ACADEMIA ROMÂNĂ
Revue Roumaine de Chimie
http://web.icf.ro/rrch/

Rev. Roum. Chim.,
2015, *60*(2-3), 219-226

Dedicated to Professor Zeno Simon
on the occasion of his 80th anniversary

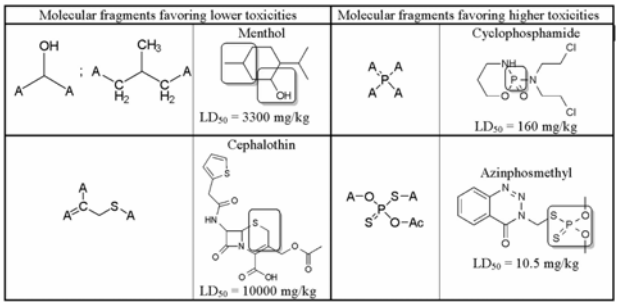
CHEMICAL AND BIOLOGICAL DESCRIPTOR INTEGRATION IMPROVES COMPUTATIONAL MODELING OF *IN VIVO* RAT TOXICITY

Cristian G. BOLOGA,[a] Oleg URSU,[a] Liliana HALIP,[b] Ramona CURPĂN[b] and Tudor I. OPREA[a,*]

[a] Translational Informatics Division, Department of Internal Medicine, MSC09 5025,
University of New Mexico School of Medicine, Albuquerque, NM 87131, USA
[b] Roumanian Academy – Institute of Chemistry, 24 Mihai Viteazul Avenue, RO-300223, Timişoara, Roumania

Received September 3, 2014

Computational toxicology is a new discipline in the area of computational molecular sciences, which is rapidly developing as a result of the public interest stirred by several European and US initiatives. Here, we report the use of primary high throughput screening (HTS) data as biological descriptors to complement the chemical descriptors for the modelling of the acute toxicity. The combination of biological and chemical descriptors was performed on the median lethal dose following oral administration in rats (rat LD50). The hybrid model developed based on chemical and biological descriptors is superior to models based on the chemical or biological description alone. Using this model, besides the accurately prediction of a compound's toxicity we also identified molecular fragments whose presence may contribute to increase or decrease of the toxicity.



INTRODUCTION

Chemical-induced toxicity is a major concern for healthcare professionals, cosmetic industry, flavour and fragrance, as well as lawmakers and chemical safety regulators. It is of particular concern in pharmaceutical drug discovery and development, and its evaluation is mandatory for the approval of new drugs for human use. The impact of toxicity and safety related events on the development of new chemicals is substantial, whether it relates to medicines[1], environmental chemicals or other chemicals. The United States passed the Toxic Substances Control Act (TSCA) into law[2] in 1976, whereas the European Union adopted the Restriction of Hazardous Substances Directive[3] in 2003, which became law in all member states in 2006. In addition to costs and societal impact, however, toxicity and safety limit the benefit of using chemicals, in particular therapeutics, by significantly lowering the cost/benefit ratio for certain sub-populations that tolerate exposure to a given chemical (or therapy), and by limiting the amount (or dose) such that the most useful amount/dose and thereby maximal effect are not reached. Lowering toxicity impact

* Corresponding author: TOprea@salud.unm.edu

and thus maximizing the cost/benefit ratio are an essential goal in chemical research.

Computational toxicology[4] is a growing field in the area of computational molecular sciences that is poised to gain significance and impact due to several European and US initiatives. These include for example the REACh (Registration, Evaluation, Authorization and Restriction of Chemicals) regulation[5] implemented in the EU in 2007, part of this initiative being to create community and expert driven computational models of toxicity in the context of OpenTox online community.[6] Tox21 program[7] in the US has similar goals and aims at identification of better toxicity assessment methodology both experimental and computational.

One of the key objectives of *in silico* toxicology assessment is the prioritization of chemicals for toxicity evaluation, thus reducing the experimental burden and the need to evaluate compounds in animal models. This is often accomplished by highlighting chemical substructures, or structural alerts[8], which are associated with harmful effects. Several categories of chemicals are flagged in this manner by means of expert systems such as DEREK Nexus[9] and machine learning or QSAR, Quantitative Structure-Activity Relationships.[10,11] Many computational toxicology tools are now freely available on the internet.[12]

Here, we explore the possibility of adding primary high throughput screening (HTS) endpoints as biological descriptors, to complement the molecular descriptors derived from chemical structures. The combination of biological and chemical descriptors was performed on the median lethal dose following oral administration in rats (henceforth termed rat LD50). Since LD50 measures the lethal effect of the exposed chemical on a given population, in this case rats, this particular endpoint is used to evaluate acute toxicity. LD50 values are usually expressed in mg/kg; lower values indicate high toxicity, while higher values are observed for less harmful substances.[13] Our hypothesis is that, by combining biological and chemical descriptors, one could develop enhanced, more predictive, QSAR models. Despite the limitations of rat oral LD50 as an endpoint, one that has since been abandoned due to its mechanisticy complexity, we hereby illustrate the power of the joint descriptor system using NIH Roadmap endpoints, combined with structural alerts.

MATERIALS AND METHODS

Compound selection: The Hazardous Substances Data Bank (HSDB)[14] was leased from the National Library of Medicine in XML format and converted to tabular format. CAS identifiers from HSDB records were used to lookup chemical structures from PubChem and NCI Chemical Structure Lookup Service using the web services public API. As in vivo toxicity data, only LD50 rat oral data were used from the HSDB dataset, where manual curation was done to convert all dose values to mg/kg units. All toxicity values were converted to logarithm of dose values for QSAR models. A set of 428 unique compounds having one or more measured experimental rat LD50 values was initially selected. The following procedure was used: the fixed values of oral LD50, were included twice in the training set; for experimental LD50 values reported as an interval, both the minimum and maximum values were included; and for cases where LD50 values were reported as larger than a specific value the threshold value was included, respectively.

Biological descriptor assembly and curation: All molecules from the oral rat LD50 set were subject to automated queries in PubChem Assays. The query was limited to primary screens from the Molecular Libraries Initiative (MLI),[15] where each screen had to originate from MLI, and where at least 20,000 chemicals were evaluated (HTS screens only). All molecules were mapped to the Molecular Libraries-Small Molecules Repository (MLSMR) library using BABEL2 (Openeye)[16,17] generated canonical SMILES. The HTS screening results (active/inactive) for those substances were retrieved for 822 MLI assays. Physico-chemical profiling assays (solubility, aggregation, fluorescence, etc) were excluded.

Chemical descriptor assembly: Circular ECFP fragment fingerprints[18] of radius 2-10 were generated for the selected substances and used as chemical descriptors, using an in-house implementation of ECFP. JChem library[19] was used for parsing chemical structures input.

Statistics: All multivariate modelling of the LD50 were performed by PLS[20] (projection to latent structures) using the Simca package.[21] The estimation of principal component significance was performed by the cross-validation (CV) procedure[22] and reported as Q^2. Overfit was investigated by perturbation of the response values and the related loss of Q^2. Model predictivity was

estimated using an independent test set: Starting from 265 LD50 values (for 223 approved drugs) extracted from the Wombat PK database,[23] all duplicates (present in the training set) were removed. A total of 102 unique drugs associated with 122 oral rat LD50 values were retained as a completely independent external set, i.e., not used to evaluate the QSAR model. This test set was submitted to the same biologic and chemical descriptor evaluation, and used to evaluate model predictivity.

RESULTS

The initial set of 428 unique molecules resulted in 1155 objects, due to processing of LD50 values, as outlined earlier, and were used as training set for multivariate modeling. The initial PLS model, based on a combination of biological and chemical descriptors, had 6 latent variables (see Table 1).

This PLS models was analyzed in terms of the variable importance contribution (VIP) for each of the starting descriptors. This comparison helps to eliminate those descriptors that did not relate to oral rat LD50, and to focus on those descriptors that consistently contribute to the PLS model. Only descriptors having a VIP value above 0.8 are discussed in this paper, and included in the "best" models. The sequential descriptor elimination process via the VIP criterion reverted to previous models if the new dataset resulted in a loss of too many objects (over 25%) due to zero columns, or if the model resulted in a significant Q^2 drop following cross-validation. Descriptors that had VIP above 1.0 (this being deemed as highly important) are highlighted below. For comparative purposes, models based on chemical or biological descriptors only, are also reported in Table 1. The loss in objects and descriptors is either due to VIP selection, or to the complete loss of descriptors per row.

The comparative study that is summarized in Table 1 suggests that the "hybrid" model, that is the model based on a combination of chemical and biological descriptors, is superior to models based on chemical or biological description alone. When used separately, biological descriptors result in a PLS model with lower statistical significance, compared to those based on chemical description. However, the "best" model appears to be the hybrid one, based on CV-derived Q^2. The significance of the Hybrid Model (score plot in Fig. 1, top) was investigated by redoing the regression with scrambled Y data. The procedure shows a complete loss of explained variance, as measured by Q^2, with the degree of correlation between the true response vector and the perturbed vector (Fig. 1, bottom). Thus, we can safely state that the PLS model used to model rat LD50 after oral exposure is highly significant.

We further investigated the relationship between the test set and the training set, as measured by the overlap in the score plot between objects present in the training set (Fig. 2, light gray) and test set (Fig. 2, black). This plot suggests that chemicals predicted in our model are within the applicability domain, and therefore estimates of model predictivity are reliable.

DISCUSSIONS

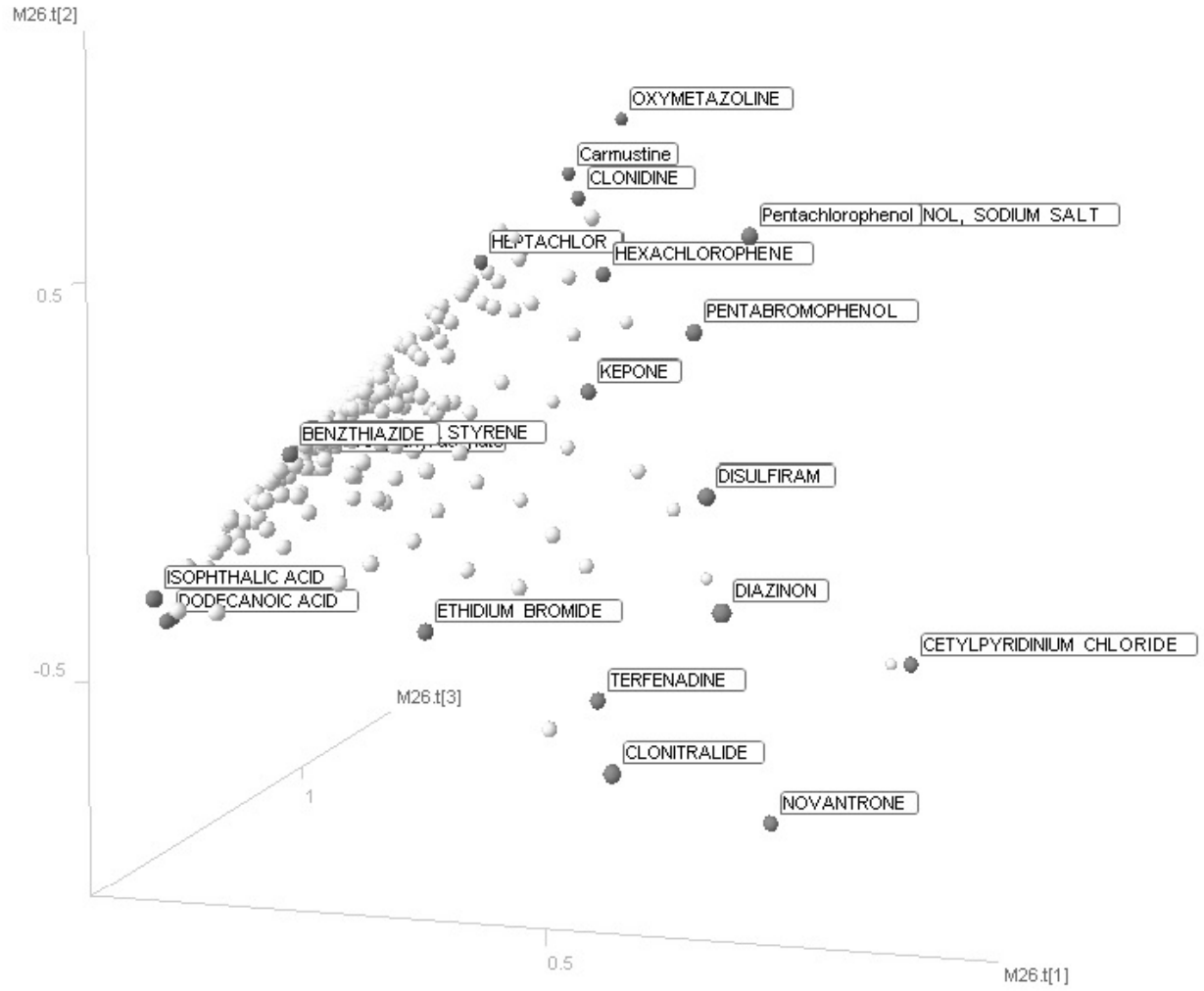
Although toxicity endpoints are successfully used to build QSAR models that pass statistical criteria, these models are likely to require temporal updates, as illustrated by developing a global model for cardiac toxicity mediated by the hERG potassium channel.[24] This illustrates not only the need to periodically re-evaluate a model's applicability domain,[25] but further highlights the inherent limitations of model predictivity for novel chemical structures. One possible approach in addressing this limitation, inherent to all QSAR models, is to introduce descriptors that are less (or not at all) dependent on chemical structures. Here, we explored the use of primary HTS data as biological descriptors, in combination with chemical descriptors, to model LD50 data.

Table 1

Comparative summary and overview of the statistical models

No.	Type	R^2Y	Q^2Y	A	Observations & Variables
1	Initial Model	0.684	0.483	6	1155 objects; 499 bio and 548 chem descriptors
2	Hybrid Model	0.668	0.565	7	899 objects; 205 bio and 181 chem descriptors
3	Chemistry-based Model	0.49	0.406	7	1155 objects; 181 chem descriptors
4	Bioactivity-based Model	0.376	0.267	5	603 objects; 106 bio descriptors

[*] "Objects" refers to the entries in the multivariate model, which are not the same as unique molecules (see Materials and Methods). "A" refers to the number of latent variables.



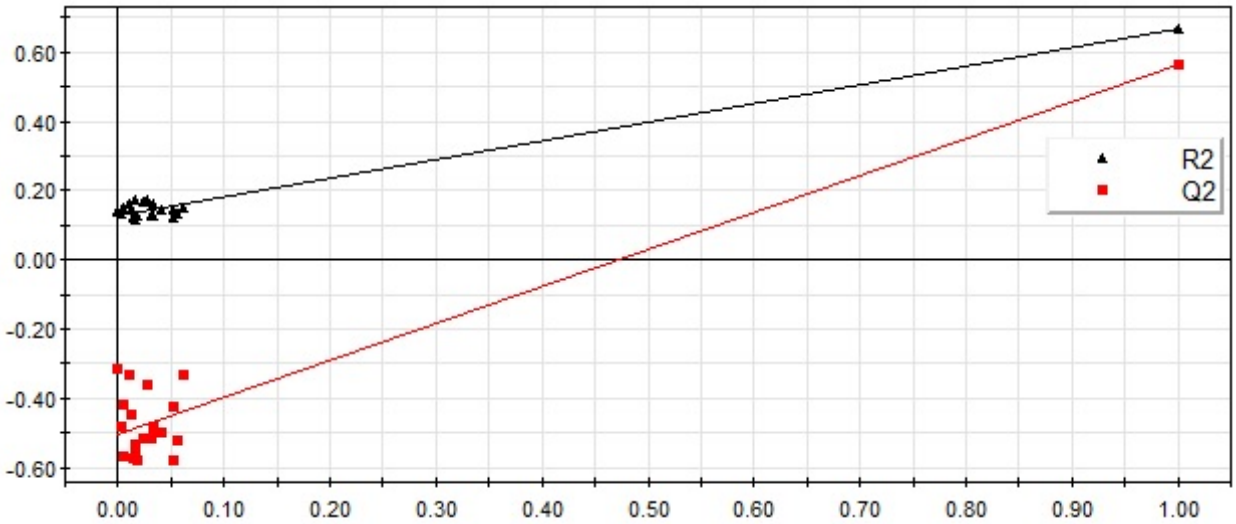


Fig. 1 – Three-dimensional scatter plot of the Hybrid PLS Model (top). Some of the compounds from the training set are highlighted for comparative purposes. PLS model validation using Y – scrambling (bottom).

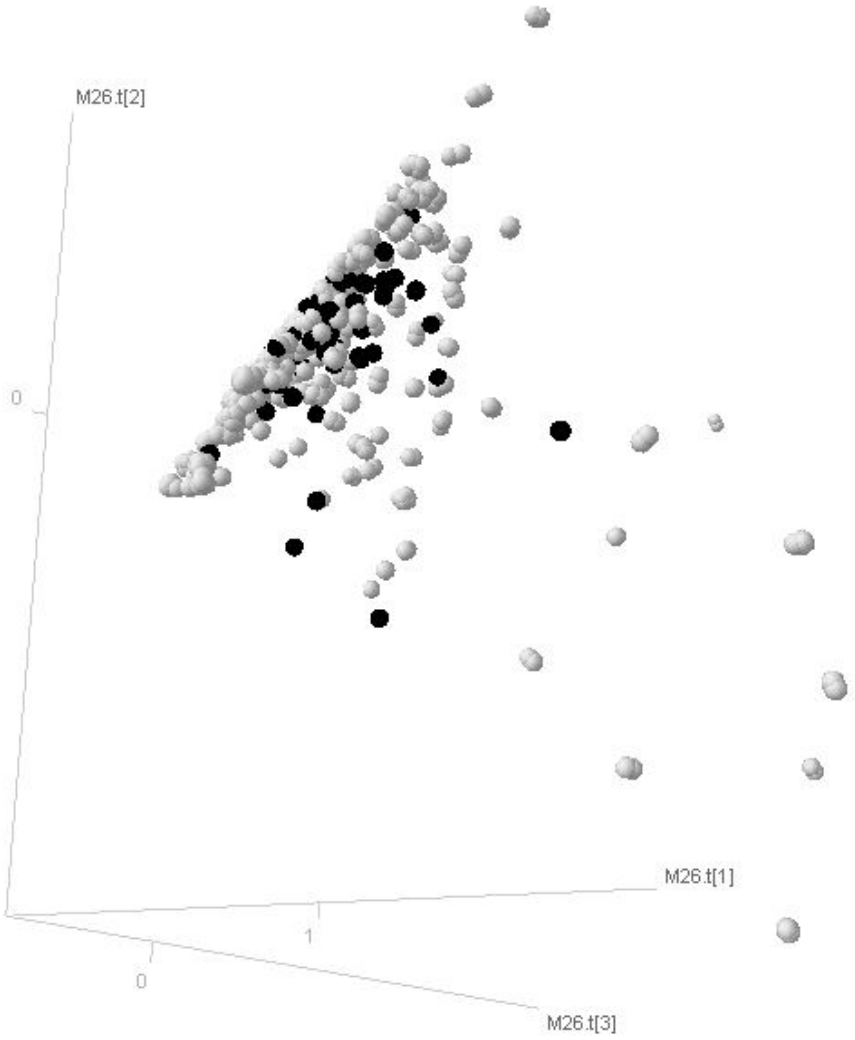


Fig. 2 – PLS model applicability domain: Compounds included in the test set (black) are well within the domain representation for the training set (light grey).

A previous report used cellular toxicity endpoints from the Molecular Libraries Initiative as HTS descriptors which, in combination with molecular descriptors, were used to model rat LD50 data by Tropsha and collaborators.[26] Although the model published by Tropsha et al. uses a similar approach, that of combining chemical descriptors and qHTS in manner similar to the work described here, there are several differences between the two approaches: i) In this study, all HTS assays (>20000) published in PubChem run on MLSMR library were used; ii) All processed data from HSDB were included in the training set, not just min/max values; iii) Our chemical descriptors are molecular patterns and can be used as structural alerts.

The main advantages brought by these differences are discussed below.

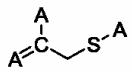
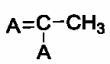
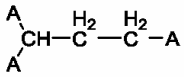
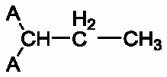
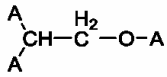
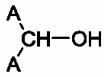
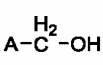
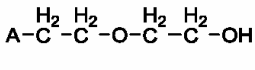
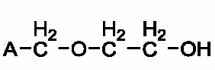
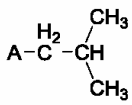
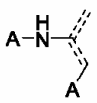
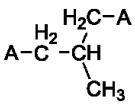
i) It is possible to formulate hypotheses concerning the compound's mechanism of toxicity.

Using all available (fitting our selection criteria) HTS data from PubChem provided additional insights into mechanism of toxicity (Table 2). Compounds active in multiple assays with unrelated targets indicate a pattern of non-specific or promiscuous binders most likely through covalent binding to proteins like in the case of organomercury compounds which are likely to bind to cysteine residues. On the other hand activities in one assay or assay with related or identical targets indicates a more specific mechanism of toxicity for example oxymetazoline is active only on HTR2A receptor, and literature data indicate that it is also active on other GPCR class A/amine subclass receptors suggesting that toxicity is related to GPCR activities.

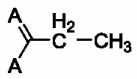
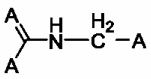
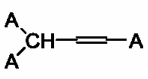
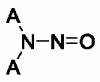
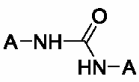
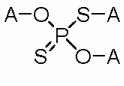
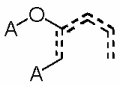
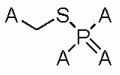
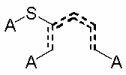
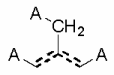
Table 2

HTS descriptors and mechanism of toxicity

Chemicals	PubChem assay target(s)	Mechanism of toxicity explained by HTS descriptors
Oxymetazoline	5-HT2A	GPCR class A/amine
Organophosphoric pesticides/insecticides	>250	covalent binders, important target acetylcholinesterase
Cycloheximide	~65	non specific protein synthesis inhibitor
Thimerosal (organomercury)	~50	non specific; protein covalent binder
Pyriminil, Verapamil	mitochondrial permeability transition pore (regulates ion passage like ion channels)	Interference with mitochondrial function

			
[C;H0D3][C;H2D2][S;H0D2]	[C;H0D3][C;H3D1]	[C;H0D3][O;H1D1]	[C;H0D3][c;H0D3](:[c;H1D2]):[c;H1D2]
			
[C;H1D3][C;H2D2][C;H2D2]	[C;H1D3][C;H2D2][C;H3D1]	[C;H1D3][C;H2D2][O;H0D2]	[C;H1D3][O;H1D1]
			
[O;H0D2][c;H0D3](:[c;H1D2]):[c;H1D2]	[C;H2D2][O;H1D1]	[C;H2D2][C;H2D2][O;H0D2][C;H2D2][C;H2D2][O;H1D1]	[C;H2D2][O;H0D2][C;H2D2][C;H2D2][O;H1D1]
			
[C;H2D2][C;H1D3]([C;H3D1])[C;H3D1]	[N;H1D2][c;H0D3](:[c;H0D3]):[c;H1D2]	[O;H0D1]=[S;H0D4]	[C;H2D2][C;H1D3]([C;H2D2])[C;H3D1]
			
	[S;H0D4]	[c;H0D3][C;H0D3]([O;H1D1])=[O;H0D1]	

Scheme 1 – Molecular fragments associated with lower acute toxicity, dashed bonds indicate presence of aromatic ring. Molecular fragments are depicted along with SMARTS patterns.

			
[C;H0D3][C;H2D2][C;H1D3]	[C;H0D3][Cl;H0D1]	[C;H0D3][N;H1D2][C;H2D2]	[C;H1D3][C;H1D2]=[C;H1D2]
A=S		A–N=A	
[S;H0D1]	[c;H0D3][C;H2D2][C;H0D3]	[N;H0D2]	[N;H0D2]=[O;H0D1]
			
[N;H0D3]	[N;H0D3][N;H0D2]=[O;H0D1]	[N;H0D3][c;H0D3](:[c;H1D2]):[c;H1D2]	[N;H1D2][C;H0D3]([N;H1D2])=[O;H0D1]
			
[O;H0D1]=[P;H0D4]	[O;H0D2][P;H0D4]([O;H0D2])([S;H0D2])=[S;H0D1]	[O;H0D2][c;H0D3](:[c;H0D3]):[c;H1D2]:[c;H1D2]:[c;H1D2]	[P;H0D4]
			
[P;H0D4]=[S;H0D1]	[C;H2D2][S;H0D2][P;H0D4]	[S;H0D2][c;H0D3](:[c;H0D3]):[c;H1D2]:[c;H1D2]:[c;H0D3]	[C;H2D2][c;H0D3](:[c;H0D3]):[c;H0D3]
			
	[c;H1D2]:[c;H1D2]:[n;H0D2]	[c;H1D2]:[n;H0D2]:[c;H1D2]	

Scheme 2 – Molecular fragments associated with higher acute toxicity, dashed bonds indicate presence of aromatic ring. Molecular fragments are depicted along with SMARTS patterns.

ii) The model proposed here is more general and has a higher applicability domain, and is less biased towards structures with "marginal" toxicity. For example, there are 48 chemical structures with marginal toxicity in our validation set that would fall outside the applicability domain of previously published QSARs.

iii) Structural alerts list can be used directly by other investigators, even if PubChem data were not available. By identifying the structural patterns from chemical descriptors (ECFP fingerprints) which have the largest contribution to decreasing/increasing of the acute toxicity value it is possible to classify molecular fragments that implicitly cause the decreasing or increasing of LD50 values (Schemes 1 and 2). For example, molecular fragments found to contribute to the lowering of toxicity are usually alcohols, sugars or ethers (Scheme 1), which are generally accepted to be safe. On the contrary, amines, phosphorus derivatives or halogen-containing compounds (Scheme 2) are as associated with chemicals that may be toxic.

CONCLUSIONS

A PLS model with high significance and good predictivity was generated to estimate the acute oral toxicity in rats. The model uses a combination of chemical and biological descriptors, and is statistical more powerful than the model generated

based on chemical descriptors only. Biological data alone (HTS data) did not offer sufficient information to obtain a PLS model with satisfactory statistical significance. Chemical descriptors expressed as circular ECFP fingerprints help identify molecular fragments that are likely to be associated with increased, or decreased oral rat LD50 values, respectively. Also, we illustrated how biologic descriptors can assist in developing novel mechanistic insights into complex toxicological endpoints such as LD50. Despite inherent limitations (e.g., lack of predictivity for novel chemicals due to the termination of the Molecular Libraries Program), the methodology described here can be extended to novel endpoints, as well as novel chemical libraries, provided that biomolecular screening is used as supplement to informatics-based systems.

REFERENCES

1. J. A. DiMasi, R. W. Hansen and H. G. Grabowski, *J. Health Economics,* **2003,** *22*, 151-185.
2. The Toxic Substances Control Act (15 U.S.C. 2601–2692) consists of Public Law 94–469 (Oct. 11, 1976; 90 Stat. 2003) and the amendments made by subsequent enactments. See also http://www.epw.senate.gov/tsca.pdf
3. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:037:0019:0023:EN:PDF
4. W. Muster, A. Breidenbach, H. Fischer, S. Kirchner, L. Müller L and A. Pähler, *Drug Discov. Today*, **2008**, *13*, 303-310.
5. http://echa.europa.eu/regulations/reach
6. http://www.opentox.org/
7. http://www.ncats.nih.gov/research/reengineering/tox21/tox21.html
8. D. M. Sanderson and C. G. Earnshaw, *Hum. Exp. Toxicol.,* **1991**, *10*, 261-273.
9. http://www.lhasalimited.org/products/derek-nexus.htm
10. N. Green, *Adv. Drug Deliv. Rev.,* **2002**, *54*, 417-431.
11. S. Ekins, *J. Pharmacol. Toxicol. Methods.,* **201**4, *69*, 115-140.
12. N. Jeliazkova, *Expert Opin. Drug Metab. Toxicol*., **2012**, *8*, 791-801.
13. http://www.ccohs.ca/oshanswers/chemicals/ld50.html
14. The Hazardous Substances Data Bank, http://toxnet.nlm.nih.gov/newtoxnet/hsdb.htm, accessed at May 2014
15. C. P. Austin, L. S. Brady, T. R. Insel and F. S. Collins, *Science*, **2004**, *306*, 1138-1139.
16. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, *J. Cheminf*, **2011**, *3*, 33.
17. The Open Babel Package, version 2.3.1, http://openbabel.org
18. D. Rogers and M. Hahn, *J. Chem. Inf. Model,* **2010** *50*, 742-754.
19. JChem Base was used for structure searching and chemical database access and management, JChem 5.9.4, 2012, ChemAxon (http://www.chemaxon.com).
20. S. Wold, A. Ruhe, H. Wold and W. J. Dunn, *J. Sci. Stat. Comput.,* **1984**, *5*, 735-743.
21. Simca is available from Umetri AB, Umeå, Sweden, http://www.umetri.com/
22. S. Wold, *Technometrics,* **1978**, *20*, 397-405.
23. M. Olah, R. Rad, L. Ostopovici, A. Bora, N. Hadaruga, D. Hadaruga, R. Moldovan, A. Fulias, M. Mracec and T.I. Oprea, "Chemical Biology: From Small Molecules to Systems Biology and Drug Design", New-York, **2007**, Schreiber SL, Kapoor TM, Wess G (Eds), Wiley-VCH, pp 760-786.
24. C. L. Gavaghan, C. Hasselgren Arnby, N. Blomberg, G. Strandlund G and S. Boyer. *J. Comput.-Aided Mol. Design,* **2007**, *21*, 189-206.
25. S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G Patlewicz, J. Niemela and O. Mekenyan, *J. Chem. Inf. Model,* **2005**, *45*, 839-849.
26. A. Sedykh, H. Zhu, H. Tang, L. Zhang, A. Richard,I. Rusyn and A. Tropsha. *Environ. Health Perspect,* **2011**, *119*, 364-370.