Food for Thought ... on *In Silico* Methods in Toxicology

Thomas Hartung[1] *and Sebastian Hoffmann*[2]

[1]CAAT, Johns Hopkins University, Baltimore, USA, and University of Konstanz, Germany; [2]TÜV Rheinland BioTech GmbH, Germany

This series of articles has already addressed *in vitro* and *in vivo* tools (Hartung, 2007; Hartung, 2008b). With this contribution addressing computational toxicology we now try to complete considerations on the main methodological approaches in toxicology. However, we do still plan to address specific aspects, such as endpoints (omics, image technologies), high-throughput testing or physiology-based pharmaco-(toxico-)kinetic modelling (PBPK), in later issues. All of these aspects have major *in silico* components, which already shows how difficult it is to discuss *in silico* methods on their own. Indeed, their integration and interplay with *in vivo* and *in vitro* approaches is critical, at least in the way their development often depends critically on the input of either *in vitro* or *in vivo* data. This is a major difference to experimental approaches. An important consideration will thus be whether *in silico* methods are limited by the limitations of their input and whether we have any hope of overcoming their weaknesses or can only approximate them…

There are some excellent introductions to and reviews of computational toxicology (Durham and Pearl, 2001; van de Waterbeemd, 2002; Greene, 2002; Veith, 2004, Helma, 2005; Simon-Hettich et al., 2006; Kavlock et al., 2008; Merlot, 2008; Nigsch et al., 2009; Greene and Naven, 2009). In addition, the ex-ECB website (http://ecb.jrc.ec.europa.eu/qsar/), hosted by Andrew Worth and his team (chronically understaffed given the high expectations) who act as key promoters of computational toxicology, is an excellent resource. The same holds true for the US-based International QSAR Foundation (http://www.qsari.org/) around Gil Veith, non-profit research organization devoted solely to creating alternative methods for identifying chemical hazards without further laboratory testing, and Angelo Vedani's Biographics Laboratory 3R in Basel, Switzerland (see his article in this issue). As usual this article aims less at summarizing the state of the art than at feeding discussions and showing up the critical issues faced by a central element of this increasingly important approach to toxicology.

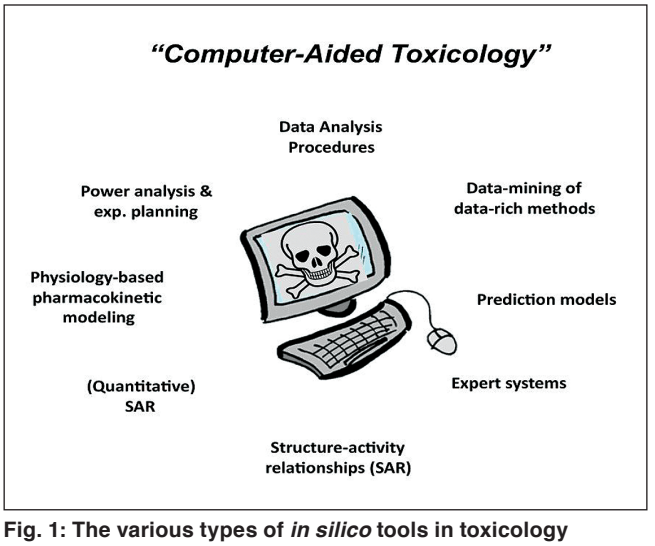

Fig. 1: The various types of *in silico* tools in toxicology

Consideration 1: *In silico* methodologies comprise a number of very heterogeneous approaches

If we define "*in silico* methodologies" as anything we can do with a computer in toxicology, there are indeed few tests that would not fall into this category, as most make use of computer-based planning and/or analysis. Thus, the first types of *in silico* approaches (Fig. 1) are certainly:

1. Planning of experiments and power analysis, i.e. tools to improve the design of our *in vivo* and *in vitro* experiments (Puopolo, 2004). This tool is at best not being fully exploited in toxicology. Especially for *in vitro* tests we have rarely seen that even the reproducibility of a test system has been systematically addressed in order to establish the number of replicates necessary to align biological relevance and statistical significance. However, this is a must for human clinical trials, and the number of replicates should be requested by animal use committees.
2. Data analysis procedures (DAP) – every experiment requires analysis: calculating a result, a number, a percentage, a threshold met and, usually, statistical analysis. All this is typically done *in silico*. There is astonishingly little guidance on this for those working *in vitro* or *in vivo*. In consequence, too often we see that analysis is crude, not substantiated by statistics and, even worse, that significance is mistaken for relevance. The result is often that each significant effect is reported as a result – a key problem in our field, where everything outside the norm is too easily interpreted as a threat. Together with a bias toward the publication of positive results, an impression

of risk and danger is created, not only for the lay audience but also for professionals exposed to the constant mantra of hazard and risk.

3. The most advanced (or the one most lacking) use of DAP is certainly DAP for omic and image analysis technologies. On the one hand, sophisticated data-mining techniques are available, but often the essential step leading to the derivation of a result, in our case the prediction of a hazard, is lacking. This becomes especially complex when physiological and biochemical knowledge on pathways and modes of toxic action are used in systems biology approaches, in our case the area of systems toxicology (Hartung and Leist, 2008; Leist et al., 2008). When does a derangement of a pathway become adverse, how do we interpret the concomitant activation of damaging and protective pathways?
4. Prediction models – these might be considered specific variants of DAP for alternative methods: Early on in the validation of alternative methods it was recognized that we need to translate the *in vitro* outcome into the outcome expected for *in vivo* tests. For example, cytotoxicity rates need to be converted into estimates of toxicity classifications, IC_{50} values need to predict whether significant organ toxicity is to be expected, etc. Prediction models are thus algorithms used to convert a result into an estimate of the result of the reference method. This often but not always requires computation.

The next series of *in silico* tools formalize what is called the "art of toxicology", i.e. ("eminence-based") expert knowledge. In the worst case "expert knowledge" only represents gut feelings of the evaluators. In expert systems they are at least systematically defined and applied by the software to compounds of interest:

5. Expert systems formulate rules to give guidance for decision-making. Rules such as structural alerts are often not explicitly formulated, compiled, combined or are too complex to be applied without a computer. The nature of these rules is often empirical, but they offer the advantage of challengeable definitions in contrast to most "expert judgements" in risk assessment. Expert systems have a major advantage over QSAR methods in that the prediction is related to a specific mechanism (Durham and Pearl, 2001).
6. This expert guidance can take the form of a structure activity relationship (SAR), which means that structural alerts are formalized, e.g. reactive groups such as aldehydes suggest mutagenicity, etc. An excellent example of this is the rule-based system for skin sensitization hazard developed by the German Federal Institute for Risk Assessment (BfR) (Gerner et al., 2004).

The most prominent *in silico* tools at the moment are the (quantitative) structure/activity relationships, i.e.

7. (Q)SAR. These aim to describe chemical structures by certain descriptors to correlate (typically biological) effects or properties. If there is a dominant physicochemical property this makes a lot of sense: for example, diffusion through membranes or accumulation from an aqueous environment both depend largely on lipophilicity, a property reasonably well predicted from structure by estimated octanol/water partition coefficients. However, often such relationships are based entirely on correlation, sometimes even trying various descriptors until a fit is found. Here, the problem of multiple testing arises: The more descriptors we try to include, the more likely we are to find one that correlates well, whether it makes biological sense or not. And a certain percentage will even show up again when "validating" with a second data set, even more if there is a partial connection, e.g. with a descriptor. For example, if lipophilicity is a contributing factor to a certain toxicity, it is likely to show up as a descriptor for a (Q)SAR, because it correlates with whether the substance can reach its target. This will result in some correlation, whether the other descriptors make sense or not. In addition, too often limited structural variety of substances feeding into the generation of a (Q)SAR will create an unrealistic predictivity if the limited applicability in turn is not understood.

The next type of *in silico* tool is the modelling tool:

8. Modelling tools, originating from computer-aided drug design (CADD) approaches to model a receptor and test the fit of new structures to this, are increasingly being used. Typical examples come from protein modelling, such as models of the estrogen receptor or various P-450 enzymes, where crystal structures are used to model the fit of the test compound into the reactive site of the receptor and to determine the likelihood of its triggering a response. These models are typically three-dimensional, but notably there are also 4D-models, which accommodate the induced fit of the substance on the receptor (Vedani et al., 2007) or pseudo-receptor models (Vedani et al., 2005). In case of the latter, in the absence of a real receptor, available data on a training set of ligands are used to emulate a hypothetic receptor to then be applied to new, untested substances. We might consider this a specific form of (Q)SAR, since the structures of training compounds are used for prediction. Thus the approach is clearly placed somewhere in-between (Q)SAR and receptor modelling.
9. Models of kinetics of substances aim to predict the fate of substances over time in the human organism, i.e. absorption, distribution, metabolism and excretion (ADME) (van de Waterbeemd and Gifford, 2003, Dearden, 2007). By 1997 it was reported (Kennedy, 1997) that poor ADME properties and toxicity accounted for 60% of failures of chemicals in the drug development process. These models are often based on models of blood flow, compartment sizes, organ characteristics, etc. on the one hand and physico-chemical properties of the test compound on the other hand, which can be either default values or actual measurements. Such measurements feeding into the models might in the simplest cases be lipophilicity (octanol/water distribution), solubility and charge, while more complex input could be metabolism or specific pathways triggered, etc.

These nine groups of *in silico* tools are neither exhaustive nor entirely distinct, as we have already seen in the last example. Table 1 lists the main approaches and some pertinent commercial products. In this article we shall especially deal with the

Tab. 1: Principal classes of *in silico* approaches and pertinent commercial offers

	Expert Systems, SAR	Statistical Systems, QSAR	Molecular Modelling	Tools
Basis	Expert knowledge, rules, decision trees	Data driven, molecular descriptors,	Protein structure, ligands	Biometry, data mining, data analysis
Variants	Structural alerts (2D) vs. pharmacophores (3D)	Correlation vs. artificial neural networks	Pseudo receptor modelling	
Prominent (commercial) products	Oncologic DEREK METEOR META HazardExpert COMPACT	TOPKAT MULTICASE* CASE LAZAR	VirtualToxLab	SAS SPSS jmp MatLab GraphPad

* includes an expert system component

"non-testing methods", a terminology used in the REACH legislation to describe ways to generate results to satisfy the data requirements without testing. This applies to approaches 5-8, but a clear focus will be on (Q)SAR as the most prominent technique. Only consideration 2 will address approaches 1-4, which can be understood as *in silico* tools used together with *in vitro* and *in vivo* methods. Approach 9 (PBPK) shall be addressed in a separate article soon, since it is a key technology necessary to link *in vitro* findings (or their *in silico* estimates) with dose/exposure on the whole organism level.

In toxicology and environmental health sciences in general, *in silico* tools boast a remarkably dichotomous following: While some colleagues uncritically embrace them, others are reluctant, sceptical and avoidant, such as Orrin Pilkey and Linda Pilkey-Jarvis in their book "Useless arithmetic" (2007). The cartoonist Scott Adams (The Dilbert Future) put it like this: *"There are many methods of predicting the future. For example, you can read horoscopes, tea leaves, tarot cards, or crystal balls. Collectively, these methods are known as 'nutty methods'. Or you can put well-researched facts into sophisticated computer models, more commonly referred to as 'a complete waste of time'."* Moderates recognising these methods as useful tools with limitations are slowly appearing. The authors count themselves as belonging to this species of method-critical users with high expectations as to the future of these approaches. This certainly has to do with our formal university training in mathematics, informatics and statistics, but also with impressions gained in some years of experience in the formal validation of methods, which intriguingly exposes the shortcomings of all models. We have proposed earlier (Hoffmann and Hartung, 2006) to translate evidence-based medicine to toxicology, which might be considered in a nutshell as the marriage of toxicology and biometry, i.e. applying objective and quantitative assessments to toxicology (Hartung, 2009a). We are aware of the problems this proposal is creating on both sides of the spectrum, for believers in *in silico* tools and for those "basing their toxicology on bloody evidence". The former fear that the basis of their modelling is de-valued, highlighting the limitations of the input of their modelling. The latter fear (consciously or unconsciously) that these authoritative tools will challenge or even destroy their traditional tools, endanger their comfort zone. And, not to forget, many have a fear of simply being overwhelmed by complex mathematics – a fear captured nicely by the statement: *"When used improperly, mathematics becomes a reason to accept absurdity."* (Pilkey and Pilkey-Jarvis, 2007, p. xiii).

Consideration 2: We must embrace more *in silico* tools in our current approaches

Even if we hesitate to use *in silico* methods as stand-alone non-test methods, we can profit from integrating more *in silico*, especially biometry, into our *in vivo* and *in vitro* approaches. Why so?

1. Because they allow us to standardise how we analyse and express results, enabling us to compare and share them. This is an objective and transparent way of handling results, reducing the reliance on individual expertise – something becoming increasingly important when more and more substances are being dealt with by more and more regulators in more and more economic and geographical regions.
2. Because we become less likely to fool ourselves into seeing what we want to see. We might even identify our prejudices, mere traditions, beliefs and fears that are interfering with the best possible judgement.
3. Because it will help us deal with complex situations and methodologies where non-formal analysis is not possible.
4. Because we need fair (biometry-based) assessments of traditional and new approaches to compare them and decide on their usefulness, their compatibility and transition to the new.
5. Because we need higher through-put to allow proper risk assessment of the multitude of substances, an exercise that is not feasible with traditional approaches as shown recently for REACH (Rovida and Hartung, 2009b, Hartung and Rovida, 2009). This can ideally be done with *in silico* approaches embedded in an over-all strategy.

What does this comprise? First of all, we need to establish the reproducibility of any tool. This sounds simple but it is by far not standard for tests used in toxicology. Secondly, this will allow power analysis to define which number of replicates will actually allow showing significance of a given difference (effect size). How often do we miss significance because group sizes

are simply too small… and (forgetting the mantra "absence of evidence is no evidence of absence") conclude that there is no effect? How often do we cherry-pick the results where by chance the variations in groups where small, suggesting significant differences, although group sizes and inherent variability of the model do not really allow the determination of significant results? The third need is to apply (appropriate) biometry. Too many tests (even those done according to test guidelines) have no statistics or use inappropriate methods. Fourthly, this requires a formalised DAP for each test. Often this will include formally establishing the limit of detection. This, fifthly, means moving the focus from significant to relevant: A significant effect is far from equivalent to a relevant effect (though relevant effects may be missed just as easily if the sample size is too small, i.e. the test is "under-powered").

Notably, these considerations hold true far beyond (regulatory) toxicology. Many areas of the life sciences could benefit from entering into such a discourse on appropriate biometrics (as well as proper documentation and other good practices). It is our strong belief that the spirit of quality assurance and validation, which were piloted in the field of toxicology and the validation of alternatives, will be instrumental in raising standards in biomedical research.

But we do not yet have all the tools we may need to support toxicology at our disposal. Appropriate biostatistical tools are pivotal for the development of prediction models and the data analysis of validation studies. In addition, application of sophisticated biostatistical methods can reduce the number of animals required for regulatory tests (Hoffmann et al., 2005).

For the mid-term the opportunity lies in moving away from stand-alone tests for each problem to integrated testing strategies (Hartung, 2009a; Nigsch et al., 2009; Greene and Naven, 2009). Here, *in silico* approaches can play a major role, not only as pieces of evidence used in a weight-of-evidence evaluation but as components and decision points in formalised strategies. Easy examples are prioritisations, which direct substances toward a certain test method or not. In the simplest case this can be seen as a screening approach, which is followed by a specific confirmatory one. This can also extend to the combination of two *in silico* methods, which are rendered one sensitive and one specific (McDowell and Jaworska, 2002). First examples show that the combined use of *in vitro* and *in silico* data actually improves predictions (Helma, 2005).

Consideration 3: Non-testing "*in silico*" methods are only emerging and are continuously being adjusted

Cell culture started more than 100 years ago and provided relevant contributions to toxicology from the 1960s onwards. Formal validation of *in vitro* systems for regulatory use has developed to maturity over the last 20 years. *In silico* toxicology started in the 70s and 80s[1] and is entering formal validation right now. Thus, *in silico* toxicology is at the stage *in vitro* tests were at twenty years ago. Given the enormous speed of development of all informational technologies and the fact that many experiences gained in the *in vitro* field can be adopted, a much quicker development can be anticipated.

Much development with regard to the robustness of modelling systems occurred in the 1990s (Durham and Pearl, 2001), but major developments were stimulated especially by the REACH legislation. Increasing consideration and experience from validation (Worth et al., 2004 a and b) in these years also serves as a sparring partner for model development.

The dual problem of the *in silico* technologies for regulatory use is that they are both "emerging" and "volatile". They are emerging because they are the new kids on the block; experience with them in (non-)regulatory use is short and they are emerging more quickly than other technologies because of the dynamics of the informatics revolutions and their ease of generation and application. In consequence the accumulated experiences are limited and often relate to earlier states of development, letting prejudices persist. It is part of their nature at the same time that they are "volatile", i.e. permanently changing. It would most probably be a waste of opportunities if we did not continuously incorporate new data to fine-tune our models. *"No predictive system should be considered as the 'finished article'. All expert systems require the capability of being updated as new knowledge becomes available."* (Cronin, 2000). However, any quality assessment such as validation requires freezing the methods in time. What is the value of an assessment when the method has been changed in the meantime? We cannot even assume that changing tests will make them better. In case of *in vitro* tests the effort of redoing the validation part might be considered reasonably possible, but this creates the risk of an inbred development, where the arbitrary selection of a validation set of compounds directs further developments. For *in silico* methods such re-validation might be more feasible, but who should keep track of the respective status and ensure the avoidance of inbred developments?

Much of the *in silico* technology now available was developed for pharmaceutical industry, where substances are developed for discrete targets, with certain physicochemical properties (allowing for example bioavailability), with well-understood kinetics and metabolism and comparatively unlimited resources. Very different from this, the safety assessment of environmental chemicals requires identifying unknown modes of action, usually in the absence of ADME data, for many diverse chemicals with rather limited resources.

Consideration 4: Once again, the problem of multiple testing...

The multiple test problem crops up repeatedly in toxicology: we test a lot and report only some results, typically the positive

[1] The authors are aware that already around 1860 correlations of structural modifications to functional properties of chemicals were reported; the first structure activity relationships can be traced back to Corwin Hansch 1963, who connected logP with activity. A broader movement, however, started only with the increasing availability of computational power in the 1970s.

(toxic) findings. If we do test for significance, we often forget the many tests we did to arrive there. For non-testing methods there are many multiple-testing traps: More than 2,500 descriptors for chemicals have been reported – Todeschini and Consonni (2000) alone list about 1800 descriptors. Of course, normally only a subset is considered for the development of a (Q)SAR, and there are recommendations that the number of datapoints should be greater than the number of descriptors evaluated. However, for any given data-set it is likely that we will arrive at a correlation when testing a substantial number of these. We have referred elsewhere (Hartung, 2009b) to the impressive example of demonstrating connections between zodiac birth signs and certain diseases, which works if only we test enough hypotheses. This appears to be of limited concern until now: 70% of all (Q)SAR include logP (Autti Poso, University of Kuopio, Finland, personal communication), showing that usually rather simple and physiologically meaningful descriptors are used. The desire to use "reversible descriptors" has been stated (Eriksson et al., 2003), which means that the descriptor can be translated into understandable chemical properties. However, there are examples (Yuan et al., 2007) of bizarre descriptors, which were identified from a multitude of ones tested such as "mean atomic Sanderson electronegativity (scales on carbon atom)", "lowest eigenvalue n. 1 of Burden matrix/weighted by atomic masses", "path/walk 2sRandic shape index" or "Narumi harmonic topological index" to name only the first few listed. Here a mechanistic interpretation appears to be most difficult.

There is also a substantial risk of using too many (Q)SAR – it is too easy to create a battery and run things through. It will be most important to require all results of (Q)SARs to be reported in a notification, e.g. a dossier on a substance for REACH, to avoid that only favourable results are included in notifications. This resembles somewhat the "intent to treat" requirement for clinical studies: once enrolled the patient is part of the study even if he drops out later. We need an "intent to test" attitude, i.e. to report the multiple modelling which took place, whatever the result was.

Consideration 5: Trash in, trash out? Or can *in silico* tools be better than the *in vivo* and *in vitro* data they are based on?

The development of *in silico* methods depends first of all on the availability and quality of usually *in vivo* data (Hartung and Daston, 2009) and other sources for modelling (Fig. 2) for the respective endpoint (principle "trash in – trash out" or "GIGO – garbage in – garbage out"). In principle, this is exactly the same problem as is faced in the validation of *in vitro* systems (Hartung, 2007; Hartung, 2008a). However, while for *in vitro* tests only 20 to 100 substances are required for validation, a much larger training set of data is necessary for *in silico* methods. In many cases this is not available, especially since many animal experiments are ill-defined and carried out in different variants which cannot be compared. Our own experience in one case might illustrate the problem: We chose acute fish toxicity, which is considered one of the most promising areas for QSAR because of the simplicity of the toxic mechanism. We used the

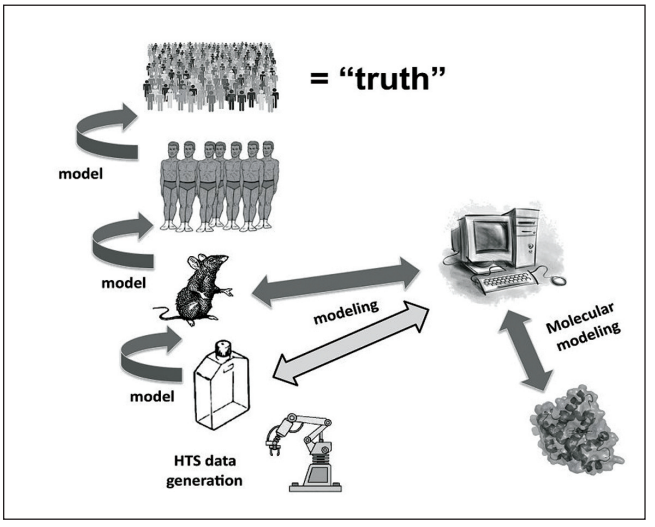

Fig. 2: The model of the model of the model... *in vitro* models animals models models human volunteer data models population.

probably best database available, i.e. the New Chemicals Database of ECB, which contains high-quality data (obtained over the last 25 years according to test guidelines and under the quality regimen of Good Laboratory Practice). Of 3100 substances in the end only about 150 could be used for modelling purposes (Lessigiarska, 2004) when stringent quality criteria were used. Since these tests are normally only performed once, the key factor of their reproducibility cannot even be assessed. Most recently, Hrovat et al. (2009) have shown an incredible variability of test results for exactly this area: For 44 compounds they found at least 10 data entries with variability exceeding several orders of magnitude. Of the analysed 4,654 test reports, 67% provided no information on fish life stage used, no information was provided on water temperature, hardness and pH in 20%, 48% and 41% of the reports, respectively. It is difficult to imagine how any computational approach should align this.

Data availability also requires that there is a certain homogeneity and absence of influential outliers as well as a span of the chemical domain of interest (Eriksson et al., 2003) – not easy to test as long as the definition of the chemical domain only becomes clear after the modelling or to correct if the dataset does not fulfil this requirement. This includes the problem of optimally defining the chemical space beforehand, when the type and combination of descriptors to be found useful for the modelling are not yet known. This problem will be smaller when more data spanning a larger chemical space are available.

Without doubt there are large reservoirs of data within pharmaceutical companies, but it is likely that they represent largely more or less closely defined series of compounds not covering broad regions of chemical space (Dearden, 2007). An analysis of the World Drug Index led to the famous Lipinski's rule of five (van de Waterbeemd and Gifford, 2003): most (oral) drugs have a molecular mass below 500 daltons, a calculated octanol/water partition coefficient <5, a number of hydrogen-bond donors <5 and a number of hydrogen-bond acceptors <10. This

characterises the chemical space in which most drug candidates for which data were generated in industry lie. This can certainly only overlap with the world of industrial chemicals to a very limited degree.

Another aspect is that, obviously, standard *in silico* models have similar or even stronger limitations than *in vitro* methods, as they do not reflect metabolism of compounds and their contribution to biological activity. However, exactly here specific *in silico* tools can come in to predict metabolites and subject these to further analysis. Such systems are desperately needed (Nigsch et al., 2009).

But, *in silico* predictions can be better than their input – if we are dealing with random errors and not systematic errors. In principle we are averaging the input from all data-points. Occasional errors will diminish with regard to their influence the more data is entered as input. However, any systematic error, such as species differences or abnormal reactions of chemical classes, cannot be sorted out this way.

QSAR are typically based on (multiple) linear correlations of descriptors of chemicophysical properties of substances. Linear correlations, however, are extremely rare in biological test systems. Given alone the complexity of uptake, distribution, metabolism and excretion of substances, it is clear that we cannot expect quantitative estimates (potency of toxicants) from linear correlations. Non-linear correlations for QSAR, however, require a much larger collection of good data. The problem would be less serious for mechanistically based QSAR (Netzeva et al., 2005), but as the paper states *"There are relatively few regulatory endpoints for which 'mechanistic QSAR' have been proposed, due to gaps in our understanding of underlying mechanisms of action and the scarcity of high-quality data sets suitable for hypothesis testing."*

Another limitation on the quality of *in silico* tools is the need for structural similarities for extrapolation. The confidence in an *in vitro* system is usually based not only on correct results in the validation exercise but also on the fact that the pathophysiology of the human health effect is modelled *in vitro*. This is not typically the case *in silico*. Therefore, it is much more important here that very similar substances are included in the training and validation of the model. This means that these substances form an "applicability domain" for new, similar compounds. The chemical universe of more than 100,000 commercial substances is difficult to cover. Measures for similarity of substances are only emerging (Nikolova and Jaworska, 2003; Netzeva et al., 2005).

A problem that should not be underestimated is the annotation of chemicals. There are many ways to name and describe a chemical – the (automatic) retrieval of biological data requires some efforts, also called "Toxico-Cheminformatics" (Kavlok et al., 2008) *"to integrate the disparate and largely textual information available on the toxicology and biological activity of chemicals"*. Excellent reviews on the ongoing annotation activities and remaining needs are given by (Richard et al., 2006 and 2008).

However, not necessarily (Q)SAR have to model existing data. Gil Veith has considered this "test now – model later" perception as a primary barrier for progress and belittles it as *"tangential research after-thought to the major programs developing testing methods and molecular biology"*. Much progress can be made if datasets are systematically produced for the purpose of modelling and the impact of the approach comes from producing guidance on what to test in the future. It might well be that the measure of "full replacement" is simply not adequate. *"The goal of QSAR is not to produce a series of models to be used in place of laboratory tests, but rather to improve both the design and strategic use of test methods."* (Veith, 2004).

Consideration 6: The validation dilemma

Validation can be seen as the process of building up confidence in (test) methods for regulators based on scientific evidence. These in the end have to take responsibility by agreeing on regulatory implementation. Over the last decade it has been possible to convince regulators (to different extents) of the merits of *in vitro* tests. Still, they are often only accepted for the identification of hazards, while the confirmation of negatives in animals is requested. Nevertheless, several alternative methods have made it to international acceptance over the last decade.

In case of *in silico* tools, especially those considered as non-testing methods to substitute for testing, a similar process of trust building will be necessary. When a review on *in silico* methods (Dearden, 2007) starts quoting Aristotle: *"It is the mark of an instructed mind to rest easy with the degree of precision which the nature of the subject permits, and not to seek an exactness where only an approximation of the truth is possible"*, this is counterproductive: Sure, we need to stay realistic about the limitations of any method, but we also need to ask whether the method is fit for its purpose. This holds especially true where other methods exist or where it is better to express the need for new approaches than to create the illusion of giving an answer. As much as we desire *in silico* models to work, we must be careful not to push them beyond their capabilities by lowering standards, such as by requesting "valid" instead of "validated" tests in the REACH legislation. If the experts are not clear on the difference between these two concepts, how should others understand it? Already in 1971 J. H. Chessire and A. J. Surrey remarked: *"that because of mathematical power of the computer, predictions of computer models tended to become 'imbued with a spurious accuracy transcending the assumptions on which they are based. Even if the modeller is aware of the limitations of the model and does not have a messianic faith in its predictions, the layman and the policymakers are usually incapable of challenging the computer predictions.'"* (Pilkey and Pilkey-Jarvis, 2007, p. 186).

We are not aware of any internationally accepted *in silico* alternative in the sense of a full replacement for all testing of a certain hazard. Regulatory use so far has been limited mainly to the prioritisation of test demands or to filling data gaps in some specific cases, especially for low risk chemicals. US EPA is responsible according to the Toxic Substances Control Act for assessing risks of new chemicals before they are marketed; initial screening is done using (Q)SAR models to find out whether more thorough assessments are needed (National Research Council, 2007). Notably, about 2,000 sub-

stances have to be evaluated per year, prompting the use of *in silico* tools. Some data collection schemes (such as the US EPA high-production volume chemical program) make use of non-testing *in silico* methods but usually not with regulatory consequences, i.e. *in silico* results are typically reported and accepted when they fit the expectations for the substance anyway. REACH for the first time gives *in silico* methods, and especially (Q)SAR, a broad room. This can only be understood when one knows that some of the persons involved in drafting the legislation where strong proponents of this approach. The subsequent political discussion of the Commission draft by Council and Parliament left these "technical issues" astonishingly untouched. (Q)SAR largely escaped the validation paradigm in the legislation, only asking for "valid" but not for validated methods, and in contrast to the *in vitro* approaches, the legislation makes no distinction between positive and negative results. The situation that real testing is more challenged than mere calculation will need to be resolved. Classification of substances (especially for existing high-production volume substances with their respective market value) based only on computational toxicology is, however, at this stage highly unlikely, especially since regulatory implementation is a consensus process. Broad waiving of testing for REACH because of negative *in silico* results is also rather unlikely, since not even validated cell systems are generally accepted for this. The most likely use in the mid-term will be the intelligent combination of *in vitro*, *in silico* and *in vivo* information.

Today, an impressive discrepancy exists between studies employing external evaluations, such as the Predictive Toxicology Challenge (PTC), and internal validation results: For the PTC a training set of 509 compounds from the US National Toxicology Program (NTP) with results for carcinogenic effects was used (Helma and Kramer, 2003). 185 substances with data from US FDA were used as a test set. 14 groups submitted 111 models, but only five were better than random at a significance level of p=0.05, with accuracies of predictions between 25 and 79% (Toivonen et al., 2003). Two previous comparative exercises by NTP had challenged models with 44 and 30 chemicals prospectively, i.e. with chemicals which were to be tested only (Benigni and Giuliani, 2003). The accuracy of *in silico* predictions in the first attempt was in the range of 50-65%, while the biological approaches attained 75%. The results in the second attempt (Benigni and Zito, 2004) ranged from 25 to 64%. In remarkable contrast, mere internal validations can show results of 75-89% predictivity (Matthews et al., 2006) for carcinogenicity and >80% for reproductive toxicity (Matthews et al., 2007), considered one of the most difficult areas for *in silico* predictions (Julien et al., 2004; Bremer et al., 2007).

A key question is therefore whether validation can be performed using the training dataset (leaving some data out or permutating the portion of data left out, i.e. cross-validation) or whether a challenge with a new dataset (external validation) is necessary? A first systematic investigation showed that, in general, there is no relationship between internal and external predictivity (Kubinyi et al., 1998): high internal predictivity may result in low external predictivity and vice versa. This effect, now called the "Kubinyi paradox" (Kubinyi, 2004), was also observed in other QSAR studies (Golbraikh and Tropsha, 2002), which show that in the commonly used leave-one-out cross-validation no correlation exists between the predictive ability in the training set and test set, especially when the number of descriptors considered is high relative to the number of compounds in the training set. In a retrospective investigation of about 40 different 3D QSAR models (Doweyko, 2004) similar results were obtained. Kubinyi recommends returning to the recommendations of Topliss and Costello as well as Unger and Hansch (1972 and 1973) to include only reasonable variables, selected from small numbers of variables, and to generate only models that have a sound biophysical background. This is the old misunderstanding of correlation vs. causality: The number of storks declining in parallel with the number of babies is not a proof of causality. Thus, the hypothesis to be tested should be rational and not empirical, which leads us also to the question of prior knowledge. *"It is often a valuable exercise in defining what we know about something and the basis of that belief."* (McDowell and Jaworska, 2002) calling for a Bayesian analysis of QSAR. Topliss and Costello (1972) show very nicely the risk of chance correlations when too many variables are tried on a limited dataset. Taken together, it is becoming evident that we need to validate with an external dataset (Hawkins, 2004; Worth et al., 2004 a and b; Helma, 2005); this requirement has been incorporated into the Setubal principles for QSAR validation.

A probably underestimated problem is that of errors introduced in computer programming: *"The potential for computer code error is vast, and it is very difficult to evaluate."* (Pilkey and Pilkey-Jarvis, 2007, p.26). Similarly, errors in chemical annotation might be more frequent than generally assumed: Ann Richard, US EPA, reported (personal communication) a 15% error rate of SMILES strings versus CAS numbers and a 1-2% error rate of CAS numbers in Chemfinder, all routine tools for *in silico* approaches. In conclusion, it is most important to know and quality-control the database used for model development and validation.

To date, positive experiences (>70% correct predictions, notably mostly not validated with external datasets) were reported mainly for mutagenicity, sensitisation and aquatic toxicity, i.e. areas with relatively well understood mechanisms (Simon-Hettich et al., 2006), not for complex/multiple endpoints. Hepatotoxicity, neurotoxicity and developmental toxicity cannot be accurately predicted with *in silico* models (Merlot, 2008). Here, the perspective lies in breaking down complex endpoints into different steps or pathways (Merlot, 2008) with the common problem of how to validate these and put them together to make one prediction. Unfortunately, it is exactly these endpoints that drive animal use and costs for REACH (Rovida and Hartung, 2009; Hartung and Rovida, 2009). To wait for REACH to finally deliver the data needed to create the QSARs (Simon-Hettich et al., 2006) sounds a bit like "mustard after the meal".

Validation is the prerequisite for regulatory acceptance, i.e. it is requested by OECD for both *in vitro* and *in vivo* methods. There is no reason why standards should be different for *in silico* tools, which are used in the same area with the same

consequences, i.e. regulatory decisions. Our understanding as to how to actually validate these tools is increasing (Jaworska et al., 2003; Erriksson et al., 2003; Hartung et al., 2004; Worth et al. 2004 a and b; Helma 2005), especially for (Q)SAR, where the "Setubal principles" (Jaworska et al., 2003) have been formulated, which state that (Q)SAR should:

1. be associated with a defined endpoint of regulatory importance
2. take the form of an unambiguous algorithm
3. ideally, have a mechanistic basis
4. be accompanied by a definition of domain of applicability
5. be associated with a measure of goodness-of-fit
6. be assessed in terms of their predictive power by using data not used in the development of the model.

Especially points (2) and (6) are problematic for some commercial models, which do not necessarily share underlying algorithms and datasets. Similarly, (2) is difficult to meet for artificial neural network modelling (Devillers, 2008). Aspect (3) is more difficult to satisfy because few modes of action can be traced to molecular descriptors and same mode of action does not necessarily mean the same target of interference for the chemical in the biological system. At the same time, (4), (5) and (6) are technically under debate (Eriksson et al., 2003; Veith, 2004; Netzeva et al., 2005) and will likely need individual consideration for each and every (Q)SAR to be validated.

The Setubal principles led to the definition of respective OECD principles (http://www.oecd.org/dataoecd/33/37/37849783.pdf), which require:

1. a defined endpoint
2. an unambiguous algorithm
3. a defined domain of applicability
4. appropriate measures of goodness-of-fit, robustness and predictivity
5. a mechanistic interpretation, if possible.

Noteworthy, the requirement of "regulatory importance" but more importantly also that of "external validation" (see above discussion) was abandoned. Also, the term mechanistic "basis" was weakened to "interpretation".

Interestingly, beside this approach to validation, which very strongly follows that for *in vitro* methods, (Q)SAR can also be seen as a specific form of environmental computer models (National Research Council, 2007), where somewhat divergent concepts as to evaluation and validation have emerged, especially from US EPA (Fig. 3).

Consideration 7: Applicability of QSAR

Gil Veith (2004) pointed out clearly that *"we must embrace the facts that chemical structure is complex, the metabolites associated with a chemical represent a complex array of additional agents that can vary with species and target cells, and that toxicity pathways from molecular interactions to the adverse outcomes used in risk assessment are also complex. It is inconceivable that a QSAR based on simple equations and rules of thumb based on parent structures will be accepted by the scientific and regulatory community, and rightly so."* Often overlooked, many test problems in the "real world" do not even deal with pure parent substances. A structure/activity relationship can only be meaningful if the biological effect is exerted by a single compound and the accompanying contaminations, metabolites and degradation products are not important. This is most difficult for areas like allergic reactions (skin sensitisation), where femtogram quantities of compounds can in principle produce reactions not related to the main compound under study. In general, limits of 90 to 95% purity are set for the application of (Q)SAR, but what is the sense in areas such as mutagenicity/carcinogenicity, where we do not accept threshold concentrations? Practical experience shows that even in "good" databases there are limits as to the applicability of *in silico* approaches: For example, when addressing aquatic toxicity in the New Chemical Database, i.e. compiled from the harmonised notification of chemicals in Europe after 1981, it was found in the study quoted above (Lessigiarska, 2004) that of 2,900 substances including data on the respective endpoints only about 1,400 represented substances suitable for QSAR, because they require purity and exclude mixtures, salts and metal compounds. It must be assumed that the relative high production volume chemicals in REACH will include many such substances, since new chemicals are more likely to be synthetic compounds such as dyes. Therefore, it will have to be assessed for which proportion of chemicals under REACH QSAR are actually suitable for *in silico* assessment. A preliminary evaluation of 200 low and 200 high production volume chemicals falling under REACH, which were randomly selected confirmed this percentage: Only 54% of the chemicals qualified in both tonnage classes for *in silico* approaches with regard to purity, exclusion of mixtures, salts, metals, etc. (Martin Barrat and the authors, unpublished).

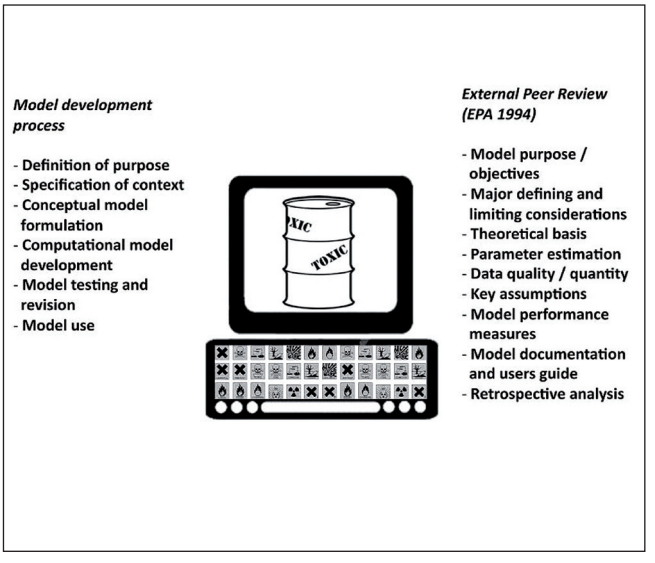

Fig. 3: Model development and peer-review steps suggested by the US National Academy of Sciences (National Research Council 2007)

A most interesting question is whether *in vitro* (high-through-put systems, HTS) models can deliver the database required for (Q)SAR development. Lombardo et al. (2003) and Dearden (2007) have argued that HTS data are generally not accurate enough for modelling purposes. This is not at all our own experience with robotised testing at ECVAM in collaboration with Maurice Wheelan and his Molecular Imaging group in our sister unit, where the variability of *in vitro* tests (here the Balb 3T3 cytotoxicity test) was considerably reduced compared to several laboratories performing it manually. The reason for this difference in perception is likely that Lombardo et al. and Dearden refer to HTS as it is conducted in pharmaceutical industry – to find the needle in the haystack: hundreds of thousands of substances, often not pure, partially decomposed, not controlled for solubility, are tested at one concentration only and without replicates – in comparison HTS for *in vitro* toxicology is typically done with concentration response curves, replicates as appropriate and quality control of the chemicals. This can indeed form the basis for modelling, and we might go as far as saying that the more simple *in vitro* systems with such large homogenous datasets qualify much better for *in silico* modelling than any *in vivo* data. To model a (validated) *in vitro* test might thus lead to "second generation *in silico* alternatives".

A key problem will be to move risk assessment away from a (pretended) black/white situation, where good and bad compounds are identified for any endpoint: *In silico* (like *in vitro*) methods will leave uncertainties ("grey areas") (Merlot, 2008). This is not really different from *in vivo* methods, but it was always too comfortable to neglect this there (Bottini and Hartung, 2009). The inherent biometrical analysis and unavoidable validation of the new tools will move us to acknowledge uncertainty. In the end, we will have to move to something like a "probabilistic risk assessment", where only a probability for a substance to exert a certain health hazard is given. This will change the way we perform risk assessments: Sorry, but we will not be able to close the books after a risk assessment is completed any longer. Instead, we will need to keep an eye on the true toxicological effects of the substances in use. For the moment, it is already some progress to introduce a grey zone, i.e. "we are not sure" – notably again: not only for *in silico* methods.

We should be clear: (Q)SAR and other *in silico* approaches are already used for regulatory purposes (Walker et al., 2002; Cronin et al., 2003; Gerner et al., 2004), rarely yet as a full replacement of testing requirements, but more commonly to identify testing needs and to prioritise. Their use is only going to expand, just as our expertise and computational power and databases are expanding.

An interesting consideration (Huynh et al., 2009) is that for the real application of *in silico* tools the "*in silico* specialist" must be groomed, someone who needs to take on the legal risks and responsibilities of a risk assessor. *"For this purpose, in silico technology should be subject to the same evaluation as clinical biology: standardization, reproducibility, precision, accuracy, detection limit, and so on."* However, before this, the full integration of modellers into project teams might already take us further, as lack of communication was identified as a major obstacle for the effective use of *in silico* predictions in practice (Merlot, 2008).

Consideration 8: *In silico* tools are in many respects a forerunner for the development of new toxicological tools

With the imprinting of some years spent in a validation body, the authors tend to look for the problems of any methodology. But we should also clearly state that the *in silico* technologies come with a much more open view as to their limitations than any other technology. Okay, some exaggerated promises are made here too, but selling is part of the business. Some aspects deserve specific appraisal:

- The concept of an applicability domain – we can use a model only for the test substances for which it is adequate. Sounds simple, but just because we can inject something into an animal or pipet it into a cell culture does not mean that it is applicable. We owe the area of *in silico* methodologies the concept of a rigorous definition of the applicability domain, which we introduced into ECVAM's modular approach to validation only in 2004 (Hartung et al., 2004).
- The continuous update of databases – the concept that models need to continuously accommodate new findings is not at all shared by *in vivo* and *in vitro* tools: public availability would already be a big step forward, but who will use the emerging information to refine the model itself?
- Central repositories of methods and guidance – the work done both at the former European Chemical Bureau and OECD to make toolboxes for QSAR available is unique in toxicology.
- Defined DAP – recall? DAP is a data analysis procedure. This makes the difference between a model and a test. Only with a standardised procedure to interpret data and deduce a result do we have a defined test. *In silico* tools do not exist without it, but the number of *in vivo* and *in vitro* models where "significant" effects (= something happened) are reported without interpreting their relevance is tremendous.
- Public domain character of many tools – the *in silico* field has many open source and public domain offers. This is a double-edged sword since it impairs business opportunities in this field, which can be a driving force (Bottini and Hartung, 2009). However, in general, this ease of availability and transparency helps implementation.

In conclusion, a new spirit is entering toxicology with the emerging *in silico* opportunities. Our science will benefit from it.

Final considerations

The *in silico* / QSAR field is facing more resistance in traditional toxicology than it deserves, though some is appropriate and necessary: *"As much as scepticism over QSAR comes from inappropriate use of QSAR for chemicals that elicit dif-*

ferent mechanisms as comes from the intentional and unintentional over-selling of the predicitive capabilities of QSAR." (Veith, 2004).

1. QSAR represent most promising complementary methods to achieve intelligent testing strategies; their use as standalone methods for regulatory purposes with broad application on a short term is unlikely.
2. Repeat-dose toxicities (chronic toxicity, reproductive toxicity, cancer) represent the largest challenge. Here, no *in silico* approaches are evident yet.
3. The fast development of QSAR requires the "sparring partner" of validation to coach developments.
4. The relative contribution of QSAR to the toxicological toolbox and especially REACH will depend on the (non)-availability of high-quality *in vivo* data, applicability to the substances (no mixtures, sufficient purity, no salts, no metal compounds, sufficient similar structure in the training set), successful validation, confidence of regulators and speed of regulatory implementation.
5. We can learn a lot from *in silico* methods in other areas of toxicology, especially a more rigorous biometric and self-critical approach to toxicological tool development.

So, where do we stand? *In silico* tools have a bright future in toxicology. They add the objectivity and the tools to appraise our toolbox. They help to combine various approaches in more intelligent ways than a battery of tests. They cannot be better than the science they are based on, "no model can overcome a series of bad assumptions." (Pilkey and Pilkey-Jarvis, 2007, p. 29). For any model (*in vivo*, *in vitro* or *in silico*), it is luck if a large part of the real world is reflected, and we will only know so after laborious validation. George Box is known for his statement *"All models are wrong, some models are useful"* (Box, 1979). We agree that it ultimately boils down to usefulness, as elegantly expressed by Gil Veith: *"The bottom line in the business world for scientific products is their usefulness, perhaps the most rigorous test that can be given to any research product."* We can help to improve their usefulness by integrating them into the toolbox of toxicology, estimating their usefulness by validation and demonstrating this usefulness finally by showing in comparison the limitations of current approaches.

References

Benigni, R. and Giuliani, A. (2003). Putting the predictive toxicology challenge into perspective: reflections on the results. *Bioinformatics 19*, 1194-1200.

Benigni, R. and Zito, R. (2004). The second national toxicology program comparative exercise on the prediction of rodent carcinogenicity: definitive results. *Mut. Res. 566*, 49-63.

Bottini, A. A. and Hartung, T. (2009). Food for thought… on economics of animal testing. *ALTEX 26*, 3-16.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer and G. N. Wilkinson (eds.), *Robustness in Statistics*, 202. New York: Academic Press.

Bremer, S., Pellizzer, C., Hoffmann, S. et al. (2007). The development of new concepts for assessing reproductive toxicity applicable to large scale toxicological programs. *Curr. Pharm. Des. 13*, 3047-3058.

Cronin, M. T. D. (2000). Computational methods for the prediction of drug toxicity. *Curr. Opin. Drug Disc. Develop. 3*, 292-297.

Cronin, M. T. D., Jaworska, J., Walker, J. D. et al. (2003). Use of QSAR in International decision-making frameworks to predict health effects of chemical substances. *Env. Health Persp., 111*, 1391-1401.

Dearden, J. C. (2007). In silico prediction of ADMET properties: how far have we come? *Expert Opin. Drug Metab. Toxicol. 3*, 635-639.

Devillers, J. (2008). Artificial neural network modeling in environmental toxicology. In D. S. Livingstone (ed.), *Artificial neuronal networks: methods and protocols*, 61-80. New York: Humana Press.

Doweyko, A. (2004). 3D-QSAR illusions. *J. Comput.-Aided Mol. Design 18*, 587-596.

Durham, S. K. and Pearl, G. M. (2001). Computational methods to predict drug safety liabilities. *Curr. Opin. Drug Disc. Develop. 4*, 110-115.

Eriksson, L., Jaworska, J., Worth, A. P. et al. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classifications- and regression-based QSARs. *Env. Health Persp. 111*, 1361-1375.

Gerner, I., Spielmann, H., Hoefer, T. et al. (2004). Regulatory use of (Q)SARs in toxicological hazard assessment strategies. *SAR and QSAR Env. Res. 15*, 359-366.

Golbraikh, A. and Tropsha, A. (2002). Beware of q^2. *J. Mol. Graphics Modelling 20,* 269-276.

Greene, N. and Naven, R. (2009). Early toxicity screening strategies. *Curr. Opin. Drug Disc. Develop. 12*, 90-97.

Greene, N. (2002). Computer systems for the prediction of toxicity: an update. *Adv. Drug Delivery rev. 54*, 417-431.

Hartung, T. (2009a). Food for thought… on evidence-based toxicology. *ALTEX 26*, 75-82.

Hartung, T. (2009b). Toxicology for the twenty-first century. *Nature 460*, 208-212.

Hartung, T. and Rovida, C. (2009). Chemical regulators have overreached. *Nature 460*, 1080-1081.

Hartung, T. and Daston, G. (2009). Are in vitro tests suitable for regulatory use? *Tox. Sci.,* in press.

Hartung, T. (2008a). Towards a new toxicology – evolution or revolution? *ATLA – Altern. Lab. Anim. 36*, 635-639.

Hartung, T. (2008b). Food for thought… on animal tests. *ALTEX 25*, 3-9.

Hartung, T. and Leist, M. (2008). Food for thought… on the evolution of toxicology and phasing out of animal testing. *ALTEX 25*, 91-96.

Hartung, T. (2007). Food for thought… on validation. *ALTEX 24*, 67-72.

Hartung, T., Bremer, S., Casati, S. et al. (2004). A Modular Approach to the ECVAM principles on test validity. *ATLA –*

Altern. Lab. Anim. 32, 467-472.
Hawkins, D. (2004). The problem of overfitting. *J. Chem. Inf. Comput. Sci. 44*, 1-12.
Helma, C. (2005). In silico predictive toxicology: the state-of-the-art and strategies to predict human health effects. *Curr. Opin. Drug Disc. Develop. 8*, 27-31.
Helma, C. and Kramer, S. (2003). A survey of the predictive toxicology challenge 2000-2001. *Bioinformatics 19*, 1179-1182.
Hoffmann, S. and Hartung, T. (2006). Towards an evidence-based toxicology. *Human Exp Toxicol 25*, 497-513.
Hoffmann, S., Luderitz-Puchel, U., Montag-Lessing, U. and Hartung, T. (2005). Optimisation of pyrogen testing in parenterals according to different pharmacopoeias by probabilistic modelling. *J. Endotoxin Res. 11*, 25-31.
Hrovat, M., Segner, H. and Jeram, S. (2009). Variability of in vivo fish acute toxicity data. Regulat. *Toxicol. Pharmacol., 54*, 294-300.
Huynh, L., Masereeuw, R., Friedberg, T. et al. (2009). In silico platform for xenobiotics ADME-T pharmacological properties modelling and prediction. Part I: beyond the reduction of animal use. *Drug Disc. Today 14*, 401-405.
Jaworska, J. S., Comber, M., Auer, C. and van Leeuwen, C. J. (2003). Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoint. *Env. Health Persp. 111*, 1358-1360.
Julien, E., Willhite, C. C., Richard, A. M. and DeSesso, J. M. (2004). Challenges in constructing statistically based structure-activity relationship models for developmental toxicity. *Birth Defects Res. 70*, 902-911.
Kavlock, R. J., Ankley, G., Blancato, J. et al. (2008). Computational toxicology – a state of the science mini review. *Toxicol. Sci. 103*, 14-27.
Kennedy, T. (1997). Managing the drug discovery/development interface. *Drug Discov. Today 2*, 436-444.
Kubinyi, H., Hamprecht, F.A. and Mietzner, T. (1998). Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem. 41*, 2553-2564.
Kubinyi, H. (2004). Validation and predictivity of QSAR models. http://www.kubinyi.de/istanbul-2004-manuscript.pdf
Leist, M., Hartung, T. and Nicotera, P. (2008). The dawning of a new age of toxicology. *ALTEX 25*, 103-114.
Lessigiarska, I., Wort, A. P., Sokoll-kluettgen, B. et al. (2004). QSAR investigation of a large data set for fish, algae and Daphnia toxicity. *SAR and QSAR Env. Res. 15*, 413-431.
Lombardo, F., Gifford, E. and Shalaeva, M. Y. (2003). In silico ADME prediction: data, models, facts and myths. *Mini Rev. Medicin. Chem. 3*, 861-875.
Matthews, E. J., Kruhlak, N. L., Cimino, M. C. et al. (2006). An analysis of genetic toxicity, reproductive and developmental toxicity and carcinogenicity data: II. Identification of genotoxicants, reprotoxicants, and carcinogens using in silico methods. *Reg. Toxicol. Pharmacol. 44*, 97-110.
Matthews, E. J., Kruhlak, N. L., Benz, D. R et al. (2007). A comprehensive model for reproductive and developmental toxicity hazard identification: II. Construction of QSAR models to predict activities of untested chemicals. *Regult. Tox. Pharmacol. 47*, 136-155.
McDowell, R. M. and Jaworska, J. S. (2002). Bayesian analysis and interference from QSAR predicitive model results. *SAR and QSAR Env. Res. 13*, 111-125.
Merlot, C. (2008). In silico methods for early toxicity assessment. *Curr. Opin. Drug Disc. Develop. 11*, 80-85.
National Research Council (2007). *Models in environmental regulatory decision making*. Washington: The National Academies Press.
Netzeva, T. I., Worth, A. P ., Aldenburg, T. et al. (2005). Current status of methods defining the applicability domain of (quantitative) structure-activity relationships – the report and recommendations of ECVAM workshop 52. *ATLA 33*, 1-19.
Nigsch, F., Macaluso, N. J. M., Mitchell, J. B. O. and Zmuidinavicius, D. (2009). Computational toxicology: an overview of the sources of data and of modelling methods. *Expert Opin. Drug Metab. Toxicol. 5*, 1-14.
Nikolova, N. and Jaworska, J. (2003). Approaches to measure chemical similarity – a review. *QSAR Comb. Sci. 22*, 1006-1026.
Pilkey, O. H. and Pilkey-Jarvis, L. (2007). *Useless arithmetic – why environmental scientists can't predict the future*. New York: Columbia University Press.
Puopolo, M. (2004). Biostatistical approaches to reducing the number of animals used in biomedical research. *Ann. Ist. Super. Sanita 40*, 157-163.
Richard, A. M., Gold, L. S. and Nicklaus, M. C. (2006). Chemical structure indexing of toxicity data on the internet: moving toward a flat world. *Curr. Opin. Drug Disc. Develop. 9*, 314-325.
Richard, A. M., Yang, C. and Judson, R. S. (2008). Toxicity data informatics: supporting a new paradigm for toxicity prediction. *Toxicol. Mech. Meth. 18*, 103-118.
Rovida, C., and Hartung T. (2009). Re-evaluation of animal numbers and costs for in vivo tests to accomplish REACH legislation requirements for chemicals. *ALTEX 26*, this issue.
Simon-Hettich, B., Rothfuss, A. and Steger-Hartmann, T. (2006). Use of computer-assisted prediction of toxic effects of chemical substances. *Toxicol. 22*, 156-162.
Todeschini, R. and Consonni, V. (2000). *Handbook of molecular properties*. Weinheim, Germany: Wiley-VCH Verlag.
Toivonen, H., Srinivasan, A., King, R. D., Kramer, S. and Helma, C. (2003). Statistical evaluation of the predictive toxicology challenge 2000-2001. *Bioinformatics 19*, 1183-1193.
Topliss, J. G. and Costello, R. J. (1972). Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem. 15*, 1066-1068.
Unger, S. H. and Hansch, C. (1973). On model building in structure-activity relationships. A reexamination of adren-

ergic blocking activity of β-halo-β-arylalkylamines. *J. Med. Chem. 16*, 745-749.

Van de Waterbeemd, H. (2002). High-throughput and in silico techniques in drug metabolism and pharmacokinetics. *Curr. Opin. Drug Disc. Developm. 5*, 33-43.

Van de Waterbeemd, H. and Gifford, E. (2003). ADMET in silico modelling: towards prediction paradise? *Nature Rev. Drug Disc. 2*, 192-204.

Vedani, A., Dobler, M. and Lill, M. A. (2005). In silico prediction of harmful effects triggered by drugs and chemicals. *Toxicology and Applied Pharmacology 207, Suppl. 1*, 398-407.

Vedani, A., Dobler, M., Spreafico, M. et al. (2007). VirtualToxLab – in silico prediction of theHartung toxic potential of drugs and environmental chemicals: evaluation status and internet access protocol. *ALTEX 24*, 153-161.

Veith, G. D. (2004). On the nature, evolution and future of quantitative structure-activity relationships (QSAR) in toxicology. *SAR and QSAR Env. Res. 15*, 323-330.

Walker, J. D., Carlesen, L., Hulzebos, E. and Simon-Hettich, B. (2002). Global government applications of analogues, SARs and QSARs to predict aquatic toxicity, chemical or physical properties, environmental fate and health effects of organic chemicals. *SAR and QSAR Env. Res. 13*, 607-616.

Worth, A. P., Hartung, T. and Van Leeuwen, C. J. (2004a). The role of the European centre for the validation of alternative methods (ECVAM) in the validation of (Q)SARs. *SAR QSAR Environ. Res. 15*, 345-358.

Worth, A. P., Van Leeuwen, C. J. and Hartung, T. (2004b). The prospects for using (Q)SARs in a changing political environment – high expectations and a key role for the European Commission's joint research centre. *SAR QSAR Environ. Res. 15*, 331-343.

Yuan, H., Wang, Y. and Cheng, Y. (2007). Local and global Quantitative Structure-Activity Relationship modeling and prediction for the baseline toxicity. *J. Chem. Inf. Model. 47*, 159-169.

Correspondence to

Prof. Thomas Hartung, MD, PhD
Johns Hopkins University
Bloomberg School of Public Health
Doerenkamp-Zbinden Chair for Evidence-based Toxicology
Center for Alternatives to Animal Testing (CAAT)
615 N. Wolfe St. W7035
Baltimore, MD, 21205, USA
e-mail: Thartung@jhsph.edu