Virtual Screening of two ligand databases using five QSAR models

Edward Francisco Mendez-Otalvaro, Daniel Alberto Barragan, Isaias Lans-Vargas

2021

Importing all datasets with the molecules

After the virtual screening with each QSAR model, let's concatenate all the outputs to analyze the molecules and ranking them according to different scores

Importing output from virtual screening of each QSAR model

Index of the models

```
indices_2 <- read.csv("indice_2.csv", stringsAsFactors = FALSE, header = TRUE, sep = ",")
indices_2 <- cbind(indices_2, c(1:310))
colnames(indices_2) <- c("Label", "Index")

indices_3 <- read.csv("indice_3.csv", stringsAsFactors = FALSE, header = TRUE, sep = ",")
indices_3 <- cbind(indices_3, c(1:465))
colnames(indices_3) <- c("Label", "Index")
```

QSAR Results

```
## LASSO VS
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
LASSO_2 <- read.csv("VS_LASSO-2.csv", stringsAsFactors = FALSE, header = TRUE, sep = ",")
LASSO_2 <- left_join(LASSO_2, indices_2, by = "Index")

LASSO_3 <- read.csv("VS_LASSO-3.csv", stringsAsFactors = FALSE, header = TRUE, sep = ",")
LASSO_3 <- left_join(LASSO_3, indices_3, by = "Index")

## LASSO-MLR VS
L_MLR_2 <- read.csv("VS_MLRLASSO-2.csv", stringsAsFactors = FALSE, header = TRUE,
    sep = ",")
L_MLR_2 <- left_join(L_MLR_2, indices_2, by = "Index")
```

```
L_MLR_3 <- read.csv("VS_MLRLASSO-3.csv", stringsAsFactors = FALSE, header = TRUE,
    sep = ",")
L_MLR_3 <- left_join(L_MLR_3, indices_3, by = "Index")

## GA-MLR VS
GA_MLR_2 <- read.csv("VS_GAMLR-2.csv", stringsAsFactors = FALSE, header = TRUE, sep = ",")
GA_MLR_2 <- left_join(GA_MLR_2, indices_2, by = "Index")

GA_MLR_3 <- read.csv("VS_GAMLR-3.csv", stringsAsFactors = FALSE, header = TRUE, sep = ",")
GA_MLR_3 <- left_join(GA_MLR_3, indices_3, by = "Index")

## PLS VS
PLS_2 <- read.csv("VS_PLS-2.csv", stringsAsFactors = FALSE, header = TRUE, sep = ",")
PLS_2 <- left_join(PLS_2, indices_2, by = "Index")

PLS_3 <- read.csv("VS_PLS-3.csv", stringsAsFactors = FALSE, header = TRUE, sep = ",")
PLS_3 <- left_join(PLS_3, indices_3, by = "Index")

## RSM VS
RSM_2 <- read.csv("VS_RSMMLR-2.csv", stringsAsFactors = FALSE, header = TRUE, sep = ",")
RSM_2 <- left_join(RSM_2, indices_2, by = "Index")

RSM_3 <- read.csv("VS_RSMMLR-3.csv", stringsAsFactors = FALSE, header = TRUE, sep = ",")
RSM_3 <- left_join(RSM_3, indices_3, by = "Index")
```

Concatenating datasets

```
## Concatenating datasets

set.seed(10)
Moleculas_2 <- as.data.frame(cbind(LASSO_2$Index, L_MLR_2$Index, GA_MLR_2$Index,
    PLS_2$Index, RSM_2$Index))
colnames(Moleculas_2) <- c("LASSO", "LASSO-MLR", "GA-MLR", "PLS", "RSM")

Nombres_2 <- as.data.frame(cbind(LASSO_2$Label, L_MLR_2$Label, GA_MLR_2$Label, PLS_2$Label,
    RSM_2$Label))
colnames(Nombres_2) <- c("LASSO", "LASSO-MLR", "GA-MLR", "PLS", "RSM")


Moleculas_3 <- as.data.frame(cbind(LASSO_3$Index, L_MLR_3$Index, GA_MLR_3$Index,
    PLS_3$Index, RSM_3$Index))
colnames(Moleculas_3) <- c("LASSO", "LASSO-MLR", "GA-MLR", "PLS", "RSM")

Nombres_3 <- as.data.frame(cbind(LASSO_3$Label, L_MLR_3$Label, GA_MLR_3$Label, PLS_3$Label,
    RSM_3$Label))
colnames(Nombres_3) <- c("LASSO", "LASSO-MLR", "GA-MLR", "PLS", "RSM")
```

Frequency of apparition of the molecules

```
## Two variants
Frecuencia_2 <- table(unlist(Moleculas_2[1:20, ]))
Frecuencia_2 <- as.data.frame(Frecuencia_2)
```

```r
colnames(Frecuencia_2) <- c("Molecule", "Occurrence frequency 2")
Frecuencia_2$Molecule <- as.numeric(as.character(Frecuencia_2$Molecule))

## Three variants
Frecuencia_3 <- table(unlist(Moleculas_3[1:20, ]))
Frecuencia_3 <- as.data.frame(Frecuencia_3)
colnames(Frecuencia_3) <- c("Molecule", "Occurrence frequency 3")
Frecuencia_3$Molecule <- as.numeric(as.character(Frecuencia_3$Molecule))
```

Molecule frequency per QSAR model

Two variants

```r
## Matching each frequency with initial index

I_INDEX_2 <- match(Frecuencia_2$Molecule, Moleculas_2$LASSO)
II_INDEX_2 <- match(Frecuencia_2$Molecule, Moleculas_2$`LASSO-MLR`)
III_INDEX_2 <- match(Frecuencia_2$Molecule, Moleculas_2$`GA-MLR`)
IV_INDEX_2 <- match(Frecuencia_2$Molecule, Moleculas_2$PLS)
V_INDEX_2 <- match(Frecuencia_2$Molecule, Moleculas_2$RSM)

Frecuencia_2_final <- as.data.frame(cbind(Frecuencia_2$Molecule, I_INDEX_2, II_INDEX_2,
    III_INDEX_2, IV_INDEX_2, V_INDEX_2))
colnames(Frecuencia_2_final) <- c("Molecule", "LASSO", "LASSO_MLR", "GA_MLR", "PLS",
    "RSM")
kable(Frecuencia_2_final[1:20, ], align = "cccccc")
```

Molecule	LASSO	LASSO_MLR	GA_MLR	PLS	RSM
1	21	21	11	22	192
2	22	22	12	27	193
5	25	25	1	118	1
6	26	26	2	135	2
7	11	11	143	8	64
8	12	12	144	3	65
11	29	29	3	121	3
12	30	30	4	63	4
21	13	13	149	6	68
22	14	14	150	1	69
29	15	15	209	17	129
30	16	16	210	21	130
35	49	49	10	20	163
42	56	56	13	23	194
46	60	60	268	12	298
55	69	69	14	13	117
71	85	85	6	14	122
72	86	86	7	10	123
82	96	96	15	16	195
85	99	99	16	18	196

Three variants

```
I_INDEX_3 <- match(Frecuencia_3$Molecule, Moleculas_3$LASSO)
II_INDEX_3 <- match(Frecuencia_3$Molecule, Moleculas_3$`LASSO-MLR`)
III_INDEX_3 <- match(Frecuencia_3$Molecule, Moleculas_3$`GA-MLR`)
IV_INDEX_3 <- match(Frecuencia_3$Molecule, Moleculas_3$PLS)
V_INDEX_3 <- match(Frecuencia_3$Molecule, Moleculas_3$RSM)

Frecuencia_3_final <- as.data.frame(cbind(Frecuencia_3$Molecule, I_INDEX_3, II_INDEX_3,
    III_INDEX_3, IV_INDEX_3, V_INDEX_3))
colnames(Frecuencia_3_final) <- c("Molecule", "LASSO", "LASSO_MLR", "GA_MLR", "PLS",
    "RSM")
kable(Frecuencia_3_final[1:20, ], align = "cccccc")
```

Molecule	LASSO	LASSO_MLR	GA_MLR	PLS	RSM
1	31	31	13	28	286
7	37	37	1	195	1
8	38	38	2	175	2
9	39	39	3	129	3
10	16	16	215	1	98
11	17	17	216	38	99
12	18	18	217	5	100
16	43	43	4	155	4
17	44	44	5	117	5
18	45	45	6	159	6
31	19	19	224	42	104
32	20	20	225	43	105
43	22	22	314	20	197
54	75	75	12	24	250
59	80	80	410	10	452
62	83	83	14	15	287
68	89	89	411	12	453
83	104	104	15	48	181
84	105	105	16	37	182
121	142	142	17	18	288

Top molecules

Top 20 molecules two variants

```
Moleculas_2_Total <- cbind(Moleculas_2, Nombres_2)
Moleculas_2_Total <- Moleculas_2_Total[, c(1, 6, 2, 7, 3, 8, 4, 9, 5, 10)]

kable(Moleculas_2_Total[1:20, ], align = "ccccccccccc")
```

LASSO	LASSO.1	LASSO-MLR	LASSO-MLR.1	GA-MLR	GA-MLR.1	PLS	PLS.1	RSM	RSM.1
115	446101	115	446101	5	118707521	22	54758613	5	118707521
116	446101	116	446101	6	118707521	125	440271	6	118707521
215	ZINC000019893747	215	ZINC000019893747	7	76325328	8	91820057	11	76325328
216	ZINC000019893747	216	ZINC000019893747	7	76325328	126	440271	12	76325328
275	ZINC000257346903	275	ZINC000257346903	2	ZINC000101361978	2	440235	147	198515
276	ZINC000257346903	276	ZINC000257346903	2	12600646	21	54758613	148	198515

LASSO	LASSO.1	LASSO-MLR	LASSO-MLR.1	GA-MLR	GA-MLR.1	PLS	PLS.1	RSM	RSM.1
277	ZINC000257347637	277	ZINC000257347637	128	12600646	128	440235	149	198514
278	ZINC000257347638	278	ZINC000257347638	96	5284365	7	91820057	277	ZINC000257347638
279	ZINC000257347639	279	ZINC000257347639	739	ZINC000101361998	739	278	ZINC000257347638	
280	ZINC000257347640	280	ZINC000257347640	46174147	72	12600646	279	ZINC000257347639	
7	91820057	7	91820057	1	134695375	96	5284365	280	ZINC000257347639
8	91820057	8	91820057	2	134695375	46	45266741	99	3052142
21	54758613	21	54758613	42	45357367	55	42612215	100	3052142
22	54758613	22	54758613	55	42612215	71	12600646	150	198514
29	51351654	29	51351654	82	7067772	86	6713972	219	ZINC000032106887
30	51351654	30	51351654	85	6713972	82	7067772	220	ZINC000032106887
125	440271	125	440271	86	6713972	29	51351654	221	ZINC000032106889
126	440271	126	440271	162	112106	85	6713972	222	ZINC000032106889
127	440235	127	440235	207	ZINC000015203840	ZINC000082111392	129618		
128	440235	128	440235	208	ZINC000015205540	46174147	154	129618	

Top 20 molecules three variants

```
Moleculas_3_Total <- cbind(Moleculas_3, Nombres_3)
Moleculas_3_Total <- Moleculas_3_Total[, c(1, 6, 2, 7, 3, 8, 4, 9, 5, 10)]

kable(Moleculas_3_Total[1:20, ], align = "ccccccccccc")
```

LASSO	LASSO.1	LASSO-MLR	LASSO-MLR.1	GA-MLR	GA-MLR.1	PLS	PLS.1	RSM	RSM.1
172	446101	172	446101	7	118707521	10	91820057	7	118707521
173	446101	173	446101	8	118707521	187	440271	8	118707521
174	446101	174	446101	9	118707521	192	440235	9	118707521
322	ZINC000198937472	322	ZINC000198937467	76325328	190	440235	16	76325328	
323	ZINC000198937473	323	ZINC000198937477	76325328	12	91820057	17	76325328	
324	ZINC000198937474	324	ZINC000198937487	76325328	172	446101	18	76325328	
412	ZINC000257346902	412	ZINC000257346903	ZINC000101361998	440271	220	198515		
413	ZINC000257346903	413	ZINC000257346903	ZINC000101361998	440235	221	198515		
414	ZINC000257346904	414	ZINC000257346903	5284365	362	ZINC000101133291	198515		
415	ZINC000257347635	415	ZINC000257347648	5284365	59	45933887	224	198514	
416	ZINC000257347636	416	ZINC000257347638	ZINC000101361998	5284365	225	198514		
417	ZINC000257347637	417	ZINC000257347648	46174147	68	45266741	415	ZINC000257347638	
418	ZINC000257347638	418	ZINC000257347639	134695375	426	ZINC000307564262	ZINC000257347638		
419	ZINC000257347639	419	ZINC000257347629	45357367	143	5284365	417	ZINC000257347638	
420	ZINC000257347640	420	ZINC000257347639	42612215	62	45357367	418	ZINC000257347639	
10	91820057	10	91820057	84	42612215	128	6713972	419	ZINC000257347639
11	91820057	11	91820057	121	7067772	129	6713972	420	ZINC000257347639
12	91820057	12	91820057	122	7067772	121	7067772	148	3052142
31	54758613	31	54758613	123	7067772	123	7067772	229	129618
32	54758613	32	54758613	127	6713972	43	51351654	231	129618

Applying an exponential consensus ranking (ECR) to the prediction to improve the selection of molecules

Recently, (Palacio-Rodríguez et al. 2019) proposed a new ranking function for virtual screening in molecular docking. According to the paper, this ranking function can be extrapolated into other ranking problems, such as this QSAR prediction. The function is:

$$\mathbf{P(i)} = \frac{\mathbf{1}}{\sigma} \sum_{\mathbf{j}} \exp(-\mathbf{r_i^j}) \tag{1}$$

Where i are the *i-th* molecule predicted with the ranking given by the *j-th* model (QSAR). σ is the expected value of the exponential distribution. This parameter takes account the number of molecules for each scoring function that will be considered (it can be seen as a threshold of the molecules that will be take account for the consensus). Finally, the sum gives the consensus ranking *P(i)* for the i-th molecule. Let's apply this strategy

ECR function

```
ECR <- function(X, Sig) {

    EC <- (exp(-X/Sig))/Sig

    output <- EC

    return(output)
}
```

ECR for the QSAR models

```
## Two variants
ECR_2 <- cbind(match(indices_2$Index, Moleculas_2_Total$LASSO), match(indices_2$Index,
    Moleculas_2_Total$`LASSO-MLR`), match(indices_2$Index, Moleculas_2_Total$`GA-MLR`),
    match(indices_2$Index, Moleculas_2_Total$PLS), match(indices_2$Index, Moleculas_2_Total$RSM))
colnames(ECR_2) <- c("LASSO", "LASSO-MLR", "GA-MLR", "PLS", "RSM")

ECR_Salida_2 <- t(as.data.frame(apply(ECR_2, 1, ECR, Sig = 20)))

Ranking_Consenso_2 <- as.data.frame(apply(ECR_Salida_2, 1, sum))
colnames(Ranking_Consenso_2) <- c("ECR")
row.names(Ranking_Consenso_2) <- as.vector(indices_2$Index)
Ranking_Consenso_2$ID <- as.vector(indices_2$Index)

Output_ECR_2 <- Ranking_Consenso_2[with(Ranking_Consenso_2, order(Ranking_Consenso_2$ECR,
    decreasing = TRUE)), ]

colnames(Output_ECR_2) <- c("ECR", "Index")
Output_ECR_2 <- left_join(Output_ECR_2, indices_2, by = "Index")
kable(Output_ECR_2[1:20, ], align = "ccc")
```

ECR	Index	Label
0.1239104	5	118707521
0.1177955	6	118707521

ECR	Index	Label
0.1096457	11	76325328
0.1063287	12	76325328
0.1039851	277	ZINC000257347638
0.0998926	8	91820057
0.0989136	278	ZINC000257347638
0.0988349	22	54758613
0.0987525	115	446101
0.0940895	279	ZINC000257347639
0.0932883	7	91820057
0.0925655	116	446101
0.0909432	21	54758613
0.0895007	280	ZINC000257347639
0.0889071	125	440271
0.0864304	215	ZINC000019893747
0.0824722	126	440271
0.0822214	216	ZINC000019893747
0.0804882	1	134695375
0.0784499	127	440235

```
## Three variants
ECR_3 <- cbind(match(indices_3$Index, Moleculas_3_Total$LASSO), match(indices_3$Index,
    Moleculas_3_Total$`LASSO-MLR`), match(indices_3$Index, Moleculas_3_Total$`GA-MLR`),
    match(indices_3$Index, Moleculas_3_Total$PLS), match(indices_3$Index, Moleculas_3_Total$RSM))
colnames(ECR_3) <- c("LASSO", "LASSO-MLR", "GA-MLR", "PLS", "RSM")

ECR_Salida_3 <- t(as.data.frame(apply(ECR_3, 1, ECR, Sig = 20)))

Ranking_Consenso_3 <- as.data.frame(apply(ECR_Salida_3, 1, sum))
colnames(Ranking_Consenso_3) <- c("ECR")
row.names(Ranking_Consenso_3) <- as.vector(indices_3$Index)
Ranking_Consenso_3$ID <- as.vector(indices_3$Index)

Ranking_Consenso_3 <- as.data.frame(apply(ECR_Salida_3, 1, sum))
colnames(Ranking_Consenso_3) <- c("ECR")
row.names(Ranking_Consenso_3) <- as.vector(indices_3$Index)
Ranking_Consenso_3$ID <- as.vector(indices_3$Index)

Output_ECR_3 <- Ranking_Consenso_3[with(Ranking_Consenso_3, order(Ranking_Consenso_3$ECR,
    decreasing = TRUE)), ]

colnames(Output_ECR_3) <- c("ECR", "Index")
Output_ECR_3 <- left_join(Output_ECR_3, indices_3, by = "Index")
kable(Output_ECR_3[1:20, ], align = "ccc")
```

ECR	Index	Label
0.1323101	172	446101
0.1108496	7	118707521
0.1054485	8	118707521
0.1003772	9	118707521
0.0945118	173	446101
0.0935430	16	76325328

ECR	Index	Label
0.0928678	10	91820057
0.0910086	322	ZINC000019893747
0.0891044	17	76325328
0.0880936	415	ZINC000257347638
0.0867441	174	446101
0.0846394	18	76325328
0.0837973	416	ZINC000257347638
0.0799349	12	91820057
0.0797105	417	ZINC000257347638
0.0779056	323	ZINC000019893747
0.0758229	418	ZINC000257347639
0.0744185	412	ZINC000257346903
0.0741063	324	ZINC000019893747
0.0740046	187	440271

Removing variants from the predictions (getting only the first variant per ligand according to each QSAR model)

Two variants

```
Moleculas_2_curado <- cbind(Moleculas_2_Total$LASSO[!duplicated(Moleculas_2_Total$LASSO.1)],
    Moleculas_2_Total$LASSO.1[!duplicated(Moleculas_2_Total$LASSO.1)], Moleculas_2_Total$`LASSO-MLR`[!du
    Moleculas_2_Total$`LASSO-MLR.1`[!duplicated(Moleculas_2_Total$`LASSO-MLR.1`)],
    Moleculas_2_Total$`GA-MLR`[!duplicated(Moleculas_2_Total$`GA-MLR.1`)], Moleculas_2_Total$`GA-MLR.1`
    Moleculas_2_Total$PLS[!duplicated(Moleculas_2_Total$PLS.1)], Moleculas_2_Total$PLS.1[!duplicated(Mo
    Moleculas_2_Total$RSM[!duplicated(Moleculas_2_Total$RSM.1)], Moleculas_2_Total$RSM.1[!duplicated(Mo

colnames(Moleculas_2_curado) <- c("Lasso_Index", "Lasso_Molecule", "Lasso-MLR_Index",
    "Lasso-MLR_Molecule", "GA-MLR_Index", "GA-MLR_Molecule", "PLS_Index", "PLS_Molecule",
    "RSM_Index", "RSM_Molecule")
kable(Moleculas_2_curado[1:20, ], align = "ccccccccccc")
```

Lasso_Index	Lasso_Molecule	Lasso-MLR_Index	Lasso-MLR_Molecule	GA-MLR_Index	GA-MLR_Molecule	PLS_Index	PLS_Molecule	RSM_Index	RSM_Molecule
115	446101	115	446101	5	118707521	22	54758613	5	118707521
215	ZINC000019893747	215	ZINC000019893747	7	76325328	125	440271	11	76325328
275	ZINC000257346903	275	ZINC000257346903	6	ZINC000101361998	91	91820057	147	198515
277	ZINC000257347638	277	ZINC000257347638	8	12600646	127	440235	149	198514
279	ZINC000257347639	279	ZINC000257347639	9	5284365	191	739	277	ZINC000257347638
7	91820057	7	91820057	35	46174147	72	12600646	279	ZINC000257347639
21	54758613	21	54758613	1	134695375	96	5284365	99	3052142
29	51351654	29	51351654	42	45357367	46	45266741	219	ZINC000032106887
125	440271	125	440271	55	42612215	55	42612215	221	ZINC000032106889
127	440235	127	440235	82	7067772	86	6713972	153	129618
1	134695375	1	134695375	85	6713972	82	7067772	17	57339290
3	129878031	3	129878031	162	112106	29	51351654	284	ZINC000307565252
5	118707521	5	118707521	207	ZINC000015235840	238	ZINC000082173892	285	ZINC000257346492
9	90659182	9	90659182	209	ZINC000015235946	35	46174147	141	312829
11	76325328	11	76325328	263	ZINC000106384102	4	134695375	155	127132
13	57376616	13	57376616	63	24779679	42	45357367	165	99461
15	57339292	15	57339292	49	44341508	208	ZINC000015235840	288	ZINC000100815672

Lasso_Index	Lasso_Molecule	Lasso-MLR_Index	Lasso-MLR_Molecule	GA-MLR_Index	GA-MLR_Molecule	PLS_Index	PLS_Molecule	RSM_Index	RSM_Molecule
17	57339290	17	57339290	163	102416	17	57339290	33	49802606
19	54758653	19	54758653	169	90220	161	112106	259	ZINC000106382957
23	52948856	23	52948856	173	82398	284	ZINC000307565252	261	ZINC000106382965

Three variants

```
Moleculas_3_curado <- cbind(Moleculas_3_Total$LASSO[!duplicated(Moleculas_3_Total$LASSO.1)],
    Moleculas_3_Total$LASSO.1[!duplicated(Moleculas_3_Total$LASSO.1)], Moleculas_3_Total$`LASSO-MLR`[!du
    Moleculas_3_Total$`LASSO-MLR.1`[!duplicated(Moleculas_3_Total$`LASSO-MLR.1`)],
    Moleculas_3_Total$`GA-MLR`[!duplicated(Moleculas_3_Total$`GA-MLR.1`)], Moleculas_3_Total$`GA-MLR.1`
    Moleculas_3_Total$PLS[!duplicated(Moleculas_3_Total$PLS.1)], Moleculas_3_Total$PLS.1[!duplicated(Mo
    Moleculas_3_Total$RSM[!duplicated(Moleculas_3_Total$RSM.1)], Moleculas_3_Total$RSM.1[!duplicated(Mo

colnames(Moleculas_3_curado) <- c("Lasso_Index", "Lasso_Molecule", "Lasso-MLR_Index",
    "Lasso-MLR_Molecule", "GA-MLR_Index", "GA-MLR_Molecule", "PLS_Index", "PLS_Molecule",
    "RSM_Index", "RSM_Molecule")
kable(Moleculas_3_curado[1:20, ], align = "cccccccccc")
```

Lasso_Index	Lasso_Molecule	Lasso-MLR_Index	Lasso-MLR_Molecule	GA-MLR_Index	GA-MLR_Molecule	PLS_Index	PLS_Molecule	RSM_Index	RSM_Molecule
172	446101	172	446101	7	118707521	10	91820057	7	118707521
322	ZINC000198937[illegible]	322	ZINC000198937[illegible]	147	76325328	187	440271	16	76325328
412	ZINC000257346[illegible]	412	ZINC000257346[illegible]	307	ZINC000101369298	190	440235	220	198515
415	ZINC000257347[illegible]	415	ZINC000257347[illegible]	648	5284365	172	446101	224	198514
418	ZINC000257347[illegible]	418	ZINC000257347[illegible]	650	46174147	362	ZINC000101146391	465	ZINC000257347638
10	91820057	10	91820057	1	134695375	59	45933887	418	ZINC000257347639
31	54758613	31	54758613	62	45357367	144	5284365	148	3052142
43	51351654	43	51351654	83	42612215	68	45266741	229	129618
187	440271	187	440271	121	7067772	426	ZINC000307565252	652	ZINC000032106887
190	440235	190	440235	127	6713972	62	45357367	331	ZINC000032106889
1	134695375	1	134695375	130	6420074	128	6713972	25	57339290
4	129878031	4	129878031	159	2733335	121	7067772	322	ZINC000019893747
7	118707521	7	118707521	242	112106	43	51351654	425	ZINC000307565252
13	90659182	13	90659182	310	ZINC000152038340	58	6420074	409	ZINC000257346492
16	76325328	16	76325328	313	ZINC000152059946	211	46174147	211	312829
19	57376616	19	57376616	394	ZINC000106389102	232	2733335	232	127132
22	57339292	22	57339292	94	24779679	1	134695375	247	99461
25	57339290	25	57339290	73	44341508	310	ZINC000152038340	352	ZINC000100815672
28	54758653	28	54758653	244	102416	25	57339290	412	ZINC000257346903
34	52948856	34	52948856	253	90220	376	ZINC000102974493	734	52944904

ECR two variants

```
Output_ECR_2_curado <- cbind(Output_ECR_2$ECR[!duplicated(Output_ECR_2$Label)], Output_ECR_2$Index[!dupl
    Output_ECR_2$Label[!duplicated(Output_ECR_2$Label)])
colnames(Output_ECR_2_curado) <- c("ECR", "Index", "Molecule")
kable(Output_ECR_2_curado[1:20, ], align = "ccc")
```

ECR	Index	Molecule
0.123910394377029	5	118707521
0.10964571955221	11	76325328
0.103985062328752	277	ZINC000257347638
0.0998926021126605	8	91820057
0.0988349376415874	22	54758613
0.0987524816241701	115	446101
0.0940895291726436	279	ZINC000257347639
0.0889070691073631	125	440271
0.0864304316353736	215	ZINC000019893747
0.0804882060518684	1	134695375
0.0784499443559222	127	440235
0.078082494327749	275	ZINC000257346903
0.0686878753936395	29	51351654
0.0670244674612893	72	12600646
0.0628736414964675	96	5284365
0.0573643004621222	35	46174147
0.0542501126233325	55	42612215
0.0493435166015162	17	57339290
0.0480181977442779	42	45357367
0.0469106652810004	82	7067772

ECR three variants

```
Output_ECR_3_curado <- cbind(Output_ECR_3$ECR[!duplicated(Output_ECR_3$Label)], Output_ECR_3$Index[!dupl
    Output_ECR_3$Label[!duplicated(Output_ECR_3$Label)])
colnames(Output_ECR_3_curado) <- c("ECR", "Index", "Molecule")
kable(Output_ECR_3_curado[1:20, ], align = "ccc")
```

ECR	Index	Molecule
0.132310100606202	172	446101
0.110849573814621	7	118707521
0.0935430282121767	16	76325328
0.0928677690607199	10	91820057
0.0910086285115274	322	ZINC000019893747
0.0880936480494446	415	ZINC000257347638
0.0758229257026819	418	ZINC000257347639
0.0744184874508252	412	ZINC000257346903
0.0740045912479688	187	440271
0.0656928404795316	190	440235
0.0596569652183867	1	134695375
0.0592038919518806	144	5284365
0.0516837253833607	43	51351654
0.0500240637735259	62	45357367
0.0450734356889405	31	54758613
0.0448522533185709	54	46174147
0.0417817679463113	121	7067772
0.0400215026402701	128	6713972
0.0352496273855809	367	ZINC000101361998
0.0352353230855984	220	198515

Removing LASSO predictions since this QSAR and LASSO-MLR have the same output prediction and are biasing ECR algorithm

Two variants

```
ECR_2_final <- cbind(match(indices_2$Index, Moleculas_2_Total$`LASSO-MLR`), match(indices_2$Index,
    Moleculas_2_Total$`GA-MLR`), match(indices_2$Index, Moleculas_2_Total$PLS), match(indices_2$Index,
    Moleculas_2_Total$RSM))
colnames(ECR_2_final) <- c("LASSO-MLR", "GA-MLR", "PLS", "RSM")

ECR_Salida_2_final <- t(as.data.frame(apply(ECR_2_final, 1, ECR, Sig = 20)))

Ranking_Consenso_2_final <- as.data.frame(apply(ECR_Salida_2_final, 1, sum))
colnames(Ranking_Consenso_2_final) <- c("ECR")
row.names(Ranking_Consenso_2_final) <- as.vector(indices_2$Index)
Ranking_Consenso_2_final$ID <- as.vector(indices_2$Index)

Output_ECR_2_final <- Ranking_Consenso_2_final[with(Ranking_Consenso_2_final, order(Ranking_Consenso_2_
    decreasing = TRUE)), ]

colnames(Output_ECR_2_final) <- c("ECR", "Index")
Output_ECR_2_final <- left_join(Output_ECR_2_final, indices_2, by = "Index")

Output_ECR_2_final <- cbind(Output_ECR_2_final$ECR[!duplicated(Output_ECR_2_final$Label)],
    Output_ECR_2_final$Index[!duplicated(Output_ECR_2_final$Label)], Output_ECR_2_final$Label[!duplicat
colnames(Output_ECR_2_final) <- c("ECR", "Index", "Molecule")

kable(Output_ECR_2_final[1:20, ], align = "ccc")
```

ECR	Index	Molecule
0.109585154534019	5	118707521
0.0979172051475202	11	76325328
0.074005672452017	22	54758613
0.0724520203079591	8	91820057
0.0687506578428166	277	ZINC000257347638
0.0675363225099268	125	440271
0.0663460395106793	72	12600646
0.0629913185963106	1	134695375
0.0626693029245443	96	5284365
0.0622081215915549	279	ZINC000257347639
0.0591128931831972	127	440235
0.0530496211371537	35	46174147
0.0526628308044291	55	42612215
0.0511910103991344	115	446101
0.0464991779285494	82	7067772
0.0453286089440921	86	6713972
0.0450695477565888	29	51351654
0.044977694613017	42	45357367
0.0433950328141207	215	ZINC000019893747
0.0406548194289939	17	57339290

Three variants

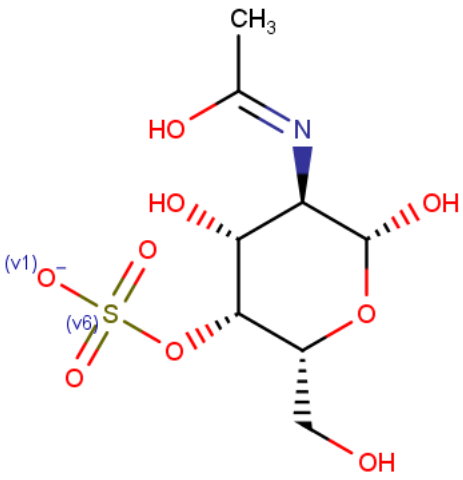
```
ECR_3_final <- cbind(match(indices_3$Index, Moleculas_3_Total$`LASSO-MLR`), match(indices_3$Index,
    Moleculas_3_Total$`GA-MLR`), match(indices_3$Index, Moleculas_3_Total$PLS), match(indices_3$Index,
    Moleculas_3_Total$RSM))
colnames(ECR_3_final) <- c("LASSO-MLR", "GA-MLR", "PLS", "RSM")

ECR_Salida_3_final <- t(as.data.frame(apply(ECR_3_final, 1, ECR, Sig = 20)))

Ranking_Consenso_3_final <- as.data.frame(apply(ECR_Salida_3_final, 1, sum))
colnames(Ranking_Consenso_3_final) <- c("ECR")
row.names(Ranking_Consenso_3_final) <- as.vector(indices_3$Index)
Ranking_Consenso_3_final$ID <- as.vector(indices_3$Index)

Output_ECR_3_final <- Ranking_Consenso_3_final[with(Ranking_Consenso_3_final, order(Ranking_Consenso_3_
    decreasing = TRUE)), ]

colnames(Output_ECR_3_final) <- c("ECR", "Index")
Output_ECR_3_final <- left_join(Output_ECR_3_final, indices_3, by = "Index")

Output_ECR_3_final <- cbind(Output_ECR_3_final$ECR[!duplicated(Output_ECR_3_final$Label)],
    Output_ECR_3_final$Index[!duplicated(Output_ECR_3_final$Label)], Output_ECR_3_final$Label[!duplicat
colnames(Output_ECR_3_final) <- c("ECR", "Index", "Molecule")

kable(Output_ECR_3_final[1:20, ], align = "ccc")
```

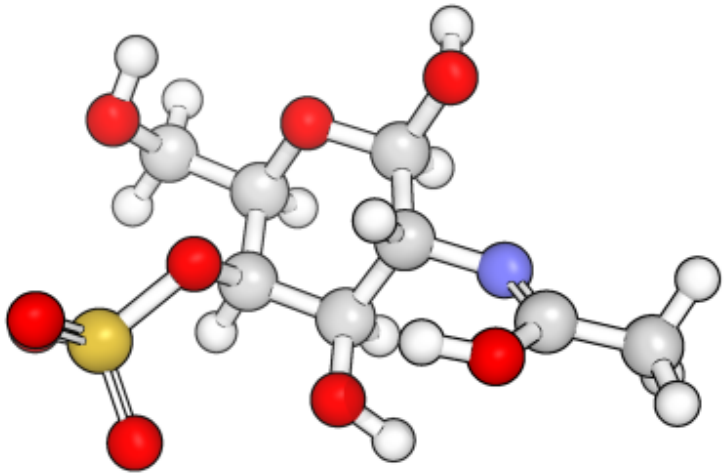
ECR	Index	Molecule
0.102987715498939	7	118707521
0.0877188203235018	16	76325328
0.0847486293811661	172	446101
0.0704013208548588	10	91820057
0.0596793514049593	187	440271
0.0591908290240155	144	5284365
0.0577671150638129	415	ZINC000257347638
0.054279319942693	192	440235
0.0500720908576283	322	ZINC000019893747
0.0497206368646311	418	ZINC000257347639
0.0492358429492832	62	45357367
0.0490445665270496	1	134695375
0.0436763660257705	54	46174147
0.0417405127001481	121	7067772
0.0399924305596604	128	6713972
0.0391840829648896	412	ZINC000257346903
0.0352496270434173	367	ZINC000101361998
0.0352348647812116	220	198515
0.0350401711984567	43	51351654
0.031900896754871	362	ZINC000101136391

Top 3 molecules for each QSAR method

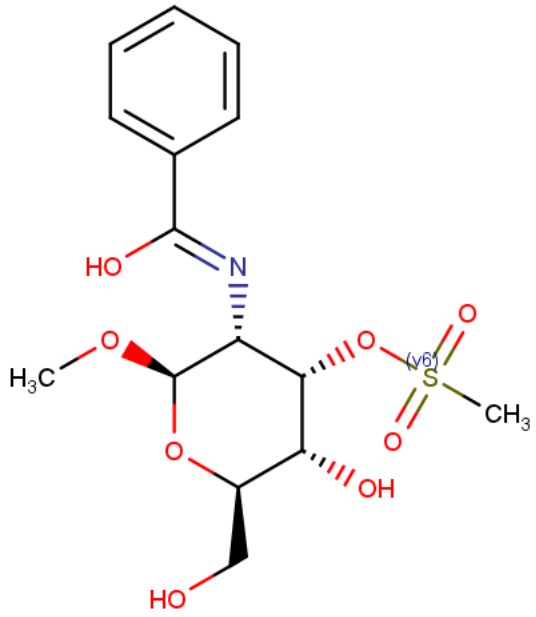
Two variants // LASSO & MLR-LASSO

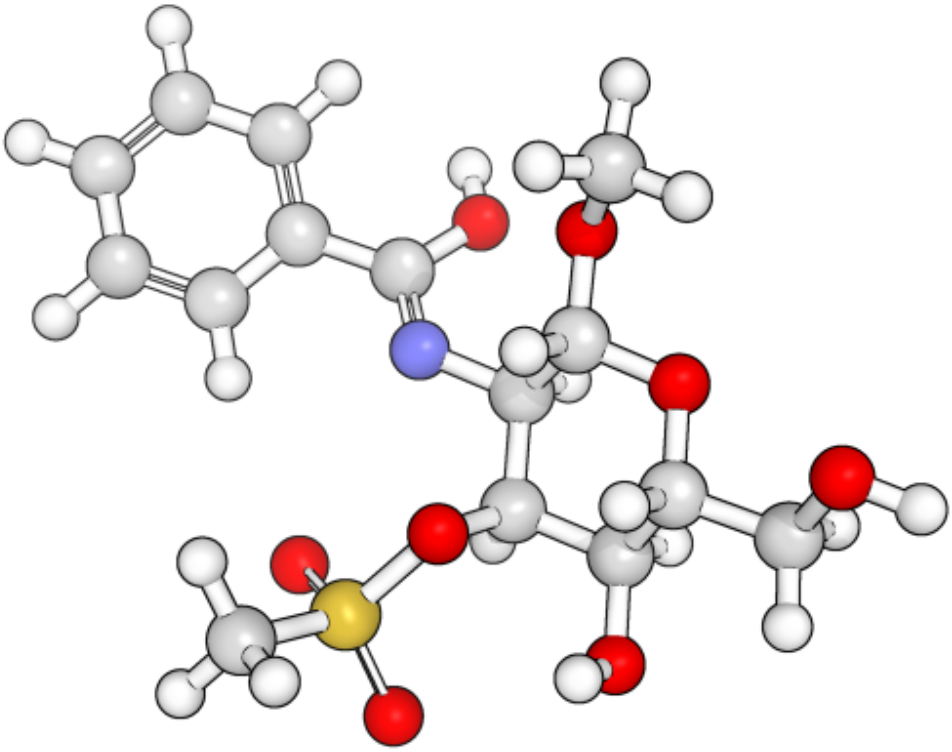
Molecule 115



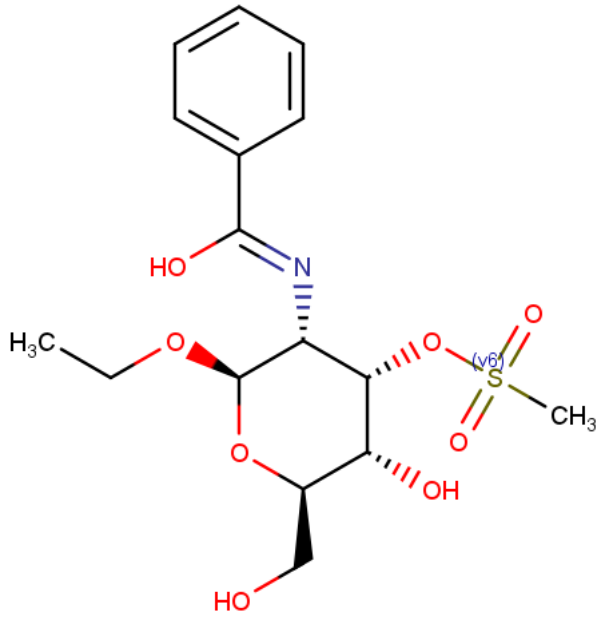


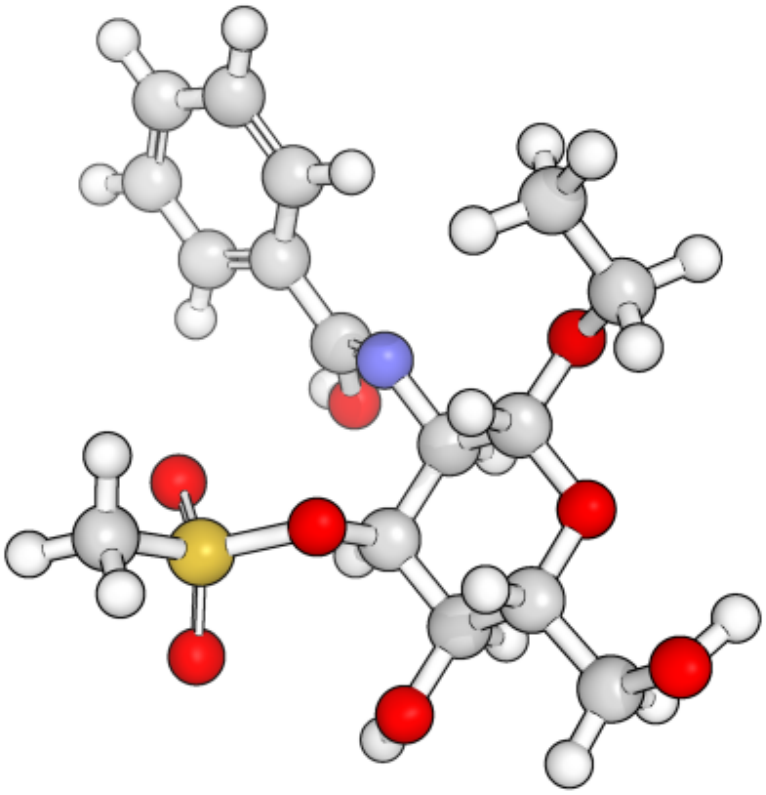
Molecule 215





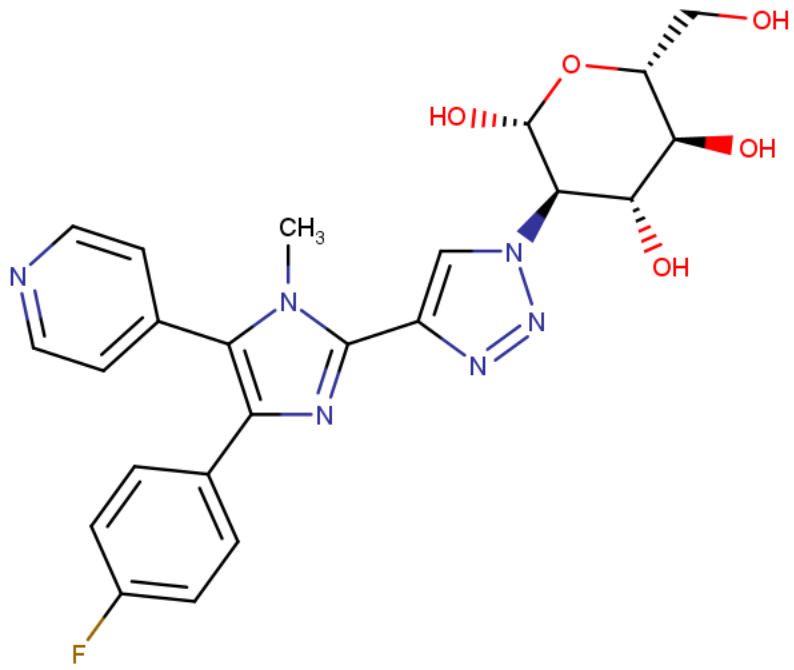
Molecule 275

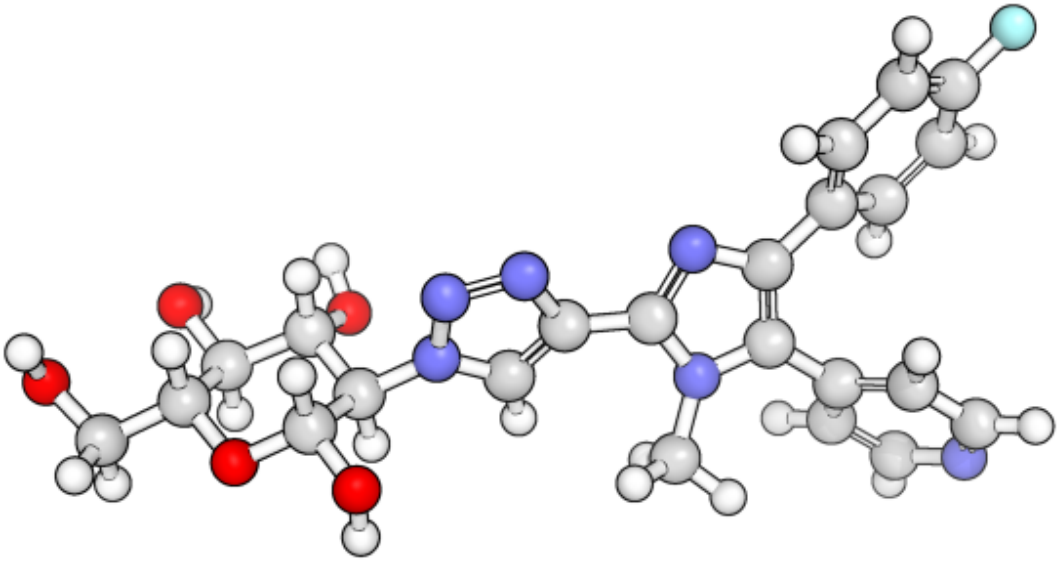

HO
N
H3C
O
O
O
(v6)
S
O
O
CH3
O
OH
HO



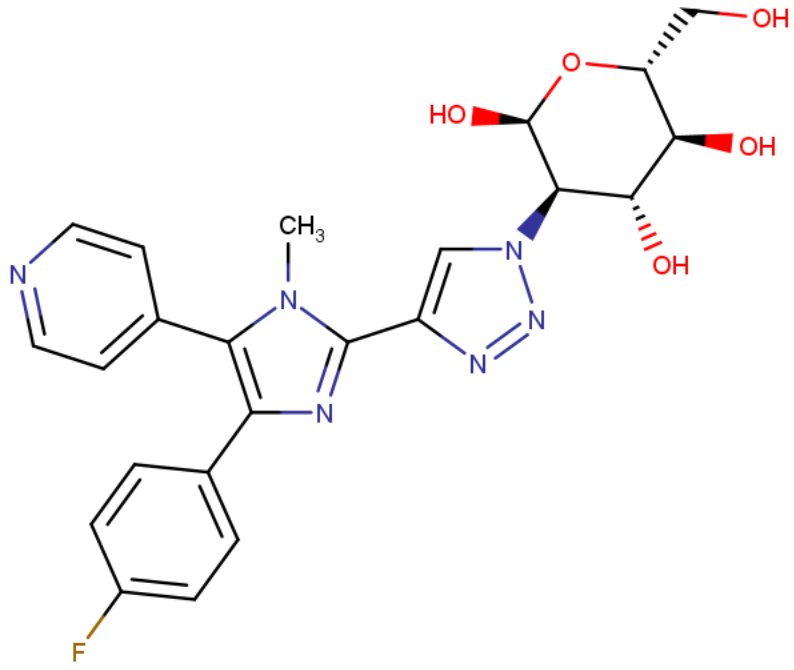
Two variants // GA-MLR

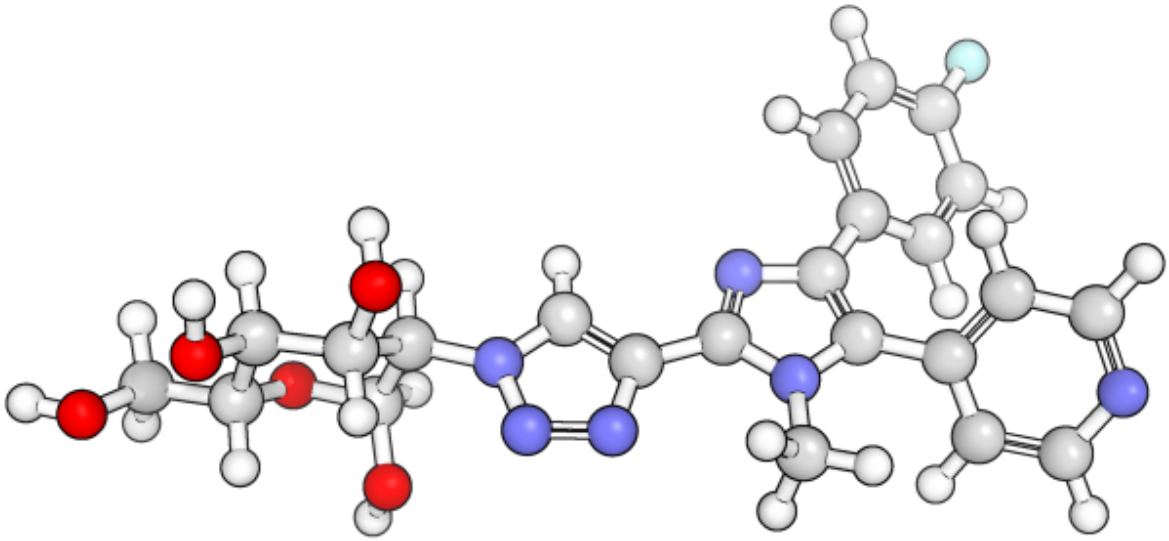
Molecule 5



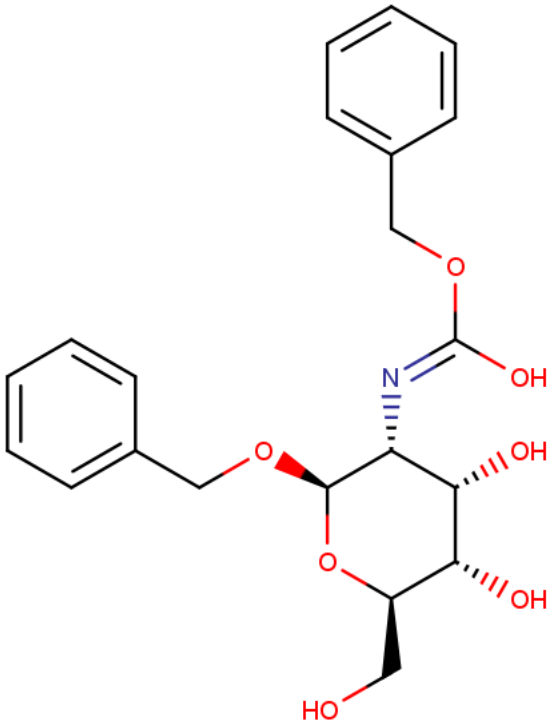


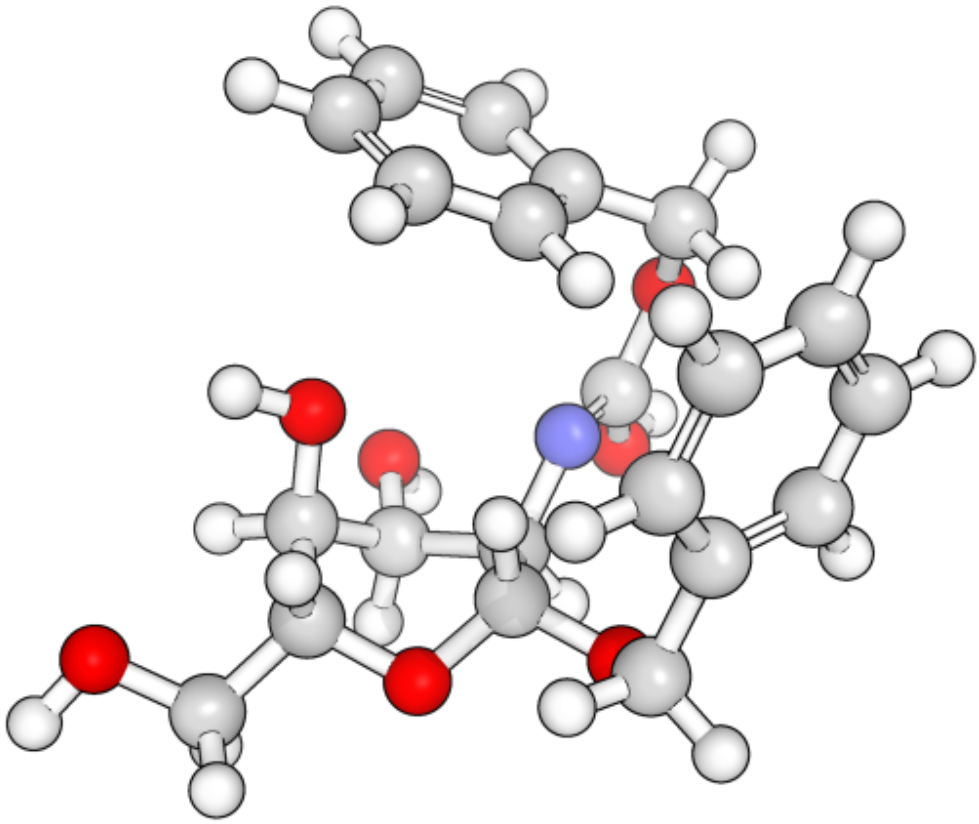
Molecule 11





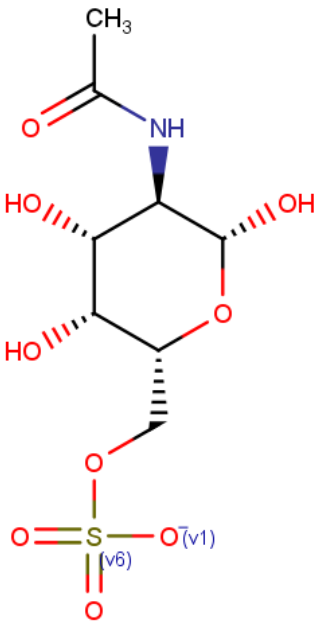
Molecule 246

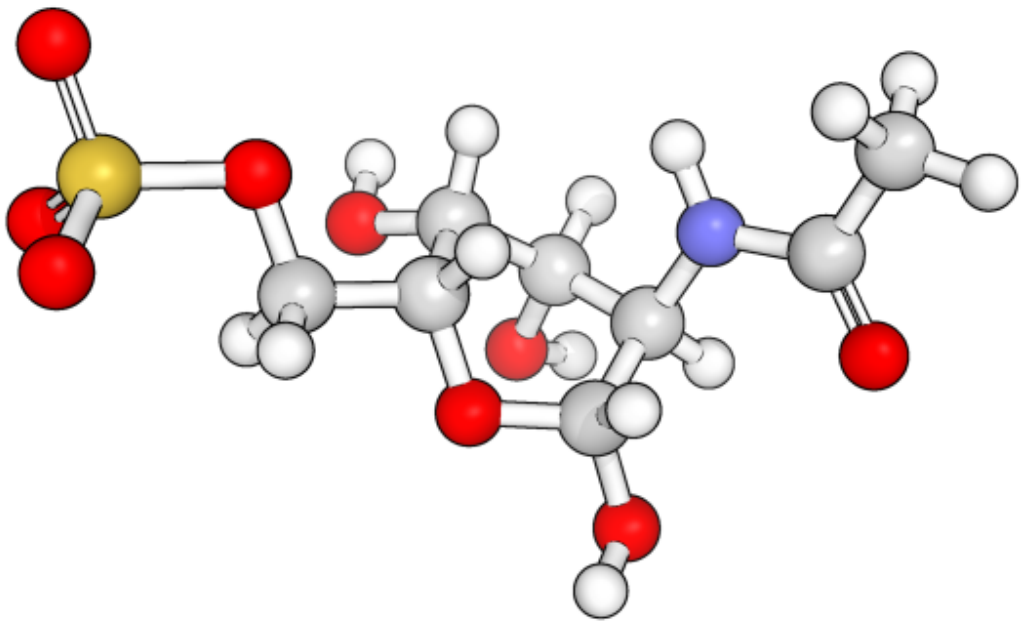

O
N
OH
O
OH
O
OH
HO



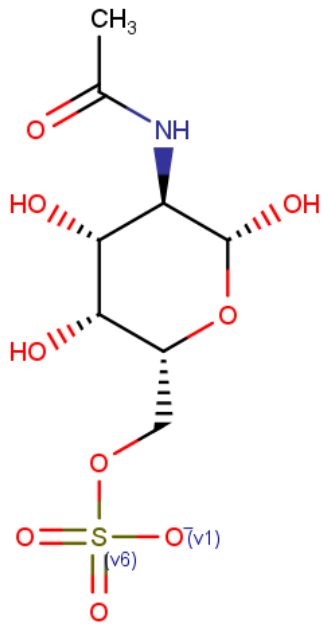
Two variants // PLS

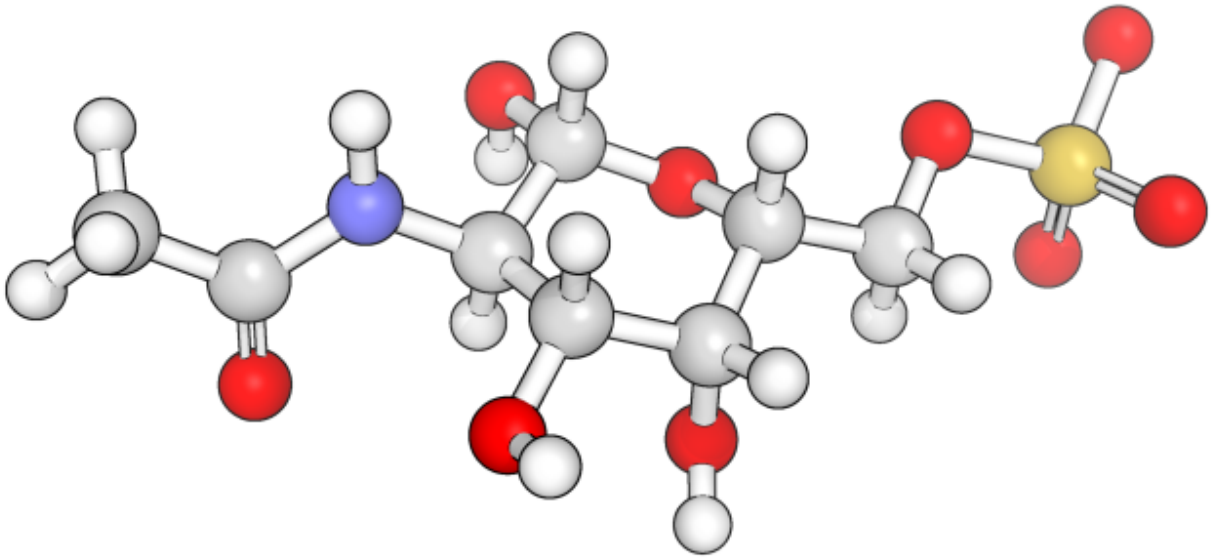
Molecule 22



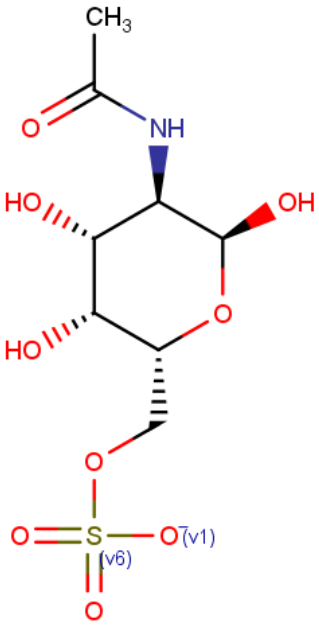


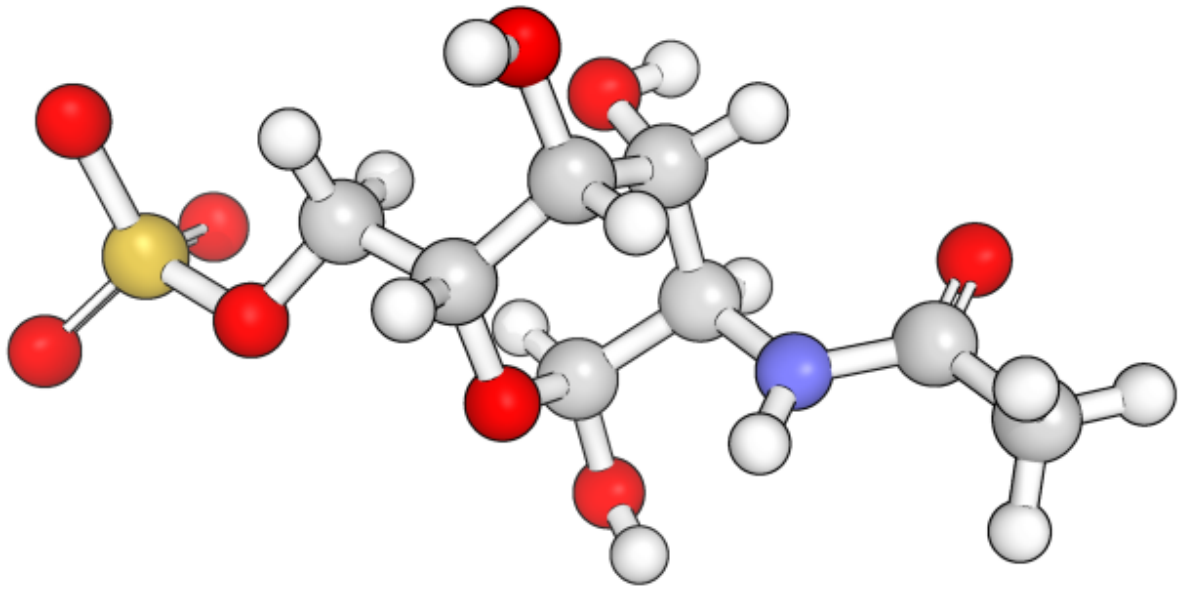
Molecule 125





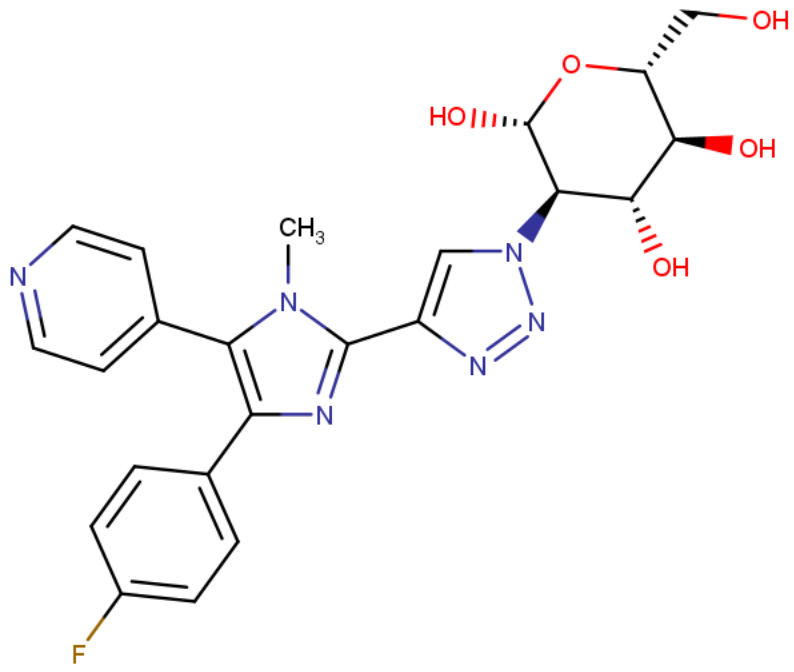
Molecule 8

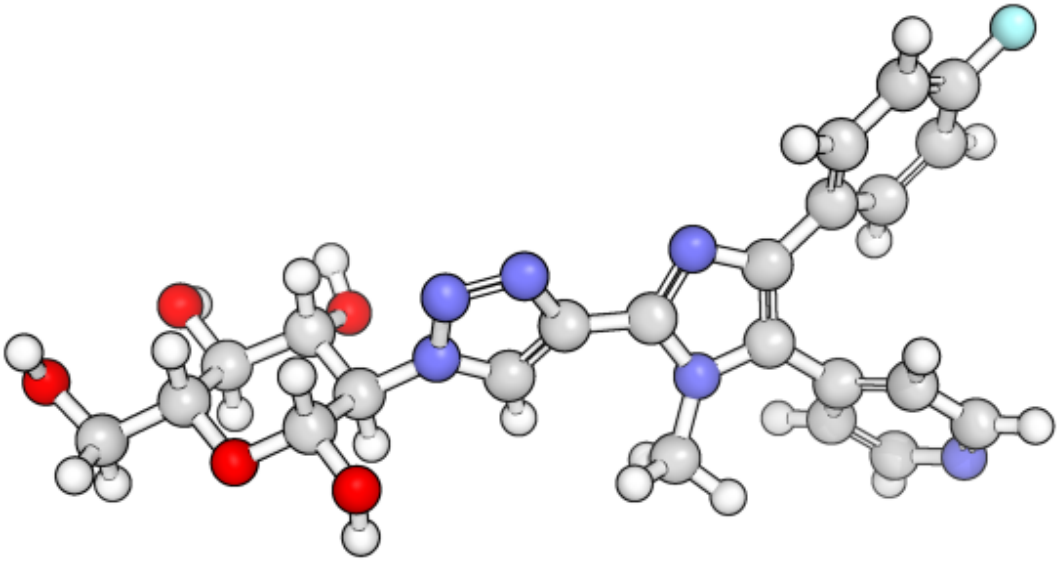

CH3
O
NH
HO
OH
HO
O
O
O
S
O(v1)
(v6)
O



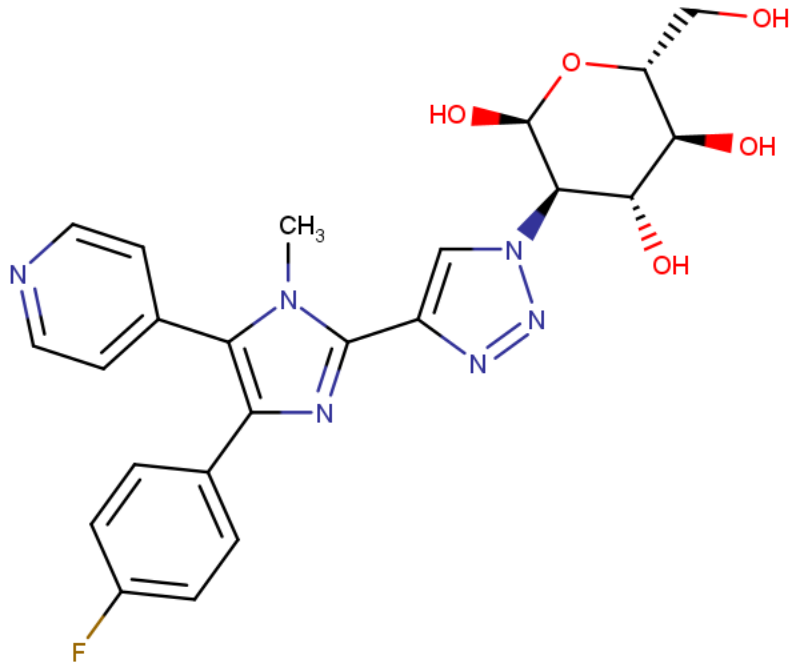
Two variants // RSM

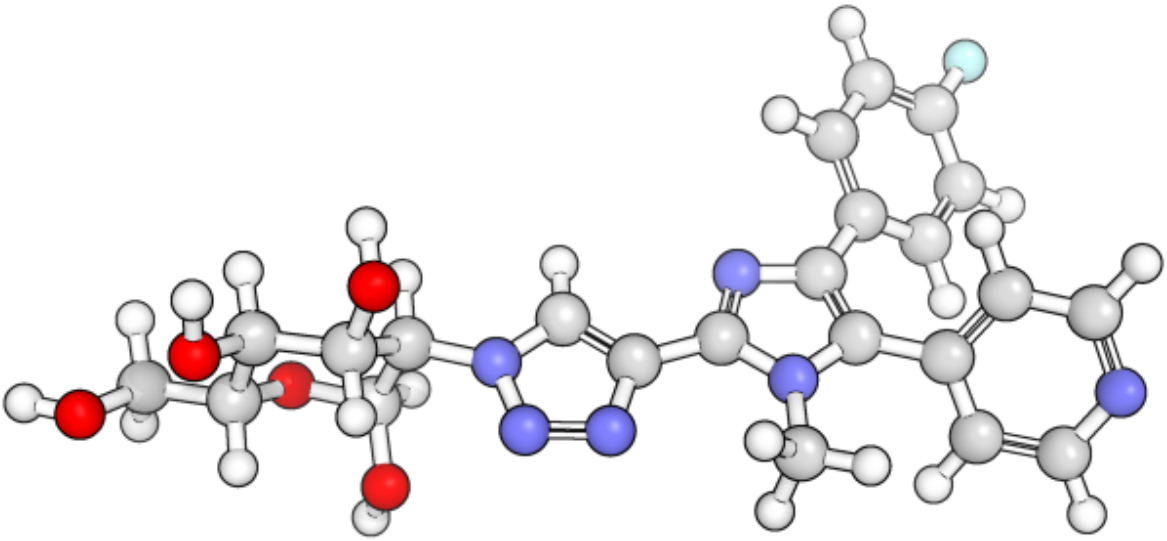
Molecule 5



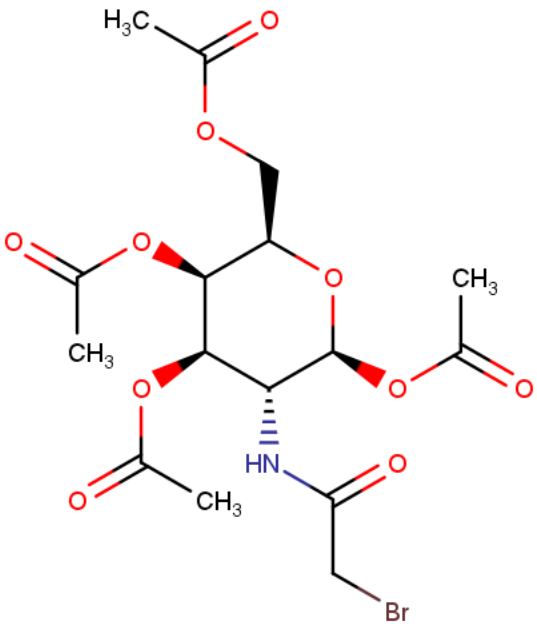


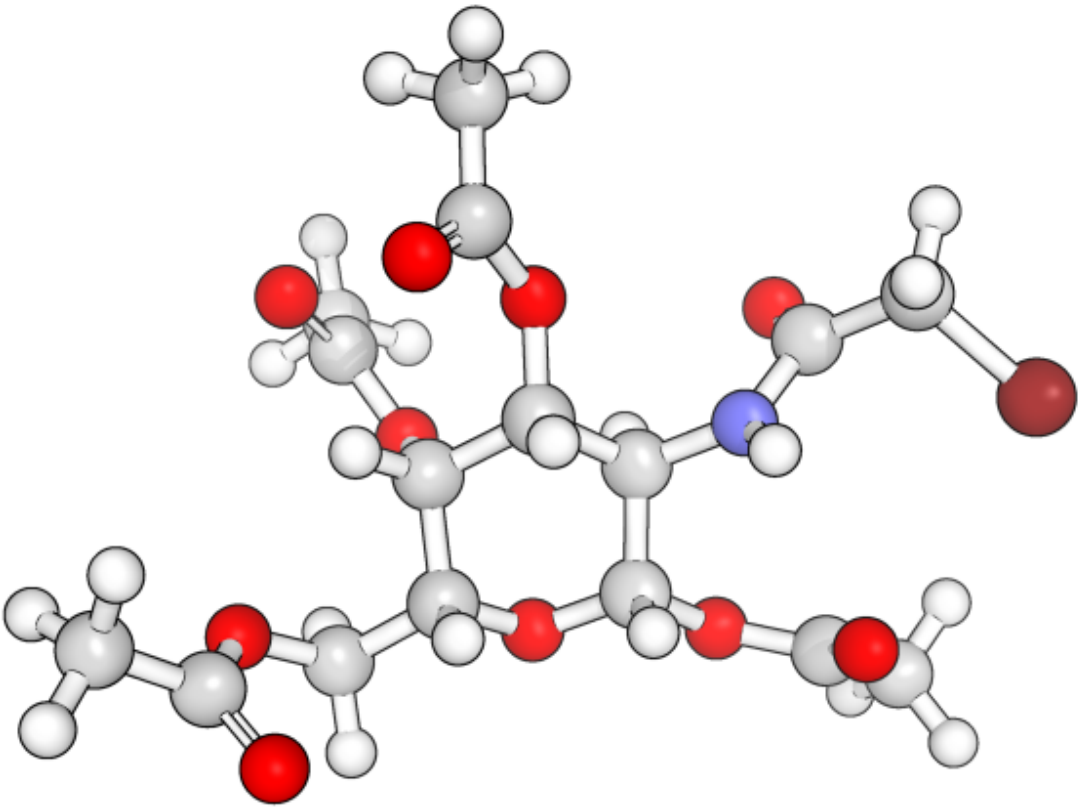
Molecule 11





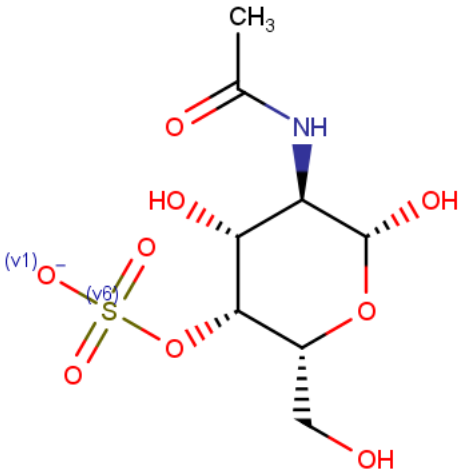
Molecule 147

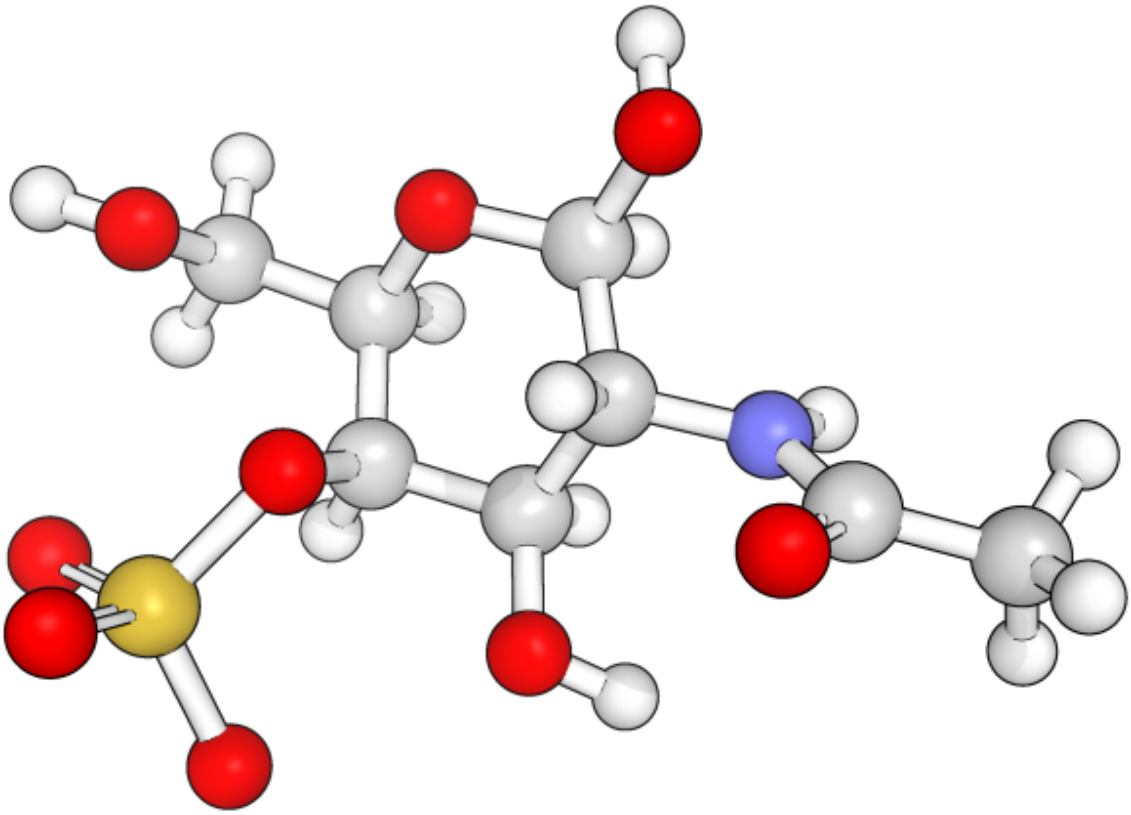




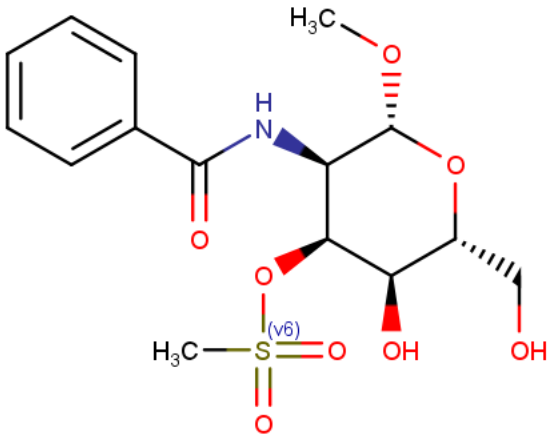
Three variants // LASSO & MLR-LASSO

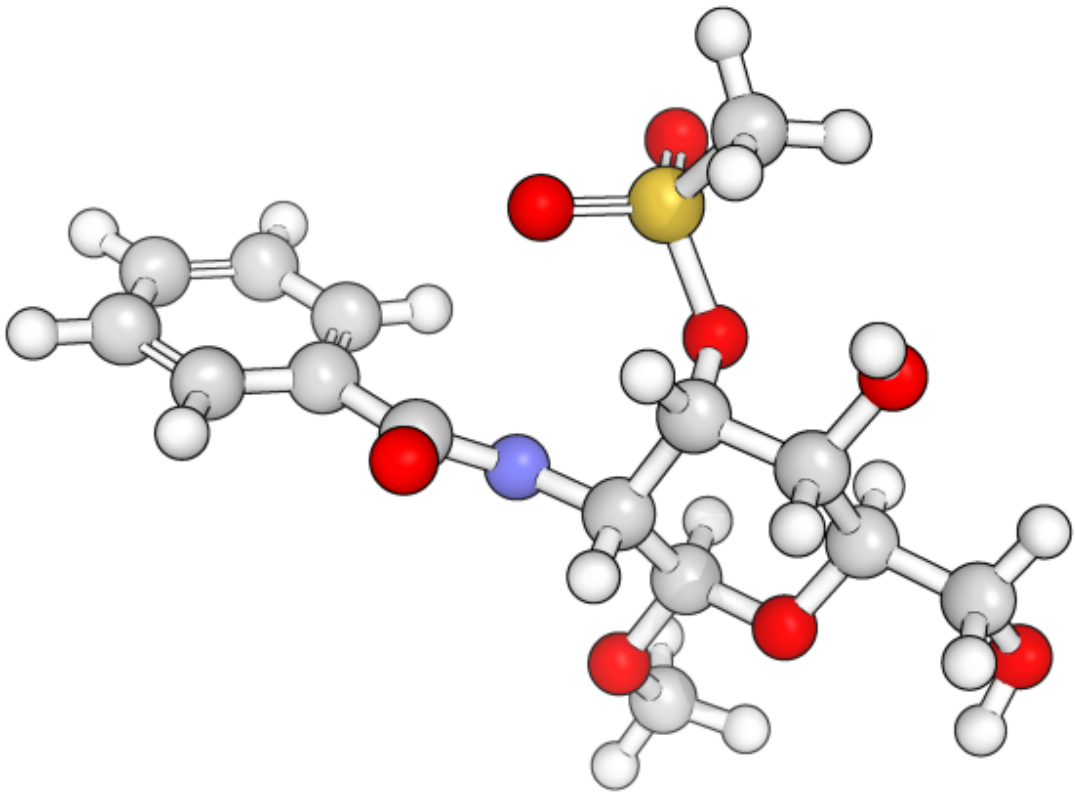
Molecule 172



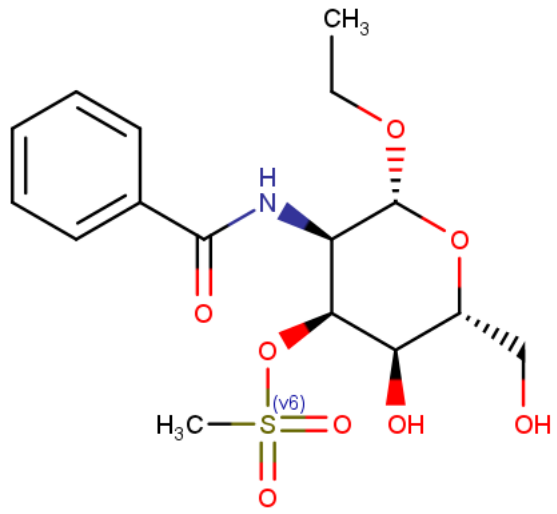


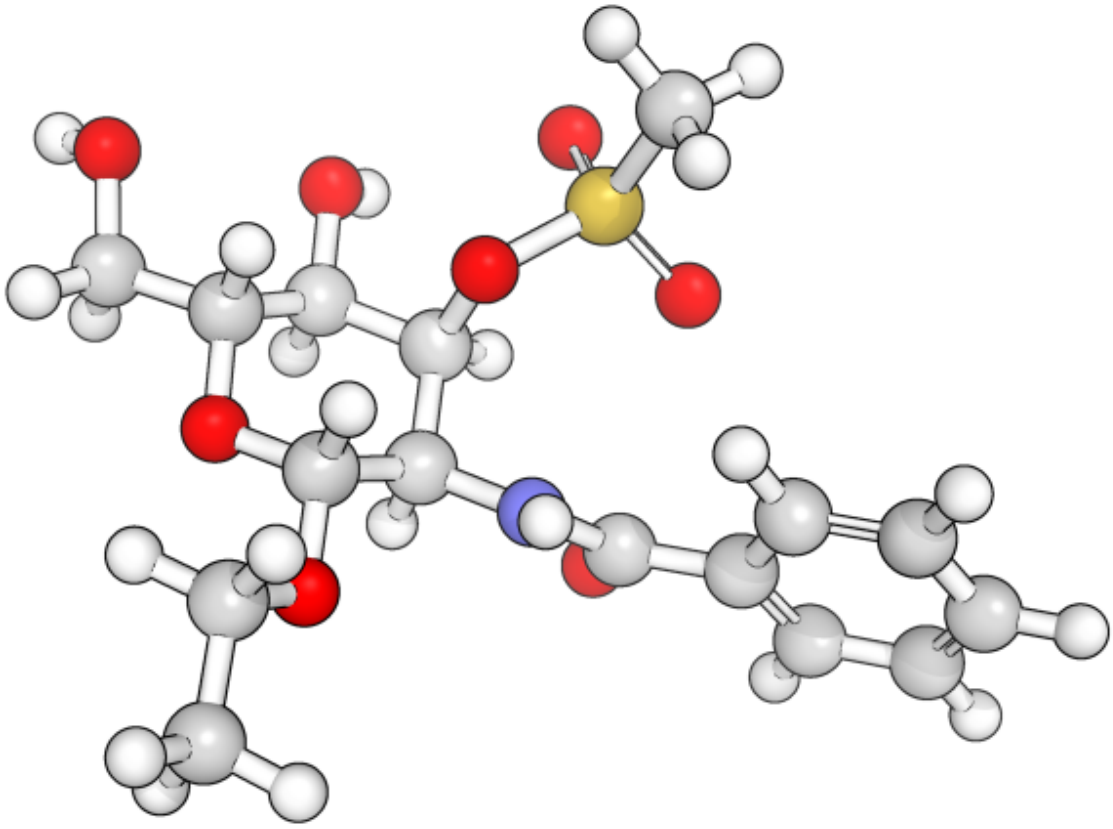
Molecule 322





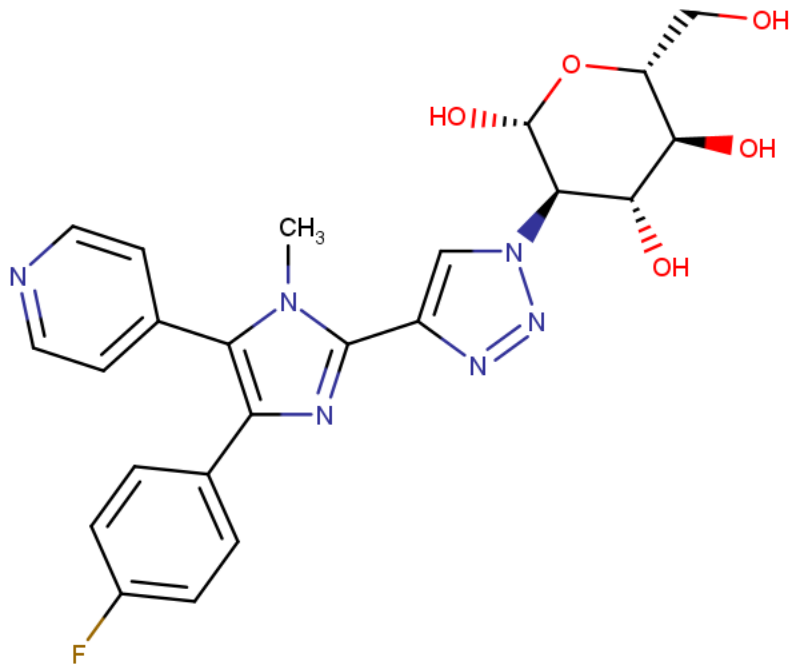
Molecule 412

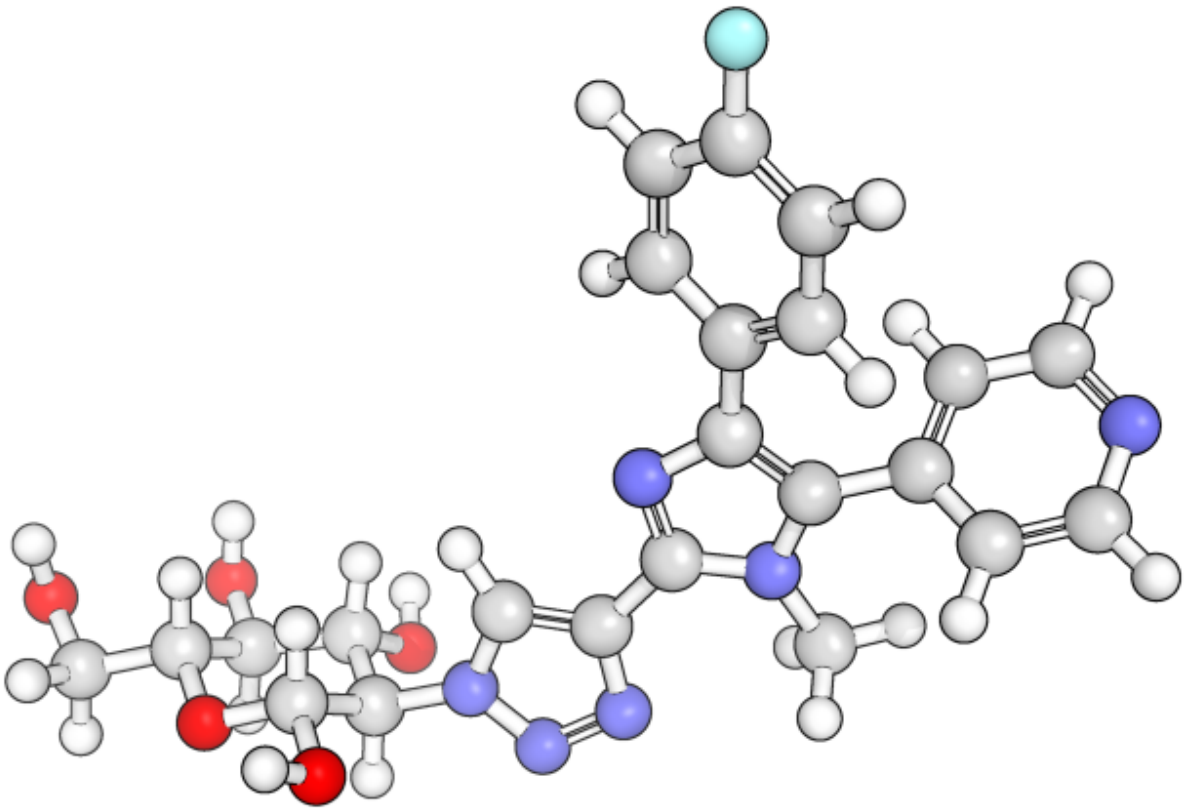




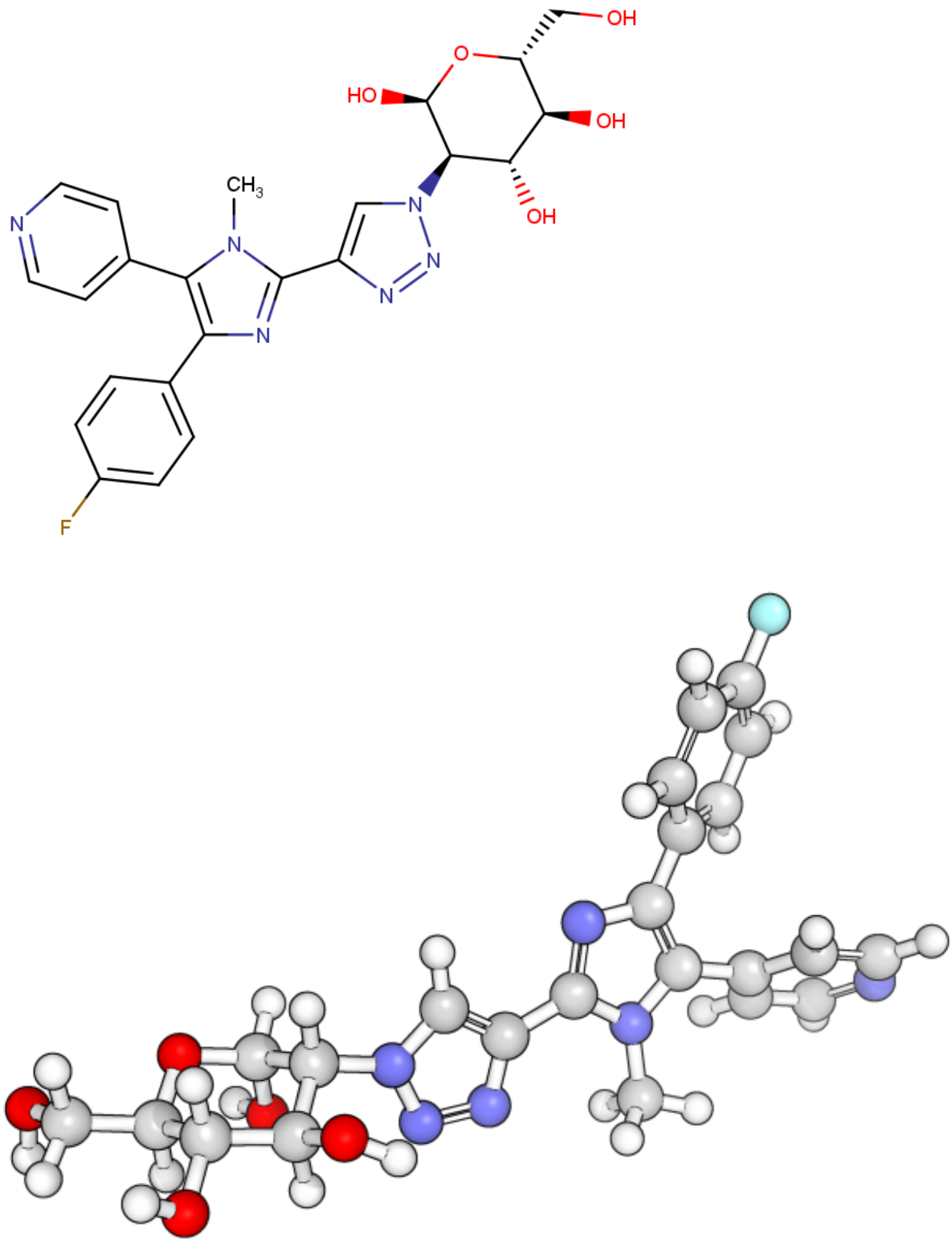
Three variants // GA-MLR

Molecule 7

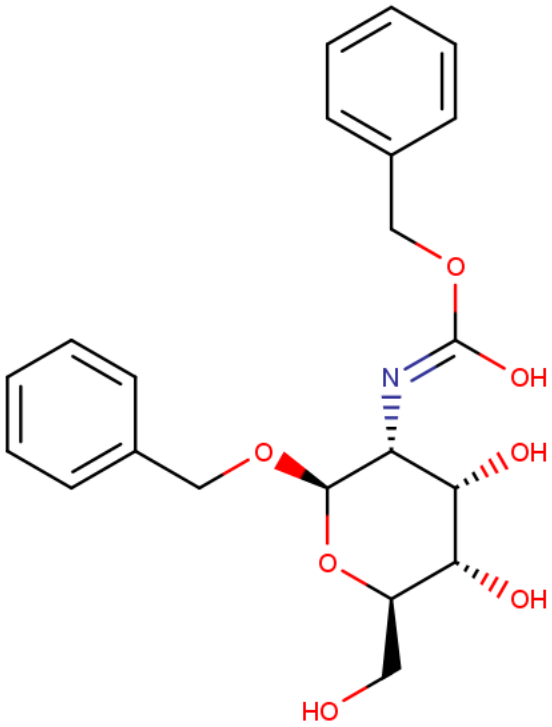


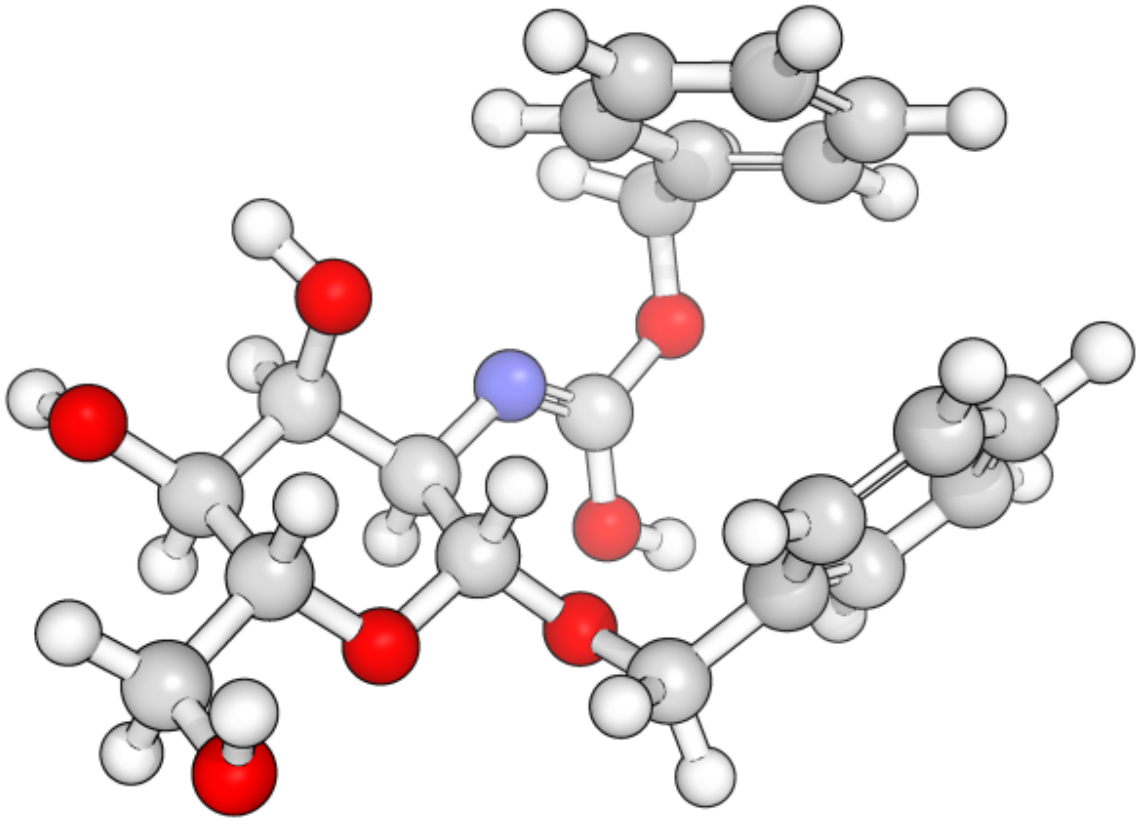


Molecule 16



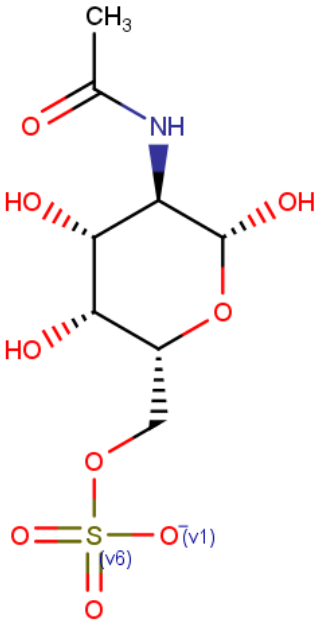
Molecule 367

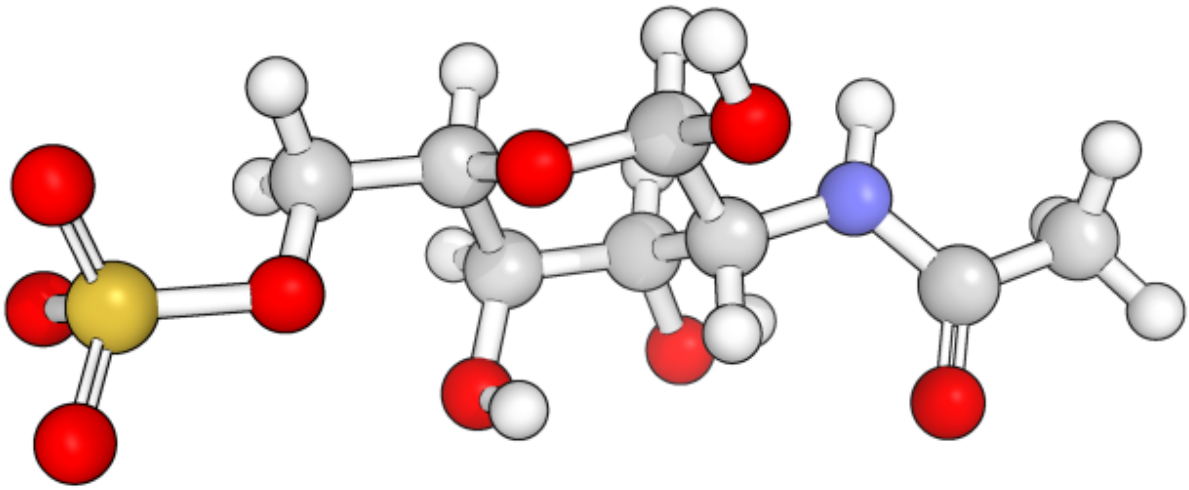

O
N
OH
O
OH
O
OH
HO



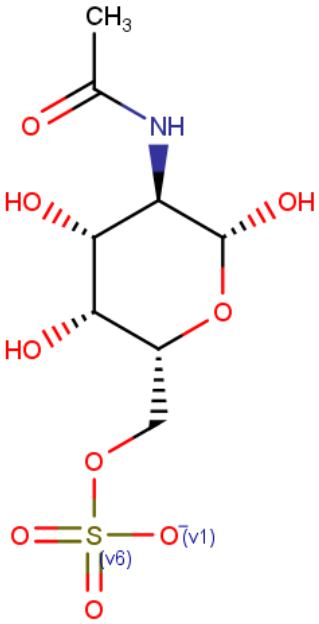
Three variants // PLS

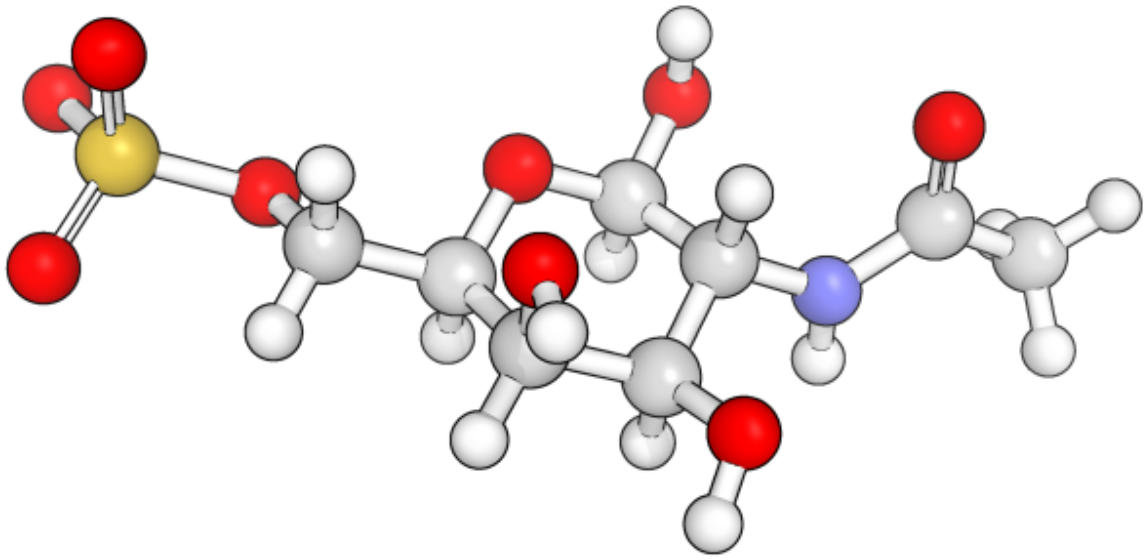
Molecule 10



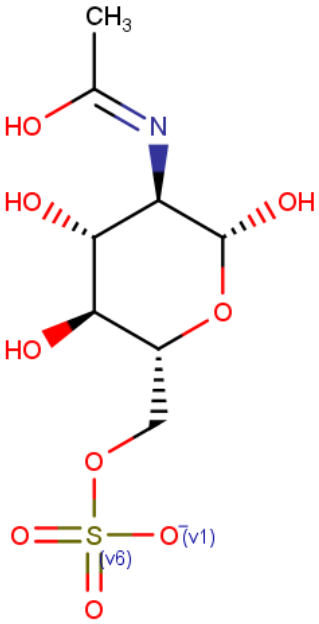


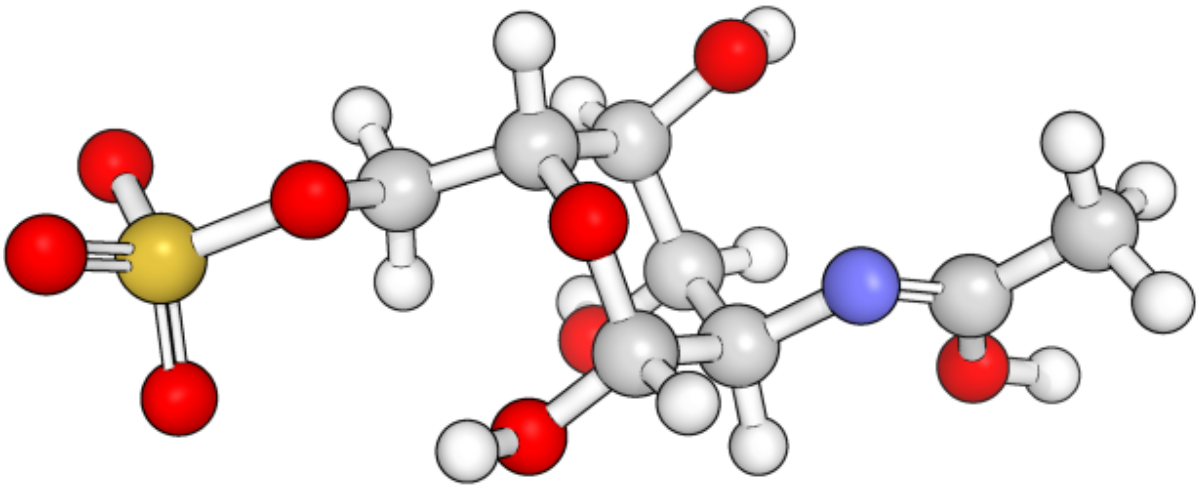
Molecule 187





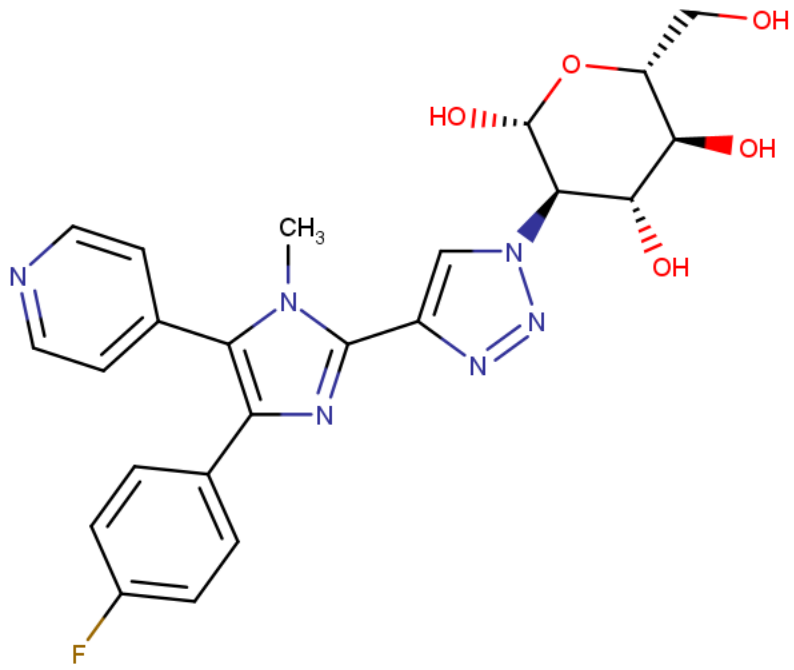
Molecule 192

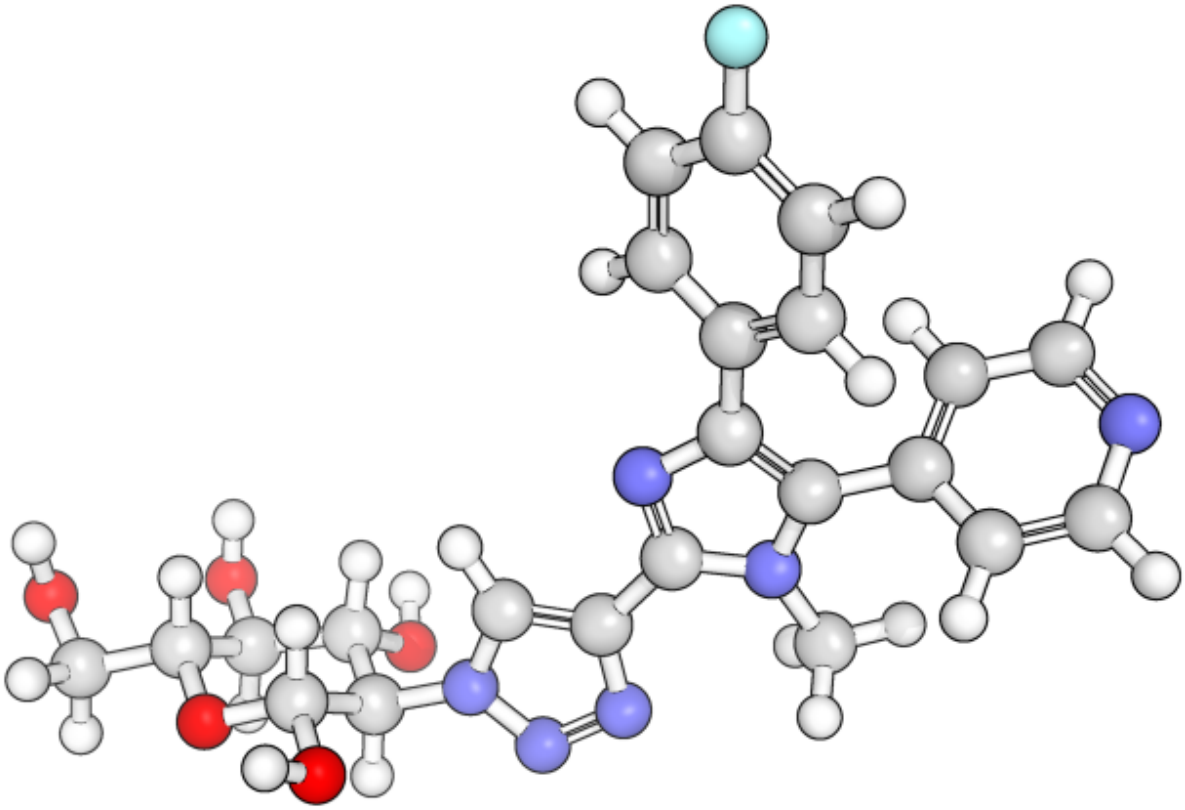

CH3
HO
N
HO
OH
O
HO
O
O
S
O(v1)
(v6)
O



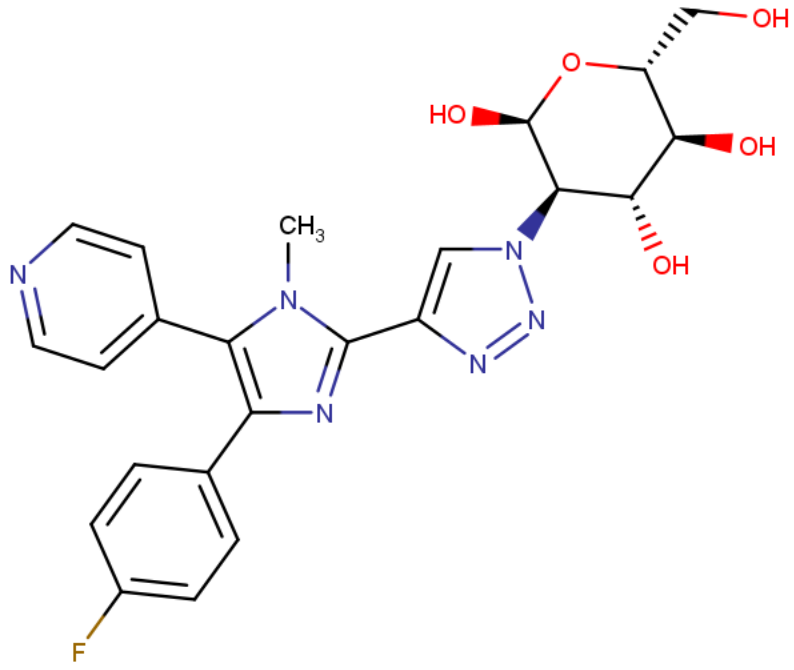
Three variants // RSM

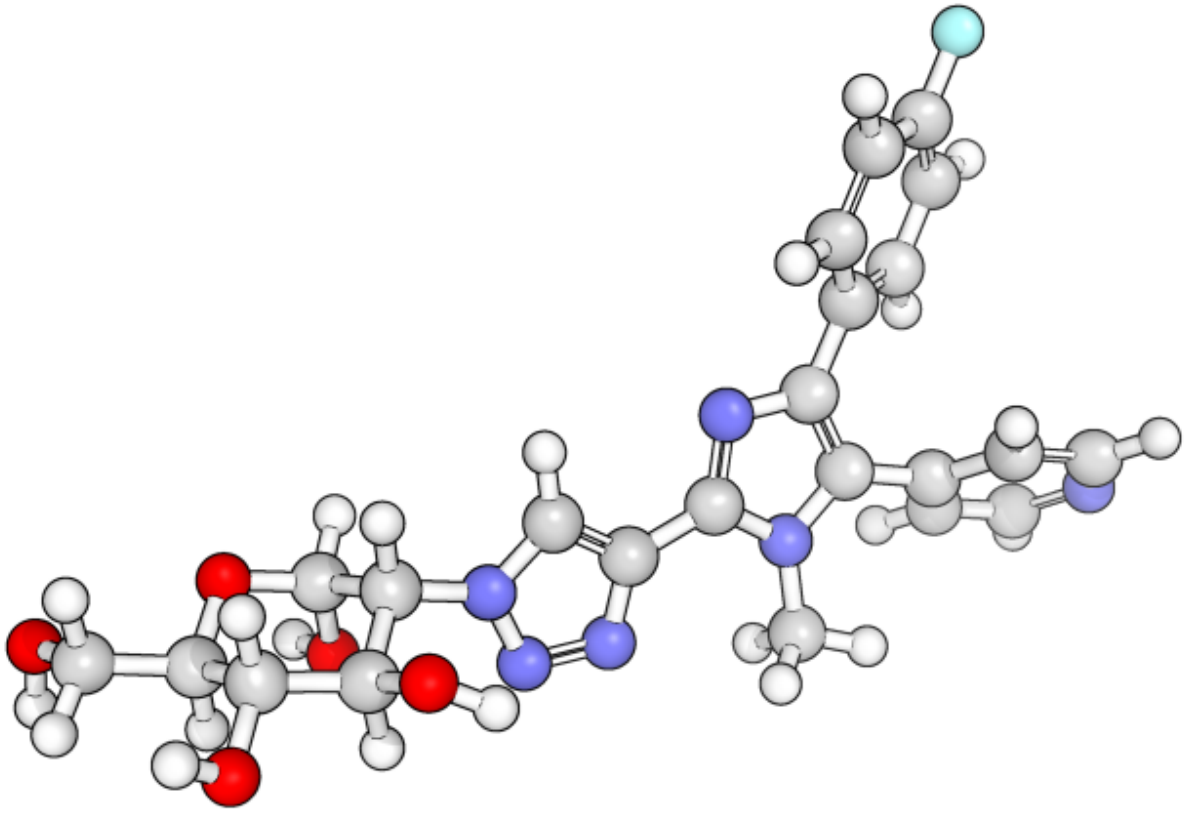
Molecule 7



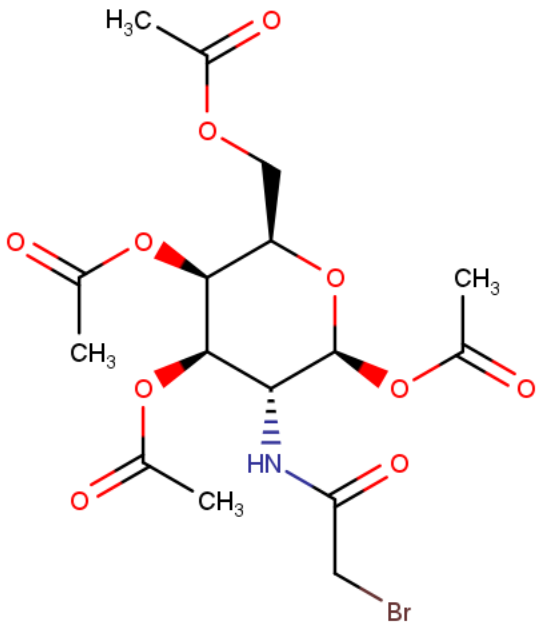


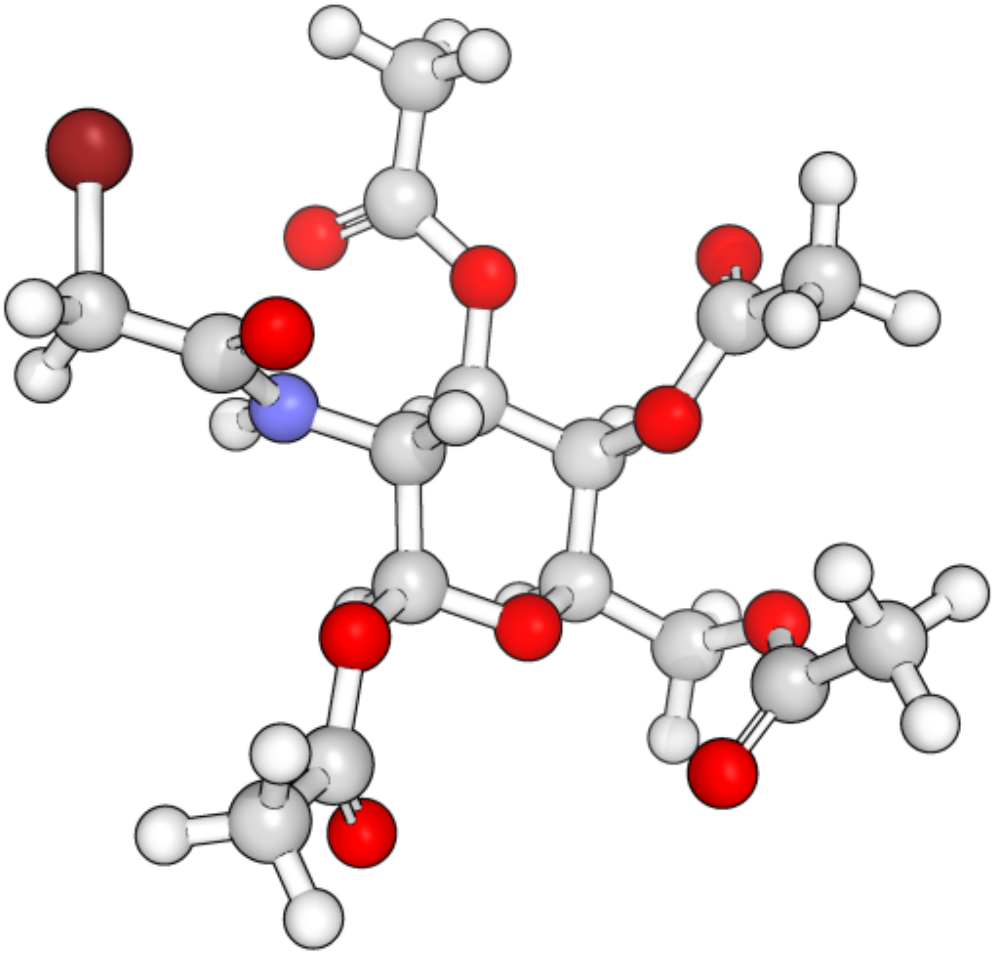
Molecule 16





Molecule 220

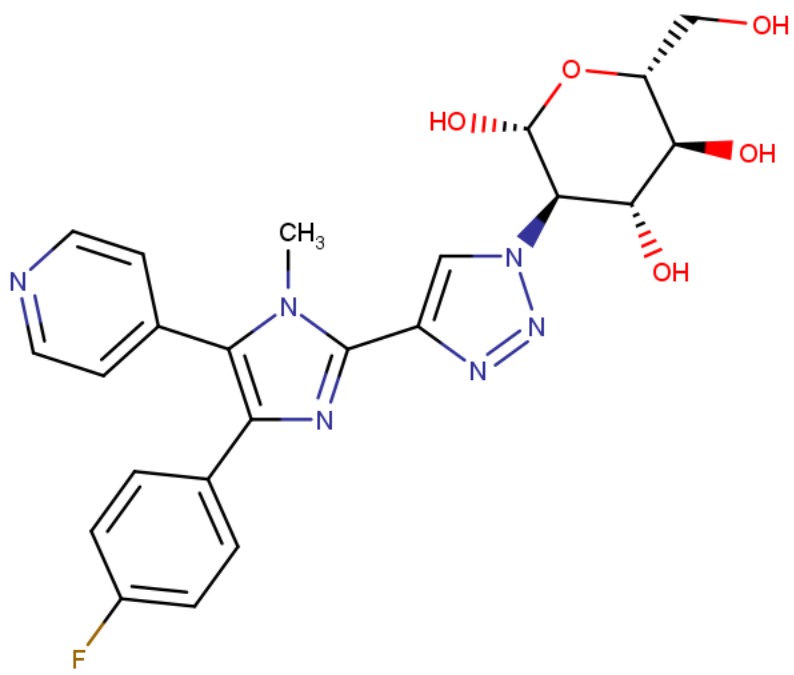


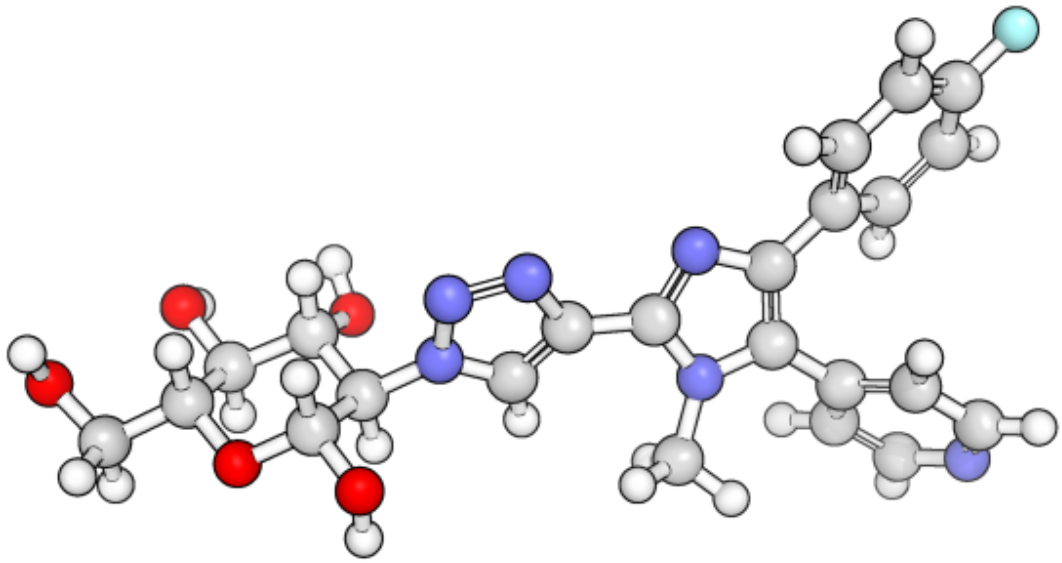


Top 3 molecules for exponential consensus ranking (ECR):

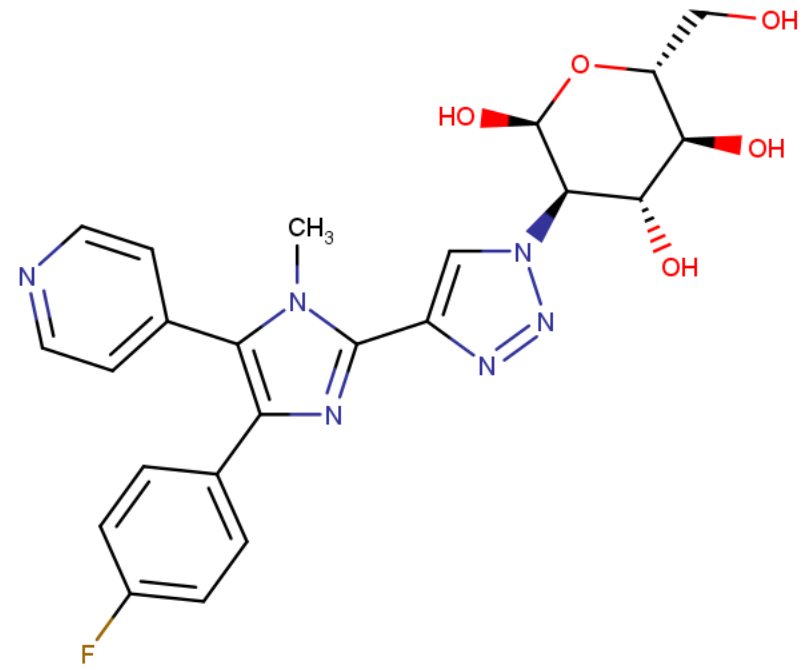
Two variants

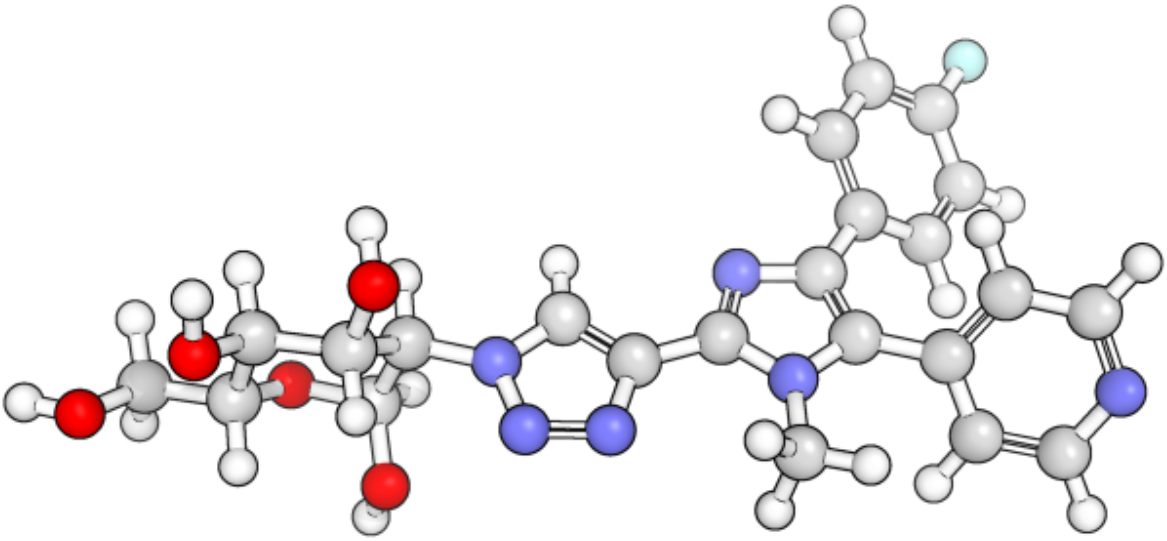
Molecule 5



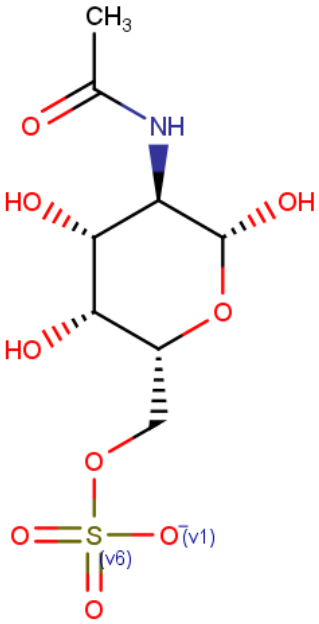


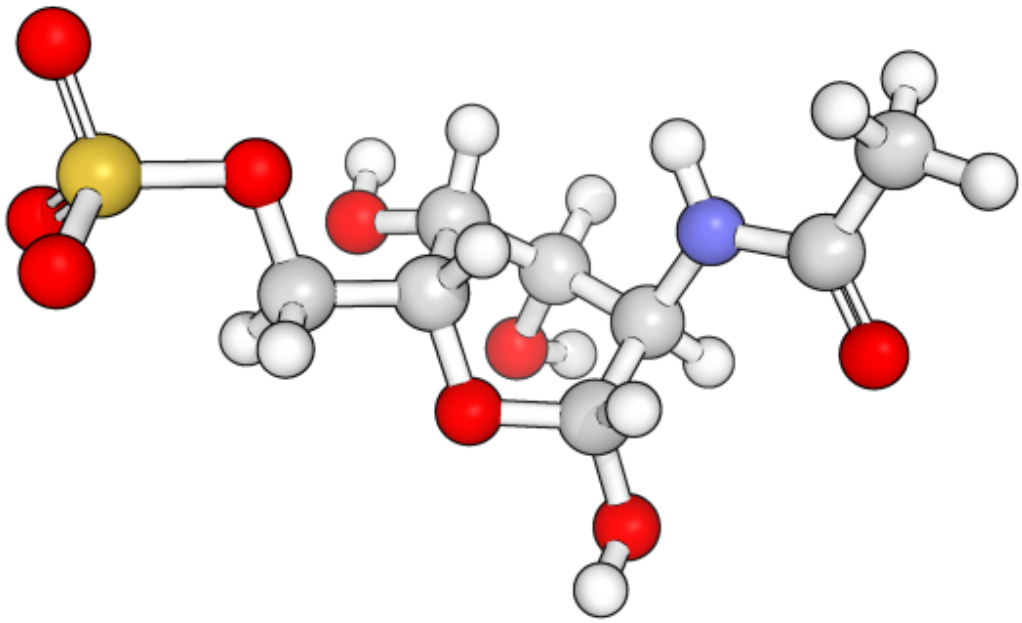
Molecule 11





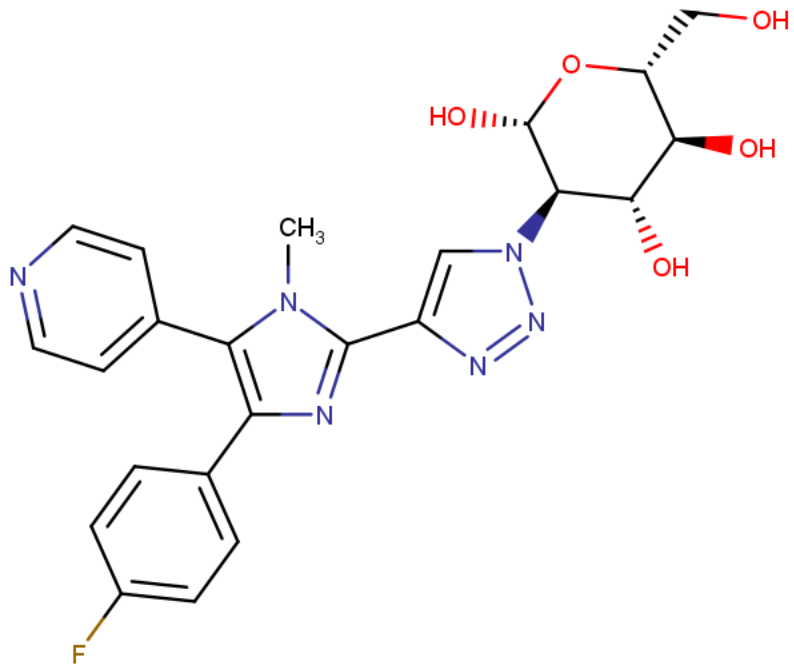
Molecule 22

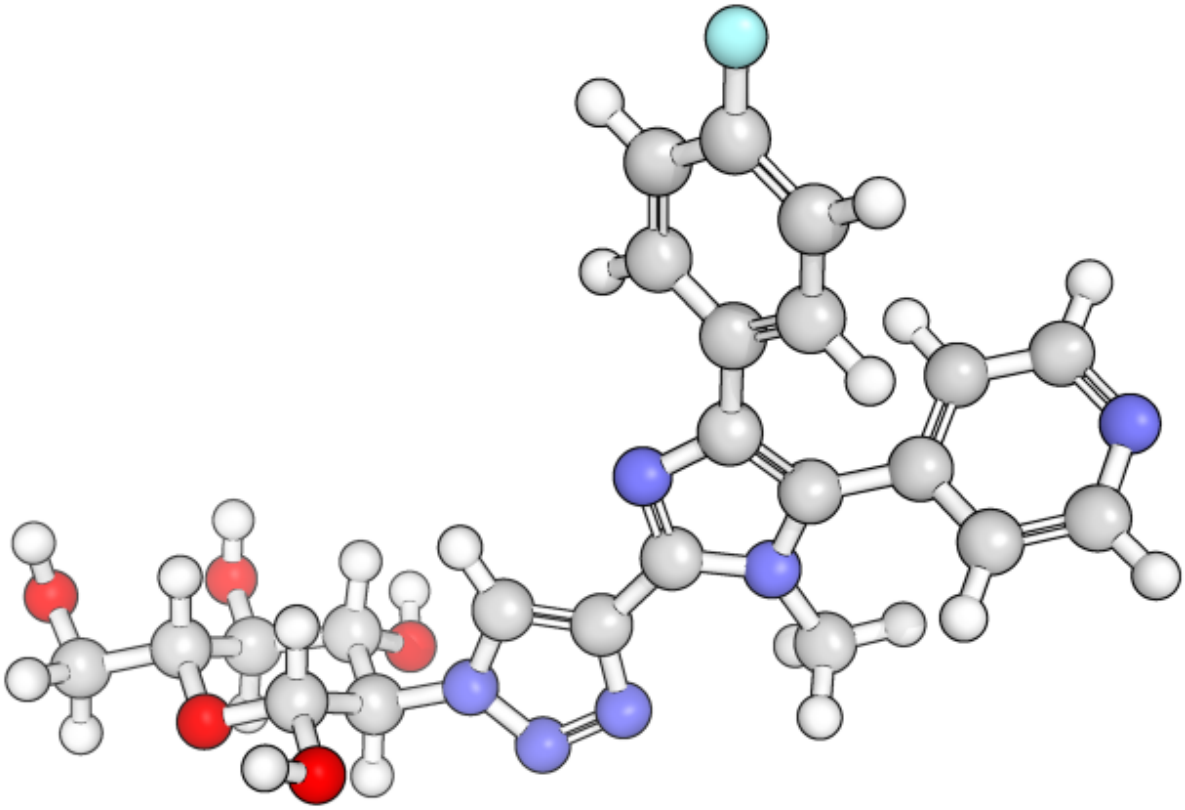

CH3
O
NH
HO
OH
HO
O
O
O
S
O
(v1)
(v6)
O



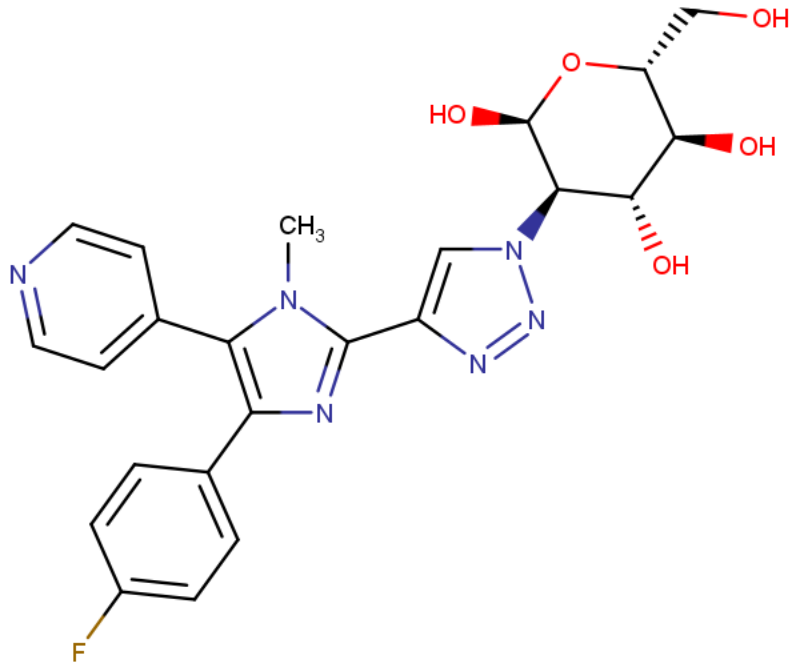
Three variants

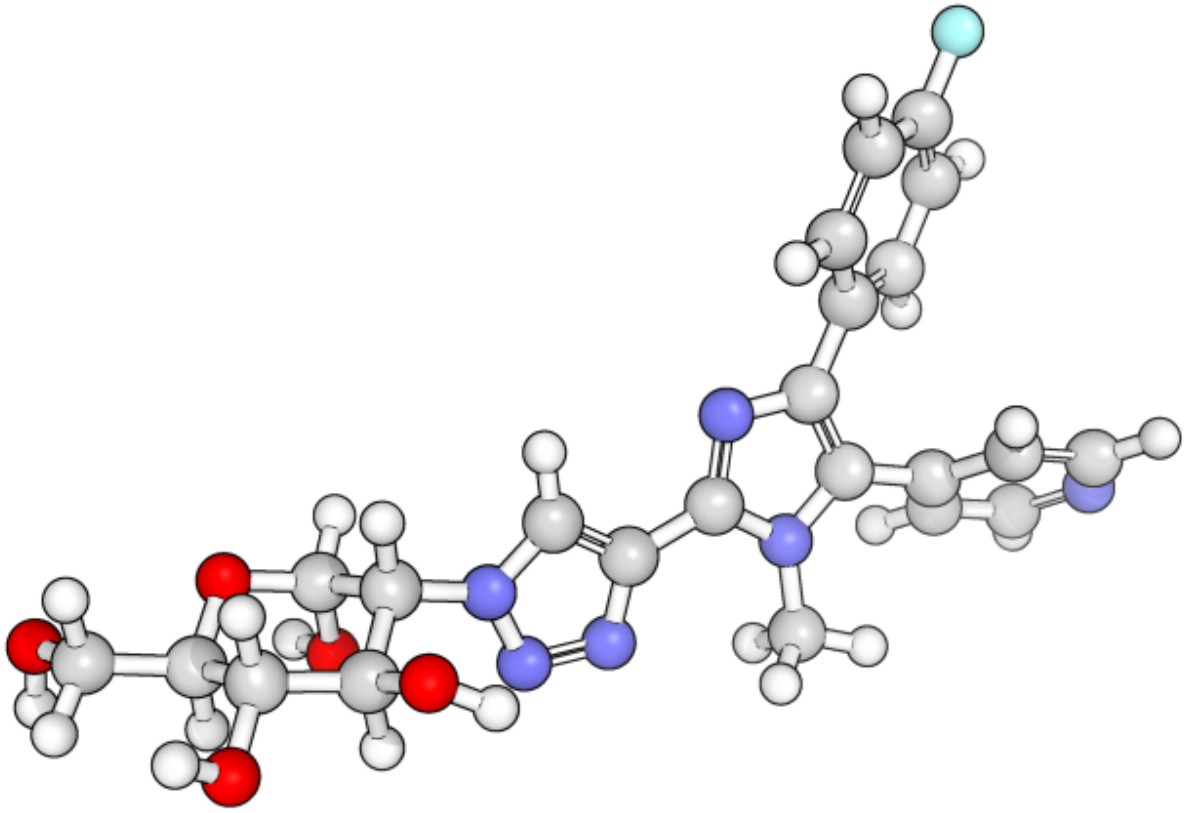
Molecule 7



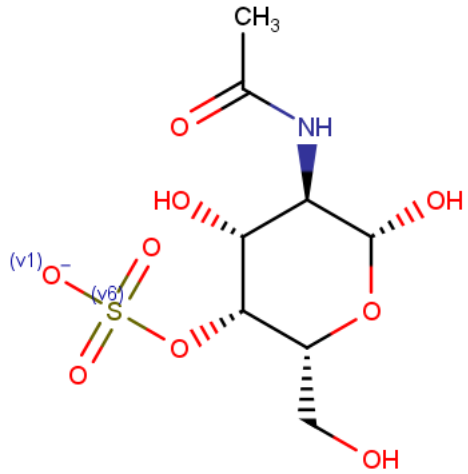


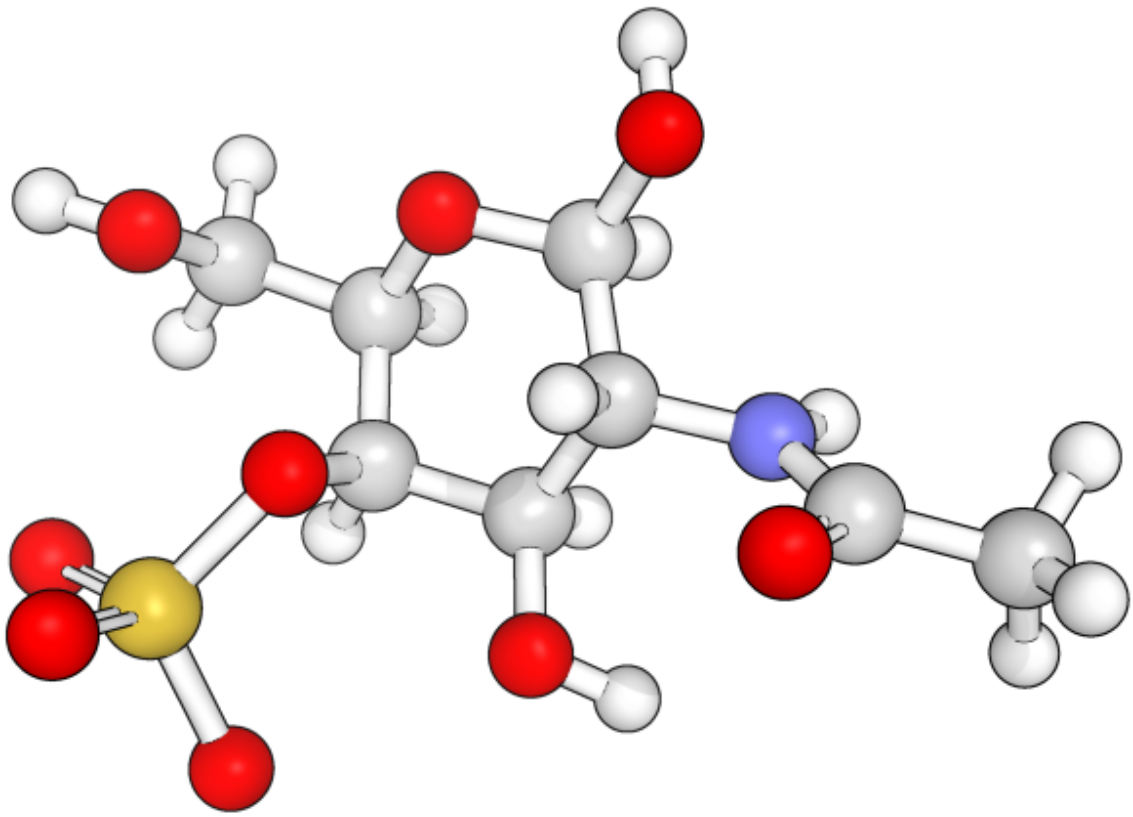
Molecule 16





Molecule 123





Molecular descriptors

One of the most important feature of these QSAR models are their physicochemical interpretation, since the correlation between these characteristics and the biological activity of the ligands can be used to propose new ligands and to get more insights about the mechanism of action.

Descriptor	Chemical interpretation	Descriptor type
F04[C-S]	Frequency of apparition of C-S atoms at topological distance of 4.	Atom-pair descriptor
B10[C-S]	Presence/Ausence of C-S atoms at topological distance of 10.	Atom-pair descriptor
CATS2D_00_LL	Distance count between lipophilic-lipophilic pair atoms at topological distance of 0.	Pharmacophore descriptor
CATS2D_08_AL	Distance count between hydrogen bond acceptor-lipophilic pair atoms at topological distance of 8.	Pharmacophore descriptor
nR06	Number of six-membered rings	Ring count descriptor
CATS2D_02_LL	Distance count between lipophilic-lipophilic pair atoms at topological distance of 2.	Pharmacophore descriptor
CATS2D_04_DA	Distance count between hydrogen bond donator-hydrogen bond acceptor pair atoms at topological distance of 4.	Pharmacophore descriptor

So, the common physicochemical features are the presence of C-S atoms at a certain topological distance, and at least one hydrogen bond donator and one hydrogen bond aceptor with a liphophilic group and at least one 6-membered ring. Please refer to the manuscript to get more information about

References

Palacio-Rodríguez, Karen, Isaias Lans, Claudio N. Cavasotto, and Pilar Cossio. 2019. "Exponential consensus ranking improves the outcome in docking and receptor ensemble docking." *Scientific Reports* 9 (1): 1–14. https://doi.org/10.1038/s41598-019-41594-3.