

Towards a unifying strategy for the structure-based prediction of toxicological endpoints

Pau Carrió, Ferran Sanz, Manuel Pastor*

Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Hospital del Mar Medical Research Institute (IMIM), carrer Dr. Aiguader 88, E-08003 Barcelona, Spain.

Corresponding Author: Manuel Pastor

E-mail: manuel.pastor@upf.edu

[Tel.] +34 933 160 512

[Fax.] +34 933 160 550

Abstract

Most computational methods used for the prediction of toxicity endpoints are based on the assumption that similar compounds have similar biological properties. This principle can be exploited using computational methods like read-across or quantitative structure-activity relationships. However, there is no general agreement about which method is the most appropriate for quantifying compound similarity neither for exploiting the similarity principle in order to obtain reliable estimations of the compound properties. Moreover, optimal similarity metrics and modeling methods might depend on the characteristics of the endpoints and training series used in each case. This study describes a comparative analysis of the predictive performance of diverse similarity metrics and modeling methods in toxicological applications. A collection of two quantitative (n=660, n=1114) and three qualitative (n=447, n=905, n=1220) datasets representing very different endpoints of interest in drug safety evaluation and rigorous methods were used to estimate the external predictive ability in each case. The results confirm that no single approach produces the best results in all instances and the best predictions were obtained using different tools in different situations. The trends observed in this study were exploited to propose a unifying strategy allowing the use of the most suitable method for every compound. A comparison of the quality of the predictions obtained by the unifying strategy with those obtained by standard prediction methods confirmed the usefulness of the proposed approach.

Keywords

In silico toxicity prediction, QSAR, QSPR, read across, chemical domain.

Introduction

Obtaining novel active ingredients in the pharmaceutical, agrochemical and cosmetic industry requires the use of diverse methods for ascertain their biological properties and assessing potential toxic liabilities at different development stages. Experimental methods have the inconvenient of consuming existing product and might have associated significant costs in terms of money and time. In vivo experimental methods are particularly costly and in some cases they can be ethically unjustifiable or not legally allowed (EC 2015). For these reasons, computational (in silico) methods are a very attractive alternative: they are fast, cheap, consume no compound and can be carried out even before the compound is synthesized. In silico methods have been in use in areas of toxicology, like ecotoxicology, for a long while (Andersson et al. 2002; Könemann 1980; Könemann and Musch 1981; Perkins et al. 2003) but their application for drug safety assessment is more recent (Muster et al. 2008; Raunio 2011; Ekins 2014).

The latest strategies for efficient risk assessment of chemicals propose the replacement of animal models by integrated approaches that incorporate a combination of in vitro and in silico methods, including quantitative structure-activity relationships (QSAR) modeling, physiologically based biokinetics (PBPK) and biodynamics (PBBD) modeling as well as information on the compound absorption, distribution, metabolism and excretion (Wilk-Zasadna et al. 2015; Yoon et al. 2015). This change of paradigm will probably show the way to go in toxicological assessment for the next decades (NRC 2007). However, its routine application in drug development pipeline is still in its early stage of development (Kramer et al. 2015). In the present work we will focus only in the hazard characterization of new drug candidates at early stages of development. Traditionally, this was carried out evaluating the concentration-effect relationship (or any point in this relation, e.g. a point-of-departure concentration) using in vitro methods. Structure-based in silico methods used in this area can be seen as a direct replacement of such in vitro methodologies, and the prediction results as an estimate of the hazard expressed either in term of the concentration-effect (predicted IC_{50} or K_d) or in categorical terms (positive, negative).

From a methodological perspective, most applications of structure-based in silico methods in drug safety assessment fall in one of the following categories: read across (RA) methods or quantitative structure-activity relationships (so called QSAR). RA methods aim to define chemical classes around previously characterized compounds assuming that any compound belonging to the same class will likely share the same biological properties (Schultz et al. 2015). QSAR methods go one step further and analyze the association between chemical structure differences and biological property differences for a collection or previously studied compounds (training series), building a mathematical model that describes this relationship (Tropsha 2010). It must be noted that both categories, in spite of their many differences, are based on the concept of bioisosterism or similar property principle (Eckert and Bajorath 2007), which states that compounds with the same chemical structure should have similar biological properties. However, this apparently simple principle is very difficult to capture (Maggiora et al. 2014) and there are many examples where it does not hold (Kubinyi 1998; Nikolova and Jaworska 2003; Roy et al. 2012; Guha 2012; Medina-Franco 2013; Golbraikh et al. 2014; Bajorath 2014). Structure-activity landscapes might have smooth domains, sets of similar compounds both in terms of structure and biological properties, where the principle holds and rough domains where it does not, often called activity cliffs (Maggiora 2006).

The structure-activity landscape smoothness depends of the biological property studied and of the structural description used to assess the compound similarity. It can be hypothesized that the use of biologically-relevant structural descriptors will help to produce smoother landscapes, but in practice the complexity of the biology involved (e.g. different ligand binding modes, diverse mechanisms or the influence of pharmacokinetics) makes almost impossible finding a single method that is optimal in all cases. Figure 1 shows three examples representing extreme situations. All compounds in series A behave in a homogeneous way and the biological property (represented in the Y axis) correlates well with the structural description (represented in the X axis). Here, a single QSAR model can produce good predictions. On the contrary, the compounds in the series B belong to diverse classes. For some classes the biological property could be modeled using local QSAR models, even if the functions describing the association between the structure and biological property are different for each of them. Moreover, for some of these classes the value of the biological property is nearly constant for all the compounds and no QSAR model is actually required; the average value of the class members could provide a good estimation. The biological property of compounds in series C does not show any

correlation with the structure but very similar compounds exhibit the same values and therefore RA methods could still be applied. These examples do not represent only theoretical scenarios and similar situations have been described in toxicology (Martin 1981; Bajorath et al. 2009; Medina-Franco 2012; Dimova and Bajorath 2014), where good QSAR models for predicting complex endpoints are uncommon and RA methods have been proposed to be more adequate (Sheridan 2014; Schultz et al. 2015). This observation can be explained by the limitations of *in silico* methods (Bajorath 2012; Cherkasov et al. 2014), in particular when *in silico* methods are applied to model the complex biological phenomena involved in toxicological endpoints like organ toxicity (Liebler and Guengerich 2005; Treinen-Moslen and Kanz 2006; Curigliano et al. 2010; Leise et al. 2014) or NOAEL/LOAEL (FDA 2005; Park and Cho 2011; Muller and Milton 2012). Observable outcomes are the result of the compound hazard and exposure and might involve an unknown number of mechanisms, justifying the presence of rough structure-activity landscapes (like cases B and C in figure 1) where the bioisosterism principle is applicable only for closely related compounds. These examples serve to illustrate the central hypothesis of this work: the smoothness of the landscape conditions the applicability of diverse methodologies (global QSAR, local QSAR and RA) that can be seen as alternative and complementary approaches for exploiting the structure-activity principle. As a consequence, the selection of the most appropriate approach based on the characteristics of the structure-activity landscape will produce the best predictions.

In this article we will test the validity of the hypothesis presented above, particularly for the prediction of toxicological endpoints, and propose an integrated strategy for taking advantage of the best approach for each particular situation. First, we will compare the performance of diverse molecular similarity metrics and study the roughness of the respective structure-activity landscapes. Then, we will build RA, global and local QSAR models on a collection of datasets annotated for endpoints of diverse complexity. The analysis of the model performance in these examples will identify trends useful for selecting the best similarity metrics and predictive methods.

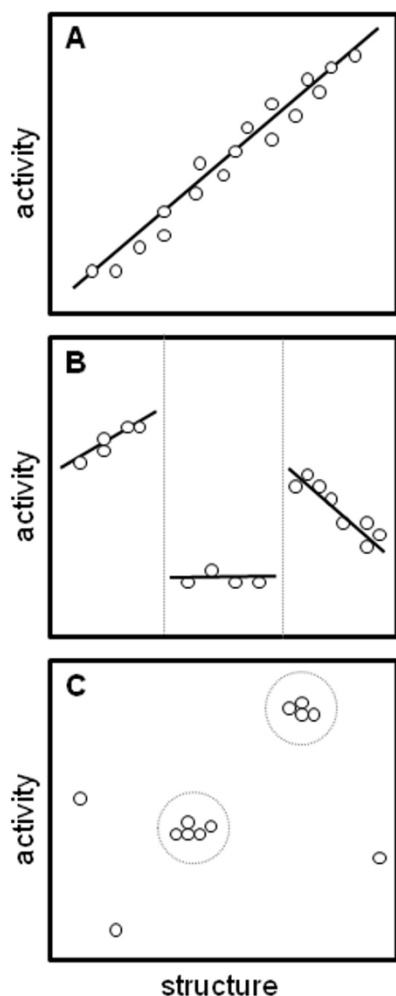


Fig. 1 Diverse extreme situations describing the relationship between the compound structures and activities: (A) there is a clear and smooth correlation between both; (B) the structure and the activity shows a clear association, but this is different for diverse classes of compounds with different structure; (C) there is no clear association, even if some structurally similar compounds have similar biological properties.

Methods

Data sets

Datasets were selected to represent endpoints commonly used in early drug development and drug safety assessment studies and to cover a wide range of complexity; from relatively simple to very complex ones. All datasets are large (min 447 compound) and structurally diverse. The collection contains five examples of quantitative and qualitative endpoints: aqueous solubility (SOLU), potassium channel hERG blocking (hERG), P-glycoprotein inhibition (ABCB1), drug-induced

phospholipidosis (DILI) and drug-induced liver Injury (DILI). The dataset characteristics are summarized in Tables 1 and 2.

The SOLU dataset contains 1144 water solubility values for low molecular weight organic compounds (Delaney 2004). The solubility values range from -11.60 to 1.58 units with a mean of -3.06 and a standard deviation of 2.10. Oral bioavailability of drugs critically depends on their water solubility. Also, poor water solubility has been associated with safety concerns and undesirable side effects (Alelyunas et al. 2010). Accurate in silico prediction of water solubility is difficult (Hua et al. 2007). Paradoxically, this endpoint represents the lowest extreme in the scale of the complexity of the studied endpoints, since it can be assumed that the physicochemical mechanisms involved in the solvation are rather similar for all the compounds and comparatively simpler than the mixture of molecular and physiological mechanisms involved in the rest of endpoints described here.

The hERG dataset contains 750 pIC₅₀ values for hERG potassium channel blockade obtained from (Obiol-Pardo et al. 2011) and (Li et al. 2008). Activities range from 2 to 9 units, with a mean of 5.51 and a standard deviation 1.2. Blockade of the hERG channel plays a major role in drug-induced QT prolongation which is associated to an increase in risk of sudden cardiac arrest (Hancox et al. 2008) and is a classical liability that needs to be tested at early development stages for any drug candidate. Ligands blocking this channel typically bind at the pore domain lining the central ion conduction pathway. This region contains several binding positions and differentiated states (Vandenberg et al. 2012). Indeed, hERG channel is a good example of polyspecific biomolecule able to interact with many classes of compounds (Thai et al. 2010a), making challenging the development of in silico predictive models. Many structure-based, pharmacophore and QSAR models, aiming to understand the structure-activity relationships that govern hERG-drug interactions have been published (Aronov 2008). The development of such models is further hampered by the low quality of training data extracted from bibliographic sources, due to the large disparity of the experimental conditions (e.g. cell line, temperature, pH) used in the original experiments.

The ABCB1 dataset (Broccatelli et al. 2012) contains 562 inhibitors and 515 non-inhibitors of ABCB1. ABCB1, also known as P-glycoprotein, is a membrane protein member of the ATP-binding cassette (ABC) transporters superfamily, which transports a variety of compounds through the membrane

against a concentration gradient (Choudhuri and Klaassen 2006). ABCB1 is of interest due to its duality as therapeutic target and antitarget (Juliano and Ling 1976; Borst and Elferink 2002; Aller et al. 2009; Broccatelli et al. 2010; Klepsch and Ecker 2010). Modeling transporter binding is difficult due to their intrinsic ligand promiscuity, since these biomolecules have evolved to transport numerous structurally and functionally unrelated compounds (Szakács et al. 2006). The case of ABCB1 is further complicated by the existence of diverse ligand binding modes (Benet 2009). The overall assumption is that ABCB1 possesses a huge binding pocket with at least two distinct binding sites (Loo and Clarke 2002).

The DIPL data set was curated by (Przybylak et al. 2014) from two US FDA data sets (Kruhlak et al. 2008; Orogo et al. 2012) and contains 215 phospholipidosis inducers and 232 non-inducers. The quality of these data is limited by the disparity of the species and tissue types used in the experiments. Typically, the ability of a compound to produce phospholipidosis is assessed using experimental methods that detect the accumulation of phospholipids within the cells of different tissues. Sometimes this is only observable when the compound is administered at high doses, therefore producing an unknown number of false negatives in the training series (Reasor et al. 2006). Furthermore, there are several possible mechanisms by which drugs can induce phospholipidosis (Sawada et al. 2005). In the original article SMARTS patterns that can be used as structural alerts for phospholipidosis were also provided.

The DILI data set contains 525 drug-induced liver injury (DILI) inducers and 234 non-inducers collected by a data mining study (Fourches et al. 2010). The original data set collected drug liver effects in different species, comprising humans, rodents, and non-rodents, but in the present work only human data was used. Compounds with lack of reported effects were classified as non-inducers. Drug-induced liver injury (DILI) is one of the main causes of attrition both for candidate and marketed drugs (Fung et al. 2001). Liver injury can be produced by a large number of mechanisms. Moreover, liver injury might be produced by a metabolic product rather than the parent drug. Hence, for all the aforementioned reasons, this dataset is on the top position in the scale of endpoint complexity and no accurate prediction can be expected. (Hewitt et al. 2013) provided SMARTS patterns that can be used as structural alerts for non-inducers and inducers.

Table 1. Characteristics of the quantitative datasets. ^a Number of compounds (n). ^b Mean, standard deviation (SD), and range of the studied property.

dataset	Endpoint	<i>n</i> ^a	mean ^b	SD ^b	range ^b	Units	source
hERG	Functional inhibition measured in patch clamp experiments	660	5.41	1.17	2 to 9	pIC ₅₀	(Li et al. 2008; Obiol-Pardo et al. 2011)
SOLU	Water solubility	1114	−3.06	2.10	−11.60 to 1.58	log ₁₀ (mol/L)	(Delaney 2004)

Table 2. Characteristics of the qualitative datasets, after data curation

dataset	Endpoint	negative/positive	source
DILI	Drug-induced liver injury	289/616	(Fourches et al. 2010)
DIPL	Drug-induced phospholipidosis	232/ 215	(Przybylak et al. 2014)
ABCB1	P-glycoprotein inhibition	567/653	(Broccatelli et al. 2011)

For all datasets, the compound structures provided as SMILES in the original sources were converted first to Mol format using RDKit (Landrum). The presence of duplicate parent structures was tested by obtaining the compounds InChIkey with RDKit and comparing all string pairs. Whenever duplicated entries were detected they were merged into a single compound and the average of their biological properties was used. The ionization status of all structures was adjusted to pH 7.4 using Moka 1.1.0-RC3 (Milletti et al. 2007; Milletti et al. 2009). When necessary, 3D structures were generated using CORINA 2.4 (Sadowski and Gasteiger 1993; Sadowski et al. 1994). Compounds that failed to pass this curation protocol for any reason were discarded. The values reported in Tables 1 and 2 make reference to the final datasets obtained after the curation process.

Molecular Descriptors

Molecular weight (MW), MACCS Fingerprints (Accelrys), MORGAN Fingerprints with radius 2 (Morgan 1965; Rogers and Hahn 2010) and Murcko scaffolds (Bemis and Murcko 1996) were generated with RDKit. GRIND Independent Descriptors of second generation (GRIND-2) (Pastor et al. 2000; Pastor 2006) were generated using Pentacle version 1.06 (Durán et al. 2008; Durán and Pastor 2010) with default parameters and used without scaling, as recommended by the authors.

QSAR and RA models

Global QSAR models were built using PLS regression (PLS-R) (Martens 2001), Random Forest regression (RF-R) (Breiman 2001) and Support Vector Machine regression (SVM-R) (Cortes and Vapnik 1995) for quantitative endpoints and PLS discriminant analysis (PLS-DA), RF classifier (RF-C) and SVM classifier (SVM-C) for qualitative endpoints. SVM used the radial kernel. SVM gamma value and PLS number of latent variables were tuned to obtain best cross-validation results. Random Forest number of trees was set to 500.

Local models were built using the metrics shown in Table 3 to define domains of similar compounds. Query compounds were assigned to the closer domain. When it contains more than five compounds, a PLS model (PLS-R or PLS-DA for quantitative and qualitative endpoints, respectively) with a number of latent variables maximizing cross-validation results is built and used to predict the query compound property, otherwise the mean (for quantitative endpoints) or majority voting (for qualitative endpoints) of the activities in the chemical domain was used as prediction.

RA models were built using either MS or MORGAN metrics. For RA-MS, all compounds with the same Murcko scaffold were assigned to the same category. RA based on MORGAN space was applied as described by (Enoch et al. 2009) in two alternative ways: RA fix nn uses the 10 nearest compounds to form a category while RA fix th uses the compounds with Tanimoto distance less than 0.6 to form a category. In all cases, the predictions obtained from RA correspond to the median (for quantitative endpoints) or majority voting (for qualitative endpoints) of the class activities.

The model building and validation was carried out using R scripts that make use of the pls (Mevik and Wehrens 2007), e1071 (Meyer et al. 2014), class (Venables and Ripley 2002) and randomForest (Liaw

and Wiener 2002) R packages. Structural alerts were implemented using an in-house developed R script that makes use of RDKit.

Molecular similarity metrics

Different metrics were used to calculate pair wise compound similarities and to build chemical classes encompassing similar compounds. Selected metrics are representative of the current state-of-the-art methods used to assess chemical similarity (Guha et al. 2006; Hua et al. 2007; Helgee et al. 2010). Some metrics rely on simple properties, like molecular weight, while others make use of molecular descriptors or fingerprints, as those described above. It is useful to distinguish between activity unbiased and activity biased metrics. Activity unbiased metrics are based only on the compound structure while activity biased includes somehow the values of the biological properties of the training series for defining the metric space. In the MW metric the distance between two compounds is defined by the absolute value of the difference between their molecular weights. MACCS and MORGAN metric spaces are defined using MACCS fingerprints and MORGAN fingerprints, respectively for computing 1-Tanimoto distance (Willett et al. 1998). MS metric is based on Murcko scaffolds: the distance between two compounds is 0 if they share the same Murcko scaffold and 1 otherwise. Pentacle-PCA and Pentacle-PLS metrics is based on Euclidean distances in the PCA and PLS scores space of GRIND-2 molecular descriptors, calculated as described above. The number of principal components and latent values was set to 2, which explained more than 50% of the variance in all instances. The main characteristics of the metrics used in this article are summarized in Table 3.

Table 3. Similarity metrics used in this article

Name	Molecular descriptors	Distance definition	Activity-biased
MW	Molecular Weight	$ MW_i - MW_j $	No
MACCS	MACCS Fingerprints	1-Tanimoto	No
MORGAN	Morgan Fingerprints	1-Tanimoto	No
MS	Murcko Scaffolds	0 if compounds share scaffold, 1 otherwise	No

Pentacle-PCA	GRIND-2	Euclidean in the PCA scores space	No
Pentacle-PLS	GRIND-2	Euclidean in the PLS scores space	Yes

Metric performance was assessed in terms of the internal validity of the bioisosterism principle in the space defined by this metric for a given set of compounds. The procedure is described in detail the next section.

Chemical domains

The so called "chemical domains" describe a collection of similar compounds, grouped according to a certain similarity criterion. Chemical domains were built as follows: for a give metric (see Table 3) the distance matrix between all compounds in the training series was computed. Then, hierarchical clustering with complete linkage method was used to build the similarity tree using the hclust function from R base package. Complete linkage was chosen since it helped to keep low distances among members in the cluster. Then we used a similarity threshold to define clusters containing only compounds with internal similarity under this given value (see Figure 2). Threshold values were selected based on the quantiles 0.05, 0.1, 0.2, 0.4 of distance pairs. For example, a threshold of 0.05 defines clusters whose maximal distance between cluster members is lower than the 5 percent of all distance pairs. In order to avoid chemical domains containing highly diverse compounds, the standard deviation of the biological property within each cluster was computed and compared with the standard deviation of the dataset. Domains for which this value was much larger (1.5 fold) were discarded. Test compounds were assigned to the chemical domain with the closest centroid, but if the distance between the closest centroid and the query compound was greater than the maximal distance among domain members the domain assignment was rejected.

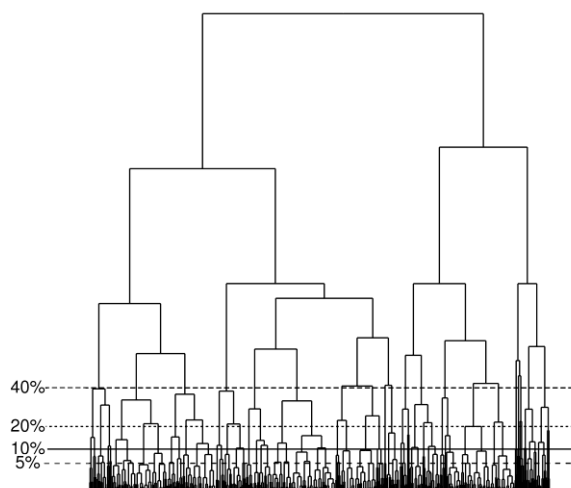


Fig. 2 Hierarchical clustering example with different similarity threshold values.

In this work, chemical domains were used both to assess the quality of different similarity metrics and to build local models on top of them. The metric quality was assessed by defining chemical domains with this similarity metric and representing them side by side, ordered by the median of the compound biological property. If the bioisosterism principle holds for this particular metric and dataset, the values of the biological property will exhibit very low dispersion within the domains (similar compounds will have similar biological properties). Conversely, the median values of the biological properties obtained for the different chemical domains will exhibit a large dispersion and therefore, if represented side-by-side (as in Figure 3) there will be a clear growing tendency from left to right (diverse compounds will have diverse biological properties).

Predictive quality assessment

The quality of all models was evaluated by comparing the model predictions with the experimental values for a set of external test compounds, not used for building the model. All the studied datasets were split randomly 10 times in training and test series containing 80% and 20% of the compounds, respectively. The final prediction scoring aggregates the results of these 10 training-test set splits. For quantitative endpoints (hERG and SOLU) the mean Standard Deviation Error of the Prediction (SDEP) was used. For qualitative endpoints (ABCB1, DILI and DIPL) it was used the mean Accuracy, calculated as:

$$\text{Accuracy} = \text{TP} + \text{TN} / (\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

where TP means True Positive; TN means True Negatives; FP means False Positives and FN means False negatives .

The source code of all the scripts used in this work for computing the chemical domains, the models and the predictive quality estimation is available as Supplementary Information.

Results and Discussion

In this section we will compare the performance of diverse molecular similarity metrics and analyze the roughness of the respective structure-activity landscapes. Then, we will build RA, global and local QSAR models, comparing their predictive performance. From these analyses we aim to identify the best metrics and modeling methods and if these choices have universal validity or if they depend on the datasets characteristics.

Performance of molecular similarity metrics

As stated in the introduction, the adequacy of molecular similarity metrics for biological applications can be judged by the roughness of the structure-activity landscape they generate. When using a good metric, structurally similar compounds will have similar biological properties (Martin et al. 2002). This property was assessed in our study by clustering all the datasets listed in Tables 1 and 2 using the metrics listed in Table 3, as described in the Methods section, using a threshold of 0.4. The grouping obtained for the quantitative datasets (SOLU and hERG), was represented in Figure 3 as a series of box plots side by side, one for each cluster, ordered from lower to higher median property value. Ideally, all the compounds belonging to the same cluster must have similar biological properties. The intra-cluster dispersion represents how smooth or rough the structure-activity surface is: the greater the dispersion, the rougher the surface and the less adequate the metric. Also, the mean activity of the diverse clusters must be different from left to right. Indeed, the comparison of the inter-cluster and the intra-cluster dispersion in these graphics gives a good idea of how much the activity can be explained in terms of molecular similarity for the different metrics and endpoints. Different metrics produce different number of clusters. MW metric produces fewer clusters since it is a very simple metric based in a single criteria while, on the opposite side, MORGAN and MACCS yield many

different clusters based on a much rich description of the structures. Metrics based on GRIND-2 are located in the middle.

The inspection of Figure 3 shows that the boxes width is highly variable for fingerprint metrics (MACCS, MORGAN) with many narrow or wide boxes. This variability is less pronounced for MW and Pentacle metrics, where the boxes width is more homogeneous. MW produces a few chemical classes of very similar compounds, something that can be explained by the fact that the small compounds tend to have low hERG blockade (Recanatini et al. 2008). For GRIND-2, the activity biased metric (Pentacle-PLS) explains better the activity differences than the non activity biased (Pentacle-PCA), particularly in the case of hERG, as can be seen by the largest differences of the chemical classes median values. The comparison of the two quantitative endpoints indicates that SOLU is slightly better explained than the hERG with larger inter-cluster differences and smaller oscillation of the intra-cluster dispersion (see for example the MACCS or Pentacle-PLS metrics). This can be justified by the lower complexity of the phenomenon explained (water solubility vs. hERG receptor blockade) and the higher accuracy of the experimental measurements.

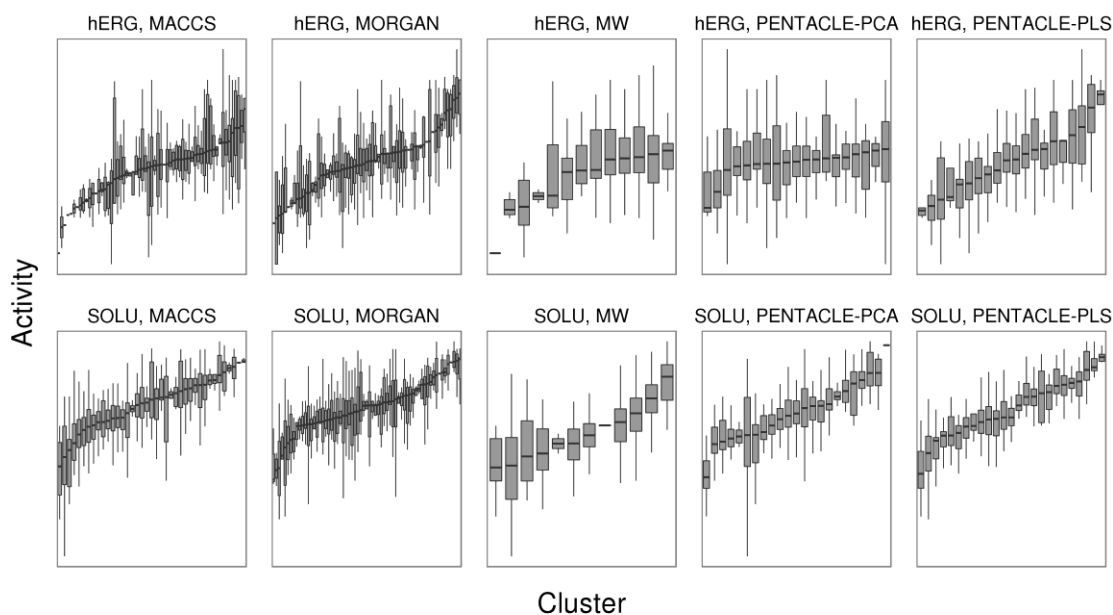


Fig. 3 Distribution of the biological property (activity) values of the compounds in the chemical domains obtained using diverse similarity metrics, for datasets hERG and SOLU. The chemical domains were ordered according to their median activity value, from left to right. See text for details.

Figure 4 shows the result of an equivalent analysis for the datasets representing qualitative endpoints (ABCB1, DIPL, DILI). In this graphic, each bar represents a cluster. Bar sizes describe the proportion of positives compounds within each cluster. Ideally, clusters must represent only positive or negative compounds. When many bars show intermediate values the metric is unable to cluster together compounds with the same biological properties. Regarding the number of clusters (bars), we can see the same influence of the metrics on the final number of clusters obtained.

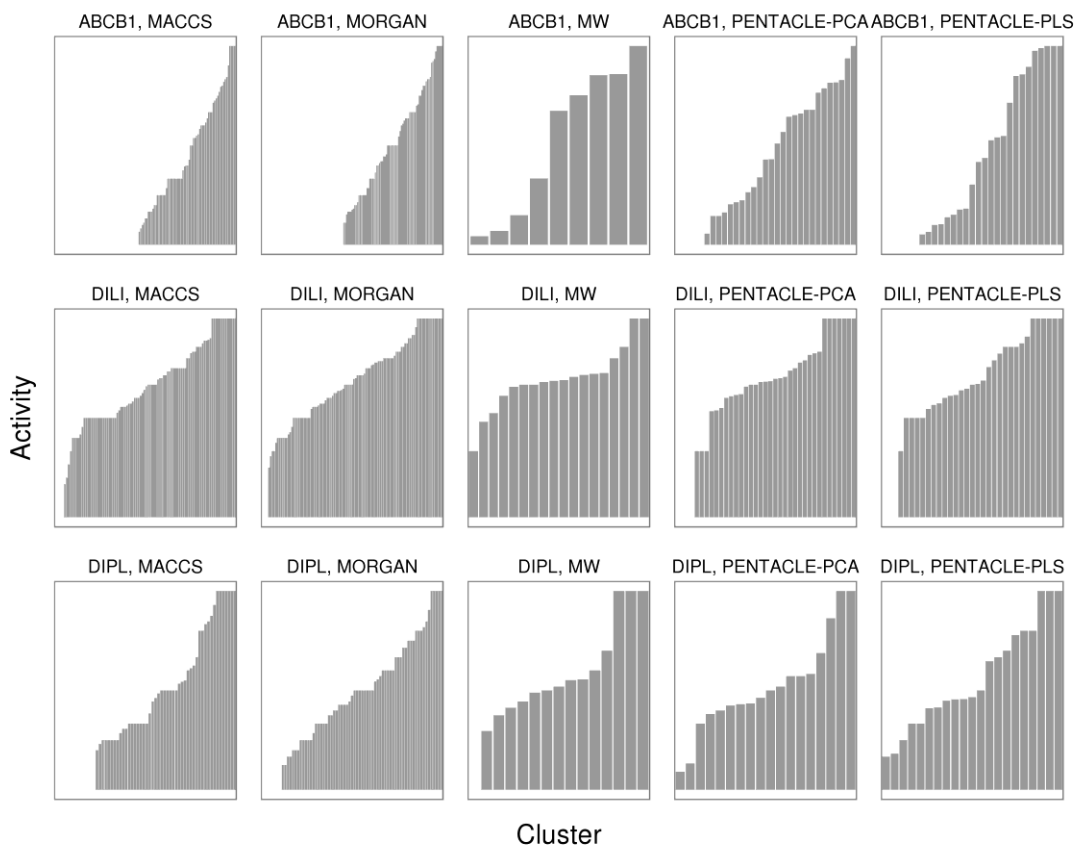


Fig. 4 Proportion of positive compounds in the chemical domains obtained using diverse similarity metrics, for datasets ABCB1, DIPL and DILI. The chemical domains were ordered according to their proportion of positive, from left to right. See text for details.

The metric comparison shows how fingerprint metrics (MACCS and MORGAN), producing the smaller clusters, are able to identify many chemical classes containing only negative compounds (wide empty areas on the left of the graphics). Conversely, they are not so good for obtaining clusters containing only positive compounds. MW produces fewer clusters and separates rather well positive from negative for some endpoints (ABCB1) and nearly not at all in some others (DILI). Pentacle metrics

produce worse results than fingerprints, particularly for clustering negative compounds but is slightly better than fingerprints for clustering the positives. As observed for quantitative endpoints, the activity biased variant (Pentacle-PLS) outperforms the non activity biased one (Pentacle-PCA).

The comparison of the endpoints clearly shows DILI as the most difficult endpoint. All metrics produce some clusters of positive compounds, but only Pentacle generates clusters of negatives. In some cases (MW) the graphic is mostly flat, indicating nearly random proportions of positive-negatives. At the opposite side is ABCB1 for which most metrics can cluster inactive compounds. Even a metric as simple as MW produces clusters enriched in negative and positive compounds corresponding respectively to small and large molecules, as can be expected by the geometry and function of this transporter (Broccatelli et al. 2011). DIPL represents an intermediate situation. Fingerprint metrics and particularly MACCS, produce many clusters of negative compounds and some of positive compounds. MW and Pentacle perform poorly in DIPL, with nearly no cluster of negative and many clusters containing the same proportions of positive and negative.

All in all, the results of this analysis indicate that the bioisosterism principle holds remarkably well in some cases. Some of the chemical classes contain compounds with homogeneous biological properties for both for quantitative and qualitative endpoints. Unfortunately, this is not a constant behavior. There are differences that can be attributed to the adequacy of the metrics, the complexity of the endpoint but also some random variability, even in the more stable situations. A general trend observed is that the more complex or difficult the endpoint is, the more difficult is to obtain homogeneous chemical domains (e.g. compare MACCS-hERG with MACCS-SOLU). Another general observation is that in nearly no case we can expect to obtain chemical domains with the same level of biological homogeneity for the entire activity spectrum. Some metric-endpoint combinations define better the domains of positive/active compounds (e.g. Pentacle-PLS-ABCB1), others favor the negative (e.g. MACCS-ABCB1) and in other cases there are high random variations (e.g. MACCS-hERG). Finally, as a general trend, the activity biased version of the Pentacle metric shows a clear advantage over the non activity biased one and for this reason the later will be removed from the analysis from now on.

Models performance

The predictive quality of the models obtained using RA and global QSAR models was estimated as described in the Methods section. The results are summarized in Table 4. For quantitative endpoints, a lower SDEP describes a better model while, for the qualitative endpoints (ABCB1, DILI and DIPL) higher Accuracy values describe better models.

Table 4. Quality metrics of the predictions obtained by read across (RA), structural alerts (SA) and global models.

		SDEP (lower is better)		Accuracy (higher is better)		
		hERG	SOLU	ABCB1	DILI	DIPL
Read across	fix nn	1.33	2.67	0.82	0.64	0.63
	fix th	0.57 (32%)	1.56 (21%)	0.95 (39%)	na ²	0.75 (4%)
	MS	0.77 (43%)	0.94 (75%)	0.87 (52%)	0.77 (23%)	0.64 (21%)
	SA	na ¹	na ¹	na ¹	0.62	0.36
Global models	PLS	1.00	1.02	0.81	0.53	0.61
	SVM	1.13	1.39	0.62	0.67	0.56
	RF	0.91	0.92	0.85	0.66	0.64

¹ Structural alerts not available. ² Too few compounds within the classes. The best method for each dataset is highlighted in bold

Read Across and Structural Alerts

Before comparing the results, it must be stressed that the conditions imposed to define chemical categories by the RA-MS and RA-fix th do not guarantee obtaining predictions for all compounds. The values shown in parenthesis in Table 4 for these methods indicate the percentage of compounds for which a prediction was obtained. Focusing only on the results obtained for these compounds, Table 4 shows that the quality of the predictions given by some RA methods is impressive and yields the best results for all datasets with the only exception of SOLU. This is particularly true for RA-fix th, even if the percentage of results is the lowest. RA-MS also produced good results for all datasets. RA-fix nn has the advantage of producing predictions for all compounds, but the prediction quality was not brilliant and in some cases (e.g. hERG, SOLU) they were surprisingly poor.

The performance of structural alerts is not high. However, they were obtained after human analysis of the datasets (Hewitt et al. 2013; Przybylak et al. 2014) and the value of the mechanistic insight gained in this exercise is difficult to reflect in a quantitative index.

Global Models

The results shown in Table 4 indicate that global models perform rather well and in some cases their performance are comparable with those obtained using RA methods. In addition, they have the advantage of providing a prediction for any query compound, unlike the RA approaches. The best results were obtained with RF, with the only exception of DILI and DIPL but in either case the differences were very small. It must be noted that the global models obtained with RF use indeed a local model strategy, since the results produced by this method can be seen as a weighted sum of a subset of activities, where the weights are defined according to the similarity to the query compound (Lin and Jeon 2006). The only drawback of this method is the need of adjusting internal parameters (see the Methods Section). The second best method was PLS, showing predictive quality indexes nearly similar to RF's in all endpoints, with the only exception of DILI. In fact all DILI models have poor performance, as can be expected after the analysis of the previous section.

Local Models based on Metrics

Local models were built as described in the Methods section, for all the datasets. The analysis of the results is rather complicated, due to the use of four metrics and four threshold values. Note that threshold values have a large effect on the definition of the chemical domains since smaller values produced smaller domains, as can be seen in Figure 2.

A preliminary analysis based on the comparison of the average quality metrics obtained after the ten random splitting did not show relevant differences. A deeper analysis showed that the training-test splitting has a significant effect on the prediction results (some splits are more difficult to predict than the others) and that more information could be obtained if we compare global (PLS global) and local models for each individual split. The results of this analysis are represented in Figure 5 (for quantitative endpoints) and 6 (for qualitative endpoints). The performance of each modeling condition for every split is shown as a grey dot, while the global model is shown for comparison as a black dot on the right hand side. When local models perform better (in average) than the global model the graphic shows lines linking the best local model with the global model. In Figure 5, lines pointing upwards (SDEP lower for local than for global models) indicate that the local models performed better and vice versa. Conversely, in Figure 6 the lines pointing downwards (accuracy higher for local than for global models) indicate that the local models performed better.

For quantitative endpoints we only observed improvements in the case of hERG. For the SOLU endpoint all local models had lower predictive ability than the global model. This can be justified by the homogeneity of the mechanism: if the compound properties affecting their solubility are similar for all compounds, there is no advantage in building separate models for diverse chemical classes. Indeed, in the analysis of metric performance for this dataset the intra-cluster dispersion was relatively large, with few clusters of highly similar compounds. For the hERG endpoints, local models obtained with MACCS-0.05, Morgan-0.2 and MW-0.4 yielded better average results than the global model. Such results are particularly significant for MW. These results can be justified by the nature of the hERG receptor, with a large binding pocket that could accommodate ligands of very different size (Thai et al. 2010b). Therefore, local models built for compounds of similar MW are likely to be more stable and predictive than the global models.

For qualitative endpoints, the best local models were obtained using fingerprint based metrics, even if the improvement with respect to global models were not large. For ABCB1, local models using MACCS and MORGAN fingerprints with any threshold produce better results than global models, but the best results were obtained for the threshold value producing smallest domains (0.05). MW and Pentacle metrics produce local models performing worse than the global one for all threshold values. A rather similar situation is observed for DIPL, and for some splits we observe a slight improvement for local models using MACCS-0.1 and MORGAN-0.1 over the global models. No improvement was obtained with MW and Pentacle metrics. DILI is a particular case. We obtained better results for local models using fingerprint metrics with low thresholds, but in this dataset the improvements obtained in some splits were large, (see Figure 6) obtaining differences in accuracy of nearly 0.2 units (0.51 for MORGAN 0.05 vs 0.31 Global). Also, only in the dataset we obtained local models based on non-fingerprint metrics that outperform the global models, in particular Pentacle-0.01.

Summarizing, the local models obtained with low (0.05-0.2) thresholds and fingerprint metrics have slightly better overall predictive performance than global models, but the effect depends on the dataset. It can be negligible for models describing rather general properties like the water solubility or rather significant for complex properties (e.g. DILI). Moreover, the effect of local model can also depend on the split and for the same dataset we observed either an increase or a decrease of the

predictive performance. Some metrics can produce good results when the chemical classes discriminate between compounds which use different mechanisms, as it was the case of MW metric for hERG, where low and high MW compounds bind different regions of the binding pocket.

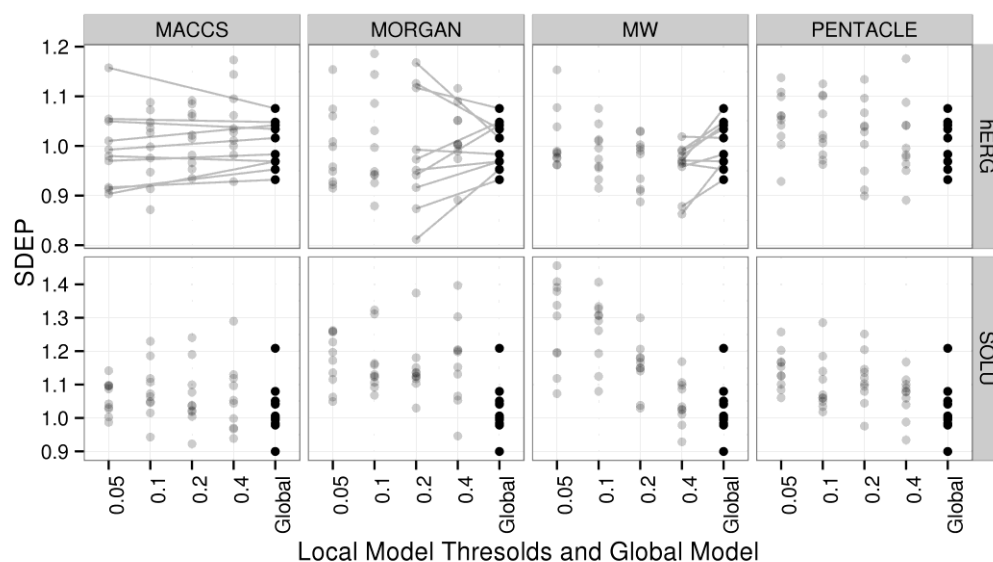


Fig. 5 Comparison of SDEP obtained with local models using diverse similarity thresholds and global models. For every condition the figure represents the values obtained for the ten random training-test set splits. When the average SDEP of the local models is lower than the obtained for the global model, the SDEP values for every split were linked by a line. Pointing up lines indicates that the local model were better than the global model and vice versa.

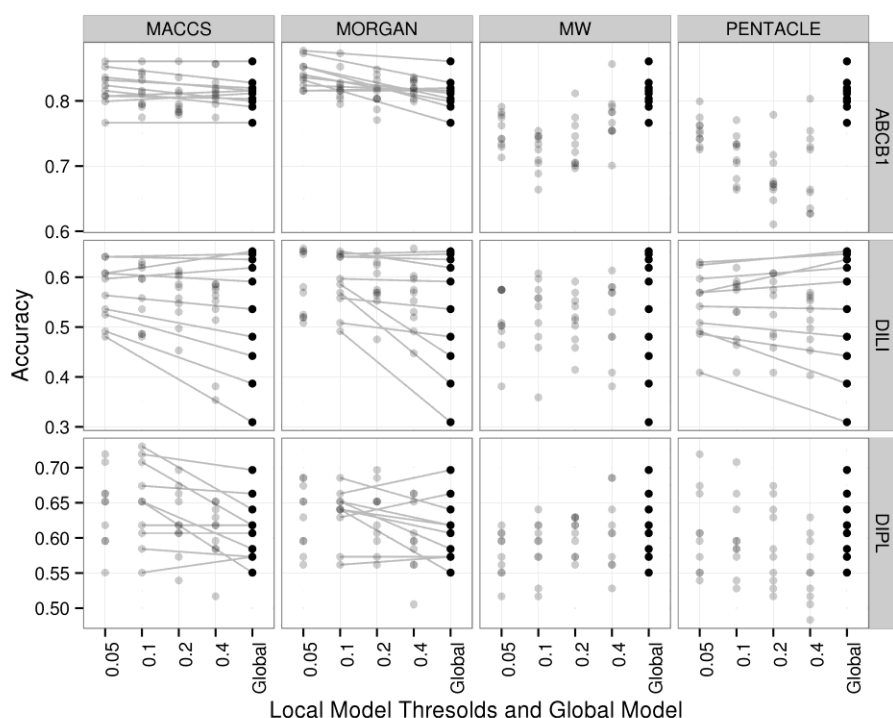


Fig. 6 Comparison of the accuracy of local models using diverse similarity thresholds and PLS global models. For every condition the figure represents the values obtained for the ten random training-test set splits. When the average accuracy of the local models is lower than those obtained for the global model, the accuracy values for every split were linked by a line. Pointing down lines indicate that the local model were better than the global model and vice versa (the opposite than in Figure 5).

Strategy

The results of this study, reported in the previous section clearly show the difficulties of selecting the best modeling method for any given dataset. Some method predict very well, but only for a small percentage of the compounds, and not in all datasets. The analysis of the molecular similarity metrics (Figure 3 and 4) shows the different performance of the methods but also the variation observed within a dataset for different ranges of the biological property. The presence of this variability has been further confirmed by the differences observed in the models predictive performance for different training-test set splits. Referring back to the Figure 1, it looks like the situations depicted there can co-exist within the same dataset for different regions of the chemical space and their relative weight can be affected by the addition or removal of a few compounds.

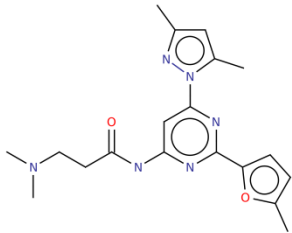
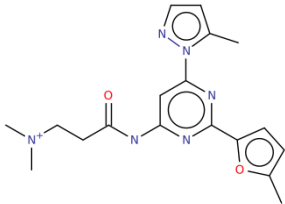
This observation led us to propose an alternative strategy, based on the selection of the best modeling method for every query compound. We consider this a "unifying strategy" because we assume that RA, local and global QSAR models are equally valid approaches and the choice of one or the other can be made on an individual, per compound basis. The rationale of this strategy is that the information content of the chemical space is not homogeneous. When the query compound is within an information-rich region, due to the presence of structurally similar compounds of known properties, this information can be exploited to our advantage. The workflow in Figure 7 illustrates the proposed strategy: first, the query compound is evaluated using a read across method, which looks for the presence of highly similar compounds. If these are found, the prediction will be based on their properties and no further steps are done. Since the similarity criterion is rather strict, it can be expected that in many cases the RA will yield no results. In these cases, the query compound is assigned to a chemical domain of the training set, using the similarity metrics described above, and predicted using a local model built within. When no suitable chemical domain can be found, a global model based on random forest was used as a fallback.

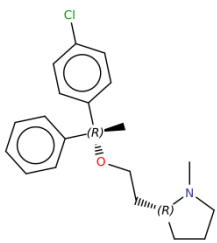
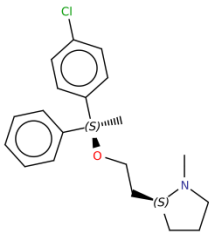
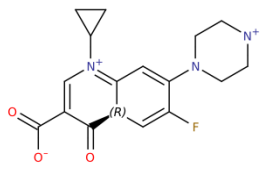
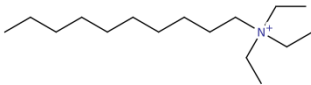
This strategy was applied to the five data sets of Table 1-2 and the quality of the predictions obtained was analyzed and compared with those obtained using global models. For all studied endpoints the average results show a clear but modest quality improvement. However we wanted to investigate more in depth if the improvement was observed in all instances or only in some cases. Figure 8 illustrates the results for each dataset split, quantified as SDEP (for quantitative endpoints) or accuracy (for qualitative endpoints) and compared with the global model prediction. For quantitative endpoints (hERG and SOLU), the advantages were clear: in hERG it always produced better results. In SOLU, only one of the splits out of ten produced worse results and another was not significantly improved. For the qualitative endpoints, the results were even better. For ABCB1 and DILI the strategy produced very relevant improvements in 9 out of 10 splits and in the remaining case, the differences were very small. In the most favorable case, the improvement means a jump from 0.31 to 0.70. For the DIPL endpoint the strategy did not produce so consistent improvements and failed in 3 out of 10 cases.

In order to further illustrate the obtained results we show in Table 5 three compounds of the hERG dataset, selected to represent extreme situations. For all compounds, we show the 2D structure of

the query compound, the structure of the most similar compound and the predictions obtained using the different methods described in the present paper. The prediction selected by our proposed strategy is highlighted in bold and in all instances it was coincident with the best option. For the first compound, the RA prediction matched exactly the experimental value, due to the presence in the dataset of a very similar structure. Local and global models also yield reasonable predictions with differences of 0.08 and 0.25 log units vs. the experimental value, respectively. For the second compounds, the presence of a diastereoisomer with similar biological properties also produced a perfect match when using RA. Conversely, the predictions produced by the QSAR models fail by more than 2 log units. For the third structure, no similar structure was found and hence the RA method produced no result, even though the structure and biological property of the closer compound are shown as a reference. In this case, the global models produced the best result and the predicted value fall within 0.69 log units of the experimental value. All in all, the compounds represented in Table 5 illustrate well how no single method produces the best results in all instances and how large the errors could be if we do so.

Table 5: Predictions obtained with RA, LM and GM for a few representatives of the hERG data set. The table contains the activity values of the closest compound if the RA and LM do not provide predictions. The predictions selected by the proposed strategy are highlighted in bold.

2D structure		Predictions			Experimental
Query compound	RA/Closest compound	RA	LM	GM	
		6.19	6.27	5.94	6.19

		8	5.4	5.71	8
		5.44*	4.1	3.71	3.02

(* RA provided no prediction. The activity of the closest compound measured with Tanimoto and MORGAN fingerprints is shown)

The same ability to select the best method, illustrated for diverse compounds of the same series in the above table, can also be observed across diverse datasets. The proposed strategy can be seen as a universal tool, applicable in highly diverse datasets, making optimal use of their heterogeneous information content. The chart at the bottom left corner of Figure 8 shows the proportion of prediction carried out using the different modeling methods (RA, local and global QSAR) for the different datasets, further illustrating the large diversity of the datasets and the advantages of using a flexible strategy, over more rigid approaches.

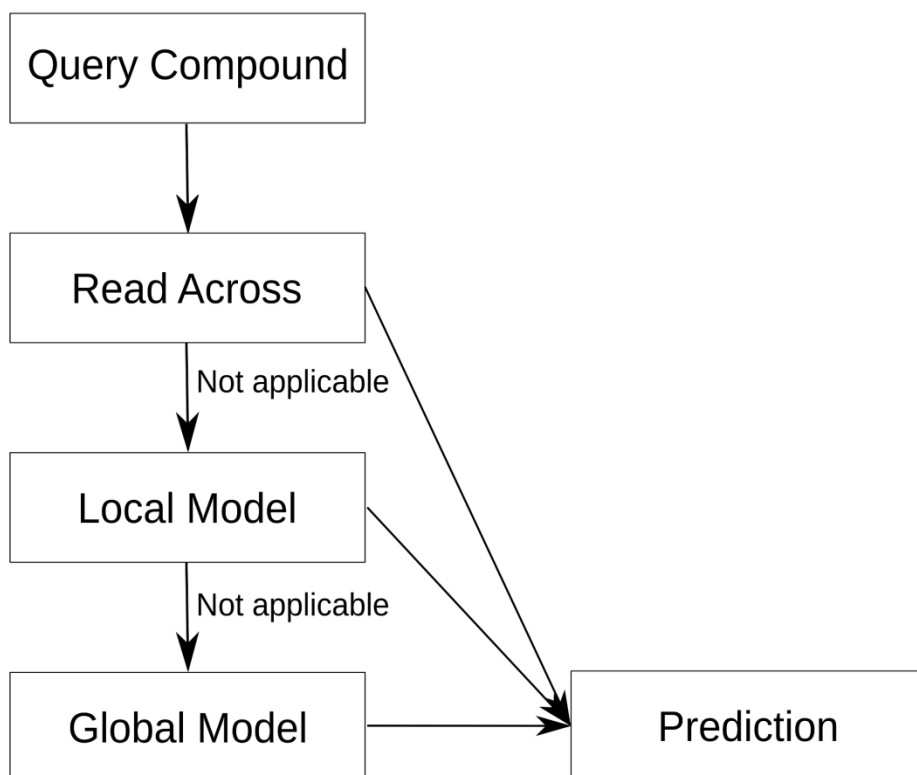


Fig. 7 Workflow of the proposed unifying strategy for the prediction

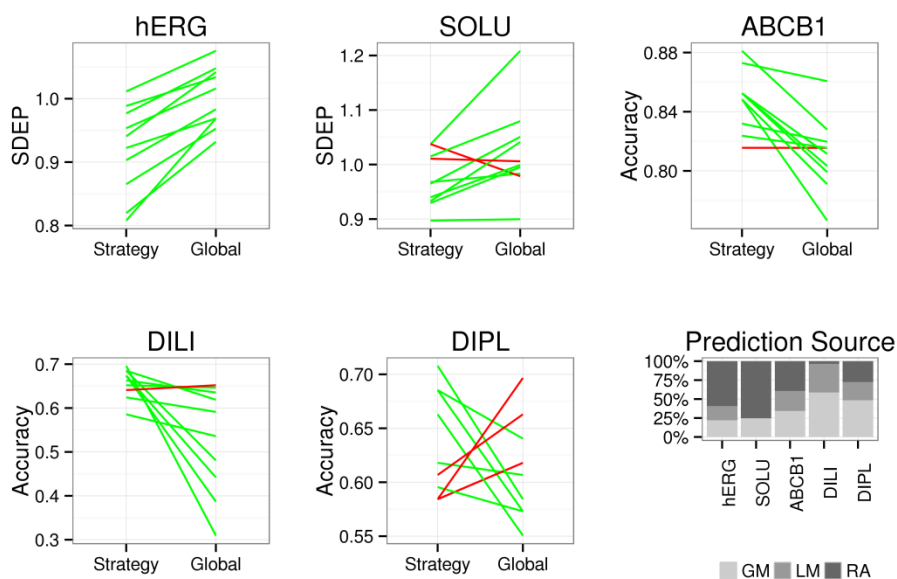


Fig. 8 Comparison of the predictive quality of the proposed strategy (see text) and of global models quantified as SDEP for quantitative endpoints (hERG and SOLU) and accuracy (ABCB1, DIPL and DILI). For every dataset we represent the results obtained for ten random training-test set splits. Green

lines indicate that the results obtained using the proposed strategy were better than the results obtained using the global model. Red lines indicate the opposite. The column chart at the bottom right corner represents the average number of predictions obtained by the different modeling methods, according to the proposed strategy, in each of the datasets.

Conclusion

The large diversity of endpoints studied in drug safety assessment makes difficult to develop a "fit for all" modeling solution. This idea was confirmed by the disparity of the predictive performance results obtained for the sample of datasets studied here. Beyond predictive quality, the choice of the best predictive method will be often imposed by practical considerations, like the lack of compounds for obtaining a truly global model or the need of predicting the properties for which no suitable chemical category can be assigned.

For these reasons, we proposed here a unifying strategy where the use of a whole spectrum of methods is considered; from structural alerts to global QSAR models, including read across and local models, in which the final decision is taken in a per-compound basis. The advantages of using such approach were illustrated above for a sample of representative datasets. However, it must be emphasized that its application in real-world situations can have additional benefits. First, in pharmaceutical research it is very common to generate clusters of closely related structures produced by chemical derivatization of candidate compounds. These situations are ideal for the application of RA or local models and can lead to much better predictions than approaches based on global models. Also, the chemical space of interest should not be seen as something static. During the life of a predictive model, new compounds can be studied and incorporated to the model training series thus changing their characteristics. The use of a flexible strategy allows adopting the most suitable method after any enrichment of the training series, thus making optimal use of all available information.

As indicated in the title, we do not claim to have the best possible solution and we consider our proposed strategy as a first step towards a new generation of predictive methodologies, better adapted to the characteristics of the structure-activity landscape where they need to operate. The examples presented here can therefore be seen as a proof of concept, the value of which needs to be confirmed by the application of this methods to a richer dataset collection. The strategy described in

the present work and illustrated in Figure 8 has been implemented as a semi-automatic procedure based on command mode scripts, which are distributed as Supplementary Information under Open Source (GNU GPL-3) license of use. Currently our group is working in the development of automatic software tools allowing to apply unifying strategies like the ones presented here without human intervention and integrate them in toxicity predictions tools like eTOXlab (Carrió et al. 2015; Sanz et al. 2015).

Acknowledgments

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking, under Grant Agreement No. 115002 (eTOX), resources of which are composed of a financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contributions.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Accelrys MACCS Structural Keys.

Alelyunas YW, Empfield JR, McCarthy D, et al. (2010) Experimental solubility profiling of marketed CNS drugs, exploring solubility limit of CNS discovery candidate. *Bioorganic Med Chem Lett* 20:7312–7316. doi: 10.1016/j.bmcl.2010.10.068

Aller SG, Yu J, Ward A, et al. (2009) Structure of P-Glycoprotein Reveals a Molecular Basis for Poly-Specific Drug Binding. *Science* 323:1718–1722. doi: 10.1126/science.1168750

Andersson PL, Maran U, Fara D, et al. (2002) General and class specific models for prediction of soil sorption using various physicochemical descriptors. *J Chem Inf Comput Sci* 42:1450–9.

Aronov AM (2008) Tuning out of hERG. *Curr Opin Drug Discov Devel* 11:128–40.

Bajorath J (2014) Exploring Activity Cliffs from a Chemoinformatics Perspective. *Mol Inform* 33:438–442. doi: 10.1002/minf.201400026

Bajorath J (2012) Computational chemistry in pharmaceutical research: At the crossroads. *J Comput Aided Mol Des* 26:11–12. doi: 10.1007/s10822-011-9488-z

- Bajorath J, Peltason L, Wawer M, et al. (2009) Navigating structure-activity landscapes. *Drug Discov Today* 14:698–705.
- Bemis GW, Murcko M a. (1996) The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 39:2887–2893. doi: 10.1021/jm9602928
- Benet LZ (2009) The Drug Transporter–Metabolism Alliance: Uncovering and Defining the Interplay. *Mol Pharm* 6:1631–1643. doi: 10.1021/mp900253n
- Borst P, Elferink RO (2002) Mammalian ABC transporters in health and disease. *Annu Rev Biochem* 71:537–592. doi: 10.1146/annurev.biochem.71.102301.093055
- Breiman L (2001) Random Forests. *Mach Learn* 45:5–32. doi: 10.1186/1478-7954-9-29
- Broccatelli F, Carosati E, Cruciani G, Oprea TI (2010) Transporter-mediated efflux influences CNS side effects: ABCB1, from antitarget to target. *Mol Inform* 29:16–26. doi: 10.1002/minf.200900075
- Broccatelli F, Carosati E, Neri A, et al. (2011) A Novel Approach for Predicting P-Glycoprotein (ABCB1) Inhibition Using Molecular Interaction Fields. *J Med Chem* 54:1740–1751. doi: 10.1021/jm101421d
- Broccatelli F, Mannhold R, Moriconi A, et al. (2012) QSAR modeling and data mining link torsades de pointes risk to the interplay of extent of metabolism, active transport, and hERG liability. *Mol Pharm* 9:2290–2301.
- Carrió P, López O, Sanz F, Pastor M (2015) eTOXlab, an open source modeling framework for implementing predictive models in production environments. *J Cheminform.* doi: 10.1186/s13321-015-0058-6
- Cherkasov A, Muratov EN, Fourches D, et al. (2014) Perspective QSAR Modeling : Where have you been ? Where are you going to ? QSAR Modeling : Where have you been ? Where are you going to ? *J. Med. Chem.*
- Choudhuri S, Klaassen CD (2006) Structure, function, expression, genomic organization, and single nucleotide polymorphisms of human ABCB1 (MDR1), ABCC (MRP), and ABCG2 (BCRP) efflux transporters. *Int J Toxicol* 25:231–59.
- Cortes C, Vapnik V (1995) Support-Vector Networks. *Mach Learn* 20:273–297.
- Curigliano G, Mayer EL, Burstein HJ, et al. (2010) Cardiac toxicity from systemic cancer therapy: a comprehensive review. *Prog Cardiovasc Dis* 53:94–104.
- Delaney JS (2004) ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J Chem Inf Comput Sci* 44:1000–1005. doi: 10.1021/ci034243x

- Dimova D, Bajorath J (2014) Extraction of SAR information from activity cliff clusters via matching molecular series. *Eur J Med Chem* 87:454–460. doi: 10.1016/j.ejmech.2014.09.087
- Durán Á, Martínez GC, Pastor M (2008) Development and validation of AMANDA, a new algorithm for selecting highly relevant regions in molecular interaction fields. *J Chem Inf Model* 48:1813–1823. doi: 10.1021/ci800037t
- Durán Á, Pastor M (2010) Pentacle. <http://www.moldiscovery.com/software/pentacle>.
- EC (2015) REACH. European Community Regulation on chemicals and their safe use. http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm.
- Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today* 12:225–33. doi: 10.1016/j.drudis.2007.01.011
- Ekins S (2014) Progress in computational toxicology. *J Pharmacol Toxicol Methods* 69:115–140. doi: 10.1016/j.vascn.2013.12.003
- Enoch SJ, Cronin MTD, Madden JC, Hewitt M (2009) Formation of structural categories to allow for read-across for teratogenicity. *QSAR Comb Sci* 28:696–708. doi: 10.1002/qsar.200960011
- FDA (2005) Guidance for Industry Starting Dose in Initial Clinical Trials Guidance for Industry Estimating the Maximum Safe. FDA 27. doi: 10.1089/blr.2006.25.697
- Fourches D, Barnes JC, Day NC, et al. (2010) Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species. *Chem Res Toxicol* 23:171–183. doi: 10.1021/tx900326k
- Fung M, Thornton A, Mybeck K, et al. (2001) Evaluation of the Characteristics of Safety Withdrawal of Prescription Drugs from Worldwide Pharmaceutical Markets-1960 to 1999. *Drug Inf J* 35:293–317. doi: 10.1177/009286150103500134
- Golbraikh A, Muratov E, Fourches D, Tropsha A (2014) Data Set Modelability by QSAR. *J Chem Inf Model* 54:1–4. doi: 10.1021/ci400572x
- Guha R (2012) Exploring uncharted territories: predicting activity cliffs in structure-activity landscapes. *J Chem Inf Model* 52:2181–91. doi: 10.1021/ci300047k
- Guha R, Dutta D, Jurs PC, Chen T (2006) Local lazy regression: making use of the neighborhood to improve QSAR predictions. *J Chem Inf Model* 46:1836–47. doi: 10.1021/ci060064e
- Hancox JC, McPate MJ, El Harchi A, Zhang Y hong (2008) The hERG potassium channel and hERG screening for drug-induced torsades de pointes. *Pharmacol Ther* 119:118–132. doi: 10.1016/j.pharmthera.2008.05.009

- Helgee EA, Carlsson L, Boyer S, Norinder U (2010) Evaluation of quantitative structure-activity relationship modeling strategies: Local and global models. *J Chem Inf Model* 50:677–689. doi: 10.1021/ci900471e
- Hewitt M, Enoch SJ, Madden JC, et al. (2013) Hepatotoxicity: a scheme for generating chemical categories for read-across, structural alerts and insights into mechanism(s) of action. *Crit Rev Toxicol* 43:537–58. doi: 10.3109/10408444.2013.811215
- Hua Y, Yongyan W, Yiyu C (2007) Local and global quantitative structure-activity relationship modeling and prediction for the baseline toxicity. *J Chem Inf Model* 47:159–169. doi: 10.1021/ci600299j
- Juliano RL, Ling V (1976) A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants. *Biochim Biophys Acta* 455:152–162. doi: 10.1016/0005-2736(76)90160-7
- Klepsch F, Ecker GF (2010) Impact of the Recent Mouse P-Glycoprotein Structure for Structure-Based Ligand Design. *Mol Inform* 29:276–286. doi: 10.1002/minf.201000017
- Könemann H (1980) Structure-activity relationships and additivity in fish toxicities of environmental pollutants. *Ecotoxicol Environ Saf* 4:415–421. doi: 10.1016/0147-6513(80)90043-3
- Könemann H, Musch A (1981) Quantitative structure-activity relationships in fish toxicity studies Part 2: The influence of pH on the QSAR of chlorophenols. *Toxicology* 19:223–228. doi: 10.1016/0300-483X(81)90131-1
- Kramer NI, Di Consiglio E, Blaauboer BJ, Testai E (2015) Biokinetics in Repeated-Dosing In Vitro Drug Toxicity Studies. *Toxicol. Vitr.*
- Kruhlak NL, Choi SS, Contrera JF, et al. (2008) Development of a phospholipidosis database and predictive quantitative structure-activity relationship (QSAR) models. *Toxicol Mech Methods* 18:217–227. doi: 10.1080/15376510701857262
- Kubinyi H (1998) Similarity and Dissimilarity: A Medicinal Chemist's View. 225–252.
- Landrum G RDKit: Open-source cheminformatics. <http://www.rdkit.org>.
- Leise MD, Poterucha JJ, Talwalkar JA (2014) Drug-induced liver injury. *Mayo Clin Proc* 89:95–106.
- Li Q, Jørgensen FS, Oprea T, et al. (2008) hERG Classification Model Based on a Combination of Support Vector Machine Method and GRIND Descriptors. *Mol Pharm* 5:117–127. doi: 10.1021/mp700124e
- Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News* 2:18–22.
- Liebler DC, Guengerich FP (2005) Elucidating mechanisms of drug-induced toxicity. *Nat Rev Drug Discov* 4:410–420. doi: 10.1038/nrd1720

- Lin Y, Jeon Y (2006) Random Forests and Adaptive Nearest Neighbors. *J Am Stat Assoc* 101:578–590. doi: 10.1198/016214505000001230
- Loo TW, Clarke DM (2002) Location of the rhodamine-binding site in the human multidrug resistance P-glycoprotein. *J Biol Chem* 277:44332–44338. doi: 10.1074/jbc.M208433200
- Maggiora G, Vogt M, Stumpfe D, Bajorath J (2014) Molecular similarity in medicinal chemistry. *J Med Chem* 57:3186–3204. doi: 10.1021/jm401411z
- Maggiora GM (2006) On outliers and activity cliffs - Why QSAR often disappoints. *J Chem Inf Model* 46:1535. doi: 10.1021/ci060117s
- Martens H (2001) Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression. *Chemom Intell Lab Syst* 58:85–95. doi: 10.1016/S0169-7439(01)00153-8
- Martin YC (1981) A practitioner's perspective of the role of quantitative structure-activity analysis in medicinal chemistry. *J Med Chem* 24:229–237. doi: 10.1021/jm00135a001
- Martin YC, Kofron JL, Traphagen LM (2002) Do structurally similar molecules have similar biological activity? *J Med Chem* 45:4350–8.
- Medina-Franco JL (2012) Scanning Structure – Activity Relationships with Structure – Activity Similarity and Related Maps : From Consensus Activity Cliffs to Selectivity Switches.
- Medina-Franco JL (2013) Activity cliffs: Facts or artifacts? *Chem Biol Drug Des* 81:553–556. doi: 10.1111/cbdd.12115
- Mevik B-H, Wehrens R (2007) The pls Package: Principal Component and Partial Least Squares Regression in R. *J Stat Softw* 18:1–24.
- Meyer D, Dimitriadou E, Hornik K, et al. (2014) e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.
- Milletti F, Storchi L, Sforza G, et al. (2009) Tautomer enumeration and stability prediction for virtual screening on large chemical databases. *J Chem Inf Model* 49:68–75. doi: 10.1021/ci800340j
- Milletti F, Storchi L, Sforza G, Cruciani G (2007) New and original pKa prediction method using grid molecular interaction fields. *J Chem Inf Model* 47:2172–81. doi: 10.1021/ci700018y
- Morgan HL (1965) The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J Chem Doc* 5:107–113.
- Muller PY, Milton MN (2012) Index in Drug Development. *Nat Rev Drug Discov* 11:751–761. doi: 10.1038/nrd3801

- Muster W, Breidenbach A, Fischer H, et al. (2008) Computational toxicology in drug development. *Drug Discov Today* 13:303–310. doi: 10.1016/j.drudis.2007.12.007
- NRC (2007) Toxicity Testing in the 21st Century: A Vision and a Strategy. The National Academies Press
- Nikolova N, Jaworska J (2003) Approaches to Measure Chemical Similarity– a Review. *QSAR Comb Sci* 22:1006–1026. doi: 10.1002/qsar.200330831
- Obiol-Pardo C, Gomis-Tena J, Sanz F, et al. (2011) A multiscale simulation system for the prediction of drug-induced cardiotoxicity. *J Chem Inf Model* 51:483–492. doi: 10.1021/ci100423z
- Orogo AM, Choi SS, Minnier BL, Kruhlak NL (2012) Construction and consensus performance of (Q)SAR models for predicting phospholipidosis using a dataset of 743 compounds. *Mol Inform* 31:725–739. doi: 10.1002/minf.201200048
- Park YC, Cho MH (2011) A new way in deciding NOAEL based on the findings from GLP-toxicity test. *Toxicol Res* 27:133–135. doi: 10.5487/TR.2011.27.3.133
- Pastor M (2006) Alignment-independent Descriptors from Molecular Interaction Fields. In: Cruciani G (ed) *Mol. Interact. Fields. Appl. Drug Discov. ADME Predict.* Wiley-VCH, pp 117–141
- Pastor M, Cruciani G, McLay I, et al. (2000) GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J Med Chem* 43:3233–3243. doi: 10.1021/jm000941m
- Perkins R, Fang H, Tong W, Welsh WJ (2003) Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. *Environ Toxicol Chem* 22:1666–79.
- Przybylak KR, Alzahrani AR, Cronin MTD (2014) How Does the Quality of Phospholipidosis Data Influence the Predictivity of Structural Alerts? *J Chem Inf Model*. doi: 10.1021/ci500233k
- Raunio H (2011) In silico toxicology - non-testing methods. *Front Pharmacol* 2:33. doi: 10.3389/fphar.2011.00033
- Reasor MJ, Hastings KL, Ulrich RG (2006) Drug-induced phospholipidosis: issues and future directions. *Expert Opin Drug Saf* 5:567–583. doi: 10.1517/14740338.5.4.567
- Recanatini M, Cavalli A, Masetti M (2008) Modeling HERG and its interactions with drugs: recent advances in light of current potassium channel simulations. *ChemMedChem* 3:523–35. doi: 10.1002/cmdc.200700264
- Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *JChelInformModel* 50:742–54. doi: 10.1021/ci100050t

- Roy K, Mitra I, Kar S, et al. (2012) Comparative studies on some metrics for external validation of QSPR models. *J Chem Inf Model* 52:396–408. doi: 10.1021/ci200520g
- Sadowski J, Gasteiger J (1993) From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem Rev* 93:2567–2581. doi: 10.1021/cr00023a012
- Sadowski J, Gasteiger J, Klebe G (1994) Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J Chem Inf Model* 34:1000–1008. doi: 10.1021/ci00020a039
- Sanz F, Carrió P, López O, et al. (2015) Integrative Modeling Strategies for Predicting Drug Toxicities at the eTOX Project. *Mol Inform* 34:477–484. doi: 10.1002/minf.201400193
- Sawada H, Takami K, Asahi S (2005) A toxicogenomic approach to drug-induced phospholipidosis: Analysis of its induction mechanism and establishment of a novel in vitro screening system. *Toxicol Sci* 83:282–292. doi: 10.1093/toxsci/kfh264
- Schultz TW, Amcoff P, Berggren E, et al. (2015) A strategy for structuring and reporting a read-across prediction of toxicity. *Regul Toxicol Pharmacol* 72:586–601. doi: 10.1016/j.yrtph.2015.05.016
- Sheridan RP (2014) Global Quantitative Structure-Activity Relationship Models vs Selected Local Models as Predictors of Off-Target Activities for Project Compounds. *J Chem Inf Model* 54:1083–92. doi: 10.1021/ci500084w
- Szakács G, Paterson JK, Ludwig JA, et al. (2006) Targeting multidrug resistance in cancer. *Nat Rev Drug Discov* 5:219–234. doi: 10.1038/nrd1984
- Thai K-M, Windisch A, Stork D, et al. (2010a) The hERG potassium channel and drug trapping: insight from docking studies with propafenone derivatives. *ChemMedChem* 5:436–42. doi: 10.1002/cmdc.200900374
- Thai K-M, Windisch A, Stork D, et al. (2010b) The hERG potassium channel and drug trapping: insight from docking studies with propafenone derivatives. *ChemMedChem* 5:436–42.
- Treinen-Moslen M, Kanz MF (2006) Intestinal tract injury by drugs: Importance of metabolite delivery by yellow bile road. *Pharmacol Ther* 112:649–67.
- Tropsha A (2010) Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol Inform* 29:476–488. doi: 10.1002/minf.201000061
- Vandenberg JJ, Perry MD, Perrin MJ, et al. (2012) hERG K⁺ Channels: Structure, Function, and Clinical Significance. *Physiol Rev* 92:1393–1478. doi: 10.1152/physrev.00036.2011
- Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*, Fourth. Springer, New York
- Willett P, Barnard JM, Downs GM (1998) Chemical Similarity Searching. *J Chem Inf Model* 38:983–996. doi: 10.1021/ci9800211

- Wilk-Zasadna I, Bernasconi C, Pelkonen O, Coecke S (2015) Biotransformation in vitro: An essential consideration in the quantitative in vitro-to-in vivo extrapolation (QIVIVE) of toxicity data. *Toxicology* 332:8–19. doi: 10.1016/j.tox.2014.10.006
- Yoon M, Blaauboer BJ, Clewell HJ (2015) Quantitative in vitro to in vivo extrapolation (QIVIVE): An essential element for in vitro-based risk assessment. *Toxicology* 332:1–3. doi: 10.1016/j.tox.2015.02.002