

Supplementary Information

GASP: A pan-specific predictor of family 1 glycosyltransferase acceptor specificity enabled by a pipeline for substrate feature generation and large-scale experimental screening

David Harding-Larsen^{1,§}, Christian Degnbol Madsen^{1,2,§}, David Teze¹, Tiia Kittilä¹, Mads Rosander Langhorn¹, Hani Gharabli¹, Mandy Hobusch¹, Felipe Mejia Otalvaro¹, Onur Kirtel¹, Gonzalo Nahuel Bidart¹, Stanislav Mazurenko^{3,4}, Evelyn Travník¹, Ditte Hededam Welner^{1,*}

¹ Technical University of Denmark, DTU Biosustain, Kemitorvet 220, Lyngby, DK 2800

² The University of Melbourne Faculty of Science, Melbourne Integrative Genomics Building 184, Royal Parade, University of Melbourne, Parkville 3010, Melbourne, VIC, AU 3052

³ Masarykova Univerzita, Department of Experimental Biology and RECETOX, Faculty of Science, Kamenice 5/A4, Brno, CZ 625 00

⁴ St. Anne's University Hospital Brno, International Clinical Research Center, Pekarska 53, Brno, CZ 656 91

[§] Contributed equally to this work

*Corresponding author: diwel@biosustain.dtu.dk

Supporting Figures

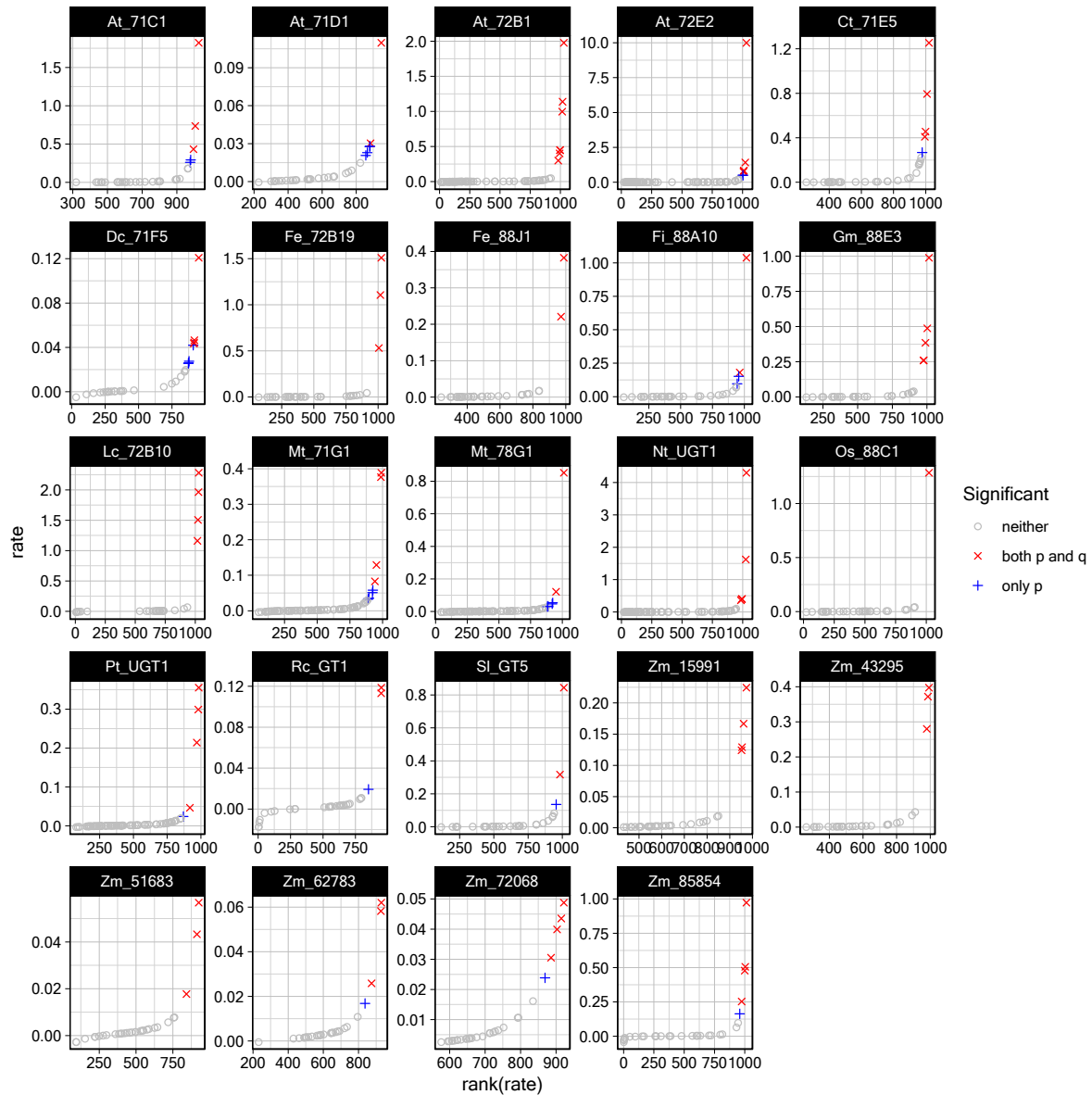


Figure S1. Reactivity classification of rate measurements. Measured rate plotted versus ranking among measurements from all enzymes. Standard significant threshold 0.05. Grey circles are measurements that have both p -value (here denoted p) and adjusted p -value (here denoted q) above the threshold, labelled as non-reactive. Red crosses are measurements with both p -value and adjusted p -value below the threshold, labelled as reactive. Blue crosses are measurements with a p -value below the threshold, but an adjusted p -value above the threshold, thereby discarded due to being regarded as inconclusive.

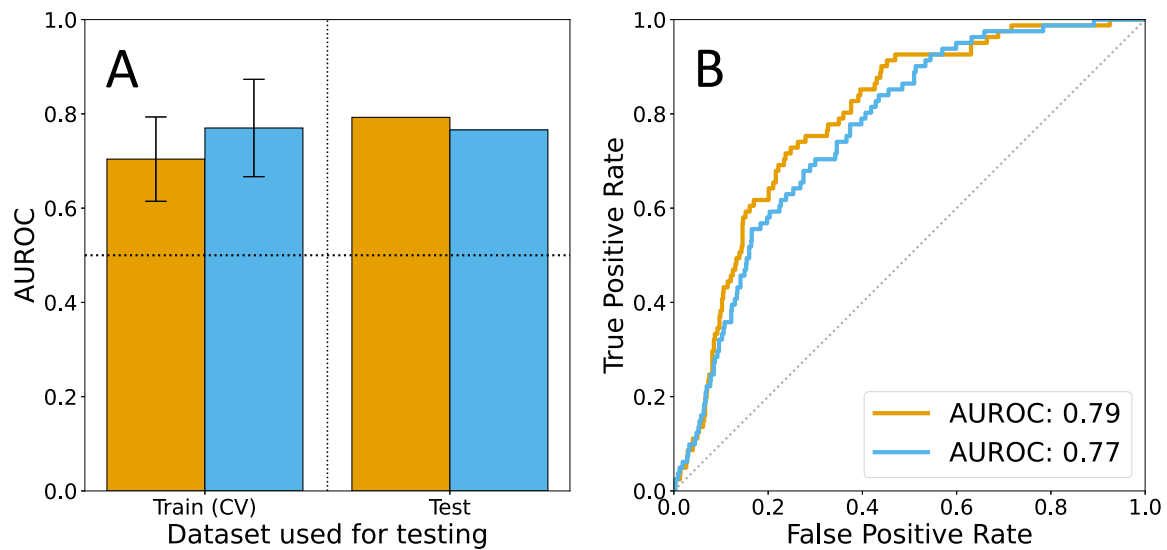


Figure S2. **A** Test results of the model based on manual hyperparameter selection (yellow) and the model with the highest cross-validation score from hyperparameter grid search (blue) using either training data (left, cross-validated) or in-house test data (right) for testing. Standard deviation is represented as error bars. The dotted line corresponds to AUROC for random predictor. **B** ROC curve for the two models tested on in-house test data with corresponding AUROC. The grey dotted line corresponds to the random predictor.

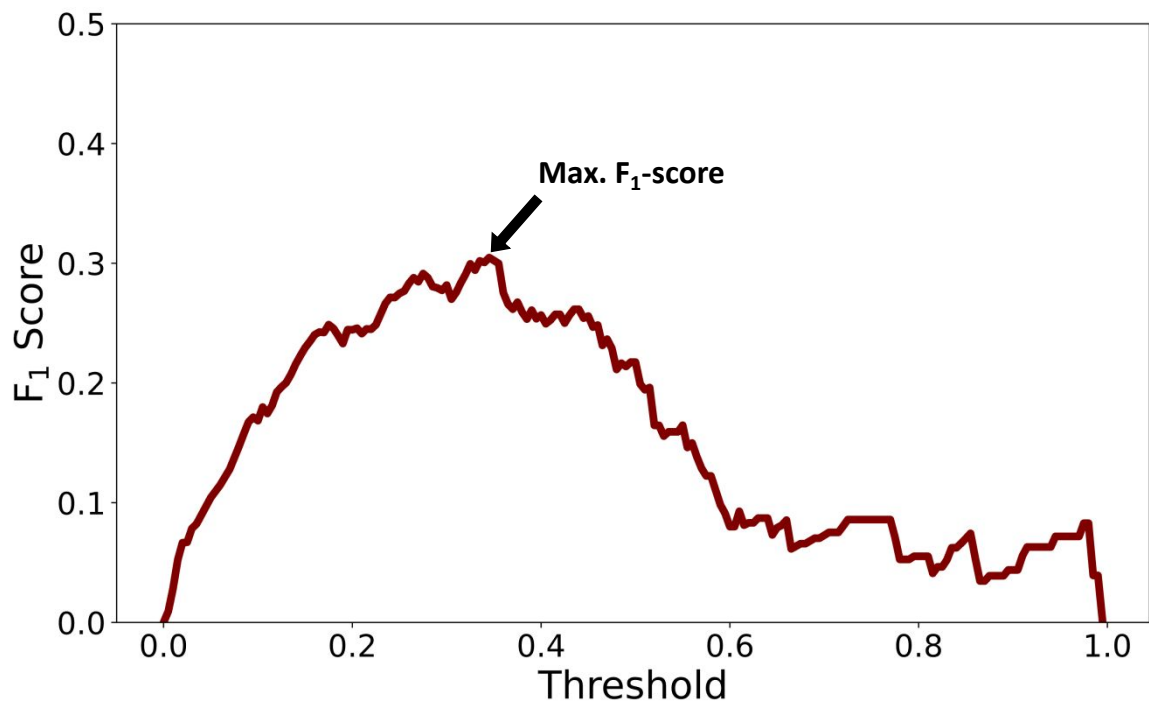


Figure S3. The optimal threshold (marked) was found by plotting the F₁-score as a function of the threshold.

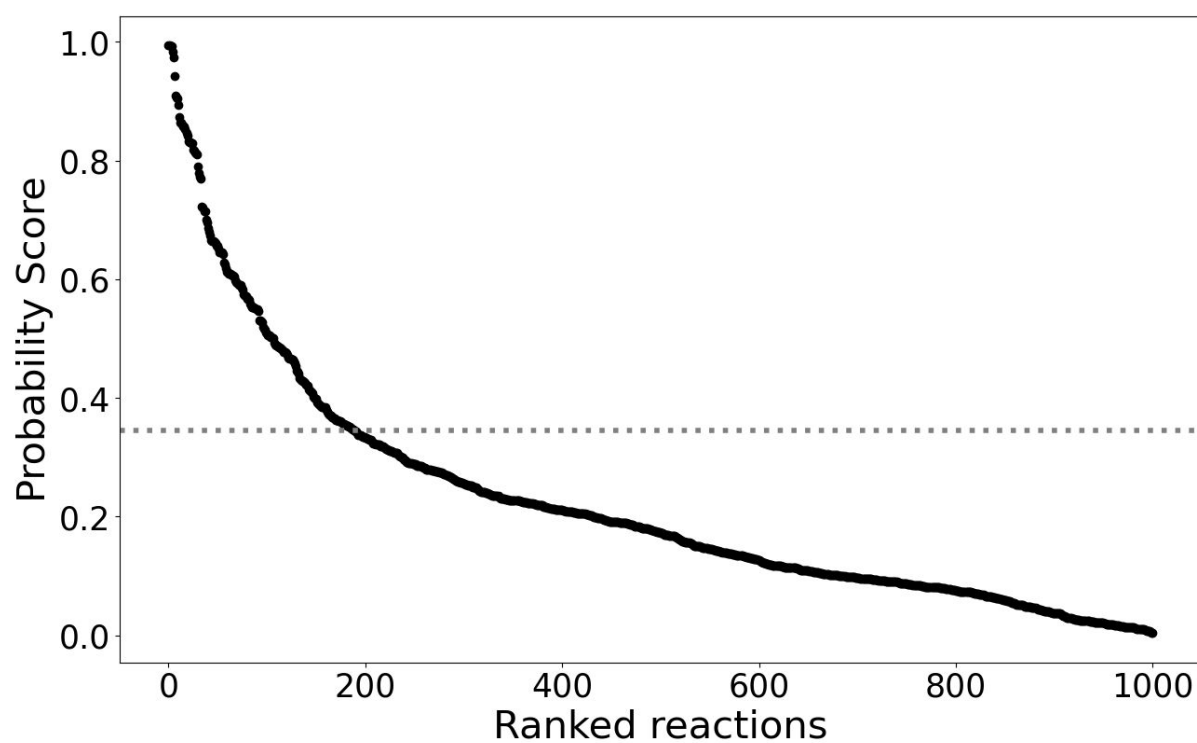


Figure S4. The predicted probability scores of all 1001 reaction data points in the in-house dataset using GASP. Grey dotted line denotes the threshold for maximum F_1 of 0.345.

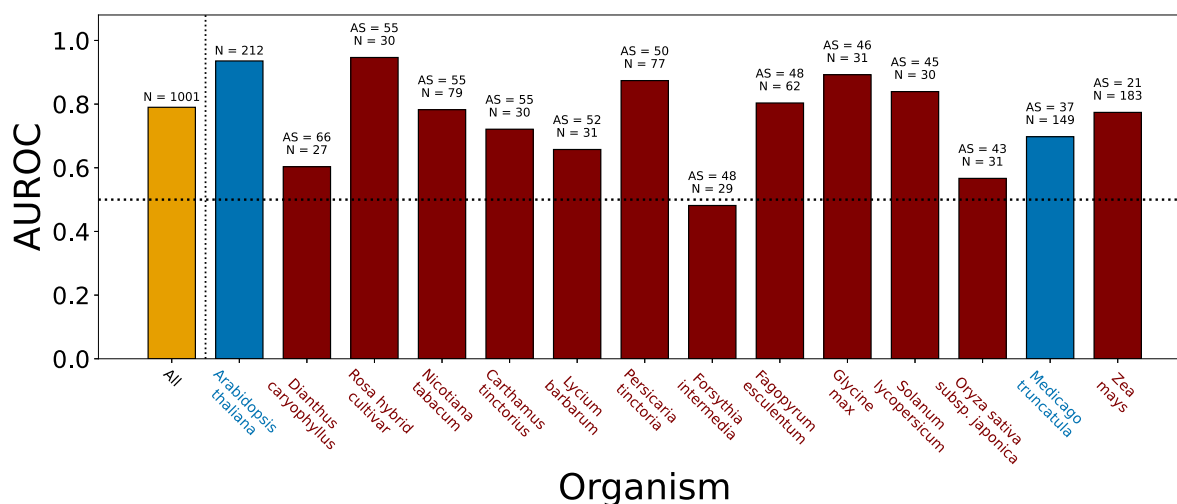


Figure S5. Performance of GASP using all test data or subsets of the test data based on organism of origin for the enzymes. Organisms are ranked on the graph from most to least similar to *Arabidopsis thaliana* based on the pairwise alignment scores (AS in figure) obtained with CLUSTALW by aligning the “internal transcribed spacer (ITS1) – 5.8S rRNA – internal transcribed spacer (ITS2)” gene region. N denotes the number of data points in the corresponding test data subset. Organisms for which enzymes are present in the training data are blue (N_{train} is 3607 for *Arabidopsis thaliana*, 78 for *Medicago truncatula*), while those not present in the training data are red. The dotted line corresponds to AUROC for the random predictor.

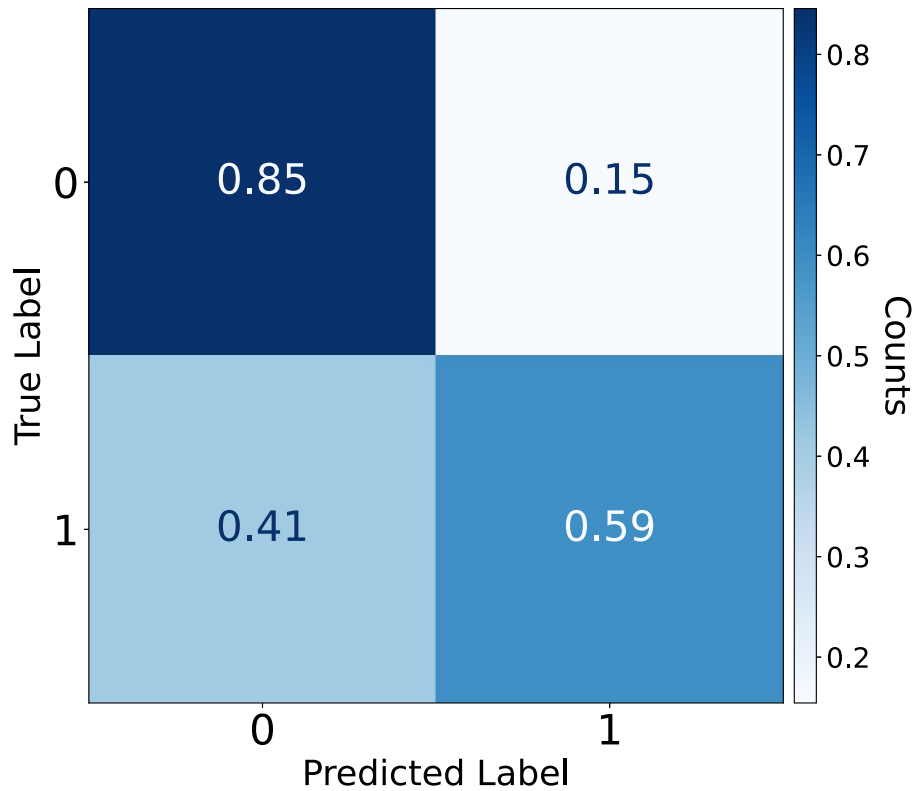


Figure S6. Confusion Matrix of the GASP model using a probability threshold of 0.345 and normalized by the true label count in each class.

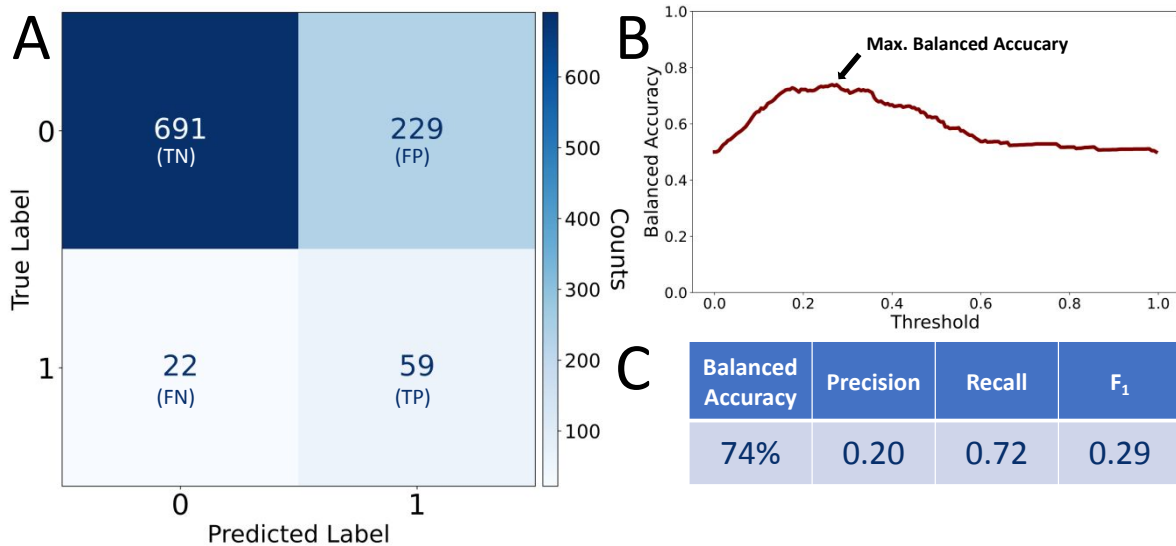


Figure S7. A Confusion matrix for GASP on in-house test data calculated using a threshold of 0.265, resulting in the maximum balanced accuracy. **B** Balanced accuracy plotted as function of threshold with maximum marked. **C** Metrics calculated using threshold of 0.265.

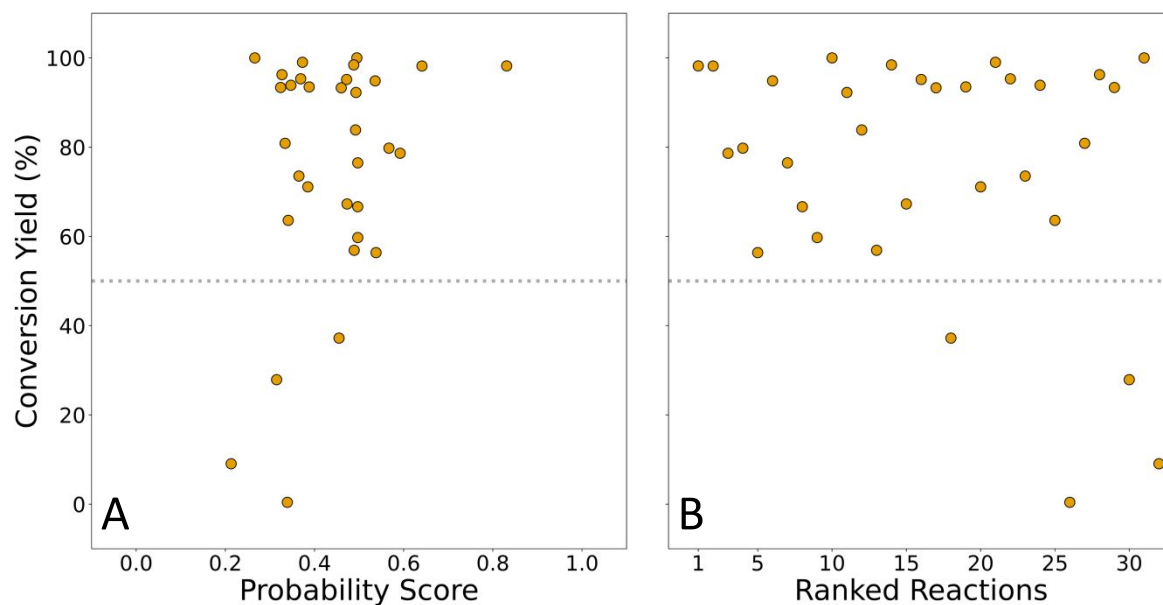


Figure S8. Correlation between substrate conversion yields and probability scores (A) or ranked reactions (B) of *GmUGT88E3*:polyphenol pairs. Grey dotted lines denote 50% conversion yield.

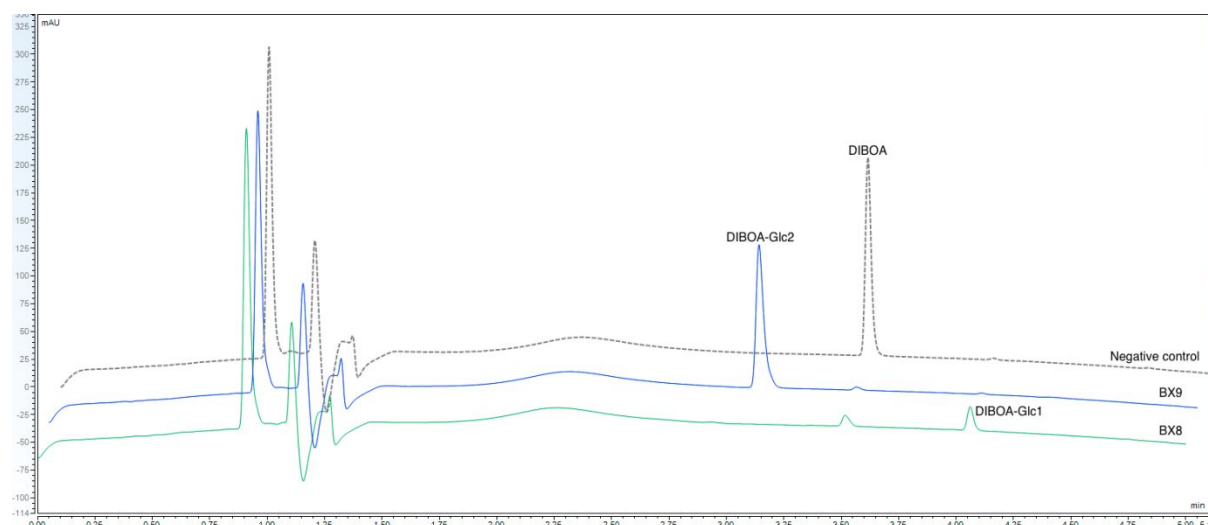


Figure S9. HPLC spectra of BX8 and BX9 (positive control) reactions with DIBOA. 100 μ g/mL of each GT1 was incubated with 2 mM of UDP-Glc and 0.5 mM of DIBOA in 100 mM of phosphate-citrate buffer (pH 7.0) at 30°C and 300 rpm for 1 h prior to HPLC analysis. DIBOA-Glc1 and DIBOA-Glc2 indicate two different DIBOA glucosides. Negative control indicates a reaction with no enzyme. Absorbances were recorded at 240 nm wavelength.

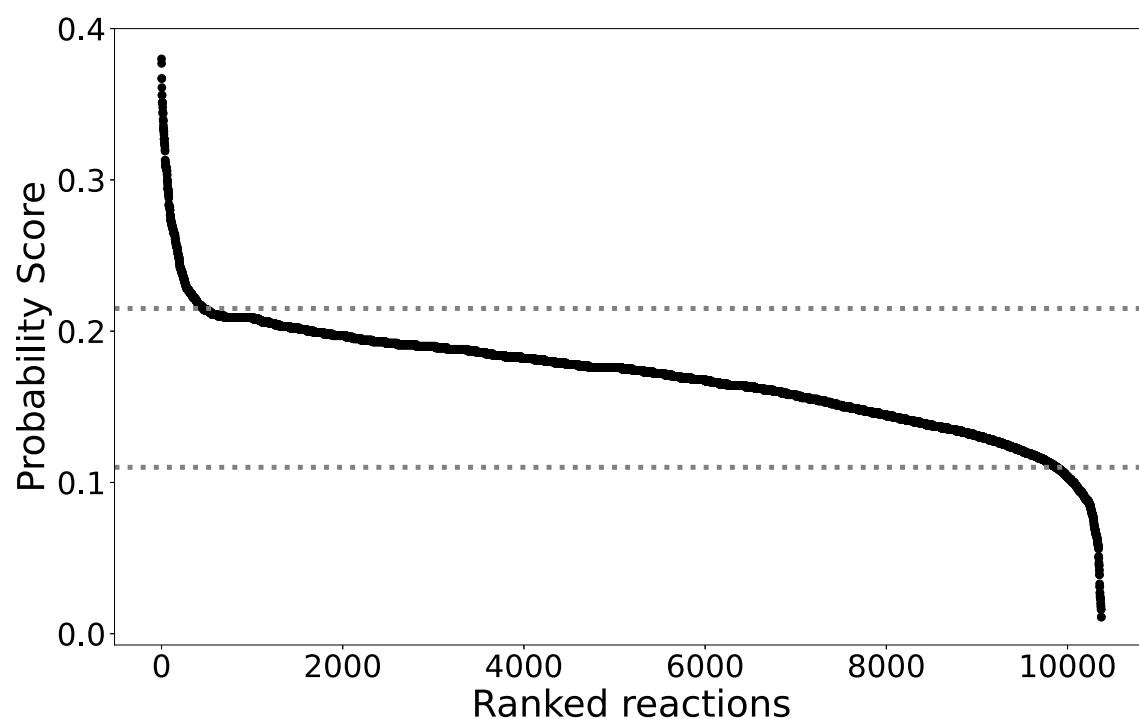


Figure S10. The probability scores of reactions with DIBOA and a filtered database of 10,374 GT1 sequences using GASP. The majority of the reactions had a prediction probability between 0.11 and 0.215 (marked by the grey dotted lines), while 456 reactions were above this range and 456 reactions were below.

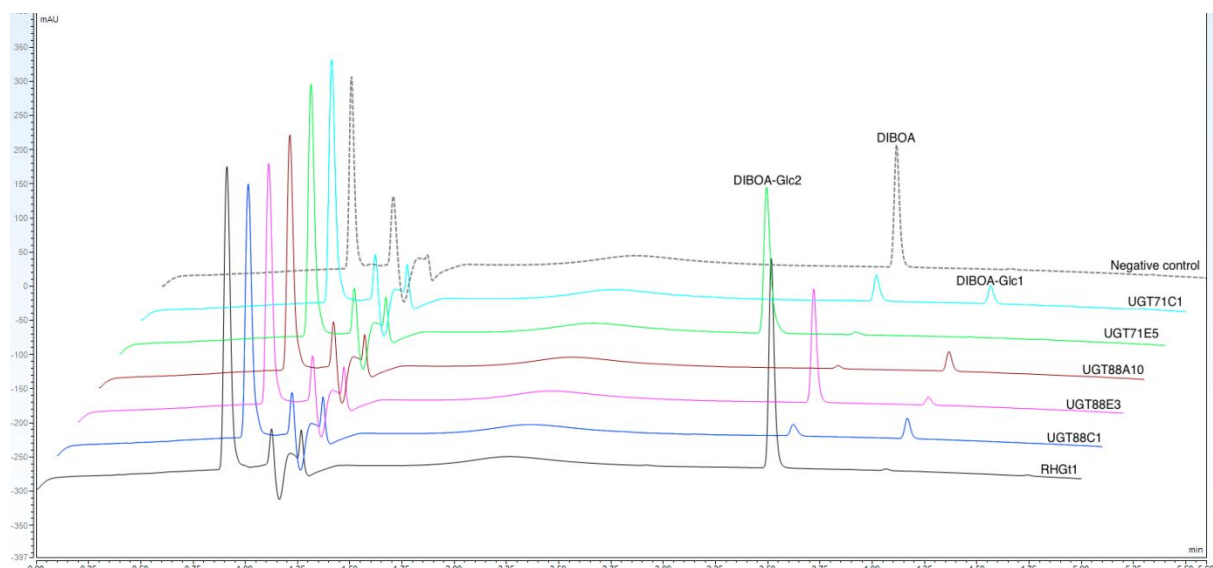


Figure S11. HPLC spectra of expert-selected GT1 reactions with DIBOA. 100 $\mu\text{g/mL}$ of each GT1 was incubated with 2 mM of UDP-Glc and 0.5 mM of DIBOA in 100 mM of phosphate-citrate buffer (pH 7.0) at 30°C and 300 rpm for 1 h prior to HPLC analysis. DIBOA-Glc1 and DIBOA-Glc2 indicate two different DIBOA glucosides. Negative control indicates a reaction with no enzyme. Absorbances were recorded at 240 nm wavelength.

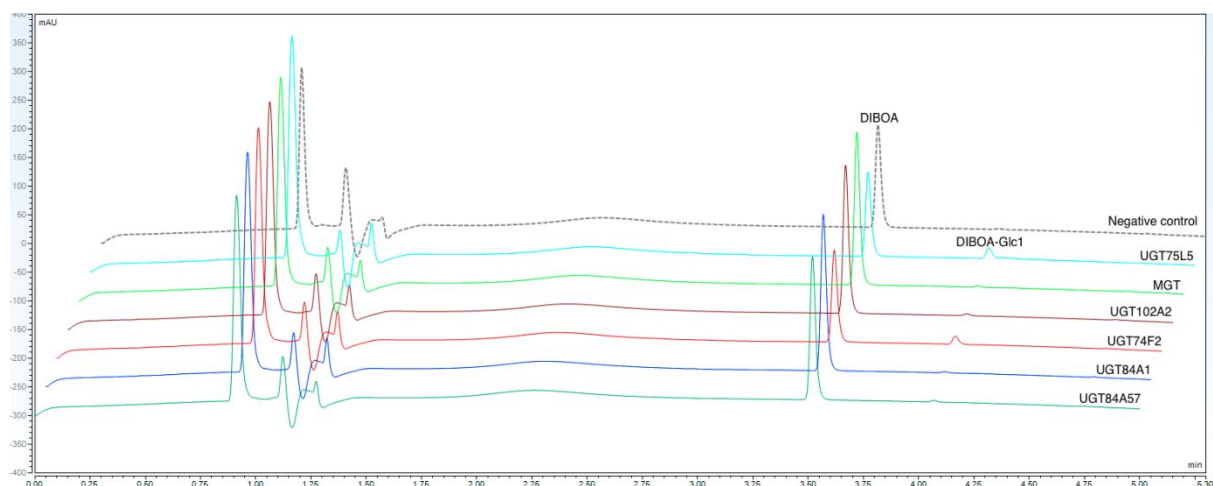


Figure S12. HPLC spectra of predicted GT1 reactions with DIBOA. 100 $\mu\text{g/mL}$ of each GT1 was incubated with 2 mM of UDP-Glc and 0.5 mM of DIBOA in 100 mM of phosphate-citrate buffer (pH 7.0) at 30°C and 300 rpm for 1 h prior to HPLC analysis. DIBOA-Glc1 indicates the DIBOA glucoside. Negative control indicates a reaction with no enzyme. Absorbances were recorded at 240 nm wavelength.

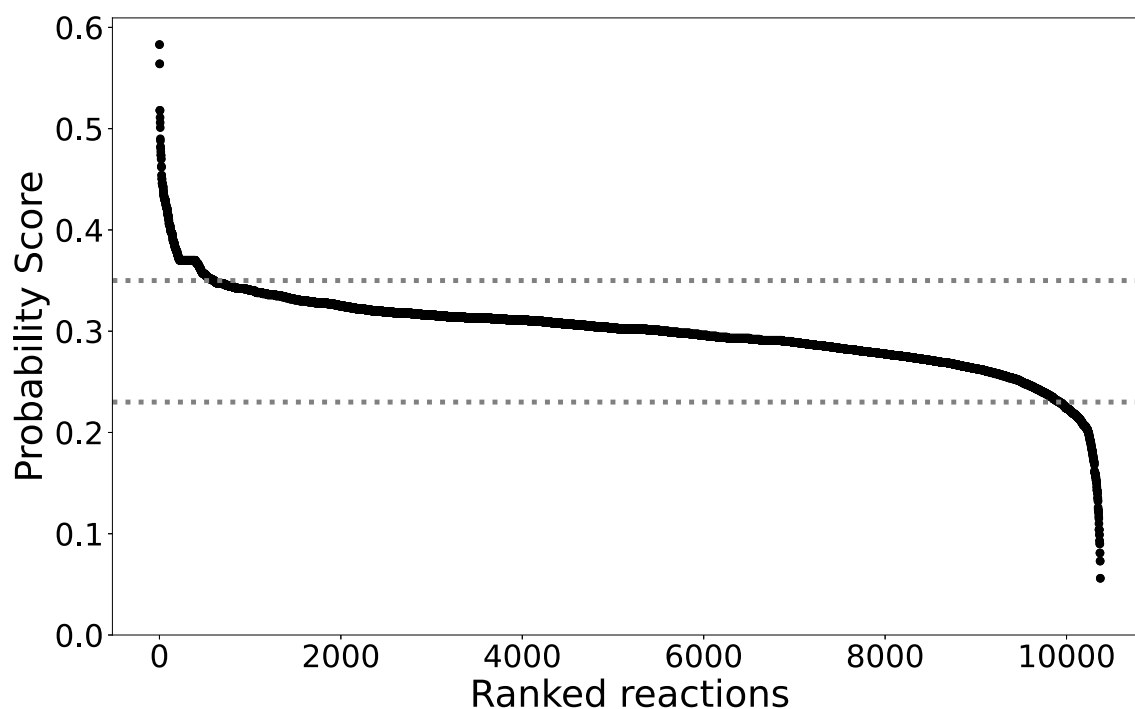


Figure S13. The probabilities scores of reactions with niclosamide and a filtered database of 10,374 GT1 sequences using GASP. The majority of the reactions had a prediction probability between 0.23 and 0.35 (marked by the grey dotted lines), while 589 reactions were above this range and 447 reactions were below.

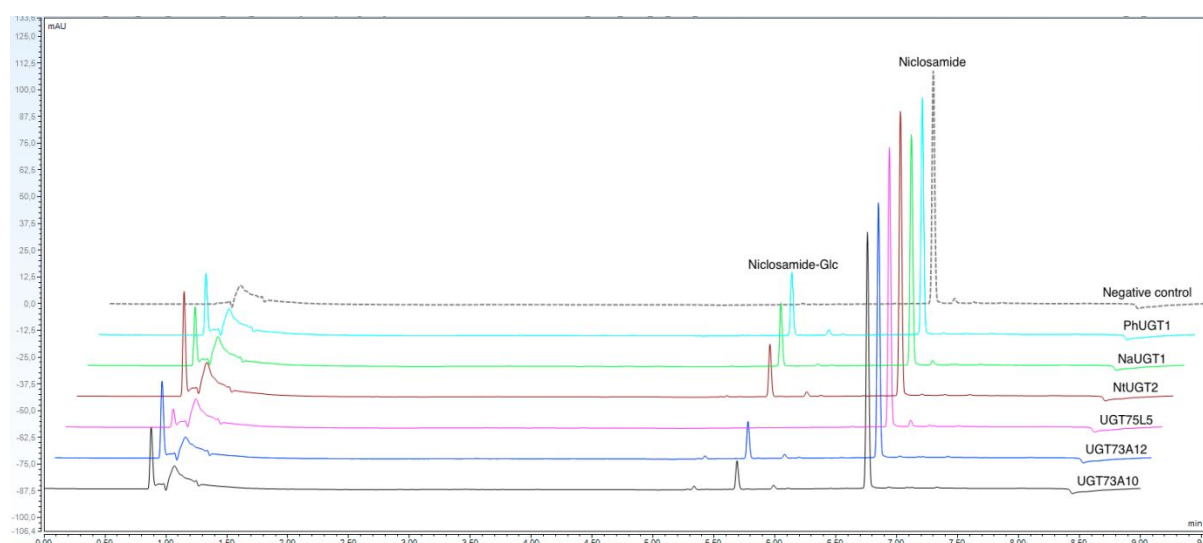


Figure S14. HPLC spectra of predicted GT1 reactions with niclosamide. 50 $\mu\text{g/mL}$ of each GT1 was incubated with 5 mM of UDP-Glc and <1 mM of niclosamide in 50 mM of sodium phosphate buffer (pH 7.45) at 30°C and 300 rpm for 2 h prior to HPLC analysis. Niclosamide-Glc indicates the niclosamide glucoside, while peak around 1 min indicates UDP. Negative control indicates a reaction with no enzyme Absorbances were recorded at 290 nm wavelength.

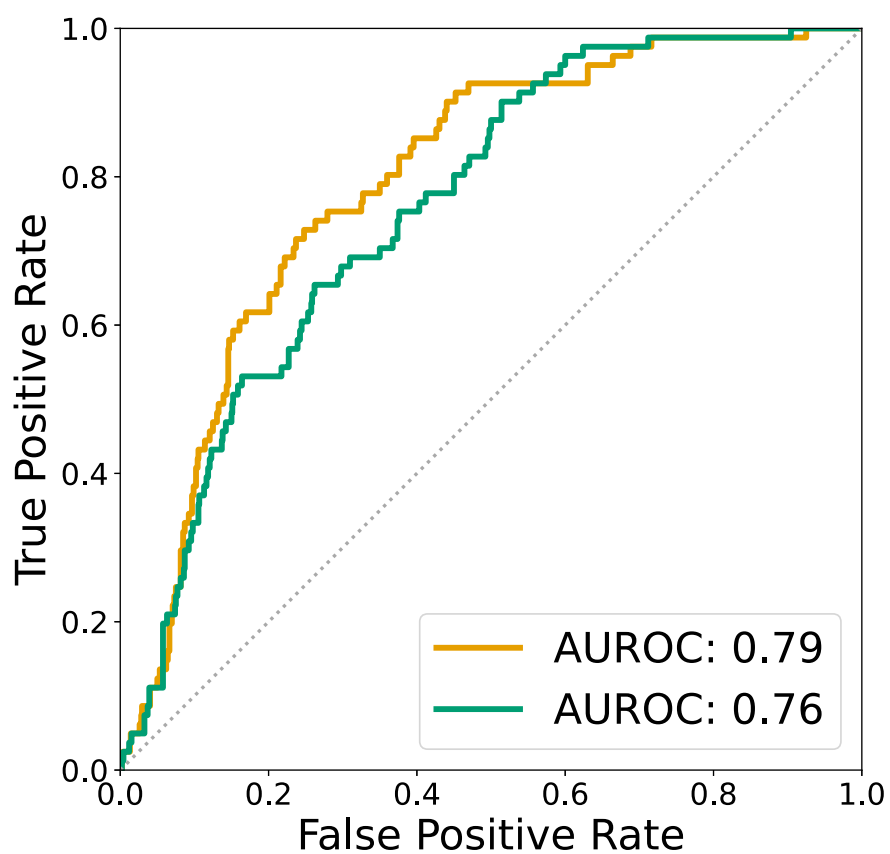


Figure S15. ROC curves and corresponding AUROC values of full GASP model (yellow) and a model trained without including MDS values in the substrate feature set (green). The grey dotted line corresponds to the random predictor.

Supporting Tables

Table S1. Hyperparameters chosen after manual search (final model), values tested during grid search of training random forest classifiers, and values resulting in the highest cross-validation (CV) AUROC.

Hyperparameter	Manual search	Values tested by grid search	Highest CV
n_estimators	1000	400, 600, 800, 1000, 1200, 1400, 1600	1000
criterion	'gini'	'gini', 'entropy'	'entropy'
max_depth	None	20, 50, 100, 200, None	100
min_samples_split	2	2, 5, 10	2
min_samples_leaf	1	1, 2, 4	1
max_features	'sqrt'	'sqrt', 'log2'	'log2'

Table S2. Summary and classifications of models used for GASP validation.

Model Name	Enzyme Representation	Substrate Representation	Model Architecture	Discovery
Baselines				
Levenshtein KNN	Sequence Levenshtein distances	None	KNN	Enzyme
Tanimoto KNN	None	Morgan fingerprints Tanimoto distances	KNN	Substrate
Random Features Ridge	None	Random (n = 80)	Ridge Regression	Substrate
Single-task				
GASP _{enz.disc.}	GASP enzyme pipeline	None	Random Forest	Enzyme
GASP _{sub.disc.}	None	GASP substrate pipeline	Random Forest	Substrate
Full				
GASP	GASP enzyme pipeline	GASP substrate pipeline	Random Forest	Both

Table S3. Overlapping enzymes and substrates between GT-Predict data and in-house dataset used for single-task and baseline model creation. Substrates used for the enzyme discovery models are represented by their CID while enzymes used for the substrate discovery are represented by their accession number.

Enzyme Discovery	Substrate Discovery
7257	A6XNC6
444539	B6EWZ3
445858	O82381
5280343	Q5IFH7
5280445	Q9M156
5280460	Q9LVR1
5280896	
5280961	

Table S4. Results of leave-one-out validation for GT-Predict and GASP.

Substrate name by GT-Predict	Substrate number By GT-Predict	Substrate CID*	Accuracy (GTP ⁶)	Accuracy (GASP)	MCC (GTP ⁶)**	MCC (GASP) **
Baicalein	2	5281605	77%	67%	0.222	0.320
Umbelliferone	3	5281426	68%	47%	0.161	0.302
4-Methyl-umbelliferone	4	5280567	74%	77%	0.197	0.574
Sinapic acid	5	637775	70%	94%	-0.446	0.834
4-hydroxyl-benzoic acid	6	135	79%	83%	0.291	0.556
α -cyano-4-hydroxyl- cinamic acid	7	5328791	96%	79%	0.679	N/A
3,4-dichloroaniline	8	7257	91%	76%	0.024	0.040
3,4-dihydroxylbenzoic acid	9	72	91%	93%	0.49	0.836
2,5-dihydroxylbenzoic acid	10	3469	91%	93%	0.601	0.824
D-glycerate	11	5460358	95%	96%	0.924	N/A
GlcNAc	12	439174	90%	100%	0.991	N/A
Indole 3-acetate	13	801	89%	94%	0.384	0.684
Gibberellin A3	14	6466	98%	100%	0.991	N/A
Gibberellin A4	15	92109	98%	100%	0.991	N/A
($\pm$)-Jasmonic acid	16	12304694	98%	100%	0.99	1.000
($\pm$)- cis. Trans Absciscic acid	17	5280896	98%	90%	0.989	0.507
Kinetin	18	3830	95%	100%	0.127	N/A
Zeatin	19	449093	96%	94%	0.792	0.000
Luteolin	20	5280445	87%	83%	0.325	0.463
Quercetin	21	5280343	87%	84%	0.595	0.515
Fisetin	22	5281614	91%	92%	0.527	0.677
Kaempferol	23	5280863	92%	94%	0.353	0.732
Cinnamic acid	24	444539	81%	86%	0.232	0.385
4-hydroxy cinnamic acid	25	637542	92%	82%	0.514	0.417
3,4-dihydroxy cinnamic acid	26	689043	85%	82%	0.318	0.639
4-hydroxy 3-methoxy cinnamic acid	27	445858	78%	94%	-0.328	0.833
2-hydroxy cinnamic acid	28	637540	75%	80%	0.112	0.408
3-hydroxy cinnamic acid	29	637541	81%	76%	0.312	0.354
Scopoletin	30	5280460	56%	45%	0.991	0.182
Esculetin	31	5281416	87%	64%	0.239	0.393
Threonine	32	6288	98%	100%	0.989	N/A
Glucose	33	5793	98%	100%	0.239	N/A
Dihydrojasmonic acid	34	107126	98%	100%	0.989	1.000
Ser-Phe	35	7009598	98%	100%	0.991	N/A

Ser-Leu	36	7015695	98%	100%	0.991	N/A
BocCysThrOMe	37	N/A	98%	100%	0.991	N/A
1-Thio-S-cyanomethyl-N-acetyl-D-glucosamine	38	101926211	98%	100%	0.991	N/A
MUGlcNAc	41	2733787	98%	100%	0.991	N/A
Trans-dihydroquercetin	42	439533	79%	86%	0.273	0.717
7-hydroxycoumerin 3-carboxylic acid	47	5337757	74%	62%	0.156	0.277
7-hydroxycoumerin 4-acetic acid	48	5338490	72%	72%	-0.023	-0.160
Chloramphenicol	49	5959	96%	100%	0.694	N/A
β -GlcOBn	52	188977	91%	100%	0.426	N/A
α -ManOBn	53	10956572	98%	100%	0.991	N/A
α -ManOPh	54	89049	98%	100%	0.991	N/A
α -ManOCH ₂ Bn	55	49855777	98%	100%	0.991	N/A
α -ManOPMP	56	18396036	98%	100%	0.991	N/A
α -ManOBn(pNO ₂)	57	10781615	98%	100%	0.991	N/A
α -ManOPhF ₅	58	N/A	98%	100%	0.991	N/A
α -ManOBnF ₅	59	N/A	98%	100%	0.991	N/A
ManSTol	60	11161970	98%	100%	0.991	N/A
Catechin	63	9064	77%	74%	0.431	0.505
Genistein	64	5280961	77%	69%	0.348	0.515
N ⁶ -Benzyladenine	65	62389	98%	96%	0.991	N/A
Trans-Zentin-Glucose	66	9842892	93%	100%	0.067	N/A
DihydroZeatin	67	32021	98%	96%	0.989	0.692
Olomoucine	68	4592	96%	94%	0.693	0.000
N ⁶ -isopentenyladenine	69	92180	98%	96%	0.991	N/A
Spectinomycin	72	15541	98%	100%	0.991	N/A
Average			90%	90%	0.398	0.497
Std. Dev			10%	13%	0.361	0.285

*Substrates where PubChem entries were not available denoted with "N/A".

**Substrates for which an MCC value could not be calculated for GASP due to the lack of positive labels in the corresponding subset were omitted in MCC average and standard deviation calculation (denoted with "N/A").

Table S5. Activities of expert-selected and predicted GT1s on DIBOA. Prior to HPLC analysis, 100 µg/mL of each GT1 was incubated with 2 mM of UDP-Glc and 0.5 mM of DIBOA in 100 mM of phosphate-citrate buffer (pH 7.0) at 30°C and 300 rpm for 1 h.

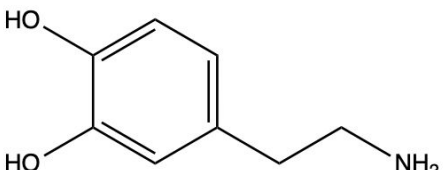
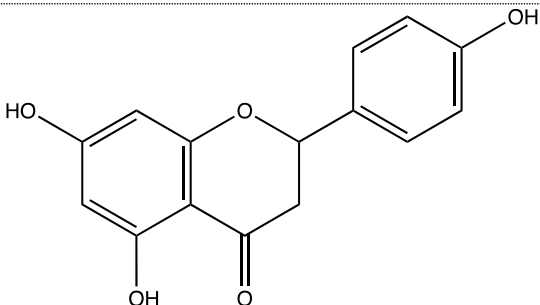
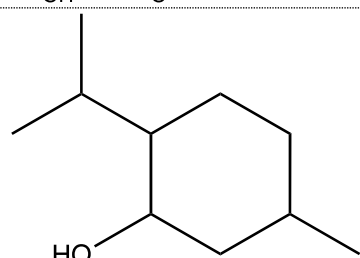
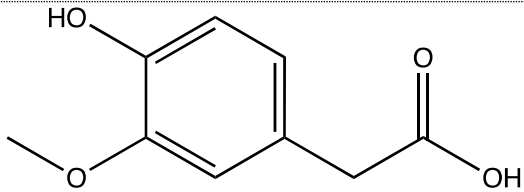
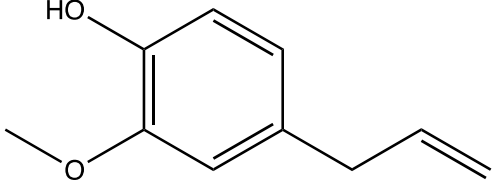
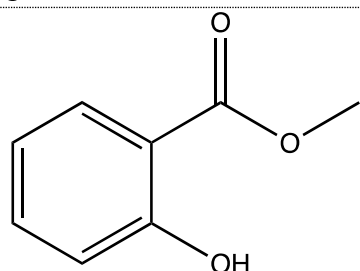
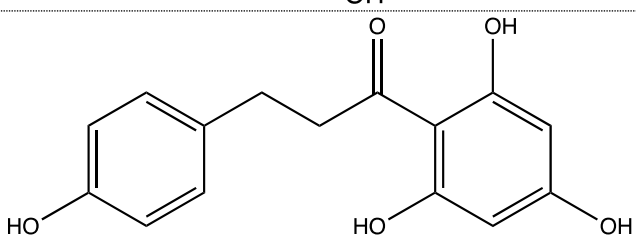
Enzyme name	Source organism	Accession number	Probability score	DIBOA activity
Expert-selected GT1s				
GT188E3	<i>Glycine max</i>	BAF64416.1	0.156	Yes
RHGt1	<i>Rosa hybrid</i>	BAE72452.1	0.137	No
GT171E5	<i>Carthamus tinctorius</i>	AOC55048.1	0.134	Yes
GT171C1	<i>Arabidopsis thaliana</i>	AAC35226.1	0.127	Yes
GT188A10	<i>Forsythia x intermedia</i>	BAI65910.1	0.119	Yes
GT188C1	<i>Orzya sativa</i> subsp. <i>japonica</i>	BAC10743.1	0.098	Yes
Predicted GT1s				
GT184A1	<i>Arabidopsis thaliana</i>	AAN72025.1	0.377	No
GT1102A2	<i>Streptomyces antibioticus</i>	ABA42119.2	0.367	No
MGT	<i>Streptomyces lividans</i>	AAA26780.1	0.348	No
GT175L5	<i>Lycium barbarum</i>	BAG80544.1	0.327	Yes
GT184A57	<i>Eutrema japonicum</i>	BBI55602.1	0.300	Yes
GT174F2	<i>Arabidopsis thaliana</i>	AAB64024.1	0.203	Yes

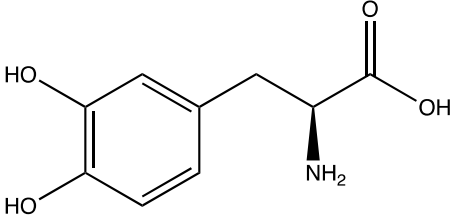
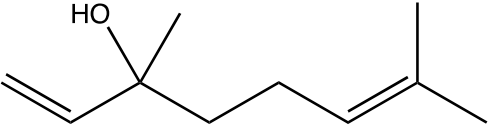
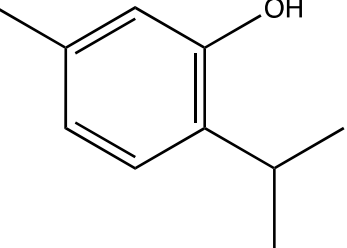
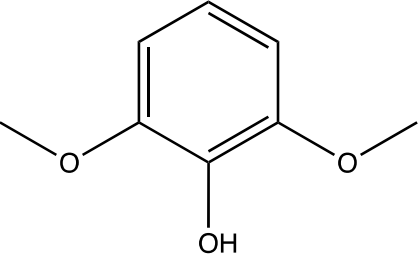
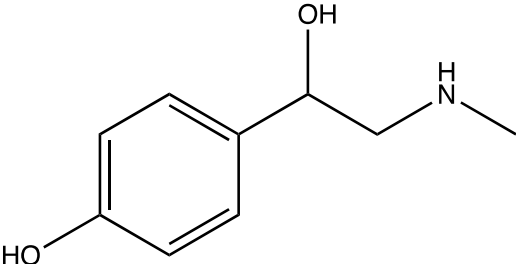
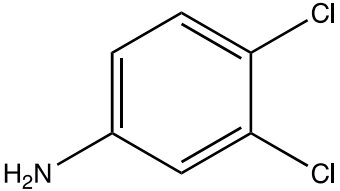
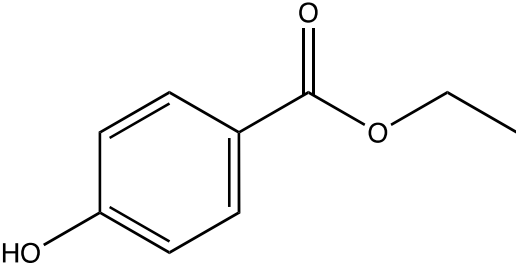
Table S6. Activities of random selection of in-house GT1s and GASP predicted GT1s on niclosamide.

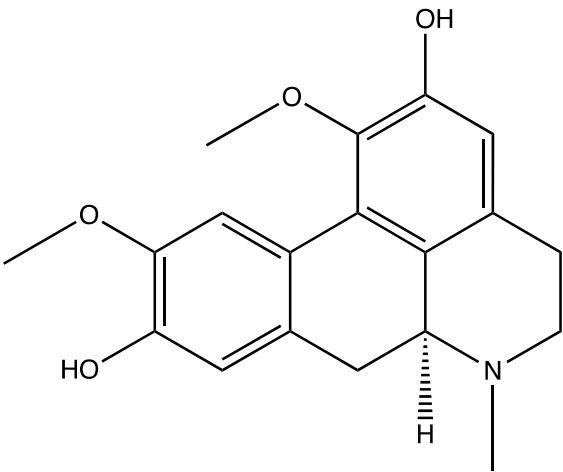
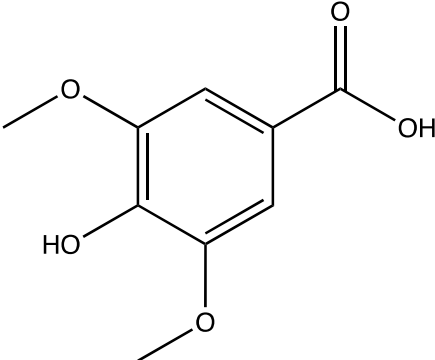
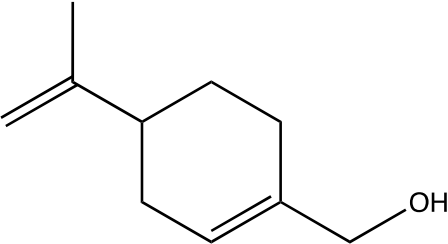
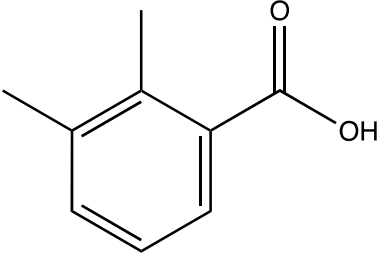
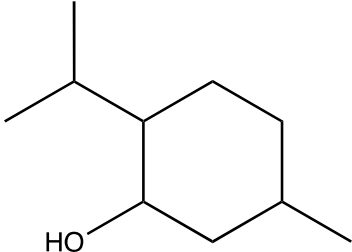
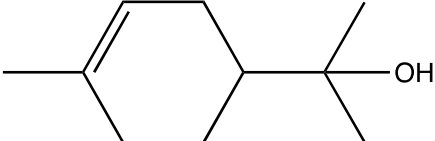
Enzyme name	Source organism	Accession number	Probability score	Niclosamide activity*
Randomly selected UGTs				
PtUGT1	<i>Persicaria tinctoria</i>	BBK15460.1	0.320	No
RhGT1	<i>Rosa</i> hybrid cultivar	BAD99560.1	0.279	Yes
UGT706F8	<i>Zea mays</i>	NP_001132331.1	0.268	No
UGT708A6	<i>Zea mays</i>	NP_001132650.2	0.256	Yes
UGT71B1	<i>Zea mays</i>	NP_001148465.2	0.273	Yes
UGT71C1	<i>Arabidopsis thaliana</i>	AAC35226.1	0.414	Yes
UGT71D1	<i>Arabidopsis thaliana</i>	AAC35239.1	0.308	Yes
UGT71E5	<i>Carthamus tinctorius</i>	AOC55048.1	0.292	No
UGT71F5	<i>Dianthus caryophyllus</i>	BAF75880.1	0.274	Yes
UGT72B10	<i>Lycium barbarum</i>	BAG80556.1	0.309	No
UGT72B19	<i>Fagopyrum esculentum</i>	BAP90371.1	0.265	No
UGT72G3	<i>Zea mays</i>	NP_001147461.1	0.242	No
UGT72G4	<i>Zea mays</i>	NP_001130518.1	0.237	No
UGT72J1	<i>Oryza sativa</i> subsp. <i>japonica</i>	BAG98542.1	0.283	Yes
UGT88A10	<i>Forsythia intermedia</i>	BAI65910.1	0.295	Yes
UGT88C10	<i>Zea mays</i>	NP_001159290.1	0.219	No
UGT88D7	<i>Perilla frutescens</i>	BAG31948.1	0.287	No
UGT88E3	<i>Glycine max</i>	BAF64416.1	0.287	Yes
UGT88J1	<i>Fagopyrum esculentum</i>	BAP90366.1	0.303	Yes
GASP-predicted UGTs				
UGT73A10	<i>Lycium barbarum</i>	BAG80536.1	0.583	Yes
UGT73A12	<i>Lycium barbarum</i>	BAG80550.1	0.564	Yes
UGT75L5	<i>Lycium barbarum</i>	BAG80544.1	0.518	No
UGT84A1	<i>Arabidopsis thaliana</i>	AAN72025.1	0.506	N/D
WsUGT1	<i>Withania somnifera</i>	ACO44747.1	0.501	N/D
UGT73A25	<i>Nicotiana benthamiana</i>	UXR81504.1	0.490	N/D
UGT73A24	<i>Nicotiana benthamiana</i>	UXR81503.1	0.488	N/D
NtUGT1	<i>Nicotiana tabacum</i>	AAB36652.1	0.480	N/D
NtUGT2	<i>Nicotiana tabacum</i>	CAC07382.1	0.474	Yes
PhUGT1	<i>Petunia hybrida</i>	AAW00305.1	0.474	Yes
NaUGT1 (TOGT1_1)	<i>Nicotiana attenuata</i>	AQQ16675.1	0.462	Yes
UGT84A59	<i>Eutrema japonicum</i>	BBI55604.1	0.463	N/D

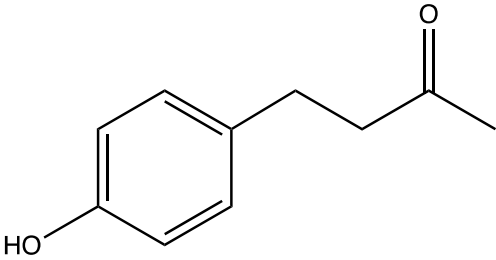
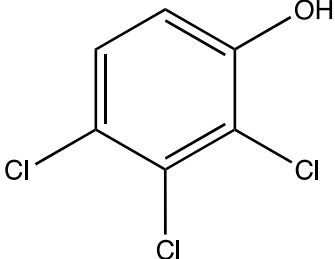
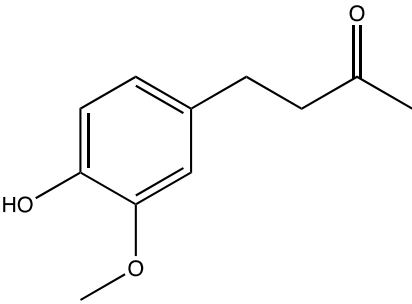
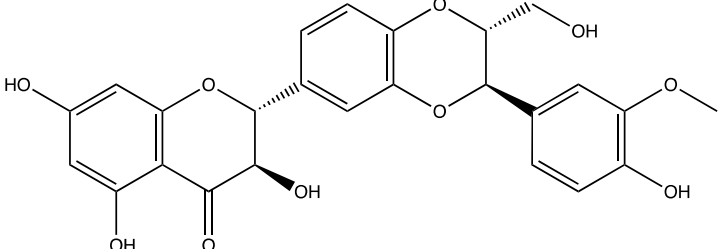
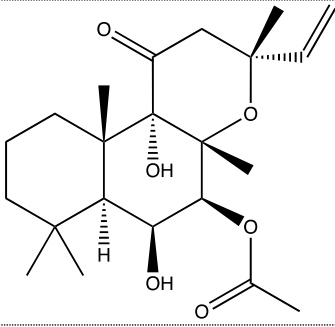
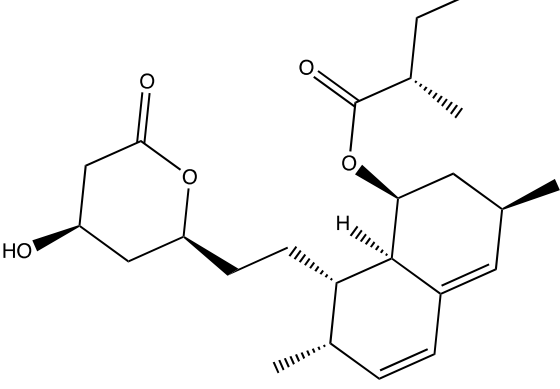
*N/D – no data due to very low or no expression in *E. coli* or insoluble expression in the case of UGT73A25.

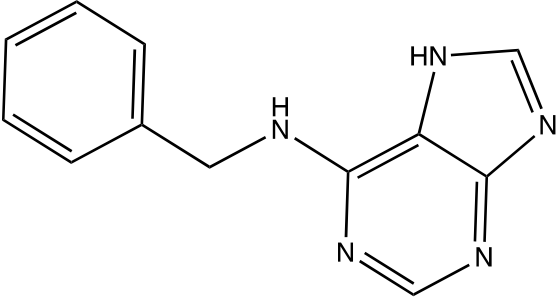
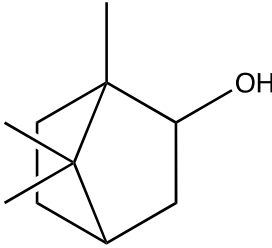
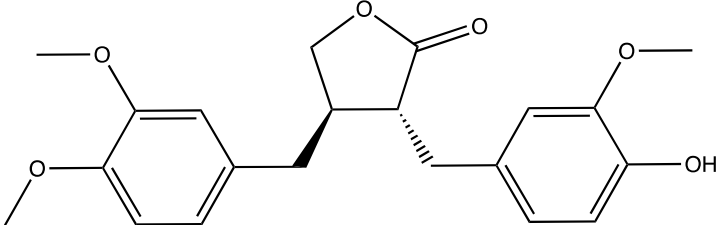
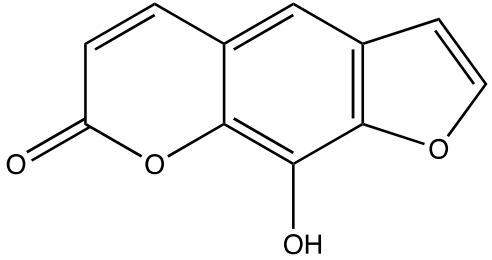
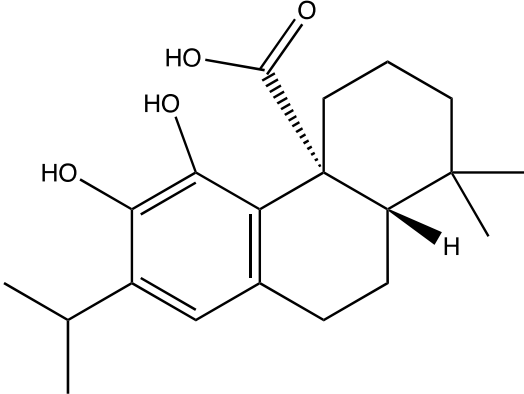
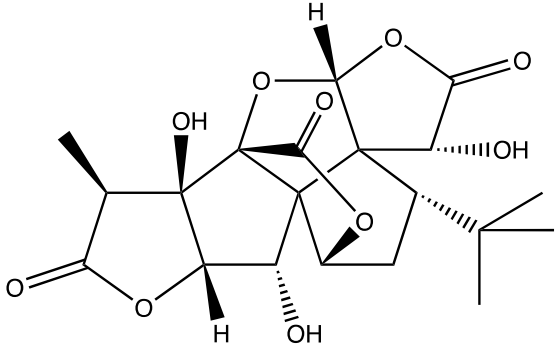
Appendix

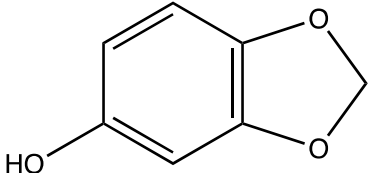
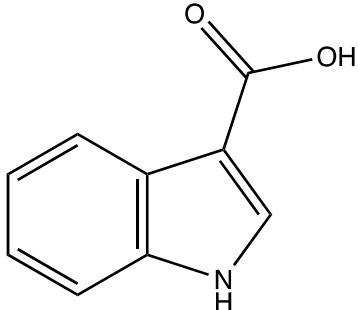
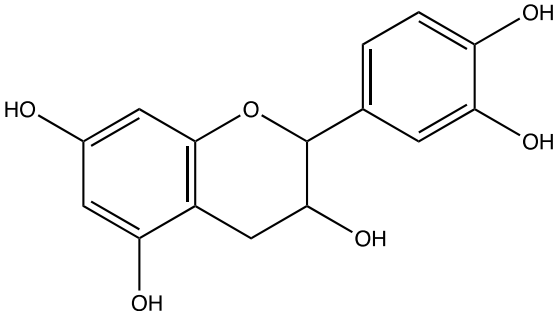
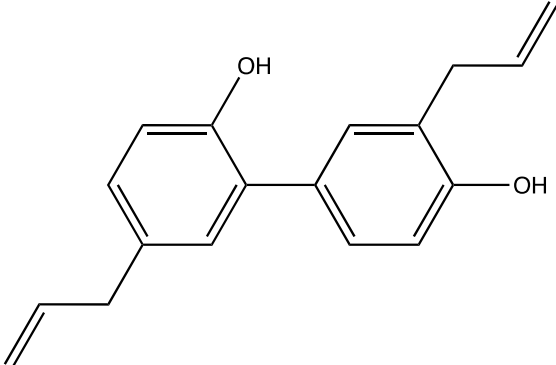
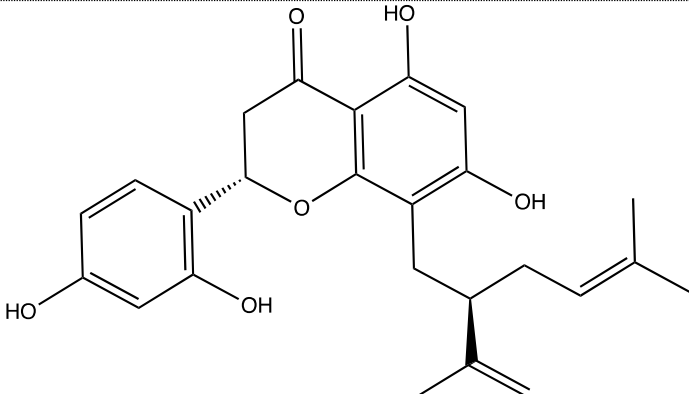
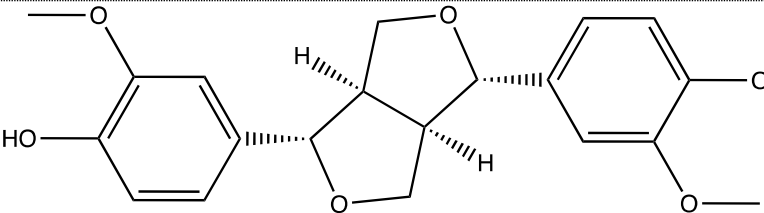
Acceptor	CID	Structure
Dopamine	681	
Naringenin	932	
Menthol	1254	
Homovanillic acid	1738	
Eugenol	3314	
Methyl Salicylate	4133	
Phloretin	4788	

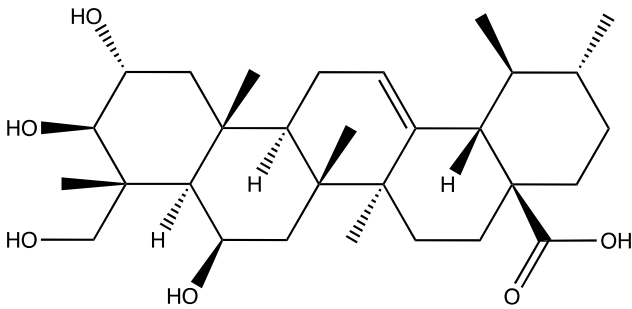
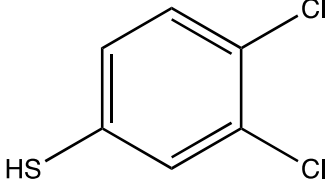
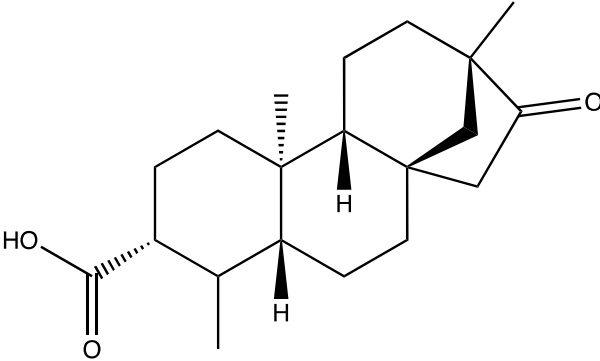
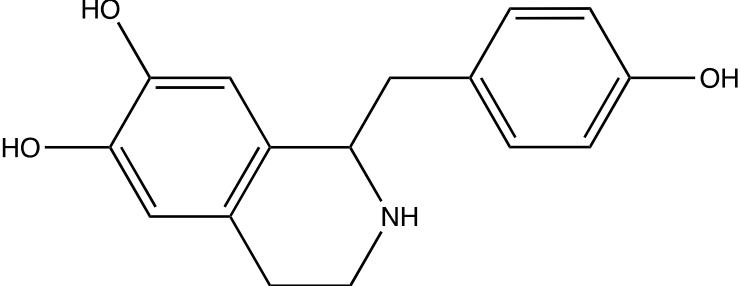
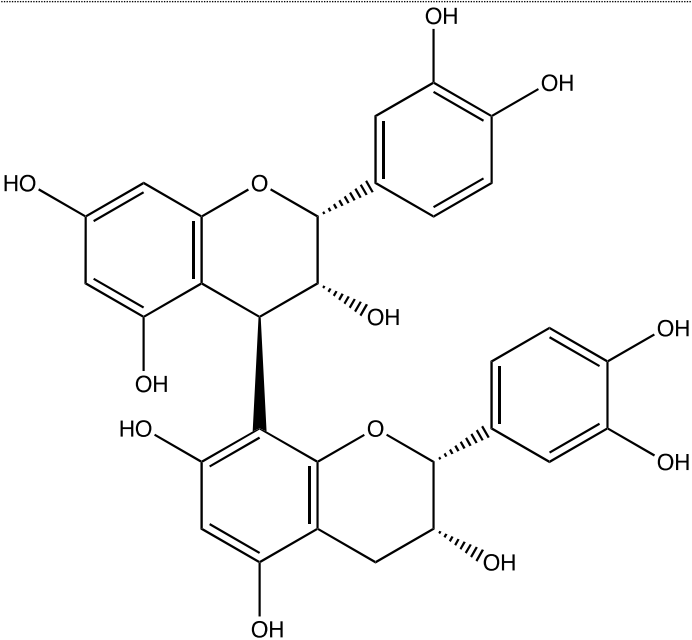
Levodopa	6047	
Linalool	6549	
Thymol	6989	
2,6-Dimethoxyphenol	7041	
Synephrine	7172	
3,4-Dichloroaniline	7257	
Ethylparaben	8434	

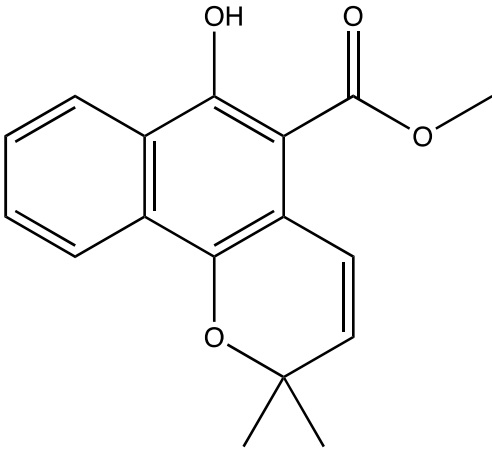
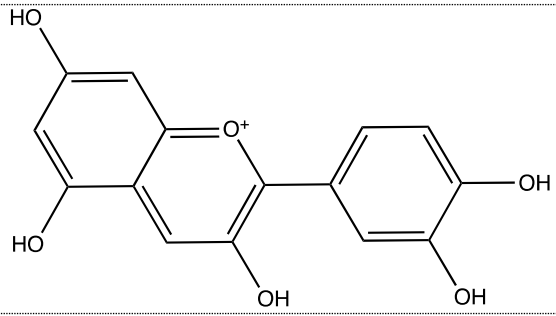
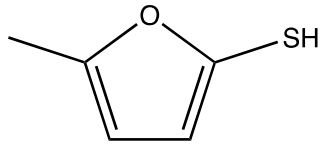
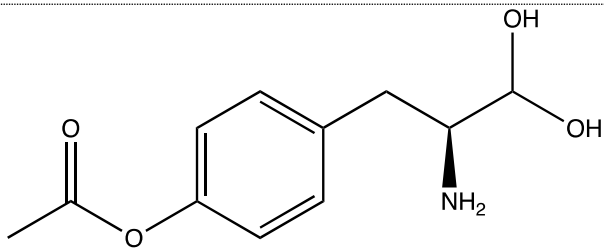
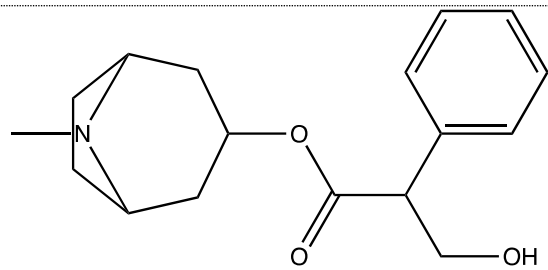
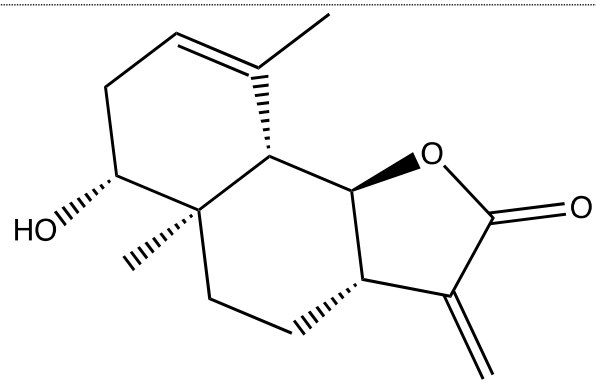
Boldine	10154	
Syringic acid	10742	
Perillyl alcohol	10819	
2,3-Dimethylbenzoic acid	11782	
I-Menthol	16666	
alpha-Terpineol	17100	

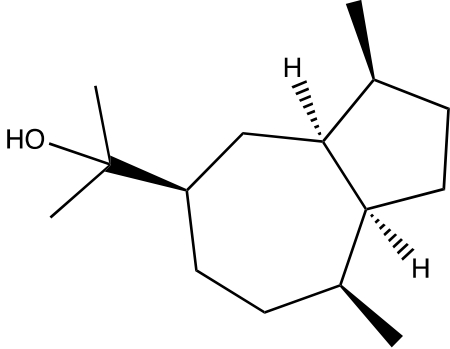
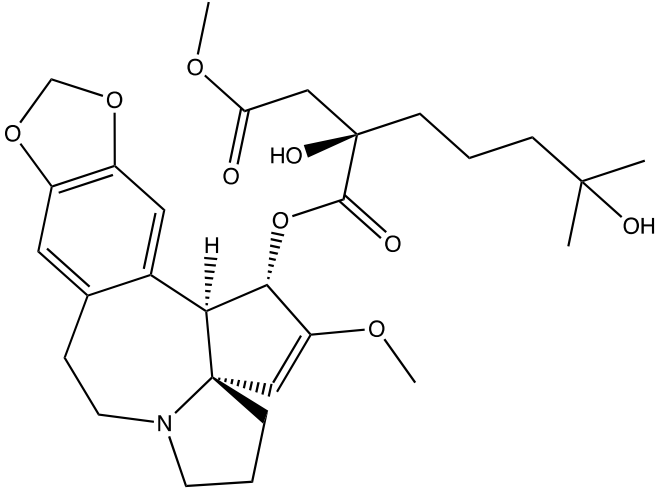
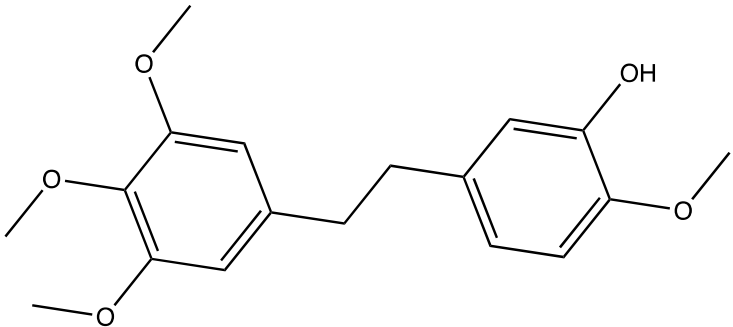
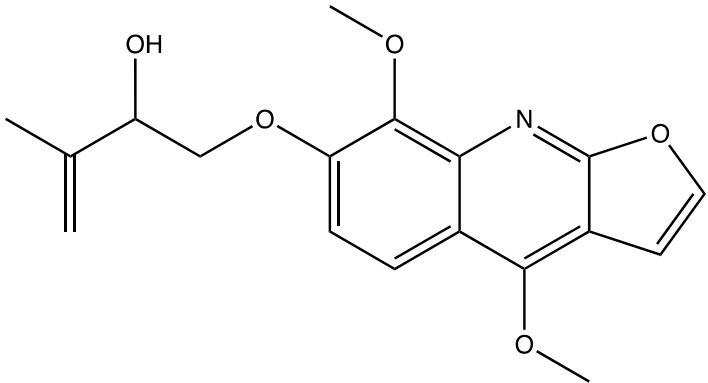
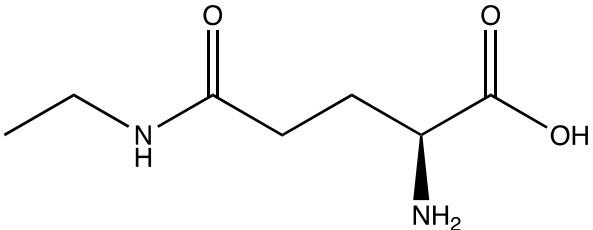
Frambinone	21648	
2,3,4-Trichlorophenol	27582	
Zingerone	31211	
Silibinin	31553	
Forskolin	47936	
Lovastatin	53232	

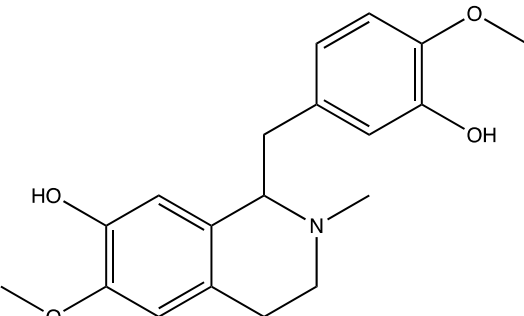
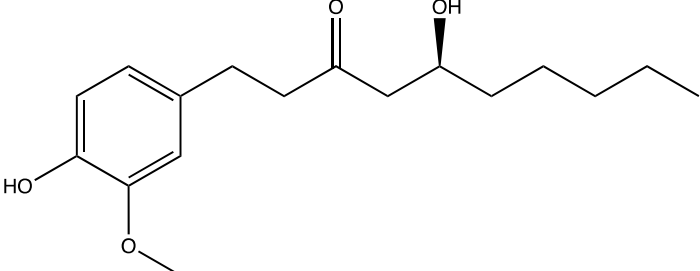
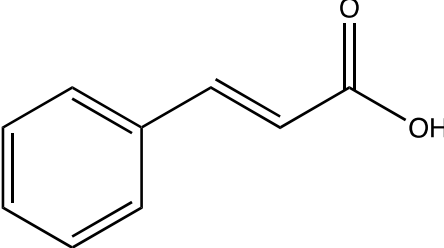
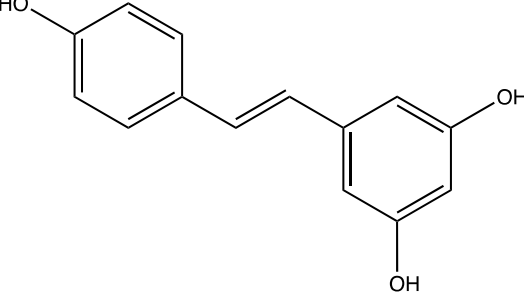
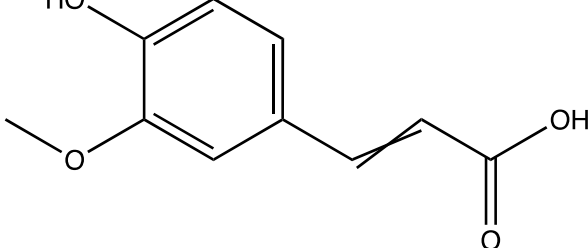
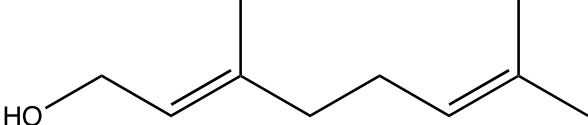
6-Benzylaminopurine	62389	
Isoborneol	64685	
Arctigenin	64981	
Xanthotoxol	65090	
Carnosic acid	65126	
Ginkgolide B	65243	

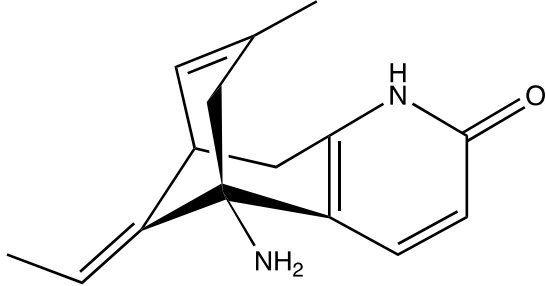
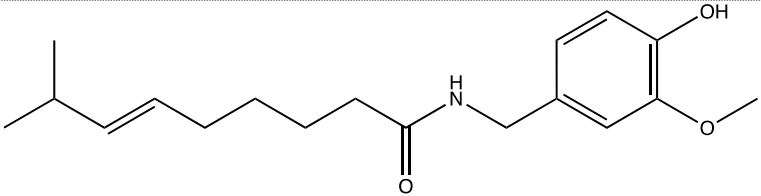
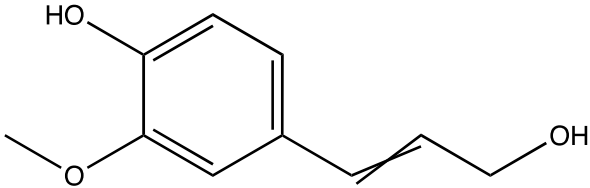
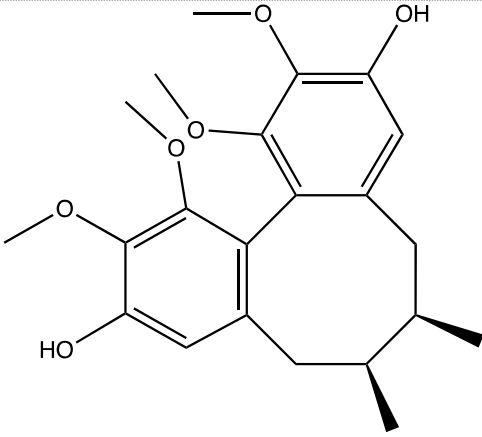
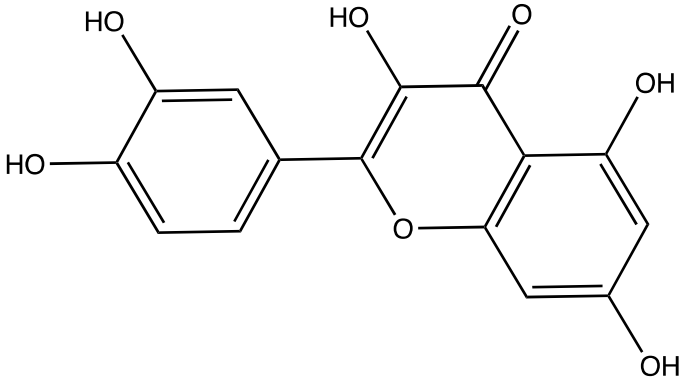
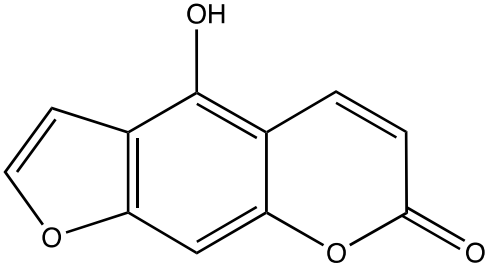
Sesamol	68289	
Indole-3-carboxylic acid	69867	
(-)-Epicatechin	72276	
Honokiol	72303	
Sophoraflavanone G	72936	
Pinoresinol	73399	

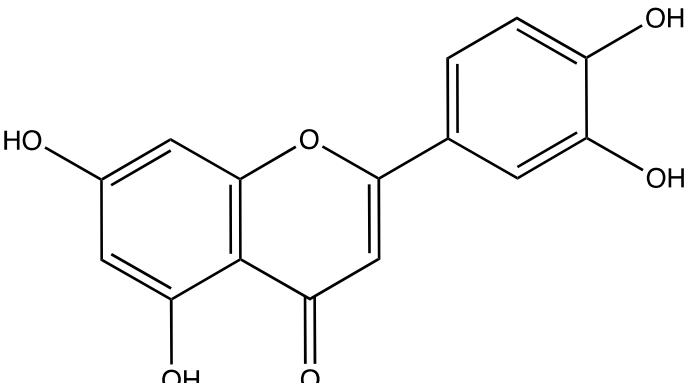
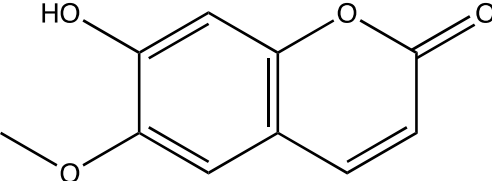
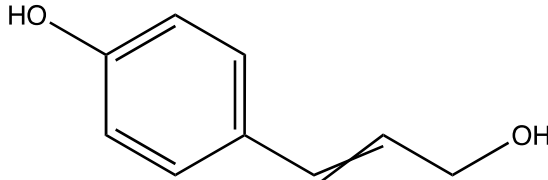
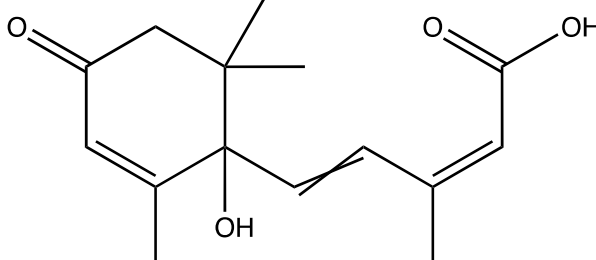
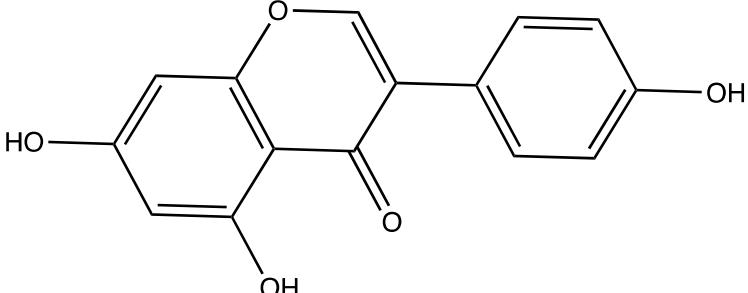
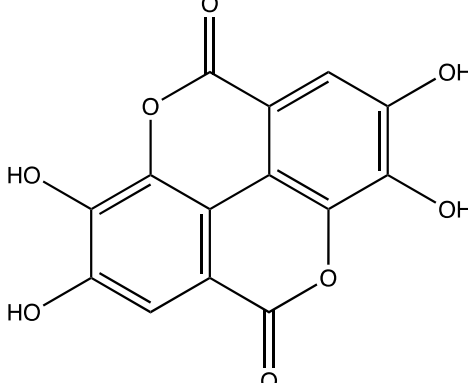
Madecassic acid	73412	
3,4-Dichlorothiophenol	79964	
Isosteviol	99514	
Higenamine	114840	
Procyanidin B2	122738	

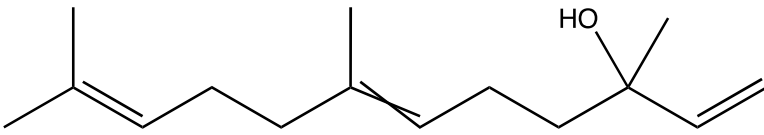
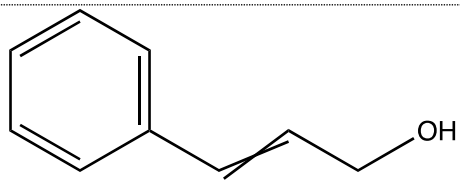
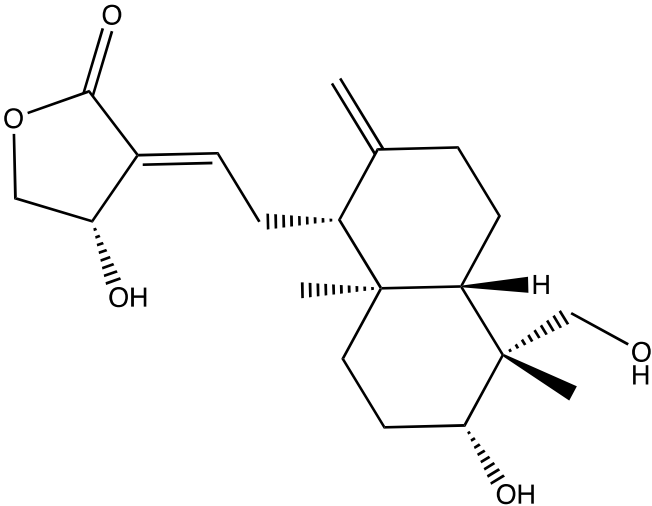
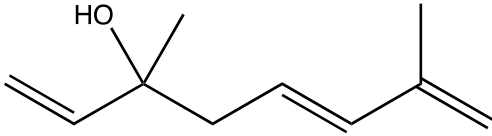
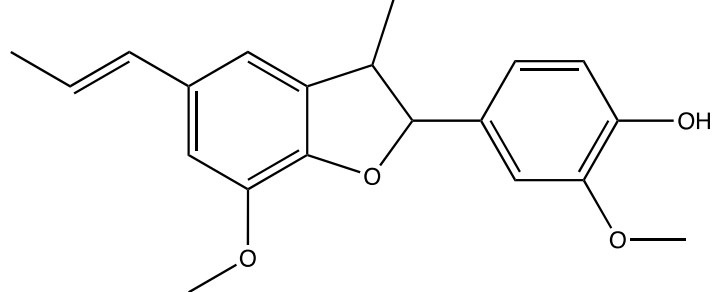
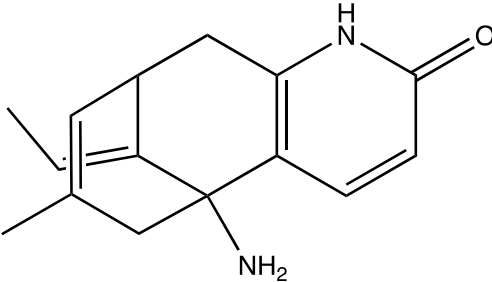
Mollugin	124219	
Cyanidin	128861	
5-Methyl-2-furanthiol	143754	
Hydroxytyrosol Acetate	155240	
Atropine	174174	
Santamarine	188297	

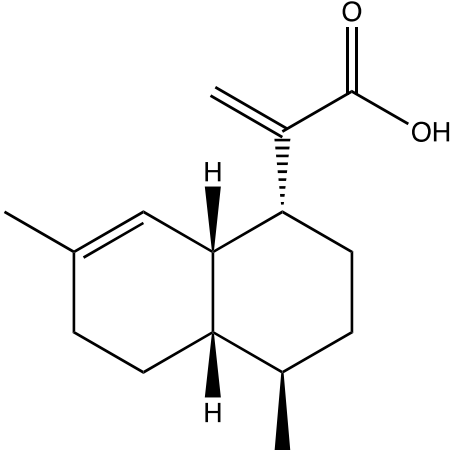
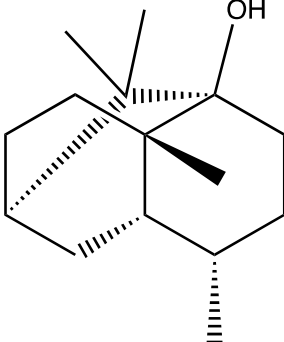
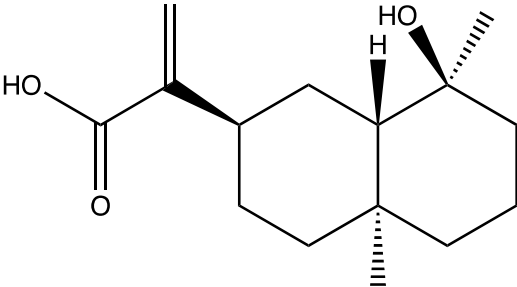
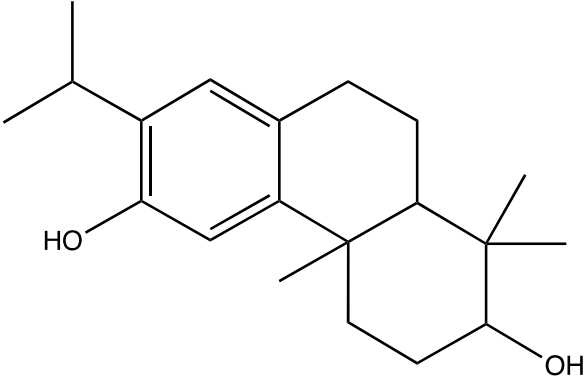
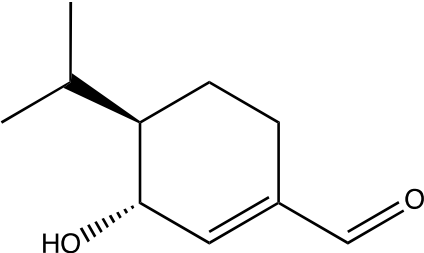
Guaiol	227829	
Omacetaxine mepesuccinate	285033	
Erianin	356759	
Evodine	398789	
L-Theanine	439378	

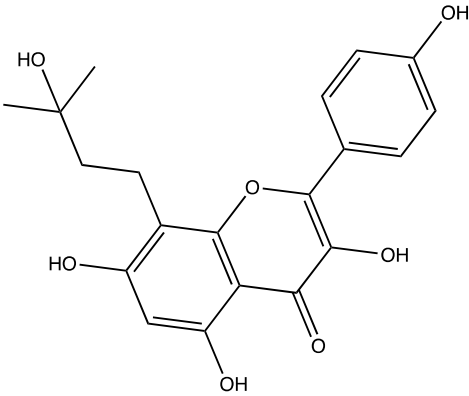
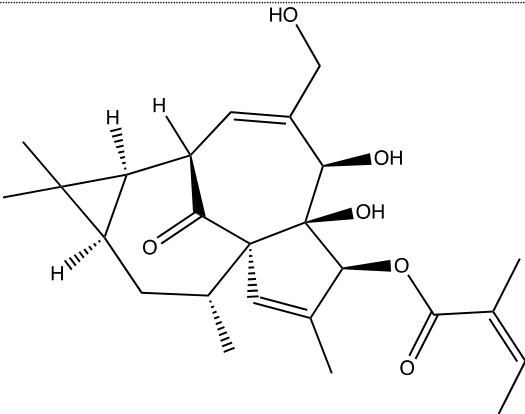
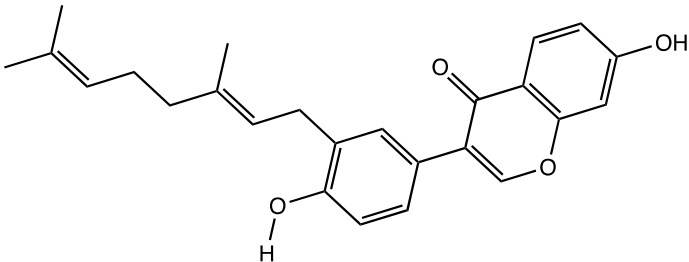
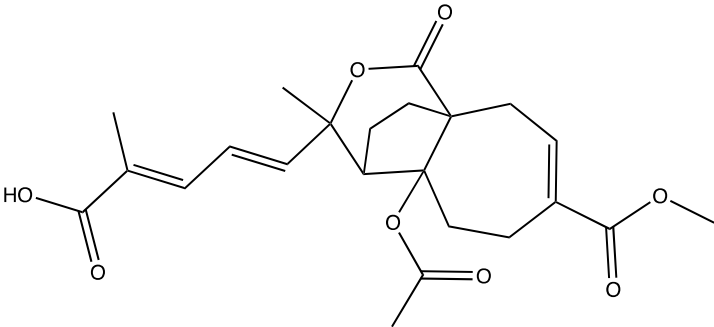
Reticuline	439653	
Gingerol	442793	
Cinnamic acid	444539	
Resveratrol	445154	
Ferulic acid	445858	
Geraniol	637566	

Huperzine A	854026	
Capsaicin	1548943	
Coniferyl alcohol	1549095	
Gomisin J	3001686	
Quercetin	5280343	
Bergaptol	5280371	

Luteolin	5280445	
Scopoletin	5280460	
p-Coumaryl alcohol	5280535	
Absciscic acid	5280896	
Genistein	5280961	
Ellagic acid	5281855	

Nerolidol	5284507	
Cinnamyl alcohol	5315892	
Andrographolide	5318517	
Hotrienol	5366264	
Dehydrodiisoeugenol	5379033	
(+/-)-Huperzine A	5912039	

Artemisinic acid	1092246 5	
Patchouli alcohol	1095517 4	
Illicic acid	1187619 5	
Hinokiol	1231049 2	
Eucamalol	1242623 9	

Noricaritin	1442742 3	
Ingenol mebutate	2358194 6	
Corylifol A	2505640 7	
Pseudolarix acid B	5337739 0	
Theaflavine	1354037 98	