RSC Advances

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
RSC Advances
www.rsc.org/advances
ROYAL SOCIETY OF CHEMISTRY

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. This *Accepted Manuscript* will be replaced by the edited, formatted and paginated article as soon as this is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

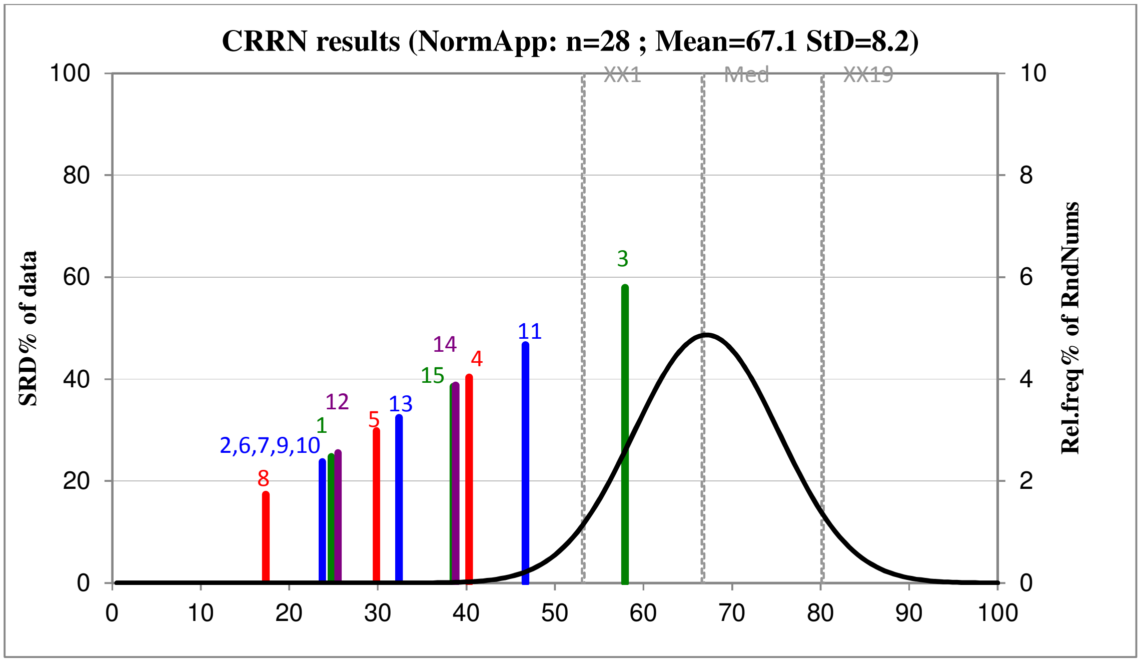
www.rsc.org/advances

Graphical Abstract

Classification Study of Solvation Free Energy of Organic Molecules with Machine Learning Techniques

N.S. Hari Narayana Moorthy*, Silvia A. Martins, Sergio F Sousa, Maria J. Ramos, Pedro A. Fernandes

Classification models to predict the solvation free energies of organic molecules were developed using different machine learning approaches (decision tree, random forest and support vector machine). MACCS fingerprints, MOE descriptors and PaDEL descriptors were used to construct the models.



RSC Advances Accepted Manuscript

Classification Study of Solvation Free Energy of Organic Molecules with Machine Learning Techniques

N.S. Hari Narayana Moorthy*, Silvia A. Martins, Sergio F Sousa, Maria J. Ramos, Pedro A. Fernandes

REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, s/n, Rua do Campo Alegre, 4169-007 Porto, Portugal.

RSC Advances Accepted Manuscript

**Corresponding author*

E-mail address: hari.moorthy@fc.up.pt, hari.nmoorthy@gmail.com, pafernan@fc.up.pt

Tel/Fax: +351-220 402 506

Abstract:

In this work, we have developed a list of classification models to categorize the organic molecules with respect to their solvation free energies using different machine learning approaches (decision tree, random forest and support vector machine). Those solvation free energy values (experimental values obtained from literatures) were splitted into highly favourable (<-3 kcal/mol) and less favourable (>-3 kcal/mol) solvation free energies of molecules, which was set as threshold value for the classification model development. The MACCS fingerprint along with a category of physicochemical descriptors such as atom count, topological, vdW surface area (volsurf) and subdivided surface area were contributed in the classification models. The validation studies by test set and 10-fold cross-validation methods provide statistical parameters such as accuracy, sensitivity and specificity with >90% significance. The sum of ranking difference (SRD) analysis reveals that the support vector machine models are comparatively significant, while the MACCS fingerprints possessed models are ranked as good models in all approaches. The MACCS fingerprints explained that the presence of halogen atoms causes less favourable solvation free energy generation. However, the presence of polar atoms/groups and some functional groups such as heteroatoms, double bonded branched aliphatic chains, C=N, N-C-C-O, NCO, >1 heterocyclic atoms, OCO, etc cause highly favourable solvation free energy generation. The results derived from these investigations would be used along with some quantitative models for the prediction of solvation free energies of organic molecules and to design novel molecules with acceptable solvation free energies.

Key words: Solvation free energy, MOE descriptors, MACCS fingerprints, random forest, support vector machine, decision tree.

Introduction:

Solvent accessible surface area is an important analysis tool for biologists to characterize the hydrophobic and/or hydrophilic nature of the exposed molecular surface. These surface area properties are used to calculate the solvation free energy of the molecules[1]. The majority of the biological processes take place in solution. The solvation effects are thus an essential part in the analysis of reactions that occurs in liquid phase, the water being the solvent par excellence. Solvation free energy (ΔG_{solv}) is the amount of energy necessary to transfer a molecule from gas to a solvated environment[2,3].

Protein-ligand binding and the transport of drugs across membranes are closely connected to the solvation free energy as it is an important component of binding free energy. The

RSC Advances Accepted Manuscript

molecules with importance to chemical, biological and pharmaceutical sciences are usually polyfunctional (ex. drug molecules). The exposure or protection of chemical groups from solvent influences the binding process and this involves desolvation of the ligand in the thermodynamic process. Therefore, the determination of ΔG_{solv} is a valuable objective, with major weight in the study of chemical/biochemical processes, pursued since the beginnings of computer-aided drug design[4,5]. The free energy of solvation (ΔG_{solv}) is an important thermodynamical property and the free energy created in the molecules from effect of the constitution of groups and the physicochemical features of the molecules[2]. Earlier experiments explained that the contribution of the electrostatic and the nonpolar parts of the molecules cause the solvation free energy of a molecule[6-11]. The nonpolar contribution is usually modelled as proportional to the solvation surface area. The electrostatic term dominates the total solvation free energy of the molecules, while it does not always mean a high affinity[12]. This showed that the physicochemical features in the molecules cause variation in the free energy of solvation for the molecules. Hence, the analysis was carried out to investigate the important physicochemical properties and topological structural features responsible for the formation of free energy of solvation. Further, the classification analysis (qualitative analysis) is used to categorize the molecules with their solvation free energy using different machine learning approaches (the quantitative analysis needs more computational cost, time consuming, precise experimental activities, etc). The reported quantitative models on the solvation free energy prediction were developed with high computing powers[13,14]. In order to simplify the analysis, the initial qualitative models developed with the same data set, can support the development of quantitative models with less time and precision.

Machine learning is a field of artificial intelligence, extract characteristic of interest from the data set of their unknown underlying probability distribution. Machine learning focuses on prediction, based on known properties learned from the training data[15]. These methods use different algorithms for the classification and are evaluated on its generalization capability, which is its ability to apply successfully the learnt knowledge to unseen data. Generally, supervised and non-supervised machine learning methods are available for classification analysis[16]. In the present study, we have used some supervised machine learning methods to classify highly favourable and less favourable solvation free energy of the organic molecules.

Computational methods

Data set

RSC Advances Accepted Manuscript

A data set comprised of 241 organic molecules with their experimental solvation free energy was retrieved from the literatures (Table S1)[3,4,12,13,17-22]. The Molecular Operating Environment (MOE) software was used to calculate the physicochemical descriptors of the molecules. The semi-empirical MOPAC program with Hamiltonian Austin Model 1 (AM1) force field with 0.05 RMS gradients of MOE software was used to optimize the molecules for vsurf descriptors calculation[23,24]. Additionally, 2D descriptors of the molecules were calculated using the PaDEL software[25].

MACCS fingerprints

MACCS fingerprints for the data set compounds were calculated using PaDEL software. It indicates one of the 166 MACCS structural keys computed from the molecular graph and it represents as a spare list of keys present in the molecules.

Machine learning methods

In this study, the support vector machine, random forest and decision tree approaches were used for the classification analysis with the help of Weka software[26]. Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning. It is a classification method, which predicts the value of a dependent attribute (variable) through the given values of the independent (input) attributes (variables). Decision tree classifies instances by sorting them down the tree from the root to some leaf node.

Breiman (2001) proposed an ensemble learning method for classification (and regression) is called random forests that operated by constructing a multitude of decision trees at training time. Random forests change how the classification or regression trees are constructed using different bootstrap samples. In standard trees, each node split using the best split among all variables. In a random forest, each node split using the best among a subset of predictors randomly chosen at that node[27,28].

Support vector machines are supervised learning models with associated learning algorithms for classification study, which are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class of memberships[29].

Sum of ranking difference (SRD):

The ranking analysis was performed using the softwares named CRRN_DNA and SRDrep (SRD with ties) (downloadable from: http://aki.ttk.mta.hu/srd or http://goliat.eik.bme.hu/~kollarne/CRRN. The calculated performance parameters (statistical

RSC Advances Accepted Manuscript

parameters) such as specificity, sensitivity, precision, accuracy, G-mean, F-measure and MCC were used for the SRD analysis of the developed models[30,31].

Results and Discussion:

Classification models

The ability to predict the solvation free energy of the molecules, classification models were developed on a data set comprised of 241 molecules using different approaches such as decision tree, random forest and support vector machine. The physicochemical descriptors calculated from MOE and PaDEL softwares and the MACCS fingerprints of the molecules were used as independent variables in the classification studies. There are 5 classification models using each approach (algorithm) and each model comprised of different descriptors such as MOE descriptors in model 1, fingerprints in model 2, PaDEL descriptors in model 3, fingerprint and MOE descriptors in model 4 and all the descriptors (MOE, PaDEL and fingerprints) in models 5 were developed. Initially, the descriptor pool was reduced using stepwise regression and principal component analysis methods. The pruned descriptors were used for the classification studies and the descriptors contributed in the models are provided in Table 1. Many models were constructed with different descriptors and approaches, because single model/method does not give best result for any data set, hence multiple models/methods are needed to construct classification models and also required to compare their results.

In this analysis, solvation free energy values such as <-3 kcal/mol as highly favourable and >-3 kcal/mol as less favourable of the molecules were set as threshold values for the classification studies. The results derived from all three approaches are provided in Table 2 and 3. However, the analysis also performed with other threshold values (<-1 kcal/mol and >-1 kcal/mol), unfortunately that data set did not provide balanced number of compounds as highly favourable and less favourable solvation free energies. The abovementioned threshold value (<-3 kcal/mol and >-3 kcal/mol) have yielded better classification models with significant statistical parameter, hence those models are discussed here. The physicochemical descriptors contributed in the models significantly classified the solvation free energies of molecules as highly favourable and less favourable.

All the developed classification models were validated by 10 fold cross-validation and test set methods. In test set method, 30% and 40% of the molecules in the data set were considered as test set to validate the models. The classification performances of the models constructed through all the methods were observed through the confusion matrix. These models classified >95% of the molecules exactly as highly favourable and less favourable solvation free

RSC Advances Accepted Manuscript

energies containing molecules. The true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) of the classified molecules are provided in Table 3-4. The statistical parameters such as sensitivity, specificity, precision and negative predicted values calculated from the confusion matrix are >0.9 for all the models in response to the training set, the test set and the 10 fold cross-validation analyses. Matthew's correlation coefficient (MCC) measures the quality of a classification model by calculating the value between -1 and +1. MCC value 0 is average or random prediction, -1 is worst prediction, and +1 is perfect prediction[32,33]. An MCC value above 0.4 is considered to be predictive in classification studies. In our analysis, the MCC values are >0.95 for all the models (through all the methods) reveal that the models are significant. In random forest method, some models provided the MCC values are 1 represent that those models classified the molecules perfectly. The G-mean is a statistical parameter measures the overall performance of models and used to check the balanced prediction of highly favourable and less favourable solvation free energy of the molecules. These models yielded the G-mean values >0.95 explain that the models predict the molecules in balanced way. As the discussed statistical parameters, the models provided significant accuracy of >0.95 against all the studied classification methods. These results reveal that the developed models with the physicochemical descriptors and fingerprints have classified the molecules perfectly as highly favourable and less favourable solvation free energies of the molecules.

In order to compare the performance of each method (and models), the SRD values was calculated for all the models. This ranking is intended to compare models, methods, techniques, etc with some scaled (calculated) variables. SRD values provide a refined scale for ranking even with very small differences among the results (methods, models, etc). A value closer to zero (golden standard) indicates the better is the model, when it has proximity value indicates similarity of the variables. The results derived from the SRD analysis are provided in Table 4 and graphically represented in Figure 1. These results reveal that the support vector machine approach provided significant results and these models ranked as better models. It is interesting that the models developed with only fingerprint descriptors provided better SRD values (with all the studied approaches). Hence, it is important to investigate the kind of fingerprints (substructures, atoms, groups, etc) present in highly favourable and less favourable solvation free energy containing molecules, which are useful to design novel molecules with appropriate solvation free energies.

MACCS fingerprint analysis:

RSC Advances Accepted Manuscript

In order to understand the structure property relationship, the MACCS fingerprint was calculated for all the molecules based on their molecular structure. The frequency of each fingerprints (substructure) appearing in the molecules with highly favourable and less favourable solvation free energy was calculated. This provides the important substructure/functional group/atoms responsible for the observed (increased and decreased) variation of solvation free energies of the molecules. Some substructures are present in both kind of molecules (highly favourable and less favourable), while limited fingerprints present in highly favourable or the less favourable solvation free energies pertaining molecules. The details of those fingerprints present in the molecules with different solvation free energies (threshold) are graphically represented in figure 2. The graphs explained that the molecules have the solvation free energy of <-5 kcal/mol exhibited specifically the following substructures such as heteroatoms, double bonded branched aliphatic chains, C=N, N-C-C-O, NCO, >1 heterocyclic atoms, OCO, etc. These substructures are absent in those molecules have the free energy values >-5 kcal/mol.

Interestingly, it has been observed that the presence of halogen atom in the molecules cause increased solvation free energies (>-3 kcal/mol) to the molecules. The presence of aliphatic long chains in the molecules has high solvation free energy than aromatic or aliphatic ring containing compounds. These results described that the presence of these substructures in the molecules have variation of solvation free energies.

Description of the contributed descriptors

The classification models generated in this analysis possessed descriptors from different categories to classify molecules according to their highly favourable and less favourable solvation free energies. Those descriptors are categorized below.

Atom count descriptors (a_nH, A_nN and nN): These atom count descriptors count number of nitrogen and hydrogen atoms in the molecules[24].

Topological descriptors (KierA3, SHdsCH and ETA_AlphaP): The KierA3 descriptor describes shape of the molecules with third alpha shape index. It calculated by $(s-1)(s-3)^2/p_3^2$ for odd n and $(s-3)(s-2)^2/p_3^2$ for even n, where s = n+a. However, Kier and Hall kappa molecular shape indices compare the minimal and maximal molecular graphs and are intended to capture different aspects of molecular shape[24].

The electrotopological state descriptors are designated through E-state symbol that composite of three parts. The first part is "S" which stands for the sum of E-state values for all atoms of the same type in the molecule. The second part is a string representing the bond types

RSC Advances Accepted Manuscript

associated with that atom (“s” for single bond, “d” for double, “t” for triple and “a” for aromatic). Finally, there is a symbol for the set of atoms in the hydride group, such as CH_3, CH_2, OH, Br, or NH. The SHdsCH is the atom-type hydrogen electrotopological state index for =CH- groups[34-36]. Another descriptor present in this category is ETA_AlphaP, an extended topochemical atom (ETA) indices, which are a group of topological descriptors from modification and refinement of the topologically arrived unique (TAU) scheme parameters of the 1980s[37,38].

Polar descriptors (Lip_acc, Lip_don, nHBAcc_Lipinski, nHBDon, TopoPSA and Vsa_pol): These descriptors explain the number of hydrogen bond acceptor and hydrogen bond donor atoms/groups present in the molecules. The TopoPSA and Vsa_pol descriptors describe the polar properties on the van der Waals (vdW) surface area of the molecules[24].

Subdivided surface area descriptors (SlogP_VSA0, SMR_VSA1 and SMR_VSA7)

The subdivided surface area descriptors are based on an approximate accessible vdW surface area (VSA) calculation (in Å^2) for each atom, v_i along with other atomic property, P_i (either partition coefficient or molar refractivity). The v_i values are calculated using a connection table approximation. The properties (P_i) of small molecules can be calculated as the sum of the contributions of each of the atoms in the molecule as per (1).

$$P\_VSA_k = \sum V_i \delta(P_i \Sigma(a_k\text{-}1,\ a_k)) \qquad k = 1,2,3,\ \ldots,\ n. \qquad (1)$$

where $a_o<a_k<a_n$ are interval boundaries such that (a_o, a_n) bound are values of P_i in any molecule. Each VSA type descriptor can be characterized as the amount of surface area with P in a certain range. SlogP_VSA and SMR_VSA descriptors explain the partition coefficient and molecular refractivity respectively on vdW surface area of the molecules. These are defined to be the sum of the v_i over all atoms *i*. P_i denotes the contribution to partition coefficient molar refractivity for atom *i* as calculated in the SlogP or SMR descriptor, calculated in a specified range[24,39].

Volsurf descriptors (Vsurf_CW1, Vsurf_CW2, Vsurf_CW3, Vsurf_W2, Vsurf_Wp3, Vsurf_EWmin1 and Vsurf_A): The vsurf like descriptors depend on the structural connectivity and the conformation (dimensions are measured in Å) of the molecules. It generally describes the hydrophobic and hydrophilic properties mediated by surface properties such as shape, electrostatic, hydrogen-bonding and hydrophobicity. The vsurf_CW descriptor describes the capacity factor of the molecules and is calculated in different energy levels. It provides information on the amount of hydrophilic regions per unit surface[40,41].

RSC Advances Accepted Manuscript

The vsurf_Wp descriptor describes the polar volume (either polarizability and dispersion forces or hydrogen bond acceptor-donor regions) of the molecule and are calculated at eight different energy levels (-0.2, -0.5, -1.0, -2.0, -3.0, -4.0, -5.0 and -6.0 kcal/mol) and may be defined as the molecular envelope accessible by solvent (water) molecules. Other vsurf descriptors such as Vsurf_A and vsurf_EWmin1 represent the amphiphilic moment and lowest hydrophilic energy of the molecules respectively.

MACCS fingerprints: The MACCS fingerprints explain the presence or absence of particular functional groups, atom or fragments on different molecules. Those contributed fingerprints explain the following structural information of the compounds. MACCSFP49 (charge on the molecule), MACCSFP88 (presence of sulphur atoms), MACCSFP103 (presence of chlorine atom), MACCSFP104 (hetero atom with hydrogen and connected with CH_2 through any other atom), MACCSFP107 (halogen atom connected with branched atoms (any atom (any atom)+ any atom)), MACCSFP121 (nitrogen containing heterocycles), MACCSFP127 (any atom+ ring bond + any atom + non ring bond connected with O_2), MACCSFP134 (halogens), MACCSFP139 (hydroxyl group), MACCSFP151 (-NH group), MACCSFP156 (N connected with branched atom as any atom (any atom)+ any atom), MACCSFP157 (C-O) and MACCSFP161 (nitrogen atom)

Conclusion:

This study concluded that all the developed models provided >90% significance on the statistical parameters such as sensitivity, specificity, MCC, accuracy, G-mean, etc. The frequency of appearance of MACCS fingerprints in the molecules explained the substructures/groups/atoms responsible for the change of solvation free energy of the molecules. Multiple methods and models are reported in the study because a single method/model can't provide significant prediction. The SRD values showed that all the models have similar performance on the dataset classification, however, the support vector machine showed slightly better performance than other methods.

Our analysis has performed with easily calculable descriptors and freely available modelling tools. Earlier report for the prediction of solvation free energies of the molecules are quantitative models, which was calculated with high computational algorithms[13,14]. The results obtained from our studies are also significant and can be improved with sophisticated methods and algorithms, which will be used along with other quantitative studies to reduce the computing power and time consumption. Further, it supports the investigation of quantitative models for the prediction of solvation free energies of organic molecules and to design novel molecules with acceptable solvation free energies.

RSC Advances Accepted Manuscript

Acknowledgement

N.S.H.N. Moorthy is grateful to the Fundaçao para a Ciencia e Technologia (FCT), Portugal for a Postdoctoral Grant (SFRH/BPD/44469/2008).

References

1. L. Cavallo, J. Kleinjung, F. Fraternali, *Nucleic Acids Res.,* 2003, **31,** 3364–3366.
2. P. F. B. Gonçalves, H. Stassen, *Pure Appl. Chem.,* 2004, **76,** 231–240.
3. S. Lee, K. H. Cho, C. J. Lee, G. E. Kim, C. H. Na, Y. In, K. T. No, *J. Chem. Inf. Model.,* 2010, **51,** 105–114.
4. R. C. Rizzo, T. Aynechi, D. A. Case, I. D. Kuntz, *J. Chem. Theory Computat.,* 2006, **2,** 128–139
5. D. S. Palmer, V. P. Sergiievskyi, F. Jensen, M. V. Fedorov, *J. Chem. Phys.,* 2010, **133,** 044104. doi: 10.1063/1.3458798.
6. B. Honig, A. Nicholls, *Science,* 1995, **268,** 1144–1149.
7. M. K. Gilson, B. Honig, *Proteins,* 1998, **4,** 7–18.
8. C. J. Cramer, D. G. Truhlar, *Chem. Rev.*, 1999, **99,** 2161–2200.
9. C. J. Cramer, D. G. Truhlar, *Science,* 1992, **256,** 213–217.
10. D. Sitkoff, K. A. Sharp, B. Honig, *J. Phys. Chem.,* 1994, **98,** 1978–1988.
11. R. Luo, J. Moult, K. Gilson, *J. Phys. Chem. B,* 1997, **101,** 11226–11236.
12. J. Wang, W. Wang, S. Huo, M. Lee, P. A. Kollman, *J. Phys. Chem.* B, 2001, **105,** 5055–5067.
13. V. N. Viswanadhan, A. K. Ghose, U. C. Singh, J. J. Wendoloski, *J. Chem. Inf. Comput. Sci.,* 1999, **39,** 405–412.
14. L Bernazzani, C. Duce, A. Micheli, V. Mollica, A. Sperduti, A. Starita, M. R. Tine, *J. Chem. Inf. Model.,* 2006, **46,** 2030–2042.
15. I. H. Witten, E. Frank, M. A. Hall. Data mining: Practical machine learning tools and techniques. Burlington, MA: Morgan Kaufmann, 2011.
16. J. Han, M. Kamber, Data mining:concepts and techniques. San Francisco: Morgan Kaufmann Publishers, 2001.
17. S. Cabani, P. Gianni, V. Mollica, L. Lepori, *J. Solution Chem.,* 1981, **10,** 563–595.
18. R. Wolfenden, L. Andersson, P. M. Cullis, C. C. G. Southgate, *Biochemistry* 1981, **20,** 849–855.
19. E. Gallicchio, L. Y. Zhang, R. M. Levy, *J. Comput. Chem.,* 2002, **23,** 517–529.

RSC Advances Accepted Manuscript

20. W. L. Jorgensen, J. P. Ulmschneider, J. Tirado-Rives, *J. Phy. Chem. B,* 2004, **108,** 16264–16270.

21. A. V. Marenich, C. J. Cramer, D. G. Truhlar, *J. Phy. Chem. B,* 2009, **113,** 4538–4543.

22. E. O. Purisima, C. R. Corbeil, T. Sulea, *J. Chem. Theory Computat.,* 2010, **6,** 1622–1637

23. MOE 2012, Chemical Computing Group Inc. Montreal, H3A 2R7, Canada, 2012.

24. A. Lin, QuaSAR-descriptors. Chemical Computing Group Inc. Montreal, H3A 2R7 Canada, 2002.

25. C. W. Yap, *J. Computat. Chem*., 2011, **32**, 1466–1474.

26. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, *SIGKDD Explorations,* 2009, **11,** 10–18.

27. L. Breiman, *Machine Learning,* 2001, **45(1),** 5–32.

28. L. Breiman. *Machine Learning,* 1996, **24(2),** 123–140.

29. C. Cortes, V. Vapnik, *Machine Learning,* 1995, **20(3),** 273–297.

30. K. Heberger, K. Kollár-Hunek, *J. Chemometr.,* 2011, **25,** 151–158.

31. K. Kollár-Hunek, K. Héberger, *Chemom. Intell. Lab. Syst.,* 2013, **127,** 139–146.

32. A. Jurik, R. Reicherstorfer, B. Zdrazil, G. F. Ecker, *Mol. Inf.*, 2013, **32,** 415–419.

33. K. M. Thai, G. F. Ecker, *Bioorg. Med. Chem.*, 2008, **16,** 4107–4119.

34. N. S. H. N. Moorthy, M. J. Ramos, P. A. Fernandes, *J. Enz. Inhibit. Med. Chem.,* 2011, **26(6),** 755–766.

35. N. S. H. N. Moorthy, M. J. Ramos, P. A. Fernandes, *Chemom. Intell. Lab. Sys.,* 2011, **109,** 101–112.

36. N. S. H. N. Moorthy, M. J. Ramos, P. A. Fernandes, *Lett. Drug Des. Discov.,* 2011, **8,** 14–25.

37. K. Roy, R. N. Das, *J. Hazardous Materials,* 2013, **254–255,** 166–178.

38. K. Roy, G. Ghosh, *Chemosphere,* 2009, **77(7),** 999–1009.

39. N. S. H. N. Moorthy, S. F. Sousa, M. J. Ramos, P. A. Fernandes, *J. Enz. Inhibit. Med. Chem.,* 2011, **26(6),** 777–791.

40. N. S. H. N. Moorthy, M. J. Ramos, P. A. Fernandes, *RSC Adv,* 2011, **1,** 1126–1136.

41. N. S. H. N. Moorthy, M. J. Ramos, P. A. Fernandes, *SAR QSAR Environ. Res.,* 2012, **23(5-6),** 521–136.

RSC Advances Accepted Manuscript

Figure 1: SRD-CRRN results of the of the classification models

1-5=Decision Tree models 1-5; 6-10=Support Vector Machine Models 1-5; 11-15=Random Forest models 1-5; XX1—first icosaile (5%), Q1—first quartile, Med—median, Q3—last quartile, XX19—last icosaile (95%).

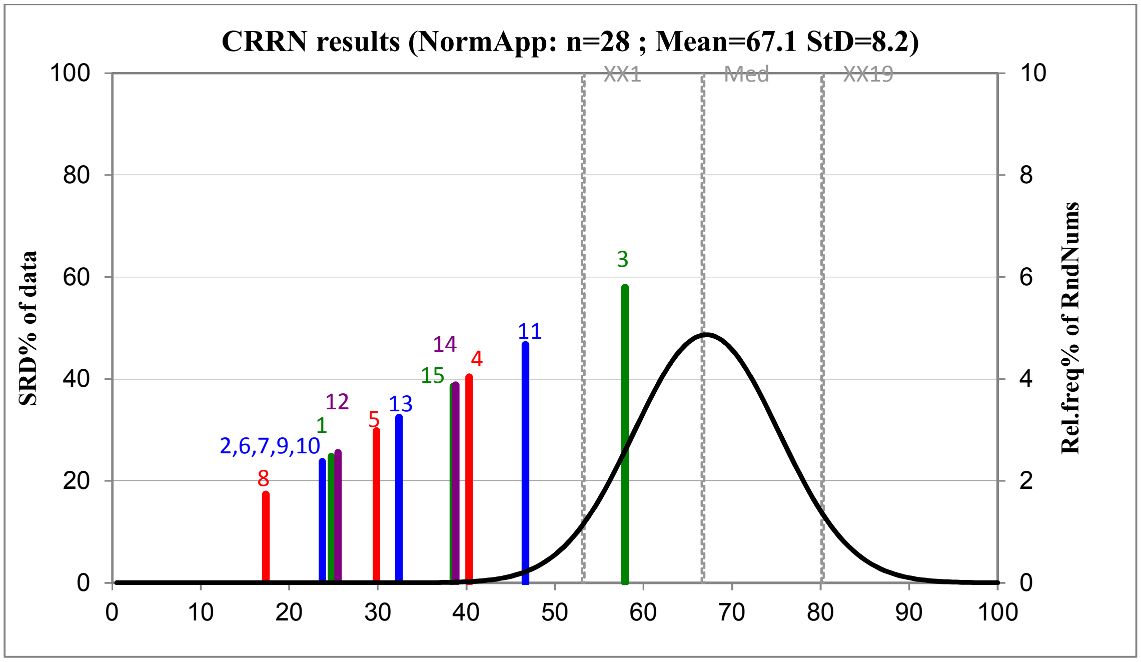
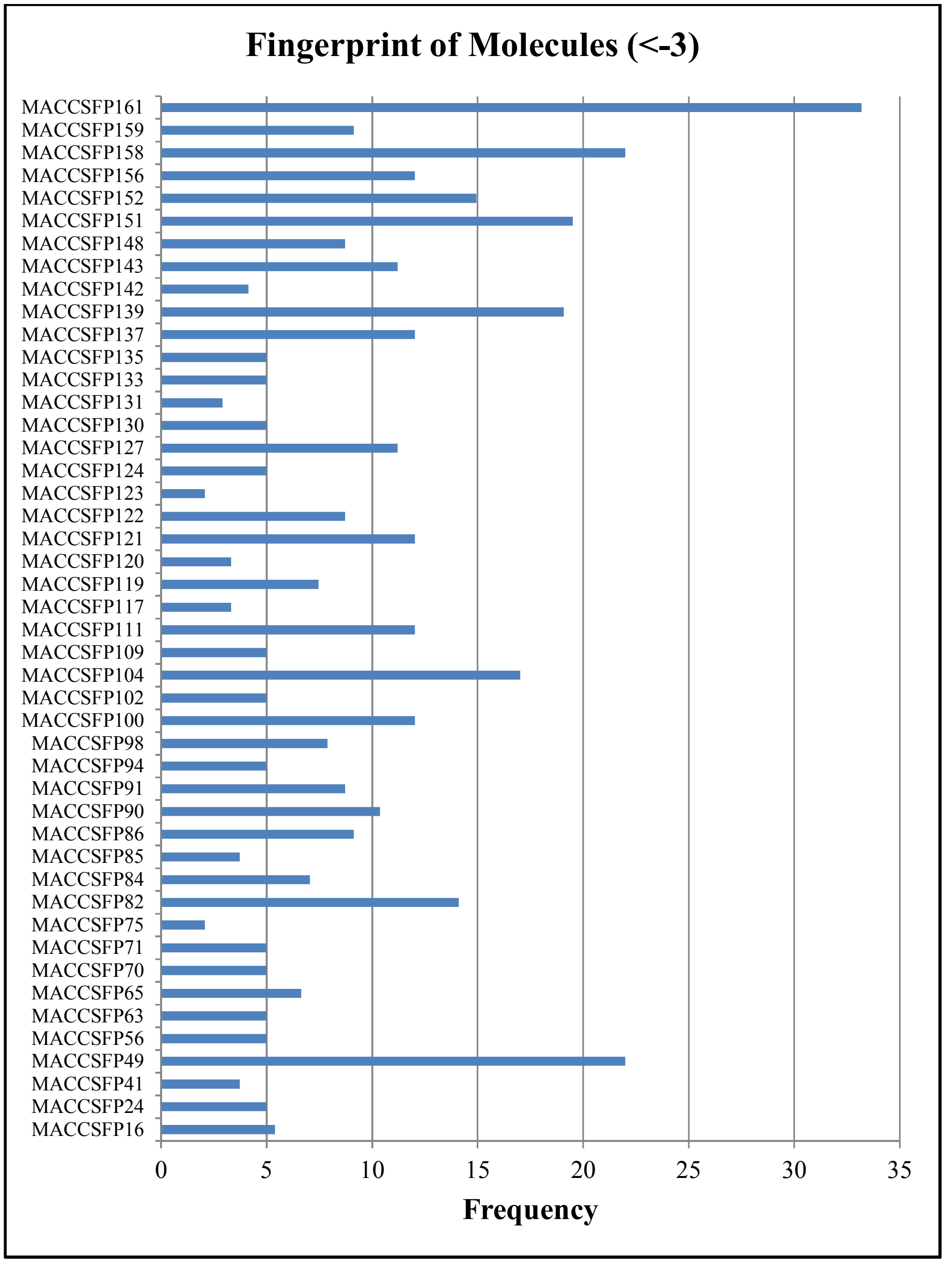


Figure 2: Graphical representation of frequency of fingerprints on the molecules

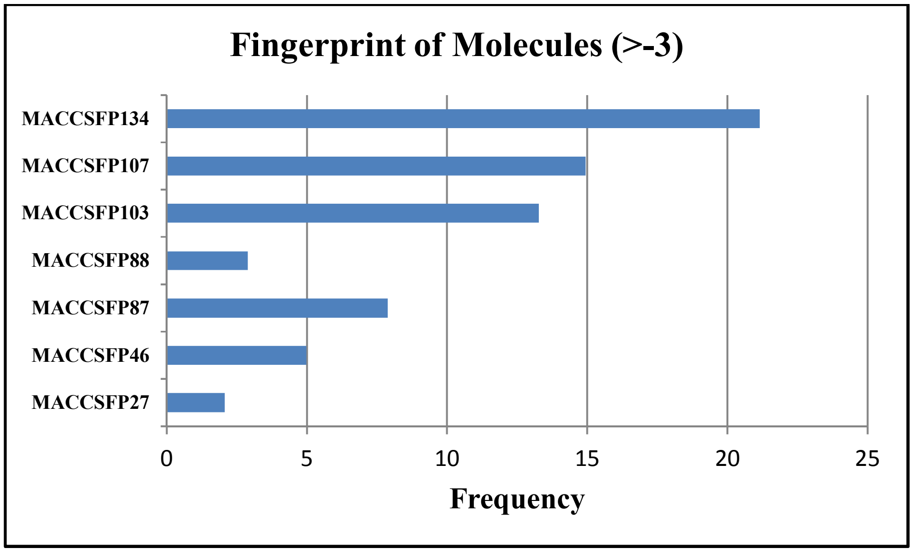
A

RSC Advances Accepted Manuscript



RSC Advances Accepted Manuscript

B



RSC Advances Accepted Manuscript

Table 1: Physicochemical and fingerprint descriptors contributed in each models

Model 1	Model 2	Model 3	Model 4	Model 5
a_nH	MACCSFP49	nN	MACCSFP134	nN
a_nN	MACCSFP88	SHdsCH	MACCSFP161	nHBAcc_Lipinski
KierA3	MACCSFP103	ETA_AlphaP	a_nN	nHBDon
Lip_acc	MACCSFP104	nHBAcc_Lipinski	Lip_acc	MACCSFP134
Lip_don	MACCSFP107	nHBDon	Lip_don	MACCSFP161
SlogP_VSA0	MACCSFP121	TopoPSA	SlogP_VSA0	SlogP_VSA0
SMR_VSA1	MACCSFP127		SMR_VSA7	SMR_VSA7
Vsurf_CW3	MACCSFP134		Vsurf_EWmin1	Vsa_pol
Vsurf_W2	MACCSFP139		Vsurf_W2	Vsurf_A
	MACCSFP151		Vsurf_Wp3	Vsurf_CW1
	MACCSFP156			Vsurf_CW2
	MACCSFP157			
	MACCSFP161			

RSC Advances Accepted Manuscript

Table 2: Confusion matrix of the classification models

Mod			Decision Tree		Support vector Machine		Random Forest	
el			Total P	Total N	Total P	Total N	Total P	Total N
No	Dataset	Total	(TP/FN)	(TN/FP)	(TP/FN)	(TN/FP)	(TP/FN)	(TN/FP)
1	**Training**	241	129 (128/1)	112 (111/1)	129 (122/7)	112 (110/2)	129 (129/0)	112 (112/0)
	Test 30%	72	41 (40/1)	31 (31/0)	41 (39/2)	31 (30/1)	41 (41/0)	31 (31/0)
	Test 40%	96	53 (51/1)	43 (43/0)	53 (51/2)	43 (42/1)	53 (52/1)	43 (42/1)
	10-fold	241	129 (124/5)	112 (109/3)	129 (122/7)	112 (110/2)	129 (123/6)	112 (109/3)
2	**Training**	241	129 (122/7)	112 (110/2)	129 (122/2)	112 (110/2)	129 (121/8)	112 (111/1)
	Test 30%	72	41 (39/2)	31 (30/1)	41 (39/2)	31 (30/1)	41 (39/2)	31 (30/1)
	Test 40%	96	53 (51/2)	43 (42/1)	53 (51/2)	43 (42/1)	53 (51/2)	43 (42/1)
	10-fold	241	129 (122/7)	112 (110/2)	129 (122/7)	112 (110/2)	129 (121/8)	112 (111/1)
3	**Training**	241	129 (127/2)	112 (110/2)	129 (122/7)	112 (110/2)	129 (127/2)	112 (112/0)
	Test 30%	72	41 (41/0)	31 (30/1)	41 (39/2)	31 (30/1)	41 (40/1)	31 (30/1)
	Test 40%	96	53 (53/0)	43 (42/1)	53 (51/2)	43 (42/1)	53 (53/0)	43 (42/1)
	10-fold	240	129 (127/2)	111 (109/2)	129 (123/6)	112 (107/2)	129 (125/4)	112 (109/3)
4	**Training**	241	129 (129/0)	112 (110/2)	129 (122/7)	112 (110/2)	120 (120/0)	112 (112/0)
	Test 30%	72	41 (39/2)	31 (30/1)	41 (39/2)	31 (30/1)	41 (40/1)	31 (31/0)
	Test 40%	96	53 (52/1)	43 (40/3)	53 (51/2)	43 (42/1)	53 (52/1)	43 (41/2)
	10-fold	241	129 (124/5)	112 (106/6)	129 (122/7)	112 (110/2)	129 (123/6)	113 (110/3)
5	**Training**	241	129 (127/2)	112 (109/3)	129 (122/7)	112 (110/2)	129 (129/0)	112 (112/0)
	Test 30%	72	41 (39/2)	31 (30/1)	41 (39/2)	31 (30/1)	41 (41/0)	31 (30/1)
	Test 40%	96	53 (51/2)	43 (41/2)	53 (51/2)	43 (42/1)	53 (52/1)	43 (43/0)
	10-fold	241	129 (125/4)	112 (103/9)	129 (122/7)	112 (110/2)	129 (125/4)	112 (107/5)

Total P : Total Positives, Total N: Total Negatives, TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives

RSC Advances Accepted Manuscript

Table 3: Statistical parameters calculated through classification analysis

Model	Data	Specificity			Sensitivity			Accuracy			Precision			G-mean			F-measure			MCC		
No	set	DT	SV	RF	DT	SV	RF	DT	SV	RF	DT	SV	RF	DT	SV	RF	DT	SV	RF	DT	SV	RF
1	Training	0.99	0.98	1.00	0.99	0.94	1.00	0.99	0.96	1.00	0.99	0.98	1.00	0.99	0.96	1.00	0.99	0.96	1.00	0.98	0.93	1.00
	Test (30%)	0.97	0.98	1.00	1.00	0.94	1.00	0.99	0.96	1.00	1.00	0.97	1.00	0.99	0.96	1.00	0.99	0.96	1.00	0.97	0.92	1.00
	Test (40%)	0.98	0.98	0.98	1.00	0.95	0.98	0.99	0.97	0.98	1.00	0.98	0.98	0.99	0.97	0.98	0.99	0.97	0.98	0.98	0.94	0.96
	10-fold	0.96	0.98	0.95	0.98	0.94	0.98	0.97	0.96	0.96	0.98	0.98	0.98	0.97	0.96	0.96	0.97	0.96	0.96	0.93	0.93	0.93
2	Training	0.94	0.98	0.93	0.98	0.94	0.99	0.96	0.96	0.96	0.98	0.98	0.99	0.96	0.96	0.96	0.96	0.96	0.96	0.93	0.93	0.93
	Test 30%	0.94	0.98	0.94	0.98	0.94	0.98	0.96	0.96	0.96	0.97	0.97	0.97	0.96	0.96	0.96	0.96	0.96	0.96	0.92	0.92	0.92
	Test 40%	0.95	0.98	0.95	0.98	0.95	0.98	0.97	0.97	0.97	0.98	0.98	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.94	0.94	0.94
	10-fold	0.94	0.98	0.93	0.98	0.94	0.99	0.96	0.96	0.96	0.98	0.98	0.99	0.96	0.96	0.96	0.96	0.96	0.96	0.93	0.93	0.93
3	Training	0.98	0.98	0.98	0.98	0.94	1.00	0.98	0.96	0.99	0.98	0.98	1.00	0.98	0.96	0.99	0.98	0.96	0.99	0.97	0.93	0.98
	Test 30%	1.00	0.98	0.97	0.98	0.94	0.98	0.99	0.96	0.97	0.98	0.97	0.97	0.98	0.96	0.97	0.99	0.96	0.98	0.97	0.92	0.94
	Test 40%	1.00	0.98	1.00	0.98	0.95	0.98	0.99	0.97	0.99	0.98	0.98	0.98	0.99	0.97	0.99	0.99	0.97	0.99	0.98	0.94	0.98
	10-fold	0.98	0.96	0.96	0.98	0.95	0.98	0.98	0.95	0.97	0.98	0.96	0.98	0.98	0.95	0.97	0.98	0.96	0.97	0.97	0.91	0.94
4	Training	1.00	0.98	1.00	0.98	0.94	1.00	0.99	0.96	1.00	0.98	0.98	1.00	0.99	0.96	1.00	0.99	0.96	1.00	0.98	0.93	1.00
	Test 30%	0.94	0.98	0.97	0.98	0.94	1.00	0.96	0.96	0.99	0.97	0.97	1.00	0.96	0.96	0.99	0.96	0.96	0.99	0.92	0.92	0.97

RSC Advances Accepted Manuscript

	Test 40%	0.98	0.98	0.98	0.95	0.95	0.96	0.96	0.97	0.97	0.94	0.98	0.96	0.96	0.97	0.97	0.96	0.97	0.97	0.92	0.94	0.94
	10-fold	0.95	0.98	0.95	0.95	0.94	0.98	0.95	0.96	0.96	0.95	0.98	0.98	0.95	0.96	0.96	0.96	0.96	0.96	0.91	0.93	0.93
5	**Training**	0.98	0.98	1.00	0.98	0.94	1.00	0.98	0.96	1.00	0.98	0.98	1.00	0.98	0.96	1.00	0.98	0.96	1.00	0.96	0.93	1.00
	Test 30%	0.94	0.98	1.00	0.98	0.94	0.98	0.96	0.96	0.99	0.97	0.97	0.98	0.96	0.96	0.98	0.96	0.96	0.99	0.92	0.92	0.97
	Test 40%	0.95	0.98	0.98	0.96	0.95	1.00	0.96	0.97	0.99	0.96	0.98	1.00	0.96	0.97	0.99	0.96	0.97	0.99	0.92	0.94	0.98
	10-fold	0.96	0.98	0.96	0.93	0.94	0.96	0.95	0.96	0.96	0.93	0.98	0.96	0.94	0.96	0.96	0.95	0.96	0.97	0.89	0.93	0.92

DT: Decision Tree, SV: Support Vector Machine, RF: Random Forest

RSC Advances Accepted Manuscript

Table 4: Sum of ranking difference (SRD) and p% interval of the variables of the classification analyses

Ranking results			%p	
Model No (Original)	Model code	SRD	x < SRD > =x	
SV-3	8	68	6.46E-08	7.69E-08
DT-2	2	93	6.01E-06	7.22E-06
SV-1	6	93	6.01E-06	7.22E-06
SV-2	7	93	6.01E-06	7.22E-06
SV-4	9	93	6.01E-06	7.22E-06
SV-5	10	93	6.01E-06	7.22E-06
DT-1	1	97	1.18E-05	1.41E-05
RF-2	12	100	1.96E-05	2.28E-05
DT-5	5	117	2.73E-04	3.20E-04
RF-3	13	127	1.14E-03	1.32E-03
RF-5	15	151	2.43E-02	2.75E-02
RF-4	14	152	2.75E-02	3.06E-02
DT-4	4	158	5.40E-02	6.00E-02
RF-1	11	183	0.63	0.70
	XX1	209	4.94	5.25
DT-3	3	227	13.08	13.76
	Q1	240	24.61	25.61
	Med	262	49.91	51.15
	Q3	284	74.25	75.26
	XX19	315	94.70	95.02

DT: Decision Tree, SV: Support Vector Machine, RF: Random Forest, XX1—first icosaile (5%), Q1—first quartile, Med—median, Q3—last quartile, XX19—last icosaile (95%).

RSC Advances Accepted Manuscript