Current trends in quantitative structure–activity relationship validation and applications on drug discovery

First draft submitted: 22 April 2017; **Accepted for publication:** 28 April 2017; **Published online:** 14 July 2017

Keywords: big data • CADD • drug discovery • QSAR

Quantitative structure–activity relationship history, current status & the importance of validation

Methods to correlate biological activity and chemical structure of compounds have been employed since 1868 [1]. Early on, quantitative structure–activity relationship (QSAR) analyses were performed using experimentally determined physicochemical properties, such as logarithm of water/*n*-octanol partition coefficient (log *P*), hydrophobic constant (π) and Hammett electronic constant (σ), which were then correlated with the biological activity of the tested compounds [2].

Nowadays, there are several prerequisites to construct and apply QSAR models. From the applicability point of view, QSAR requires a compound set that has been tested against an identified molecular target, cell, tissue, or even microorganism, under the same experimental conditions, and possesses the minimum variance in the observed responses [3]. Once an appropriate dataset has been selected, the main steps of a QSAR modeling require molecular/physicochemical properties, followed by variable selection, model generation from different algorithms and, most importantly, a validation process using internal and external datasets.

After the generation and validation of a QSAR model, the model can be employed in predictions of biological activity of new samples and a physicochemical interpretation of the observed phenomena could be also conducted, providing insights for the design of

new bioactive chemicals and/or their molecular mechanism of action. Subsequently, QSAR models have been widely employed in several steps of the drug design process, for instance, in order to understand and predict the compound binding affinity for a specific molecular target. QSAR modeling is also applied to better comprehend general phenomena such as pharmacokinetics and toxicity-related end points, which generally are standard measurements and for which there are available large datasets [4].

Complementarily, although animal testing is still considered crucial to the evaluation of chemical safety, toxicity testing is moving toward a greater understanding of the disease at multiple biological levels, so as to develop alternative methods [5,6]. From the biological point of view, various models representing several layers of human physiology and metabolism, both healthy and disease based, have been developed. In order to address general problems, such as skin sensitization, hepatotoxicity and DNA harming agents, some research groups have built local QSAR models using mechanistic information [6,7]. However, in the future, we can easily expect that improvements in the validation techniques and data quality will follow the growth in the data generation, leading to broader QSAR models.

There are several relevant articles in the QSAR field describing methods and validation techniques [8], besides extensive literature on troubleshooting, which makes this field still highly attractive for research and

Vinicius G Maltarollo^{*1},
Thales Kronenberger²,
Carsten Wrenger² & Kathia M
Honorio^{3,4}

¹Department of Pharmaceutical Products, Faculty of Pharmacy, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

²Unit for Drug Discovery, Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil

³Center for Natural & Human Sciences, ABC Federal University, Santo André, São Paulo, Brazil

⁴School of Arts, Sciences & Humanities, University of São Paulo, São Paulo, Brazil

*Author for correspondence:

viniciusmaltarollo@gmail.com

FUTURE
SCIENCE

part of

fsg

applications. In this sense, the Organisation for Economic Co-operation and Development has a main role, providing guidelines for the evaluation of the generated models and their quality [9].

Trends in scientific computation & big data in the QSAR field

With the advances in scientific computation and theoretical methods to calculate chemical structures and their properties, QSAR approaches have become physicochemically and statistically more complex. As an example, multidimensional QSAR approaches (such as 3D- and 4D-QSAR, among others) were developed as an attempt to describe the mechanism related to lock-and-key and induced-fit theories, respectively [10].

Perspectives from the last decade expected that, aside from the dramatic improvements in the computer processing unit (CPU) speed and algorithm development, the greatest technological impact on modern QSAR would be the unbridled generation of new molecular descriptors. These descriptors could encompass new approximations for known ones, as well as new methods to calculate free energy and advances in the quantum chemical calculation [11].

These molecular descriptors would rather be calculated than experimentally determined; this leap in the descriptor calculation and a decrease in the quality of QSAR models are not accidental. More than ever we have new high-quality predictions of physicochemical properties, which has enabled the development of complex QSAR models with a larger number of correlated variables [12]. Still, the association between occasional correlations together with the misuse of the number of parameters has led to a misunderstanding of QSAR as a scientific technique. In the last few years, the modeling community has been more concerned about the criterion for the evaluation of the model's predictability. Now, a consensus has been reached considering the external validation as the best way to establish a reliable QSAR model, which is a positive novelty in a field once dominated by the use of q^2 [8,13].

In addition to various parameters, the increased rate in the data generation poses a new paradigm where data from different experimental domains have to be analyzed, integrated and, mainly, contextualized [14]. Although this approach is common in big pharma, the general public still suffers from the absence of efforts to integrate the data and mostly individual decisions are taking place. This is a general issue since, despite a large amount of compound-target data in the public domain, they still require a substantial amount of manual effort before proper use, such as proper dataset curation, preliminary physicochemical hypotheses and manual decision making.

Machine-learning techniques have been widely employed in the QSAR field to construct regression and classification models using public domain large datasets and/or using large sets containing calculated descriptors. Recently, a new proposed trend involves the use of combinatorial approaches, since the combination of equivalent models can reduce the variance in the predictions [15,16]. Also, models generated with the most frequently selected descriptors by the equivalent algorithms or full models combining numerous homogeneous or heterogeneous (*e.g.*, generated by different learners) individual models provide better performance [11,17].

Currently, most of the QSAR models generated by medicinal chemists are based on an interdisciplinary team's intuition, who, through several structure-activity relationship (SAR) steps, suggest chemical modifications. Since at this time most of the experimental drug-discovery projects are automated in an integrated fashion, the decision making is the main bottleneck and a shift toward algorithm-based decisions would be more cost effective.

Pharmacoinformatic methods require the largest possible chemical space to improve the applicability domain of a QSAR model. However, the fact that most of the datasets derive from small SAR campaigns and the lack of connection between these compound pools makes the chemical space difficult to explore. Pressure on academics to publish new compounds with positive results has shifted the proportion in the public datasets, leading to a lack of true inactive compounds compared with active ones. This is a fact that, together with the low exploitation of the chemical space, inserts structural gaps in modeling [18].

Future perspective in new technologies & large-scale QSAR

Along the years, several publications have discussed how to conduct an adequate and valid QSAR study and, now more than ever, a large amount of both data and computational power allow large-scale QSAR studies.

Technologies such as graphical processing units and the use of servers and cloud computing are already implemented in other fields of computer-aided drug design and are routinely integrated in drug discovery campaigns, in both academia and big pharma. Big pharma companies have comprehensive internal databases of unique synthesized compounds, and recent initiatives of data sharing by partnerships among companies and even with academic research groups are helping to overcome obstacles such as intellectual property barriers.

Due to all the aforementioned factors, we expect a shift in the QSAR role. QSAR would move from being applied to lead optimization processes toward a role as a tool integrating early stages of computer-aided drug design and experimental drug-discovery pipelines [19,20]. However, there are still challenges to be overcome, such as the implementation of robust validation processes and automation of different parts involved in the generation of predictive QSAR models.

Financial & competing interests disclosure

The authors would like to acknowledge the support from the São Paulo state funding agency – FAPESP (grants

2014/03644-9, 2014/27313-1 and 2015/26772-8) and CNPq (project number 456984/2014-3). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Open access

This work is licensed under the Creative Commons Attribution 4.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

References

- Selassie C. History of Quantitative Structure-Activity Relationships. In: *Burger's Medicinal Chemistry and Drug Discovery*. Abraham DJ (Ed.). John Wiley & Sons, Inc., CA, USA, 3–46 (2003).
- Hansch C, Fujita T. p - σ - π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* 86(8), 1616–1626 (1964).
- Kunal Roy, Supratik Kar RN Das. Statistical Methods in QSAR/QSPR In: *A Primer on QSAR/QSPR Modeling - Fundamental Concepts*. Hawkins E (Ed.). Springer International Publishing, New York Dordrecht, London, 37–59 (2015).
- Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* 112(12), 1249–1254 (2004).
- Pradeep P, Povinelli RJ, White S, Merrill SJ. An ensemble model of QSAR tools for regulatory risk assessment. *J. Cheminform.* 8(1), 48 (2016).
- Patlewicz G, Casati S, Basketter DA *et al.* Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? *Regul. Toxicol. Pharmacol.* 82, 147–155 (2016).
- Tomic S, Nilsson L, Wade RC. Nuclear receptor-DNA binding specificity: a COMBINE and free-wilson QSAR analysis. *J. Med. Chem.* 43(9), 1780–1792 (2000).
- Roy K, Das RN, Ambure P, Aher RB. Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom. Intell. Lab. Syst.* 152, 18–33 (2016).
- Organisation for Economic Co-operation and Development. Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models. www.oecd.org/officialdocuments
- Lill MA. Multi-dimensional QSAR in drug discovery. *Drug Discov. Today* 12(23–24), 1013–1017 (2007).
- De Benedetti PG, Fanelli F. Multiscale quantum chemical approaches to QSAR modeling and drug design. *Drug Discov. Today* 19(12), 1921–1927 (2014).
- Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discov. Today* 21(8), 1291–1302 (2016).
- Golbraikh A, Tropsha A. Beware of q^2 ! *J. Mol. Graph. Model.* 20(4), 269–276 (2002).
- Richter L, Ecker GF. Medicinal chemistry in the era of big data. *Drug Discov. Today Technol.* 14, 37–41 (2015).
- Maltarollo VG, Gertrudes JC, Oliveira PR, Honorio KM. Applying machine learning techniques for ADME-Tox prediction: a review. *Expert Opin. Drug Metab. Toxicol.* 11(2), 259–271 (2015).
- Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* 20(3), 318–331 (2015).
- Baurin N, Mozziconacci J-C, Arnoult E, Chavatte P, Marot C, Morin-Allory L. 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database. *J. Chem. Inf. Comput. Sci.* 44(1), 276–285 (2004).
- Maggiora GM. On outliers and activity cliffs—why QSAR often disappoints. *J. Chem. Inf. Model* 46(4), 1535 (2006).
- Melo-Filho CC, Dantas RF, Braga RC *et al.* QSAR-driven discovery of novel chemical scaffolds active against *Schistosoma mansoni*. *J. Chem. Inf. Model.* 56(7), 1357–1372 (2016).
- Andrade CH, Pasqualoto KFM, Ferreira EI, Hopfinger AJ. 3D-Pharmacophore mapping of thymidine-based inhibitors of TMPK as potential antituberculosis agents. *J. Comput. Aided. Mol. Des.* 24(2), 157–172 (2010).