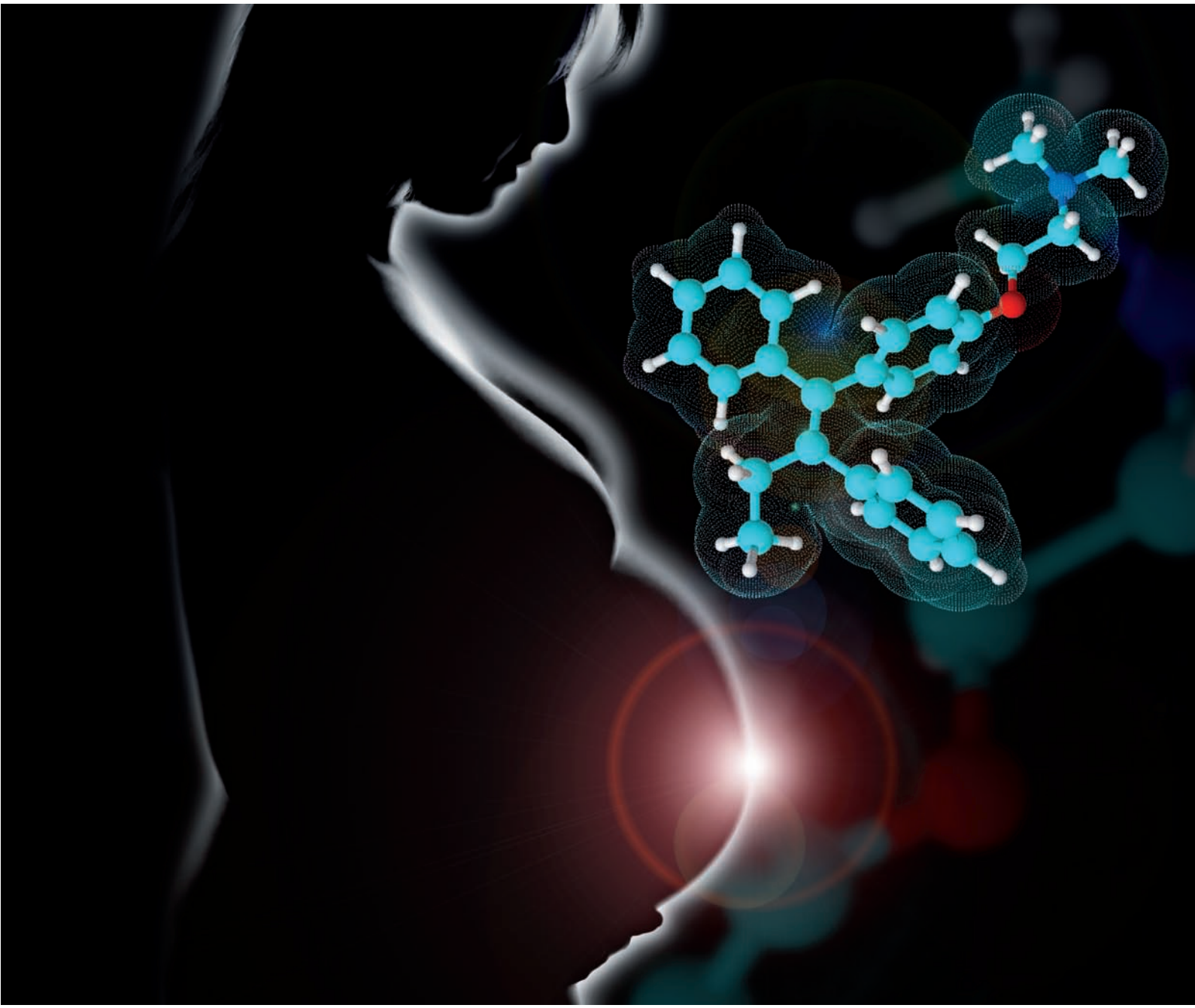
Chem Soc Rev

Chemical Society Reviews

www.rsc.org/chemsocrev

Volume 37 | Number 3 | March 2008 | Pages 433–612



ISSN 0306-0012

TUTORIAL REVIEW
Alessandra Roncaglioni and Emilio Benfenati
In silico-aided prediction of biological properties of chemicals: oestrogen receptor-mediated effects

CRITICAL REVIEW
Floris Chevallier and Florence Mongin
Functionalization of diazines and benzo derivatives through deprotonated intermediates

RSC Publishing

0306-0012(2008)37:3;1-A

In silico-aided prediction of biological properties of chemicals: oestrogen receptor-mediated effects

Alessandra Roncaglioni and Emilio Benfenati*

Received 17th September 2007
First published as an Advance Article on the web 8th October 2007
DOI: 10.1039/b616276m

In silico methods are a valid tool for analysing the properties of chemical compounds and interest in computational modelling techniques to predict the activity of chemicals is constantly growing. Many computational methods can be used to analyse the toxicity or biological activity of chemicals, particularly as regards their interactions with biological macromolecules (*e.g.* receptors) and other physico-chemical properties. An overview of these methods is provided in this *tutorial review*, with some examples of their application to predict oestrogen receptor (ER)-mediated effects. Nuclear receptors, particularly ER, have been studied with *in silico* tools since concern is growing about substances, called endocrine disrupters, that can interfere with hormone regulation. Molecular modelling techniques such as Quantitative Structure–Activity Relationships (QSAR), related methods like 3D-QSAR, and virtual docking have been used to investigate these phenomena and are described here. Implications about regulatory acceptance and use of these methods and the resulting models for identifying hazards and setting priorities are also addressed.

In silico tools as an alternative to animal testing

Interest in computer-aided methods for investigations in the biological field has increased significantly in recent years. Analogously to the expressions *in vivo* (referring to methods using animals) and *in vitro* (referring to methods using mainly cellular systems), the expression *in silico* has been introduced referring to silicon, as a metaphor for computers. *In silico* tools are becoming more accessible to researchers as their cost drops and the speed of computational calculation increases, so interest in their application can spread to a wide range of biological problems. Many different approaches can be listed within the *in silico* tools. Data mining techniques are frequently used to analyse biological data such as genomic or proteomic findings, for example. Docking studies involve a detailed modelling of the interactions between the ligands and the receptor. QSAR, instead, do not analyse the receptor, but only small molecules to characterize their properties, using information based on chemical structures.

A challenging application of these methods is modelling and characterizing the (bio)activity profiles of chemicals. Many studies have addressed, for example, ecotoxicity, human health, physico-chemical and Absorption Distribution

Laboratory of Environmental Toxicology and Chemistry, Istituto di Ricerche Farmacologiche ''Mario Negri'', via La Masa 19, 20156 Milan, Italy. E-mail: benfenati@marionegri.it; aroncaglioni@marionegri.it; Fax: +39 02 39014735; Fax: +39 02 39014735; Tel: +39 02 39014420 Tel: +39 02 39014595

Alessandra Roncaglioni

Alessandra Roncaglioni graduated in environmental science in 2001. Since 2001 she has been a researcher in the Laboratory of Environmental Chemistry and Toxicology, at the ''Mario Negri'' Institute. She attended a three years specialization course in pharmacological research (2001–2004) focusing her research on the use of computational chemistry methods applied to environmental and health issues. She is now completing her PhD studies, started in 2004.

Emilio Benfenati

Emilio Benfenati, head of the Laboratory of Environmental Chemistry and Toxicology, at the ''Mario Negri'' Institute, Milan, is working in predictive toxicology, environmental toxicology and environmental analysis. He coordinates/coordinated a dozen of international projects and participates/participated in more than 20 EC funded projects. He is author or co-author of about 200 articles in international journals and books. He received his diploma in Chemistry at the University of Milan. He obtained a post-doc diploma at ''Mario Negri'' Institute. He has worked in Stanford and Berkeley, California, on biochemical research.

Metabolism Excretion (ADME) properties. Nowadays, we still have no complete (eco)toxicological characterization of the risk associated with the use of industrial chemicals. Two elements are required to address this: an evaluation of the intrinsic properties of the chemical – hazard assessment – and an estimate of the exposure. The main obstacle to completing hazard assessment is the lack of adequate experimental data for chemicals, required to cover all the major effects relating to human health or ecological safety.[1] In the near future this situation will change with the REACH (Registration, Evaluation and Authorisation of Chemicals) legislation recently approved in the EU. REACH will affect the current authorization scheme for marketing chemicals in the EU, introducing new requirements and this is expected to boost *in silico* studies.

Previous legislation distinguished between ''existing'' and ''new'' chemicals (placed on the market after 1981); the latter had to be tested before being placed on the market, but there was no such provision for ''existing'' chemicals. Thus, although there is some information on the properties and uses of existing substances, the old system has not produced sufficient information about the effects of the majority of these chemicals on human health and the environment. REACH should help fill this data gap but the increased need for experimental tests raises a number of problems, ranging from economic – in terms of either time or money – to ethical ones, in view of the large number of laboratory animals needed for testing. It is estimated that 3.9 million laboratory animals might potentially be utilized for testing under the REACH requirements.[2] The economic costs of implementing REACH are also very high: over the next 11 years it might cost nearly 2.5 billion euros to test the approximately 30,000 substances of which more than 1 tonne per year is produced.[1]

In silico techniques could be a valuable complement to *in vivo* and *in vitro* studies in assessing hazards of chemicals. Of course they cannot replace ''wet'' experiments but they can be integrated with them. They can set priorities for compounds needing deeper *in vitro* and *in vivo* investigations, for a more rational use of resources by planning experimental testing better.

To avoid unnecessary testing REACH provides some indications for the use of existing information, techniques such as QSARs, read across and analogue identification. Some studies have estimated that these alternatives, including QSARs, will reduce the additional costs due to implementing REACH by about one billion euros[1] and these alternatives could potentially save more than a million animals.[2]

Some efforts are still needed to facilitate regulatory acceptance of QSAR as an alternative[3] by increasing the transparency and reproducibility of the models generated with QSAR and taking account of regulators' needs. The OECD has identified some principles on QSAR validation to satisfy these aspects.

Even though REACH legislation substantially boosted interest on computational chemistry to replace experimental testing, this is not the only case where QSAR is accepted in a regulatory framework. These methods are already accepted and used in the USA, Canada, Japan and some EU countries at a national level.[3]

In silico technology can also help improve the safety of newly synthesized chemicals or even chemicals in a pre-synthesis phase since only the chemical composition is required. This can help in evaluating safer alternatives to be inserted later in the marketing stage.

The present paper illustrates the main techniques for evaluating the properties of chemicals *in silico*. With a theoretical introduction of the various approaches, practical proof of the concept will be given, using as a working example *in silico* studies conducted in the field of endocrine disrupters (EDs). The models here presented are not intended to be exhaustive but simply show the range of possible different applications.

Computational modelling methods

This section describes the three main approaches for computationally evaluating the properties of chemical compounds: the classical QSAR approach, 3D-QSAR and virtual docking. After an introduction to the theory of the various approaches some practical examples will be given later, on *in silico* studies in the field of EDs. This serves to identify the pros and cons of the different approaches and gives a background about what is already known in this area and the most promising approaches, with emphasis on the regulatory perspective.

(Q)SAR/(Q)SPR

Many attempts have been made in the past to relate, in a qualitatively or quantitatively manner, molecular characteristics to some observed properties.[4] In the '60s a fundamental contribution was given by Corwin Hansch. Hansch's paradigm was based on the study of congeneric series of chemicals. The activity, expressed in the logarithmic form, was assumed to depend on the substituents' contributions to the parent compound in terms of hydrophobicity, electronic and steric terms. The biological relevance of these terms was correlated with the compound's ability to penetrate the biosystem and to reach the target site for interaction.[4] To move on from the requirement of congeneric series a variety of molecular descriptors have been proposed over the years to encode the structural features of chemicals.[5]

The workflow of the QSAR process is schematically represented in Fig. 1. The assumption behind the development of a QSAR model is that there is a quantitative relation between molecular features and the biological activity. To find this relationship the following steps are needed:

1) Calculation of chemical descriptors

As already mentioned, several types of descriptors can be used to encode different properties of chemicals (*e.g.* electrostatic, hydrophobic, steric, topological, *etc.*). They can include experimentally obtained physico-chemical properties (*e.g.* boiling point) but the majority of the studies employ a molecular description computationally obtained on the basis of the chemical structure.

A more detailed description about the types of descriptors can be found elsewhere[5] but for the purposes of the present paper a very practical approach for grouping them is to

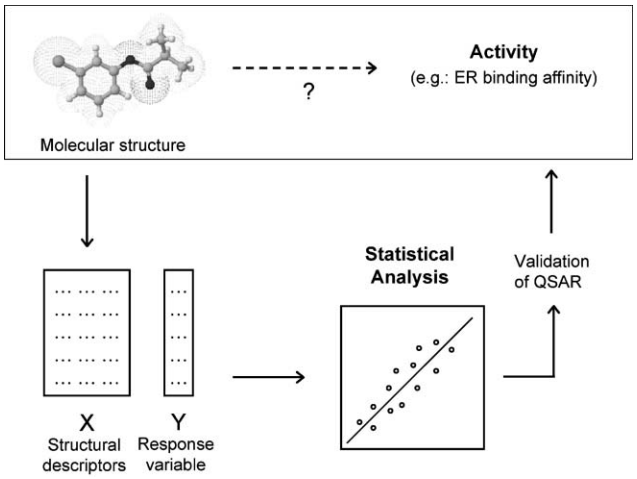


Fig. 1 Schematic representation of the steps for developing (Q)SAR models.

consider the structural information required to calculate them. Some descriptors encode very simple features of the molecules that do not depend on the three-dimensional (3D) conformation of the molecule and can be easily computed on the basis of the bi-dimensional (2D) structure alone. Other characteristics, such as energy terms, require knowledge of the 3D structure. This is not straightforward because each molecule can exist in multiple conformations with different levels of stability and occurrence. The solution commonly adopted is to use the energetically most stable conformation as a reference. To obtain it the process must sample the conformational space (conformational search) and find the global minimum structure by optimising the geometry *via* common force fields, semi-empirical or *ab-initio* methods. Then 3D descriptors are computed on this structure.

For 2D descriptors it is relatively easy to set up a procedure to make them reproducible, so they can be reliably calculated for new compounds later on, and to apply the QSAR model; however this is more complex for 3D descriptors. First of all, the optimisation can include non-deterministic steps; secondly, sometimes small changes in the 3D conformation can have a larger effect on some 3D descriptors. Moreover, only a reference 3D structure can be obtained, not necessarily representing the bioactive conformation of a molecule in its interaction with the biological environment.

Despite these limitations, however, 3D descriptors have been used successfully in many studies as shown later, even though on large and heterogeneous datasets they have sometimes proved to be as good as 2D ones.

2) Preparation of the Y-block variable

With the X matrix containing the independent variables, a Y matrix containing the target properties to be studied has to be collected. If the target is a single activity this matrix consists of a single column with the activity value for each chemical. It can be a continuous variable modelled quantitatively or a categorical one modelled with classification techniques. A major problem in preparing activity data is that any algorithm adopted during the modelling relies on these data to extract rules describing the activity trend so if the data are unreliable the model will be misleading. Great attention must therefore be paid in preparing a dataset suitable for modelling purposes, pruning all potentially ambiguous data. Despite all efforts to use only reliable data there is always an intrinsic uncertainty that cannot be avoided, especially if the data take biological systems into account. A certain degree of noise is introduced into the system and the modelling step must distinguish between the relevant information in the data and the noise and redundancy introduced with X or Y-block variables. Beside the mathematical definition, the Y variable represents the target to be modelled and can be of heterogeneous nature; in principle any physicochemical or biochemical properties can be used (and normally referred as Quantitative Structure–Property Relationship, QSPR) as well as more complex biological activities such as toxicity (those properly named QSAR).

3) Statistical analysis

This is the central step of the modelling task. It includes pre-processing the data matrix, variable selection to include only relevant descriptors, and the application of specific algorithms to find the relationship between variables and the target property.

Pre-processing is a preliminary but essential stage where the data matrix is pruned of redundant information, incomplete variables, and the scaling procedure is applied to the dataset.

Important variables can be selected in two ways: hypothesis-driven, including only variables considered relevant *a priori* to model the endpoint, or statistically driven using mathematical algorithms to search for the most important solutions. A stepwise approach or multivariate data exploratory methods such as Principal Component Analysis (PCA) may be used. However if there are too many initial variables these methods do not efficiently explore all possible combinations and more sophisticated tools are required. Genetic Algorithms (GA) have turned out to be one of the most promising algorithms. They are based on the Darwinian evolutionary theory. The best individuals in a population of models are crossed over, merged, mutated, then iteratively evaluated against a fitness function which gives a statistical evaluation of the model's performances.

A wide variety of methods is available to derive a model. A first distinction can be made between QSAR and Structure–Activity Relationships (SAR). While QSAR searches for a quantitative relationship, SAR typically is a qualitative relation between a molecular substructure and the presence or absence of a certain activity or the ability to modulate that activity. Altogether QSAR and SAR methods are indicated as (Q)SAR. The algorithms used for modelling also vary depending on the type of study: dealing with either categorical target properties, which employ classification tools, or continuous variables, which use regression approaches. The increasing number of descriptors commonly calculated required the introduction of different tools compared with multi-linear regression (MLR) to cope with correlated variables and with matrixes constituted by more numerous variables than the chemicals in the data set. Tools such as

PLS (Partial Least Squares) and PCA, had introduced the way to deal with this kind of problem, based on the use of "latent variables" generated by a linear combination of the original set of descriptors. Thus, in the last decades the simpler models with a few variables based on homogeneous set of chemicals have been replaced with studies using more heterogeneous data sets and a high number of variables. Beside the classical multivariate techniques based on linear methods, it is now common to use neural networks (NN) as a non-linear statistical data modelling tool. In the last years these non-linear tools have been introduced to take into account relationships where the Y variable does not depend in a linear way with a combination of the independent variables. NN is inspired by the way biological nervous systems, such as the brain, process information. These are systems of interconnecting neurons in a network functioning together to produce an output algorithm.[6] This trend is a further step towards more complex systems. This fact, and the request about the possible use of QSAR methods as predictive tools for regulatory purposes, underlined the need of a more robust basis for the validation of the model.

4) Validation of the model

Whatever the technique chosen to model a specific dataset, one of the most important issues in the QSAR field is validation of the model. To assess a qualitative or quantitative model, several characteristics have to be analysed, focusing in particular on three aspects: (i) internal validation, (ii) prediction ability and (iii) applicability domain.

(i) Internal validation is based on the assessment of goodness-of-fit and robustness. The first concept applies to the model's ability to describe variation in the training set, while the latter provides an indication on the model's stability in terms of how sensitive it is to perturbation in the training set.

Commonly for models based on continuous responses the main statistical parameter for assessing the goodness-of-fit is the coefficient of determination R^2, reported in eqn (1):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{\sum_{i=1}^{n} (y_i - \bar{y})} \quad 0 \le R^2 \le 1 \qquad (1)$$

where y, $\hat{y}_i$ and $\bar{y}$ are respectively the observed, calculated and mean values of the Y dependent variable. The closer R^2 is to 1 the better the model fits.

For classification models the quality of the model can be assessed by measuring its accuracy: the ratio of correctly classified compounds to the total number of compounds in the dataset.

Robustness is usually assessed by the cross-validation procedure: the training set is iteratively perturbed by excluding one or more compounds and the other compounds are used to generate a model predicting the excluded chemicals with this sub-model. This procedure is repeated for all compounds. Statistical parameters similar to accuracy or R^2 (usually called Q^2 or R^2_{cv}) are then calculated based on the predicted values and should maintain a considerable value compared with R^2. The Y-response can also be randomised to evaluate whether with a dataset containing Y-scrambled responses the model statistics decrease significantly, as expected, or if not, this is an indication of chance correlation.

(ii) Traditionally QSAR models have been developed to describe a phenomenon suitable to identify a rational relationship between a given parameter and the property. However, later the emphasis was moved to use these relationships to predict the properties of unknown compounds. Consequently different tools become necessary to avoid over-fitting. In many cases, especially when there are many descriptors and complex algorithms are used, there is a risk of obtaining an over-fitted model that too closely follows the behaviour of the training set and cannot capture the trend of activity in a more general way. Statistical tools must therefore prove the model's ability to be valid in a general sense, *i.e.* to be predictive for compounds not used in developing the model. A debate is still ongoing in the scientific community on the best way to assess the robustness and predictive performance of a model.[7,8] An external set for validation has been proposed in some cases as the most appropriate way to assess the predictive power of a model[8] even though some difficulties may arise in designing it as representative of the set of chemicals to be addressed, covering all its main structural and physico-chemical characteristics.

(iii) The concept of applicability domain (AD) has attracted increasing attention in the QSAR field because of the need for a better definition of areas where the models can be used in practice with greater confidence about the prediction obtained. AD is based on the assessment of similarity for the new chemical to be predicted with the group of compounds used to develop the model. Approaches include chemometric tools based on comparisons of the descriptors used to develop the model for the new molecules to be tested, with the descriptor distribution for the molecules in the training set. Another approach involves comparison of the structural features of the compounds *a priori*, without necessarily using the descriptors selected in the models. In this case structures are encoded in fingerprints or on the basis of major fragments and are used to assess the similarity with the training set. A review of these approaches has been recently published.[9]

Overall the validation task should ensure that the model is statistically significant, reliable and robust to noise and data perturbation, and maintains its validity when the relationship is extrapolated to compounds sharing similarity with the training data, at least within a defined chemical space. The validity of a model is therefore judged on the basis of a series of aspects, summarized here, sometimes assessed with different methods.

3D-QSAR

3D-QSAR includes a variety of methods which basically differ from the classical QSAR analysis in the descriptor types used in the modelling. These methods are based on the concept of Molecular Interaction Fields (MIF).[10] Molecular features are obtained by mapping the environment surrounding the molecules in terms of energy interactions of various nature, mainly steric and electrostatic, but sometimes hydrophobic or hydrogen bonding potential may be included. This is done by

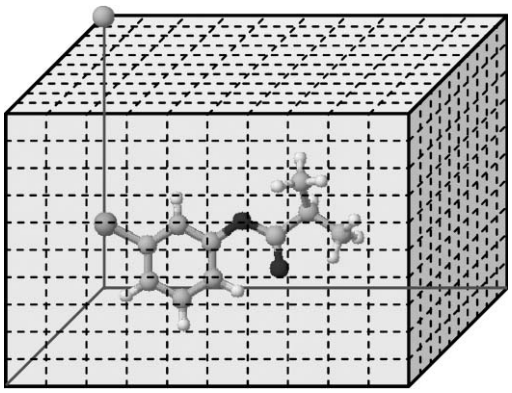


Fig. 2 Representation of the lattice used to calculate field-based descriptors. The probe used to calculate the energy terms in each point of the grid is in the upper left-hand corner.

placing the molecule within a lattice and calculating the interaction energies of that molecule with a probe (*e.g.* an sp^3 carbon atom) in each point of the grid, as shown in Fig. 2.

Since field-based descriptors are directionally dependent, a critical step in 3D-QSAR analysis is the alignment in the space of all the molecular 3D optimized structures in the dataset, according to various methodologies of superposition. The alignment can be based on the electrostatic and/or steric field overlap, based on a common skeleton superposition, evaluating docking or crystallographic information, or based on a pharmacophoric hypothesis.

Each descriptor column contains the values of the interaction field assumed for the compounds in the dataset at a certain point of the grid. Then these field energy terms are used as a very large pool of descriptors – hundreds or thousands – to search for a relationship with the property of interest, usually by PLS analysis, so that the differences from molecule to molecule in the fields they generated in some areas of the grid are usually related to the differences in the modelled activity. Since these interactions are clearly placed in the 3D space surrounding the molecules, a regression map can be created by mapping the regression coefficient of the model back into the box. Therefore the method identifies the most important regions in the molecule responsible for modulating the target properties.

The most popular method in 3D-QSAR studies is CoMFA[11] but other methods have been developed based on different force fields adopted to calculate the energy terms, and considering more heterogeneous probe definitions to capture more complex interactions, for example CoMSIA or GRID/PLS,[10] where GRID MIF is coupled with PLS analysis.

One of the most attractive features of 3D-QSAR compared to classical QSAR is that the biological environment surrounding the molecules is taken into account even if only implicitly. This happens especially when direct interaction with a target macromolecule is considered (*e.g.* activity against a specific receptor) and a hypothesis on these interactions can be drawn from the resulting relationship. The model can be interpreted so that the conformations of ligands are representative of the bioactive conformation in the binding pocket of the receptor and the alignment represents the different poses of the molecules binding the receptor. Choosing the bioactive conformation and proper alignment are therefore essential phases, and often information on these characteristics comes from crystallographic or docking studies.

Traditionally, the linear PLS method has been used to derive CoMFA models since it tolerates the inclusion of a large number of variables in the final model although other methods, such as neural networks, can be used to investigate nonlinear relationships. Including a large number of variables in the model, even if condensed in a few principal components, can, however, increase the risk of chance correlation. The validation procedure is thus essential to test the model performance, also for the 3D-QSAR method. Commonly a cross-validation procedure is adopted and sometimes a randomisation test is included. Often the model performances are verified by evaluating the prediction for new compounds, constituting a test set. However, either results from cross-validation or performances on the test set compounds cannot give a proper estimate of model reliability if these parameters have been used to select the best architecture of the models. These issues are discussed in more detail in a review by Y. Martin.[12]

Growing attention to the biological environment and the easiness in terms of interpreting the chemical features through the regression maps are two of the main advantages of 3D-QSAR. The principal drawback is the increasing complexity of the models which requires 3D conformations, their alignment and a large number of variables. This can make it more difficult to reproduce a model or at least to apply it to new compounds if the alignment rules are too specific or are not suitable for other chemical classes, limiting the range of chemicals that can be analysed.

To overcome the limitations due to the superposition procedure some alignment-independent extensions of 3D-QSAR descriptors have been developed that do not require an aligned structure, such as VolSurf and GRIND derived from GRID/PLS.[10]

Virtual docking

Virtual docking computationally predicts the binding between two molecules, usually a protein and another macromolecule (protein or DNA) or a small molecule (ligand). Here we focus on protein–ligand docking as a tool to estimate the reactivity of chemical compounds with biological target sites.

For this kind of study the chemical composition and 3D spatial organization of the protein must be known, with identification of the cavity defining the binding site of the protein whose position and shape is used in the docking process. Usually the best source is the structure provided by the X-ray crystallography. If this is not available, a structure determined with NMR spectroscopy or by homology modelling may be used. The latter method involves reconstructing the 3D shape of the protein of interest from other proteins whose structure is known, that have similarities in the aminoacidic sequences.

Using the crystal structure of a ligand–receptor complex as starting point means beginning from a single, low-energy snapshot of an actual dynamic biological system. The first task

is to sample the conformational space of possible energetically reasonable poses constituted by the protein–ligand complexes. This is not a trivial matter since the conformational space to be examined is huge due to the intrinsic flexibility of both the ligands and the protein, including further forms of plasticity that can be introduced on the mutual recognition between protein and ligand by the induced fit process. Nowadays the majority of docking programs take into account ligand flexibility in contrast with rigid docking (where both the ligand and the protein are considered as rigid bodies), but protein flexibility has not yet been fully integrated into docking protocols and is often considered only marginally. Among the methods employed for the searching strategy there are molecular dynamic simulations, Monte Carlo methods, genetic algorithms and fragment-based methods.[13]

Once a pool of ligand–protein complexes has been generated, scoring functions are used by docking programs to indicate the likelihood that the pose offers a favourable binding interaction. The scoring function provides an estimate of the Gibbs free energy of binding, released when ligand and receptor bind, to evaluate their stability as a complex (Fig. 3). For each pose $\Delta G_{\text{binding}}$ can be used to assess whether a favourable binding opportunity exists through eqn (2):

$$\Delta G_{\text{binding}} = RT\ln K_{\text{D}} \quad (2)$$

where R is the gas constant, T the absolute temperature and K_{D} the dissociation constant. The dissociation constant can then be directly linked with the inhibition constant (K_{i}) or the inhibition concentration (IC_{50}) obtained *in vitro* in a binding assay. Proper scoring for the docked poses is the second step in the docking process. The scoring functions may rely on force fields to calculate energies or on knowledge-based or empirical functions (including QSAR relationships).[14]

Although fairly accurate ways exist for estimating these energies, based on free energy perturbation or thermodynamic integration methods, this accuracy level in practice can be used only in a few, very focused studies. On the contrary, the most attractive application of virtual docking is in virtual high throughput screening (HTS), where large virtual libraries of compounds are reduced to a subset which, if successful, includes molecules with high binding affinities to a target receptor. This approach is often used in the drug discovery process to identify new lead compounds. The scoring function used to approximate the energies must be fast enough to accomplish this task in a reasonable time, but this tends to give low accuracy. The typical level of precision reached by docking programs in virtual HTS does not allow direct correlation of their scores with the binding affinity. They are also hardly able to qualitatively rank the compound order properly in relation to the binding strength.

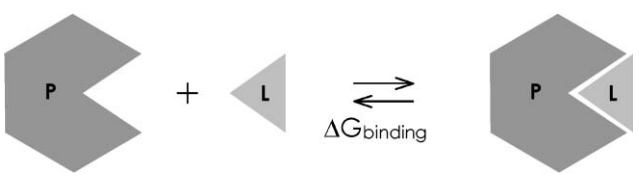


Fig. 3 Mutual recognition between a protein (P) and a ligand (L) is governed by the Gibbs free energy of binding, released when ligand and receptor bind.

Normally the outcomes of this task can be measured with the enrichment factor so that the subset selected with the docking procedure contains a larger amount of compounds showing affinity for the receptor studied.

For a practical comparison of performances and features of the most popular docking programs the reader is referred to the literature.[15]

The calculation complexity of virtual docking, which also encodes the protein structure, is surely larger than for ligands alone, since more atoms are involved in the calculation and proper force field parameterisation for aminoacid residues is essential; however, once the experimental protocol is set, docking methods are fast enough to screen very large libraries of chemical compounds in a reasonable time.

Some limitations concern the inaccurate energy estimation. Docking still remains effective in drug design, since it gives better performances than random selection of possible hits, as indicated by the enrichment factor. Its application as a pre-screening tool for the hazard assessment of less enhanced binding activities has been tested less often.

The endocrine disrupters issue and the oestrogen receptor

Endocrine disrupters (EDs) form an emerging field that is attracting attention from scientists and political institutions. It deals with a number of exogenous substances interfering with the function of the endocrine system producing consequences on the homeostasis of all the process controlled by this system in humans and wildlife. Effects on reproductive, developmental, immunological and neurological functions may arise, such as cancer, behavioural changes and reproductive abnormalities.[16]

The EDs issue is highly complex on account of the wide range of mechanisms of action they can interfere with.[16] The targets include receptors belonging to the nuclear receptor (NR) superfamily. This group of ligand-inducible transcription factors mediates the effects of hormones and other endogenous ligands to regulate the expression of specific genes. They include receptors for hormones like steroids, retinoic acid and thyroid hormones.

Among them the NR more extensively investigated to account for endocrine disrupting effects is the oestrogen receptor (ER) that mediates the effects of the steroid hormone 17β-estradiol in males and females. It is needed for the development, growth and maintenance of reproductive tissues but is also present in a number of non-reproductive tissues, such as bone, liver, brain, the CNS, cardiovascular and immune systems in the physiological situation.

The ER is present in humans in two isoforms (ER alpha and beta) whose specific chemical structures and functions have been elucidated by crystallographic studies and biochemical analyses.[17]

To detect endocrine disrupting effects different *in vivo* and *in vitro* experiments have been set up. Since the system itself and the variety of targets is very broad and varies in different species, a battery of tests has been proposed to analyse

different possible interactions with the endocrine system, including ER, organized in a tier approach. Some *in vitro* tests have been designed to detect direct binding with the receptor ligand binding domain (receptor binding assay), or the transcriptional activation of DNA (cell proliferation and reporter gene assays).

Despite all the experimental methods under investigation, only a few experimental protocols are fully standardized and validated, so computational chemistry offers a complementary tool to characterize the binding to ER and more in general chemicals that interfere with the endocrine system. This is becoming ever more necessary because so many of these substances are used in industry and the data gap also needs to be filled to satisfy REACH legislation. In particular EDs are mentioned as substances requiring a more detailed control within REACH framework together with other group of chemical of particular concern.

Computational models on ER

QSAR and other *in silico* tools are very suitable for addressing the direct interaction of chemicals with receptors, since both the ligands and the receptors can be characterized in their chemical structure. The ER was one of the first targets studied with computer-aided methods addressing specifically EDs.

On the basis of the complexity of the EDs issue, it has to be understood that the scientific target is huge, and *in silico* tools can provide a strong help in the research, but not solve all open problems. There are some critical aspects in the application of *in silico* tools to EDs. The description of the phenomenon is done through a series of experimental models, and only a few of them have been standardised. Thus, the data are very heterogeneous and their comparison and integration critical. This is a general aspect, not only specific for *in silico* methods. However, *in silico* models are based on data in an extensive way, and data heterogeneity represents a heavy limit for most of the *in silico* methods. Related to this qualitative assessment of the data there is a second aspect: the quantitative limitation of the data. Many *in silico* methods need a high number of data and this affects the result. Finally, most of the data refer to the binding measurements, while there are less data for the more complex targets addressing the whole phenomenon.

Thus, the complexity of the task demands a complex strategy of the *in silico* tools, and in perspective different methods integrated each other, including *in vitro* and *in vivo* ones. Referring to the *in silico* tools it has to be underlined that they are different on the approach and consequently on the addressed phenomena. Docking studies are superb methods to describe the binding to the receptor. However, they fail in the description of other phenomena, for instance when the chemical is metabolised, or the final effect of the ED is mediated by post-transcriptional activity. *Vice versa*, QSAR models can implicitly encode different phenomena, but to deal efficacy with such a complex scenario many data are necessary.

Models on ER with (Q)SAR

ER has been widely studied with QSAR techniques in relation to the EDs issue. Many studies were (Q)SPR focused on the binding assay data. The datasets were relatively heterogeneous in terms of the number of compounds used – from a few dozen up to a few hundred – and the source of binding activity data: different species (rat, mouse, human, calf) and subtypes (alpha, beta or mixed ones).

Some studies employed structural features to discriminate binders and non-binders in a SAR. Fang *et al.* associated the ability of compounds to bind ER qualitatively with the presence of certain characteristics (*e.g.* a phenolic ring).[18] These kinds of studies can give useful tools both as mathematical method, easily used, and as a single, transparent set of rules related to the presence of certain chemical features. Other examples have been proposed by Klopman *et al.* where the occurrence of certain groups among active or inactive chemicals were used to recursively characterize the main fragments in the two groups.[19] Classification methods were also used to divide the data into two classes, employing different cut-offs to discriminate binders from non-binders or chemicals with marginal or strong activity.[20] Another approach involves the use of multiple conformations for each chemical.[21] Then the distribution in the population of values for the descriptors is used to derive a classification model.

QSAR/QSPR studies proposed instead quantitative equations using chemical descriptors and different algorithms. Linear regression models have been produced using MLR and PLS.[22,23] Non-linear techniques have been explored. One of them is a non-linear technique based on the concept of molecular similarity and K-nearest neighbour principle.[24] Other non-linear methods relied on different NN techniques.[23]

In these studies SAR and QSAR methods use approaches which are independent of the ED topic. The same approaches are used for other properties. In the specific case of EDs, the utility of these techniques is as follows. SAR methods identify parts of the molecule responsible for the effect, with the aim to characterize a compound as active or not. *Vice versa*, QSAR/QSPR methods adopt algorithms and descriptors suitable to modulate the activity. The chemical information, introduced as chemical descriptors, can be useful to highlight a possible mechanistic basis, which however has to be proved, and for some chemical descriptors can be not intuitive. For instance, SAR and QSAR/QSPR models may be an independent way to confirm key factor for the ER ligand interactions derived from other types of study. Often in these models some descriptors are frequently selected and their presence can be related to the biochemical mechanism of receptor binding. This is the case of descriptors encoding for the presence of the phenolic ring or for hydrogen bonding abilities since H-bonds are considered essentials for the interactions within the binding site.

Some studies have taken a different endpoint for oestrogenicity: instead of developing models focused on binding affinity data, the ER transactivation properties were investigated in terms of reporter gene assay or cell proliferation assay. Classification models have been produced, for instance, by using classification trees.[25] This kind of endpoint could not be addressed through docking methods.

This is a major difference between QSAR models and 3D-QSAR or docking. QSAR models can virtually address any phenomenon and also for this reason they have been widely used. They, in principle, can encode a number of mechanisms,

eventually within the same model. Indeed multiple descriptors can be used which in principle refer to different biochemical processes. Thus not only very specific properties can be addressed, such as the binding to a specific receptor, but also more general biological effects that can be more relevant in the context of EDs. In this sense the method is flexible, because it simply makes association between some chemical features and the effects. However, the disadvantage of this approach is on its premises. Since it is a statistical based method, it needs sufficient examples to extract the correct knowledge. The lessons which can be derived are quite general, and there is the risk that the population of examples used to derive the model is not well representative.

It is difficult to compare the different QSAR approaches since these models often relied on different datasets or at least different validation parameters. Regarding the chemical information, several of the papers mentioned used 2D descriptors or even compared performances with 3D ones. The majority found that with 2D descriptors, simpler to calculate and allowing faster analysis, it is possible to obtain results comparable to those involving more complex 3D descriptors.[20,26] This finding is surely an advantage for models that have to process a large number of compounds, since they are faster. 2D descriptors are also preferable for regulatory purposes, since they do not need 3D optimization which is manually done in the typical case, and thus affected by subjective factors. However, the interpretability of the selected descriptors is sometimes less explicit, especially if the models use a large number of variables.

Some attempts have been made to improve the regulatory acceptance of these models, particularly by starting to address the question of defining the applicability domain better.[20,25]

Models on ER with 3D-QSAR

3D-QSAR methods have been widely applied to study receptor–ligand interactions since the defined biological reaction site makes it easier to detect a proper alignment.

Often modelling exercises have used both 3D-QSAR and classical QSAR approaches and compared the outcomes based on the investigation of a common dataset.[26,27] For instance a comparative study was carried out using CoMFA and CoMSIA, classical 2D/3D descriptors, and fingerprint-type descriptors – characterizing chemical structures in a string of bits indicating the presence or absence of specific 2D or 3D structure characteristics.[26] Another study proposed CoMFA combined with 2D-QSAR methods based on fingerprint descriptors (HQSAR, and FRED/SKEYS) providing a helpful comparison of their predictive power.[27]

Compared to QSAR, 3D-QSAR is more focused on the chemical parts of the molecules which are related to the effects, identifying the location and nature of the interaction (for instance steric or electronic) with the macromolecule responsible for the given effect. *Vice versa*, as we have seen, QSAR may implicitly encode different mechanisms. Thus 3D-QSAR is less suitable for heterogeneous sets of compounds, which quite likely are active through different mechanisms. Furthermore, 3D-QSAR requires a similar skeleton for the alignment, which is a further element of focalization into a defined set of chemicals and mechanisms. Compared to docking methods, 3D-QSAR do not involve the receptor into the modelling phase, thus they are simpler. However the interaction with the macromolecule is indirectly assessed. 3D-QSARs are more limited in their possibilities to embrace a series of processes simultaneously, but are more understandable, for the specificity of the interaction. Indeed some models proved to be useful to study the receptor subtype selectivity that is a peculiar characteristic of natural phytoestrogens.

Overall these studies demonstrated that for heterogeneous datasets 3D and classical QSAR approaches offer similar performances, not necessarily justifying the use of 3D-QSAR with its greater complexity. This kind of assessment is often based on a limited set of validation parameters, provided in the original studies, so the conclusion may be different from a larger pool of validation factors. On the other hand 3D-QSAR provides a more easily interpretable model in terms of chemical features especially compared to other less intuitive and transparent molecular descriptors. Indeed, 3D-QSAR defines the position in the molecule which is related to the biological activity through certain factors and this is a very clear way to identify biochemical factors to be related to a certain mechanism. *Vice versa* QSAR do not indicate the part of the molecule which is responsible for a given effect; an exception is for instance the use of a phenolic group as a relevant molecular feature in the model. In the case of QSAR many descriptors have been criticized to be of poor help in understanding the mechanism, as they have been only selected on the basis of statistical relationships.

3D-QSAR can be used to highlight the differences in the receptor affinity and for modelling the selectivity of the ligands to some receptor subtypes. Tong *et al.*[28] employed CoMFA maps to identify and differentiate the structural features of ligands responsible for selective binding to ER alpha and beta. Although the receptor crystal structure is available, CoMFA provides additional information about the receptor from the perspective of the ligands.

Other successful applications involve series of relatively homogeneous compounds where proper alignment rules can be detected more easily. This approach is commonly used in the pharmaceutical industry for optimising the characteristics of a lead compound in a homogeneous series. CoMFA, CoMSIA and HQSAR were used to investigate a series of bisphenol A analogues considering not only the binding but modelling other properties too, such as transactivation potency. Again, the statistical performances of the different methods seem similar even if their outcomes are more complementary than alternative. This study provides an example where instead of the minimum conformation, the most probable bioactive conformer identified by another *in silico* simulation involving virtual docking was used.[29]

Models on ER with docking methods

Some docking studies have addressed the issue of EDs more or less specifically. Some of these works provided a quantitative assessment of the binding with the docking approach for a limited number of compounds. Some encouraging outcomes of docking simulations have been obtained to measure the

interacting energies fairly precisely and link them with the experimental binding,[30] also recognizing different selectivity for the two subtypes.[31]

In another study with a larger set of heterogeneous compounds, the binding affinities correlated less precisely with docking scores but the approach can be considered successful for evaluating enrichment factors in screening. To account for receptor flexibility a subset of receptor crystal structures were used in parallel in the docking process and this approach increased the precision compared with a single complex.[32]

Other studies compared or combined docking and 3D-QSAR and found that the docking method failed to give enough accuracy in estimating binding strength, as explained previously, but confirmed its validity as a complementary tool for developing more powerful 3D-QSAR models. In particular, the use of docking conformers generated a biologically more plausible alignment, used in 3D-QSAR.

Although in one of these studies docking itself did not quantitatively predict the binding affinities, the use of docked poses for the ligand and – with less influence – the inclusion of scores as additional parameters, gave a better model than the classical CoMFA model using the lowest energy conformer for ER alpha, but completely failed for ER beta.[33]

There are also examples where direct calculation of binding energies based on docking runs gave poor results compared to classical CoMFA models, but the use of docked poses instead of minimum energy conformation significantly improved the performance of 3D-QSAR, including promising prediction of the activity of new chemicals.[34]

The docking approach has been used as a starting point to develop multi-dimensional QSAR models for ER.[35] The multiple dimension was achieved by the inclusion of multiple conformations (4D), induced fit (5D) and solvation effects (6D). Receptor surrogates were obtained by mapping the different properties on a surface surrounding the molecules and selecting the most appropriate ones to assess the binding affinity using GA. This method gave very good quantitative results for a fairly large dataset of heterogeneous compounds.

Conclusions

We have illustrated different ways for assessing the properties of chemical compounds *in silico*. The different approaches are summarized in Table 1. A first distinction can be made between receptor-dependent and independent methods.

Traditionally, in environmental safety and health, the phenomena under evaluation are more general and do not necessarily represent the explicit interaction with a well-characterized, specific receptor (*e.g.* systemic toxicity or carcinogenic process), so receptor-based methods can be applied only in some circumstances such as for investigating NR interactions or metabolic processes mediated by the cytochrome P450 family. When more general multi-step toxic effects of chemicals are studied, or no defined mode of action can be recognized, the modelling can rely on the chemical structure alone. In these cases QSAR and 3D-QSAR can be applied even when the target macromolecule is not known or toxic effects cannot be linked with a specific receptor.

Moving from classical QSAR to 3D-QSAR to docking more attention is being paid to the biochemical mechanism side and this increases the "biological plausibility" of the results given by statistical methods. QSAR has been criticized in some cases to rely solely on statistical factors. 3D-QSAR more clearly supports some findings related to important molecular areas, which modulate activity making the mechanism more easily understandable. Docking is strongly convincing thanks to the background knowledge on the receptor structure and chemical properties.

Explicitly introducing a description of the biological receptor increases the precision of the biological environment description, and can provide more insight into the mechanism.

Table 1 Methods for studying the biological effects of chemical compounds

	Techniques		
	Receptor-independent		Receptor-dependent
	QSAR	3D-QSAR	Docking
Description	(Q)SAR searches for a qualitative or quantitative relationship describing the influence of small molecule features, encoded in molecular descriptors, in producing a certain biological effect through a statistically significant model.	It uses field-based descriptors to represent the energy surrounds of the molecules. It requires 3D conformers which must be properly aligned.	The binding affinity of a ligand with a receptor is inferred by an energy evaluation of the complex through scoring functions.
Pros	Quite fast and reproducible especially if descriptors depend only on 2D characteristics of the molecules.	It considers the biological environment surrounding small molecules, although only implicitly.	Closer to reality: it explicitly includes a description of the biological macromolecules responsible for activity.
	Widely applicable to new compounds. Although the model itself can be complex to generate, easy rules can be formulated.	It allows for visualization of the main molecular characteristics through regression maps.	It gives a deeper mechanistic understanding.
Cons	Less realistic: it ignores the 3D biologically active conformation of the ligand and the chemical structure of biological macromolecules responsible for the effects.	It requires 3D conformers and is very sensitive to the alignment procedure.	It can be used to model the binding strength of ligands with proteins but not more complex and global biological effects.
			The accuracy often permits only qualitative output or ranking compounds for their activity.

On the other hand, the increased complexity (*e.g.* computational), although it may help in understanding the process, often does not produce more significant models.

It is not worth trying to estimate *a priori* whether one technique is superior to another. Frequently, depending on the problem to be addressed, some instruments are better than others. An interesting perspective is that complementary outcomes can be reached, so that a more complete description of a phenomenon can be obtained by integrating different techniques and each method reinforces the others. However, the target phenomena which can be addressed by the different methods are different, since docking is limited to receptor modelling, while methods such as QSAR can be broader.

Due to the high complexity of EDs problem it has to be recognized that addressing some specific nuclear receptors may be not sufficient to assess the possible hazard posed by chemicals. A wise battery of *in silico* models, supported by experimental tests focusing on different aspects of this phenomenon, seems to be the most promising direction to provide a more complete overview of endocrine disruption.

A further aspect to be considered is the intended use of the model. Docking studies are more complex, while QSAR typically are the easiest. If a certain model has to be easily and quickly used by non expert users, docking models cannot be chosen. Thus, in the aim to offer easy, fast tools for regulatory purposes, QSAR are the most suitable ones, while for research an advanced, complex integration of different models is more promising.

In any case, a main critical aspect will be larger data availability, necessary to obtain further, improved models.

Acknowledgements

The authors acknowledge partial financial support by JRC (CCR.IHCP.C430414.X0) and EC (CASCADE project, FOOD-CT-2004-506319).

References

1 F. Pedersen, J. de Bruijn, S. Munn and K. van Leeuwen, *Assessment of additional testing needs under REACH. Effects of (Q)SARS, risk based testing and voluntary industry initiatives*, JRC Report, EUR 20863 EN, 2003.
2 K. van der Jagt, S. Munn, J. Tørsløv and J. de Bruijn, *Alternative approaches can reduce the use of test animals under REACH. Addendum to the report: Assessment of additional testing needs under REACH Effects of (Q)SARS, risk based testing and voluntary industry initiatives*, JRC Report, EUR 21405 EN, 2004.
3 J. Jaworska, M. Comber, C. Auer and C. Van Leeuwen, *Environ. Health Perspect.*, 2003, **111**, 1358–1360 and references therein.
4 H. Kubinyi, *Quant. Struct. Act. Relat.*, 2002, **21**, 348.
5 R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, ed. R. Mannhold, H. Kubinyi, and H. Timmerman, WILEY-VCH, Weinheim, 2000.
6 U. Burkard, in *Chemoinformatics*, ed. J. Gasteiger and T. Engel, WILEY-VCH, Weinheim, 2004, ch. 9, pp. 439–485.
7 D. M. Hawkins, *J. Chem. Inf. Modell.*, 2004, **44**, 1.
8 A. Golbraikh and A. Tropsha, *J. Mol. Graphics Modell.*, 2002, **20**, 269.
9 J. Jaworska, N. Nikolova-Jeliazkova and T. Aldenberg, *ATLA, Altern. Lab. Anim.*, 2005, **33**, 445.
10 *Molecular Interaction Fields*, ed. G. Cruciani, WILEY-VCH, Weinheim, 2006.
11 R. D. Cramer, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959.
12 Y. C. Martin, *Perspect. Drug Discovery Des.*, 1998, **12–14**, 3.
13 R. D. Taylor, P. J. Jewsbury and J. W. Essex, *J. Comput. Aided Mol. Des.*, 2002, **16**, 151.
14 T. I. Oprea and G. R. Marshall, *Perspect. Drug Discovery Des.*, 1998, **9–11**, 35.
15 G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff and M. S. Head, *J. Med. Chem.*, 2006, **49**, 5912.
16 T. Damstra, S. Barlow, A. Bergman, R. Kavlock and G. van der Kraak, in *Global assessment of the state-of-the-science of endocrine disruptors*, WHO 2002, http://www.who.int/ipcs/publications/new_issues/endocrine_disruptors/en/index.html.
17 S. Nilsson, S. Mäkelä, E. Treuter, M. Tujague, J. Thomsen, G. Andersson, E. Enmark, K. Pettersson, M. Warner and J.-Å. Gustafsson, *Physiol. Rev.*, 2001, **81**, 1535 and references therein.
18 H. Fang, W. Tong, L. M. Shi, R. Blair, R. Perkins, W. Branham, B. S. Hass, Q. Xie, S. L. Dial, C. L. Moland and D. M. Sheehan, *Chem. Res. Toxicol.*, 2001, **14**, 280.
19 G. Klopman and S. K. Chakravarti, *Chemosphere*, 2003, **51**, 445.
20 W. Tong, Q. Xie, H. Hong, L. Shi, H. Fang and R. Perkins, *Environ. Health Perspect.*, 2004, **112**, 1249.
21 P. K. Schmieder, G. Ankley, O. Mekenyan, J. D. Walker and S. Bradbury, *Environ. Toxicol. Chem.*, 2003, **22**, 1844 and references therein.
22 T. Ghafourian and M. T. D. Cronin, *SAR QSAR Environ. Res.*, 2005, **16**, 171.
23 F. Marini, A. Roncaglioni and M. Novič, *J. Chem. Inf. Modell.*, 2005, **45**, 1507.
24 A. H. Asikainen, J. Ruuskanen and K. A. Tuppurainen, *SAR QSAR Environ. Res.*, 2004, **15**, 19.
25 T. I. Netzeva, A. Gallegos Saliner and A. P. Worth, *Environ. Toxicol. Chem.*, 2006, **25**, 1223.
26 N. Brown, B. McKay and J. Gasteiger, *QSAR Comb. Sci.*, 2005, **24**, 480.
27 C. L. Waller, *J. Chem. Inf. Modell.*, 2004, **44**, 758.
28 W. Tong, R. Perkins, L. Xing, W. J. Welsh and D. M. Sheehan, *Endocrinology*, 1997, **138**, 4022.
29 P. Coleman Kelly, A. Toscano William, J. Thomas and E. Wiese, *QSAR Comb. Sci.*, 2003, **22**, 78.
30 M. M. H. vanLipzig, A. M. ter Laak, A. Jongejan, N. P. E. Vermeulen, M. Wamelink, D. Geerke and J. H. N. Meerman, *J. Med. Chem.*, 2004, **47**, 1018.
31 R. K. DeLisle, S.-J. Yu, A. C. Nair and W. J. Welsh, *J. Mol. Graph. Modell.*, 2001, **20**, 155.
32 S. Yoon and W. J. Welsh, *J. Chem. Inf. Modell.*, 2004, **44**, 88.
33 P. Wolohan and D. E. Reichert, *J. Comput. Aid. Mol. Des.*, 2003, **17**, 313.
34 W. Sippl, *J. Comput. Aided Mol. Des.*, 2000, **14**, 559.
35 A. Vedani, M. Dobler and M. A. Lill, *J. Med. Chem.*, 2005, **48**, 3700 and references therein.

