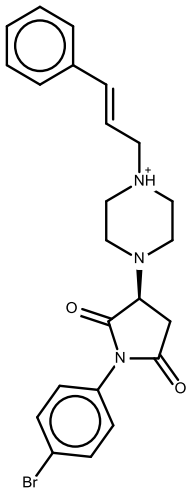
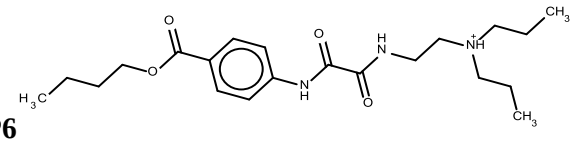
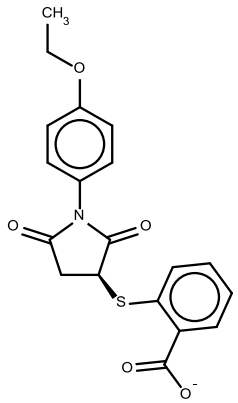
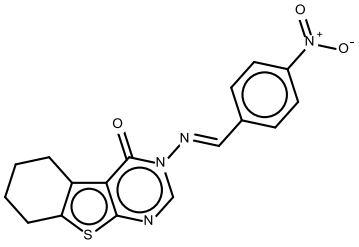
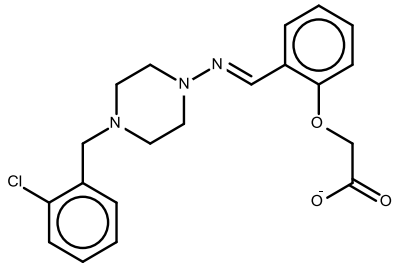
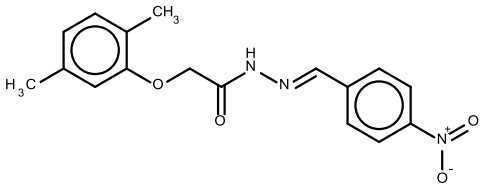
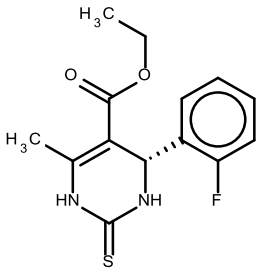
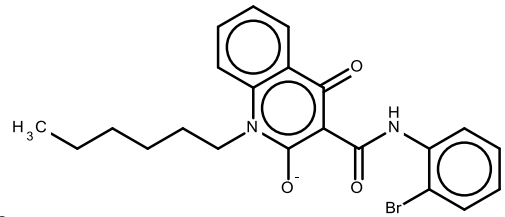


Table S3: FDA Approved compounds ordered by parasite activity Bayesian prediction score (2 μM model). The top 2.1% of compounds showing the known βH inhibiting antimalarials (bold), a list of non-βH antimalarials with their scores and the bottom 2.1%.

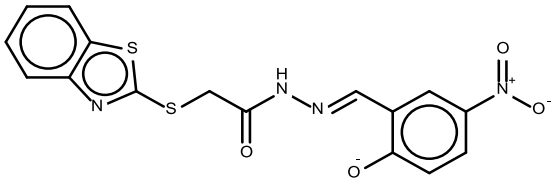
Rank #	Generic name	Parasite activity Bayesian prediction score	βH activity Bayesian prediction score	Antimalarial activity indication from PubChem
Top ranked 2.1%				
1	Lapatinib	121.907	81.4034	Active
2	**Amodiaquine**	**55.9673**	**4.97018**	**Clinical Antimalarial**
3	Imatinib	55.1797	17.2137	Active
4	Nafarelin	50.2008	-19.0656	ND
5	Nilotinib	48.8765	18.0575	ND
6	Antrafenine	43.3054	-12.1206	ND
7	Vapreotide	41.9603	16.0751	ND
8	Gefitinib	41.7064	15.5785	Active
9	**Quinine**	**40.4845**	**-11.527**	**Clinical Antimalarial**
10	**Quinidine**	**40.4845**	**-11.527**	**Clinical Antimalarial**
11	Quinidine barbiturate	39.6127	-15.6597	Active
12	Rilpivirine	38.4368	15.1667	Active
13	Thiabendazole	37.0463	15.0677	Active
14	Erlotinib	34.2159	14.7457	Inactive <10 μM
15	Gonadorelin	33.8182	-18.7343	ND
16	Goserelin	32.5423	-17.1688	Inactive <10 μM
17	Afatinib	30.9591	11.0776	ND
18	Vandetanib	30.2873	2.8095	Active
19	Lomitapide	29.8482	-15.6564	ND
20	Hydroxychloroquine	29.579	-7.90028	Active
21	**Chloroquine**	**28.9237**	**-7.38423**	**Clinical Antimalarial**
22	**Quinacrine**	**26.9148**	**0.964427**	**Clinical Antimalarial**
23	Terazosin	24.2283	-20.9691	Inactive <10 μM
24	Dasatinib	23.8589	3.44577	Active
25	Pazopanib	23.8397	13.2397	Inactive <10 μM
26	Lansoprazole	23.4241	-1.71826	Active
27	Cetrorelix	22.9426	-26.9078	ND
28	Sorafenib	22.8016	6.40403	Active
29	Regorafenib	22.1709	10.9474	Inactive <10 μM
30	Octreotide	21.8772	0.0651364	ND
31	Avanafil	21.125	-6.335	ND
32	**Halofantrine**	**20.8674**	**3.41694**	**Clinical Antimalarial**

Non-βH inhibiting antimalarials				
317	Primaquine	-2.66276	-8.36974	Clinical Antimalarial
469	Doxycycline	-5.95531	-5.4519	Clinical Antimalarial
718	Pyrimethamine	-10.3092	-10.695	Clinical Antimalarial
873	Sulfadoxine	-13.429	-3.24919	Clinical Antimalarial
964	Proguanil	-15.1309	-6.73212	Clinical Antimalarial
1172	Atovaquone	-19.699	-10.4775	Clinical Antimalarial
Bottom ranked 2.1%				
1479	Drotaverine	-38.3959	-22.9142	ND
1479	Ethopropazine	-38.6606	-27.2162	ND
1480	Cefonicid	-38.7062	-13.6207	ND
1481	Niclosamide	-39.3053	-22.1728	Active
1482	Cisatracurium Besylate	-39.4671	-30.1177	Inactive <10 μM
1483	Ceforanide	-39.6463	-8.66601	ND
1484	Nicardipine	-39.6929	-31.0469	Contradicting Data
1485	Chlorpromazine	-40.0968	-30.5136	Inactive <10 μM
1486	Furazolidone	-41.1267	-25.7797	Inactive <10 μM
1487	Acepromazine	-41.2932	-26.769	Inactive <10 μM
1488	Nilutamide	-42.0274	-14.8976	Inactive <10 μM
1489	Aceprometazine	-42.3818	-32.5787	ND
1490	Diltiazem	-42.5729	-17.633	Inactive <10 μM
1491	Sildenafil	-43.0273	-33.237	Inactive <10 μM
1492	Clonazepam	-43.6032	-24.5517	Inactive <10 μM
1493	Nitrofurantoin	-43.9518	-20.7281	Inactive <10 μM
1494	Nifedipine	-44.2094	-15.5129	Inactive <10 μM
1495	Propiomazine	-44.4947	-15.9686	ND
1496	Clobazam	-45.7832	-20.2821	Inactive <10 μM
1497	Nisoldipine	-46.6286	-34.8889	Inactive <10 μM
1498	Apixaban	-47.7752	-12.0871	Inactive <10 μM
1499	Probenecid	-48.0048	-21.7798	Inactive <10 μM
1500	Cefmetazole	-48.6828	-26.0458	Inactive <10 μM
1501	Entacapone	-49.5917	-17.3079	Inactive <10 μM
1502	Vardenafil	-50.4405	-35.2596	Inactive <10 μM
1503	Nitazoxanide	-52.1584	-6.27099	Inactive <10 μM
1504	Cefazolin	-52.5087	-22.5625	Inactive <10 μM
1505	Dantrolene	-56.3814	-18.768	Inactive <10 μM
1506	Nitrendipine	-56.4016	-25.1557	Inactive <10 μM
1507	Flunitrazepam	-56.5882	-14.6742	ND
1508	Nimodipine	-56.659	-21.5478	Inactive <10 μM
1509	Nilvadipine	-56.8842	-26.4166	Inactive <10 μM
1510	Acenocoumarol	-57.1474	-29.3544	Inactive <10 μM

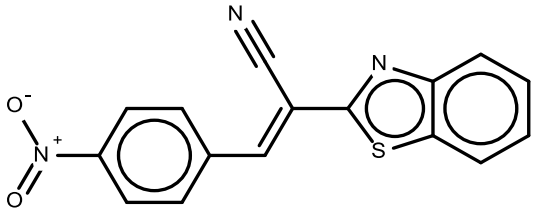
Table S4: Purchased drug-like compounds

Predicted inactive against βH formation	
 **P1**	 **P5**
 **P2**	 **P6**
 **P3**	 **P7**
 **P4**	 **P8**
	 **P9**
	 **P10**

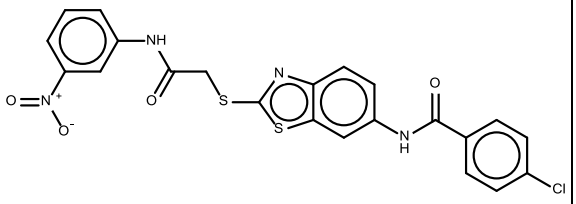
Predicted active against βH formation



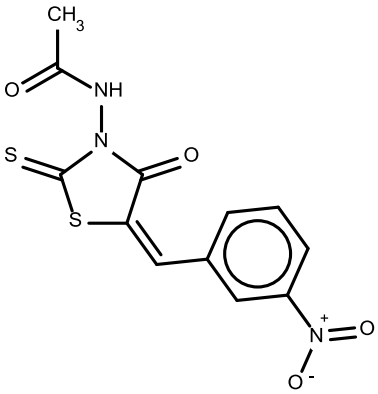
P11



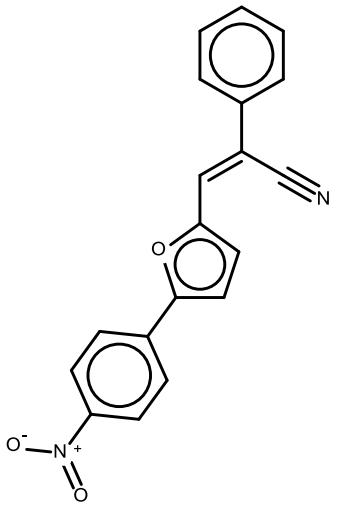
P12



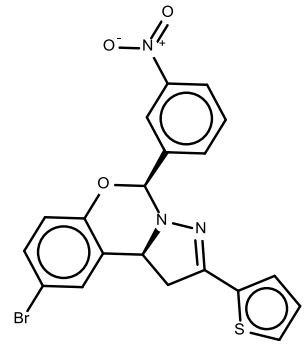
P13



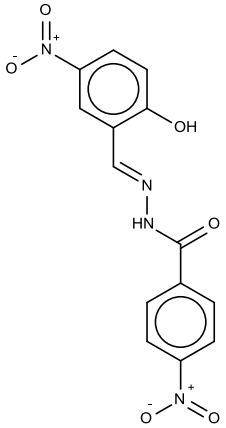
P14



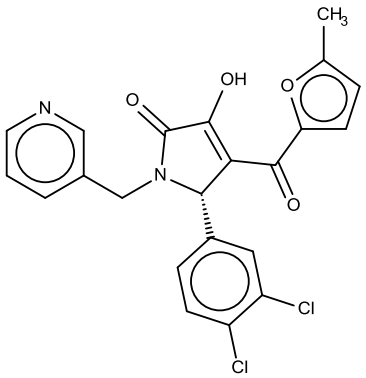
P15



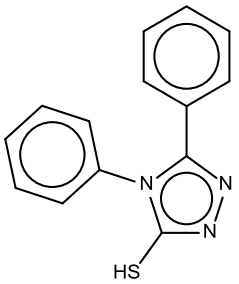
P16



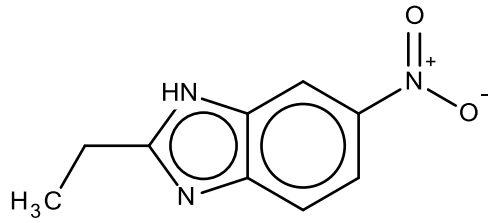
P17



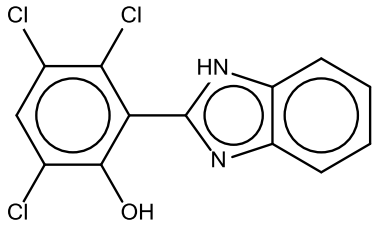
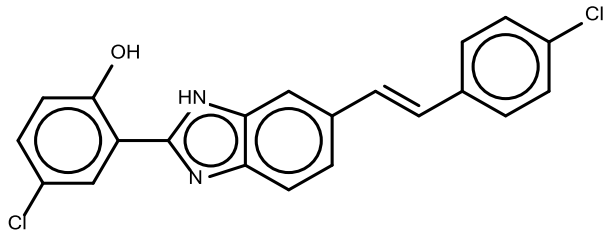
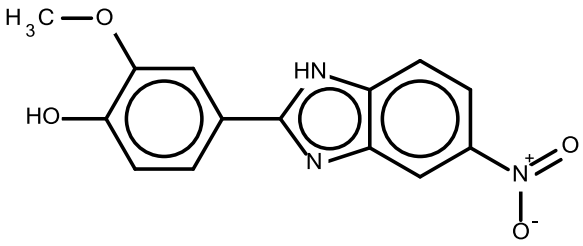
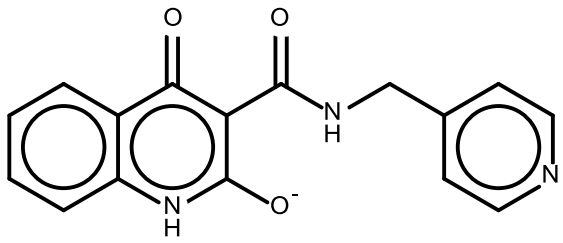
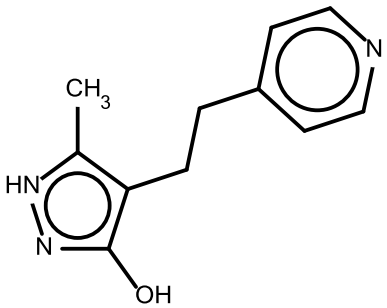
P18

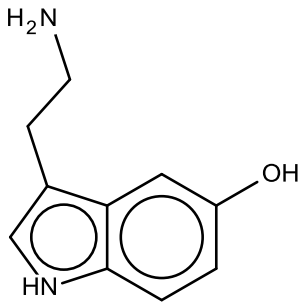
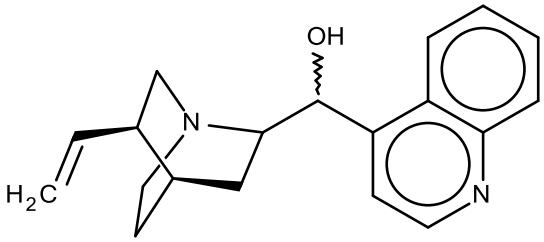
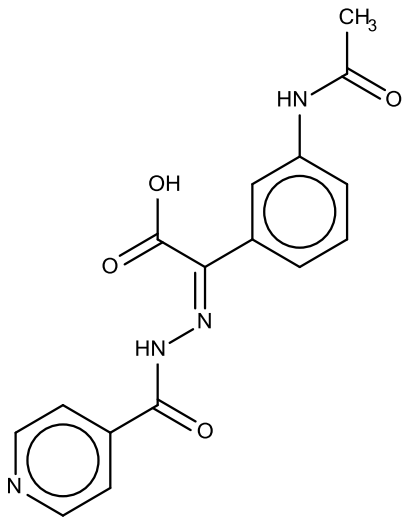
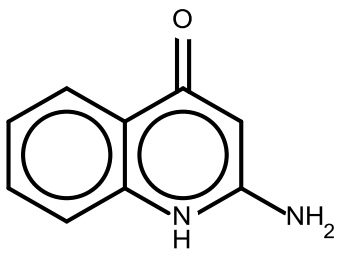


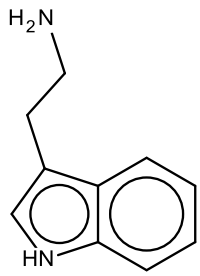
P19



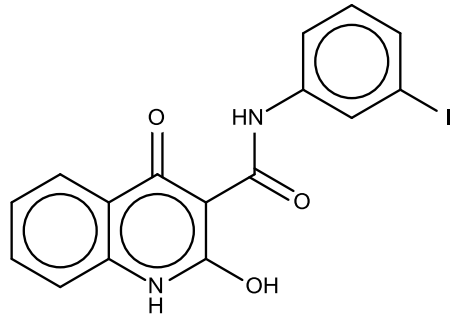
P20

Predicted bioactive (whole-cell)
 **P21**
 **P22**
 **P23**
 **P24**
 **P25**

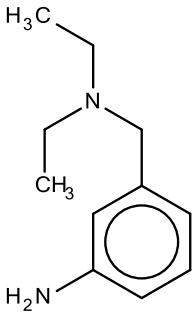
 **P26**
 **P27** (corrected structure)
 **P28**
 **P29**



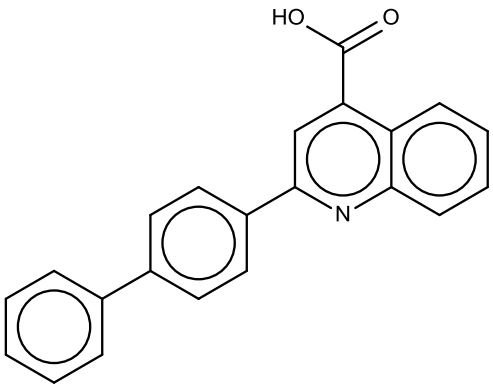
P30



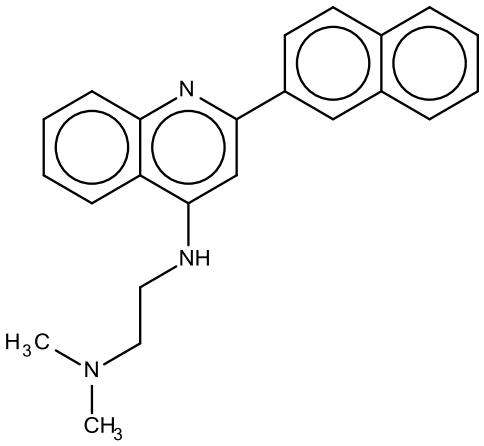
P31



P32



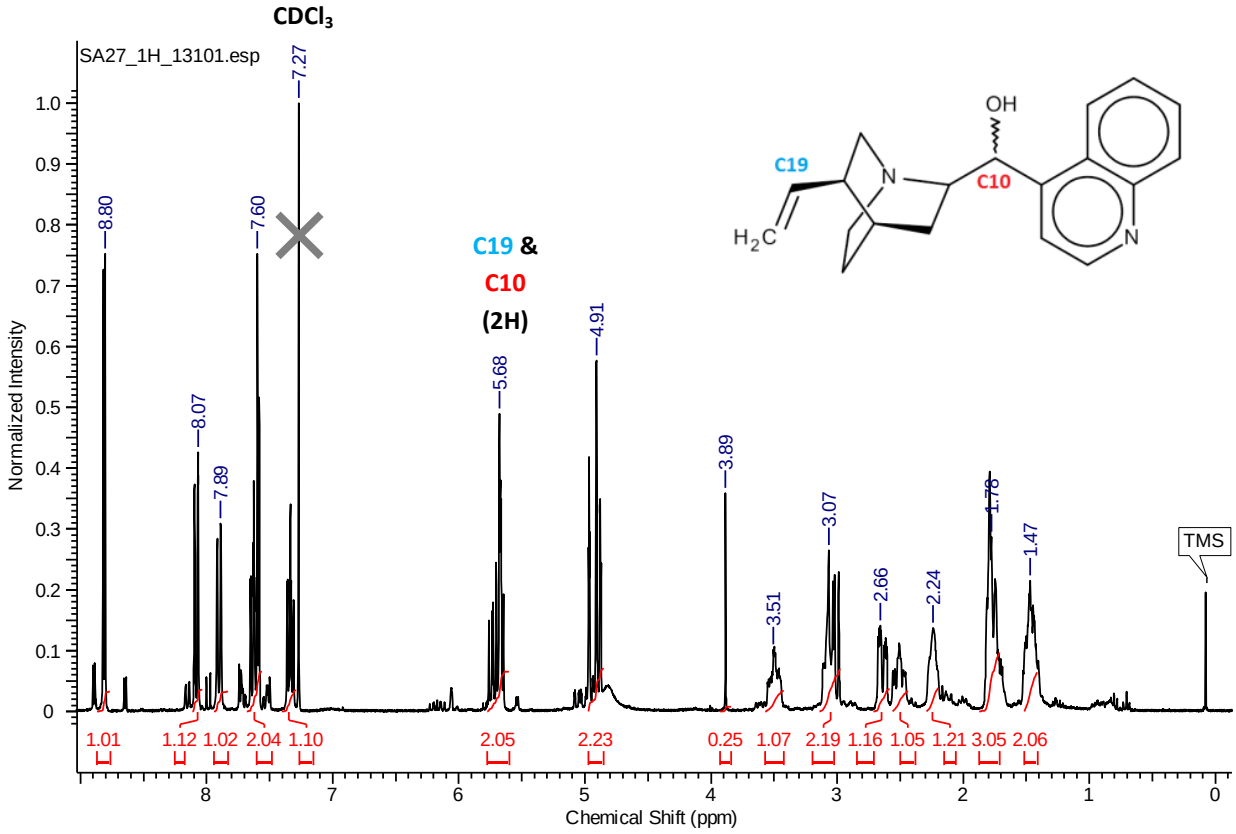
P33



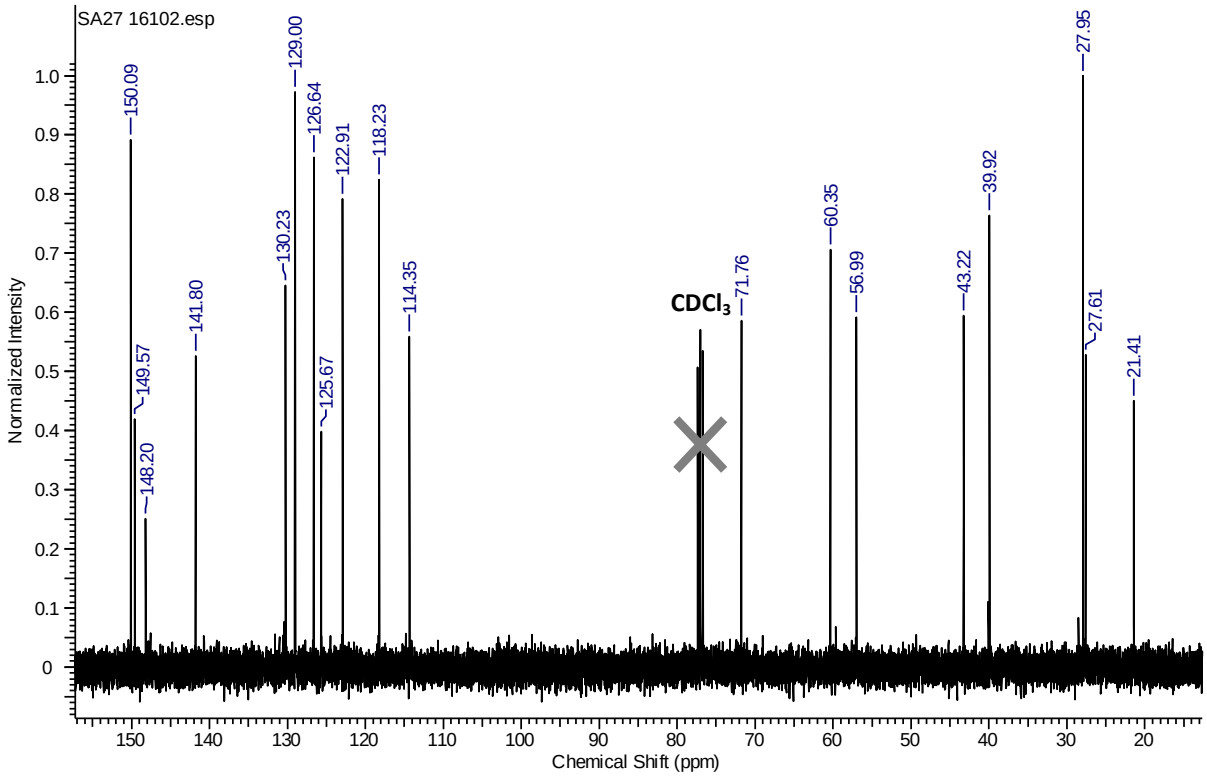
P34

Figure S4: (a) ^{1}H NMR spectrum of **P27**. The protons attached to C19 and C10 overlap in the ^{1}H NMR spectrum at 5.7 ppm integrating for 2H. (b) ^{13}C NMR spectrum of **P27** and (c) ^{1}H and ^{13}C two-dimensional heteronuclear correlation HSQC spectrum. The proton peak at 5.7 ppm couples to carbon C19 at 142 ppm and C10 at 71 ppm. The downfield chemical shift of C10 relative to that expected for an aliphatic carbon nucleus is a result of the electron withdrawing effects of the hydroxyl substituent. NMR spectra were collected using a Bruker Ultrashield 400 Plus spectrometer (at 399.95 MHz for ^{1}H and 100.64 MHz for ^{13}C in deuterated chloroform ($CDCl_3$). (d) Mass spectrum of **P27** from a JEOL GC mate II single magnetic mass spectrometer (theoretical mass for CD/CN = 294.43 g/mol).

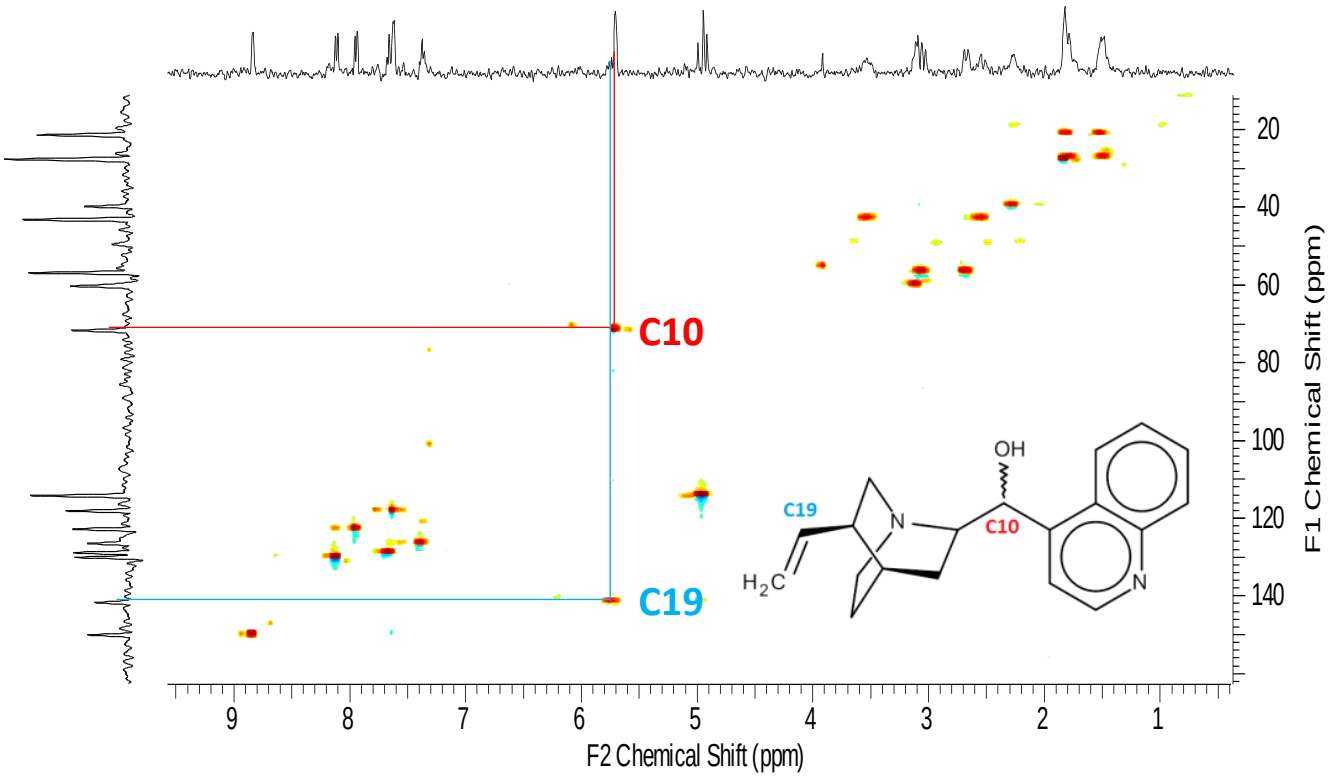
(a)



(b)


SA27 16102.esp
Normalized Intensity
Chemical Shift (ppm)
150.09
149.57
148.20
141.80
130.23
129.00
126.64
125.67
122.91
118.23
114.35
CDCl3
71.76
60.35
56.99
43.22
39.92
27.95
27.61
21.41

(c)


C10
C19
F2 Chemical Shift (ppm)
F1 Chemical Shift (ppm)
OH
N
C19
C10
H2C

(d)

File: 14J15-SA 27 **Date Run: 10-15-2014 (Time Run: 14:27:02)**
Sample:
Instrument: JEOL GCmateII
Inlet: Direct Probe **Ionization mode: EI+**

Scan: 62 **R.T.: 1.23**
Base: m/z 136; 41.1%FS TIC: 5060012 **#Ions: 218**