Predictive models based on molecular structure

(the following is a personal judgement of the authors)

pp-LFERs give accurate results and are applicable for a wide and diverse set of organic chemicals but the required compound descriptors are tabulated for only about 2000 chemicals and the experimental determination of additional descriptors is tedious. This is not an option when thousands of compounds have to be screened (Wittekindt, C. and K.-U. Goss, 2009, Chemosphere, **76**: 460-464 Journal link Download pdf) or when chemicals have to be assessed that have not even been synthesized (e.g. in the first stages of the design of a new pesticide). In these cases models are needed that can predict partitioning based solely on the molecular structure of the chemical.

Fragment models: The principal problems of fragment models have been outlined already (see Fragment models). In order to account for non-additive behaviour of neighboring functional groups in a molecule appropriate correction factors have to be derived from the calibration data. Therefore, virtually thousands of calibration data are needed to construct a fragment model whose applicability domain is sufficiently large to be of interest for environmental chemistry purposes. Such a well calibrated fragment model only exists for the prediction of log K_{ow} values (e.g. the commercial ClogP software or the KOWWIN ^{TM} module of the EPI Suite ^{TM} which is publicly available via the internet http://www.epa.gov/oppt/exposure/pubs/episuite.htm). However, since the use of log K_{ow} for predicting environmental partition processes is questionable, these fragment models are of little use here. In addition to the module KOWWIN ^{TM}, the EPI Suite ^{TM} contains two additional fragment models: HenryWIN ^{TM} for predicting air/water partitioning and KOCWIN ^{TM} for predicting K_{oc} values. Both have only been calibrated with a few hundred data and cannot be expected to have a wide applicability domain. The calibration data are not disclosed to the user so that a more specific judgement is not possible. A preliminary evaluation (Goss, K.-U., H.P.H. Arp, G. Bronner, C. Niederer, 2009 Environ. Tox. Chem., **28**: 52-60 Journal link Download pdf, and Wittekindt, C. and K.-U. Goss, 2009, Chemosphere, **76**: 460-464 Journal link Download pdf) seems to support our scepticism.

Quantitative Structure Activity Relationships (QSAR) or Quantitative Structure property Relationship (QSPR)
The terms „QSAR“ and “QSPR” are defined very broadly and cover any kind of mathematical relationship between one or several chemical descriptors than can be derived from the molecular structure of a chemical and its biological activity (e.g. toxicity) or it physico-chemical properties. Chemical descriptors used for such relationships may be very simple ones like molecular volume and mass, the number of double bonds or aromatic rings; or they may be the outcome of quite complex quantum chemical modelling. There are commercial software packages (e.g. CODESSA Pro or Dragon) with a statistical routine to automatically search for such QSAR or QSPR models. To this end, the user just has to have a calibration data set at hand for which he/she can enter the 3D structure and the corresponding data for the target variable. The software will then automatically calculate around 1000 different descriptors for each chemical from its 3D structure. In a second step a statistical regression is performed in order to search for one or more of these descriptors that best describe the variability of the target variable. Obviously there is no mechanistic understanding behind this approach. This results in severe limitations: a) it is very difficult to come up with a meaningful definition of the application range of such a model if only statistical tools but no mechanistic insight is used; b) there is a

considerable chance that no reasonable model is identified although it does exist. The latter situation is best illustrated with an example. Imagine a set of activity or property data that is indeed described to 100% by a linear combination of 5 descriptors for the respective chemicals. These 5 descriptors are part of a set of 1000 descriptors available for model building. The number of possible combinations of 5 descriptors out of 1000 descriptors is so high that it would take decades to calculate and check them all. Instead the software first searches for the 10 descriptors with the best predictive power for the target variable when used on their own. In a second step 10 additional variables are searched that – in combination with the 10 descriptors selected first – yield the best prediction based on two combined decriptors. This procedure is repeated until the "best" 5-descriptor model is found. If each of the 5 descriptors that ideally describe the data set only accounts for 20% of the variability then they will never be chosen by the algorithm because in the first step for example there are typically always several descriptors among the 1000 that cover more than 30% of the statistical variability of the data set. Thus the five ideal descriptors will never be identified by the algorithm each single one of them on its own performs inferior then others that are available. In praxis we have found this situation to happen regularly for many partition data sets: while the 5 descriptors used for building pp-LFER models are able to describe more than 90% of the variability the QSAR software offers models with 5 other descriptors that describe less than 70% of the variability.

All models discussed so far (fragment models, sp-LFER, pp-LFER and QSAR) treat the considered partition system as a black box; i.e. the partition properties of the system are accounted for by a calibration with experimental data. This has the advantage that any partition phase - no matter how complex it may be on a molecular level (e.g. humic matter) - can be treated with these models. The disadvantage is that a reliable calibration is needed for every new system that one is interested in.
The following two models (**SPARC** and **COSMOtherm**) that predict partitioning based on the molecular structure of the solute (just like the fragment method) are different in that they also require the molecular structure of the partition phases instead of a calibration. Obviously, this has the advantage that the user does not have to bother about finding reliable calibration data in the literature or perform measurements him/herself. Hence, these models can be used right away for partitioning between phases such as air, water, organic solvents and their mixtures, fuel, crude oil, well defined polymers. However these models are not readily applicable to humic matter, biological tissues, aerosols or other complex phases. Recently we have discovered, though, that it is possible to identify surrogate molecular structures for humic matter and the organic fraction of the aerosols. It seems likely that the same may be possible for biological tissues and other phases of environmental interest. This together with a number of other advantages makes these models highly attractive for environmental chemistry purposes.

SPARC is available free of charge via the internet http://ibmlc2.chem.uga.edu/sparc/. SPARC is based on the cavity concept and explicitly accounts for the intermolecular interactions between all molecules (solutes and phase molecules) that are involved in the partition process. SPARC depends on a calibration with experimental data for the intermolecular interactions but since it uses a very general and fundamental concept calibration only had to be done once by the software developers and could be based on a huge set of available partition data. Thus we expect SPARC to be a very robust model with a wide application range. As a special feature SPARC does provide information about possible tautomeric forms of a molecule which can be very relevant for partitioning.

COSMOtherm is a commercial software based on quantum-chemical calculations. Due to its very fundamental nature it can be expected to be the most robust method i.e. the one with the widest range of applicability. The fundamental nature of this approach allows predictions (without additional theoretical effort) that no other model can provide:

- adsorption to surfaces
- influence of various conformers of a molecule and stereochemistry
- concentration dependence of partitioning
- salting out effect
- partitioning into ionic liquids

Both, SPARC and COSMOtherm also allow to predict temperature dependence of partitioning and the pK_a value of organic chemicals, features that are very relevant for environmental partitioning.

A preliminary evaluation suggests that SPARC and COSMOtherm can be expected to outperform fragment models for all partition processes other than K_{ow} especially for complex, multifunctional molecules:

Arp, H.P., C. Niederer, and K.-U. Goss, 2006 Environ. Sci. Technol., **40**: 7298-7304. Journal link Download pdf
Wittekindt, C. and K.-U. Goss, 2009, Chemosphere, **76**: 460-464. Journal link Download pdf
Goss, K.-U., H.P.H. Arp, G. Bronner, C. Niederer, 2009 Environ. Tox. Chem., **28**: 52-60. Journal link Download pdf

As a rough guidance one can expect SPARC and COSMOtherm to predict partition coefficients with a root mean square error (rmse) of 0.6 to 0.8 log units.

Download this page as a pdf