HackBio Machine Learning Project Report

Prediction of the antimalarial activity of a group of dihydroorotate dehydrogenase analogs.

Paschal Ugwu ugwupaschal@gmail.com

November, 2022.

Data Warrior software was used to extract compounds with good bioactivties. Componds were retrieved from Chembl database. I ensured that the compounds used all have same scafolds. They also shared same fragments. I saved my files in .sdf format. Chemopy Calculator was used to generate chemical descriptors.

The performance of the ezqsar package in an example data set that is provided by the package after installation was demonstrated in the study.

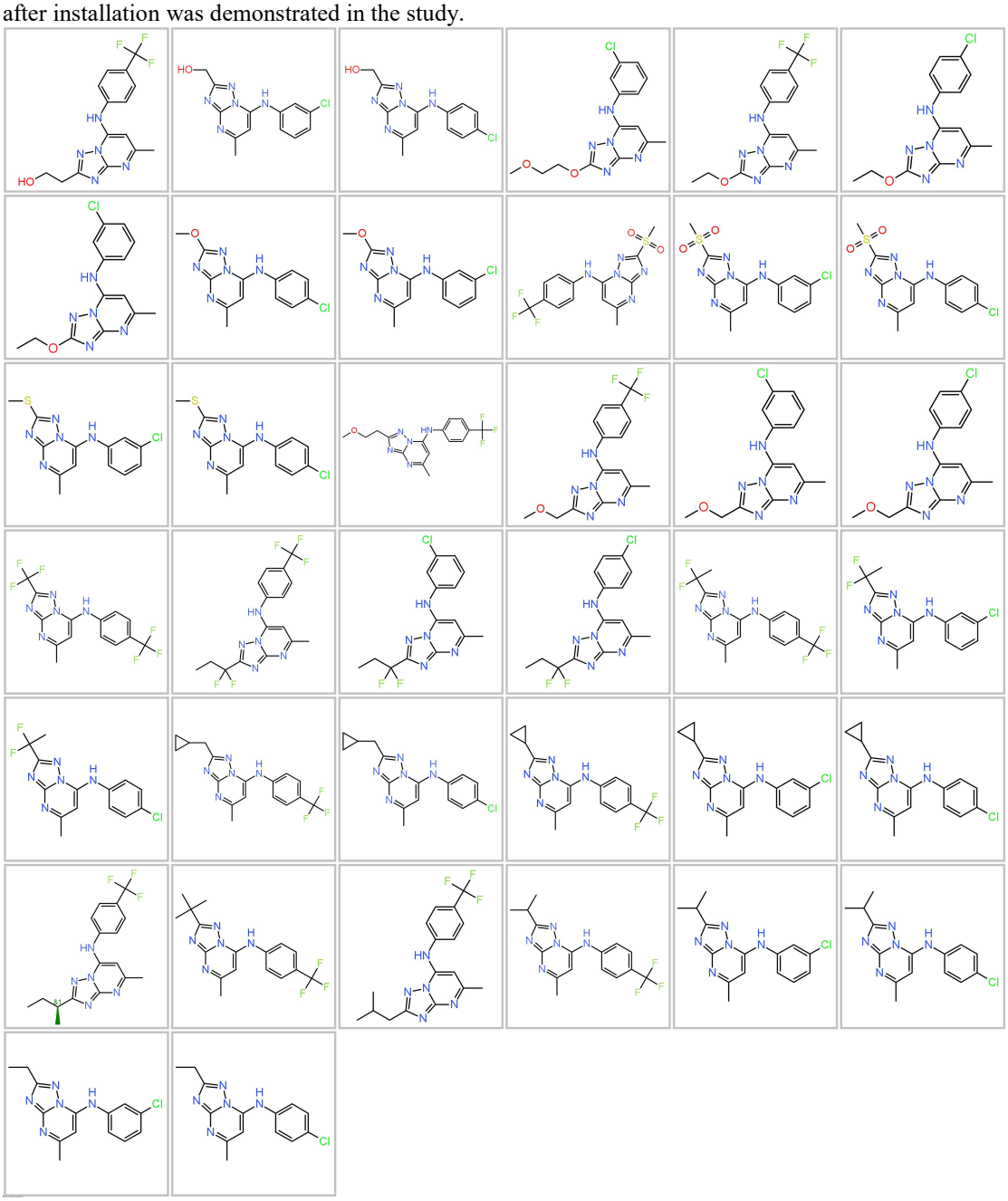

Fig. (1). Structures of the 38 dihydroorotate dehydrogenase inhibitors used.

An overview of the ezqsar_f workflow is demonstrated in Fig. (3). All of the molecules were collected in a single SDF file. Activities were provided in a separate csv file rank ordered same as the SDF file.

The activities were expressed as pIC50, however, they also can be expressed as IC50. Data set contains 38 dihydroorotate dehydrogenase inhibitors. There is also 79 de novo syntesized molecules. In this analysis, the dependent variable is the bioactivity (pIC50) while the independent variables are the descriptors generated.

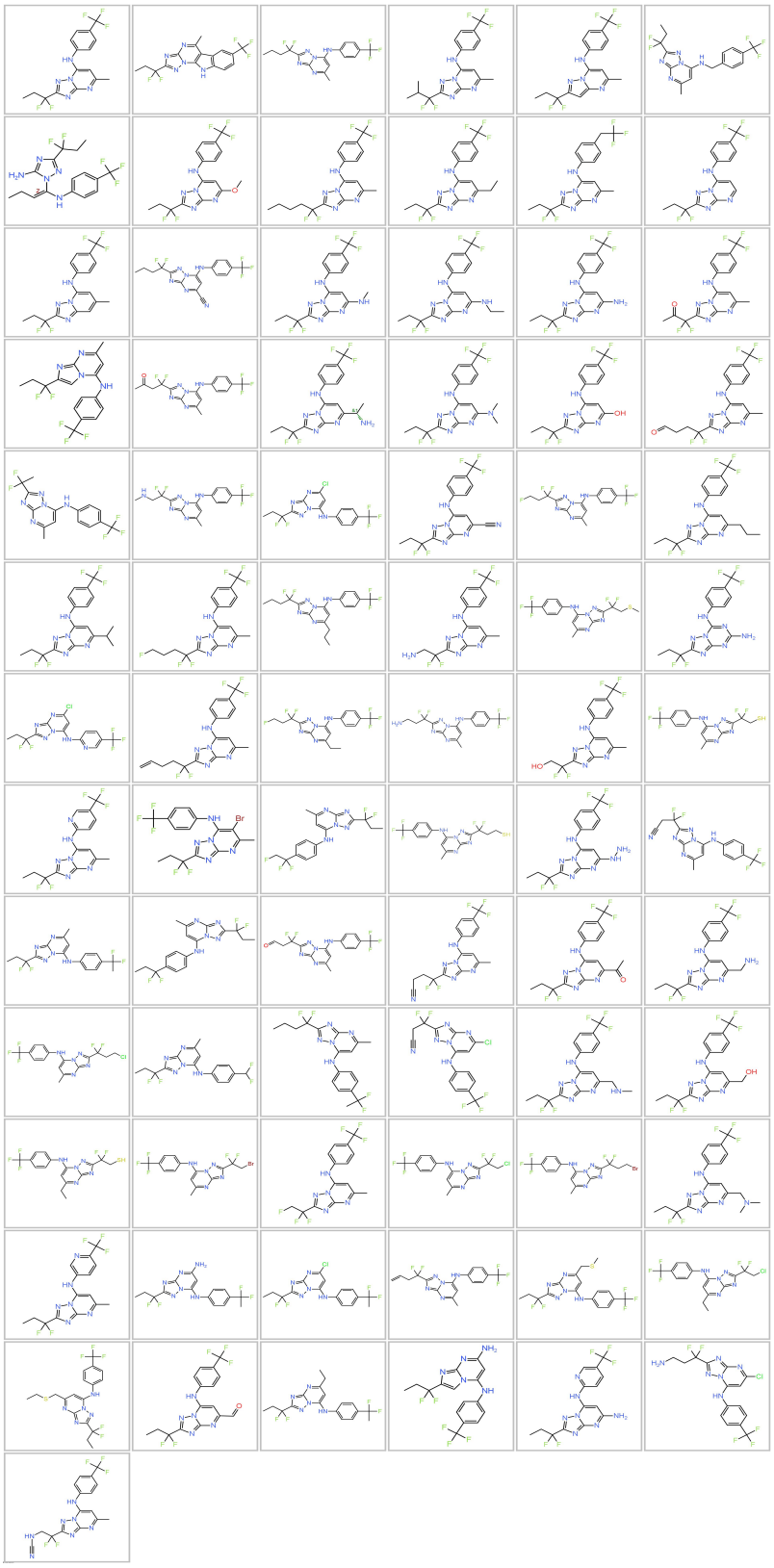

Fig. (2). De novo Synthesized compounds.

I performed the analysis thus:

```
# Load required libraries
>library(ezqsar)
#Import datasets
>file1 <- ("C:/Users/ugwupaschal/Downloads/converted_raw data.sdf")
>file2 <- ("C:/Users/ugwupascha/Downloads/IC50.csv")
>file3 <- ("C:/Users/ugwupascha/Downloads/de novo data.sdf")
# Fit model
>model<-ezqsar_f (SDFfile=file1, activityfile=file2, newdataset=file3,
testset=c(2,4,6,8,10,12,14,16,18,20,22,24,26))
>attributes (model)
$names
 [1] "Q2"                    "R2"
 [3] "adjusted_R2"           "RMSE"
 [5] "F_statistics"          "Descriptors"
 [7] "Model"                 "train"
 [9] "test"                  "newset"
[11] "newset2"               "R2_pred"
[13] "standardized_des_train"   "standardized_des_test"
[15] "standardized_des_newset"  "standardized_des_newset2"
[17] "AD_outlier_train"         "AD_outlier_test"
[19] "AD_outlier_newset"        "AD_outlier_newset2"
[21] "Tanimoto_train"           "Tanimoto_test"
[23] "Tanimoto_newset"          "Tanimoto_newset2"
[25] "Tanimoto_train_sum"       "Tanimoto_test_sum"
[27] "Tanimoto_newset_sum"      "Tanimoto_newset2_sum"

> a<-print (model$Q2)
[1] 0.5840896
> b<-print (model$R2)
[1] 0.829861
> c<-print (model$test)
           test_observed test_predicted
Compound 2           5.21       5.746110
Compound 4           5.31       6.090550
Compound 6           6.57       6.383552
Compound 8           6.34       6.160545
Compound 10          7.05       7.170920
Compound 12          7.35       6.722691
Compound 14          7.52       7.158979
Compound 16          6.39       5.950965
Compound 18          5.66       5.497065
Compound 20          7.72       6.948119
Compound 22          7.47       6.602815
Compound 24          7.27       7.378074
Compound 26          5.77       5.540850
> d<-print (model$R2_pred)
       26
0.6676623
> e<-print (model$Tanimoto_test_sum)
           Maximum Tanimoto similarity index Minimum Tanimoto similarity index
Compound 2                          1.0000000                          0.6000000
Compound 4                          0.8461538                          0.5588235
Compound 6                          1.0000000                          0.5890411
Compound 8                          1.0000000                          0.6271186
Compound 10                         0.9508197                          0.5441176
Compound 12                         1.0000000                          0.5200000
Compound 14                         1.0000000                          0.5873016
Compound 16                         0.9444444                          0.6000000
```

```
Compound 18                1.0000000                0.6376812
Compound 20                0.9534884                0.6031746
Compound 22                1.0000000                0.6349206
Compound 24                1.0000000                0.6557377
Compound 26                0.9361702                0.5441176
          Average Tanimoto similarity index
Compound 2                 0.7274071
Compound 4                 0.6476895
Compound 6                 0.7123553
Compound 8                 0.7206833
Compound 10                0.6140800
Compound 12                0.6229989
Compound 14                0.7499481
Compound 16                0.7333404
Compound 18                0.7418575
Compound 20                0.8126867
Compound 22                0.8231059
Compound 24                0.8264060
Compound 26                0.7660747
> f<-print (model$AD_outlier_test)
     Row number in test set Row number in entire set Molecule name
[1,] "7"                    "0"                     "Compound 14"
[2,] "1"                    "0"                     "Compound 2"
[3,] "1"                    "0"                     "Compound 2"
     Out of range descriptor Standardized value
[1,] "khs.ssS"               "4.8"
[2,] "nHBDon"                "3.32264954516723"
[3,] "RNCS"                  "3.17233480711894"
> g<-print (model$newset)
      name          newset_predicted
 [1,] "Compound 1"  "6.92334753967643"
 [2,] "Compound 2"  "4.17725587235477"
 [3,] "Compound 3"  "6.40454143934573"
 [4,] "Compound 4"  "7.00897731125113"
 [5,] "Compound 5"  "6.36996610057841"
 [6,] "Compound 6"  "6.46255506953823"
 [7,] "Compound 7"  "8.25697835435487"
 [8,] "Compound 8"  "6.52140115494184"
 [9,] "Compound 9"  "5.22626812961143"
[10,] "Compound 10" "6.16620811791867"
[11,] "Compound 11" "6.09607676275711"
[12,] "Compound 12" "7.54927072451032"
[13,] "Compound 13" "6.40623816896407"
[14,] "Compound 14" "4.78683008585607"
[15,] "Compound 15" "7.80155381012227"
[16,] "Compound 16" "7.32235332518759"
[17,] "Compound 17" "7.8143424805104"
[18,] "Compound 18" "6.05780109400099"
[19,] "Compound 19" "6.32975649537761"
[20,] "Compound 20" "5.79089558380561"
[21,] "Compound 21" "7.00581069491717"
[22,] "Compound 22" "6.57948988042102"
[23,] "Compound 23" "7.06936241265437"
[24,] "Compound 24" "4.76160747064327"
[25,] "Compound 25" "7.35195533392697"
[26,] "Compound 26" "7.65374542231445"
[27,] "Compound 27" "7.12275974015403"
[28,] "Compound 28" "5.33830250194748"
[29,] "Compound 29" "5.48842760071952"
[30,] "Compound 30" "5.60980914939212"
```

```
[31,] "Compound 31" "5.8818163505265"
[32,] "Compound 32" "4.8938059695174"
[33,] "Compound 33" "5.03933423042551"
[34,] "Compound 34" "7.57023399957753"
[35,] "Compound 35" "7.45751742450957"
[36,] "Compound 36" "8.32847993600228"
[37,] "Compound 37" "7.51363321702414"
[38,] "Compound 38" "6.04443168884848"
[39,] "Compound 39" "4.76200329529469"
[40,] "Compound 40" "7.00494643518677"
[41,] "Compound 41" "6.72105545516893"
[42,] "Compound 42" "7.12455320019038"
[43,] "Compound 43" "7.3618882099569"
[44,] "Compound 44" "5.58759135444769"
[45,] "Compound 45" "6.32238860499002"
[46,] "Compound 46" "6.45021967253293"
[47,] "Compound 47" "8.41157887947661"
[48,] "Compound 48" "5.29845601182477"
[49,] "Compound 49" "6.49124951699116"
[50,] "Compound 50" "5.93251052514749"
[51,] "Compound 51" "5.25623079364261"
[52,] "Compound 52" "4.43504695526077"
[53,] "Compound 53" "5.3205181872174"
[54,] "Compound 54" "7.27145048959211"
[55,] "Compound 55" "6.51791847302171"
[56,] "Compound 56" "6.83950774775901"
[57,] "Compound 57" "5.88805763208429"
[58,] "Compound 58" "5.41607499777085"
[59,] "Compound 59" "6.834945300613"
[60,] "Compound 60" "6.63623524449296"
[61,] "Compound 61" "6.38052984313018"
[62,] "Compound 62" "6.97533082158821"
[63,] "Compound 63" "7.38316949971043"
[64,] "Compound 64" "7.26763238891109"
[65,] "Compound 65" "6.07948930417214"
[66,] "Compound 66" "6.07685034653695"
[67,] "Compound 67" "7.05703837241348"
[68,] "Compound 68" "7.31251870594805"
[69,] "Compound 69" "6.63910241832598"
[70,] "Compound 70" "6.87165435307235"
[71,] "Compound 71" "6.380658560697"
[72,] "Compound 72" "6.50835868218139"
[73,] "Compound 73" "4.93554789725356"
[74,] "Compound 74" "5.59885694638211"
[75,] "Compound 75" "5.77413765801027"
[76,] "Compound 76" "7.15909708682315"
[77,] "Compound 77" "8.26632195943631"
[78,] "Compound 78" "7.17211682246574"
[79,] "Compound 79" "7.90997634922459"
> h<-print (model$Tanimoto_newset_sum)
```

	Maximum Tanimoto similarity index	Minimum Tanimoto similarity index
Compound 1	0.9534884	0.6031746
Compound 2	0.8636364	0.5468750
Compound 3	0.9130435	0.5757576
Compound 4	1.0000000	0.5937500
Compound 5	0.9302326	0.5873016
Compound 6	0.8723404	0.5671642
Compound 7	0.7115385	0.4722222
Compound 8	0.8771930	0.5774648
Compound 9	0.8936170	0.5671642

Compound 10	0.9111111	0.5846154
Compound 11	0.9534884	0.6031746
Compound 12	0.9500000	0.5873016
Compound 13	0.9302326	0.5873016
Compound 14	0.8723404	0.5522388
Compound 15	0.8913043	0.6000000
Compound 16	0.8367347	0.5507246
Compound 17	0.8837209	0.5606061
Compound 18	0.9069767	0.6349206
Compound 19	0.8636364	0.5468750
Compound 20	0.8666667	0.5909091
Compound 21	0.9111111	0.5846154
Compound 22	0.8750000	0.5820896
Compound 23	0.8305085	0.5270270
Compound 24	0.8723404	0.5757576
Compound 25	1.0000000	0.6229508
Compound 26	0.8125000	0.5652174
Compound 27	0.9767442	0.6190476
Compound 28	0.9268293	0.5781250
Compound 29	0.8913043	0.5606061
Compound 30	0.8750000	0.5588235
Compound 31	0.9555556	0.5757576
Compound 32	0.8913043	0.5606061
Compound 33	0.8750000	0.5588235
Compound 34	0.8444444	0.5441176
Compound 35	0.8297872	0.6212121
Compound 36	0.8837209	0.5606061
Compound 37	0.9767442	0.6190476
Compound 38	0.8541667	0.5441176
Compound 39	0.8367347	0.5362319
Compound 40	0.8392857	0.5138889
Compound 41	0.9074074	0.5571429
Compound 42	0.8444444	0.5909091
Compound 43	0.9534884	0.6031746
Compound 44	0.9545455	0.6093750
Compound 45	0.9111111	0.5846154
Compound 46	0.8214286	0.5652174
Compound 47	0.8260870	0.5588235
Compound 48	0.9268293	0.5781250
Compound 49	0.9545455	0.5937500
Compound 50	0.9545455	0.5937500
Compound 51	0.9047619	0.5937500
Compound 52	0.8913043	0.5606061
Compound 53	0.8723404	0.6153846
Compound 54	0.7916667	0.5211268
Compound 55	0.8723404	0.5757576
Compound 56	0.9534884	0.6031746
Compound 57	0.9347826	0.5671642
Compound 58	0.8888889	0.5937500
Compound 59	0.7924528	0.5416667
Compound 60	0.8909091	0.5342466
Compound 61	0.7916667	0.5652174
Compound 62	0.9047619	0.5692308
Compound 63	0.9268293	0.5781250
Compound 64	0.9047619	0.5937500
Compound 65	0.8723404	0.5522388
Compound 66	0.8301887	0.5342466
Compound 67	0.9534884	0.6031746
Compound 68	0.8913043	0.5757576
Compound 69	1.0000000	0.6349206

Compound 70	0.9047619	0.5692308
Compound 71	0.8076923	0.5942029
Compound 72	0.8723404	0.5671642
Compound 73	0.8039216	0.5797101
Compound 74	0.8837209	0.6093750
Compound 75	0.9130435	0.5757576
Compound 76	0.7954545	0.5074627
Compound 77	0.8837209	0.5606061
Compound 78	0.7931034	0.5277778
Compound 79	0.8444444	0.5441176

	Average Tanimoto similarity index
Compound 1	0.8126867
Compound 2	0.7347698
Compound 3	0.7861360
Compound 4	0.8005951
Compound 5	0.7918776
Compound 6	0.7657791
Compound 7	0.6152189
Compound 8	0.7132641
Compound 9	0.7719575
Compound 10	0.7908496
Compound 11	0.8126867
Compound 12	0.8023399
Compound 13	0.7918776
Compound 14	0.7611263
Compound 15	0.7724960
Compound 16	0.7394038
Compound 17	0.7567397
Compound 18	0.7692896
Compound 19	0.7347698
Compound 20	0.7577776
Compound 21	0.7821808
Compound 22	0.7593588
Compound 23	0.6792039
Compound 24	0.7590804
Compound 25	0.8160044
Compound 26	0.7309676
Compound 27	0.8129593
Compound 28	0.7856805
Compound 29	0.7769408
Compound 30	0.7664284
Compound 31	0.7973267
Compound 32	0.7769408
Compound 33	0.7664284
Compound 34	0.7358412
Compound 35	0.7445942
Compound 36	0.7567397
Compound 37	0.8129593
Compound 38	0.7346676
Compound 39	0.7511645
Compound 40	0.7091065
Compound 41	0.7273381
Compound 42	0.7403054
Compound 43	0.8126867
Compound 44	0.8010216
Compound 45	0.7943495
Compound 46	0.7162474
Compound 47	0.7149777
Compound 48	0.7879850
Compound 49	0.8057947

```
Compound 50                0.8057947
Compound 51                0.7727364
Compound 52                0.7734528
Compound 53                0.7652758
Compound 54                0.7131826
Compound 55                0.7820174
Compound 56                0.8126867
Compound 57                0.7798402
Compound 58                0.7784842
Compound 59                0.7094767
Compound 60                0.7041628
Compound 61                0.7173449
Compound 62                0.7754385
Compound 63                0.7915171
Compound 64                0.7966312
Compound 65                0.7616443
Compound 66                0.7068800
Compound 67                0.8126867
Compound 68                0.7666765
Compound 69                0.8231059
Compound 70                0.7610465
Compound 71                0.7219703
Compound 72                0.7694898
Compound 73                0.7255309
Compound 74                0.7704638
Compound 75                0.7844452
Compound 76                0.6839956
Compound 77                0.7567397
Compound 78                0.7018856
Compound 79                0.7358412
> i<-print(model$Model)

Call:
lm(formula = .outcome ~ ., data = dat, verbose = TRUE)

Coefficients:
(Intercept)    khs.ssCH2      khs.ssS        VP.5       nHBDon       PNSA.1
  12.680232    -0.156935     1.286296    -4.995860     0.943373     0.003803
      RNCS
  -0.325238

> train<-model$train
> test<-model$test
> write.csv(train, "train.csv")
```

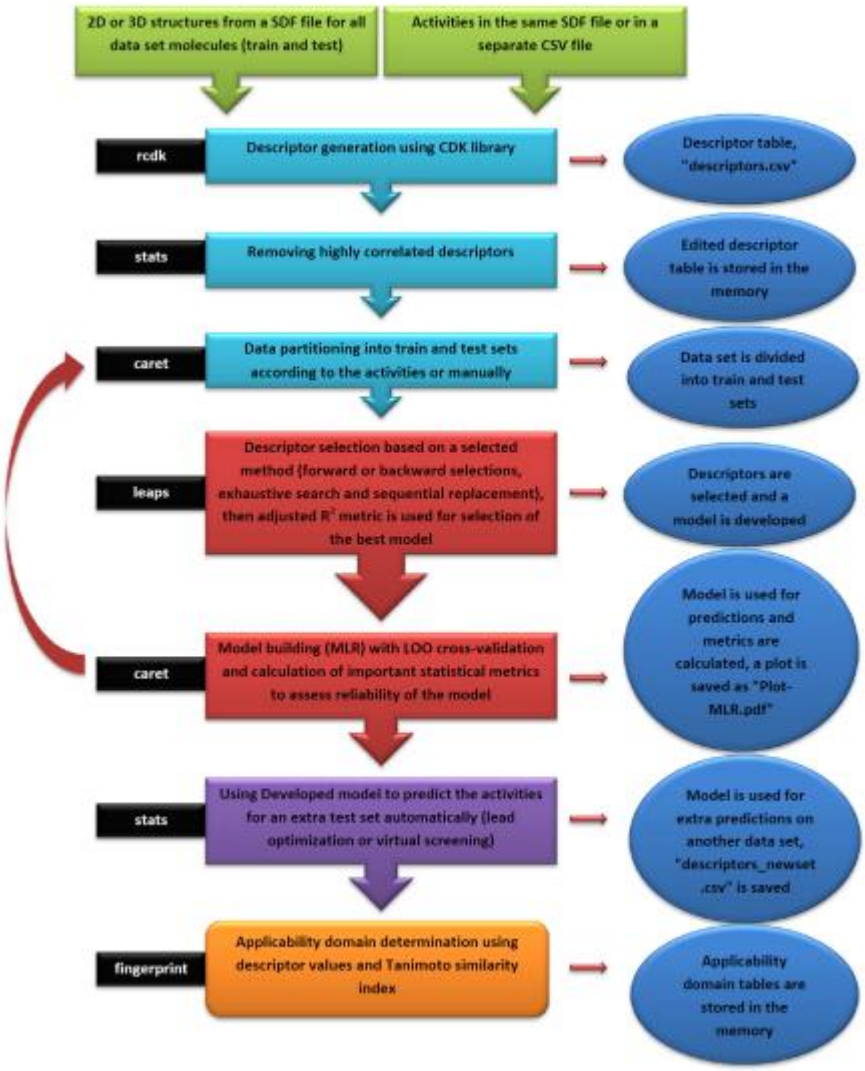

Fig. (3). The workflow of ezqsar_f function from ezqsar package.

The input files were stored as files 1, 2 and 3 by executing lines 1, 2 and 3. The model was developed by executing line 4. This ran the function with default parameters and stored the results as an object called, model. Highly correlated descriptors (correlation coefficient over a defined threshold) could be removed from the descriptor table before further processing. By default, this threshold is 1 (correlated descriptors are not removed), test set ratio is 20% and the selection is based on the activities; descriptors were selected by forward selecting method before final MLR model development. The calculated descriptors are reported as an csv file named “descriptors.csv”. A plot (“Plot-MLR.pdf “) was available in the working directory (can be changed by setwd () in R) after a successful run regarding the observed vs predicted activity values for both train and test sets using the developed model Fig. (4). Available outputs are shown after executing line 6 and some of the most important ones were printed out by remaining lines of the example (lines 6 to 13).

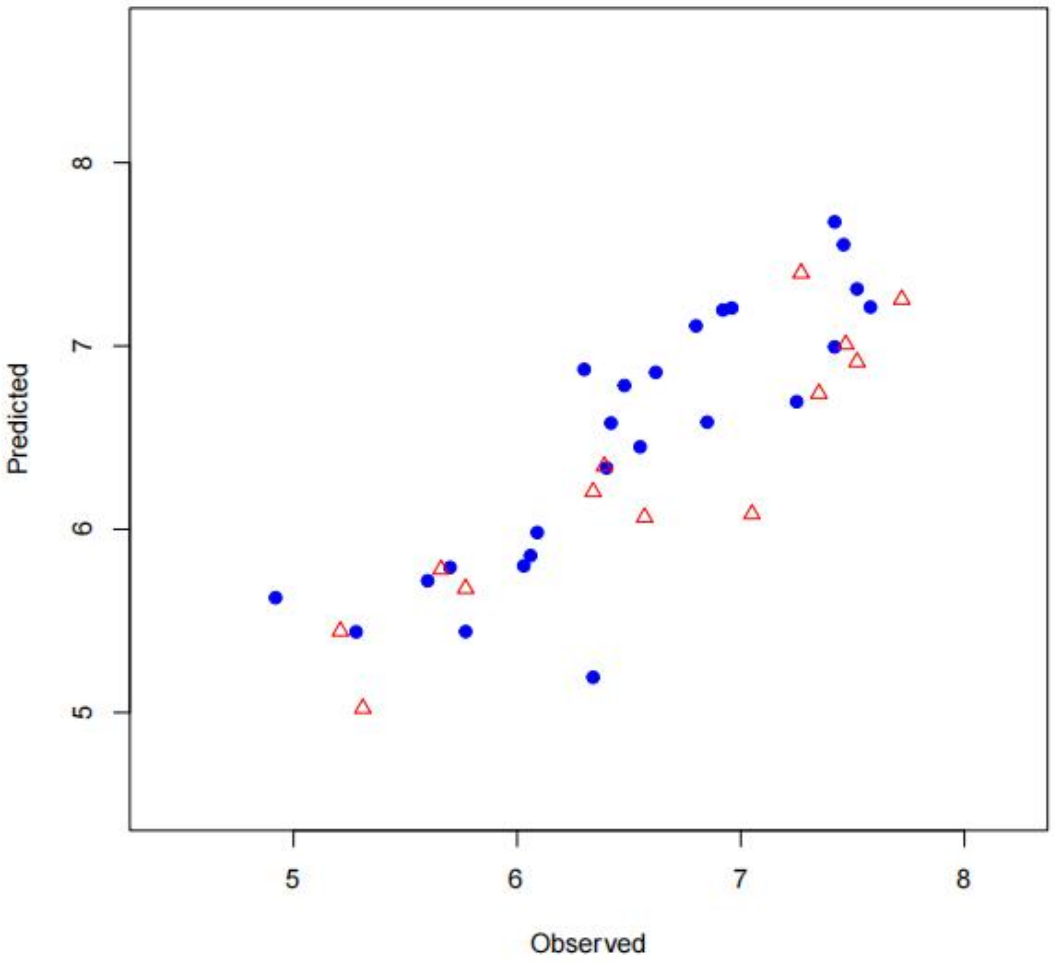

Fig. (4). Plot of observed versus predicted activities obtained from model for training (blue circles) and test (red triangles) sets.