International Journal of Quantitative Structure-Property Relationships
Volume 5 • Issue 3 • July-September 2020

Modeling Binary Fingerprint Descriptors With the Superposing Significant Interaction Rules (SSIR) Method

Emili Besalú, University of Girona, Girona, Spain
https://orcid.org/0000-0003-0093-5714

ABSTRACT

Recently, the superposing significant interaction rules (SSIR) method has been applied in several fields of QSPR to model and establish molecular rankings that correlate dichotomous properties. The origin of the method is in the field of combinatorial chemistry, but it has been shown that the procedure is fast, versatile, and that it can be applied in many other fields. In particular, an example is phospholipidosis modeling taking, as primary descriptors, the binary fingerprints of the molecules. This is the first time SSIR is used to treat this kind of descriptors. The performance achieved is similar to other results found in the literature and, in particular, to the results obtained by authors who considered the same molecular set and descriptors. One of the main advantages of SSIR is that the method acts as an automated variable selector. This allows it to be used almost immediately without prior selection of variables.

KEYWORDS

AU-ROC Curve, Binary Fingerprints, Consensus, Phospholipidosis, Ranking, Rules, SSIR Method, Votes

1. INTRODUCTION

This paper constitutes a brief demonstration of how the Superposition of Significant Interaction Rules (SSIR) method works. In this perspective article, it is not intended to review the details and specifications of the SSIR procedure, as they are explained in several places (Besalú, 2016; Besalú et al., 2016; Besalú et al., 2017; Besalú et al., 2018). Specifically, the objective is to promote the method by showing a simple application example and its performance when handling binary fingerprints. It is the first time the method is used to rank compounds being described by binary parameters.

The molecular descriptors chosen are treated in order to model Phospholipidosis (PLD), a disorder that can be induced by several drugs. PLD is characterized by the accumulation of the inducing drug and phospholipids in the lysosomes of the affected tissues. Pharmaceutical companies need to conduct regular screenings on their drug candidates in order to avoid this side effect (Goracci et al., 2013). Several experimental and *in silico* methods have been described in order to predict the potential capacity of some drugs to be inducers or non-inducers (Ploemen et al. 2004; Tomizawa et al. 2006;

DOI: 10.4018/IJQSPR.20200701.oa2

Pelletier et al. 2007; Kruhlak et al. 2008; Hanumegowda et al. 2010; Lowe et al 2010; Fisher et al. 2012; Lowe et al 2012; Sun et al. 2012; Orogo et al. 2012). The work of the scientific community is still in progress (Przybylak et al., 2014) regarding the definition of good predictors and even the correct classification of drugs as potential inducers. As described in the literature, it is difficult to build quality models mainly due to the fact that the phospholipidosis induction mechanism is not well known. In addition, a single curated reference molecular set is not available. It may be possible that the set proposed by the Goracci's team is the first successful attempt to obtain it. One of the goals of this perspective article is to show how SSIR performs in this field and how it can be used to quickly rank compounds. The rankings can be used alone or can be used as preliminary molecular filters.

2. MATERIAL AND METHODS

The SSIR method is a variable selector based on the hypergeometric experiment (Mendenhall & Sincich, 1995). Briefly, given an urn containing red and green marbles, this probabilistic experiment consists of randomly selecting some of those marbles and, subsequently, assessing the probability associated with the distribution of red and green marbles which has been picked up. Specifically, the urn originally contains *a* marbles (*b* of them are green and the rest are red) and *c* are randomly selected. After the selection, one realizes that *d* of the marbles extracted are green. The probability for the described event follows the hypergeometric probability distribution:

$$P\left(d,c;b,a\right)=\frac{\begin{pmatrix}b\\d\end{pmatrix}\begin{pmatrix}a-b\\c-d\end{pmatrix}}{\begin{pmatrix}a\\c\end{pmatrix}}\text{ with } d\leq c\leq a \text{ and } d\leq b\leq a \tag{1}$$

where the minimum allowed value for *d* is max(0,*c*+*b*-*a*), and the maximum is min(*b*,*c*).

The SSIR method works similarly according to the following analogies: the urn is a molecular database, the marbles are molecules and the condition of being a 'green' or a 'red' marble is equivalent to being a 'interest' molecule or 'of no interest' (or vice versa, since this classification is arbitrary). Note that the procedure applies to dichotomized molecular sets. When dealing with continuous variables, the dichotomization is defined by the user establishing a cut-off value. The process of random extraction of balls is equivalent to selecting *a priori* some descriptors (one or several, as it will be explained below) and, at the same time, specifying which are the levels or range of values each descriptor must have. This tandem of descriptor/s and respective levels constitutes a rule. So, a rule defines a molecular condition: a molecule will conform (fulfill) with the rule or not. All molecules conforming with the rule are (virtually) being 'extracted' (i.e., the rule is the extractor agent). Then, within the set of 'extracted' or selected molecules some will be 'green' and some will be 'red'. Ultimately, the hypergeometrical Formula (1) provides the probability of obtaining the final proportion of collected molecules of each type. The 'difficulty' inherent to the extraction of at least a minimum number of molecules of interest is measured by a *p*-value equivalent to the following addition:

$$p\left(d+,c;b,a\right)=p\left(d:\min\left(b,c\right),c;b,a\right)=\sum_{i=d}^{\min\left(b,c\right)}P\left(i,c;b,a\right)=1-\sum_{i=\max\left(0,c+b-a\right)}^{d-1}P\left(i,c;b,a\right) \tag{2}$$

where the notations *d*+ or, equivalently, *d*:min(*b*,*c*) stand for the event that involves the extraction of *d* or more molecules of interest.

The SSIR algorithm virtually performs a lot of successive extractions. Each extraction is performed without replacement of marbles/molecules (as in the hypergeometric theory framework). Due to the many possible rules (i.e. combinations of variables and levels) that can be generated given a set of descriptors, after each extraction has been inspected, there is a return of the elements in the urn. This leaves the urn unchanged to repeat the process with another rule, that is, perform another hypergeometric experiment and evaluate a new rule. The rules are so-called of order 1 if they involve a single descriptor. The rules of order 2 are those that combine the levels of two descriptors, and so on. Normally, only rules of, at most, the order 3 are to be considered. The SSIR user specifies a threshold or limit of *p*-value, p_c, and each rule that exhibits a *p*-value equal or less than this cut-off is declared 'significant' and is tagged with a positive vote. Ultimately, SSIR provides a consensus model formed by all the significant rules. In addition, a totally non-significant rule, i.e., its *p*-value satisfies that $1\text{-}p \leq p_c$ (*p*-value near to the unit) is also included in the consensus pool, but this time tagged with a negative vote (as it will be explained below). Then, given a molecule (even a compound external to the database), its descriptors face the rules of the consensus pool. It can be established if the compound conforms to the rules, one at a time. Each time a molecule conforms to a significant rule, it receives a positive (or negative) vote, the vote attached previously to the rule. The sum of positive and negative votes constitutes the molecular score. Finally, the molecules are ranked according to their scores.

SSIR can also be considered an automated variable selector. This is because when SSIR does not select a descriptor (because it is linked to a high *p*-value) the method is making a selection of variables *in situ*, in this case discarding the variable. That is one of the reasons why SSIR has been used successfully without previously performing any type of data monitoring.

3. APPLICATION EXAMPLE: PHOSPHOLIPIDOSIS INDUCERS

The database used in this example comes from QSARData package (Quantitative Structure-Activity Relationship Data Sets for R. http://qsardata.r-forge.r-project.org/QSARdata/QSARdata_Package.html), a particular R repository maintained by Max Kuhn. The set is version 2013-07-16 and collects a selection of 324 molecules coded by Goracci and coworkers using fingerprint strings (BIOVIA Pipeline Pilot Overview. http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot) of 2862 binary characters.

In this library 124 (38.3%) molecules are PLD inducers and are defined herein as molecules of interest. Specifically, the binary classification of a molecule to be a PLD inducer (1) or a non-inducer (0) is encoded in the following string of 324 characters long:

1100011000000000000000100111101101000001000011111011100101100000000000010111
1000111101110001100110000000110000100001110000000100100100100101100000000101
0000001110011001010100110100001010001000101111010100000000110000001000011000
0101100000000111110100101001000110111010100100010101100111101000000100100001
0010010011111001100000001001

3.1. Rules of Order 1

As an example, the fourth descriptor in the virtual database is the following 324-bit string (one per molecule):

0000010011111111111111100111101011110111011001111100111001111011101100000
0101001011000100101001111110001010111111111111011101011011011001011101010
1101110001010010101011011010011111001111111001011010111110011011011110111
0010011011110000110110111101000010001001110110010010111111101111100111110
1111100000011110010111110111

For instance, if a rule is defined as R = "fourth descriptor set at level 0", it can be said that 122 molecules meet this condition, that is, they conform to the rule. Of these, 79 are of interest. According to (2), the probability to pick up 79 or more items of interest by selecting 122 at random from the whole database is $p(79+,122;124,324) = 2.5\cdot 10^{-14}$. This is a very significant rule. In fact, due to the complementarity of the two descriptor levels, one reaches the same conclusion if the rule is defined as the negation of the previous one. This negative rule takes the form S = "fourth descriptor fixed at level 1". In this case, 202 molecules meet the condition and 45 are of interest. Now, according to (2), the probability to pick up 45 or more items of interest by selecting 202 at random is almost the unity. This points to define as significant rule the negation of this negative rule, i.e., the previous rule. When building the consensus pool, it is equivalent to store the R rule with a positive vote than to store the S rule but tagged with a negative vote.

A total of 2862 rules of order 1 can be explored, one per each descriptor. As seen, in the case of binary variables, it is equivalent to consider either level. A systematic calculation gave that, if the p-value threshold is set to $p_c = 10^{-5}$, a total of 57 rules are declared significant. It should be noted that this calculation only takes a few seconds.

Once each molecule adds up all the votes collected from the rules, a distribution of number of votes is obtained. In our example, the list of votes ranges from three molecules collecting 27 votes up to a molecule collecting 30 negative votes. The sorting established by this voting system gives an area under the receiver operating characteristic curve (AU-ROC) (Egan, 1975; Mason & Graham, 2002; Forlay-Frick et al., 2005; Besalú et al., 2010) of 0.831. In this training calculation, in the best classifying spot along the ranking obtained (that is, defining the cut of the binary classifier in the position where the probability (2) gives a minimum value), it is found that Accuracy = 73.5%, Sensitivity = 83.9%, Specificity = 67.0%, Precision = 61.2%, Matthews CC = 49.5%, Hit rate @5% = 95.5% (21 of the first 22 ranked molecules are of interest) with an enrichment factor of 2.49 (the maximum reachable here is 2.61). Along the first set of 100 ranked molecules, 72 of them are PLD inducers and this corresponds to an estimated probability of $p(72+,100;124,324) = 9.5\cdot 10^{-17}$. As explained in the literature, the method SSIR can easily implement the leave-one-out (L1O) cross-validation process. In this case it is obtained an AU-ROC = 0.822 (Accuracy = 74.4%, Sensitivity = 80.6%, Specificity = 70.5%, Precision = 62.9%, Matthews CC = 49.7%, Hit rate @5% = 95.5%).

A test revealed that, for training, an almost optimal p-value cutoff is 0.1. In this case, 1137 rules of order 1 became significant. This new training provides AU-ROC = 0.913, a value comparable to the ones reported by Goracci and coworkers. The new classification parameters are Accuracy = 86.4%, Sensitivity = 80.6%, Specificity = 90.0%, Precision = 83.3%, Matthews CC = 71.1%, Hit rate @5% = 93.8% with an enrichment factor of 2.45, Hit rate @10% = 96.9% with an enrichment factor of 2.53. All the first 68 ranked items are PLD inducers but the first one. For the first set of 100 ranked molecules, 86 are PLD inducers and this corresponds to a probability of $p(86+,100;124,324) = 3.3\cdot 10^{-33}$.

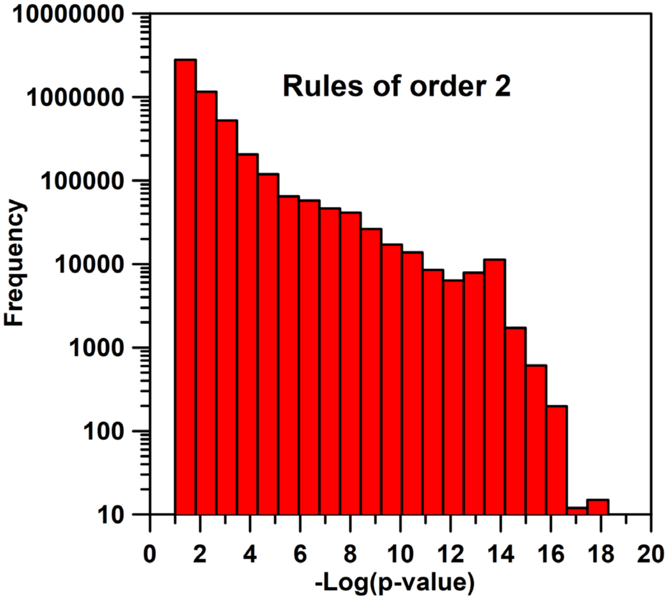
3.2. Rules or Order 2

A rule or order 2 is obtained when combining two different descriptors and, at the same time, specifying a level for each one. So, for this molecular set a total of $C(2862,2)\cdot 2^2 = 16376364$ rules of order 2 can be defined. Of these, 313419 have a p-value less than 10^{-5} (this exhaustive calculation took two hours in a desktop computer with an Intel i7-8700 processor running under Windows 10 with 16GB of RAM). Figure 1 shows the distribution of p-values when starting at $p = 0.1$ (note the logarithmic scale in both axes). Full training fit gives AU-ROC = 0.840. A similar result can be obtained if the calculation is not exhaustive, i.e., only a subset of the possible rules is randomly selected and processed.

3.3. Cross-Validation

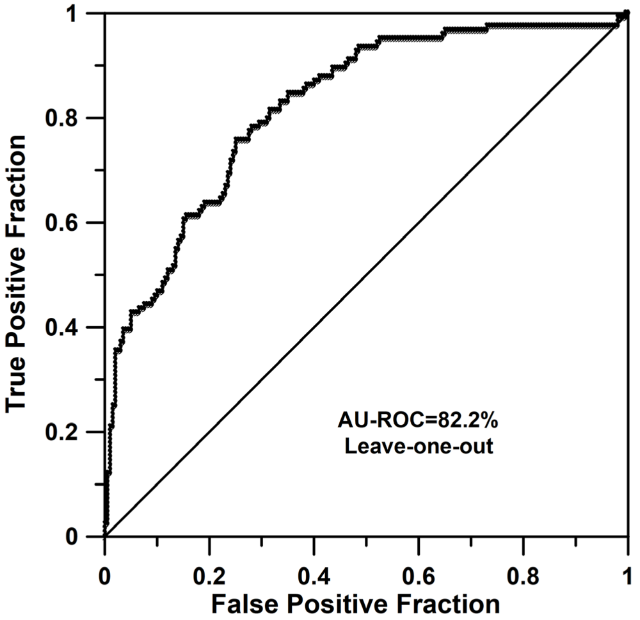
As it is explained in the literature, complete L1O procedures can be easily implemented in SSIR, and the whole procedure does not need to increase the computation time by a factor proportional to

Figure 1. Distribution of the *p*-values found for the rules of order 2 attached to the set of PLD inducers



the number of molecules in the database. In this example, when generating the rules of order 1, the L1O procedure gave in a few seconds an AU-ROC of 0.906. When considering the rules of order 2, the L1O procedure took 30h, including the previous 2h of total training time, so the extra time for each molecule left out is of about 5 minutes. For this calculation, AU-ROC = 0.822 was found (see Figure 2). The result of the cross-validation shows Accuracy = 75.3%, Sensitivity = 75.8%,

Figure 2. ROC curve corresponding to the L1O models that rank the set of PLD inducers. SSIR rules are of order 2.



Specificity = 75.0%, Precision = 65.3%, Matthews CC = 49.7%, Hit rate @5% = 93.8% with and enrichment factor of 2.45.

The previous calculations reveal that the model is quite robust. This aimed to redo a calculation, but this time selecting some compounds to act as test. Several series of 5-fold calculations were performed. Throughout the five loops, each had about 259 items for training and 65 for predicting, all chosen at random. Of these, 99 and 25 were PLD inducers, respectively. Surprisingly, the calculations involving only rules of order 1 gave quite acceptable results considering that each calculation took only a few seconds. The calculation using rules of order 2 is more time consuming (each run took about 8.5 hours) but, as expected, the results improved. Table 1 shows the statistical parameters found for the case of $p_c = 0.5$. In all cases, the hit rate fraction found along the first 5% of classified molecules is 93.8% (15 actives out of 16), the enrichment factor being 2.45 (for the first 5% of classified compounds with a maximum attainable value of 2.61). The corresponding parameters and the average values are comparable to those reported by Goracci and coworkers found for several databases and based on a second latent variable of a PLS calculation. We can keep in mind that those procedures are more sophisticated than SSIR. The most noticeable difference is found for Sensitivity (Goracci's mean value of 0.74) and Matthews correlation coefficient (Goracci's mean value of 0.64): the results shown here are poorer (mean values of 0.58 and 0.59, respectively). On the other side, the AU-ROC, Accuracy and Hit rate values are similar: Goracci's 0.82 vs. here 0.85, 0.77 vs. 0.81, and 94% vs. 94%, respectively. Here, improved values for Specificity and Precision are reported: 0.77 vs. 0.95, and 0.74 vs. 0.88. The best parameter reproduced is the Specificity associated with a low false positive rate. It must to be said that all the compared values obtained by SSIR are not attached to simple fittings but to predictions (at the level of internal cross-validation calculations).

Table 1. Classification parameters of PLD inducers for predictions obtained after ten repetitions of complete 5-fold calculations (rules of order 2, p_c = 0.5). See text for details.

Run	AU-ROC	Accuracy[a]	Sensitivity	Specificity	Precision	MCC[b]
1	0.855	0.821	0.621	0.945	0.875	0.618
2	0.852	0.796	0.532	0.96	0.892	0.570
3	0.854	0.821	0.629	0.94	0.867	0.618
4	0.865	0.83	0.758	0.875	0.790	0.638
5	0.838	0.79	0.556	0.935	0.841	0.549
6	0.856	0.796	0.500	0.980	0.939	0.579
7	0.838	0.802	0.597	0.930	0.841	0.576
8	0.858	0.799	0.516	0.975	0.928	0.583
9	0.864	0.802	0.500	0.990	0.969	0.598
10	0.839	0.796	0.548	0.950	0.872	0.567
Mean[c]	0.852	0.806	0.576	0.948	0.881	0.589
SD[d]	0.013	0.018	0.103	0.041	0.065	0.034

[a] In the literature sometimes is referred as Concordance. [b] Matthews correlation coefficient. [c,d] Corresponding to a Gaussian fit of the data.

Krstajic et al. (2014) focused on the performance of several cross-validation algorithms. One of the molecular assemblies studied is the same set of 324 compounds of Goracci. The descriptors used were a selection of the PipelinePilotFP set, but they curated the database and selected 308. In this article, the authors classified the potential PLD inducers with ridge logistic regression. Along 50 repeats of 10-fold cross-validation processes they report that the average of erroneous classifications is of 17.68%. This parameter can be compared here with the value of 100-Accuracy which, in our case, has a mean value of 19.4% for the series of 5-fold cross-validation experiments (see Table 1).

Other *in silico* models for the prediction of PLD inducers have been developed in the past considering various molecular databases, descriptors and methodologies: Ploemen et al. (2004) described a simple inequality model based on two physicochemical properties (ClogP and a calculated pK_a) to be applied to 41 compounds. Only two compounds did not follow the classification rule. Tomizawa et al. (2006) modified this and increased prediction performance. Tomizawa considered 63 compounds (33 for training/test and 30 for validation). Then, Pelletier et al. (2007) obtained an improved set of benchmark compounds (125 compounds, 84 of which being PLD inducers) that served to validate the Ploemen model (Sensitivity 58%, Specificity 87%, Precision 77%, Accuracy 75%), the modified Ploemen one (Sensitivity 79%, Specificity 80%, Precision 74%, Accuracy 80%) and build an improved Bayesian model (Sensitivity 92%, Specificity 77%, Precision 74%, Accuracy 83%). Hanumegowda et al. (2020) included other parameters in the model, further improving prediction capabilities (Sensitivity 82%, Specificity 94%, Accuracy 88%). Their particular molecular set included 53 inducers and 50 non-inducers. Fisher et al. (2012) showed that the methods based on the amphiphilic moment (replacing the lipophilicity descriptor, according to the currently accepted mechanism of action) perform even better. This author employed an expanded molecular set of 422 compounds and obtained Accuracy 86%, Sensitivity 80% and Specificity 90%. The accuracy was similar to the Ploemen model. Fischer's team also performed a validation with an FDA PLD data set of 91 compounds (56 of which were inducers) obtaining Accuracy 85%, Sensitivity 84% and Specificity and Precision of 86% and 90%, respectively.

Kruhlak et al. (2008) considered a database of 583 compounds (190 inducers) and reported several results obtained with different methods, including internal cross-validation and external validation. Lowe et al. (2010) reported MCC values of, at most, approximately 0.7. Lowe is one of the authors who considered a large dataset (Lowe et al., 2012). Sun et al. (2012) focused attention to the use of Support Vector Machines and obtained AU-ROC values of about 88-90%. Orogo et al. (2012) considered a library containing 743 compounds (385 positive and 358 negative) and obtained Sensitivity 61% but a Specificity of 58.2%. These results were improved by a model created with Leadscope Predictive Data Miner. This model gave good cross-validation statistics: Sensitivity 79.0% and Specificity 78.0%.

Other complex models have also been reported, but an improvement in classification performance is still needed for practical applications. It is worth of mention the work of Przybylak et al. (2014). It constitutes a very interesting approach to the problem to model PLD. The authors considered depurated databases (consisting of 736, 331 -a variant of the Goracci's database-, and 185 structures) and obtained an updated model whose figures of merit are Sensitivity 60.7%, 83.5%, and 90.2% respectively for each one of the libraries; Specificity of 80.5%, 78.9%, and 84.3%; and Accuracies of 74.7%, 80.4%, and 87.6%. The authors used descriptors developed from the SMILES description of the molecules (SMARTS patterns). Structural alerts were obtained from this codification that seemed to be related to the PLD activity.

CONCLUSION

It has been shown how SSIR, a systematic and combinatorial procedure, is useful to rank series of molecules described by binary fingerprints. The consensus models that were generated provided solid statistical parameters attached to the established molecular rankings. Through cross-validation calculations (leave-one-out and 5-fold ones) it has also been demonstrated how the procedure, after a proper training, is stable and robust. The results have been shown to be comparable to those obtained in the literature that employs the same molecular set and the same descriptors but more sophisticated methods. The SSIR method can be taken as a useful and quick tool to rank compounds either at the full study level or as a complementary QSAR filter tool.

REFERENCES

Besalú, E. (2016). Fast Modeling of Binding Affinities by means of Superposing Significant Interaction Rules (SSIR) method. *International Journal of Molecular Sciences*, *17*(6), 827. doi:10.3390/ijms17060827 PMID:27240346

Besalú, E., De Julián Ortiz, J. V., & Pogliani, L. (2010). *Quantum Frontiers of Atoms and Molecules* (M. V. Putz, Ed.). New York: NOVA Publishing Inc.

Besalú, E., Pogliani, L., & De Julián-Ortiz, J. V. (2016). Superposing Significant Interaction Rules (SSIR) method: A simple procedure for rapid ranking of congeneric compounds. *Croatica Chemica Acta*, *89*(4), 481–492. doi:10.5562/cca3027

Besalú, E., Pogliani, L., & De Julián-Ortiz, J. V. (2017) The Superposing Significant Interaction Rules (SSIR) method. In A.K. Haghi, L. Pogliani, E.A. Castro et al. (Eds.). Applied Chemistry and Chemical Engineering (Vol. 4). Apple Academic Press.

Besalú, E., Pogliani, L., & De Julián-Ortiz, J. V. (2018). Fast Qualitative Inspection of Designed Experiments by means of the Superposing Significant Interaction Rules (SSIR) method. In A.V. Vakhrushev, R. Haghi, J.V. De Julián-Ortiz et al. (Eds.), Physical Chemistry for Chemists and Chemical Engineers. Multidisciplinary Research Perspectives (Vol. 9). Apple Academic Press.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. New York: Academic Press.

Fischer, H., Atzpodien, E. A., Csato, M., Doessegger, L., Lenz, B., Schmitt, G., & Singer, T. (2012). In silico assay for assessing phospholipidosis potential of small druglike molecules: Training, validation, and refinement using several data sets. *Journal of Medicinal Chemistry*, *55*(1), 126–139. doi:10.1021/jm201082a PMID:22122484

Forlay-Frick, P., Van Gyseghem, E., Héberger, K., & Vander Heyden, Y. (2005). Selection of orthogonal chromatographic systems based on parametric and non-parametric statistical tests. *Analytica Chimica Acta*, *539*(1-2), 1–10. doi:10.1016/j.aca.2005.02.058

Goracci, L., Ceccarelli, M., Bonelli, D., & Cruciani, G. (2013). Modeling Phospholipidosis Induction: Reliability and Warnings. *Journal of Chemical Information and Modeling*, *53*(6), 1436–1446. doi:10.1021/ci400113t PMID:23692521

Hanumegowda, U. M., Wenke, G., Regueiro-Ren, A., Yordanova, R., Corradi, J. P., & Adams, S. P. (2010). Phospholidosis as a function of basicity, liphophilicity and volume of distribution of compound. *Chemical Research in Toxicology*, *23*(4), 749–755. doi:10.1021/tx9003825 PMID:20356072

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, *6*(1), 10. doi:10.1186/1758-2946-6-10 PMID:24678909

Kruhlak, N. L., Choi, S. S., Contrera, J. F., Weaver, J. L., Willard, J. M., Hastings, K. L., & Sancilio, L. F. (2008). Development of a phospholipidosis database and predictive quantitative structure-activity relationship (QSAR) models. *Toxicology Mechanisms and Methods*, *18*(2-3), 217–227. doi:10.1080/15376510701857262 PMID:20020916

Lowe, R., Glen, R. C., & Mitchell, J. B. (2010). Predicting phospholipidosis using machine learning. *Molecular Pharmaceutics*, *7*(5), 1708–1714. doi:10.1021/mp100103e PMID:20799726

Lowe, R., Mussa, H. Y., Nigsch, F., Glen, R. C., & Mitchell, J. B. (2012). Predicting the mechanism of Phospholipidosis. *Journal of Cheminformatics*, *26*, 1186–1758. PMID:22281160

Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, *128*(584), 2145–2166. doi:10.1256/003590002320603584

Mendenhall, W., & Sincich, T. (1995). *Statistics for Engineering and the Sciences*. Englewood Cliffs, NJ: Prentice-Hall.

Orogo, A. M., Choi, S. S., Minnier, B. L., & Kruhlak, N. L. (2012). Construction and consensus performance of (Q)SAR models for predicting phospholipidosis using a dataset of 743 compounds. *Molecular Informatics*, *31*(10), 725–739. doi:10.1002/minf.201200048 PMID:27476455

Pelletier, D. J., Gehlhaar, D., Tilloy-Ellul, A., Johnson, T. O., & Greene, N. (2007). Evaluation of a published in silico model and construction of a novel Bayesian model for predicting phospholipidosis inducing potential. *Journal of Chemical Information and Modeling*, *47*(3), 1196–1205. doi:10.1021/ci6004542 PMID:17428028

Ploemen, J. P., Kelder, J., Hafmans, T., van de Sandt, H., van Burgsteden, J. A., Saleminki, P. J., & van Esch, E. (2004). Use of physicochemical calculation of pKa and CLogP to predict phospholipidosis-inducing potential: A case study with structurally related piperazines. *Experimental and Toxicologic Pathology*, *55*, 347–355. PMID:15088636

Przybylak, K. R., Alzahrani, A. R., & Cronin, M. T. D. (2014). How Does the Quality of Phospholipidosis Data Influence the Predictivity of Structural Alerts? *Journal of Chemical Information and Modeling*, *54*(8), 2224–2232. doi:10.1021/ci500233k PMID:25062434

Sun, H., Shahane, S., Xia, M., Austin, C. P., & Huang, R. (2012). Structure based model for the prediction of phospholipidosis induction potential of small molecules. *Journal of Chemical Information and Modeling*, *52*(7), 1798–1805. doi:10.1021/ci3001875 PMID:22725677

Tomizawa, K., Sugano, K., Yamada, H., & Horii, I. (2006). Physicochemical and cell-based approach for early screening of phospholipidosis inducing potential. *The Journal of Toxicological Sciences*, *31*(4), 315–324. doi:10.2131/jts.31.315 PMID:17077586

Emili Besalú is a Lecturer in Physical Chemistry at the University of Girona since 2001. He has contributed more than 120 international papers and book chapters on Theoretical Chemistry, mainly devoted to methodologies in SAR and QSAR fields. He is referee for various journals. His preliminary interests are related to molecular quantum similarity, perturbation methods, and linear numerical methods. Actual ones are focussed on the treatment and ranking of congeneric molecular database families, and especially the interplay between statistically-based and computational procedures.