QSAR prediction models.

Two major strategies were used:

1. building models using descriptors as they are
2. crossterms, PCA, clusters generated based on the genuine descriptors.

Descriptors considered in model building were:

(a) descriptors included in commercially available software {MOE, Schrodinger]

(b) topological descriptors computed using ,,in home,, software used in conjunction with a CoMFA augumented technique.

All strategies implied don’t consider a receptor structure.

Models are built using secondary derived data[[1]](#endnote-1) .

Data set analyzed is composed of 95 compounds. All compounds were considered.

Structures formats used are mol mol and hin.

Dependent variable was toxicity expressed in log/nanomol.

Number of maximum descriptors used in models was set to 19 (n/5).

A future selection technique was considered in selecting the receptors. This algorithm selects best receptors for building the model ordering by r2.

Variability, collinearity and multicollinearity of descriptors was analyzed. Two major type of models were computed as discussed above: 1. Where collinearity and multicollinearity is considered 2. Here collinearity and multicollinearity is not considered. In characterizing descriptors Tolerance and VIF ( variance inflation factor ) were used[[2]](#endnote-2).

1. (ia)

Descriptors were generated using 2D, 3D and ADME descriptors( descriptors that simulate behavior of compounds in culture medias – used for toxicity) All descriptor used are commercially available descriptors ( used in MOE and Schrodinger software)

Multiple linear regression model

1. Model was built using a set of 15 descriptors : E\_nb, E\_stb, GCUT\_PEOE\_2, GCUT\_SLOGP\_0, SMR\_VSA6, SlogP\_VSA9, logs, opr\_nring, opr\_nrot, opr\_violation, radius, vsurf\_CW5, vsurf\_DD13, vsurf\_HL1, vsurf\_IW6 Contribution of each descriptor is shown in table below.

Table 1 Descriptors used in model computation.

|  |  |  |
| --- | --- | --- |
| Pearson r2 | Descriptor |  |
|  |
| 0.900 | E\_nb E\_stb GCUT\_PEOE\_2 GCUT\_SLOGP\_0 SlogP\_VSA9 vsurf\_HL1 vsurf\_IW6 SMR\_VSA6 logS opr\_nring opr\_nrot opr\_violation radius vsurf\_CW5 vsurf\_DD13 | 15 |
| 0.892 | E\_nb E\_stb GCUT\_PEOE\_2 GCUT\_SLOGP\_0 SMR\_VSA6 logS radius vsurf\_HL1 vsurf\_DD13 vsurf\_IW6 SlogP\_VSA9 opr\_nring opr\_nrot opr\_violation | 14 |
| 0.887 | E\_nb E\_stb GCUT\_PEOE\_2 SMR\_VSA6 SlogP\_VSA9 logS opr\_nring opr\_nrot opr\_violation vsurf\_DD13 vsurf\_HL1 vsurf\_IW6 radius | 13 |
| 0.880 | E\_stb GCUT\_PEOE\_2 SMR\_VSA6 SlogP\_VSA9 logS opr\_nring opr\_nrot opr\_violation radius vsurf\_DD13 vsurf\_HL1 vsurf\_IW6 | 12 |
| …….. | ………………………………………………………………………………………………………………… | … |
| 0.759 | E\_stb logS opr\_nring opr\_nrot | 4 |
| 0.743 | E\_stb logS opr\_nring | 3 |
| 0.723 | logS opr\_nring | 2 |
| 0.690 | opr\_nring | 1 |
|  | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A model was computed using the first solution. Model equation: r2=0.900, p=0.946501, q2=0.900, RMSD=0.604, y=-0.0196617+0.9002748x Observedtoxicity. Notice that descriptor opr\_ring has a contribution of 76.66% to model.  Descriptors analysis: tolerance[[3]](#endnote-3)[[4]](#endnote-4)[[5]](#endnote-5) and VIF were correlated[[6]](#endnote-6)[[7]](#endnote-7)[[8]](#endnote-8)[[9]](#endnote-9)[[10]](#endnote-10)[[11]](#endnote-11)   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Descriptor | R2 | Tolerance(1-r2) 0.20 minimum value | VIF 1/Tolerance  4-20 max value |  | | E\_nb | 0.1756 | 0.824 | 1.213 |  | | E\_stb | 0.5473 | 0.452 | 2.212 |  | | GCUT\_PEOE\_2 | 0.7116 | 0.288 | 3.372 |  | | GCUT\_SLOGP\_0 | 0.3399 | 0.660 | 1.515 |  | | logS | 0.8998 | 0.100 | 10.000 |  | | opr\_nring | 0.8967 | 0.103 | 9.708 |  | | opr\_nrot | 0.7543 | 0.247 | 4.048 |  | | opr\_violation | 0.3205 | 0.679 | 1.472 |  | | radius | 0.7516 | 0.243 | 4.115 |  | | SlogP\_VSA9 | 0.7427 | 0.257 | 3.891 |  | | SMR\_VSA6 | 0.1871 | 0.812 | 1.231 |  | | vsurf\_CW5 | 0.9698 | 0.032 | 31.250 |  | | vsurf\_DD13 | 0.4878 | 0.512 | 1.950 |  | | vsurf\_HL1 | 0.9628 | 0.037 | 27.027 |  | | vsurf\_IW6 | 0.6796 | 0.320 | 3.125 |  | |  |  |  |  |  |   (A model computed excluding non-concordant values of Tolerance and VIF turned an r2=0.67)  In figure 1 MLR model discussed is represented |
|  |
| Figure 1 MLR model with 15 descriptors.  Using same descriptor selection models using support vector regression(SVR), partial least square9PLS), artificial neural networks (ANN), K nearest neighbor( KNN), support vector cclassification (SVC) figures 2-6. |
|  |

Figure 2 SVR model with 15 descriptors(19) descriptors E\_oop E\_strain GCUT\_SLOGP\_2 MNDO\_IP PEOE\_RPC- PEOE\_VSA-2 ast\_fraglike b\_rotR lip\_don npr1 opr\_nring opr\_nrot radius std\_dim2 vsurf\_DW13 vsurf\_DW23 vsurf\_ID7 vsurf\_IW6 vsurf\_W2,

R2= 0.951929, p=0.971662, q2=0.94

Here opr\_ring represents 73,02% of the model.

Figure 3 PLS model with 15 descriptors E\_nb E\_stb GCUT\_PEOE\_2 GCUT\_SLOGP\_0 SMR\_VSA6 SlogP\_VSA9 logS opr\_nring opr\_nrot opr\_violation radius vsurf\_CW5 vsurf\_DD13 vsurf\_HL1 vsurf\_IW6

r2= 0.899, p=0.94, q2=0.8899.

opr\_nring=77.536

Figure 4 ANN model with 15 descriptors BCUT\_SMR\_1 MNDO\_LUMO PEOE\_RPC- PEOE\_VSA\_FPOS PEOE\_VSA\_PNEG SMR\_VSA4 SlogP\_VSA2 b\_1rotR lip\_don opr\_brigid opr\_violation std\_dim2 vsurf\_CW2 vsurf\_DW23 vsurf\_IW6, y=-0.057657+0.927357x Observedtoxicity.

r2=0.92, p=0.966, RMSD=0.513, q2= 0.928

Figure 5 KNN model with 15 descriptors E\_nb E\_stb GCUT\_PEOE\_2 GCUT\_SLOGP\_0 SMR\_VSA6 SlogP\_VSA9 logS opr\_nring opr\_nrot opr\_violation radius vsurf\_CW5 vsurf\_DD13 vsurf\_HL1 vsurf\_IW6

Figure 6 SVC model with 15 descriptors. E\_nb E\_stb GCUT\_PEOE\_2 GCUT\_SLOGP\_0 SMR\_VSA6 SlogP\_VSA9 logS opr\_nring opr\_nrot opr\_violation radius vsurf\_CW5 vsurf\_DD13 vsurf\_HL1 vsurf\_IW6

Interaction model

**PCA**

Far al 264 used descriptors PCA was performed. PC94 was computed for all 264 ligand set. Future selection was applied.

A model with 19 descriptors involving PCA using **MLR** resulted in a r2 of 0.957. Descriptors used were : GCUT\_PEOE\_0 MNDO\_HOMO PC 11 PC 23 PC 24 PC 25 PC 33 PC 37 PC 45 PC 48 PC 62 PC 9 PEOE\_PC+ PEOE\_RPC+ b\_double logS opr\_nring vsa\_don vsurf\_W8

Starting form a number bigger that 33 descriptors r2 values were 0.99 and from 86 descriptors r2=1.

Using **PLS** r2 =0.993047 was obtained usin 33 descriptors , for 19 descriptors r2=0.957 was obtained.

Using **SVR** r2=0.72857 was ontained for 7 descriptors.

Using **ANN** networks overfeeted models result in r2=1 .

**Cross terms**

Using 264 descriptors 34.980 new cross descriptors were generated.

1. Debnath, A.K.; Debnath, G.; Shusterman, A.J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Muagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella typhimurium TA98 and TA100. Environ. Mol. Mutagen. 1992, 19, 37-52. [↑](#endnote-ref-1)
2. [Baguley, T. (2012). Serious stats: A guide to advanced statistics for the behavioral sciences. Basingstoke: Palgrave.](http://www.amazon.co.uk/gp/product/0230577180/ref=as_li_tf_tl?ie=UTF8&camp=1634&creative=6738&creativeASIN=0230577180&linkCode=as2&tag=psychologic05-21)  [↑](#endnote-ref-2)
3. Menard, S. (1995). Applied Logistic Regression Analysis: Sage University Series on Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage. [↑](#endnote-ref-3)
4. Huber, E. & Stephens, J. D. (1993). Political Parties and Public Pensions: A Quantitative Analysis, *Acta Sociologica*, *36*, 309-325. [↑](#endnote-ref-4)
5. Tabachnick, B. G., & Fidell, L. S. (2001). Using Multivariate Statistics (4th ed.). Boston, MA: Allyn and Bacon [↑](#endnote-ref-5)
6. Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). Multivariate Data Analysis (3rd ed). New York: Macmillan.  [↑](#endnote-ref-6)
7. Kennedy, P. (1992). A Guide to Econometrics. Oxford: Blackwell. [↑](#endnote-ref-7)
8. Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics, 12, 591–256. [↑](#endnote-ref-8)
9. Neter, J., Wasserman, W. & Kutner, M. H. (1989). Applied Linear Regression Models. Homewood, IL: Irwin. [↑](#endnote-ref-9)
10. Pan, Y, & Jackson, R. T. (2008). Ethnic difference in the relationship between acute inflammation and and serum ferritin in US adult males. Epidemiology and Infection, 136, 421-431. [↑](#endnote-ref-10)
11. Rogerson, P. A. (2001). Statistical methods for geography. London: Sage. [↑](#endnote-ref-11)