Chapter 20

The Use of In Silico Models Within a Large Pharmaceutical Company

Alessandro Brigo and Wolfgang Muster

Abstract

The present contribution describes how in silico models are applied at different stages of the drug discovery process in the pharmaceutical industry. A thorough description of the most relevant computational methods and tools is given along with an in-depth evaluation of their performance in the context of potential genotoxic impurities assessment.

The challenges of predicting the outcome of highly complex studies are discussed followed by considerations on how novel ways to manage, store, share and analyze data may advance knowledge and facilitate modeling efforts.

Key words Drug discovery, Genotoxicity, TTC, Lead optimization

1 Introduction

Computational methods (in silico models) are widely used in the pharmaceutical industry for optimizing molecules during early drug development, not only for efficacy, but in parallel with regard to their toxicological as well as drug disposition properties. It is the fine balance of target potency, selectivity, favorable ADME (absorption, distribution, metabolism, excretion), and (pre)clinical safety properties that will ultimately lead to the selection and clinical development of a potential new drug [1, 2]. As a clinical candidate needs rigorous preclinical optimization in various aspects, multidimensional optimization (MDO) is a term often used to describe the intensive investigations during the first 3–4 years of drug discovery from the identification of the target to the selection of the best drug development compound. The current MDO process comprises the use of in silico, in vitro, as well as in vivo techniques. In general, in silico tools have the intrinsic advantages to be fast and not to need the physical presence of the test compounds and can therefore be applied very early in drug development. Theoretically, in silico models can be developed for all end points

Emilio Benfenati (ed.), *In Silico Methods for Predicting Drug Toxicity*, Methods in Molecular Biology, vol. 1425,
DOI 10.1007/978-1-4939-3609-0_20,

and organisms, but the availability of large enough, balanced, and high-quality datasets is the main drawback for reliable predictions. An excellent correlation with the in vitro/in vivo data, that is, high-sensitivity as well as high-specificity, easy-to-use, and easy-to-interpret in silico model, is a key requirement for its usefulness. In the past few years, computational toxicology prediction systems tremendously increased their predictive power for end points like genotoxicity, carcinogenicity, phototoxicity, phospholipidosis, GSH adduct formation, hERG inhibition, and CYP inductions, but still have not achieved the major breakthrough due to lack of sufficiently large datasets covering more complex toxicological end points (e.g., liver-, kidney-, cardiotoxicity). These are the critical toxicity end points, which needs to be addressed in the next years to weed out potential safety issues in the clinics. Recent initiatives and consortia (e.g., IMI/eTOX, ToxCast, and ToxBank) dealing with data sharing of preclinical in vivo toxicology studies and computational approaches have the potential of significantly improving these end point predictions and filling the data gaps [3–5].

This review will outline general considerations on the mainly applied expert systems—rule-based and statistical-based models—in toxicology and ADME for pharmaceuticals and their application in the early drug development process as well as their regulatory impact on the assessment of potential impurities arising in the manufacturing process. Recent improvements and future perspectives on the main challenge of predicting complex in vivo end points will be summarized and discussed.

2 In Silico Methods for the Prediction of Toxicity

As already described in Subheading 1 of this chapter, the thorough characterization of the safety profile of drug candidates is of great importance to ensure that no harm is posed to healthy volunteers and patients during and after clinical development throughout the entire compound lifecycle.

Drug toxicity can manifest itself in a number of ways and may interest one or more target organs or biological processes. In particular, carcinogenicity and liver, renal, cardiovascular, reproductive, and genetic toxicities are among the most significant safety issues that can prevent drug candidates to progress through clinical development or can cause the withdrawal of already marketed products. Overall, between 20 and 30 % of failures can be attributed to safety reasons [6–8].

Over that past few years, predictive computational approaches have found a significant role within drug discovery in helping scientists rank compounds classes and prioritize in vitro and in vivo experiments. A number of factors contributed to the increased importance of in silico methods in drug discovery: (1) wider avail-

ability of high-quality datasets (public domain, focused data sharing initiatives), (2) robust computational models that can provide reliable predictions[9], (3) pressure to reduce animal testing, (4) need to bring new drugs to the market faster and cheaper, (5) legislation on the assessment of potential genotoxic impurities, and (6) greater number of commercially available and open-source software tools.

The most widely used computational methods for the prediction of toxicity end points can be roughly divided into two main categories, rule-based and statistical-based systems, depending on what type of methods they use to make their classifications.

2.1 Rule-Based Systems

Computational tools included in this category store and manipulate knowledge to interpret information. They are often referred to as expert systems, which make use of a set of explicit rules (i.e., not implicitly embedded in a code) to make deductions and classifications. Such systems have the advantage that rules can be easily represented and developed by experts in the field of toxicology (or of any discipline the systems are applied to), rather than by information technology (IT) specialists. In addition, solid expert rules can be derived from limited amounts of data, as long as they are sufficiently representative of specific chemical and biological spaces.

Both commercial and open-source systems are available within the rule-based methodologies, and they include, among others, Derek Nexus [10–13], Toxtree [14], CASE Ultra Expert Rules [15], and Leadscope Expert Alerts System [16].

Derek Nexus is an expert, knowledge base system which contains structural alerts (SAs) and expert knowledge rules (derived from both public and proprietary data by scientists at Lhasa Ltd.) for a wide range of toxicological end points and applies these to make in silico predictions about the toxicity of chemical entities. The knowledge-based expert rules represent knowledge from literature, academic, industrial, and Lhasa Ltd. scientific experts and are regularly updated according to newly available experimental data and publications. In making predictions, the expert rules take into account not only the presence or absence of a structural alert but also the species and a few calculated physicochemical parameters (where applicable) in a process akin to the human-based logic of argumentation. Proprietary data donated, by Lhasa Ltd. members, has been used in the development of approximately 25 % of the bacterial in vitro (Ames test) mutagenicity alerts in Derek Nexus, and proprietary datasets are used to validate the performance of alerts for this, and other end points, to provide an indication of predictive performance within the chemical space of highest interest to users. In addition proprietary and customized alerts can be defined by users and implemented through the Derek Knowledge Editor.

The most recent version of Derek Nexus contains expert-derived functionality to provide negative predictions for bacterial in vitro mutagenicity in order to give more confidence on nonpositive predictions. If a query compound does not match a structural alert for mutagenicity, then it is compared to a Lhasa reference set of Ames test data, and a negative prediction is provided based on the features within the query compound [17]. In case of absence of alerts for end points other than mutagenicity, negative calls should be made with caution as alerts that are not part of the rulebase, hence unknown to the system, can still be relevant in the induction of certain toxicities.

Since Derek is an expert system, it has no training set in a strict sense as in QSAR-based systems, but there are example compounds for the alerts stored in its knowledge base.

Toxtree [14, 18] is a Java-based, freely available, open-source application for toxicity prediction. It was developed by IDEAconsult Ltd. (Sofia, Bulgaria) under the terms of a contract with the European Commission Joint Research Centre. The program is mainly based on structural alerts but also provides QSAR models for distinct chemical classes to refine the predictions. For mutagenicity, Toxtree implements the Benigni-Bossa rulebase [19] for carcinogenicity and mutagenicity. The alerts are only differentiated into genotoxic and a small number of non-genotoxic ones, without distinction between carcinogenicity and mutagenicity. Additionally, this module offers QSAR models for aromatic amines and α,β-unsaturated aldehydes, which should improve the predictivity for these specific chemical classes. However, the mutagenicity QSARs refer to *Salmonella typhimurium* TA100 only. With regard to structures that do not trigger any alert, the same considerations on negative predictions made for Derek Nexus apply.

CASE Ultra Expert Rules: As of version 1.5.2.0 of CASE Ultra, an *expert-rule system* is built using rules from expert knowledge or scientific literature for the prediction of bacterial mutagenicity [15]. A detailed description of the software is given in the section describing the statistical-based systems.

Leadscope Expert Alerts System: Leadscope Inc. produces several software modules applicable in the context of toxicological forecasting, particularly in the field of QSAR models. Recently, Leadscope developed a rule-based expert system for the prediction of mutagenicity, using an extensive high-quality genetic toxicity database containing the results of the bacterial mutagenesis assay along with chemical structures [20]. Firstly, the chemical structures were merged using a chemical registration system to assign a unique identifier to each chemical and merging entries on the basis of this identifier. Next, the graded end points for *Salmonella* and *E. coli* were combined from the different sources, resulting in a database of over 7,000 chemicals each with a positive/negative overall bacterial mutation call. The reference set also covers a

diverse collection of compounds since they have been derived from many different sources, including pharmaceuticals, pesticides, industrial chemicals, and food additives. Clustering led to 1,220 clusters with two or more examples and 1,049 singletons (clusters with one example). Once substructures are identified for alert definitions, the selected alerts are consolidated and organized hierarchically (i.e., parent/child). This helps in establishing a mechanistic explanation particularly where any child alert is lacking or has limited mechanistic information, as it may be inherited from the parent alert. When the expert alerts are used to make prediction, a score is calculated reflecting the precision of the alert [20]. In addition to the primary alert, it is also important to define any factors that would deactivate the alerts as a result of electronic or steric effects or by blocking an important metabolic step. In this context, the Leadscope software identified and quantitatively assessed deactivating factors using the 27,000 predefined structural features in Leadscope and generated new chemical scaffolds associated with negative bacterial mutagenicity. Any deactivating fragments identified were quantitatively evaluated using the reference set.

2.2 Statistical-Based Systems

Quantitative structure-activity relationship (QSAR) models are regression or classification models used in the chemical and biological sciences and other disciplines. Like other regression models, QSAR regression models relate a set of "predictor" variables (X) to the potency of the response variable (Y), while classification QSAR models correlate the predictor variables to a category value of the response variable.

The QSAR approach can be generally described as an application of data analysis methods and statistics to model development that could accurately predict biological activities or properties of compounds based on their structures. Any QSAR method can be generally defined as an application of mathematical and statistical methods to the problem of finding empirical relationships (QSAR models) in the form $\mathrm{Pi} = k'(\mathrm{D1}, \mathrm{D2}, \ldots, \mathrm{Dn})$, where Pi are biological activities (or other properties) of molecules; D1, D2, …, Dn are calculated (or, sometimes, experimentally measured) structural properties (or molecular descriptors) of compounds; k' is some empirically established mathematical transformation that should be applied to descriptors to calculate the property values for all molecules. The goal of QSAR modeling is to establish a trend in the descriptor values, which parallels the trend in biological activity [21].

Both commercial and open-source systems are available within the QSAR-based methodologies, and they include, among others, Sarah Nexus [22], CASE Ultra [15], Leadscope Model Applier [23], OECD Toolbox [24], Bioclipse [25], admetSAR, and Prous Institute Symmetry [26].

Sarah Nexus is a statistical system which utilizes a self-organizing hypothesis network (SOHN) model to generate predictions for mutagenicity [27]. This hierarchical model not only retrieves matching fragments, it also further refines these results by considering the structure's similarity to the query structure. The methodology retains those fragments that are perceived to be of greater value; fragments may be of various sizes and can even overlap, ensuring greater accuracy in predictions. Fragments are generated from the provided training set of molecules and not selected from lists of predetermined fragments. Both global (broad coverage, not adequately sensitive to local variations) and local (more accurate for fragments that fall inside their chemical space, narrower in scope) models are available in Sarah Nexus. If the query structure is not an exact match to a compound within the training set (for which a prediction of 100 % confidence is generated), the structure is fragmented and the software will select the most appropriate model for each fragment.

The structural explanation for the prediction provided by Sarah Nexus is conveyed by highlighting those fragment(s) that the model considers meaningful. Sarah Nexus provides a confidence score and a structural explanation for each prediction alongside direct access to supporting data to aid expert analysis [28].

CASE Ultra: CASE Ultra's algorithm is mainly influenced by the original MCASE methodology [29, 30], a traditional QSAR system, which can automatically generate a predictive model from a training set of non-congeneric compounds with associated biological or toxicity data. The training set ideally should contain examples of both active and inactive chemicals in a non-overly skewed ratio.

CASE Ultra can identify alerts that are not limited to linear paths of limited size or limited branching pattern, and the training sets could be larger than 8,000 molecules [31]. To build a model, CASE Ultra picks up one active chemical at a time from the training set and systematically generates a list of fragments for that chemical. Each fragment's relevance for activity is then determined using a two-objective criteria comprised of Shannon's entropy [32] as a fitness measure and the number of the active training set molecules containing this fragment (fragments that are optimal based on the two objectives, i.e., the ones that cannot be replaced by any other fragment without degrading one or both objectives, are selected and then sorted in descending order of the number of their active chemicals). A top few fragments (based on the aforementioned two-objective criteria, e.g., fragments that have low entropy as well as supported by higher number of active training chemicals) are selected. These fragments are considered as potential positive alerts. The fragment generation procedure produces simple linear chains of varying lengths and branched fragments as

well as complex substructures generated by combining simple fragments. When the algorithm has finished scanning all the active chemicals, a search is made in the accumulated list of the potential positive alerts to find the alert that covers the highest number of active chemicals, and it is added to the final list of positive alerts. This step is repeated until enough positive alerts were identified to cover all the active chemicals in the training set. Once a final set of positive alerts is identified, CASE Ultra attempts to build separate local QSARs for each positive alert in order to explain the variation in activity within the training set chemicals covered by that alert. In addition, deactivating alerts are found using a very similar process but by scanning inactive chemicals and finding fragments that occur mainly in inactive chemicals. This collection of positive and deactivating alerts constitutes a model for a particular end point and can be used for predicting activity in test chemicals. During prediction, a test chemical is scanned against the list of the model's positive and deactivating alerts, and if no positive alerts could be identified in it, the chemical is considered inactive. In general, if the test chemical contains one or more positive alerts, it is predicted as "active." However, this active prediction call can be changed if the local QSAR of the positive alert modifies the prediction. The presence of a deactivating alert alongside a positive alert renders the prediction call as "inactive." If more than one positive alert is present, then the one with the highest number of active chemicals is used, and in the case of more than one deactivating alert, the one with the highest number of inactive chemicals is used. If a test chemical contains a positive alert that has been seen in just one or two active training set chemicals, the prediction result is considered "inconclusive" because of the alert low statistical confidence. CASE Ultra recognizes unusual features/fragments in test chemicals that do not match training data (unknown structural fragments). The presence of more than three unknown structural fragments in the test chemical results in an "out of domain" call.

Leadscope Model Applier: The Leadscope software employs a fragment-based QSAR paradigm; however, the fragments are not paths of distinct lengths but are predefined in a hierarchically organized dictionary that is closely related to common organic/medicinal chemistry building blocks. For binary classification problems, such as the Ames test results, the algorithm identifies toxicity modulating fragments using a χ^2-test. Furthermore, the software is able to build superstructures from smaller fragments if they improve predictivity. Together with eight global molecular properties, the set of fragments is then used as a descriptor set in a partial least squares (PLS) logistic regression model of the activity class. Therefore, the predictions from this algorithm are continuous probabilities of class membership rather than binary outputs. The program also assesses the applicability domain by measuring the

distance to training set molecules. Typically, probabilities greater than 0.5 can be used to give an "active" prediction and probabilities smaller than 0.5 an "inactive" prediction, which is the standard procedure used by the Model Applier for pretrained models. The system can also annotate compounds as "out of domain" or with "missing descriptors" when a conclusive prediction cannot be made [23].

OECD Toolbox: The OECD Toolbox [24, 33] represents a free source of various models. The Toolbox is a software application intended to the use of governments, chemical industry, and other stakeholders in filling gaps in (eco)toxicity data needed for assessing the hazards of chemicals. The Toolbox incorporates information and tools from various sources into a logical workflow. Crucial to this workflow is grouping chemicals into chemical categories. The seminal features of the Toolbox are (1) identification of relevant structural characteristics and potential mechanism or mode of action of a target chemical, (2) identification of other chemicals that have the same structural characteristics and/or mechanism or mode of action, and (3) use of existing experimental data to fill the data gap(s). The Toolbox includes a number of models predicting several toxicological end points, such as skin sensitization, Ames mutagenicity, acute and repeat-dose toxicity, aquatic toxicity, and others [34].

Bioclipse [25]: It is an open-source cheminformatics toolkit with a wide array of toxicity models integrated, such as carcinogenicity, mutagenicity (Ames), hERG, aquatic tox (Daphnia), and a wide array of models from OpenTox [35]. The Ames mutagenicity model in Bioclipse is built using the dataset published by Kazius et al. in 2005 [36] containing 4337 chemical structures of which 2401 were classified as mutagen and 1936 non-mutagen. The datasets can be downloaded, and the software can be used to generate many molecule descriptors (using the CDK) [37, 38] and then QSAR models (through integration with the R statistical software). The software is considered not as user friendly as some commercial tools [39].

admetSAR: admetSAR [40] is a free website (http://lmmd.ecust.edu.cn:8000/) [41] that enables a single input SMILES structure to be used to rapidly predict scores against a wide range of ADME/Tox models (at the time of writing, 26 qualitative classification and 5 quantitative regression models). These datasets can also be downloaded as most are based on other publications. Each model has some statistics describing the model as well as a probability to provide more confidence in the result. The software is simple to use, and drawbacks appear to be the lack of batch processing operation, the "black box" nature of the models, and the lack of capability to build or update the models on the website [39].

Symmetry: Symmetry [26] is a platform that applies advanced machine learning techniques to a variety of structural features and physico-chemical properties of small molecules to provide quality predictions about biological effects. Available Symmetry algorithms include binary classification for active/inactive datasets, meta-classifiers to achieve consensus predictions for sets of binary models, and multi-label learning that yields ranking and probabilistic estimates of the possible outcomes. Symmetry offers a wide range of predictive models, including mechanism of action and phenotypic models, toxicity [42], and human adverse effects.

3 Assessment of Potential Genotoxic Impurities

3.1 ICH M7 Guideline

3.1.1 Background

The European Medicines Agency Committee for Medicinal Products for Human Use (CHMP) released in 2006 [43] a "Guideline on the Limits of Genotoxic Impurities," which describes an approach for assessing genotoxic impurities of unknown carcinogenic potential based on the TTC concept. In 2007 a question and answer document was published on the EMA website addressing several aspects of the practical implementation of the recommendations contained in the Guideline.

Genotoxicity is a broad term that typically describes a deleterious action on cellular genetic material. Chemicals may induce DNA damage by directly interacting with it (e.g., alkylating agents) or by acting on non-DNA targets (e.g., mitotic spindle poisons, inhibitors of topoisomerase, etc.). For DNA-reactive genotoxins, the mechanism by which they induce genetic damage is assumed to follow a linear no-threshold model; on the other hand, for molecules not interacting directly with DNA, the existence of a threshold concentration required to induce the damage is by and large accepted [44]. Impurities that belong to the second category of substances can be regulated according to the ICH Quality Guideline Q3C [45] which includes class 2 solvents. The thresholds or permissible daily exposures (PDE) are calculated from the no-observed-effect level (NOEL) obtained in the most relevant animal studies with the use of conservative conversion factors used to extrapolate the animal data to humans.

The CHMP Guideline suggests that the TTC concept should be applied to those genotoxic impurities that do not have sufficient evidence of a threshold-related mechanism of action. The reference values are taken from Kroes et al. [46], where a TTC of 0.15 μg/day is proposed for impurities presenting a structural alert for genotoxicity, corresponding to a 10^{-6} lifetime risk of cancer. In the case of pharmaceuticals, the Guideline suggests a 1 in 100,000 risk be applied, resulting in a TTC of 1.5 μg/day.

For drug substances, the identification thresholds above which impurities are required to be identified are within the range of 0.05

and 0.1 %. ICH Guidelines Q3A(R) [47] and Q3B(R) [48] state that even though the identification of impurities is not necessary at levels lower than or equal to the identification threshold, "analytical procedures should be developed for those potential impurities that are expected to be unusually potent, producing toxic or pharmacological effects at a level not more than the identification threshold." The Guideline recommends carrying out a thorough evaluation of the synthetic route along with chemical reactions and conditions, with the aim of identifying reagents, intermediates, starting materials, and readily predicted side products which may be of potential concern. Once all potential impurities are theoretically identified and listed, an initial assessment for genotoxicity is carried out by a scientific expert using computer tools such as QSAR and knowledge base expert systems. A thorough literature and internal archive (when applicable) search also needs to be completed, as a number of intermediates and reagents have often been tested in genotoxicity or carcinogenicity assays. The potential genotoxic impurities which may be present in an API are then classified into one of five classes described by Müller et al., in 2006 [49]; the purpose is to identify those impurities that pose a high risk and need to be limited to very low concentrations.

In 2006, a task force established under the umbrella of the US Pharmaceutical Research and Manufacturers of America (PhRMA) for the first time proposed the "staged TTC" concept to be applied to pharmaceuticals [49]. The task force was established as a response to various clinical holds imposed by the FDA on investigational drugs in clinical trial phases based on suspicions to contain genotoxic impurities at levels potentially associated with a risk for the volunteers or patients involved in these trials [50]. The staged approach allows levels of daily intake of mutagenic impurities higher than 1.5 μg as defined by the lifetime TTC, namely, 10 μg (for a 6–12-month duration), 20 μg (3–6 months), 40 μg (1–3 months), and 120 μg for not more than 1 month. The EMA adopted the staged TTC approach for limits of genotoxic impurities in clinical trials in the 2007 Q&A document (EMA 2010), but to be more conservative, it reduced the staged TTC limits proposed in the PhRMA paper by a factor of 2.

In 2008, the FDA issued a draft "guidance for Industry on Genotoxic and Carcinogenic Impurities in Drug Substances and Products: Recommended Approaches" (FDA 2008) which was largely similar to the EU guidance. However, this document has not been finalized because in 2009 the topic "genotoxic impurities" was adopted by ICH for development of a new internationally harmonized guideline. Since the topic was considered to include both safety and quality aspects, the projected Guideline was assigned to the M (multidisciplinary) series of the ICH process and designated as ICH M7 with the title "Assessment and Control

of DNA-Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk" [51].

In February 2013 a draft of the M7 Guideline was published in the three ICH regions for public consultation (step 3 of the ICH process). The document was adopted as a step 4 ICH Harmonised Tripartite Guideline in June 2014 (ICH 2014) and is currently on step 5, adopted by CHMP on 25 September 2014 and issued as EMA/CHMP/ICH/83812/2013 [51].

3.1.2 Key Aspects of the ICH M7 Guideline

The ICH M7 Guideline combines many of the principles set by the EU and the draft FDA Guidelines on genotoxic impurities. Some aspects, though, have been updated and clear recommendations can be identified. A thorough description of all key aspects of the ICH M7 Guideline, which are described elsewhere [50], is beyond the scope of the present contribution. It is nonetheless worthwhile mentioning few of the critical aspects that the ICH M7 Guideline does enforce:

1. Structure-based assessment of potentially mutagenic impurities has to be carried out using two in silico systems that complement each other: one should be a rule-based and one a statistics-based method (*see* Subheading 2 in this chapter).
2. The impurities classification system proposed by the ICH M7 Guideline has been derived from the scheme proposed by Müller et al. in 2006 [49], which identifies five classes of impurities as a function of data availability for the characterization of their mutagenicity and carcinogenicity potential.
3. ICH M7 replaced the term "genotoxic impurities" as applied by the EU Guideline on the Limits of Genotoxic Impurities with the term "DNA-reactive impurities" in order to specify that DNA-reactive compounds (i.e., that typically covalently bind to DNA-generating adducts, which, if unrepaired, can lead to point mutations and/or strand breakage) are those that fall within the scope of the Guideline. There is also the assumption that DNA-reactive (Ames-positive) compounds are likely carcinogens with no threshold mechanism.
4. For DNA-reactive (Ames-positive) compounds lacking rodent carcinogenicity data, a generic TTC value would be applied as an acceptable intake level that poses a negligible risk of carcinogenicity.
5. If rodent carcinogenicity data is available for a (potentially) mutagenic impurity, the application of the TTC concept is not warranted, and a compound-specific calculation of acceptable levels of impurity intake is recommended as is described in more detail in the Note 4 of the Guideline [51].
6. Compound-specific calculations for acceptable intakes can be applied case-by-case for impurities which are chemically similar

Table 1
Acceptable intakes for an individual impurity

Duration of treatment	≤1 month	>1–12 months	>1–10 years	>10 years
Daily intake (μg/day)	120	20	10	1.5

Table 2
Acceptable total daily intakes for multiple impurities

Duration of treatment	≤1 month	>1–12 months	>1–10 years	>10 years
Daily intake (μg/day)	120	60	30	5

to a known carcinogen compound class (class-specific acceptable intakes) provided that a rationale for chemical similarity and supporting data can be demonstrated (Note 5) [44, 51].

7. The acceptable intakes derived from compound-specific risk assessments can be adjusted for shorter duration of exposure. The TTC-based acceptable intake of 1.5 μg/day is considered to be protective for a lifetime of daily exposure. To address less-than-lifetime (LTL) exposures to mutagenic impurities in pharmaceuticals, a formula is applied in which the acceptable cumulative lifetime dose (1.5 μg/day × 25,550 days = 38.3 mg) is uniformly distributed over the total number of exposure days during LTL exposure. This allows higher daily intakes of mutagenic impurities than would be the case for lifetime exposure and still maintain comparable risk levels for daily and non-daily treatment regimens.

   Table 1 summarizes the levels for different duration.

8. As far as multiple impurities are concerned, when there are more than two mutagenic (i.e., Ames-positive) or alerting impurities, total mutagenic impurities should be limited as described in Table 2 for clinical development and marketed products.

3.2 Performance of Commercial Systems on Proprietary Compounds

In silico methods for the prediction of mutagenic activity have been available for many years, and they have been continuously improved in terms of technology and prediction results, also for greater availability of high-quality data.

The specific use of such in silico tools in the pharmaceutical industry, in the context of the evaluation of genotoxic impurities, has been recently summarized and reviewed by Sutter et al. [52]. The authors, representing a total of 14 pharmaceutical companies, compared the predictive value of the different methodologies analyzed in two surveys conveyed in the US and European

pharmaceutical industry: most pharmaceutical companies used a rule-based expert system as their primary methodology, yielding negative predictivity values of ⩾78 % in all participating companies. A further increase (>90 %) was often achieved by an additional expert review and/or a second statistics-based methodology. Also in the latter case, an expert review was encouraged, especially when conflicting results were obtained. The conclusion was that a rule-based expert system complemented by either expert knowledge or a second (Q)SAR model is appropriate. Overall, the procedures for structure-based assessment presented in the article by Sutter et al. [52] were already considered appropriate for regulatory submissions within the scope of ICH M7, which mandates the use two different methodologies: one expert-rule based and one statistical-based.

In order to comply with such Guideline specification, additional commercial in silico tools and novel models have been recently made available to the scientific community. Brigo *et al.* [53] evaluated three expert-rule systems (*Derek Nexus v.4.0.5* [13], *Toxtree v.2.6.6* [14], *Leadscope Expert Alerts v.3.2.4-1* [16]) and three statistical systems (*Sarah v.1.2.0* [22], *Leadscope Model Applier v.3.2.4-1* [23], *three models of CASE Ultra v.1.5.1.8* [15]—*GT1_7B, SALM2013, SALM2013PHARMA*) in an individual and combined fashion.

The evaluation was carried out using a large validation set of Ames mutagenicity data comprising over 10,000 compounds, 30 % of which are Roche proprietary data (Table 3). The Roche datasets include the vast majority of compounds (not only impurities) tested in the Ames Standard [54] and Microsuspension [55] protocols.

All programs have been applied as commercially available, without internal customization or follow-up expert knowledge.

Individual systems showed adequate performance statistics with public domain datasets (concordance, 74–95 %; sensitivity, 58–99 %; specificity, 51–96 %; *see* Fig. 1); however, there was a consistently significant drop in sensitivity with the Roche datasets,

Table 3
External validation sets

Dataset	Number of compounds	Positive	Negative
Roche Ames Standard	1,335	254	1,081
Roche Ames Microsuspension	1,785	190	1,595
LSDB	4,699	2,068	2,631
Hansen [56]	2,647	1,773	874
Total	10,466	4,285	6,181

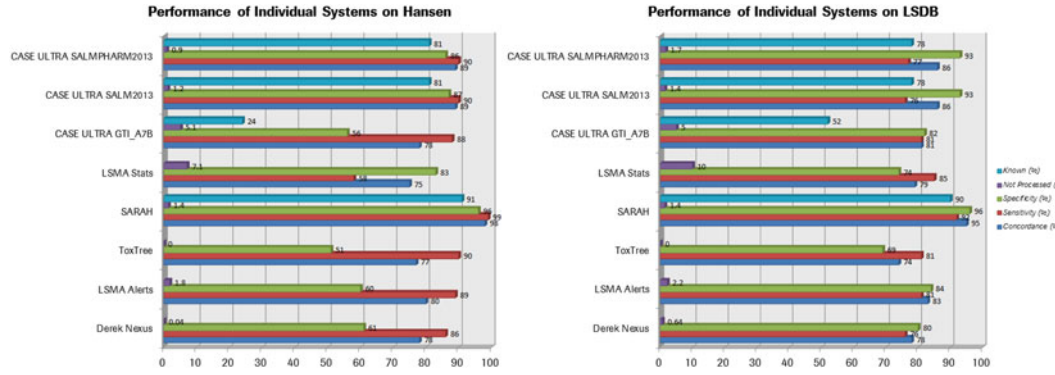


Fig. 1 Performance of individual systems on public datasets Hansen [56] and LSDB

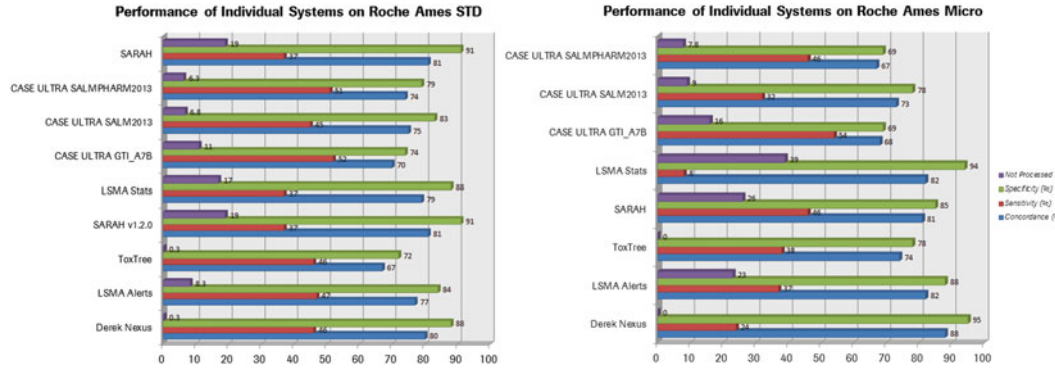


Fig. 2 Performance of individual systems on Roche Ames Standard and Ames Micro datasets

down, in one case, to single digit (concordance, 66–88 %; sensitivity, 8–54 %; specificity, 69–95 %; *see* Fig. 2). All systems showed good performance with "public validation sets," also due to the training set overlap, which went up to 91 % for Sarah (Fig. 1).

Expert-rule-based tools showed lower specificity with public domain datasets versus the statistical-based programs. Statistical tools showed a much higher number of compounds (up to 39 % in one case) outside of their applicability domains and, hence, not predicted (Fig. 2).

To evaluate the performance of the combined approach recommended by the ICH M7 Guideline, the Roche validation sets have been submitted to all possible combinations of one expert-rule-based and one statistical-based system (Figs. 3 and 4).

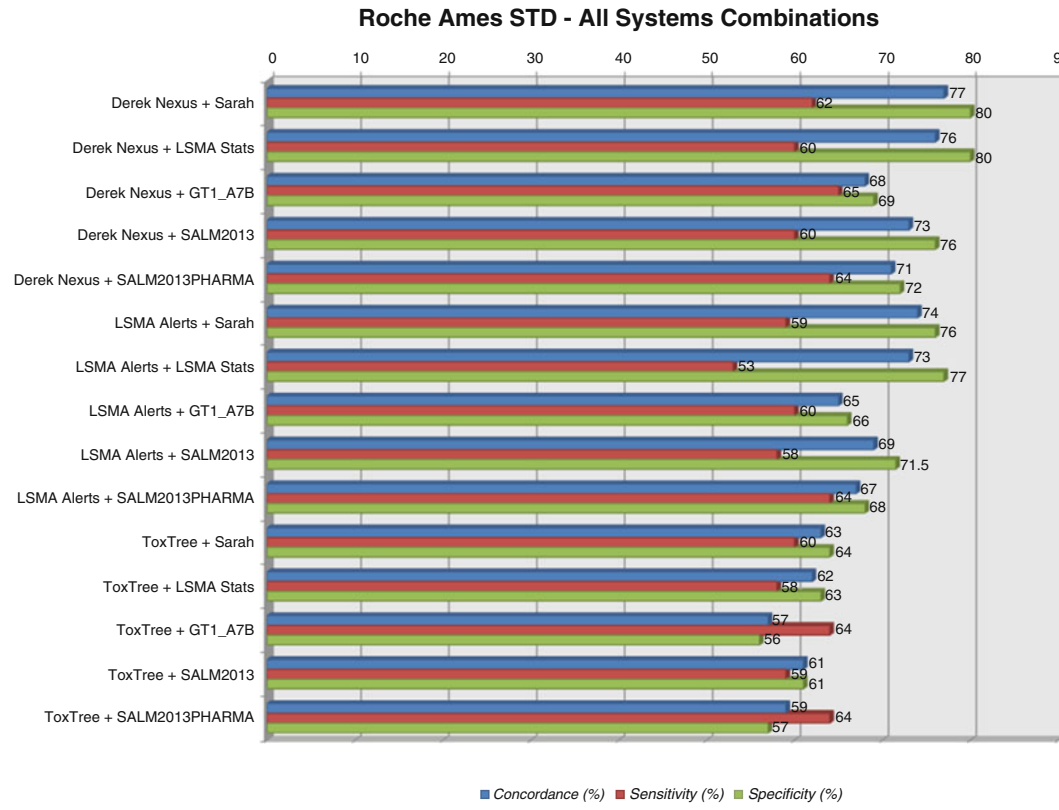


Fig. 3 Performance of combined systems on the Roche Ames Standard dataset

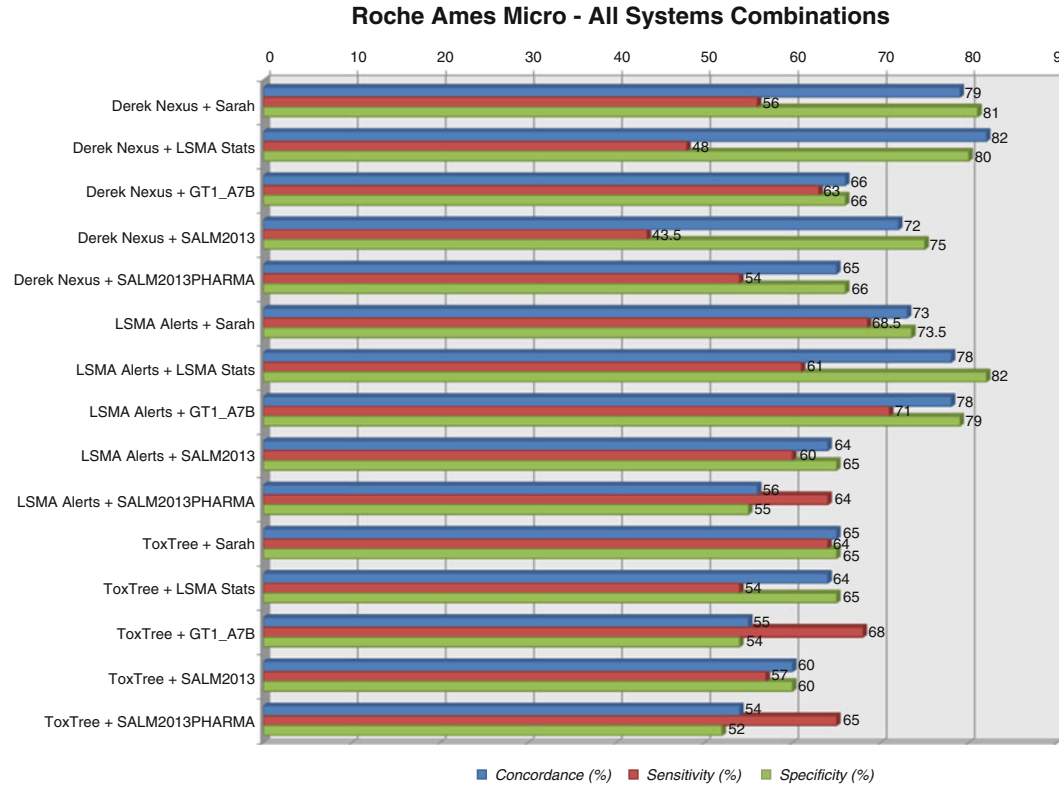


Fig. 4 Performance of combined systems on the Roche Ames Micro dataset

The combinations of all systems, compared to their individual performance with both Roche validation sets, improve the sensitivity to consistently above 50 %, up to 71 % for the combination "LSMA Alerts + SALM2013." As expected, specificity is generally lower than with individual systems, but its reduction is limited for the majority of combinations.

In order to assess the prediction tools with chemicals that fall within the potential genotoxic impurities chemical space, four subsets of both Roche validation sets have been generated with molecular weights (MW) ≤400, ≤350, ≤300, and ≤250. Such subsets cover the chemical space of the large majority of the potential genotoxic impurities tested in Roche over the past decade.

All programs have been tested against these subsets individually (Figs. 5 and 6) and in combination (Figs. 7 and 8)[53].

With individual systems, sensitivity shows a clear trend to increase proportionally to the decrease of MW. For example, in the Roche Ames Microsuspension set, sensitivity improves as follows: Derek from 27 to 64 %, Sarah from 51 to 76 %, Toxtree from 42 to 85 %, and CASE Ultra SALMPHARM2013 from 45 to 71 %. LSMA Alerts and LSMA Stats show an increase in sensitivity to 60 % up to MW ≤300, but there is a flection down to ~55 % for both programs for MW ≤250. In general, sensitivity increases significantly with low-MW subsets with almost all programs and models (Figs. 5 and 6). The only exception is CASE Ultra SALM2013 model, which keeps the same sensitivity values throughout all subsets (between 29 and 33 %) [53].

The evaluation of combined systems with low-MW Roche subsets shows a significant increase in sensitivity, up to over 90 % for sets with MW ≤300 and ≤250 with several combinations (Figs. 7 and 8). The increase in sensitivity is proportional to the decrease in MW; at the same time, there is a considerable decrease in specificity (<30 % in some cases). Such deltas are generally more pronounced in the Ames Micro dataset (Fig. 8) compared to the Roche Ames Standard dataset (Fig. 7). In the Ames Standard subsets, specificity and sensitivity values are consistently comprised between 70 and 80 % in nearly all Derek Nexus and LSMA Alerts combinations. In the latter combinations, values are a bit lower than 70 % at higher MW. Toxtree combinations show lower sensitivity and specificity values at higher molecular weights and greater gaps between sensitivity and specificity within the subsets MW ≤300 and MW ≤250 [53].

As far as the Roche Ames Micro set is concerned, the sensitivity is in the range of 90 % in the subset with MW ≤250 with several combinations, such as Derek Nexus + Sarah and Derek Nexus + GT1_A7B; LSMA Alerts + Sarah; LSMA Alerts + CASE Ultra models. Nearly all combinations with Toxtree gave sensitivity in the range of 90 %. Nearly all combinations of LSMA Alerts showed high sensitivity also in the subset with MW ≤300.

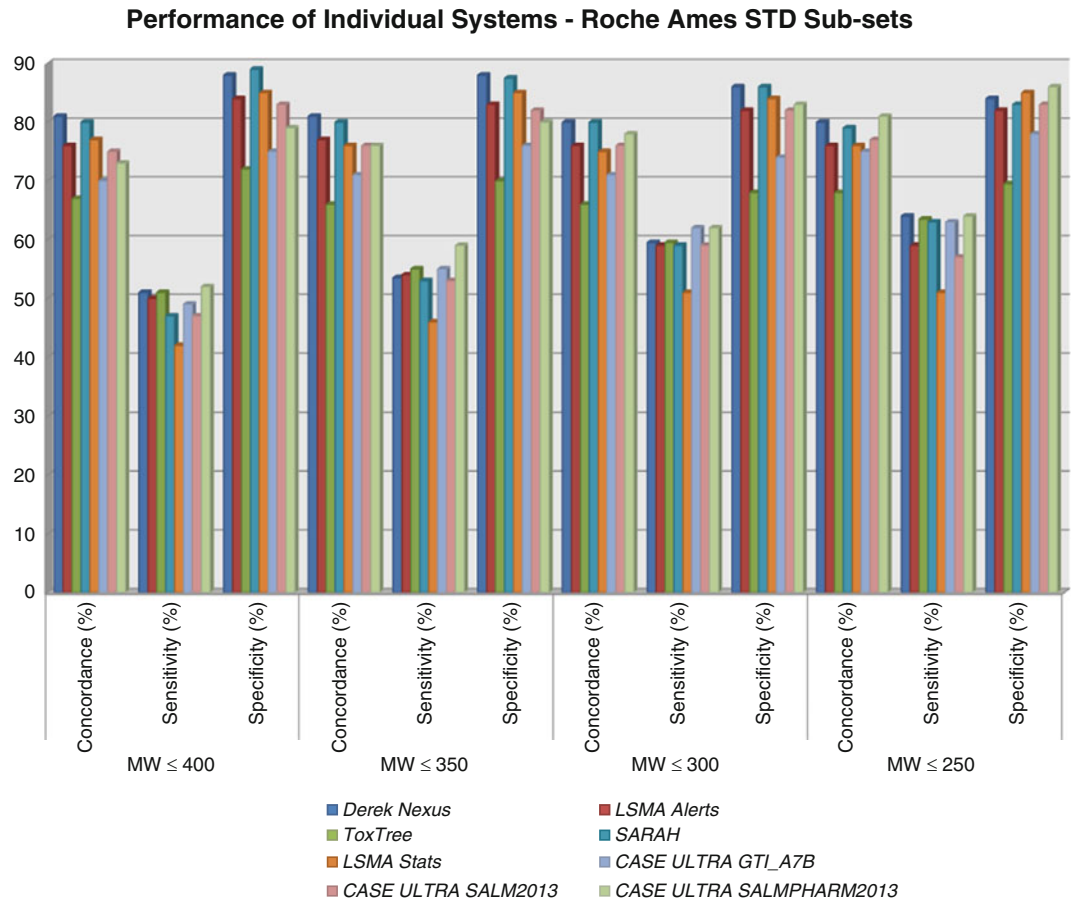


Fig. 5 Performance of individual systems on the Roche Ames Standard dataset filtered by MW

Looking at the plots in Fig. 8, it is evident that more balanced results are obtained with all Derek Nexus combinations: in other words, the sensitivity increases proportionally to the decrease of the MW at a moderate expense of specificity. Compared to this, LSMA Alerts combinations have overall lower specificity than Derek combinations. At the same time, Toxtree combinations, despite showing good sensitivity, have a greater corresponding decrease in specificity.

3.3 Improvement of In Silico Predictions with Proprietary Data

Validation exercises such as those described in Subheading 3.2 for mutagenicity or for other end points are typically very useful for the identification of specific gaps in the chemical space represented by the assessed models and tools.

In particular, when proprietary data are used as external validation sets, false predictions represent a valuable opportunity to improve the models and expand their overall applicability domain.

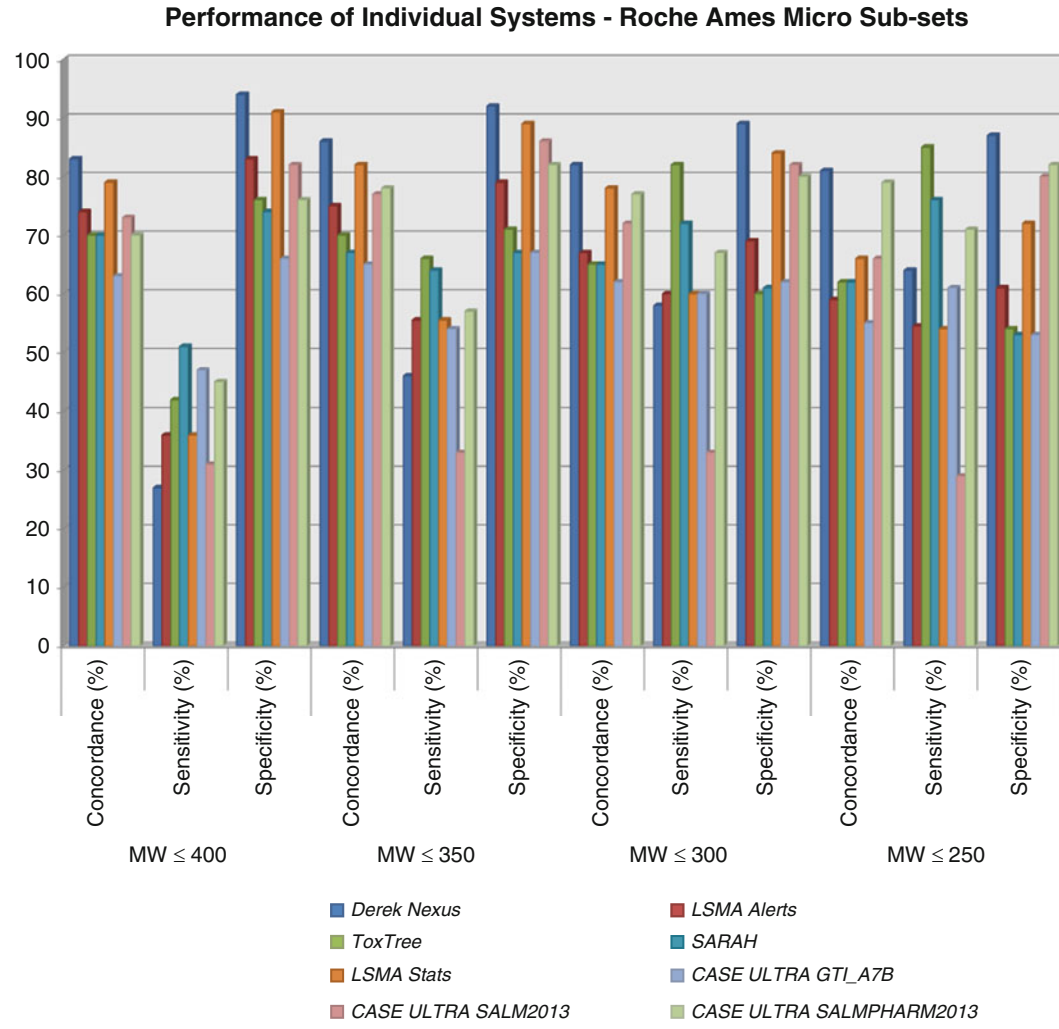


Fig. 6 Performance of individual systems on the Roche Ames Micro dataset filtered by MW

Roche recently undertook a similar exercise with Lhasa Ltd. in order to systematically include proprietary knowledge into the in silico prediction tools that are routinely used for early safety assessment. Data collected from Ames test, embryonic stem cell assay (teratogenicity), hERG inhibition in vitro screening, and micronucleus in vitro (chromosome damage) have been used to fill the gaps identified in the models adopted within the company.

These collaborative efforts, aimed at incorporating proprietary knowledge in prediction models, quickly translated into a significant increase in the prediction metrics (*see* Table 4), with sensitivity values that showed up to 60 % improvements.

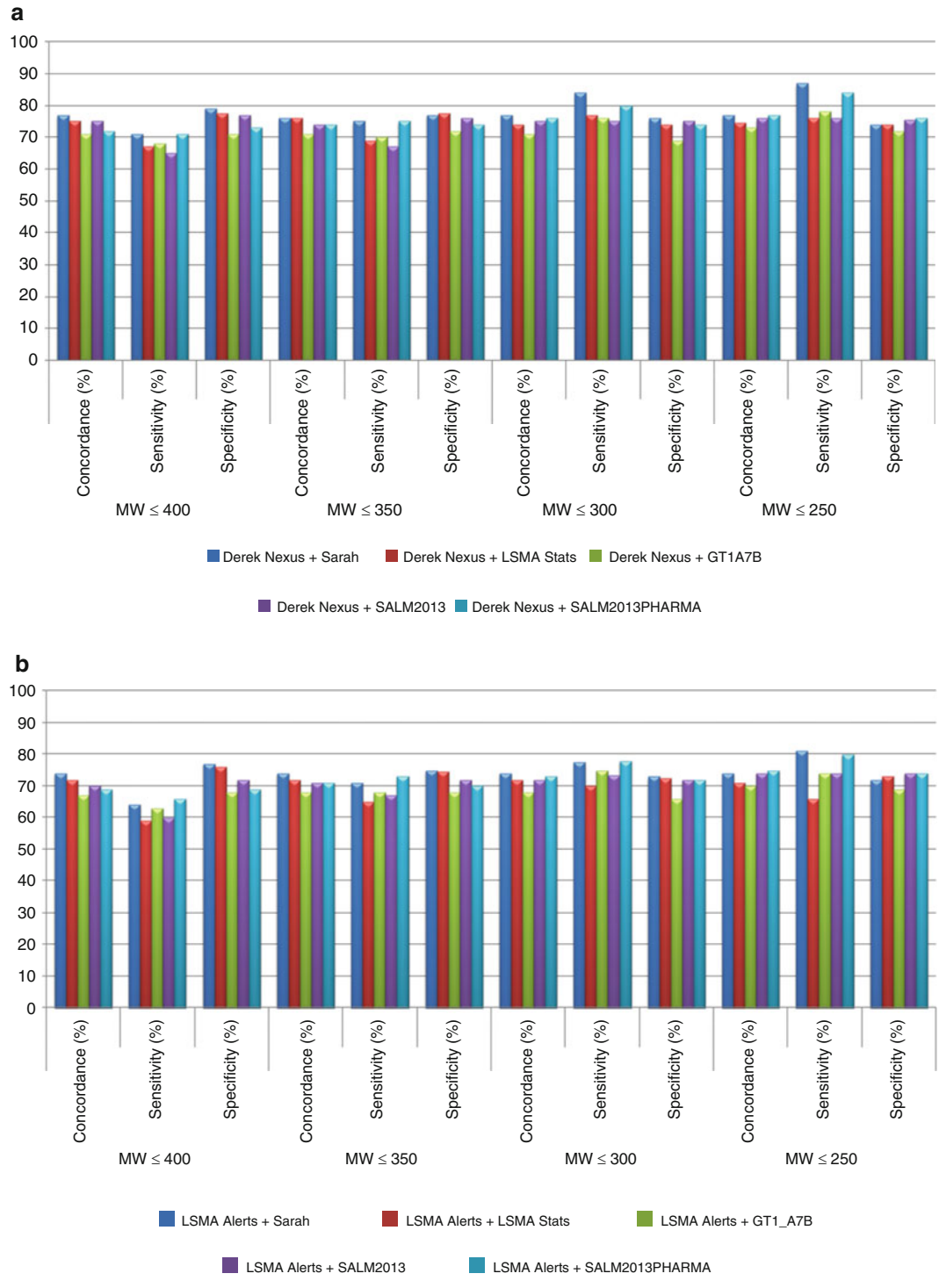


Fig. 7 Performance of combined systems on the *Roche Ames Standard* dataset filtered by MW

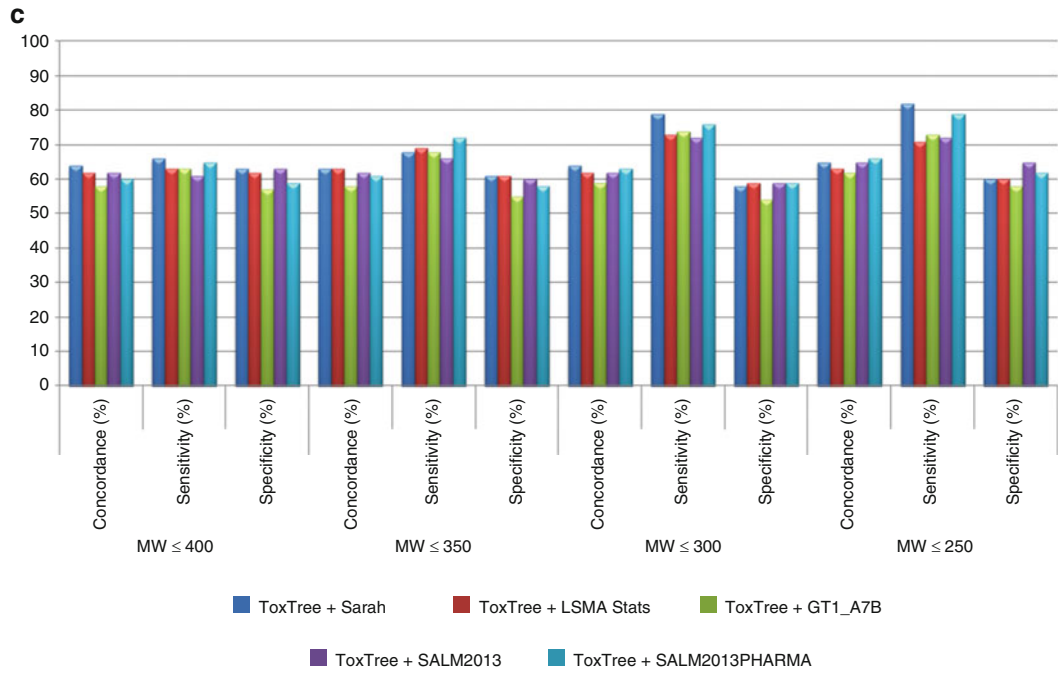


Fig. 7 (continued)

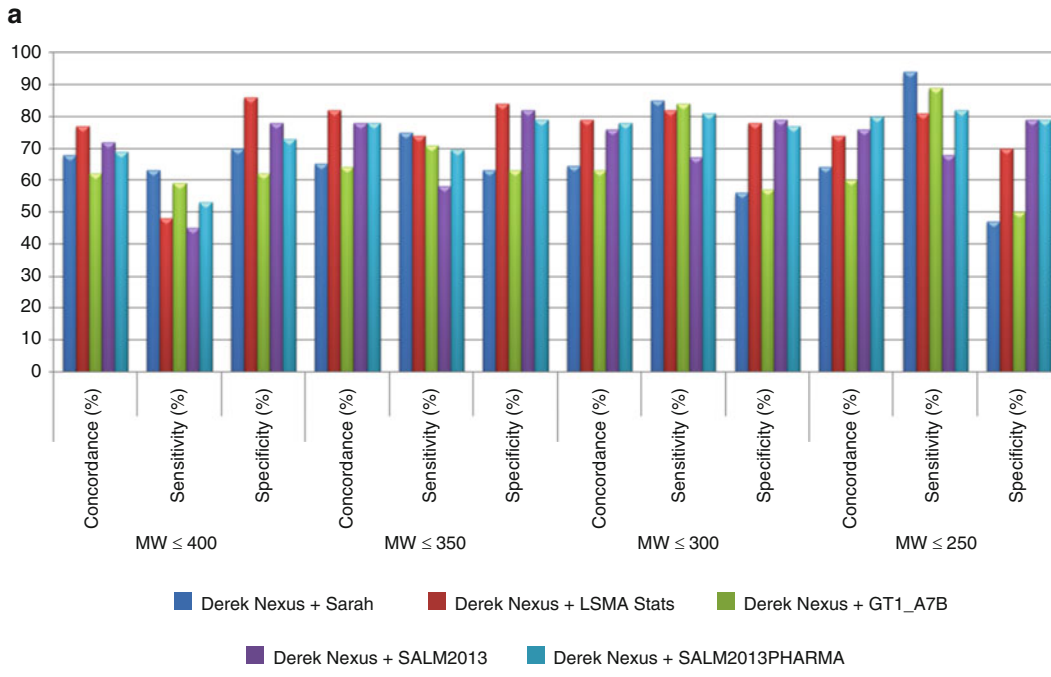
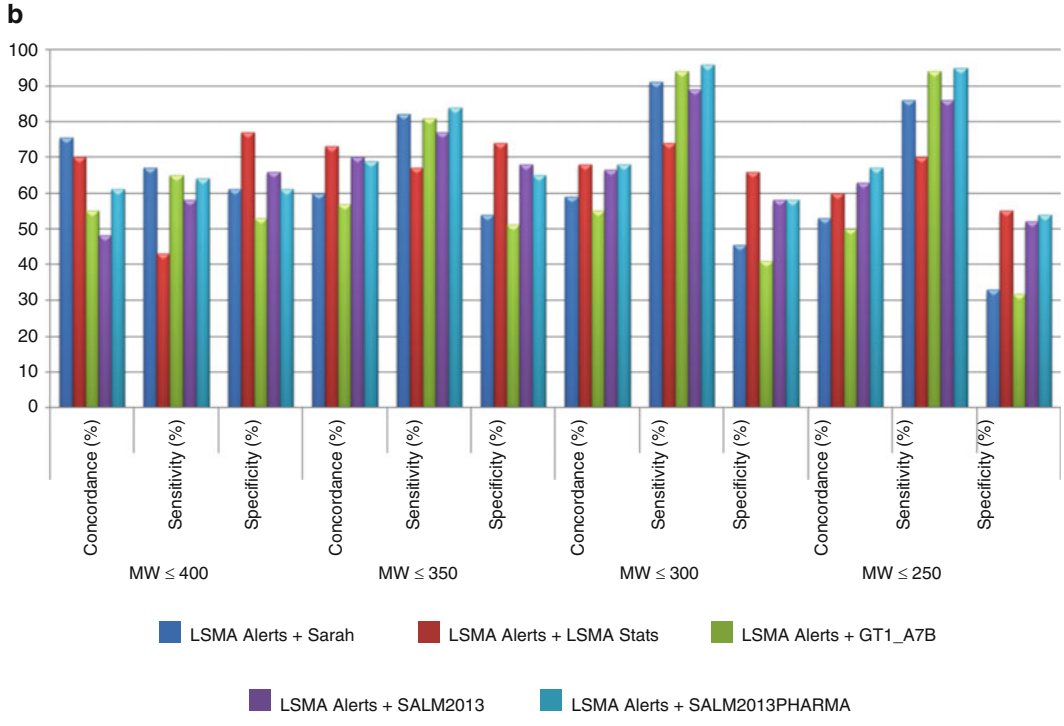


Fig. 8 Performance of combined systems on the *Roche Ames Micro* dataset filtered by MW



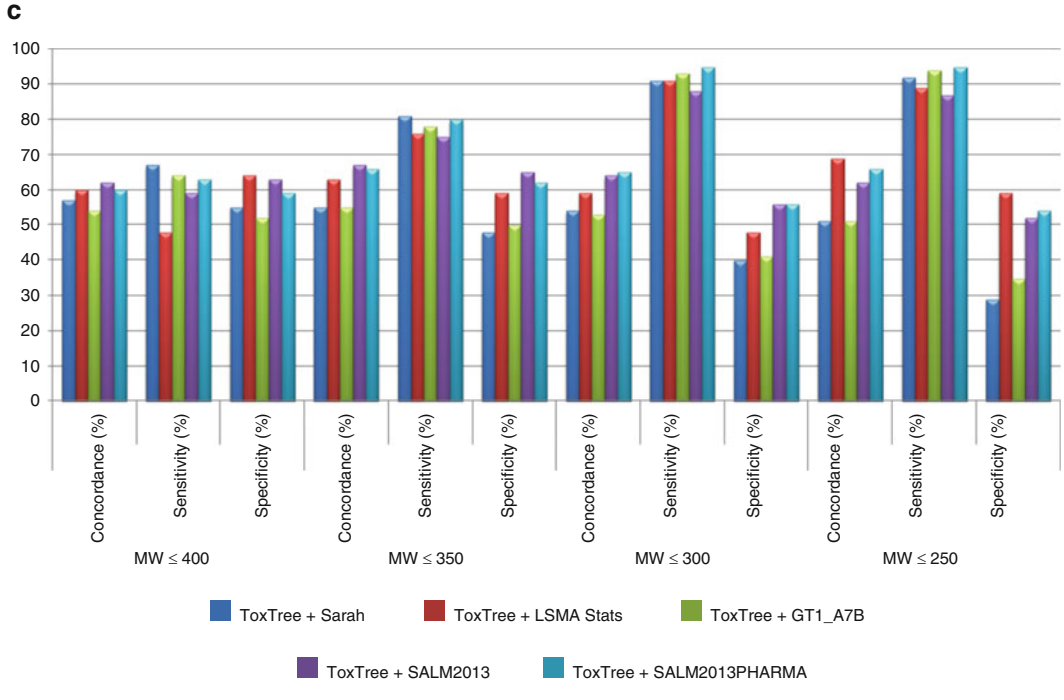


Fig. 8 (continued)

Table 4
Improvement in predictive performance of an in silico prediction tool including Roche proprietary data

End point		Sensitivity (%)	Specificity (%)	Positive predictivity (%)	Negative predictivity (%)	Balanced accuracy (%)
Mutagenicity	Previous	36	92	45	89	64
	Updated	69	89	55	94	79
Chromosome damage	Previous	5	97	34	76	51
	Updated	65	92	72	89	78
hERG inhibition	Previous	21	90	70	50	55
	Updated	63	67	69	61	65
Teratogenicity	Previous	3	96	17	79	50
	Updated	59	92	66	90	76

4 Role of In Silico Models in the Prediction of Toxicity in Drug Discovery

In silico approaches to predict potential toxicities and drug metabolism on the basis of the chemical structure are of particular interest to the pharmaceutical industry as having the potential to impact the early drug discovery process as well as in the candidate selection phase. Prediction models for the identification of metabolic soft spots and potentially toxic substructures can be easily applied to a large number of chemical structures and are therefore integrated already during HTS (high-throughput screening) or even earlier as an automatically attributed alert for all new chemical entities. At this early stage, only a basic in silico profiling can be done, as only the most well-validated end points can be reliably applied automatically and generated on the fly without an expert intervention. Later in the development, at the latest before the final candidate is selected, a more detailed in silico profiling also considering the whole profile of the compound is thoroughly conducted. According to the development scheme provided in Fig. 9, the further in silico tools and in vitro downstream activities are conducted.

4.1 Target Identification (TI), Target Assessment (TA), and Exploratory Work

The first step after the target has been identified as potential development opportunity is a target assessment (TA) conducted by nonclinical safety experts, using appropriate databases and public sources. A proper target/functionality assessment in healthy and diseased status contains pathway mapping, information from knockout and transgenic models, a target expression profile in relevant species, as well as a critical evaluation of potential off-target

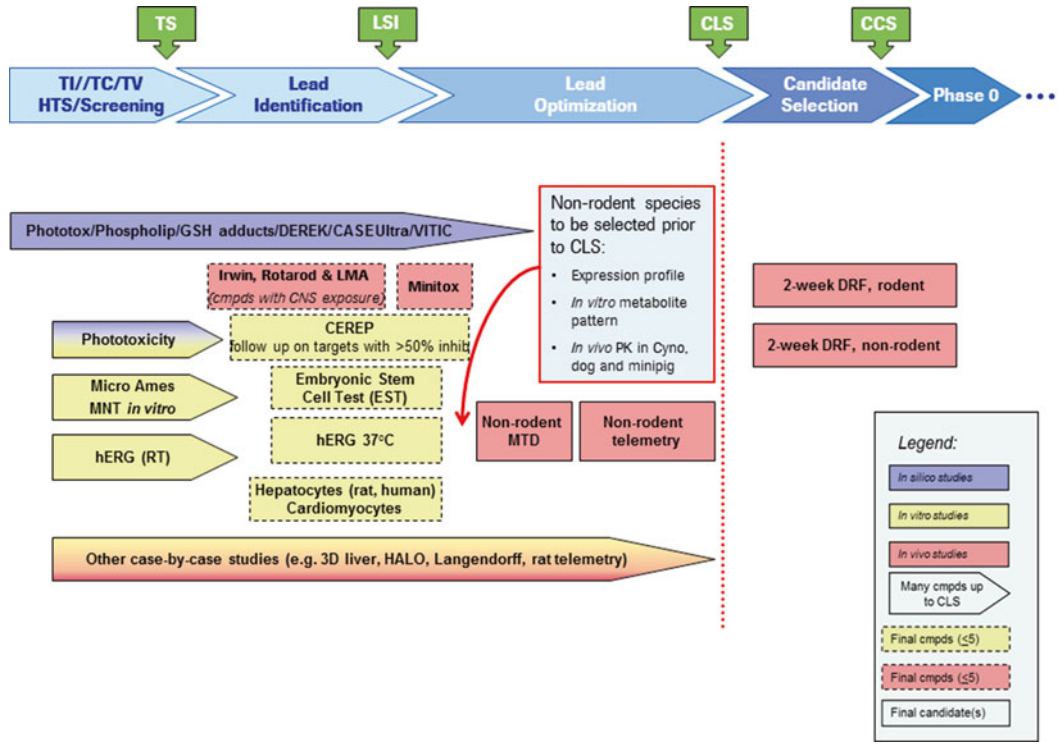


Fig. 9 Use of in silico tools and safety screening during the early drug development process

safety alerts (selectivity). Various software systems are available to assist the experts in these assessments (e.g., MetaCore [57], Symmetry [26]).

Modern in silico prediction software is able to calculate thousands of chemical structures on the fly and can be therefore applied very early in the drug development process. Immediately after the chemical structure is known, meaning chemical libraries are added to the companies' chemical database, a basic in silico prediction panel is applied using reliably validated toxicological end points like genotoxicity and carcinogenicity. As always, a large, homogeneous, and high-quality database is the prerequisite for reliable predictions. Therefore, in vitro screens which have been used within pharmaceutical companies for years containing data generated often in one single lab are the best sources for the development of highly predictive models. For example, an in silico model predicting the potential of drug-induced phospholipidosis (a reversible storage disorder characterized by accumulation of phospholipids within cells) has been developed. Based on more than 600 in vitro assay, an accuracy of 86 % led to a replacement of the in vitro by the in silico method. The model is calculating the free energy of amphiphilicity ($\Delta\Delta G_{AM}$) and log P value [58] of cationic amphiphilic drugs and can be applied in a high-throughput mode.

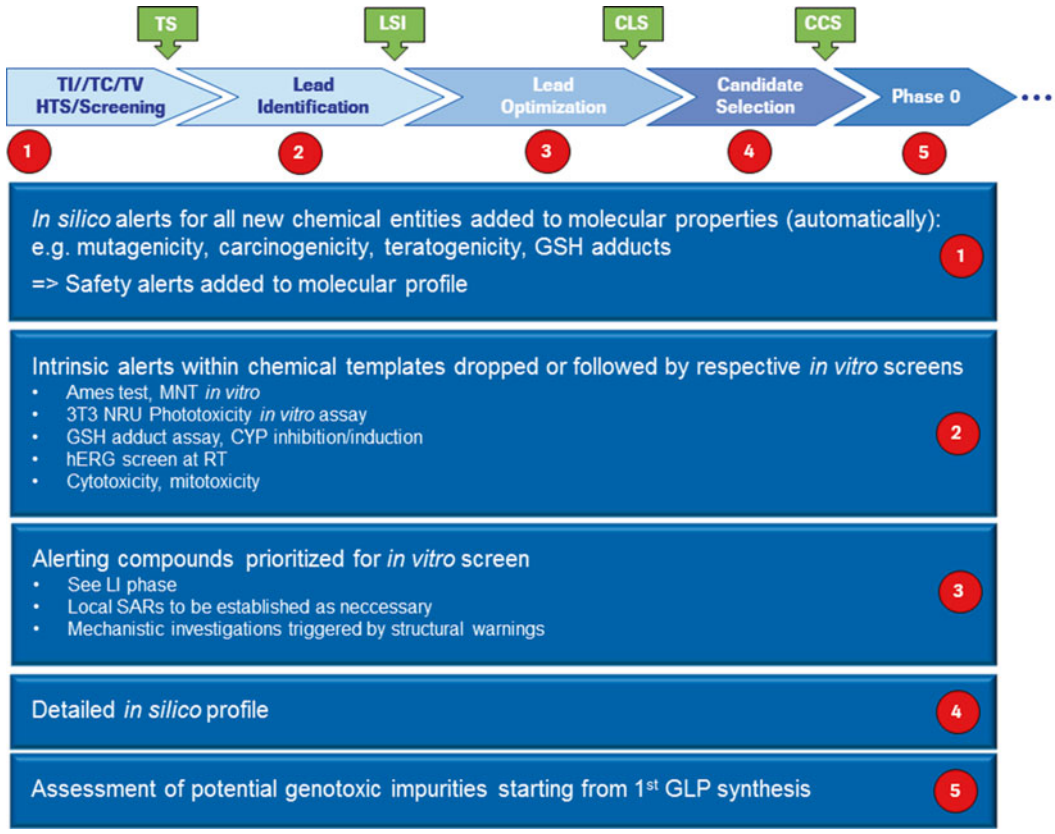


Fig. 10 Downstream activities following in silico alerts in the drug development process

Further end points, which can be used for on-the-fly predictions, are teratogenicity, GSH adduct formation, irritation, and skin sensitization.

At this early stage of development, the potential safety hazards identified by the application of an expert system in combination with a set of statistical models contribute to the overall compound profile, but are not used as a decision pathway (*see* Fig. 10).

4.2 Lead Identification (LI) Phase: Target Selected (TS) to Lead Series Identified (LSI)

The main goal during lead identification is to identify valid chemical templates for further optimizing the efficacy and selectivity on the target, ideally multiple discrete series. Besides computational chemistry tools to calculate physicochemical properties, virtual screening, structure-based design, QSAR analysis of both the desired target and off-target activities, and chemical structures are analyzed continuously in silico for possible structure-related safety concerns to identify major issues with the templates. Insights into the toxicological potential of a scaffold or series of structures early in the drug discovery process could help medicinal chemists to

prioritize particular scaffolds. Components of early avoidance of chemical structure safety liabilities include predictions for genotoxicity, carcinogenicity, hERG channel blockade, reactive metabolite formation, phospholipidosis, structural similarity to problematic molecules, CYP inductions, GSH adduct formation, and DMPK properties (cell penetration, microsomal stability, CYP3A inhibition). The in silico tools offer good guidance on what additional tests may be necessary or whether further characterization is warranted; however, they also have limitations [59].

Drug, metabolism and pharmacokinetics (DMPK) properties play a major role during lead identification. Numerous commercially available tools for the prediction of metabolites exist, such as METEOR [11, 17], MetabolExpert [60], and MetaSite [61]. Most software packages correctly predict metabolites that are detected experimentally. However, a relatively high incidence of false predictions of metabolites is common to most unspecified computerized systems. In the hand of drug metabolism experts, these software packages have a certain value for hypothesis generation and guiding to experimental approaches for the identification of drug metabolites. However, the generation of additional new local rules, intended to predict the activity of a single enzyme (and often only within a chemical series), can significantly improve the prediction accuracy.

Experimental follow-up of potential issues are conducted to build/refine safety plans moving forward. Even if in vitro assays clearly disprove identified in silico alerts, further spot-checking of the distinctive end point will be conducted to avoid creeping in of a structural liability. The in vitro results always overrule the in silico warnings provided that the corresponding assays could have been conducted under reliable conditions (e.g., solubility, stability). Chemical templates with identified and confirmed intrinsic metabolic and/or safety concerns will be eliminated (*see* Fig. 10).

4.3 Lead Optimization (LO) Phase: Lead Series Identified (LSI) to Clinical Candidate Selected (CLS)

The task of the LO phase is to take a lead and convert it into a candidate for preclinical evaluation. This phase is intensively accompanied by early safety in vitro screening in various areas: genotoxicity, hERG and other ion channels, cytotoxicity, hepatic toxicity, bone marrow toxicity, transporters, metabolite identification, metabolic stability, CYP induction/inhibition, reactive metabolites, off-target pharmacology/secondary pharmacology, and cross-species comparisons, where applicable. Further screens might be applied based on the target liabilities or already identified potential safety issues. If adverse in vitro activities appear, specified structure-activity relationships (SARs), so-called local SARs, will be established to support the discovery projects in optimizing the clinical candidates toward safety/DMPK in parallel to efficacy.

At this stage, first, fit for purpose in vivo studies are conducted to address early target or chemistry related safety concerns. The first general toxicology studies are maximal tolerated dose (MTD) and dose-range finding (DRF) studies generally performed in rodents and non-rodents. The value of performing exploratory drug safety studies before candidate nomination is to identify unwanted toxicities evident in a study of up to 14 days duration, as well as any potential toxicities anticipated based on a known cause for concern. In the absence of findings or the presence of findings that are judged manageable, these studies provide a greater comfort in the selection of a molecule for advancement into development with higher likelihood of success. Additional benefits of these studies are the identification of target organs to monitor in development and the selection of doses for the GLP toxicology studies. In addition, identification of the toxicity profile of a lead compound can be useful for the backup program where the goal is often an improved safety margin. In silico safety concerns might be included as part of pharmacokinetics/pharmacodynamics (PK/PD) characterization in vivo (disease) models to extract safety-relevant information and to build confidence in safety before expanding into larger regulatory animal studies.

During LO, every in silico alert is immediately followed up by the corresponding in vitro screen, in case, even in vivo studies might be frontloaded. To avoid late failures of optimized candidates, spot-checking of the potential development candidates without alerts is conducted if the resources and throughput of the assay allows. In case of screening alerts, creation of local SARs can result in significant acceleration of project by optimizing the chemical improvement rounds. Specific and tailor-made local models normally have a significantly higher accuracy, if continuously updated with new incoming screening results. Learnings and newly identified alerting substructures should be implemented in general rules and models (customized systems) to continuously improve the performance of the computational tools used for drug optimization (*see* Fig. 10).

4.4 Phase 0: After Clinical Candidate Selected (CLS) to Entry into Humans (EIH)

The main usage of in silico tools after the final candidate has been selected encompasses the assessment of potentially genotoxic impurities according to the ICH M7 Guideline as described in Chapter 3, as well as cross-reading and pathway analysis following an unexpected event in preclinical studies. Furthermore, a backup or fast-follower program will trigger dedicated in silico profiling and screening of the new molecules, based on experiences and identified issues of the frontrunner compound.

Apart from the use of in silico tools to assess genotoxic impurities, there are no computational assessments which are mandatory requirements from regulatory agencies, but in case in silico models have been applied during drug development and influenced the

testing strategy or triggered addition investigations, the information should be included in regulatory documents and adequately described.

5 Prediction of Complex End Points

5.1 Challenges in the Prediction of the Outcome of In Vivo Safety Studies

When the goal is the prediction of the outcome of certain assays, such as the Ames assay [54], in which the results can be roughly considered as binary, i.e., "yes" or "no" answer, in silico models have a higher chance to give a better performance if compared to more complex assays and studies.

The mechanism of action of a molecule leading to a specific readout plays a critical role in the predictive performance of in silico models as it is one of the biggest challenges of, for instance, QSAR models. "Do the descriptors have any physicochemical interpretation that is consistent with a known biological mechanism? [62]" is often a very difficult question to answer. In vitro chromosome damage (an assay used to establish the clastogenicity potential of test compounds) can also be considered binary (i.e., the test item is "clastogenic" or "not clastogenic"). However, the mechanisms of action that may lead to clastogenicity are manifold and may involve the interaction of the compound with a number of proteins or enzymes, the disruption of one or more biological pathways that ultimately lead to a clastogenicity outcome. This complexity is reflected in the performance of the in silico prediction tool described in Table 4 for the chromosome damage end point. Before the update of the model based on internal data and structures, the sensitivity was in the single digit, showing that the model was practically unable to identify any clastogenic compound within the validation set used. The update was successful in increasing the sensitivity value to 65 %; nonetheless, we need to bear in mind that due to the various mechanisms of action that can lead to clastogenicity, minor structural changes within a chemical class can have a large impact on the mechanism of action (e.g., the interaction with one or more proteins may be hampered, hence changing the final outcome of the assay).

Even greater challenges are offered to the prediction of the outcome of single-dose and repeat-dose toxicity in vivo studies. In the pharmaceutical industry, such studies are typically used to identify a maximum tolerated dose (MTD) and the NOAEL (non-observed-adverse-effect level) for a test compound, in addition to the identification of a general toxicity profile and significant target organs that may show toxicity upon exposure to the compound tested.

Since animal models are very complex and the number of readouts collected in such studies is extremely wide, the development of in silico models that can reliably predict such outcomes is

extremely challenging. For example, a typical repeat-dose study requires the use of a control group plus three dose groups: each animal is then carefully examined for clinical observations throughout the in-life part, including body weight and food consumption measurements as well as some behavioral evaluations; clinical pathology values are collected at different time points; urine analysis is performed; macroscopic and microscopic examinations are carried out on a number of selected organs; toxicokinetics values are then calculated using the test item concentrations measured in blood from the samples collected throughout the study, which could be of different durations, from 5 days till 39 weeks (up to 2 years for the rodent bioassays for the evaluation of carcinogenicity), and in different species.

The variations, permanent or transient, of the parameters and values briefly described above may depend on the pharmacological target, on the chemical structure and related physicochemical properties, and on background incidences due to adaptations or other factors, such as major differences in plasma exposures. Because of this variability, building an in silico model capable of predicting all these different "degrees of freedom" or "dimensions" is extremely challenging, in particular due to the fact that the identification of unequivocal mechanisms of action for whatever findings have been identified is not trivial. In addition, the development of robust SARs using the outcome of such studies is difficult because of the limited amount of publicly available data, and, even within large pharmaceutical companies, the number of chemically similar compounds tested in such long and expensive studies for each investigated pharmacological target is small (less than 5). This, of course, hampers the possibility to even develop local models since the number of similar compounds, designed for the same target, undergoing the same type of studies is rather limited.

Even if some sophisticated in silico models may become available for the prediction of the potential findings identified in, for example, repeat-dose studies, all the limitations described above and the difficulties to conduct a proper validation would make very difficult, within the pharmaceutical industry, to accept them for decision making on compounds prioritization or as guidance for chemical optimization.

5.2 Data Collection, Organization, Availability, and Interpretation for In Vivo Toxicity Studies

Within the industry, it has been recently recognized that the consolidation of the results of in vivo toxicity within appropriate tools making use of the right technology would allow the full exploitation of the knowledge that such data can provide.

Large pharma organizations can typically count on many years of drug discovery and research conducted across several sites on a significant number of therapeutic areas, pharmacological targets, and molecules. This translates into a large amount of complex datasets, stored in different repositories or Laboratory Information

Management Systems (LIMS) each designed specifically to accommodate the data type of interest (e.g., histopathology, clinical observation, clinical pathology, PK, etc.). The organization of such wealth of information to generate specific knowledge from the integration of all of these data types has been considered several times in the past by many pharmaceutical companies. However, due to limited resources or inadequate technology, the outcome of such initiatives has often been disappointing.

In more recent years, there has been a tremendous focus across industries, not only pharma, to extract knowledge and identify patterns or trends from large amounts of data, being either omics, market research, public preferences on digital movies rental [63], airplane estimated times of arrivals, or others. A lot of these initiatives often fall under the term "Big Data," generally underlying the intention of large organizations to look deeper into their databases and assess whether an improvement in the way such data are organized, stored, made accessible, and mined may provide any advantage for the business in terms of saving resources or increasing efficiency via surfacing hidden value.

Along this line, Roche has been working on a number of "Big Data" projects across several areas of research and IT. One of them had the goal to integrate all in vivo nonclinical safety data generated by the company over the past 30 years across three research sites, two in the USA and one in Switzerland. The goal was to ensure that all different data types that are part of in vivo studies (i.e., histopathology, clinical observations, PK, clinical pathology, etc.) were brought together electronically in such a way that they could all be searched and made available at the same time to the user community. The scope for such a platform, internally called SDI (i.e., Safety Data Integration), is to allow scientists to identify specific patterns of findings across species and their historical relevance and correlations between molecular structures and toxicological effects and, eventually, use the data to generate more reliable prediction algorithms. The application of a semantic data integration approach [64] for the harmonization of terms, formats, units, and taxonomy allowed the implementation of a nonclinical study warehouse including approximately 5,000 studies of different types which can be interrogated with very complex queries such as "Which compounds showed spleen hyperplasia and liver necrosis and lung leukocytosis and an AST increase >50 %?", returning an answer in a matter of seconds. The identification of studies and compound matching the query above, in the absence of properly designed data integration efforts, would have been extremely labor intensive and time consuming, if possible at all.

In addition, the SDI platform has been interfaced with other, already existing, internal databases, such as the chemical structures and the in vitro biology data repositories to further expand the data integration beyond toxicology allowing the users to assess the compound profile in almost its entirety.

5.3 Possible Model Generation

As far as model development is concerned, the advantage of the platform described in Subheading 5.2 is the high data granularity available, down to the single animal level.

One of the challenges in the development of predictive models for complex end points, such as hepatotoxicity, is that the modeler is forced to make a generic classification (hepatotoxic vs. non-hepatotoxic), often neglecting safety margins (vs. pharmacological activity), doses at which specific toxicity is seen, and ignoring the specific findings and whether it is transient or not. This is because, more often than not, such information is not easily available. All these factors make such classification relatively inaccurate: for example, paracetamol (or acetaminophen), an over-the-counter mild analgesic, commonly used to relieve headaches and reduce fever, is commonly classified as hepatotoxic (as its overdose can cause fatal liver damage [65]). However, at doses as high as up to 4 g per day in adults, paracetamol is regarded as totally safe and can comfortably be used (at lower doses, of course) even in infants. This example explains how critical and challenging a correct classification is: it is correct to classify paracetamol as hepatotoxic, since an overdose would likely cause a fatal liver failure? However, in drug development settings, what type of decision can be made on a compound predicted to be hepatotoxic by a model based on the information gathered, among others, from paracetamol? Should this molecule be discontinued and any further investigation stopped before knowing what safety margins might there be with regard to its intended therapeutic indication? Disregarding this molecule immediately after a positive prediction bears the risk of losing a potentially valuable compound. Continuing the investigations to further profile the molecule for future clinical development may be the best option to get to a more solid data-driven decision on its potential to become a drug. The bottom line is that, in this context, the prediction model will have a negligible impact on the decision.

In order to strengthen the reliability of in silico models for the prediction of complex end points, all information generated by in vivo single- and repeat-dose studies should be made available in a clear and searchable way at the highest possible level of details. This would allow experts to generate very specific models by making the correct compound classifications for very specific findings via a preliminary and careful data analysis. For example, it will be possible to have models for AST and ALT increases above 50 % vs. control groups or for the prediction of bilirubinemia, moving away from a nonspecific, for example, "hepatotoxicity" classification. This approach would, in principle, also make the identification of sound mechanisms of action for the specific observed toxicities a bit easier to address.

6 Future Perspectives

6.1 SEND Model and Data Exchange with FDA

On December 18, 2014, FDA issued the binding guidance titled "Providing Regulatory Submissions In Electronic Format—Standardized Study Data" [66] that requires Investigational New Drug (IND), New Drug Application (NDA), Abbreviated New Drug Application (ANDA), and Biologics License Application (BLA) submissions to be made in a standardized electronic format. The Clinical Data Interchange Standards Consortium (CDISC) Standard for Exchange of Nonclinical Data (SEND) is intended to guide the structure and format of standard nonclinical datasets for interchange between sponsors and contract research organizations (CROs) and for submission to the US FDA.

The current version of the SEND Implementation Guide (SENDIG v.3.0) is designed to support single-dose general toxicology, repeat-dose general toxicology, and carcinogenicity studies.

The guidance requires submission of nonclinical safety studies in SEND format for the study types currently supported. In the near future, the standard will be expanded to include additional study types, such as safety pharmacology (cardiovascular and respiratory) and developmental and reproductive toxicology, which will also be required.

The guidance further stipulates that published FDA-specific SEND validation rules will be enforced for all submitted datasets. The agency may refuse to file (for NDAs and BLAs) or refuse to receive (for ANDAs) an electronic submission that does not have study data in conformance to the required standards.

Under the guidance, supported studies (included in NDA, ANDA, and certain BLA submissions) starting after December 18, 2016, must be submitted in SEND.

For IND submissions supported studies starting after December 18, 2017 must be submitted in SEND.

Currently nonclinical safety data is provided as tabulated data within PDF study reports. Original electronic data, generated in-house, is normally stored on the originating LIMS systems until it is archived. In the case of CRO studies, original electronic data is typically not made available unless explicitly requested by the sponsor. The FDA now requires that, in addition to the PDF reports, the original electronic data also be submitted in SEND format.

While it is possible to build a SEND dataset manually, the process is labor intensive, error prone, and very difficult to validate. Given the fact that data comprising a study may come from multiple data sources, the challenge becomes unworkable.

An automated or semiautomated computerized system that can accurately and consistently transform original non-SEND data from multiple sources to the SEND standard and validate SEND

data following published rules is required. Oversight, tools, and processes for ensuring that source datasets are collected, curated, transformed to SEND, and made available for submission in an effective manner are also required.

Currently, FDA pharm/tox reviewers analyze the submitted study reports by manually extracting the tabulated data contained in the appendices of the PDF documents and loading them into any number of tools they see fit for visualizing and reviewing it. This first step is labor intensive and time consuming.

With the recently issued guidance for e-submissions, FDA reviewers have the opportunity to receive the study data directly in the appropriate format into one single platform called Nonclinical Information Management System (NIMS). FDA will use NIMS also to visualize the data, run their analyses, and draw their conclusions on the studies under review.

This approach will allow FDA reviewers to save time on data curation and formatting aspects and free resources for more in-depth scientific analyses, also leveraging the large amount of information and knowledge that NIMS will be capturing over the coming years.

Since clinical data is also electronically exchanged via standardized models (www.cdisc.org), it can be expected that one day, clinical and nonclinical data will be integrated under one single platform, which would represent a significant milestone in translational medicine arena.

6.2 Data Sharing Initiatives

Analysis of reasons for previous failures and exploitation of them should help in improving the efficiency of clinical development of new drugs and their safety profiles. So far, preclinical study reports have been rarely stored in a format that supports data mining or statistical analysis. Some pharmaceutical companies have realized these hidden treasures in their archives and started internal work to improve retrievability of their report data. It would clearly be of benefit to the whole industry to analyze these data across multiple companies in order to expand the chemical and biological space. However, extracting these data from the reports and building such a database requires considerable investment. Recent advances achieved in international initiatives, including IMI's eTOX project, have shown that sharing of preclinical data, both private and public, is achievable through the combination of legal (IP), IT, and honest broker concepts ([3, 67]; *see* Fig. 11).

The eTOX project aims to collect, extract, and organize preclinical safety data from pharmaceutical industry legacy study reports and publically available toxicology data into a searchable database to facilitate data mining and the development of innovative in silico models and software tools to predict potential safety liabilities of small molecules. The eTOX consortium consists of 13 pharmaceutical companies, 11 academic institutions, and 6 SMEs

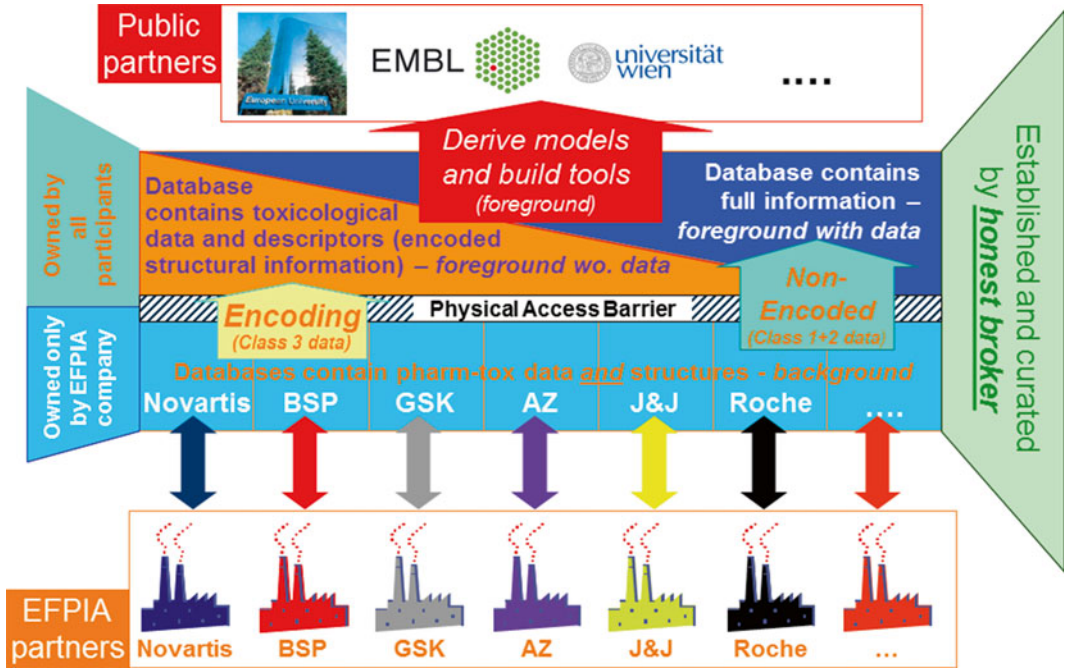


Fig. 11 The overall setup of the eTOX project

working together under the sponsorship of the Innovative Medicines Initiative (IMI) since 2010. The participating partners embrace expert knowledge in computational modeling, toxicology, pathology, and database design, liaising within the project in an integrative working environment.

After establishing an effective data sharing intellectual property (IP) protection within an "honest broker" approach (*see* Fig. 11), the project was able to compile a unique, well-curated dataset of currently more than 6,000 study reports, corresponding to ca. 1800 test compounds. The concept to divide the results from the legacy reports of the pharmaceutical companies in different "confidentiality classes" was fundamental to facilitate data sharing and overcome IP and legal hurdles. Public data (class 1) are accessible to the public on request, nonconfidential data (class 2) are open for eTOX consortium members, confidential data (class 3) are only accessible within the consortium with an additional secrecy agreement, and private data (class 4) are only for EFPIA data owners, but can be shared for model generation on request.

Treatment-related findings have been classified within the database, reflecting the interpreted study outcome of every report. A suite of ontologies, built through OntoBrowser now released by eTOX to the public domain, enables the user to directly compare observed effects or toxicities of chemically similar structures (read-across).

A new in silico tool—eTOXsys—has been developed with a single user interface, which manages search queries on the high-quality preclinical database and organizes requests to a steadily growing collection of independent prediction models. Aspects of IP rights for data sharing, definition of ontologies, design of database structure, development of in silico models, data analysis, validation, and sustainability are key aspects of the eTOX project.

References

1. Muller L, Alexander B, Christoph F, Wolfgang M, Axel P (2008) Strategies for using computational toxicology methods in pharmaceutical R&D. In: Ekins S (ed) Computational toxicology: risk assessment for pharmaceutical and environmental chemicals. Wiley, Hoboken, NY, pp 545–579
2. Muster WG et al (2008) Computational toxicology in drug development. Drug Discov Today 13(7–8):303–310
3. Cases M et al (2014) The eTOX data-sharing project to advance in silico drug-induced toxicity prediction. Int J Mol Sci 15(11): 21136–21154
4. Kavlock R (2009) The future of toxicity testing—the NRC vision and the EPA's ToxCast program national center for computational toxicology. Neurotoxicol Teratol 31(4):237–237
5. Kohonen P et al (2013) The ToxBank Data Warehouse: supporting the replacement of in vivo repeated dose systemic toxicity testing. Mol Inform 32(1):47–63
6. Arrowsmith J (2011) Trial watch: phase III and submission failures: 2007–2010. Nat Rev Drug Discov 10:87
7. Arrowsmith J, Miller P (2013) Trial watch: phase II and Phase III attrition rates 2011–2012. Nat Rev Drug Discov 12:569
8. Arrowsmith J (2011) Trial watch: phase II failures: 2008–2010. Nat Rev Drug Discov 10:328–329
9. Hillebrecht A et al (2011) Comparative evaluation of in silico systems for Ames test mutagenicity prediction: scope and limitations. Chem Res Toxicol 24:843–854
10. Sanderson DM, Earnshaw CG (1991) Computer prediction of possible toxic action from chemical structure; the DEREK system. Hum Exp Toxicol 10:261–273
11. Greene N, Judson PN, Langowski JJ, Marchant CA (1999) Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. SAR QSAR Environ Res 10:299–314
12. Judson PN (2006) Using computer reasoning about qualitative and quantitative information to predict metabolism and toxicity. In: Testa B, Kramer SD, Wunderli-Allespach H, Volkers G (eds) Pharmacokinetic profiling in drug research: biological, physicochemical, and computational strategies. Wiley, New York, pp 183–215
13. Derek Nexus (2015) http://www.lhasalimited.org/products/derek-nexus.htm
14. ToxTree version 2.6.6 (2015) https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/qsar_tools/toxtree
15. CASE Ultra version 1.5.2.0 (2015) http://www.multicase.com/case-ultra
16. Leadscope Expert Alerts version 3.2.4-1 (2015) http://www.leadscope.com/expert_alerts/
17. Limited L (2015) Derek Nexus: negative predictions for bacterial mutagenicity. http://www.lhasalimited.org/products/negative-predictions-for-bacterial-mutagenicity.htm
18. Pavan M, Worth AP (2008) Publicly-accessible QSAR software tools developed by the Joint Research Centre. SAR QSAR Environ Res 19:785–799
19. Benigni R, Bossa C (2008) Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. Mutat Res 659:248–261
20. Leadscope® Genetox Expert Alerts White paper (2014) http://www.leadscope.com/white_papers/Leadscope_alerts_white_paper.pdf
21. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inform 29:476–488
22. Sarah Nexus (2015) http://www.lhasalimited.org/products/sarah-nexus.htm
23. Leasdcope Model Appliers (2015) http://www.leadscope.com/model_appliers/
24. van Leeuwen K, Schultz TW, Henry T, Diderich B, Veith GD (2009) Using chemical categories to fill data gaps in hazard assessment. SAR QSAR Environ Res 20:207–220
25. Bioclipse (2015) http://www.bioclipse.net/
26. Prous Institute Symmetry (2015) http://symmetry.prousresearch.com/about-symmetry/

27. Hanser T, Barber C, Rosser E, Vessey JD, Webb SJ, Werner S (2014) Self organising hypothesis networks: a new approach for representing and structuring SAR knowledge. J Cheminform 6:21
28. Sarah Nexus Methodology (2015) http://www.lhasalimited.org/products/methodology-confidence-and-interpretation-of-predictions.htm
29. Klopman G (1992) A hierarchical computer automated structure evaluation program. Quant Struct Act Relat 11:176–184
30. Klopman G (1984) Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. J Am Chem Soc 106:7315–7321
31. Chakravarti SK, Saiakhov RD, Klopman G (2012) Optimizing predictive performance of CASE ultra expert system models using the applicability domains of individual toxicity alerts. J Chem Inf Model 52:2609–2618
32. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(379–423):379
33. The OECD QSAR Toolbox (2015) http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm
34. OECD Toolbox guidance document (2015)
35. OpenTox (2015) http://www.opentox.org/
36. Kazius J, McGuire R, Bursi R (2005) Derivation and validation of toxicophores for mutagenicity prediction. J Med Chem 48:312–320
37. Kuhn T, Willighangen EL, Zielesny A, Steinbeck C (2010) CDK-Taverna: an open workflow environment for cheminformatics. BMC Bioinformatics 11:159
38. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. Curr Pharm Des 12:2111–2120
39. Ekins S (2014) Progress in computational toxicology. J Pharmacol Toxicol Methods 69:115–140
40. Cheng A, Li W, Zhou Y, Shen J, Wu Z, Liu G et al (2012) admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. J Chem Inform Model 52:3099–3105
41. admetSAR (2015) http://lmmd.ecust.edu.cn:8000/
42. Valencia A, Prous J, Mora O, Sadrieh N, Valerio LG Jr (2013) A novel QSAR model of Salmonella mutagenicity and its application in the safety assessment of drug impurities. Toxicol Appl Pharmacol 273(3):427–434
43. Guideline E (2006) Guideline on the limits of genotoxic impurities. http://www.emea.europa.eu/pdfs/human/swp/519902en.pdf
44. Brigo A, Müller L (2011) Development of the threshold of toxicological concern concept and its relationship to duration of exposure. In: Teasdale A (ed) Genotoxic impurities. Wiley, New York, pp 27–63
45. ICH, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) (1997) Topic Q3C. Impurities: residual solvents
46. Kroes R, Renwick AG, Cheesemann M, Kleiner J, Mangelsdorf I, Piersma A, Schilter B, Schlatter J, van Schothorst F, Vos JG, Wurtzen G (2004) Structure-based thresholds of toxicological concern (TTC): guidance for application to substances present at low levels in the diet. Food Chem Toxicol 42:65–83
47. ICH, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) (2002) Topic Q3A(R). Impurities testing guideline: impurities in new drug products (Revision)
48. ICH, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) (2002) Topic Q3A(R). Impurities testing guideline: impurities in new drug substances (revision)
49. Müller L, Mauthe RJ, Riley CM, Andino MM, Antonis DD, Beels C, DeGeorge J, De Knaep AG, Ellison D, Fagerland JA, Frank R, Fritschel B, Galloway S, Harpur E, Humfrey CD, Jacks AS, Jagota N, Mackinnon J, Mohan G, Ness DK, O'Donovan MR, Smith MD, Vudathala G, Yotti L (2006) A rationale for determining, testing, and controlling specific impurities in pharmaceuticals that possess potential for genotoxicity. Regul Toxicol Pharmacol 44:198–211
50. Kasper P, Müller L (2015) Genotoxic impurities in pharmaceuticals. In: Graziano MJ, Jacobson-Kram D (eds) Genotoxicity and carcinogenicity testing of pharmaceuticals. Springer, Switzerland
51. M7 I. Assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenicity risk (2014)
52. Sutter A, Amberg A, Boyer S, Brigo A, Contrera JF, Custer LL, Dobo KL, Gervais V, Glowienke S, van Gompel J, Greene N, Muster W, Nicolette J, Reddy MV, Thybaud V, Vock E, White AT, Müller L (2013) Use of in silico systems and expert knowledge for structure-based assessment of potentially mutagenic

impurities. Regul Toxicol Pharmacol 67(1): 39–52

53. Brigo A, Muster W, Singer T (2015) Comparative assessment of several in silico systems and models to predict the outcome of the Ames mutagenicity assay. In: Toxicology SO (ed) Society of Toxicology Annual Meeting 2015. San Diego, California, USA
54. Ames BN, Durston WE, Yamasaki E, Lee FD (1973) Carcinogens are mutagens: a simple test system combining liver homogenate for activation and bacteria for detection. Proc Natl Acad Sci U S A 70:2281–2285
55. Escobar PA, Kemper RA, Tarca J, Nicolette J, Kenyon M, Glowienke S, Sawant SG, Christensen J, Johnson TE, McKnight C, Ward G, Galloway SM, Custer L, Gocke E, O'Donovan MR, Braun K, Snyder RD, Mahadevan B (2013) Bacterial mutagenicity screening in the pharmaceutical industry. Mutat Res 752:99–118
56. Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, Steger-Hartmann T, Heinrich N, Müller KR (2009) Benchmark data set for in silico prediction of Ames mutagenicity. J Chem Inf Model 49:2077–2081
57. Reuters T (2015) Metacore—data-mining and pathway analysis http://thomsonreuters.com/en/products-services/pharma-life-sciences/pharmaceutical-research/metacore.html
58. Fischer H, et al (2001) Prediction of in vitro phospholipidosis of drugs by means of their amphiphilic properties. In: Rational approaches to drug design, p 286–289
59. Kruhlak NL et al (2007) Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. Adv Drug Deliv Rev 59(1):43–55
60. CompuDrug. MetabolExpert (2015) http://www.compudrug.com/metabolexpert
61. Discovery M (2015) MetaSite http://www.moldiscovery.com/software/metasite/
62. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem 57:4977–5010
63. Piatetsky-Shapiro G (2012) Big data hype (and reality). https://hbr.org/2012/10/big-data-hype-and-reality
64. Sciences PL (2015) http://www.point-crosslifesciences.com/
65. James LP, Mayeuy PR, Hinson JA (2003) Acetaminophen-induced hepatotoxicity. Drug Metab Dispos 31(12):1499–1506
66. FDA (2014) http://www.fda.gov/downloads/Drugs/Guidances/UCM292334.pdf
67. Briggs K et al (2012) Inroads to predict in vivo toxicology—an introduction to the eTOX project. Int J Mol Sci 13(3):3820–3846